UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

MACHINE LEARNING ALGORITHMS AND APPLICATIONS IN INVESTMENT ANALYSIS

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

ZHEN ZHANG
Norman, Oklahoma
2016

MACHINE LEARNING ALGORITHMS AND APPLICATIONS IN INVESTMENT ANALYSIS

A DISSERTATION APPROVED FOR THE
SCHOOL OF INDUSTRIAL AND SYSTEMS ENGINEERING

BY

_____
Dr. Theodore B. Trafalis, Chair

_____
Dr. Kash A. Barker

_____
Dr. Mark Shafer

_____
Dr. Charles D. Nicholson

_____
Dr. Maher Maalouf

To my future

ACKNOWLEDGEMENTS

I sincerely thank my advisor, Dr. Theodore B. Trafalis. Without his advice, patience, kindness and unconditional support, I definitely could not have progressed this far. His trust and encouragement have been a crucial part of the completion of my dissertation and his involvement in my educational experience will continue to motivate me to further my career in the investment analysis field. I learned a lot from him, not just how to conduct rigorous research, but also how to cultivate young people to learn and to maintain a positive attitude towards difficulties in life.

I am also very thankful to Dr. Mark Shafer, Dr. Kash A. Barker, Dr. Charles D. Nicholson, and Dr. Maher Maalouf who served as my Ph.D. advisory committee members. I appreciate their time and effort in helping me to refine this dissertation. Their experience and advice helped me to build a solid foundation in data science and analytics. Moreover, I would like to show my appreciation to other ISE faculty, staff members and students for their help during my time in the Carson Engineer Center.

I cannot say enough “thanks” to Dr. Randa L. Shehab who provided me with generous financial assistance and a fantastic internship opportunity. I appreciate Dr. Suleyman Karabuk for his instructions inside and outside the classroom and his advice on both of my work and life. Furthermore, I would like to give thanks to all my friends in China, United States, Mexico, Norway and Canada for their friendships. Especially to the closest ones: Heng Fan, Eric Russell, Yuqing Chang, Samantha Moguel, Allison Krause, Sara Walizer, Chase Stevens, Morgan Furgeson, Katie Bruggeling, Mingtin Xiao and Yuan Ye, who are truly the shining lights in a dark tunnel for me. I also want

to thank my co-workers at Mercy Hospital, who dedicate their time to patients and have motivated me to be a benevolent person in the field of healthcare.

My greatest appreciation extends to my parents GuoHua Zhang and JianPing Wang, who provide me with lots of freedom, trust, and help. Every day, I feel blessed that they are healthy and happy, so they can enjoy and enrich their lives. They have not only provided me the opportunity to experience the world but have also guided me towards being a person who is honest, has the will of integrity and a hardworking spirit. I also must give my greatest appreciation to my grandpa who passed away years ago. He was the principal of a local community college in my hometown. He always encouraged me to read voraciously and think critically. He named me “Zhen”, which means “Honesty” in Chinese.

In short, this dissertation would not have been possible for me to finish without my determination and all the loving, kind and supportive people in my life. I am excited to present it and cannot wait to take new challenges in my life!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

We can simplify investment analysis as filtering out speculative stocks, bonds, derivatives and other financial products. This area is very challenging yet extremely critical since individual investors', large institutions' and the public's understanding of investment and investing behaviors determine the long-term stability of the U.S. and the interconnected global economy. This dissertation focuses on how to utilize Machine Learning (ML) techniques to facilitate investment analysis, what are the challenges in practice, and how to bridge the gap by choosing appropriate algorithms and modify them to mitigate the risk of significant financial losses. A general work path and investigated topics are as follows:

1. Demonstrate a comprehensive understanding of the U.S. economy, its key industries, and the traditional investment analysis principles.
2. Develop a thorough knowledge of several widely applied ML algorithms and gain hands-on experience through applications. We simulate these algorithms and its derivations in different scenarios and test it with correspondent accuracy and efficiency measures.
3. Present cases to explain why and how irregularities or uncertainties affect algorithms performance, propose and implement solutions to improve classification results.

This dissertation uses the Mergent or Securities Exchange Commission (SEC) published datasets. By connecting algorithms with real word datasets, this dissertation successfully demonstrates how crucial it is to understand your data and how ML algorithms can facilitate similar decision-making processes.

CHAPTER 1: INTRODUCTION

1.1 Introduction

One of the most successful ML applications in the field of investment is credit rating: Bill Fair and Earl Isaac founded FICO that uses past performance, current debt, and the length of credit records to measure a borrower's credibility and give appropriate lending guidance (Mena, 2011). Real estate agencies use ML to estimate housing prices based on raw material expenses, labor costs, housing inventory, and historical consumer demand data (Huang, Zhu, & Siew, 2004). Most recently, ML algorithms help the GE consumer finance sector and Japanese researchers to design and improve lending strategies (Pyle & Jose, 2015).

In this dissertation, an interdisciplinary work in Economics, Finance and Industrial System Engineering, we apply and modify several ML algorithms to analyze public companies and successfully improve the prediction results compare to several existing techniques. The way we look at individual companies is very similar to how economists look at the economy through various established economic indicators. Although any indicator is far from perfect, when we piece those together and evaluate them on a long-term base they can provide a decent picture of current conditions and prospects. This understanding is not just crucial for government intervention and policy design, but also reminds us that any investment analysis cannot be isolated from knowing the general economy and industries. Therefore, we take a review of the U.S. economy from 2002 to 2011 and assemble several key economic indicators together first. Then, we present a brief discussion of several key industries and the principles of

evaluating individual companies. At the end of this chapter, we summarize some main findings and discuss research limitations.

1.2 Investment Analysis

Benjamin Graham defines investment as a long-term financial planning based on the concepts of "*margin of safety*" and "*satisfied returns*" (Graham B. , 2009, p. 512). He believed that this strategy would save investors from brokerage fees, certain taxations and mental anxiety caused by market errors. He also pointed out a very important principle: "*Quantitative data are useful only to the extent that they are supported by a qualitative survey of the enterprise*" (Graham & Dodd, 1934, p. 474). Ideally, a solid investment analysis consists of four elements: the global economy, the macro economy, the industry and the company (Graham , 2009). Here, we focus heavily on the latter three due to the research constraints.

The macro economy can be defined as weak, strong or recession. The perception of the economy is a result of a combination of various economic indices such as gross domestic product (GDP), consumer price index (CPI), consumer confidence index (CCI), interest rates, unemployment rate, retail growth rate, and many other indices. The industry can be classified as cyclical versus recession-proof. Food, energy, education and healthcare are considered as recession-proof since they are essential to society. Cyclical industries move along market tides, such as high-tech, oil or fashion industry. However, even in the same industry, companies' performance can vary significantly.

1.3 The Analysis of the Economy: Why Stability is the Key?

It is a well-known fact that the U.S. economy was fueled by an unprecedented real estate boom and consumer debts from 2003 to 2006. Normally, housing booms and busts are regional. They only occur when local lenders relax their lending standards and let unqualified borrowers get access to the credit pool. Even when those borrowers default later, only regional financial institutions end up failing. However, during 2003 to 2007, the U.S. home mortgages were diced, bundled, securitized and sold to the global investors that include big hedge funds, pension funds, foreign banks, and many other key financial organizations. Around the summer of 2007, there were early signs of subprime lending failure. But it was not until the mid of 2008 that massive home foreclosures and financial instability showed up after a series of bad news: Countrywide Financial delinquency, French BNP Paribas bank run, Washington Mutual in limbo, IndyMac bank ran after $1.3 billion deposits withdrawn and Lehman Brothers bankruptcy (Financial Crisis Inquiry Commission, 2011). In the 1990s, the entire U.S. subprime mortgage market was estimated around $35 billion. In just a decade, it grew up to $665 billion with a shocking 1700% growth rate (Schloemer, Li, Ernst, & Keest, 2006). The New Century made about $60 billion subprime loans in 2006, increasing its lending volume about ten times in just five years. In 2007, American Home Mortgage Investment made about $34 billion mortgages in just the first six months (Ashcraft & Schuermann, 2008). The subprime lending based Mortgage Backed Securities (MBSs) model was based on some faulty assumptions (Diamond & Rajan, 2009):

Assumption 1: Homes are always in demand and the home values will always increase.

Assumption 2: Interest rates would be low for an extended period.

Assumption 3: Risks can be avoided through diversification.

Assumption 4: Rating agencies are impartial and always do their due diligence (Calomiris, 2009).

The ensuing crisis gave us a painful lesson that this model failed to simulate the worst scenario, leading to a catastrophic system failure. Moreover, the series of bank failures caused widespread fear that demanded an immeasurable risk premium from all kinds of lending behaviors. For instance, MBS's default quickly contaminated student loans, credit cards, and corporate credit markets; facing rising short-term interest rate, banks raised minimal credit card monthly payments and APR rates to compensate losses; and the confidence among inter-bank lending deteriorated rapidly, triggering a worldwide lending freeze (Acharya, Afonso, & Kovner, 2016). At the peak of the crisis, funds and credits were no longer as available to mass consumers and businesses, beating down the economy even further (Brunnermeier, 2009).

Two conclusions can be drawn here: first, a highly interconnected and intricate global financial system was the main reason why the U.S. housing crisis triggered a global economic crisis. Second, stability is the key to social development because dubious growths often bring in long-term detrimental impacts. **Figure 1.1** is created to show the interconnections among different sectors, the vicious cycle of losing credit confidence, and the terrifying consequence of speculative lending behaviors to the economy and society.

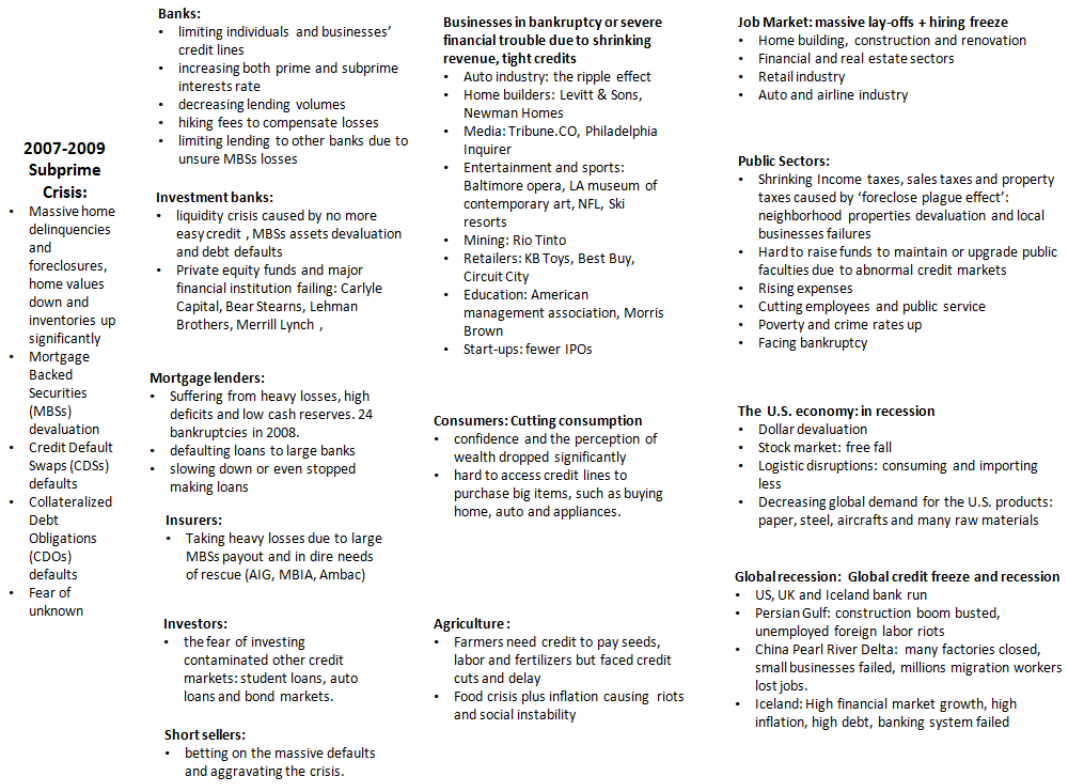


Figure 1.1 Why stability is the key? The ripple effects of the 2007-2009 housing crisis (Source: The Wall Street Journal, New York Times, and National Public Radio).

1.3.1 Several Key Economic Indicators

Many economic indicators can be used to evaluate an economy. Conventional and established indices included here are Gross Domestic Product (GDP) to measure economic growth, Unemployment Rate to track the movement of labor market, S&P Retail Select Industry Index (SPSIRE) and Consumer Confidence Index (CCI) to measure consumer confidence, Consumer Price Index to reflect inflation, ISM Manufacturing Production Index (PMI) to display economic growth, and S&P/Case-Shiller Home Price Indices to evaluate the housing market (**Table 1.1)**.

Table 1.1: Several key economy indicators (Source: The World Bank, United States Department of Labor, Federal Reserve Bank, Economic Research, Case-Shiller Index (2003-2015))

	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
GDP %	2.8	3.8	3.8	2.7	1.8	-0.3	-2.8	2.5	1.6	2.3	2.2	2.4
Unemp %	6	5.5	5.1	4.6	4.6	5.8	9.3	9.6	8.9	8.1	7.4	6.2
CPI %	2.3	2.7	3.4	3.2	2.9	3.8	-0.4	1.6	3.2	2.1	1.5	1.6
PMI	51.7	59.1	54.5	53.2	51.2	45.5	46.4	57.3	55.2	55.7	51.7	55.6
SPSIRE	-	-	-	1946	2154	868	1549	2228	2676	3121	4426	4775
S&P/Case-Shiller	179.4	197.5	216.8	215.1	195.5	171.9	160.9	152.0	141.8	148.5	162.0	168.1

A closer look at the monthly job statistics from 2004 to 2009 would give us a better understanding of how job markets rapidly deteriorated after the rising short-term interest (**Figure 1.2**). Although the estimated number of job loss or gain data and unemployment rate are only based on a poll of 60,000 selected household reports and cross-industry surveys, from a long-term perspective, they have been proven to consistently reflect the general labor market condition (United States Department of Labor, 2009).

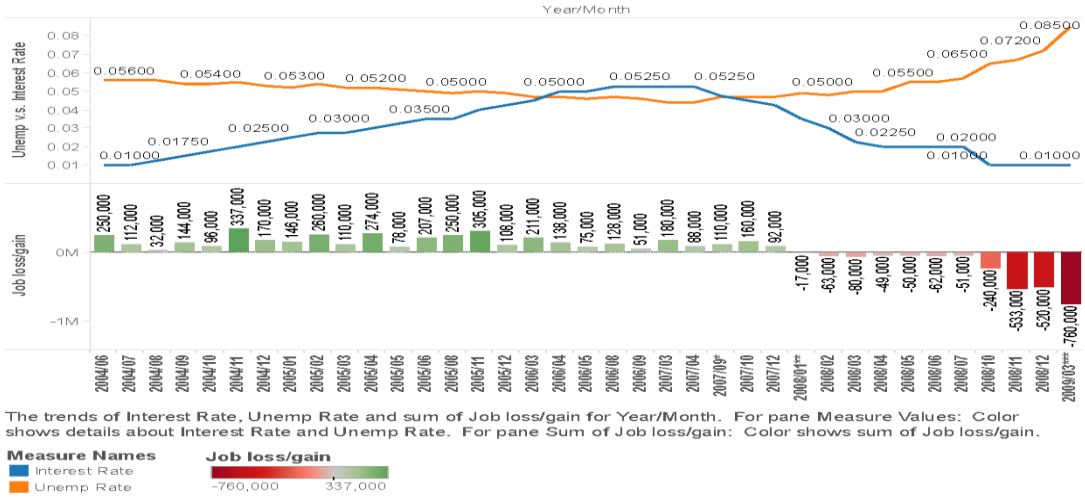


Figure 1.2 The 2004-2009 U.S. labor market and the short-term interest rate (Source: United States Department of Labor, U.S. Department of the Treasury (2004-2009)).

Many interesting yet not quite established indices can be used to test the strength of the economy (**Table 1.2**). For example, trucking contributes to 70% of U.S. freight

transportation. Between 2007 and 2009, thousands of small trucking companies went out of business, and there was minimal or no growth in the total number of trucks (American Trucking Associations, 2008). As a leader in mailing and package delivery, USPS's annual volume, costs, and revenue all went down during that period (USPS, 2014). The number of Initial Public Offerings (IPOs) reflects investors' confidence in the economy. Researchers found out that there were sharp IPO reductions in both value and volumes around 2008-2009 (Ritter, 2014). The amount of the charity donation measures the public confidence and disposable income level decreased significantly during the same time (National Philanthropic Trust, 2015). In short, after witnessing a severe economic downturn during 2007 to 2009, we treat companies' financial statistics during this period with discretion.

Table 1.2 Other economic indicators (Source: United States Department of Transportation, United States Postal Service, and National Philanthropic Trust).

U.S.	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Trucks (in Millions)	95	100	104	108	110	110	110	110	118	133	133	-
USPS Total Shipping (In Billion)	-	-	-	-	-	3.3	3.1	3.1	3.3	3.5	3.7	4
Number of IPOs	63	173	159	157	159	21	41	91	81	92	157	207
Charitable Giving (in Millions)	225	230	260	300	300	325	300	250	270	325	330	358

1.4 The Understanding of Several Key Industries

A basic understanding of different industries is indispensable to investment analysis. This section first illustrates how government plays a central role in industrial development. Then it elaborates more on the energy, auto, airline, financial, healthcare, housing and agriculture sectors.

1.4.1 The Central, Local Government and the Industry

The government plays a crucial role in a country's economic development, so its stability should be the first thing to examine when we think about investing. For instance, in 2010, India had a micro-financing crisis. Local lenders used to extend small credits to residents who had limited income and borrowing ability. However, without proper lending standards and a rigid payback system, it went out of control. One state Andhra Pradesh even called on borrowers to stop repaying. This political intervention put the whole micro lending industry in jeopardy, since millions of borrowers responded to it and the repayment rate dropped from 98% to 10% (Tadele & Rao, 2014). Fortunately, the U.S. has a relatively strong and stable political and social system so it can maneuver the economy more effectively with budget control, interest rates adjustment, credit supplies, inflation control, and industry regulations. It is worth pointing out that among all these strategies, budget control has two sides: revenue collecting and budget planning which help the country to achieve both long-term (such as job creation and living standard improvement) and short-term (such as stabilizing the system) goals. Government revenue largely depends on its tax base and tax rates. Government spending typically covers short-term incentives and stimulus programs for emergency needs, long-term defense, pension, healthcare, education, infrastructure, and other social expenditures.

Extensive literature reviews are presented here to show how government actions shape the general economy and affect related industry. For instance, the U.S. government aims to achieve greater oil independence and create a cleaner environment by promoting electric vehicles (Wirasingha, Schofield, & Emadi, 2008), which has led

to huge interest in Lithium-ion batteries research and development and a high demand for lithium (Scrosati & Garche, 2010). The government also uses tax rebates for energy-efficient products to promote conservation (Gillingham, Newell, & Palmer, 2006), compelling Federal Communications Commission (FCC) to speed up the broadband internet and lower the costs for consumers (Cambini & Jiang, 2009), and establishing the Troubled Assets Relief Program (TARP) to avoid devastating financial system collapse (Nguyen & Enomoto, 2009). Regulation is a powerful tool in guiding industry toward sustainable economic growth. For instance, the Environmental Protection Agency (EPA) regulates greenhouse gases emission and credits trading activities (Ellerman & Harrison Jr, 2003), the Food and Drug Administration (FDA) examines take-in products to ensure product quality and safety (Hilts, 2003), and the most recent financial regulation overhaul constructed a collaborative international insurance negotiation agency, a better consumer protection agency and a larger financial institutions liquidation panel to prevent and deal with large-scale financial crisis (Davies & Green, 2013). Furthermore, each state and local government has a certain degree of freedom to set up regional laws, attract investment and stimulate growth. More and more states realize that only a well-diversified economy can provide a stable tax base to weather economic storms. Like Charlotte, NC, once a rising new financial center, lost most jobs in this area during the crisis and now engages in healthcare development (Florida, 2009). In reality, if examining companies in depth, we need to consider government policies. However, in this dissertation, we focus more on companies' long-term performance under the same economic settings without adding political influences to complicate the case.

1.4.2 The Energy Industry

To understand business in the U.S. energy sector, it is the best to know each key energy source's global demand, supply, inventory, related safety issues, the federal and state regulations, weather, speculative activities, and other competitive energy sources' statutes. Here we focus on the petroleum industry that affects every aspect of society and notice that operational safety is extremely critical but often underestimated until some catastrophic events surface. For instance, in the 2010 Gulf of Mexico oil spill case, British Petroleum (BP) caused severe environmental damages, went through a lengthy and expensive litigation process that incurred billions in losses, and pushed its market value down to half of its peak. There were also enormous negative impacts on the whole oil and gas industry, since this incidence triggered a six-month offshore drilling moratorium, pushing all related companies operational costs up and impeding the progress of the coastal economy, such as local tourism, fishing, shipping, and the industry related downstream businesses (Mason, 2010).

Specifically, during the 2003 to 2008 oil boom, the public suffered directly from higher heating bills, transportation costs and fluctuating investment returns (Kilian & Park, 2009). For manufacturers, inflation was imminent, since daily production, operation, transportation and raw manufacture materials such as plastics, parts, and tools are petro chemistry based. Therefore, the only way to sustain operation and development was to pass down the cost to mass consumers. For the airline industry, the revenue increase was far behind the increase in fuel costs during the price-hike period. Therefore, airlines had to reduce workforce, change schedules and rely more on fees

(Carter, Rogers, & Simkins, 2006). The agriculture business is also energy concentrated. The costs of fertilizers, farming process and transportation all highly depend on the oil price. Corn prices were up significantly in 2007 due to higher transportation costs, contributing to the world hunger and food crises (Headey & Fan, 2008). Last but not the least, persistently high oil prices keep people away from leisure and other life enriching activities (Becken & Lennox, 2012).

On the other hand, there are many businesses benefitted from the 2003-2008 oil boom. For instance, oil and gas companies had more funds to do new exploration, generating new jobs in drilling and downstream businesses. Natural resource-abundant states such as Texas, Colorado, Utah, and Wyoming all experienced an extended period of booming economy, low unemployment rates, and budget surpluses (Weber, 2012). Meanwhile, companies focused on alternative energy development and energy efficient products were thriving too, such as scooters, motorcycles, and bikes manufacturers. The public transportation system and the railway sector experienced significant revenue growth (Gilbert & Perl, 2013). Moreover, the booming industry created a huge demand for both field laborers and highly skilled workers, benefiting related school enrollment, training programs and research activities (Timilsina, Mevel, & Shrestha, 2011).

Furthermore, research on the crude oil price (the primary revenue source of petrol-businesses) is based on a simple economic principle: the demand and supply curve (Kilian L. , 2006). The demand is affected by the global economic growth (Asif & Muneer, 2007), the seasonal demand (Alvarez-Ramirez, Alvarez, & Rodriguez, 2008), and a rising speculative force triggers a large amount of capital in and out of commodity markets (Kilian & Murphy, 2014). The supply is affected by oil inventory (Ye, Zyren,

& Shore, 2002),safety related incidences (Carson, Mitchell, Hanemann, Kopp, Presser, & Ruud, 2003), natural disasters such as hurricanes and storms (Kilian L. , 2006), government intervention (Tyner, 2008), refinery capacity constraints (Kaufmann, Dees, A., & Mann, 2008), and geopolitical conflicts (Omofonmwan & Odia, 2009). However, a comprehensive oil price model is hard to be constructed and related business forecasts are less consistent.

1.4.3 The Financial Sector

The financial system is based on trust and confidence. The stability of the society, the fairness of the judicial system, and the supply of money jointly determine the strength and the prosperity of its financial system. Many significant financial events happened in the last two decades and greatly impacted companies in this industry. For instance, early 2000s dot-com bubble (Ofek & Richardson, 2003), information asymmetry based MBSs failure (Schwarcz, 2008), new derivative credit default swap crisis (Weistroffer, Speyer, S., & Mayer, 2009), Madoff Ponzi scheme (Smith F. , 2010), Lehman Repo 105 financial manipulation (Jeffers, 2010), 2010 system errors triggered DJIA freefall (Pengelly, 2010), and the large flow of global human and investment capital (Milesi-Ferretti & Tille, 2011). In a nutshell, almost every financial turmoil originates from deregulation and incentives based imprudent practices which are the exact opposite of stable and healthy growth.

1.4.4 The Housing Sector

The housing market is determined by housing demand, mortgage rates, and credit

accessibility. This sector directly affects construction, financing, and mass consumer markets. Historically, regional and nationwide housing market collapses are linked to housing oversupply and high unemployment rate. According to Robert Shiller, the average annual house appreciation should be in alignment with the costs of construction and land, the growth rate of population and housing formation. Moreover, in his opinion, irrational lending is preventable if the credit issuing agencies obey safety and stability lending standards (Shiller, 2015).

1.4.5 The Auto Industry

Between 2005 to 2009, the U.S. Auto industry suffered multiple hits such as high oil prices, overseas competition, and the credit crisis. Typically, when gasoline prices go up, the demand for hybrid and fuel efficient cars goes up, while the sales of trucks, SUVs and all the other gas-guzzling vehicles go down (Klier & Linn, 2010). When oil prices rose above $60 per barrel and gas prices approached to $3 per gallon during the summer of 2005, General Motor and Ford's middle to large size vehicles sales plummeted, which were the main profit sources for these two leading U.S. auto companies. Many other problems followed suits, such as huge fixed expenses of pensions, rising healthcare costs for employees, and slow to build new model assembly lines (McManus, 2005). Moreover, the ripple effect of the auto industry was far reaching. For example, it was estimated that the big three would cause 3 million job losses if they received no government assistance since one auto job is estimated to be tied with 5 to 6 jobs in a local economy (Maloney, 2009). For instance, when one U.S. major auto part supplier Delphi announced bankruptcy, the surrounding neighborhood

businesses and families took the biggest hit. Residents were overwhelmed by owning inefficient big vehicles with high fuel prices, rising adjustable mortgage payments, increasing credit card interests and unemployment (Examining The Delphi Bankruptcy's Impact on Workers and Retirees, 2009). Therefore, a forecast of any auto related business must consider the consumer, financial and the world oil markets.

1.4.6 The Airline Industry

Airlines convey leisure, business travelers, and flight cargos. Airline businesses are affected by the general economic condition, labor issues, competitions, the weather and many disruptive events. Even in the same field, airlines tend to perform heterogeneously and can be categorized into two classes: traditional versus new airlines. For traditional airlines, they face same issues of the rising costs of labor and hubs (Gittell, Nordenflycht, & Kochan, 2004), fluctuating fuel prices (Abdelghany, Abdelghany, & Raina, 2005), unpredictable weather, various disruptions (Janić, 2005) and the cutthroat competitions from low-cost airlines. New ultra-low-cost airlines often do not have hefty pension obligation like traditional ones and often utilize hedging to stabilize fuel costs (Morrell & Swan, 2006). However, unexpected or uncontrollable events level up both classes' costs and operational risks, such as the global recession, severe winter storms, the 2010 Iceland volcano eruption (Air travel disruption after the 2010 Eyjafjallajökull eruption, 2015), and the ten-year litigation of Concorde crash (Air France Flight 4590). Our research finding is consistent with Graham's conclusion that airlines are rarely investment choices due to its high operational risks.

1.4.7 The Agriculture Business

For businesses in agriculture, several things deserve attention. For instance, agriculture derivatives may absorb risks but intensify speculation (Musshoff, Odening, & Xu, 2011), climate changes and random weather variations have different impacts on products and operation (Seo, 2013), agriculture-related patent issues often are hidden risks (Smith T. , 2009), and natural disasters can disrupt the whole supply chain system, causing widespread fear and instability, such as berry borer invasion (Jaramillo, Borgemeister, & Baker, 2006).

1.4.8 The Healthcare Sector

Healthcare is a prominent field that promises great job opportunities and growth, mainly due to the large aging baby boomers and rising health care costs. Companies in this booming industry encounter many serious operational issues, such as failing to meet drug safety standards (Hazell & Shakir, 2006) or engaging in illegal behaviors which damage its reputation and finance (Sparrow, 2000). The most striking case would be Turing Pharmaceuticals, which bought Daraprim and increased its price from $13.5 to $750 per pill, a more than 5000% price hike overnight (Lorenzetti, 2015). Moreover, the record case would be Pfizer, the world's largest drug company by sales, made the largest $2 billion settlement due to illegal marketing of Bextra (Harris, 2009).

In conclusion, the above brief review shows that different industries are highly interconnected and actively interact with each other, and detrimental events are not uncommon in any sector. Therefore, facing enormous uncertainness, only companies

that possess certain intrinsic stability can survive various disruptions and be considered as investment in the long-term.

1.5 The Analysis of Individual Companies

In search of a company's intrinsic stability, we look into its past records. The qualitative analysis of a company usually begins with a quantitative financial statement analysis, which consists of three main components and key ratios such as return on equity, leverage, operational efficiency and many others. We present a simplified version here:

Table 1.3 A simplified version of the Income Statement

OPER costs	Fees
✓ **Labor** ✓ **Materials** ✓ **Annual depreciation and amortization** ✓ **Unities** ✓ **Rent**	✓ **Marketing** ✓ **Management** ✓ **Accounting** ✓ **Research and development** ✓ **Exporting/importing** ✓ **Legal** ✓ **Technology**

Table 1.4 A simplified version of the Balance Sheet

Assets		Debt	
Current Assets	**Non-Current Assets**	**Current Debt**	**Non-Current Debt**
✓ **Cash** ✓ **Marketable securities** ✓ **Receivables** ✓ **Inventories (30%)** ✓ **Paid in advance**	✓ **Intangible assets** ✓ **Other financial assets** ✓ **Fixed assets (50%)**	✓ **Short-term debt** ✓ **Long-term due debt** ✓ **Payables**	✓ **Long-term debt** ✓ **Long-term loans**

Table 1.5 A simplified version of the Cash Flow Statement

Net OPER cash flow	Net investment cash flow	Net Financing cash flow
Revenue – Costs of labors & materials - Tax	**Investment income + Selling fixed assets – Capital expenses – long-term assets investment**	**Stocks + bonds + long-term and short-term debts – Stocks buyback– long-term short-term debt deduction – Interests - Dividends**

Based on extensive case studies and literature review, we summarize four principles in selecting investment grade companies (Graham, 2009):

Principle 1: The source of income should be diversified in both products and markets.

Principle 2: The financial structure should be strong.

Principle 3: The rate of return should be stable and satisfying.

Principle 4: The background and history should be solid.

Caterpillar is a great example to support **Principle 1**. During 2007 to 2009, although its domestic market shrunk significantly due to the slowdown of the U.S. commercial and home construction, the demand for its services and products from other sectors, such as mining, oil, gas exploration, power generation and fast growing oversea markets like China, Russia, Middle East and South American, pumped up its sales. Furthermore, multinational corporations like Caterpillar do not only benefit from its diversified revenue sources but also get hurt less by currency fluctuation during economic turbulences (The Associated Press, 2008). An opposite example would be Sycamore Network, which earned $11 per share at its peak. However, since Williams Communications was its only major income source, so after Williams Communications went bankrupt, Sycamore lost most of its revenue and market value (History of Sycamore Networks, Inc., 2001).

For **Principle 2**, we investigated cases such as Carlyle Capital, Cerberus, Bear Stearns and Lehman Brothers in searching for an answer that why these big private equity funds and investment banks collapsed in just a few days between 2008 to 2009. It turns out that they all shared three traits: overly concentrating on one type of risky

investment, relying heavily on borrowing and depending largely on investors' confidence (Case Study: The Collapse of Lehman Brothers, 2008).

Principle 3 indicates that other than the risky business nature, the speculative force behind prices poses an even greater danger. In 2006, Crocs had an IPO around $13/share, in just one year the share price climbed to $70/share. Its annual stock appreciation rate was 400% but without any decent rate of increase in revenue or sales. Two years later, Crocs' stock price went down to only $1/share, losing 98% of its peak value (CROX: summary for Crocs, Inc (2009)).

To validate the **Principle 4**, other than the well-known Dot.com bubble during which hundreds of new tech companies mushroomed and evaporated, we dug into Fisker, an electronic green tech car company, which perfectly illustrates why new companies rarely qualify for investment. Fisker was founded in 2009 with a rising demand for fuel-efficient cars and great assistance from the government to promote renewable energy industry. Although having many talented people, bright ideas and real products, it encountered numerous setbacks in three years. For instance, an uncertain consumer market, major assets loss caused by natural disasters, technology deficiency and lengthy litigation battles with both insurance company and competitors. Fisker ceased operation and filed for bankruptcy in 2012 (FiskerAutomotive's road to ruin: How a "Billion-Dollar Startup Became a Billion-Dollar Disaster", 2013).

In short, a distinct line should be drawn between investment and speculation. Non-experts should focus on investment grade companies for inherent stability, safety, and tax benefits. Moreover, an ideal selection proposed by Graham Benjamin is as follows (Graham , 2009):

1. A diversified portfolio consists of many stocks.
2. Companies should have a large but conservative financial background measured by its market capitalization and net assets. And the key for small businesses to survive in economic downturns is to have a flexible operational budget.
3. Companies should have been traded publicly for least ten years with a consistent dividend payment.
4. P/E ratio should be less than 25 times which is more than 4% as E/P ratio.
5. Be cautious of companies that have a high leverage or high interests on debt.

More strict quantitative criteria are given by Graham as follows (Graham, 2009, pp. 348-354):

1. Size: an industry company should generate no less than $100 million sales, and the total assets should be no less than $50 million if it is a public utility. Evidence during the 2008 crisis is many small companies could not weather the storm but failed due to the triple hits from the credit crunch, the shrinking revenue, and depreciated assets.
2. Financial condition: a company's current assets should be at least two times of its current liabilities.
3. Stability: a company should have at least 8 to 10-year earnings record without any deficit.
4. Dividend: a company should have a non-interrupt dividend payment for the past 20 years.
5. Growth: Earning per Share (EPS) growth in 10 years should be at least 30%.

6. P/E ratio or E/P the multiplier, a current price should be no larger than 15 times of its most recent 3-year average earnings.
7. P/BV ratio: price to book value should be less than 1.5 times.

As we can see some of the criteria are either outdated or arbitrary in numbers, so in Chapter 4 we expect a two stage Neural Network (NN) with feature selection to identify more generalized criteria and key attributes based on Graham's intrinsic stability concept.

1.6 Summary and Limitations

This dissertation follows the above guidelines and yields some results consistent with past findings:

- By sector: public utility companies do have great advantages in borrowing money, raising capital and increasing charges. However, they are subject to close government monitoring and regulations, providing small but stable return.
- By traits: good companies do occupy at least one of three competitive advantages. For instance, monopoly or near monopoly position in either supply or demand or both, intangible non-replica assets, none or only a few alternatives such as power, gas, telephone, water and electricity companies.
- By price: According to Benjamin Graham (2009), the price of a stock does not always align with its value but is jointly determined by the issuer's financial status, the general economy and the speculative interests. Therefore, forecasting stock prices with ML algorithms is theoretically sound but virtually intangible due to asymmetric, unquantifiable, and hidden information.

It is worth mentioning that both quantitative and qualitative analysis can be flawed due to data errors, uncertainties, market irrationality, and other limitations. For instance:

- Earnings manipulation: By changing the way of calculating depreciation and amortization charges, allocating special charges or varying actual labor and material costs, a company can easily inflate or deflate its earnings. In this dissertation, with at least ten years verified annual record for each company, the possibility of accounting manipulation is greatly minimized.
- Earnings stability: Due to the time and data limitation, we could not dissect, verify the sources of incomes or calculate the Margin of Profit (MOP) to decide whether a company is a leader in its industry.
- Debt versus earnings: we focus on the EPS growth and price advance over an extended period, without further investigating the relationship between debt increase and the earnings growth. Typically, the faster a company's debt grows, its earnings are less credible due to the potential interest rate hike, over-expansion, and unpredictable economic setbacks.
- Competitive advantages: we did not include the three most common competitive advantages into the proposed classification consideration.
- Management: we did not check Proxy or 4-K to compare their promises versus reality and to calibrate its capability.
- Inside transition: It is impossible to detect inside fundamental changes and foresee its failure or progress by just looking at its previous financial statements.

CHAPTER 2: UNIVARIATE LINEAR REGRESSION

2.1 Introduction

Machine Learning (ML) applications began when Arthur Samuel (1959) used computers to conduct repeated tests and practices to improve gaming strategies. Then computers demonstrated their superior ability to follow instructions and accumulated experience in relatively short amount of time to outperform human players. Later, Tom Mitchell (1998) defines ML as a way to improve a task performance based on past experiments and certain measures. ML can be categorized into two classes by its nature: Supervised (Kotsiantis, Zaharakis, & Pintelas, 2007) versus unsupervised learning (Alpaydin, 2014). ML algorithms can also be categorized into two classes by its output: regression with the continuous valued output (Witten & Frank, 2005) versus classification with the discontinuous valued output (Michie, Spiegelhalter, J., & Taylor, 1994). Building a model, applying an algorithm and evaluating performance are an integrated process in ML. Typical steps are as follows:

1. Data pre-processing and visualization.
2. Feature normalization: When certain features' magnitude or measurement significantly differ from others, feature scaling becomes necessary, which facilitates the convergence.
3. Cost function formulation: the goal of a cost function is to minimize the sum of estimated prediction errors. Two widely used algorithms are Batch Gradient Descent (GD) (Wilson & Martinez, 2003) and Normal Equations (NE) in search of the parameters (Levenberg, 1994).

4. Learning pace: By observing changes along the number of iterations and chosen an appropriate learning rate, practitioners can find a set of parameters to converge quickly (Jacobs, 1988).
5. Prediction: performance evaluation.

In this dissertation, I focus on four classical algorithms' application and modification in investment analysis: Linear Regression (LR), Multiple Regression (MR), Logistic Regression (LogR) and Neural Network (NN).

2.2 Simple Linear Regression

Regression is a way to explore and identify the relationships between various inputs and outputs: Do these inputs significantly affect outputs? How exactly the changes of these inputs affect outputs? Can we make credible new sample analysis based on the regression model we proposed? LR has the longest history, which is the foundation of virtually all advanced regression analysis. Several assumptions are presumed in LR application: predictors are relatively dependable free from measurement errors and other possible misrepresentation; a linear relationship can be established; prediction errors are heteroscedastic and uncorrelated; and no multicollinearity among predictors is assumed (Yan, 2009). However, LR only proposes a potential link between a set of inputs and output, and further analysis is usually required to verify the strength and possibility of this relationship (Weisberg, 2005).

The earliest LR applications can be traced back to James D. Forbes (1857) who used a small dataset to draw an almost perfect line between temperature and pressure. Then Pearson and Lee (1903) collected more than 1000 pairs of mother and daughter's

height data to study the height inheritance problem, trying to find out whether an inter-generation relationship of height can be established. Nowadays, LR applications can be found in almost every aspect of life, such as biochemistry (Liu, et al., 2016), engineering (Toh, Yeoh, Teoh, & Chin, 2016), sociology (Thomas, Amburgey, & Ellis, 2016), medical (Porto, Cardoso, & Sacomori, 2016) and finance areas (Umar & Sun, 2016). Also there are abundant improvements that made LR more adaptive to different situations, such as smoothing techniques enhanced LR (Schimek, 2013), nonparametric LR formation (Faraway, 2016), and distribution based LR (Gramacy & Lee, 2012).

LR works the best when both input and output can be mapped perfectly along a straight line. However, the tendency of over-simplification becomes LR's biggest weakness. Anscombe (1973) presented four different datasets which have different scatter plot layouts, but all can be fitted by the same simple linear model and evaluation results. It shows that different datasets may share same statistics and an identical regression line, failing to grasp more complicated relationships. Another major drawback is that LR might disguise interconnections among a set of input variables embedded in a complex system; therefore we need to either reevaluate the inputs or conduct a commonality analysis (Ray-Mukherjee, Nimon, Mukherjee, Morris, Slotow, & Hamer, 2014)

In the following sections, we focus on LR applications on financial datasets. The main goal is to understand the strength and limitation of LR and propose solutions to facilitate the modeling process.

2.2.1 The Estimated Linear Relationship between Earnings and Prices

This section starts with an investigation of the relationship between the earnings and prices of 159 public companies. We start with the scatterplot visualization process as the first step to begin a regression analysis, and then we focus on outlier detection for model improvement.

Table 2.6 A list of attributes used in this chapter (Source: Mergent's Handbook of Common Stocks. 2001-2008)

Def 01-08	Number of deficits between 2001-2008
No Div 01-08	Number of 0 dividend payout between 2001-2008
E_Avg 01-08	Average earning from 2001-2008(USD)
E_Var 01-08	Earning variance from 2001-2008
P_Avg 01-08	Average price form from 2001-2008 (USD)
P_Var 01-08	Price variance from 2001-2008
Div% 01-08	Total dividend payout between 2001-2008
P08/E_Avg	Price earnings ratio in 2008

2.2.2 Outlier Detection

Outliers in LR are defined as extremely uncommon pairs of (X_{ij}, Y_i), where X_{ij} are inputs and Y_i are output (Aggarwal, 2015). Getting rid of outliers helps to improve the model fitting ability, but it might also cause problems like shrinking the sample size or missing valuable information. Plotting and statistical analysis can help to find outliers, but it rarely explains what exactly happened and how important those outliers could be (Weisberg, 2005).

Typically, there are two ways to detect outliers: mapping all points out to find patterns and distance-based outlier detection approaches. Distribution Deviation (DD) method filters out outliers as samples deviate from common standard distributions such as Gaussian and Poisson distribution (Leroy & Rousseeuw, 1987).The major drawback of this approach is that in reality we often encounter datasets with an unknown

distribution. Clustering algorithms focus more on classification instead of circling out outliers (Jain, Murty, & Flynn, 1999). The most widely used and visualized approach is proposed by Knorr, Ng and Tucakov (Knorr, Ng, & Tucakov, 2000), which allows users to look into the composition of a dataset and decide the percentage of outliers. Another popular K-Nearest Neighbor (KNN) method (Ramaswamy, Rastogi, & Shim, 2000) and its various derivations (Angiulli & Pizzuti, 2002) (Chen, Miao, & Zhang, 2010) all use the distance ranking method to expose outliers. KNN method displays a high accuracy by calculating the distance between all data points, but the drawback is that it suffers from the curse of dimensionality. Another unique local density measure relies more on the closest neighborhood points instead of the whole group (Breunig, Kriegel, Ng, & Sander, 2000).

Based on the KNN method, we develop and test the Relative Distance Measure (RDM) in searching for the point which has the maximum distance among all distances to the Center Point (CP). CP is defined as the intersection of the mean and the regression function by assuming all points are potential outliers. RDM is dynamic in detecting extreme outliers in a simple linear regression, requiring less computational power comparing to KNN and therefore might suffer less from the curse of dimensionality. RDM is based on a distance function $D(p_1, p_2)$, which is assumed to satisfy several conditions such as non-negativity, symmetric and triangular inequality. There are three commonly used distance measurements: Manhattan, Euclidean and Chebyshev Distance (Arora, Singh, & Kaur, 2015). We adopt the Euclidean Distance measure. Assume there are two points p_1, p_2 in Euclidean *m*-space, the distance can be calculated as below:

$$D(p_1, p_2) = (\sum_{i=1}^{m} (p_{1i} - p_{2i})^2)^{\frac{1}{2}} . \quad (2.2.2.1)$$

$$D(p_c, p_n) = (\sum_{i=1}^{m} (p_{ci} - p_{ni})^2)^{\frac{1}{2}} \quad (2.2.2.2)$$

We run both Ordinary Least Squares (OLS) and Mean function first, using these two lines to locate the intersection point p_c , then we calculate and rank the distances of all points to p_c . After that, we exclude the point which has the largest value among all $D(p_c, p_n)$, indicating that the point is the farthest from the CP among all data points. Meanwhile, we update the sample size, OLS and mean function repeatedly until we are satisfied with the filtering process by monitoring and evaluating the parametric model based on the Standard Deviation (σ^2), Residual Sum of Squares (RSS), the sum of squares errors ($SSreg$) , and the graph. This approach displays the following advantages: it adheres to the majority data points, requires less computation power, and runs efficiently especially with a low dimensional dataset.

Figure 2.3a displays an extreme outlier versus the rest majority data points. The CP's vertical axis value is fixed by the Mean Function to fit the majority, which is affected by outlier 1 but in alignment with the majority. Although LR tilts more towards the outlier 1, outlier's distance to CP is the largest among all $D(p_c, p_n)$. Therefore, we can easily spot and exclude it. In short, since the Mean Function keeps the CP vertical axis value in aligning with the majority, so even LR tends to fit both the majority and the outliers, RDM can be used to expose outliers by ranking distance from CP to all points. **Figure 2.3e** and **Figure 2.3f** display the cost function's convergence tendency before and after excluding several outliers, which match with the variance analysis table

and demonstrate that the convergence tendency has been improved after outlier exclusions.

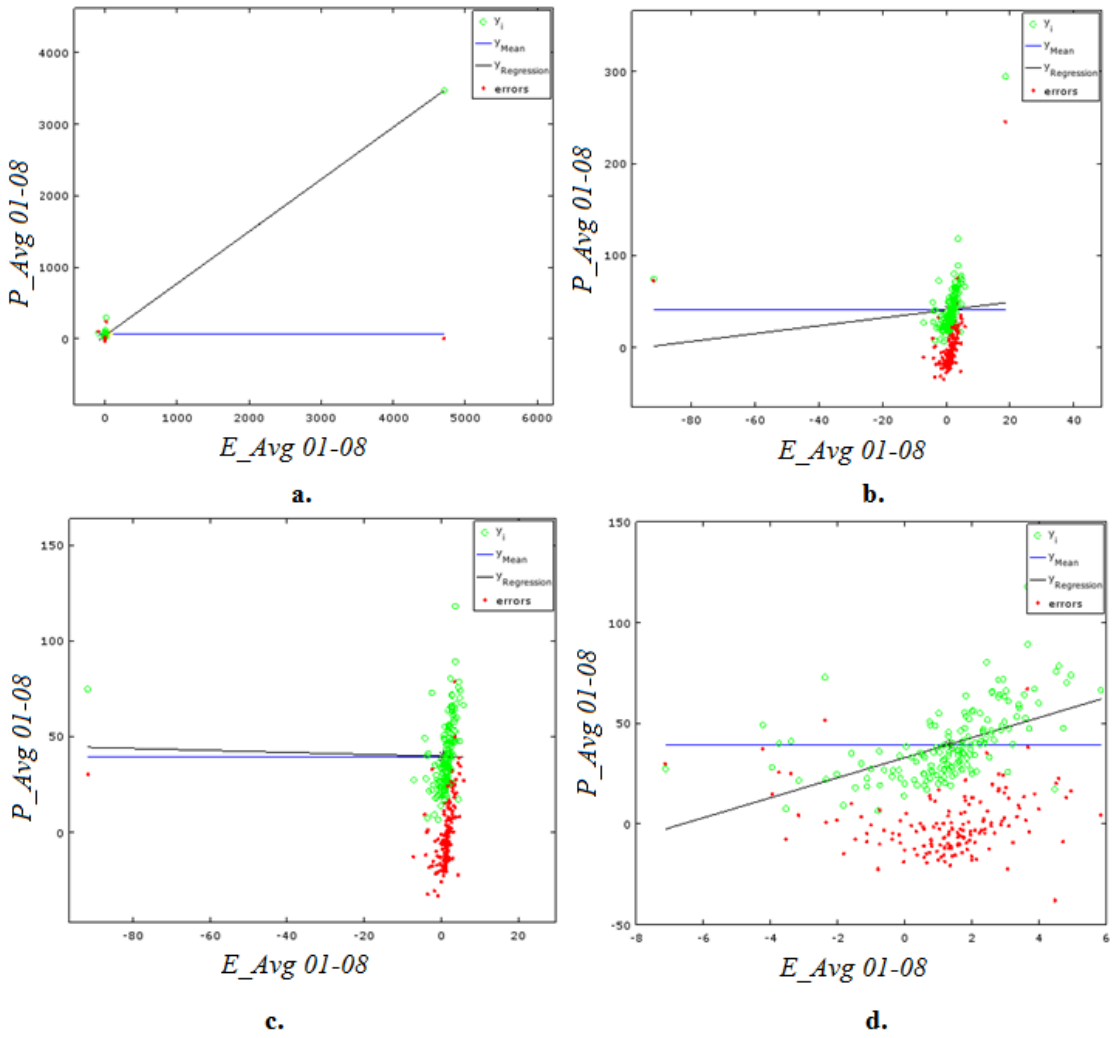


Figure 2.3 (a-d) RDM sequentially exposes and excludes outliers

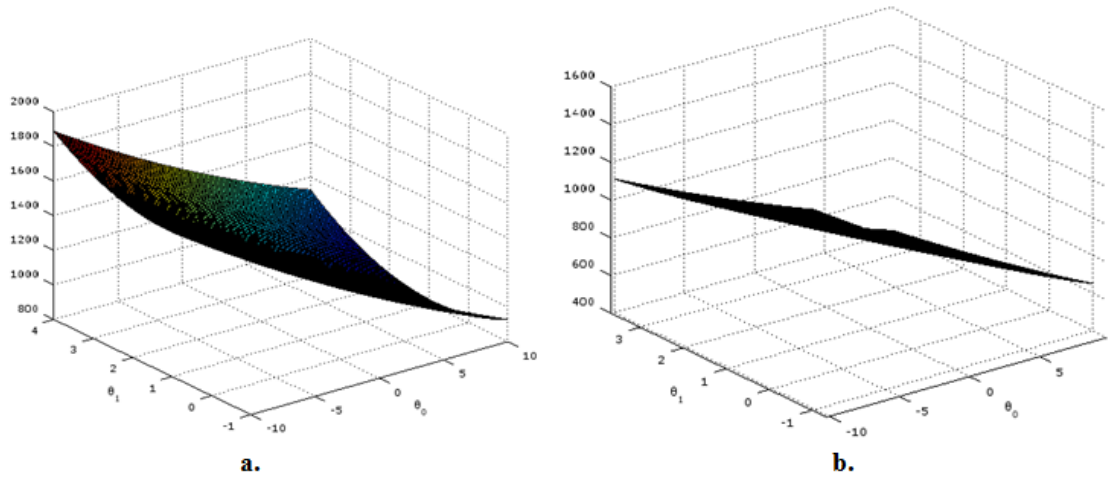


Figure 2.4 Convergence tendency before and after

Table 2.7 The analysis of variance table for regression

Original 159	df	θ_0	θ_1	$\sum_{i=1}^{n}\varepsilon_i^2$	σ^2	F-test:	R^2	**Outlier:**
Mean	n-1	62.85		11801719				**BRK-A:** **E_Avg: $4706.5** **E-Var: 45140798** **P_Average: $3471**
Residual	n-2	40.69	0.73	111871	712			
Regression	1			11689848		16405	0.99	
GD		NaN	NaN					
Revised 1								
Mean	n-1	41.28		112731				**Y:** **E_Avg: $18.56** **E-Var: 709** **P_Average: $294.3**
Residual	n-2	40.93	0.43	111006	711		0.02	
Regression	1			1725		2		
GD		40.64	0.43					
Revised 2								
Mean	n-1	39.67		48305				**AIG:** **E_Avg: $-91.73** **E-Var: 505523** **P_Average: $74.75**
Residual	n-2	39.70	-0.05	48281	311			
Regression	1			24		0	0.00	
GD		39.43	-0.048					
Revised 3								
Mean	n-1	39.44		47067				**BLK:** **E_Avg: $ 3.64** **E-Var: 30** **P_Average: $ 118.13**
Residual	n-2	32.98	4.98	33173	215			
Regression	1			13894		64	0.30	
GD		31.77	5.32					

As **Table 2.7** indicates, without excluding outlier 1 BRK-A, GD could not even converge for a solution with 500,000 iterations and various alphas. After excluding outliers step by step, we easily get the estimated slope and intercept by running 1,000 to 5,000 iterations with alphas ranging from 0.01 to 0.1. Note that the sample shrunk from

159 to 156. In this case, BRA-K was issued by Berkshire Hathaway (BH), a multinational conglomerate company, which engages in housing, insurance, transportation, manufacturing, food, energy and various businesses. By looking into its financial statement from 1977 to 2000, we grasp two important concepts. First, BRA-K is an outlier due to the fact that conventional accounting rules cannot accurately reflect BH's true value by excluding its non-controlled entities' undistributed earnings and generalizing its complicated capital structure (Berkshire Hathaway , 1997-2000); second, we witness repeated occurrences and tragic endings of imprudent lending or other risky business conduct. For insurance companies, keeping a sound financial structure should always be the key. However, many insurers completely ignored the safe underwriting principles during a good time, ending up with catastrophic failures in the economic downturn such as the outlier 3 AIG (McDonald & Paulson, 2015).

2.3 Dual Objective Minimization Cost Function Based Linear Classification

2.3.1 Introduction

Logistic Regression, Neural Network, Principal Component Analysis (PCA) and Independent Component Analysis (ICA) have been used extensively in classification problems with higher dimensional datasets or multiple features (Kurt, Ture, & Kurum, 2008) (Terrin, Schmid, Griffith, D'Agostino, & Selker, 2003) (Hinton & Salakhutdinov, 2006). On the contrary, LR is rarely considered as a classification approach due to its output characteristics. The most recent development in LR classification is a newly proposed Linear Regression-based Classification (LRC) approach and Distance-based

Evidence Fusion (DEF) algorithm which have been applied to image detection by forming many subspace classes (Naseem, Togneri, & Bennamoun, 2010). Similar Locally Linear Regression (LLR) is used to differential non-frontal and frontal face images by decomposing and extracting key features from training sets in order to form feature classifiers, so new images can be transformed and then obey the minimum distance decision rule to fit into different classes (Chai, Shan, Chen, & Gao, 2007).

Our research motives originate from a common misunderstanding of LR application (the misusage of multivariate/multivariable regression) and evolves into LR based classification technique. The multivariate analysis consists of multiple outputs, but multivariable is based on multiple inputs with only one output. Bertha Hidalgo and Melody Goodman (2013) found out that this misconception frequently appears in healthcare research and the first approach attracts more attention. For instance, a group of medical researchers used it to find patients' common features of both vertebral artery injury and blunt cervical spine injury (Lebl, Bono, Velmahos, Metkar, Nguyen, & Harris, 2013). The methodology behind this approach is to minimize two estimated outputs' errors with same inputs, same coefficients and predetermined weights of errors. Moreover, cost function and Gradient Descent (GD) are modified to solve for parameters. Our research contribution here is we form a line based on multivariate regression for classification. Our approach is similar to Multivariate Linear Regression (MVLR) which tries to predict multiple dependent variables simultaneously (Rencher & Schaalje, 2008), but differentiates in purpose and the underneath mechanism.

In the following sections, we first present the Dual Objective Minimization Cost Function Based Linear Classification (DOMCF) algorithm and related experimental

results based on a real world financial dataset. Then we describe a potential discoordination issue and prescribe a solution. By the end of this section, conclusions and future research extension are presented.

2.3.2 Dataset Description

Our initial objective was to test whether linear regression can be used to handle a real world predicament: limited inputs with multiple outputs, where two or more sets of outputs share one set of inputs. We collected hundreds companies' key financial information from 2001 to 2008 and gathered data from 2009-2014 as the performance index. We limited our test on one input **x** which is a set of companies' average earnings from 2001-2008 and two outputs: $\mathbf{y_1}$, a set of companies' average price from 2001 - 2008; $\mathbf{y_2}$, a set of companies' average earnings from 2009-2014.

We started our experiment with one crucial underlying assumption which is that $\mathbf{y_1}$ and $\mathbf{y_2}$ share the same input **x** and coefficients. Then we found out that because the two sets of output significantly map away from each other, it is impossible to have a line that reasonably represents both at the same time according to multivariate regression. On the contrary, it is more feasible by using the line to differentiate them. Therefore, by adding and adjusting weights in a newly modified objective function, we can successfully lay out a line to separate these linear separable datasets.

2.3.3 Methodology and Computational Results

The methodology behind this approach is as follows: we first form two hypotheses (h_1, h_2) that share the same parameters (θ_0, θ_1) and then we add correspondent weights

(w_1, w_2) into the cost function. Originally, LR is set to find the best line to represent one dataset, and MVLR is set to find the line that best represents two datasets. According to our observation, by adding and varying the weights to minimize the sum of errors of both classes, similar to the classic tug-of-war, we can successfully turn MVLR into a classification method. We let the algorithm identify a pair of weights that generate a set of parameters to form a good separation line by counting the number of points which are correctly classified. The algorithm is defined as follows:

The Dual Objective Minimization Cost Function Algorithm (DOMCF)

Inputs: One feature m samples X_i where $i = 1...m$

Output: Class 1: y_1, Class 2: y_2

1. Hypothesis and modified cost function:

Hypothesis 1: $h_\theta{}^1(x) = \theta_0 + \theta_1 X_1$ (2.3.3.1)

Hypothesis 2: $h_\theta{}^2(x) = \theta_0 + \theta_1 X_1$ (2.3.3.2)

Minimize cost function:

$$J(\theta_0, \theta_1) = \frac{1}{2m}(\omega_1 \sum_{i=1}^{m}(h_\theta{}^1(x) - y_1)^2 + \omega_2 \sum_{i=1}^{m}(h_\theta{}^2(x) - y_2)^2) \quad (2.3.3.3)$$

2. Vary weights and solve parameters by taking derivatives based on modified Gradient Descent:

$$\frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1) = \frac{\partial}{\partial \theta_j} \frac{1}{2m}(\omega_1 \sum_{i=1}^{m}(\theta_0 + \theta_1 X_1 - y_1)^2 + \omega_2 \sum_{i=1}^{m}(\theta_0 + \theta_1 X_1 - y_2)^2) \quad (2.3.3.4)$$

$$\frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1) = \frac{\partial}{\partial \theta_0} \frac{1}{m}(\omega_1 \sum_{i=1}^{m}(\theta_0 + \theta_1 X_1 - y_1) + \omega_2 \sum_{i=1}^{m}(\theta_0 + \theta_1 X_1 - y_2)) \quad (2.3.3.5)$$

$$\frac{\partial}{\partial\theta_1}J(\theta_0,\theta_1)=\frac{\partial}{\partial\theta_1}\frac{1}{m}(\omega_1\sum_{i=1}^{m}(\theta_0+\theta_1X_1-y_1)X_1+\omega_2\sum_{i=1}^{m}(\theta_0+\theta_1X_1-y_2)X_1) \qquad (2.3.3.6)$$

We use the dataset **'epe.txt'** that contains 100 companies' current, future earnings and price records. Here we have chosen two predetermined classes: Class 1(*P_Avg 01 -08*) and Class 2 (*E_Avg 09 -14*), each class has the same coordinating input feature X (*E_Avg 01 -08*). They show the tendency of clustering but also apart away from the opposite class. To form a line, we predefined two parameters (θ_0,θ_1), standing for the intercept and slope respectively.

We start our iteration by setting the range of the weight w_1 from 0 to 1 at the rate of 0.001. After trying different set of weights, we get different sets of parameters and estimated output based on updating two functions in Octave: **dualObjcostFunction (**X, y1, y2, theta, w1,w2**)** and **dualObjGD (**X, y1,y2,w1,w2, theta, alpha, iterations**).** Parameters (θ_0,θ_1) are solved by **dualObjGD** function for which we need to specify the feature vector, classes, initial weights, initial parameters, α and iteration time in advance. To use GD we set up three steps:

3. Initialized estimation of all outputs as: $y^{est}=\theta_0+\theta_1X$
4. Simultaneously update parameters according to the modified GD (2.3.3.5) and (2.3.3.6) with descending rate at α until it converges.
5. Keep records of all the combinations of thetas, weights and costs calculated by **dualObjcostFunction** function based on (2.3.3.3).

Therefore, each set of weights will ultimately reach a set of output estimation. In this case, based on the plot and predetermined classes we set the classification rule as:

Class 1: $y_i^1 > \theta_0+\theta_1X_i$ (2.3.3.7)

Class 2: $y_i^2 \leq \theta_0 + \theta_1 X_i$ (2.3.3.8)

Since all classes share the same horizontal coordinate values, we can easily compare and count all correctly classified points for each set of weights. After the counting and comparing process, we then single out one set of weights and assign to the DOMCF and modified GD to recall the parameters that form a line separating most points correctly. Several scenarios are plotted as follows:

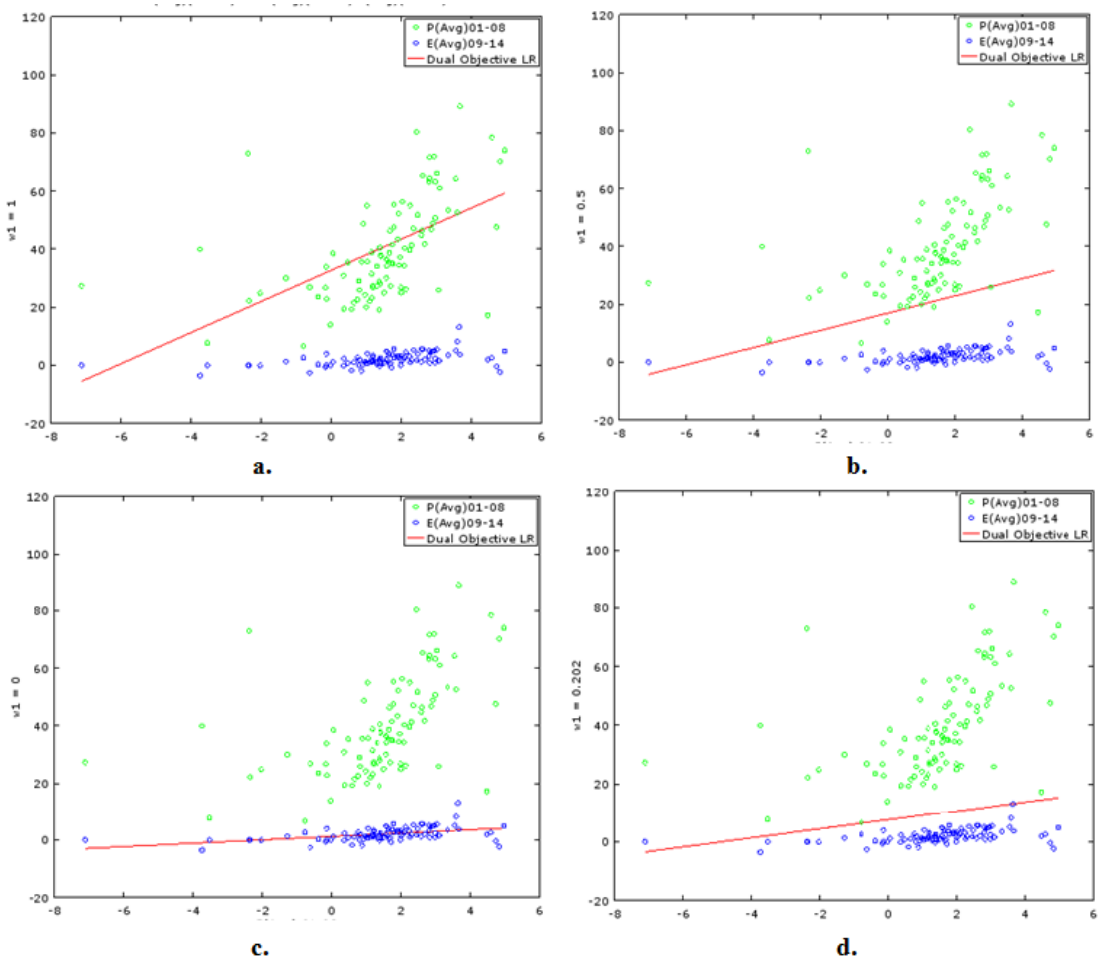


Figure 2.5 (a-d) DOMCF based classification results

Figure 2.5.a: $w_1 = 1, w_2 = 0$, DOMCF completely ignores Class 2 without even counting the sum of errors between the estimated y_2' and the real y_2. In short, the parameters in this case only forms a line best representing Class 1.

Figure 2.5.b: $w_1 = w_2 = 0.5$, it gives us a separation line which separates majority cases very well but is not the ultimate winner. Since it was set to treat both classes equally, minimizing the sum of all errors from both classes to the line based on the distance measure. The reason behind this deficiency is largely due to DOMCF's 'fairness'. Class 1 and Class 2 display very different mapping patterns. Points belong to Class 1 spread more widely while Class 2's points are more linearly condensed, which makes LR generate a less fitting line for both classes. Therefore, the well-behaved Class 2 has to make great compromises and permits the line stay closer to Class 1 in order to compensate its large in-group linear estimation errors. In short, to minimize the total estimation errors based on both classes, the algorithm chooses to balance its book by shifting away from a more linearized group. If our goal is to find the best separation line, we need to go through another step by adding a performance measure to let the algorithm search for the combination of parameters.

Figure 2.5.c: $w_1 = 0, w_2 = 1$, same case as $w_1 = 1, w_2 = 0$ but in a reverse manner. This time DOMCF completely ignored the existence of Class 1 without counting the sum of errors between the estimated y_1' and the real y_1. In short, parameters in this case solved by GD only forms a line representing Class 2 the best.

Figure 2.5.d: Our objective here is to find a line that best classifies the datasets with the highest accuracy by going through all the possible (w_1, w_2) combinations. We

get the best results with $w_1 = 0.202$ and $w_2 = 0.798$, which indicates that the algorithm put less focus on Class 1's total errors than Class 2's by heavily discounting Class 1's points' distance to the line.

2.3.4 Model Validation

Cross-validation (CV) is used to measure the classification capability and minimize the influence of overfitting (Shao, 1993). First, we separate the original dataset into different combinations of training and testing datasets. Second, we use training sets to build up models based on DOMCF. Thirdly, we test on left out data. Finally, we average the testing results to see the classification accuracy in practice. There are various ways to do cross-validation (Burman, 1989). The simplest one is partitioning the datasets randomly into 80% training set and 20% testing set. We did multiple runs and averaged out the results. Moreover, we test the procedure with different partition percentage, such as 40% versus 60%.

Table 2.8 Partition validation results

Partition: Training vs. Testing	(w_1, w_2) **Iteration: 500,1000**	(θ_0, θ_1)	**Classification Accuracy**
(80%, 20%)	(0.202,0.798)	(7.77,1.49)	99.3%
	(0.198,0.802)	(7.60,1.54)	99.3%
(70%, 30%)	(0.099,0.901)	(4.17,1.13)	96.7%
	(0.200,0.800)	(7.68,1.52)	99.2%
(60%, 40%)	(0.196,0.804)	(6.79,1.76)	98.7%
	(0.196,0.804)	(7.16,1.66)	99.1%

K-fold CV (Refaeilzadeh, Tang, & Liu, 2009) categorizes sample data randomly and equally into *k* subsets. We run *k* times CV sequentially on each single subset by using the rest *k*-1 subsets as the training set. Therefore, we have *k* lines and *k* testing results. Finally, an averaged result indicates the ultimate accuracy. In this case, after running 1000 iterations, we have the average accuracy percentage as 97.8%.

Table 2.9 Cross-validation 5 fold

Testing Set	Classification Accuracy
k1	96.9%
k2	96.9%
k3	99.4%
k4	96.9%
k5	98.8%

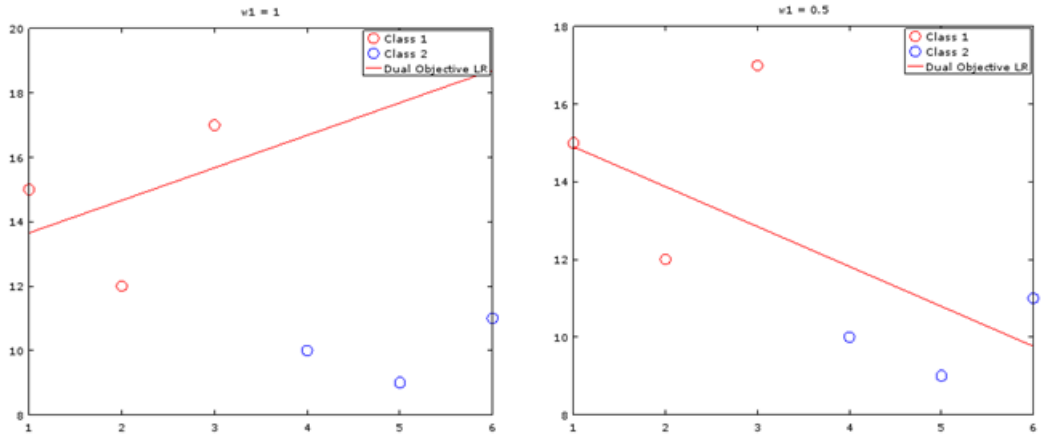
2.4 The Problem of Dis-Coordination

2.4.1 Problem Illustration and a Proposed Solution

An issue we might encounter is dis-coordinated datasets. Assuming we want to classify two linearly separable datasets that share the same feature but different values as below **Figure 2.6 (a-b)** shows. DOMCF fails to find a reasonable separation line no matter what weights we pick. The reason behind this failure is due to missing counterpart points. Since DOMCF is based on the trade-off and balance between two perfectly coordinated classes, so when there are no counterpart class points, it fails miserably by trying to identify a line to separate it.

As **Figure 2.6** shows, when $w_1 = w_2 = 0.5$, the algorithm treats both classes' errors equally. Since each of the point lacking of its class-counterpart as the algorithm

supporting pillar, so the line generated by DOMCF is solely based on each point's individual distance to the line, instead of considering simultaneously its own and counterpart's distance to the line. The classic 'Tug-of-War' element is missing here.



a. **b.**

Figure 2.6 (a-b) Classes with dis-coordinated horizontal values

Our approach is to treat this issue as a missing or incomplete data problem. We investigate several ways from easy to complex level, including interpolation (Davis, 1975), iterative value refining process (Fayyad, Reina, & Bradley, 1998), creating a symmetric dataset by deleting data, and adding more information by observing trend or using pre-defined distributions (Schafer, 1997). Some researchers work around this issue by completely ignoring the deficiency and avoiding the disadvantages of using pseudo data (Enders, 2010). However, the majority of them are in favor for filling methods such as interpolation, dissection, iteration or advanced ways to fill the gap.

Under the condition that the DOMCF algorithm will be applied to similar problems, we choose to do interpolation with pre-defined distributions (Chow & Lin, 1971). The main reason for adopting a random number generator confined by its upper and lower constraint is that it allows us to incorporate controlled uncertainty into the

experiment. Since Random number generation (RNG) intends to create numbers or samples which cannot be fully anticipated, so it gives us a certain degree of freedom but also remind us that those created points' main responsibility is to be supportive instead of being dominant. However, even RNG has been widely used in gambling, sampling, simulation and cryptography, no random generator embedded in a common platform can be viewed as the ultimately true random generator (Park & Miller, 1988). They either follow certain physics phenomenon, laws or algorithms such as Monte Carlo Simulation (Gentle, 2006), Yarrow Algorithm (Kelsey, Schneier, & Ferguson, 1999), Mersenne Twister Algorithm (Matsumoto & Nishimura, 1998). Since the number of needed supportive points is small in our case then the potential of repeated patterns should not be too concerned at this moment. Another key reason behind this procedure is to suit the algorithm's needs for matrix operation. In this case, points of Class 1 and Class 2 lack of the same horizontal values, we came up a simple yet innovative way to enable the DOMCF algorithm work properly by generating supportive data points.

First, we simply prescribe the range of each class by identifying its upper and lower bound.

Table 2.10 Sample points of each class and correspondent upper/lower bounds

Label	Point A	Point B	Point C	Lower Bound	Upper Bound
Class 1	(1,15)	(2,12)	(3,17)	12	17
Class 2	(4,10)	(5,9)	(6,11)	9	11

Second, we simulate counterpart points and add them into existing datasets respectively. New data points generation process must satisfy two conditions:

1. New points' horizontal value are set to fill the gap between existing datasets in order to let Class 1 and Class 2 have the same set of horizontal values.
2. Each new data point's vertical value is randomly generated by following a distribution and must fall within the pre-defined range.

The greatest advantage of our supportive points generation method is no need for iterative refining process, however, the success is based on one preassumed condition that the dataset distribution or pattern is known or can be assumed.

2.4.2 Methodology and Computational Results

The steps of generating supportive points are as follows: first, we compared Class 1 and Class 2's feature vector X based on its values, if there is any mismatch, we generate a new set of points for each class using random number uniformly distributed on the interval (0, 1) and bounds constraint.

$$SP_{Class1}^{new} = (x_{Class2}, \min(y_1) + (\max(y_1) - \min(y_1)) * U(0,1)) \quad (2.4.2.1)$$

$$SP_{Class1}^{new} = (x_{Class1}, \min(y_2) + (\max(y_2) - \min(y_2)) * U(0,1)) \quad (2.4.2.2)$$

Table 2.11 Counterpart points generation

Label	Exisitng Point	Couterpart Point
Class 1	(1,15)	(4, 13.37)
Class 2	(4,10)	(1, 9.98)

Parameters (θ_0, θ_1) are solved by **dualObjGD** and **dualObjcostFunction** functions which we need to specify feature vector, classes, initial weights, initial parameters, α and iteration. Here, we relax our ruling condition by setting classification

standards as below: If $y_i^1 > \theta_0 + \theta_1 X_i$ and $y_i^2 \leq \theta_0 + \theta_1 X_i$ or if $y_i^2 > \theta_0 + \theta_1 X_i$ and $y_i^1 \leq \theta_0 + \theta_1 X_i$, we consider it as a successful run.

Below are selected testing and preliminary results, cross points are supportive points generated by following the uniform distribution and being confined to the original classes' lower and upper bounds.

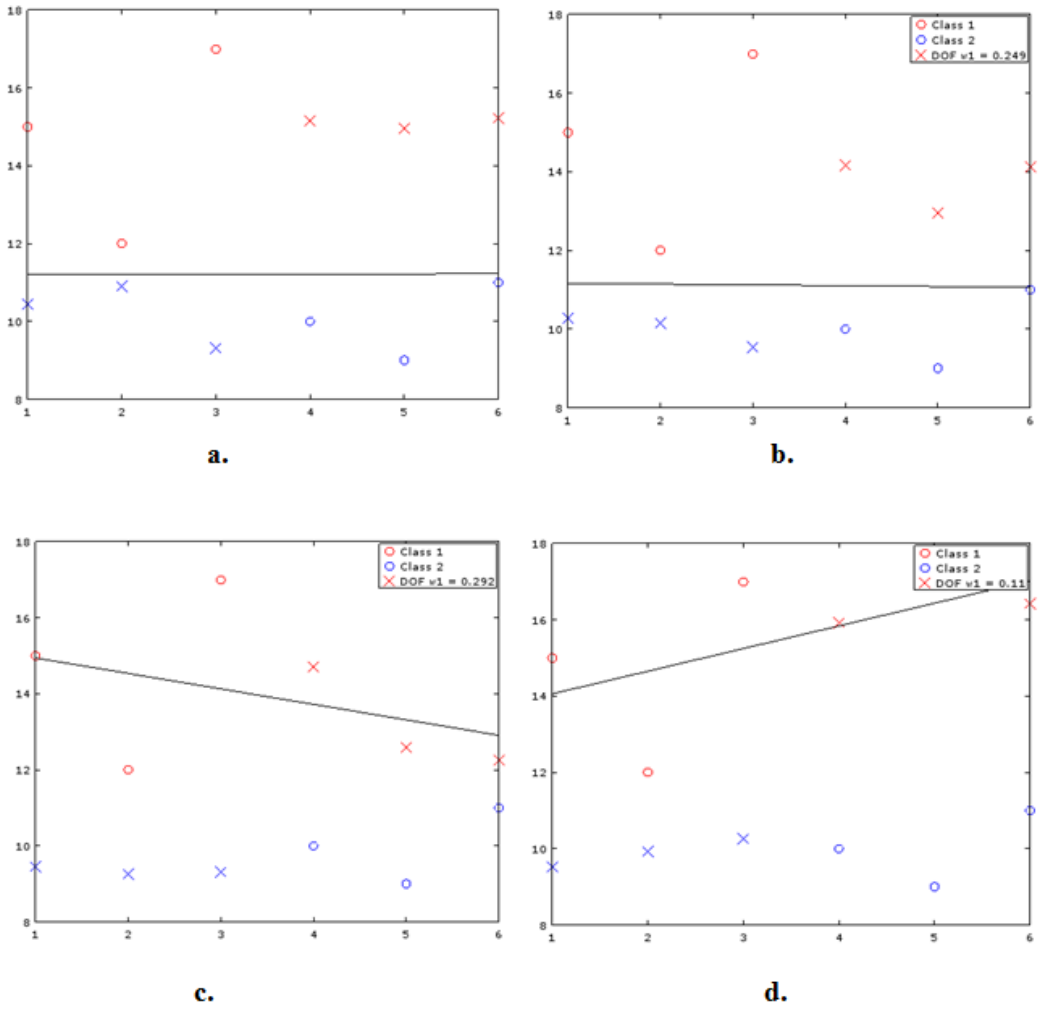


Figure 2.7 Supportive Points Generation and DOMCF Classification

Again, it only works when the datasets can be linearly separable, and so far we coded and tested it only in a two-dimensional case. As **Table 2.12** shows, we can

identify the separation line by adjusting the step of weights, the number of iterations and the value of the alpha. Insufficient iterations, too small or too big weight changing rate significantly affect our results. It works the best when the lower bound of class 1 does not cross the boundary of the upper bound of class 2. It is worth to mentioning that there can be numerous combinations of weights and parameters to specify a feasible separation line.

Table 2.12 Adjusting weights, iterations and alpha for classification

W step	Iterations	Alpha	W1	Intercept	Slope	Mis-Classficiation
0.001	5000	0.1	0.19	11.19	0.006	0
0.001	5000	0.1	0.249	11.18	-0.02	0
0.001	**1000**	0.1	0.292	15.35	-0.41	3
0.01	5000	0.1	0.11	13.46	0.59	2

2.4.3 Further Tests on Special Cases

To further validate our statement, we simulated several special cases and compared the results by displaying it on graphs to show the effects of before and after our supportive points filling steps.

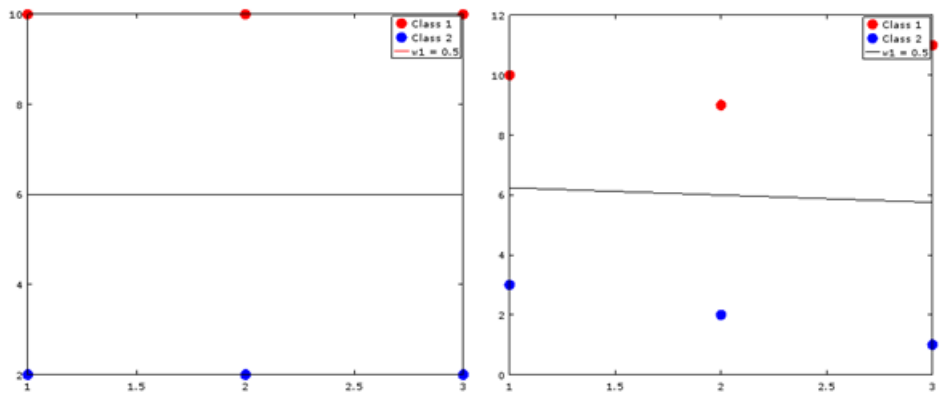


Figure 2.8 With or without supportive points based DOMCF Classification.

Figure 2.8.a Class 1 $\{(1,10)(2,10)(3,10))\}$ and Class 2$\{(1,2)(2,2)(3,2)\}$ share the exactly same feature values x_i corresponding to the same y_i value. So there is no missing pillar and no need to get help from simulating supportive points. Moreover, when $w_1 = 0.5$, DOMCF puts an equal weight on each class's sum of squared errors which represents the total distance from each class's points to the separating line. In this case, our algorithm is successfully validated by the **Figure 2.8a** which shows that the separation line is located in the center of two classes.

Figure 2.8b Class 1 $\{(1,10)(2,9)(3,11)\}$ and Class 2$\{(1,3)(2,2)(3,1)\}$ share the same horizontal values but x_i corresponding to different y_i. DOMCF still works pretty well by successfully separating both classes.

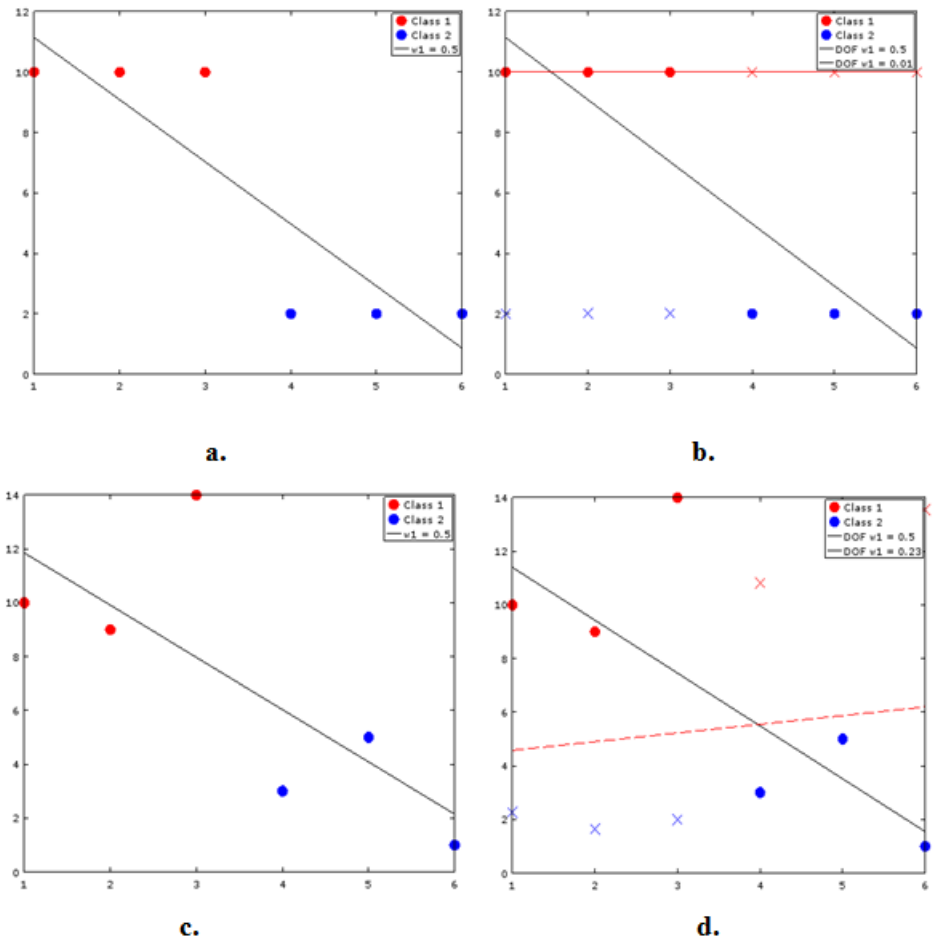


Figure 2.9 With or without supportive points based DOMCF classification

Figure 2.9a Class 1 $\{(1,10)(2,10)(3,10)\}$ and Class 2$\{(4,2)(5,2)(6,2)\}$ do not share the same horizontal feature values but each class's points share the same y value. In this case, we need to fill in new supportive points since **Figure 2.9a** shows that the separation line is based only on the goal of minimizing the sum of all classes' squared errors instead of taking the consideration of different class labels. Therefore, we use supportive points to facilitate the algorithm. **Figure 2.9b** displays a line that has an intercept = 9.999992 and slope = 0.000002 with w1 = 0.01 which successfully classifies those two classes. As we mentioned before, since we generated supportive points by

randomly simulating new points in the desired range so we expect that there could be various combinations of intercept, slope, and weights to form a separation line.

Figure 2.9c Class 1 $\{(1,10)(2,9)(3,14)\}$ and Class 2$\{(4,3)(5,5)(6,1)\}$ do not share the same horizontal feature values, and each class's points are corresponding to various y_i values, Class 1 and 2's do not cross each other's boundary. **Figure 2.9c** shows the failure of forming a separation line by DOMCF. Therefore, we had to generate supportive points to successfully apply the algorithm. **Figure 2.9d** depicts a line which consists of an intercept as 4.28; slope as 0.32 and w_1 as 0.219. Since we generated supportive points by randomly simulating new points in a predetermined range, we expect that there could be various combinations of intercept, slope and w_1 to form a separation line such as:

Table 2.13 Different combinations of intercept, slope, and weights for classification

Intercept	Slope	W1
5.63418	-0.12679	0.274
5.64162	-0.12672	0.275
5.64906	-0.12665	0.276
5.65650	-0.12657	0.277
5.66394	-0.12650	0.278
5.67138	-0.12643	0.279
5.67882	-0.12635	0.280

Figure 2.10a and Figure 2.10b shows that by running our algorithm and store as much as the possible combination of weights and parameters, we can even lay out the tangible upper and lower bound for a separation space.

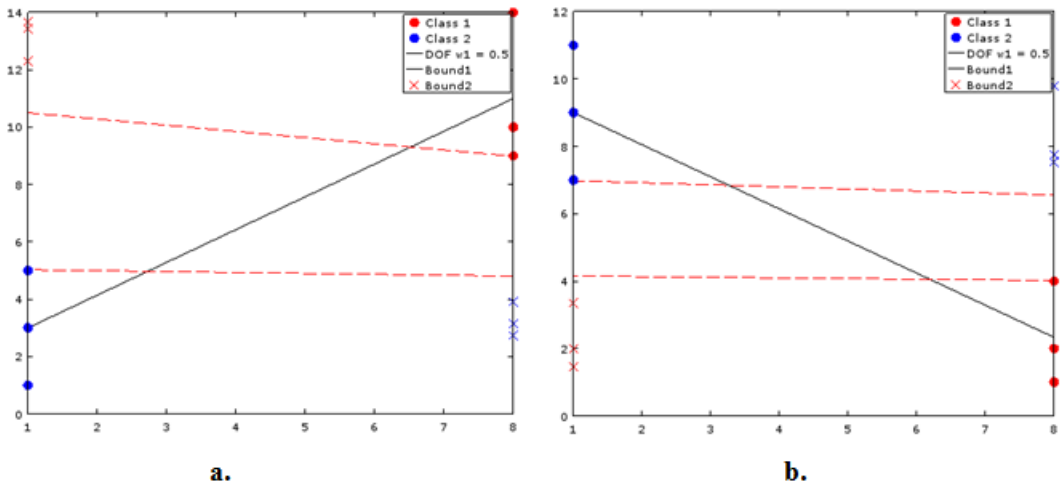


Figure 2.10 A supportive points based DOMCF separation space

Figure 2.11 displays a case that Class 1 and Class 2 cannot be linearly separated. Therefore, we need to seek other advanced algorithms to address this issue, such as polynomial regression, logistical regression or neural network.

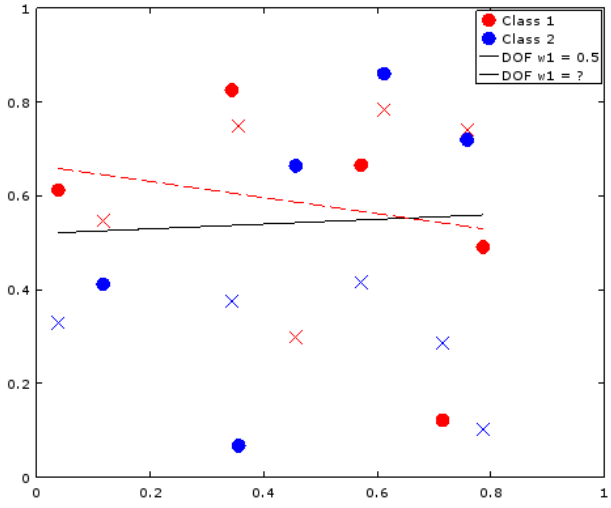


Figure 2.11 A non-linear separable dataset

2.5 Conclusions

In short, we propose and develop a new classification approach that shares the common ground with multivariate regression and shows improved classification accuracy.

Equipped with data filling techniques, this approach further tackles the imbalance issue between two dis-coordinated datasets. In future, we hope to further extend this approach to a higher dimensional setting and figure out ways to form linear subspaces to sequentially classify classes' number larger than two or with multiple features. For instance, we can conduct research on DOMCF based multi-classes classification comparing to other classification approaches. For multivariate $k > 2$ classes' issue, we can form $k-1$ lines. For every newly entering system points, it has to face the scrutiny of $k-1$'s test sequentially and will be categorized according to the ruling results. Imagine having *k* number of classes with *n* points, where each point contains *m* features and can be mapped into a high-dimensional space. Assuming that each point belongs to a certain class and each class can be separated from the other by either a line or a linear space. Similar to the multi-class logistic regression algorithm, we form the separation line or plane based on the idea of running one class against the rest of classes, so we will have *k-1* separation planes. Therefore, whenever a new point enters into the system we need to run it through all planes and determine in which class most likely belongs according to its distance from the testing plane. We can also implement other missing value filling techniques to see how different points generating process and randomness impact the classification results, observing and exploring how classes patterns affect the formation of the line.

CHAPTER 3: MULTIVARIATE LINEAR REGRESSION

3.1 Introduction

When multivariate linear regression (MLR) is applied on predictions, the common strategy is to equip it with inputs cross-examination, multicollinearity test, hypothesis testing, and statistics analysis to improve the results. Moreover, in practice, MLR often suffers from over/under-fitting. Therefore, researchers developed various ways to reduce these negative impacts by using feature selection, feature transformation, dimensionality reduction or model modification. Our approach derives from the existing and widely applied Regularized MLR method but with a different perspective. It reduces the total costs of errors if the dataset shows a consistent imbalanced tendency through a random sampling process. With observation and testing on a dataset of 156 companies over 10-year records, we believe that our modified MLR can be applied to similar cases to reduce measurement errors. A detailed discussion is as follows:

First of all, we conduct an extensive literature review on the MLR applications. Secondly, by weighing its pros and cons in practice, we narrow down our focus on how to improve the MLR performance in practice. Thirdly, we illustrate our method, the dataset's characteristics and why the modification makes sense in our case. In short, we rely on the understanding of the dataset and the algorithm to make a meaningful modification. At the end of this chapter, we discuss the potential of this approach, which can be polished further to deal with more complicated datasets.

3.2 MLR's Application History and Problems

MLR has been extensively applied in healthcare, economics, social science, and even

government policymaking process. For examples, in early healthcare, MLR was used to predict the possibility of various medical complications such as wound infection rate with given inputs such as a patient's gender, age, pre-pressing wound status (Ayliffe, Green, Livingston, & Lowbury, 1977). In manufacturing, MLR is recommended as a relatively inexpensive and efficient way to replace traditional high cost measuring equipment to estimate energy usage in factories (Cleland, Earle, & Boag, 1981). In geochemistry, MLR was used to testify the relationship between grain size and different weathering processes (Tolosana-Delgado & von Eynatten, 2009). In meteorology, MLR was used to forecast short-term ozone level and evaluated against NN, for the purpose of issuing air pollution warnings and taking quick actions to reduce the negative impacts (Moustris, Nastos, Larissi, & Paliatsos, 2012). In economics, MLR was used to evaluate the costs and benefits of funding various labor programs, recommending programs which significantly improve a country's long-term employment status (Syla, 2013). Most recently, MLR has been used to test and recommend medical treatment to combat critical health issues such as what is the best composition of nanoparticles to fight cancer (Kumar & Sawant, 2014). Since MLR related statistics tests reveal which features are more relevant, we believe that by segregating patients into two groups according to the significant ($age \leq 42$) and less significant ($age > 42$) groups, the results will be more useful. In short, it incorporates the idea that the age feature has different levels of importance instead of just being a numerical value. All above works jointly testify that MLR in practice is useful but also encounter issues such as random errors mask important factors, the pattern of a dataset affects the validity of a model, and the selection of features might pre-cap the model predictability.

Typically, a forward MLR (knowing all key features to forecast) produces more accurate prediction results than a backward MLR analysis (trying to identify potential relationships).Therefore, in order to minimize this inputs cross-intervention issue, regression analysis with more than one input normally has to go through multiple correlations analysis to minimize the possible interaction among variables and reduce the bias in the final results. However, in practice, it is rarely to have a set of features which are completely independent of each other yet all have significant impacts on the dependent variable. With all the above concerns, there are abundant ways to fine tune MLR. For example, dimensionality reduction (Cleland, Earle, & Boag, 1981), feature transformation (Ng A. Y., 1998), feature selection (Ng A. Y., 2004), and data transformation (Moustris, Nastos, Larissi, & Paliatsos, 2012). The best model is agreed to be the one which has few yet statistically significant features, teaming up with appropriate coefficients to minimize the cost of errors. We summarize the most common MLR improvement strategies as follows:

1. Forming higher polynomial terms or different models based on existing inputs, such as MARS (Friedman J. H., 1991) and Decision Trees (DTS) (Quinlan, 1987).
2. Using a set of MLRs for one or more inputs which can be categorized, such as age, gender, type, level and others (Goonatilake, 1981). Categorizing information often makes it easier for computation and interpretation. However, valuable information might be lost during this process.
3. Adding new inputs.
4. Enlarging the dataset for both training and testing purposes.

5. Regularizing parameters to suppress the over-fitting tendency (Xu & Chen, 2009); (Ogutu, Schulz-Streeck, & Piepho, 2012); (Bartlett, Mendelson, & Neeman, 2012).

For instance, Modified Multiple Adaptive Regression Splines (MARS) was one of many MLR enhancement methods used to estimate customers' responsiveness to direct marketing efforts, which incorporates inputs nonlinearity detection based on its hidden interactions (Deichmann, Eshghi, Haughton, Sayek, & Teebagy, 2002). MARS is a classic approach to increase MLR's dimensionality by evaluating all the possible combinations of inputs which are called Basic Functions (BFS). It is especially useful when dealing with sequential inputs displaying propensity or saturation effect. However, there is no solid method to pin down the optimal number of BFS. In addition, the costs of using many nonlinear BFS could be high, and the interpretation of MARS or BFS are not very explicit. In our case, since we adopted the average earnings and prices which already smoothed out the possible propensity impact so we do not think that MARS would be a good fit here. Similar in transformation but different in fundamentals, a group of DTS algorithms focuses on the possible transformation of the dependent variable.

Lee (1959) used pattern analysis to point out two main reasons why MLR underperforms in practical behavior sciences: First, the dependent variable cannot be easily defined as a single definite value, which is usually based on the combination of many different measurements. Second, users often abuse MLR without realizing the pattern of a dataset. The former issue can be tackled by using a set of MLRs, logistic regression or neural networks. The latter one demands users be more discriminative

towards a specific dataset (Lee, 1959). He suggested and applied several ways to address this issue: such as using non-additive pattern analysis. Although in his college students' GPA prediction case, the non-additive joint function failed to outperform traditional MLR. However, based on our observation and experiments, we agree with his idea that LR and MLR have no differentiating attitude towards individual cases, without even considering the difference between the majority and minority groups.

Regarding new features or constructed polynomial terms, it has been demonstrated theoretically that by adding as many features as possible, we can always find a perfect line to fit the training set, but at the same time the possibility of overfitting is rising rapidly. Therefore, researchers came up with several ways to curb the negative influence of overfitting, such as adding regularized terms into the cost function to form Regularized Linear Regression (RLR) a set of Modified Gradient Descent (MGD) (Friedman & Popescu, 2003). The regularized error function is as follows, which can be minimized by iterative procedures as (3.2.2) and (3.2.3):

$$J(\theta) = \frac{1}{2m}\left[\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda\sum_{j=1}^{n}\theta_j^2\right] \tag{3.2.1}$$

$$\theta_0 := \theta_0 - \alpha\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_0^{(i)} \tag{3.2.2}$$

$$\theta_j := \theta_j - \alpha\frac{1}{m}\left[\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)} + \lambda\theta_j\right],\ (j = 1,2,3...n) \tag{3.2.3}$$

$$\theta_j := \theta_j(1 - \alpha\frac{\lambda}{m}) - \alpha\frac{1}{m}\left[\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)}\right],\ (j = 1,2,3...n) \tag{3.2.4}$$

The logic behind the regularized approach is that by adding $\lambda\sum_{j=1}^{n}\theta_j^2$ into the

original cost function, λ works as a penalty term for the compensation of using more features. Now the goal turns into minimizing the cost of estimation errors plus the costs of using multiple features. Setting a reasonable λ will suppress the parameters put in front of each feature, if λ is big enough compared to the rest components, then each feature will function in a much more restricted way. There are many derivations to suit for different restriction needs. For example, Goonatilake (1981) noticed the imbalance data influence in its MLR experiment and pre-categorized 12,000 cases into 96 groups, using the number of patients per each group as a weight parameter to prevent minority overpowering the majority. Similarly, a Weighted Regularized Linear Regression Cost Function (WRLR) can be formed, the weight and penalization condition jointly decide when and how to penalize inaccurate estimations. For instance, if we want to penalize overestimation, then when the estimation is larger than the actual output, the difference will be multiplied by a weight parameter large than 1, so the overall costs will be larger than the original setting:

$$J(\theta) = \frac{1}{2m}\left[\sum_{i=1}^{m} W_i (h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^{n} \theta_j^2\right] \tag{3.2.5}$$

Overestimation Penalization: $W_i = a$ if $h_\theta(x^{(i)}) - y^{(i)} > 0$ (3.2.6)

Underestimation Penalization: $W_i = b$ if $h_\theta(x^{(i)}) - y^{(i)} \leq 0$ (3.2.7)

3.3 Feature Penalty MLR (FPMLR)

3.3.1 Methodology

Based on the previous discussion of Regularized MLR and the modified cost function, we try to minimize the overall cost errors by putting more focuses on the majority

group's cost errors. The underlying assumption is that we assume the training dataset has an imbalanced structure and can be categorized into two groups, which enables us to differentiate each group's contribution to the cost function. We start by looking at the histogram graphs of inputs to identify a rough boundary of the majority group and leave the rest aside for further consideration. After the categorization, we then discount the estimation errors of the minority group in cost function in order to make the outcome fits more for the majority. For instance, m_1: Group1 includes companies have average positive earnings but no more than \$5, and its average stock price is no less than \$20 but no more than \$80. m_2: Group2 includes any point that does not belong to group1, its estimated errors will be discounted with a penalty item $\lambda < 1$ since we put more focus on how to fit the MLR for the majority Group1.

FPMLR Algorithm

Hypothesis:

$$h_\theta(x) = \theta^T X = \theta_0 X_0 + \theta_1 X_1 + \theta_2 X_2 + ... + \theta_n X_n \tag{3.3.1.1}$$

Cost function with feature penalty:

$$J(\theta_0, \theta_1, ..., \theta_n) = \frac{1}{2m_1} \sum_{i=1}^{m_1} (h_\theta(x^{(i)}) - y^{(i)})^2 + \frac{\lambda}{2m_2} \sum_{i=1}^{m_2} (h_\theta(x^{(i)}) - y^{(i)})^2 \tag{3.3.1.2}$$

Batch gradient descent: Adjusting parameters θ_j to get minimized cost $J(\theta_0, \theta_1, ..., \theta_n)$, simultaneously update thetas for every $j = 0,1,...n$.

$$\theta_j := \theta_j - \alpha(\frac{1}{m_1}(\sum_{i=1}^{m_1} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}) + \frac{\lambda}{m_2} \sum_{i=1}^{m_2} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)})) \tag{3.3.1.3}$$

As we discussed it before, the essential rule of distinguishing investment stocks

from speculative stocks is stability. Therefore, we focus on the earning stability and price level as explanatory variables for the future rate of return. The steps of the algorithm are as follows:

1. We pick one or more features as the penalty feature.
2. We set rules based on feature's histogram graph to categorize them into two groups.
3. We go through all vectorization, feature scaling, and GD to solve for the thetas.
4. Finally, we evaluate the model by calculating the sum of prediction errors that is equal to the sum of the majority group's estimation errors and the discounted minority group's estimation errors.

3.3.2 Data Analysis

The choice of inputs depends on the data availability and its interconnection with the output. Past average earnings and prices have been considered to have certain relationships with the current average earnings. A correlation R^2 analysis indicates that: the variability in 2009-2014 the Averaged Earnings (*E_Avg 09-14*) can be explained separately by 2001-2008 the Averaged Earnings (*E_Avg 01-08*) (22%) and 2001-2008 the Averaged Price (*P_Avg 01-08*) (23%). However, those two variables have already been proved positively correlated to each other. **Figure 3.12** shows that how the *P_Avg 01-08* can be used to linearly estimate *E_Avg 01-08*. If we use both *E_Avg 01-08* and *P_Avg 01-08* to predict *E_Avg 09-14* then we should expect certain overlapping effects. And the total explained variations should fall into the range of [23% ~ 45%], depending on how those inputs interact with each other.

The contribution of a new feature into the model depends on how well it can be used to explain a dependent variable which the original feature cannot explain and the relationship between the added new feature and the original feature. **Figure 3.12** shows that the un-explanatory part of *E_Avg 09-14* by *P_Avg 01 -08* is related less strongly with the un-explanatory parts of *E_Avg 01-08* by *P_Avg 01-08*. So we expect that the total explained capability is no more than the current upper limit 45%. By examining the graph and slope of *E_Avg 01-08* with and without *P_Avg 0 -08* (0.56 vs. 0.89), we conclude that a higher past average earning indicates a higher future average earning. Similarly, the variability of distressed *P* ratio (the price depreciation rate in 2009) can be explained by *Deficit 01-08 (17%)* and *P/E_Avg 01-08* (16%). In conclusion, higher past average earnings with higher average prices together indicate a potential of higher average earnings in the future.

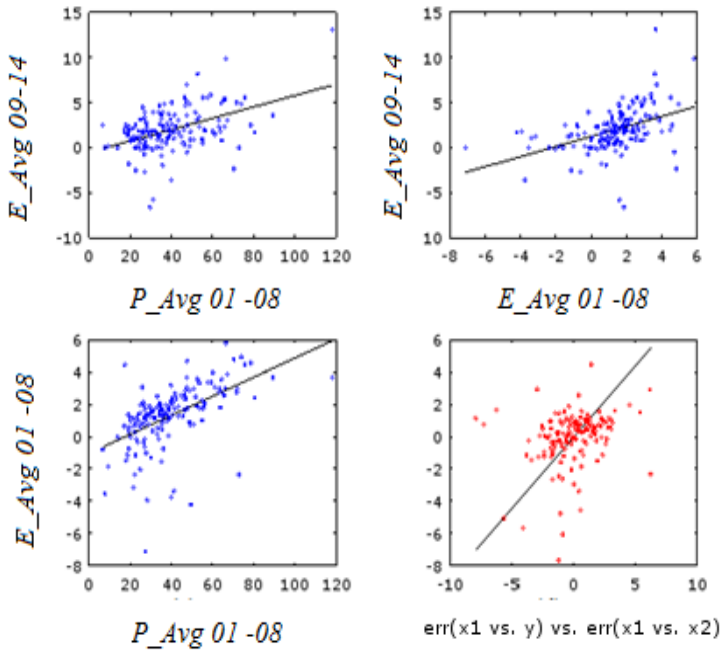


Figure 3.12 Multivariable linear regression analysis

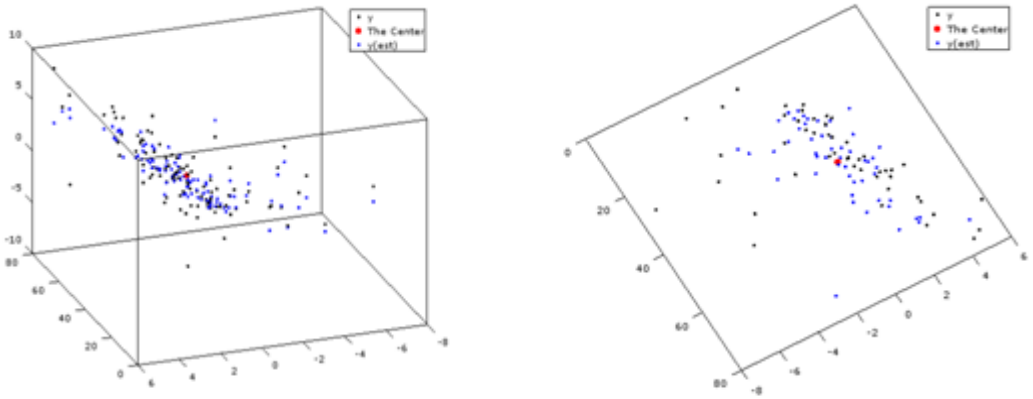


Figure 3.13 (a-b) A 3D plot of *E_Avg 09-14, P_Avg 01 -08* and *E_Avg 01 -08*

Here, we randomly select 100 data points as the training set and use the rest 56 for testing and validation sets. Then we perform feature scaling on those samples in order to facilitate convergence. Feature scaling is to get every feature into a similar range such as [-1, 1] or [-0.5, 0.5], it has been testified that a narrow contour plot based on scaled feature will make Gradient Descent (GD) take less time to converge. Here we use Mean Normalization to replace $x_i (i = 1...n)$ except x_0 with $(x_i - \mu_i) / (x_{\max} - x_{\min})$ or stand deviation to make sure all features fall into a certain range. We adopt the traditional cost function with added regularized terms and use GD to solve for thetas. We use the sum of total estimation errors on all data points to measure the model performance. Each data point's estimation error is calculated based on the same set of parameters and the same regression line, contributing equally to the cost function. Finally, we compare the basic Multiple Linear Regression (MLR) and our proposed dataset oriented Feature Penalty Multiple Linear Regression (FPMLR) based on the dataset which includes two inputs *P_Avg(01-08), E_Avg(01-08)* and one output *E_Avg(09-14)*.

The reason we chose GD here is because it works well in both small and large datasets without worrying about the non-invertible issue. And it enables us to modify and use a newly developed cost function. At the same time, the downside would be it always needs to start with a set of preset parameters, such as the rate, the penalty term, the iteration time and many others, for inexperienced users or new any new dataset, it takes time to conduct the fine tuning experiments. The benefit of using GD is by plotting out the cost function against the iteration times, we can directly see whether it works properly. If it does, then $J(\theta_0, \theta_1, ..., \theta_n)$ should decrease through every iteration and converge to a stable value easily. Andrew Ng (2013) suggests that setting an automatic convergence threshold test value as 10^{-3} will help practitioners to easily identify the moment of convergence.

3.3.3 The Analysis of Computational Results

Instead of taking 2002 to 2009 earnings and prices records individually as inputs and the 2010 to 2014 earnings as outputs, we used the averaged earning and price to simplified the illustration. This choice is not without any base since practitioners' long-term observation indicates that a single year's data is less representative than 5 or 8 years averaged data. We ran more than 20 models based on various feature criteria, and selected a few significant ones to show the improvement. Although there are many ways to evaluate a model's performance or an algorithm's effectiveness such as R^2, Mean Squared Errors (MSE) and coefficients significance testing, here we use MSE which is the sum of Euclidean Distance (ED) among all estimated and actual outputs.

Table 3.14 FPMLR vs. MLR

Iteration: 1000 **Feature 1: *P_Avg(01-08)*** **Feature 2: *E_Avg(01-08)***	**Min**	**Max**	**Average**	**Variance**
MLR_Sum(Errors^2)	637.98	721.27	650.71	10.94
FPMLR_Sum(Errors^2) with feature 1 @ condition: (1st quantile, 3rd quantile)	487.49	641.00	551.75	25.31
FPMLR_Sum(Errors^2) with feature 1 @ condition:(20,75)	553.08	665.12	571.59	11.71
FPMLR_Sum(Errors^2) with feature 2 @ condition: (1st quantile, 3rd quantile)	552.84	860.56	695.34	47.26
FPMLR_Sum(Errors^2) with feature 2 @ condition: (0,4)	481.76	579.12	520.66	15.73
FPMLR_Sum(Errors^2) with feature 1 & 2 @ condition: (20,75),(0, 4)	591.39	1019.7	729.41	76.39

The results show that when we pre-categorize the dataset into two groups either by ***P_Avg(01-08)*** to identify the majority as companies have an average price in the range of **($20 ~ $75)** or by ***E_Avg(01-08)*** to indentify the majority as companies with average earnings in the range of ***($0~ $4),*** FPMLR reduces the minimal, maximum and average MSE values without increasing MSE variances significantly. A major reason the penalty feature model performs better than basic MVLR in this case is due to the consistent imbalance tendency. For example, when we set the standard of the Group1 as companies with *E_Avg 01-08* in the range of $0 and $4, at a typical 10000 times iteration, the ratio of randomly selected Group1 to Group2 ratio is very close to the original dataset's ratio.

3.4 Conclusion

A common and traditional way to deal with outliers or imbalanced data is filtering. By simply eliminating unwanted samples, researchers can mathematically achieve better results and save computational costs as long as the elimination process is consistent with the research goal (*Trading Strategy*). In our case, we assume the consistency of imbalance and incorporate it into the cost function, which outperforms the basic MVLR and the elimination approach under the assumption that the testing and validation sets follow the pattern. In reality, a decision is made mainly based on the current constraints and comparison rarely the best selection. We notice that the learning process is an evolving process that absorbs valuable information from mistakes in order to set up standards for future reference. In future, we can include more samples by penalizing companies which have not been able to provide enough historical records (Such as companies with less than 8-year records) or companies have negative average earnings by adding new feature *Def_01-08* in order to categorize companies into two sectors: with or without deficits.

Regression problems with a relatively large number of features against small dataset size are usually solved by RLR in order to reduce over-fitting. In our case, it is totally the opposite that by limiting inputs selection as 2 comparing to 80 times sample size, we have a higher chance to encounter the issue of under-fitting. By looking into the objective cost function, we see that the traditional MLR algorithm does not take the pattern of data into consideration. Therefore, we propose a new algorithm based on the idea of regularization (L1 constraint: penalty term) but to enhance the MLR's fitting capability. Modifications occurred in the cost function, the iteration and final evaluation

process. Our experiments show that this newly modified approach consistently outperforms the traditional MLR under two conditions that the dataset shows the imbalance tendency and with reasonable pre-categorizing conditions. Multicollinearity is hard to be avoided but can be measured in reality. In our case, we notice this issue but there are two main reasons our inputs selection is viable: First, the average prices do not always exhibit a strong linear relationship with the average earnings in either our case or in historical evidence. The advantage of adding the average price into the MLR outweighs its disadvantage. Second, our focus is to show improvement based on the pattern of the dataset. Our main goal in this section is to verify that the modified MLR indeed outperforms traditional MLR due to its pre-categorizing process, instead of searching for the best model.

CHAPTER 4: NON-LINEAR SOLUTIONS FOR COMPLEX DATASETS

4.1 Introduction

In this section, the research interests originate from Neural Network (NN)'s great reputation in pattern recognition, classification, and prediction, and the main objective is combining NN with other ML techniques to uncover hidden relationships among a large number of real-world financial attributes. First, by analyzing and implementing a three-layer NN systematically in Octave, we obtained a deep understanding of NN's functionalities and characteristics. Second, after conducting an extensive literature review on NN's empirical studies in securities analysis, we found out that the majority of NN applications in this area are set to forecast the movement of stock prices instead of evaluating the associated companies from a long-term investment perspective. Although NN is an intelligent system, to maximize its performance, users must apply it on datasets that possess a certain degree of intrinsic stability or a persistent pattern to allow a successful and meaningful learning process. Therefore, keeping away from the goal of forecasting prices, we take a new route by following Graham's fundamental investment principles (Graham & Dodd, 1934) and combine with sophisticated feature selection methods to improve investment classification process.

4.2 Literature Review

NN is an intelligent system that mimics the learning process of a brain. It can also be viewed as a random yet guided exploring process by letting all inputs interact with each other in order to form a new set of critical components. Inputs, outputs, activation

nodes, and layers are predetermined and connected through the Feedforward-backpropagation processes. NN's popularity arises from its capability to quickly absorb, analyze, and use a large amount of information to form complicated non-linear boundaries in the classification territory. In short, the algorithmic advantage of NN are its high flexibility to approximate other functions (Hornik, 1991), high tolerance for linearity (Teräsvirta, Lin, & Granger, 1993), proper handling of imbalanced data (He & Garcia, 2009) or datasets with abundant features but relatively less samples (Masters, 1993), and self-autonomous critical component formation talent (Lippmann, 1988). The benefits of using NN are not just saving human labors but to eliminate certain preventable human errors (Yeung, Botvinick, & Cohen, 2004). Although being applauded as a strong and powerful technique, NN has downsides too, such as the possibility of over-fitting, high computational costs, unclear tuning processes and unstable solutions (Kaastra & Boyd, 1996).

The early research of neurons (Hebb, 1949) and the single neuron Perceptron model (Rosenblatt, 1958) laid a solid foundation for multi-layer NNs. Multivariate Adaptive Regression Splines (MARS) can be viewed as an early manual version of NN, which is a two-stage process by building up a set of Basis Functions (BFS) first and constructing a model based on these derivations (Friedman J. H., 1991).Time Lagged Recurrent Network (TLRN) is one of many most recent NN developments specializing in the area of time series forecasting (Wang & Traore, 2009) (Kelo & Dudul, 2011). A well-designed NN has been claimed to outperform other basic ML models, especially when there is an undetermined and complicated relationship between a large number of inputs and relatively few categorical outputs (Bailey & Thompson, 1990).

NN has already been applied extensively in economic and financial research: forecasting GDP (Tkacz, 2001) and unemployment rates (Johnes, 1999), assessing the risks of lending (Angelini, Tollo, & Roli, 2008), and identifying successful investing strategies (Trippi & Turban, 1992). In investment analysis, NN's applications can be seen in many categories. For instance, combining both political and psychological attributes, a four-layer NN is claimed to outperform Multiple Discriminant Analysis (MDA) in stock price prediction (Yoon & Swales, 1991). Information from letters to shareholders through content analysis were used in NN to distinguish companies by returns (Swales, 1992). Quah and Srinivasan (1999) combined many key financial and political features as inputs with NN to select stocks which generate above average performance. Lam (2004) incorporated both micro and macroeconomic factors in NN to predict stocks' rates of return and suggest equipping the NN with other data mining techniques. Chen,Weinberg and Yook (2013) adopted NNs to identify companies that have higher growth potential. Other interesting uses include utilizing distributions (Saad, Prokhorov, & Wunsch, 1998), combining with the expert system (Bergerson & Wunsch, 1991), pruning attributes (Lawrence, 1997), and building hybrid systems to seek for performance improvement (Baba & Kozaki, 1992). In short, NN applications in securities analysis area exhibit a three-stage development: from the early attempt of varying inputs and outputs selection to the second stage of comparison studies, and then forming complex hybrid systems. However, researchers in this area only agree on few things: a well-designed NN outperforms MLR, MDA, and expert systems in general; backpropagation process shows a consistency in picking good parameters; a three-layer

NN is sufficient in most cases solutions; and no rules can be followed to quickly identify an optimal NN architecture (Kaastra & Boyd, 1996).

Table 4.15 A comparison to several most cited NN securities analysis work

The Objective	Inputs	Outputs	The Structure	Comparison Analysis	Highlights
Classification and comparison (Swales, 1992)	Dividends, Price appreciation, Market value, 98 companies	High return/low return	Multi-Layer Backpropagation	Outperforms MDA	Content analysis
A review of market forecasting techniques (Lawrence, 1997)	Price, Volume, Pattern, Market indicators, Distributions	Price changes	Technical analysis, Fundamental analysis, Time series forecasting, Random walk process	Outperforms base level methods	Trend analysis, Regression, Efficient Market Hypothesis, Chaos theory, Pruning technique
Time series NN design (Kaastra & Boyd, 1996)	Technical and economic inputs	Price changes	Recurrent networks Probabilistic networks Fuzzy logic	Three layer is sufficient	A comprehensive guideline
Fundamental and technical analysis (Lam, 2004)	3-year 16 Financial and 11 macroeconomic attributes, 364 S&P companies	Above/below the market benchmarks	Multi-layer Backpropagation	Outperforms the minimal market benchmark	Post-processing with a rule extraction technique
Three investment principles (Zhang & Trafalis, 2016)	15-year 35 financial attributes, 613 public companies	Investment/ Speculative Class	A two stage multi-layer NN, Backpropagation	Outperforms LogR and KNN with feature selection	Output labeling process, Feature scaling, Feature selection, Regularization

4.3 The Problem and Proposed Solutions

A successful NN requires three components: critical inputs, clear outputs, and the right structure. The choice of layers, nodes, and architecture can be experimented with or follow certain rules. A well-designed NN has been proven to be able to approximate any measurable function (Hornik, 1991). Therefore, in practice, we believe that NN's underperformance might be due to lack of critical inputs or less distinguishable outputs. For instance, financial information is abundant but not all of them can be assembled in a timely and mathematic way. Based on our previous historical review and case analysis, numerous factors can trigger a stock to deviate from its main course. The well-known ones are economic, industrial, operational, political, meteorological, physiological factors, and interrelated markets' influence. Among these factors, some are quantifiable but many are not. Moreover, there are hidden factors which we do not know how and where to collect information. Therefore, we believe that forecasting stock prices with ML algorithms is theoretically sound but virtually intangible due to asymmetric, unquantifiable, and missing information. With that understanding, we shift away from trying to simulate the movement of prices but aim to maximize the value of available financial inputs by setting up a tangible goal in order to maximize the functionality of NN. This explains why we put extensive efforts on the output labeling process that is vastly underestimated in NN applications. Furthermore, few applications explore the impact of different feature scaling methods on NN's performance, many applications use predetermined features instead of being inclusive, and regularization deserves more investigations due to NN's over-fitting tendency with a large number of features. In sum, the main research contributions are as follows:

1. Apply NN with a different perspective by introducing a fundamental investment principle that the long-term stability embedded in the financial records is the key to achieve ML success in practical securities analysis.
2. Incorporate and experiment NN, LogR and KNN with different feature selection techniques.

Therefore, we propose a two-stage NN in investment analysis. At the first stage, our main goal is to search for an NN model that represents well for the classification process and learn how to set parameters, nodes, and weights. The output labels are jointly determined by three investment principles based on a company's historical records. Again, we have no intention to consistently predict stock prices but focus on a company's long-term prospect based on its intrinsic stability. At stage 1, the transformation process turns a 10-year 35 financial attributes dataset into a matrix that includes 413 companies with 350 features. This dataset is processed with three different scaling approaches and combined with a regularization term for potential over-fitting issue. At the second stage, we use the next 5-year (2010-2014) dataset as the forecasting set, letting NN have more freedom to identify potential key attributes. The implementation is dynamic and rigorous, which is based on Kaastra and Boyd's (1996) eight-step NN time series guideline, Andrew Ng's NN tutorial and various ML techniques. The following session consists of detailed discussions on the methodology, inputs selection, output labeling process, computational results, conclusions, and future extension.

4.4 The Methodology

A three-layer NN is commonly used and considered as the most efficient one in time series forecasting based on the evaluation results of computational costs, learning speed and classification accuracy (Kaastra & Boyd, 1996). It usually goes through a couple rounds of Feedforward (FF) and Backpropagation (BP) processes to finalize the parameters. Each layer's FF and BP processes are affected by its previous and next layers' components. Serving as the judge to compare the estimated values with the original outputs, the output layer gives feedback on how to adjust the weights and structure. There are similarities among NN and other linear and non-linear ML algorithms. For example, they are all set to minimize the total cost of errors. As Kaastra and Boyd (1996) point out: a one layer and one output NN can be considered as an LR, but a group of LRs must be combined in a right way to be close to an NN. Common NN tuning methods include varying parameters, adding, deleting nodes, increasing or decreasing layers, mapping features into higher polynomial and cross-product terms.

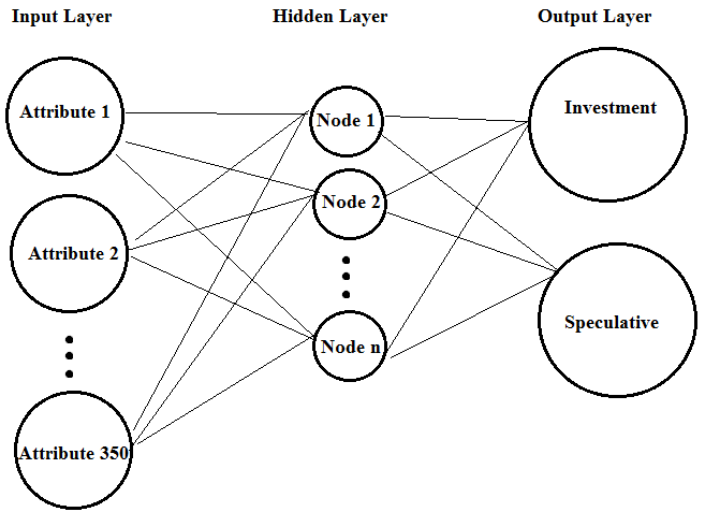


Figure 4.14 The 3-layer NN layout

In general, a three-layer NN starts with $X_{m,n}$ (a matrix of *m* samples with *n* features) and Y_m (a vector of output with *K* classes). At the input layer, a vector of bias term x_{m*1}=1 is added to the original input matrix $X_{m,n}$, turning into the first activation unit $a^{(1)}$=$X_{m*(n+1)}$. Then depending on the choice of the nodes q at the second layer, we initialize a random set of thetas $\Theta^{(1)}_{q,(n+1)}$ and multiply it with $a^{(1)}$, turning into $z^{(2)}_{m*q} = z^{(2)}_1, z^{(2)}_2, \ldots z^{(2)}_q$. After transforming through an activation function and adding another vector of bias term x_{m*1}=1, we get a whole new matrix of *m* samples *q* new features as $a^{(2)}_{m*(q+1)} = a^{(2)}_0, a^{(2)}_1, \ldots a^{(2)}_q$. Then the second random set of thetas $\Theta^{(2)}_{(q+1),K}$ will be initialized and multiplied to $a^{(2)}_{m*(q+1)}$ to form $z^{(3)}_{m*k} = z^{(3)}_1, z^{(3)}_2, \ldots z^{(3)}_k$. After putting $z^{(3)}_{m*k}$ through the activation function transformation again, the final output will be a *m* samples *K* possibilities matrix. The whole process is known as the Feedforward process (FF). For a specific sample, the highest probability among *K* values indicates that this sample is highly likely to be the correspondent class. Sigmoid function $\frac{1}{1+e^{-\alpha*\theta^T x}}$ is the most commonly used activation function in application. The black line represents a basic sigmoid function. The red and blue lines show how the function behaves with different choices of α (**Figure 4.15**).

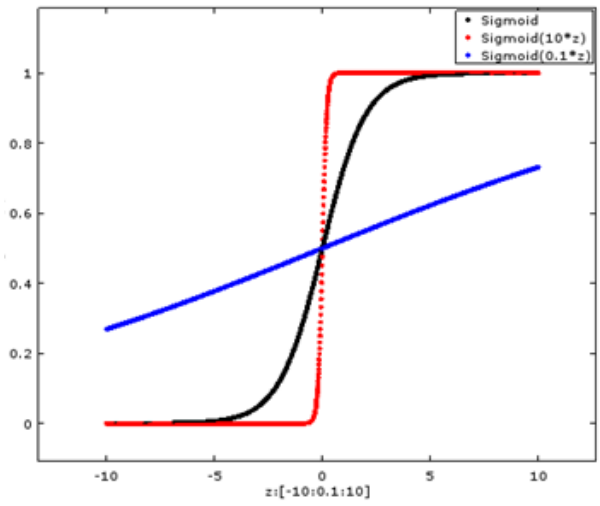


Figure 4.15 The Sigmoid transformation function

By making guesses of weights and going through several times activation transformation, the estimated outputs $y^{est} = \frac{1}{1+e^{-\alpha * \theta^T x}}$ are calculated and compared with the true outputs to determine how far away these estimations are. Each sample's activation values $a_K^{(3)}$ fall into the range of 0 to 1, with the log transformation, these values are mapped into 0 to positive infinite, depending on how close the estimated output is comparing to the true output. For instance, if the original class is 1 and the activation vector's first value is closer to 1, then the first part of the cost function is close to 0, the second part is minimal too, therefore the sum of the error costs will be very small since the estimation has a high tendency to be the true output. Here we work on two cost functions as follows, and *m* as the sample size, *L* as the layer index, f_{in} as the number of input items and f_{out} as the number of the output items:

Cost function without regularization:

$$J(\Theta) = -\frac{1}{m}[\sum_{i=1}^{m}\sum_{k=1}^{K}(y_k^{(i)}\log(h_\Theta(x)^{(i)})_k + (1-y_k^{(i)})\log(1-h_\Theta(x)^{(i)})_k] \quad (4.4.1)$$

Cost function with regularization:

$$J(\Theta) = -\frac{1}{m}[\sum_{i=1}^{m}\sum_{k=1}^{K}(y_k^{(i)}\log(h_\Theta(x)^{(i)})_k + (1-y_k^{(i)})\log(1-h_\Theta(x)^{(i)})_k] + \frac{\lambda}{2m}\sum_{l=1}^{L}\sum_{i=1}^{f_{out}}\sum_{j=1}^{f_{in}}(\Theta_{ji}^{(l)})^2 \quad (4.4.2)$$

Two sets of thetas are initialized through a random initialization process. The purpose of doing this is by randomly selecting $\Theta_{i,j}^{(layer)}$ in the range of $[-\varepsilon, +\varepsilon]$ through $\varepsilon = \frac{\sqrt{6}}{\sqrt{L_{input} + L_{output}}}$, the FF process can start with small parameters for an efficient learning process. For all samples, we perform the FF to compute all activations $a^{(l)}$ for the hidden layer. After the initial FF process, Backpropagation (BP) is implemented to adjust weights. Since after the hidden layer transformation, the cost of errors is equal to $\delta^{(OutputLayer)} = a^{(OutputLayer)} - y$. We then can use the sigmoid function's gradient g' and start with $\Delta_{ij}^{(l)} = 0$ for all layers. For hidden units, the error terms $\delta_j^{(2)}$ are based on a weighted average of the errors and activation values. The total hidden layer's error is equal to $\delta^{(2)} = \delta^{(3)} * \Theta^{(2)} * g'(z^{(2)})$, and the general form of errors formation is $\delta_j^{(l)} = \delta_1^{(l+1)} * \Theta_{1j}^{(l)} + \delta_2^{(l+1)} * \Theta_{2j}^{(l)} + \ldots + \delta_k^{(l+1)} * \Theta_{kj}^{(l)}$. Ultimately, the gradients can be calculated through weighted accumulated previous error terms according to $\frac{\delta J(\Theta)}{\delta \Theta_{ij}^{(l)}} = \Delta_{ij}^{(l)} + a_j^{(l)}\delta_i^{(l+1)}$ (Ng A. , 2013) to help the cost function to locate the minimization point along the iterative process. Here, we use Carl Edward Rasmussen's 'fmincg.m' to minimize the cost function along the calculated gradients. The starting

point is based on the initialized random weights, the cost function is designed to return the cost of errors and gradient. And the function uses Polak-Ribiere flavor of conjugate gradients (Nocedal, 1992), a line search and the Wolfe-Powell stopping criteria (Sun & Yuan, 2006) to guess step sizes and search for the solution.

Table 4.16 A flowchart for a three-layer NN

Input :

X : the training set; x : the testing set; Y : the training labels; $no.$: the number of nodes; i : the iteration time; λ :the regularization term; K : the number of classes; n : the number of features;

Random weights initialization:

1. $\Theta^{(1)}$: randWeightsInitial($n+1, no.$)
2. $\Theta^{(2)}$: randWeightsInitial($no.+1, K$)
3. $a^{(1)} = [X_0, X]$
4. $z^{(2)} = a^{(1)} * \Theta^{(1)}$
5. $a^{(2)} = [X_0, \text{sigmoid}(z^{(2)})]$
6. $z^{(3)} = a^{(2)} * \Theta^{(2)}$
7. $a^{(3)} = \text{sigmoid}(z^{(3)})$

Loop 1:

8. for i = 1…m
9. $J(\Theta) = J(\Theta) + \frac{1}{m}(-y * \log(a_i^{(3)}) - (1-y) * \log(1-a_i^{(3)}))$
10. end

Regularized Cost Function:

11. $J(\Theta) = J(\Theta) + \frac{\lambda}{2m} \sum_{l=1}^{L} \sum_{i=1}^{f_{out}} \sum_{j=1}^{f_{in}} (\Theta_{ji}^{(l)})^2$

Backpropagation:

Loop 2:

12. for i = 1…m
13. $\delta^{(3)} = a^{(3)} - y$
14. $\delta^{(2)} = \Theta^{(2)} * \delta^{(2)} * sigmoid'(z^{(2)})$
15. $\Delta^{(1)} = \Delta^{(1)} + \delta^{(2)} * a^{(1)}$
16. $\Delta^{(2)} = \Delta^{(2)} + \delta^{(3)} * a^{(2)}$

17. end

Regularized Gradients:

18. $\Delta^{(1)} = \frac{1}{m}(\Delta^{(1)} + \lambda * (\Theta^{(1)}))$
19. $\Delta^{(2)} = \frac{1}{m}(\Delta^{(2)} + \lambda * (\Theta^{(2)}))$

Cost Function Minimization:

20. [$\Theta_t^{(1)}$ $\Theta_t^{(2)}$, cost] = fmincg(@ (t)NNCostFunction(t, X , x , Y , $no.$, n , i , λ , $\Theta^{(1)}$, $\Theta^{(2)}$, options);

Output:

21. $a^{(1)} = [x_0, x]$
22. $z^{(2)} = a^{(1)} * \Theta_t^{(1)}$
23. $a^{(2)} = [x_0, \text{sigmoid}(z^{(2)})]$
24. $z^{(3)} = a^{(2)} * \Theta_t^{(2)}$
25. $a^{(3)} = \text{sigmoid}(z^{(3)})$
26. $k = \max(a^{(3)})$

4.5 The Inputs

Any successful supervised learning process must have three components: the correct outputs, the critical inputs, and a representative model between them. With sufficient information and reasonable expectations, NN serves as a powerful tool to discover hidden and complex relationships that cannot be easily detected by simple methods. How important is the input selection to the overall classification accuracy? Taking the handwritten recognition dataset as an example, each sample has 400 attributes that are precious pixel values and can be clearly displayed on a gray scale map. With clear and complete inputs, this case's classification result is extremely accurate, which is around 90% on average.

On the contrary, to forecast a company stock price's movement, there could be an endless list of quantitative and qualitative factors to be chosen from. Researchers often rely on expertise and experience, using attributes from basic financial statistics, established financial ratios and various economic indicators as **Table 4.21** shows. In our case, we absorb information from all three categories but focus on evaluating a company's long-term prospect. A 10-year record with 30 critical financial attributes and 5 refined financial ratios should be sufficient to establish a consistent learning process with a reasonable goal. In short, the fundamental difference between our work and other NN applications in this area is that the majority of researchers attempt to predict fluctuating stock prices with different combinations of inputs and NNs, instead of focusing on a company's intrinsic value and stability.

Table 4.17 The list of 35 financial attributes (Source: StockPup.com)

Share Data	Balance Sheet	Income Statement	Market Data	Financial Metrics	Derived Attributes
Shares Shares split adjusted Split factor	Assets Liabilities Shareholders' equity Non-controlling interest Preferred equity Goodwill & intangibles Long-term debt	Revenue Earnings Earnings available for stockholders EPS basic EPS diluted Dividend per share	Price Price high Price low	Return on equity Return on asset Book value of equity per share P/B P/E Cumulative dividends per share Dividend payout ratio Long-term debt to equity ratio Equity to asset ratio Net margin Asset turnover	Shares split adjusted EPS Shares split adjusted Dividends Shares split adjusted Price Shares split adjusted price high Shares split adjusted price low

4.6 Data Collecting and Processing

In this section, we discuss data sources and quality, errors checking, filling missing information, calculating and reorganizing attributes. Common ways to address missing values are using interpolation, taking the average, or simply replacing it with consistent values. We have worked on three types of datasets in this dissertation:

1. Type 1: Easy to get datasets like the *S&P 500 Constituents*, which takes a snapshot of a large number of companies at a specific point in time. Those datasets are rich in quantity but poor in quality. Stock prices, market value, and the number of shares fluctuate frequently, so it is imprudent to use one-time only statistics to carry out a rigorous investment analysis. In addition, since not all S&P 500 companies entered into the list at the same time and stayed, the composition might be heterogeneous and lead to misunderstanding. In short, a single year Earnings per Share (EPS), Book Value per Share (BVPS), Dividends, and other financial attributes cannot be used to evaluate a company's long-term prospects, since the dataset fails to meet the prerequisite of consistency and stability.
2. Type 2: A dataset of 159 companies with 10-year records built up by combining existing data sources and manual input. This dataset is excellent in quality but not superior in quantity.
3. Type 3: These datasets are both good in quantity and quality but need extensive and careful cleaning, organizing, transforming, and merging processes as follows:

Step 1: Compiling data: The raw dataset includes 612 companies' quarterly financial statistics which consist of 30 attributes ranging from 5 to 20 years. The additional 5 attributes are established according to the adjusted share number. Each attribute has its own importance, but the intricate interrelationships among them are hard to illustrate thoroughly.

Step 2: Filtering out unqualified companies: many companies do not share the same record-keeping history because they either started or ended their businesses at a different time. For example, AABV (2013 – 2015) versus ABI (1994 - 2008). To achieve a controlled experimental setting, we only selected companies with records starting from 2000 to 2014, since they do not only have sufficient records for analysis but also went through the same economic cycle.

Step 4: Identifying and separating data: We identify and separate each company's statistics into set 1 (2000 - 2009) for classification and set 2 (2010 - 2014) for forecasting.

Step 5: Cleaning and organizing data: There are countless issues embedded in the raw dataset which cause troubles in practice. To name a few, missing information such as "None" or empty spaces have to be replaced with "0" for calculation. Most quarterly attributes can be summed and averaged to become the annual record, but others like revenue, earnings, EPS, and dividends have to go through different procedures. Also dubious and inaccurate data are filtered out or recalculated. For example, AMR showed a single year $533.23 dividend per share but with a less than $1 EPS. Some given indices were illogical too, such as the ABC's P/E ratios being 72.05 and 36.5 with negative EPS $-0.91

and $-0.06 respectively. Graham (1934) repeatedly warned analysts that deficits are quality issues that should not be used to generate positive P/E ratios.

Step 6: **Evaluation and labeling**: We focus on setting up different standards to evaluate and establish classification indices. The most common one is to differentiate companies by price movement as many researchers did before. We do consider price differences as an important factor. However, with the understanding of investment principals, we believe that companies qualified for long-term investment must also keep consistent earnings and dividends records.

4.7 The Output Labeling Process

From our perspective, the output is the least discussed but the most important topic among existing NN applications. Output directs the modeling process and deserves to be treated with an extra caution. Past literature review indicates that researchers are more concerned with the choice of inputs than the formation of the output. For instance, in most work they simply use one standard which is the change of prices to differentiate stocks with NN (Hadavandi, Shavandi, & Ghanbari, 2010) (Ticknor, 2013). Some sophisticated works used smoothing techniques to limit the influence of unpredictable changes (Guresen, Kayakutlu, & Daim, 2011). In our case, by following the investment principles, the labeling process is more rigorous and reflects a company's long-term prospect. Ultimately, the underlying assumption is based on the concept of intrinsic stability that if a company has a consistent 8 to 10-year good earnings, dividends and prices records, it is more than likely to follow the same pattern for the next few years (Graham & Dodd, 1934).

4.7.1 A Test Run

For a test run with 232 companies, using a desktop with 8.00 GB RAM and Intel(R) Core(TM) i5-6500 CPU @ 3.20GHz processor, it takes about 176. 5 seconds to combine all the workbooks and only 161 companies are qualified for further analysis. After segregating, cleaning, and evaluating the dataset, we classify those companies according to the basic investment principles and test it with different combinations of standards to establish classification indices.

1. We establish a standard that an investment grade company must incur no deficits, non-interrupted dividend payouts and price appreciation. According to the 2000 -2009 records, only 44 companies can be qualified as the investment grade companies. However, based on the later 2010 -2014 records, there are 82 companies that can be categorized into the investment class. According to Graham and Dodd (1934), records during an unprecedented economic crisis should be considered as an abnormality. With the previous investigation on the general economy and key industries, we believe that the 2008 -2009 records should be taken out. By ignoring the 2008-2009 records, the result shows a significant improvement in the overall classification accuracy.
2. We tried to relax the standard by allowing one deficit per company. However, the classification and prediction results did not improve and mistakes in both categories were increased.

3. We tried to update EPS, DPS and price based on the split factor, ignoring the number of deficits and disrupted dividends but focusing on the overall performance. Mistakes in both categories are larger than the initial setting.

Table 4.18 A comparison of different classification criteria

161 companies	Investment 2000 -2009	Investment 2010 -2014	Overall classification accuracy	Mistaken as investment	Mistaken as speculation
Three basic principles	44	82	68.94%	6	38
Allowing 1 deficit	50	82	70.19%	8	40
Ignoring 2008 - 2009's records	66	82	**72.67%**	14	30
Selected features based on shares split adjustment	93	105	66.46%	21	33

In short, by allowing one deficit or ignoring the whole abnormal period, we found many cases that demonstrate the importance of understanding the general economy and the specific industry. However, companies categorized into the investment group in this dissertation are not recommended to be bought at the current price, further investigation is indispensable.

4.7.2 The Overall Dataset

Under the same configuration, it combines 612 companies in 251.98 seconds and keeps 413 qualified companies. After testing with different criteria, we summarize the main findings in **Table 4.19:**

Table 4.19 A comparison of different criteria

413 companies	Criteria 1	Criteria 2	Criteria 3	Criteria 4	Criteria 5
Specific Conditions	(2000-2008) vs. (2010-2014) Averaged EPS >0 averaged dividends >0, shares split adjusted average price appreciation > 0	(2000-2009) vs. (2010-2014) no deficits, consistent dividends, price appreciation	(2000-2008) vs. (2010-2014) no deficits, consistent dividends, price appreciation	(2000-2007) vs. (2010-2014) no deficits, consistent dividends, price appreciation	(2000-2006) vs. (2010-2014) no deficits, consistent dividends, price appreciation
Period 1 Investment	242	105	133	164	175
Period 2 Investment	313	243	243	243	243
Mistaken as Investment	34	10	16	27	38
Mistaken as speculation	105	148	126	106	106
The overall classification accuracy	66.34%	61.74%	65.62%	**67.8%**	65.13%

Table 4.18 shows that ruling out the 2008-2009 abnormal records did help to achieve a better overall classification accuracy. Moreover, we expect NN to uncover hidden relationships and critical attributes to improve the prediction accuracy.

4.7.3 Case Analysis

Investment Type 1 (Typical): Companies in this category exhibit stable EPS, dividends, and price records from 2000 to 2009 and continue the same performance till 2014. Examples are BEN, FII, EQT, CVX, BMS and many more. The requirement of an established dividends record helps to filter out companies with a relatively short

history. The only concern right now is that these companies may have a greater price appreciation far above its earnings growth.

Table 4.20 Investment type 1

EQT	EPS	DPS	Price	EMR	EPS	DPS	Price	BMS	EPS	DPS	Price
2014	2.6	0.1	96.09	**2014**	3.2	1.8	65.0	**2014**	1.9	1.1	39.9
2013	2.6	0.1	77.41	**2013**	2.8	1.7	60.0	**2013**	2.1	1.0	39.1
2012	1.2	0.9	54.04	**2012**	2.8	1.6	49.4	**2012**	1.7	1.0	31.5
2011	3.2	0.9	53.70	**2011**	3.2	1.4	52.8	**2011**	1.7	1.0	31.3
2010	1.6	0.9	40.26	**2010**	2.9	1.4	49.5	**2010**	1.8	0.9	29.7
2009	1.2	0.9	37.23	**2009**	2.3	1.3	35.5	**2009**	1.4	0.9	24.7
2008	2.1	0.9	51.21	**2008**	2.9	1.2	46.1	**2008**	1.7	0.9	25.0
2007	2.1	0.9	49.84	**2007**	2.8	1.1	48.3	**2007**	1.7	0.8	31.7
2006	1.8	0.9	36.38	**2006**	4.0	1.6	41.1	**2006**	1.7	0.8	31.7
2005	3.3	1.2	33.84	**2005**	3.7	1.7	33.9	**2005**	1.5	0.7	27.8
2004	4.5	1.4	25.11	**2004**	3.1	1.6	31.4	**2004**	1.7	0.6	26.4
2003	2.7	1.0	19.65	**2003**	2.7	1.6	26.4	**2003**	2.4	1.0	22.8
2002	2.4	0.7	16.89	**2002**	0.2	1.6	26.3	**2002**	3.1	1.0	25.2
2001	3.4	0.8	16.36	**2001**	2.2	1.5	30.1	**2001**	2.7	1.0	19.4
2000	3.2	1.2	12.68	**2000**	3.4	1.5	30.6	**2000**	2.5	1.0	16.4

Investment Type 2 (The Macro-abnormality): The 2008 -2009 financial crisis hit hard on the overall economy and various businesses. However, some companies improve their businesses quickly after the “abnormal period”. Examples are FDX, BA, CTAS and more. These companies had a rare deficit or a significant price depreciation

or both only between 2008 to 2009. Nowadays, the only concern is these companies become too popular to justify their earnings growth.

Table 4.21 The abnormality

413 Companies	% of in Deficits	% of Prices Reduction
2007	8.95%	18.9%
2008	18.4%	**74.8%**
2009	19.9%	**91.5%**
2010	8.23%	9.69%

Table 4.22 Investment type 2

BA	**EPS**	**DPS**	**Price**	**COF**	**EPS**	**DPS**	**Price**	**CTAS**	**EPS**	**DPS**	**Price**
2014	7.4	3.1	128.0	**2014**	7.7	1.2	78.2	**2014**	3.7	1.7	62.9
2013	6.0	2.2	102.8	**2013**	7.0	1.0	63.4	**2013**	2.6	0.8	46.7
2012	5.1	1.8	73.1	**2012**	6.4	0.2	54.3	**2012**	2.4	0.6	38.4
2011	5.3	1.7	69.4	**2011**	6.9	0.2	47.2	**2011**	2.1	0.5	29.7
2010	4.5	1.7	66.1	**2010**	6.0	0.2	40.5	**2010**	1.5	1.0	26.5
2009	1.8	1.7	45.2	**2009**	0.7	0.5	27.5	**2009**	1.2	0.5	24.7
2008	3.5	1.5	66.4	**2008**	-0.1	1.5	43.9	**2008**	2.1	0.5	29.0
2007	5.1	1.6	94.6	**2007**	3.6	0.1	71.3	**2007**	2.2	0.4	38.1
2006	2.8	1.3	79.7	**2006**	7.6	0.1	81.6	**2006**	2.1	0.4	41.0
2005	3.1	1.1	62.3	**2005**	6.8	0.1	78.9	**2005**	1.8	0.3	42.2
2004	2.2	0.9	47.5	**2004**	6.7	0.1	71.0	**2004**	1.7	0.3	44.4
2003	0.9	0.7	33.7	**2003**	5.2	0.1	46.9	**2003**	1.5	0.3	39.0
2002	0.6	0.7	40.4	**2002**	4.2	0.1	46.7	**2002**	1.4	0.3	22.9
2001	3.3	0.7	49.7	**2001**	3.1	0.1	55.2	**2001**	1.3	0.2	21.6
2000	2.3	0.6	48.7	**2000**	2.5	0.1	51.2	**2000**	1.2	0.2	38.6

Speculative Type 1 (Typical): Companies in this category show consistently unstable earnings, dividends, and prices. These companies are often in cyclical industries, such as mining, oil and gas, or fashion, exhibiting bizarre EPS, dividend payouts, and price records. For example, **ANF,** the clothing retailer, had few good records from 2006 to 2008, which was fueled by the housing bubble. After the bubble, ANF failed to move back to its original track. **APC**, the petroleum company, demonstrates how cyclical the oil and gas industry can be. When the oil price hiked to $140 around 2005 to 2008, APC made significant profits. After that, facing a dipping oil price, **APC** failed to balance its budget. Similarly, **CLF**, the mining company, a single year deficit -$11.88 (2014) almost wiped out the previous 10 years' profits and its price dropped from $61.3 (2010) to $15.6 (2014). Companies in a cyclical industry exhibit great speculative potential, speculators usually buy it at a low price and wait for the next unpredictable industry boom. A continuous dividend payout requirement also helps us to rule out companies that only generate negligible profits. For instance, **ADBE** had to stop its dividends payout in 2006 after serious earnings deterioration.

Table 4.23 Speculative type 1

FCX	EPS	DPS	Price	A	EPS	DPS	Price	CVC	EPS	DPS	Price
2014	-0.32	0.31	32.8	**2014**	-0.5	0.0	3.8	**2014**	1.2	0.6	18.1
2013	0.67	0.56	32.4	**2013**	-0.1	0.0	3.4	**2013**	1.7	0.6	16.3
2012	0.80	0.31	38.0	**2012**	-1.6	0.0	5.2	**2012**	0.9	0.6	14.8
2011	1.20	0.38	46.4	**2011**	0.6	0.0	7.0	**2011**	1.0	14.3	26.7
2010	1.92	0.38	40.9	**2010**	0.7	0.0	8.0	**2010**	1.2	0.5	26.5
2009	1.52	0.04	26.8	**2009**	0.4	0.0	*4.7*	**2009**	1.0	0.4	19.3
2008	-7.31	0.34	39.8	**2008**	-5.1	0.0	5.6	**2008**	-0.8	0.1	23.4
2007	2.27	0.34	40.6	**2007**	-6.3	0.0	14.0	**2007**	0.7	0.0	31.7
2006	1.87	1.19	28.0	**2006**	-0.2	0.0	27.9	**2006**	-0.5	10.0	24.1
2005	1.32	0.63	21.0	**2005**	0.4	0.0	20.1	**2005**	0.3	0.0	28.8
2004	0.22	0.28	18.6	**2004**	0.3	0.0	15.8	**2004**	-2.3	0.0	21.7
2003	0.25	0.07	13.4	**2003**	-0.8	0.0	9.4	**2003**	-1.0	0.0	20.0
2002	0.22	0.00	7.84	**2002**	-3.8	0.0	10.7	**2002**	-0.5	0.0	20.3
2001	0.13	0.00	6.13	**2001**	-0.2	0.0	20.2	**2001**	3.7	0.0	56.3
2000	0.07	0.00	5.63	**2000**	4.1	0.0	28.9	**2000**	1.4	0.0	69.8

Speculative Type 2 (Highly speculative): Companies in this category exhibit the same pattern: years of deficit, zero or close to none dividends and highly fluctuating price records. Such as **AMSC, ATI, AMR, AZO** and more.

Table 4.24 Speculative type 2

AMR	EPS	DPS	Price	AMZ	EPS	DPS	Price	AMS	EPS	DPS	Price
2014	1.00	0.05	37.9	**2014**	-0.5	0.0	333.3	**2014**	-0.9	0.0	15.9
2013	-2.26	0.00	6.40	**2013**	0.6	0.0	295.7	**2013**	-0.9	0.0	25.3
2012	-1.40	0.00	0.00	**2012**	-0.1	0.0	219.7	**2012**	-1.3	0.0	41.3
2011	-1.48	0.00	72.4	**2011**	1.4	0.0	196.1	**2011**	-6.2	0.0	132.1
2010	-0.35	0.00	113	**2010**	2.6	0.0	139.2	**2010**	0.1	0.0	315.8
2009	-1.25	0.00	94.7	**2009**	2.1	0.0	85.9	**2009**	0.3	0.0	261.9
2008	-2.18	0.00	141	**2008**	1.5	0.0	71.7	**2008**	-0.5	0.0	254.6
2007	0.52	0.00	407	**2007**	1.2	0.0	66.5	**2007**	-1.0	0.0	196.1
2006	0.29	0.00	378	**2006**	0.5	0.0	36.5	**2006**	-1.0	0.0	96.1
2005	-1.20	0.00	189	**2005**	0.9	0.0	39.2	**2005**	-0.9	0.0	101.9
2004	-1.19	0.00	164	**2004**	1.5	0.0	44.9	**2004**	-0.6	0.0	131.5
2003	-1.98	0.00	130	**2003**	0.1	0.0	37.7	**2003**	-3.6	0.0	76.9
2002	-5.67	0.00	230	**2002**	-0.4	0.0	16.3	**2002**	-3.0	0.0	59.8
2001	-2.86	0.00	453	**2001**	-1.6	0.0	12.4	**2001**	-1.8	0.0	177.9
2000	1.39	8.93	540	**2000**	-4.1	0.0	48.0	**2000**	-0.9	0.0	427.1

Unpredictable changes: Some companies have a very recent business deterioration or improvement that triggers a correspondent price depreciation or appreciation.

Table 4.25 Samples of unpredictable changes

FE	EPS	DPS	Price	ARG	EPS	DPS	Price
2014	0.18	0.36	33.36	**2014**	5.0	2.1	108.3
2013	0.24	0.41	38.60	**2013**	4.7	1.8	101.9
2012	0.46	0.55	45.16	**2012**	4.5	1.5	84.5
2011	0.54	0.55	41.52	**2011**	3.7	1.2	66.9
2010	0.65	0.55	38.45	**2010**	2.8	0.9	61.8
2009	0.83	0.55	42.71	**2009**	2.6	0.7	41.3
2008	1.10	0.55	68.82	**2008**	3.3	0.5	47.6
2007	1.07	0.50	65.72	**2007**	2.5	0.3	45.9
2006	0.95	0.45	53.92	**2006**	1.9	0.3	37.2
2005	0.66	0.43	46.17	**2005**	1.5	0.2	26.5
2004	0.67	0.38	39.01	**2004**	1.2	0.2	23.1
2003	0.37	0.38	32.98	**2003**	1.1	0.1	18.6
2002	0.47	0.38	31.91	**2002**	-0.01	0.0	16.2
2001	0.68	0.38	31.45	**2001**	0.6	0.0	11.1
2000	0.66	0.38	24.51	**2000**	0.4	0.0	6.9

In sum, several findings are consistent with the previous literature review in investment analysis. First, great purchasing opportunities always surface in the economic downturn for the first-class companies that maintain high quality through all kinds of crises and cyclical events. In a broad sense, companies with established stable EPS, DPS, and price records should be taken into consideration when temporary adversaries hit them. Second, utilities companies in general are more than likely to be qualified as investment purchases. Such as **D** (Dominion Resources) and **FPL** (Florida

Power & Light Group), each company's 10 year dividends payout already exceeds its initial stock purchasing price let alone the stock price appreciation.

Table 4.26 Samples of utilities companies

D	EPS	DPS	Price	FPL	EPS	DPS	Price	ED	EPS	DPS	Price
2014	2.3	2.4	69.9	**2014**	5.7	2.9	96.4	**2014**	3.7	2.5	57.3
2013	2.9	2.3	59.3	**2013**	4.5	2.6	80.1	**2013**	3.6	2.5	58.0
2012	0.5	2.1	52.2	**2012**	4.6	2.4	65.8	**2012**	3.9	2.4	59.9
2011	2.5	2.0	47.3	**2011**	4.6	2.2	55.0	**2011**	3.6	2.4	53.6
2010	4.8	1.8	41.0	**2010**	4.8	2.0	51.4	**2010**	3.6	2.4	45.7
2009	2.2	1.8	33.4	**2009**	4.0	1.9	52.9	**2009**	3.2	2.4	39.2
2008	3.2	1.6	42.7	**2008**	4.1	1.8	58.4	**2008**	4.4	2.3	41.6
2007	7.4	2.5	43.7	**2007**	3.3	1.6	61.4	**2007**	3.5	2.3	47.9
2006	3.9	2.8	38.2	**2006**	3.3	1.5	43.6	**2006**	3.0	2.3	45.5
2005	3.1	2.7	37.9	**2005**	2.3	1.4	42.0	**2005**	3.0	2.3	45.3
2004	3.9	2.6	32.0	**2004**	5.0	2.6	33.5	**2004**	2.3	2.3	42.3
2003	1.1	2.6	29.8	**2003**	5.0	2.4	31.2	**2003**	2.4	2.2	41.2
2002	5.2	2.6	28.5	**2002**	2.7	2.3	28.0	**2002**	3.0	2.2	41.1
2001	2.3	2.6	30.6	**2001**	4.6	2.2	29.1	**2001**	3.2	2.2	38.1
2000	1.9	2.6	24.1	**2000**	4.1	2.2	26.5	**2000**	2.8	2.2	33.1

4.8 The Two-stage NN Training and Testing Process

After transforming 613 individual workbooks into two datasets of 413 companies with the number of features based on the number of year records and putting it through the output labeling process, we began the two-stage NN training and testing process. Dataset 1 consists of 413 companies, 350 features, and 2 classes. Dataset 2 consists of 413 companies, 175 features, and 2 classes. Each company in each dataset is labeled as 0 (Speculative Class) or 1 (Investment Class), which later has been transformed into a

vector of 2 outputs with values either 0 or 1. Since the correspondent estimated output is a vector of 2 values with values ranging from 0 to 1, so the final estimated class label is determined by which value is higher and compared to the original label for classification accuracy. Stage 1 searches for a representative NN for 2000-2009 with classification labels built in its 10 years records, while stage 2 is for forecasting which identifies links between the past 10 years' records and the next 5 years' performance through NN.

4.8.1 A Three Layer NN with the Sigmoid Transformation Function

Table 4.27 The NN Feedforward process

1st layer: 413 examples, 350 features, 2 classes, adding a bias term to form the input layer. $X = \begin{bmatrix} x_{1,0} & \cdots & x_{1,350} \\ \vdots & \vdots & \vdots \\ x_{413,0} & \cdots & x_{413,350} \end{bmatrix}_{413 samples * 351 features}$
2nd layer: Theta1 correspondent to the hidden layer units: $X_{413*(350+1)} \times \Theta^{(1)}_{q*(350+1)}{}'$ and the activation function transforms the inputs matrix into a *413* samples with a *q nodes* matrix $a^{(2)}_{413*q} = sigmoid(X_{413*(350+1)} \times \Theta^{(1)}_{q*(350+1)}{}')$, then add a bias term as $a^{(2)}_{413*(q+1)}$
3rd layer: Theta2 and activation function transforms hidden layers inputs into $a^{(3)}_{413*2} = sigmoid(a^{(2)}_{413*(q+1)} * \Theta^{(2)}_{2*(q+1)}{}')$ and compares to the original output for cost of errors summation purpose.

4.8.2 Feature Scaling

Feature scaling techniques are often applied to speed up the converging process and limit the influence fro non-critical or dominant features. The downsides are possible information loss and over generalization. There are four common scaling techniques:

Differencing is used to eliminate commonality (Ramsey, 1999); Log transformation is used to smooth the distribution and transform derived ratio and polynomial features into simple subtraction/addition ones (Gelman, 2008); Standardization turns each feature into a data set with zero mean and unit variance (Shanker, Hu, & Hung, 1996); Normalization transforms each feature into a data set with values ranging from 0 to 1 (Quackenbush, 2002). We also test on a less commonly used technique by scaling the dataset with the sigmoid function, which transforms all features into values between 0 to 1 (Reed, Marks, & Oh, 1995). All testing sets are scaled differently from the training sets since the underlying assumption is that the new entering samples are unknown. As **Figure 4.16** shows, each graph contains a group of 100 randomly selected companies with 350 features, which are mapped into 100 blocks with a height of 14 features and width of 25 features. We found out that the non-scaling dataset, standardized dataset and sigmoid transformed data display the similar pattern while normalized dataset disguises the most heterogeneity. In short, it is uneasy to detect and establish meaningful relationships by looking at a large number of data, and different data scaling methods have different impacts on the classification process.

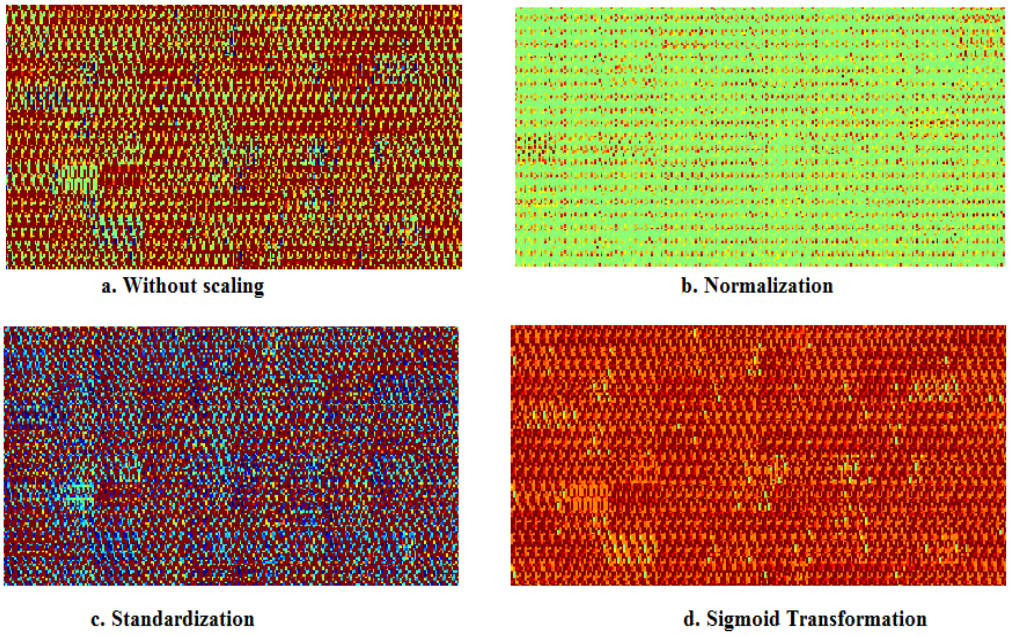


Figure 4.16 With/without feature scaling

4.8.3 The Choices of Nodes and Activation Functions

Different combinations of nodes, activation functions, and layers give researchers an infinite choice of NN architectures. The number of nodes affects the learning speed since each node has to be paired with a parameter and each parameter has to go through a random initialization and FF-BP processes (Kaastra & Boyd, 1996). Adding more nodes might be helpful in a complicated case, but it also significantly slows down the process (Baum & Haussler, 1989). So how many nodes should we choose in each layer? Theoretically, the right number of nodes with just one hidden layer can be used to approximate many sophisticated models (Hornik, Stinchcombe, & White, 1989). NNs with more than two hidden layers have not been proved to improve performance significantly other than consuming more computational efforts. Therefore, an NN with one hidden layer is often the first and best choice. Moreover, the possibility of over-

fitting will increase when a model is too complicated to fit the training set and loses its ability to generalize. According to Kaastra and Boyd (1996), the training set size and the number of nodes and layers jointly determine the over-fitting tendency, so we test our model with different number of nodes as **Table 4.28** shows. Although there are many choices of activation functions, such as linear, hyperbolic, step functions, the sigmoid function appears to be the top choice due to its simplicity and transformation ability.

Table 4.28 The number of nodes tested based on the rules of thumb

The Rule of Thumb	The Number of Nodes
(Bailey & Thompson, 1990)	150
(Ersoy, 1990)	5, 10, 20, 40…
(Klimasauskas, 1993)	40
(Masters, 1993)	28
(Ng A. , 2013)	25

4.8.4 Feature Selection and Cross Validation

Stage 2 NN mimics the situation when we have a limited understanding of the underlying connections between the inputs (the past 10 year records) and the outputs (the next 5-year performance). From previous analysis, we perceive that the three principles might not be sophisticated enough for us to make a good prediction. Therefore, at stage 2, we rely on NN with feature selection methods to search for potential critical attributes other than EPS, price, and dividend. For example, random inputs selection refers to forming and testing with extensive combinations of inputs. If the computational cost is not an issue or the size of inputs is small, then this approach is

feasible in uncovering hidden relationships (Kaastra & Boyd, 1996). On the contrary, trying different combinations of hundreds and thousands of features can be exhaustive. The most intriguing FS method is Setiono and Liu (1997)'s network pruning algorithm which is similar to regularization by adding a penalty item into the cost function and excluding attributes which made the least contribution to the overall process. Furthermore, Guyon and Elisseeff (2003) discuss variable ranking, correlation detection, wrappers and embedded methods to suit different needs. Since having a large group of features, we adopt several established and refined FS techniques from a very recent work (Pohjalainen, Räsänen, & Kadioglu, 2015). Since cross-validation (CV) has been widely accepted as an evaluation method to curb over-fitting and enhance model's generalization capability (Krogh & Vedelsby, 1995) (Setiono, 2001), so a 5-fold CV is adopted here for all experiments.

4.9 Computational Results

4.9.1 Stage 1 NN With and Without Feature Scaling

Stage 1's computational results mainly focus on different combinations of NN and feature scaling approaches, the number of nodes, with or without regularizations, and before and after feature scaling are evaluated based on efficiency and accuracy measures. The computational time mainly depends on the sample size, iteration times the number of nodes, and the randomly initiated start points. Here we set 1000 iterations as the up limit. The selection of 18 key features is based on the previous output labeling process, including 2000-2007 EPS, 2000-2007 Dividends records, and prices records of 2000 and 2007.

Table 4.29 Stage 1 350 features with/without regularization and different nodes - No Scaling

Nodes	Lambda = 0	Lambda = 100
5	57.50%	60.00%
25	60.25%	60.00%
50	54.75%	60.00%
100	60.75%	60.00%
Average Computational Time	30~90 minutes	15~30 minutes

Table 4.30 Stage 1 350 features with/without regularization and different nodes - Normalization

Nodes	Lambda = 0	Lambda = 100
5	58.50%	60.00%
25	57.75%	60.00%
50	50.00%	60.00%
100	52.00%	60.00%
Average Computational Time	60 minutes	10~20 minutes

Table 4.31 Stage 1 350 features with/without regularization and different nodes - Standardization

Nodes	Lambda = 0	Lambda = 100
5	61.22%	59.02%
25	66.59%	59.51%
50	60.73%	60.24%
100	59.27%	59.51%
Average Computational Time	50 minutes	40 minutes

Table 4.32 Stage 1 350 features with/without regularization and different nodes - Sigmoid-transformation

Nodes	Lambda = 0	Lambda = 100
5	58.54%	60.24%
25	58.78%	60.24%
50	58.78%	60.24%
100	58.78%	60.24%
Average Computational Time	30 minutes	30 minutes

Table 4.33 Stage 1 Selected features without regularization

Nodes	Non-Scaling 18 key features	Standardization 18 key features	Sigmoid Transformation
25	82.25%	75.61%	51.20%

Without scaling, the iteration runs slower and takes longer to converge. Being inclusive with all features and under regularization, sigmoid transformation seems to outperform the rest largely due to its consistence. Considering only 18 critical features, the non-scaling NN performs the best since it depends only on and preserves all critical information. Standardization ranks the second due to its close to top performance and less computational time. A group of critical features does not only significantly improve the learning process but also the NN's classification ability. Therefore, we conclude that having only critical features is the key to NN success, which increases both the running efficiency and accuracy by reducing the distractions from uncritical information. The regularization term exhibits certain influence on the results by curbing the overfitting tendency. The larger the lambda, the less the tendency to over fit the training set, which can be seen from the consistent and better NN performance after regularization. Regularization NN in general shows a consistent performance if critical features were unidentifiable. However, regularized NNs still underperform greatly comparing to NNs with only critical features. Last but not the least, the choice of the number of nodes depends on the situation. In most cases, 25 nodes are reasonable based on the size of our inputs and output.

4.9.2 Stage 2 Classification Algorithms Comparison With/Without Feature Selection

In this section, we investigate several Feature Selection (FS) methods that choose features either by joint occurrence probability, inter-feature correlation or features' influence on labels. Then we combine and compare different classifiers (NN/KNN/LogR) with and without these FS methods. An ideal FS is a well-designed process to facilitate classification/prediction by eliminating unnecessary attributes in order to reduce high computational demand, improve the results and avoid over-fitting. A typical FS consists of:

1. An approach picks a new feature or a new set of features. For instance, an algorithm can start with a random or a pre-determined choice.
2. Criteria evaluate the feature or the subset, for instance, Akaike information criteria (Symonds & Moussalli, 2011) and Bayesian information criterion (Chen & Gopalakrishnan, 1998).
3. A tipping point that is the max iteration time or a threshold halts the iterative process.

There are several types of FS: wrappers, filters, and embedded methods. Wrappers are methods that use a classifier to validate the proposed subsets by picking a set of features fits well for the specified classifier. It can be very time-consuming and highly dependent on the classifier, causing over-fitting (Kohavi & John, 1997). Filters methods look into the dataset, focusing on the relationship among features and classes, instead of trying to fit for a specific model/classifier. The advantages of using filters are the fast process and stable solutions. The shortcoming is that instead of taking advantage of a more sophisticated classification algorithm, filters are more independent from the

choice of classifier. Common filters are Mutual Information (Peng, Long, & Ding, 2005), Pearson Correlation and Distance Based Discriminate Analysis (Hall, 1999). Embedded methods integrate the feature selection and modeling process into a one-step process. Examples are regularization (LASSO) which keeps all features but penalizes everyone (Ng A. Y., 2004) and Recursive Feature Elimination (Guyon, J., S., & Vapnik, 2002). In turns of the complexity, exhaustive FS is the simplest yet the most likely infeasible one if the feature size is too large. Therefore, researchers have invented different stopping criteria for exhaustive FS, such as the feature size fluctuation threshold and the max iteration time. The most common FS methods are the Step-forward/backward (Liu & Setiono, 1995). A more advanced and complicated one is the Minimum-Redundancy-Maximum-Relevance (mRMR), which takes the consideration of the relevance between the feature and the output and the redundancy among features into feature selecting process (Ding & Peng, 2005). There are numerous research possibilities by combining different classifiers, FSs and evaluating them based on the efficiency and accuracy measures. Here, we select several filters and wrappers combined with three different classifiers for evaluation.

Table 4.34 A comparison of the most comprehensive and the most recent feature selection toolbox

Source	Classifier	Feature Selection	Advantage
Alois Schloegl NaN-toolbox (2010)	Mahalanobis distance, Gaussian radial basis, Winnow Algorithm, Augmented Naïve Bayesian, Naïve Bayesian Parzen Window Perceptron Learning Algorithm	Pearson Partial Correlation(PPC), FSDD: a distance discriminate measure	Specializes in dealing with datasets that have NaN values in inputs or outputs in various formats. Many classifiers, 4 CV approaches can be chosen from Leave-One-Out-Method (LOOM); Leave-K-Out-Method; K-fold; Group-wise (Non-independent samples).
Pohjalainen et.al (2015)	KNN	MI, SD, RSFS, SFS, SFFS	Specializes in multiple FS methods
Zhang & Trafalis (2016)	LogR NN	Regularization, MI, SD, FSDD, PPC, RSFS,SFS, mRMR	Integrates LogR and NN with multiple FS methods

Mutual Information (MI) searches for the commonalities between features and outputs (Pohjalainen, Räsänen, & Kadioglu, 2015). First, MI establishes a matrix size as the number of features with desired blocks called the edges. Then for each feature, the algorithm identifies its max/min values to construct a stepwise interval called "Quant-levels". After categorizing each feature into different quant-level bins and all training samples into different blocks, the algorithm matches features with the block by the correspondent quant-levels. For instance, the number of times of features and training samples fall into the same block will be used to determine the possibility of the joint occurrence. And a probability per feature can be calculated based on the associated outputs. Several parameters can be adjusted here. Such as the number of quantization levels defines the dimension of the edge, the step size determines the space of the interval, and the lag accommodates the sequential difference. The benefit of using this

approach is the degree of freedom to choose the top N attributes but the downside is no easy way to claim for the best combinations of parameters. **Statistical Dependency (SD)** is very similar to MI, except using a quantized feature space to rank features by the statistical relevance (Pohjalainen, Räsänen, & Kadioglu, 2015). Therefore, we combined different quantization levels with top feature lists to facilitate NN to make predictions. Below results indicate that the categorical block number and the number of top attributes jointly determine the classification accuracy. Especially in this case, including more than 50 top features clearly shows downward pressure on NN's capability to generalize.

Table 4.35 A flowchart of MI and SD

	Mutual Information (MI)	**Statistical Dependency (SD)**
Inputs	Samples: M samples *n features Labels: M samples *1 Q : the number of bins Step size: M/Q Edges: n features* (Q+1)	Same
Core	Loop 1: For i = 1 to nth feature Define a N quant-levels interval, categorizing each feature (i) into that interval; Identify boundaries for each block and per feature which must contain equal amount of samples; End Loop2: For i = 1 to nth feature Replicate the vector feature(i) Q + 1 times and the row edges(i) M times to form two metrics size as M * (Q + 1). Compare those two and see each sample's feature(i)′s value in which block and mark it with the block number End	Same except for the MI calculation SD: $\sum_{i=1}^{q}\sum_{j=1}^{k}(M_{q*k} * (M_{q*k} / SP_{q*k})) - 1$

	Loop3: For i = 1 to nth feature For each block, how many features (i) fall into that block and calculate the probability of feature(i) per block P1: feature distribution probability P2: class distribution probability SP: P1 ∗ P2 individual state per feature per class probability M: Joint occurrence per feature per block per class probability MI: $\sum_{i=1}^{q}\sum_{j=1}^{k}(M_{q*k} * \log(M_{q*k} / SP_{q*k}))$ End	
Outputs	Features ranked by connections level and a correspondent weight vector.	Same

Table 4.36 Stage 2 a three-layer 25 nodes NN with MI

Q	Top 10	Top 30	Top 50	Top 100
3	75.25%	78.25%	68.00%	62.00%
12	79.75%	79.50%	69.50%	65.25%
15	80.00%	73.00%	70.25%	61.50%

Table 4.37 Stage 2 a three-layer 25 nodes NN with SD

Q	Top 10	Top 30	Top 50	Top 100
3	77.00%	77.50%	65.75%	64.50%
12	77.50%	71.25%	70.50%	63.75%
15	78.00%	71.50%	69.00%	64.75%

A Distance Discriminant Based Feature Selection (FSDD) ranks features based on its contribution in differentiating samples by taking consideration of per class per feature's mean and variation. A feature is considered as the most important one if its score is the lowest among all features (Liang, Yang, & Winstanley, 2008). FSDD with KNN runs fast and stable in this case. On the other hand, **Pearson Partial Correlation (PPC)** starts with a complete set of features and search for which are highly correlated with the output. The score assigned to each feature indicates the relevance of the feature to the output, the higher the score the better the rank (Rao & Lakshminarayanan, 2007).

However, according to our experiments, PPC runs less efficiently and generates less credible results in this case. For instance, if we increase the feature selection to Top 100, PPC yields a better classification result with KNN but the choice of features is far above other FSs' recommendations.

Table 4.38 A flowchart of FSDD and Pearson Correlation

Component	**FSDD**	**Pearson Correlation**
Input	Samples: M samples * n features Labels: M samples *1 T: the number of highest ranked features	Transform labels [1, 2 , 3…] into a binary format [1 0 0 ; 0 1 0, 0 0 1] for the matrix calculation
Core	Loop: For class 1 : k n_k: each class's sample size $mean_k$: each class's each feature's mean Var_k: each class's each feature's variance end $mean_0$: the weighted mean based on the entire sample Var_0: the weighted variance per feature based on the entire sample s^2: the difference between the mean of the weighted squared features and the squared weighted mean Score: the weighted difference of s^2 and the mean of weighted variance	Partial correlation calculation X = classes; Y = the subset of feature(s); Z = the excluded feature(s) Loop: X = X-Z*(Z\X); Y = Y-Z*(Z\Y); r = correlation between (X,Y); score = max(sum(r^2))
Output	Ranked feature indices and correspondent scores	Ranked feature indices and correspondent scores

Table 4.39 A comparison table of NN based on FSDD/PPC with different choices of top features

Method/Iteration	Top 10	Top 30	Top 45	Top 50	Top 100
FSDD_NN	60.25%	75.75%	75.25%	71.75%	63.75%
PPC_NN	59.75%	59.25%	60.25%	59.50%	64.25%

Random Subset Feature Selection (RSFS) RSFS is an exhaustive FS method by repeatedly going through a specified classifier with various randomly chosen feature subsets in order to identify a group of good features (Räsänen & Pohjalainen, 2013). Several things can be altered to form some new perspectives, for instance, the size of subset used per testing session, the performance criterion and the classifier. Pohjalainen, Räsänen and Kadioglu (2015) used the square root number of features per session, KNN as the classifier and Unweighted Average Recall (UAR) as the criteria. Here we introduce NN and propose an efficiency improvement method. KNN with RSFS already is an extremely time-consuming process since the default iteration setting for KNN_RSFS with 200 features is 300,000. Here, with an inner and an outside iteration requirement, a typical 5 CV NN_ RSFS can take days to converge. We also derive and test the **Mutual Information/Statistical Dependence Based Random Subset Feature Selection (NN_MI/SD_RSFS**) methods. The logic behind this approach is that from previous MI/SD results, we can clearly see that using more than top 50 attributes do not help the classification process. Therefore, instead of doing an incremental test by adding 1 feature each time from top 30 to top 50, we wrap all top 50 attributes and feed it into RSFS. Therefore, without taking all 350 features into the selection pool, we start with top 50 features that significantly reduce the size of feature pool and improve the running efficiency. However, both derivations do not exhibit any superiority in this case. In sum, NN_RSFS is a comprehensive but time-consuming process, we save it as a last resort. However, we found out that under a limited number of iteration, NN_RSFS slightly outperforms the KNN_RSFS. Future extension can be made on how different

choices of the subset size, classifier and performance measures affect the classification results.

Table 4.40 A flowchart of NN_RSFS and its derivatives

	NN_RSFS	NN_MI_RSFS	NN_SD_RSFS
Inputs	Training set: $m_{11} * n$ features Development set: $m_{12} * n$ I: Maximum iteration time $\boldsymbol{\Delta}$: feature set size fluctuation threshold $\boldsymbol{R}$: Relevance matrix 350 *1 $\boldsymbol{D}$: dummy relevance matrix 100*1 $\boldsymbol{f}$: size of the random subset $\boldsymbol{d}$: size of dummy subset	Using Mutual Information to narrow down the scope of features	Using Statistical Dependence to narrow down the scope of features
Core	Randomly choose a subset of features Apply NN and test on the development set Define the performance and expected performance measures. The first run is always the baseline performance measure here. Define the relevance measure per selected feature by whether this feature is better or worse in the set. Compare to the randomly selected dummy features, true features should show higher relevance. Update relevance vector: If the current subset performance better/worse than the previous performance, then the correspondent feature relevance will add in or subtract the difference Find features that are better than random dummy features and rank them by its relevance	Same	Same
Outputs	S: Selected features indices W: Selected features relevance values AW: All features relevance values	Same	Same

Table 4.41 Stage 2 three-layer 25 nodes NN with RSFS and embedded FS methods

Method/Iteration	200	2000	20000
NN_RSFS	66.75%	68.25%	64.75%
NN_RSFS_MI	58.50%	61.25%	60.15%
NN_RSFS_SD	58.78%	59.00%	57.10%

Other than the above methods, **Sequential Feature Selection (SFS)** is similar to RSFS but with the direction of choosing features sequentially and keep or exclude them depending on whether that feature makes a significant contribution to the output measured by Mean Squared Error (Regression) or classification accuracy (Classification). RSFS, SBFS and SFFS all three approaches tend to be infeasible if the feature size is too large.

4.9.3 The Ultimate Comparison among KNN, LogR and NN with/without Feature Selection

We present a compressive review and comparison in this session. **K-Nearest Neighbors (KNN)** method is the simplest yet widely adopted classification algorithm. The benefits of using KNN are the degree of freedom by choosing any *k* nearest training samples for estimation and adjusting the weight parameter for imbalanced data. The downside came from the high dimensionality curse since running KNN requires to build a large matrix based on the size of both testing and training sets. By calculating the distance between each testing sample and each training sample, *k* nearest neighbors are exposed after distance ranking. The final estimated label is determined by the majority votes system. KNN and KNN Incremental Distance Matrix (KNN-IDM) with 5 feature selection methods have been well-integrated by Pohjalainen, Räsänen and Kadioglu (2015). **Logistic Regression (LogR)** works as a processing center that uses the sigmoid

function to transform linear regression hypothesis results into probabilistic outputs ranging from 0 to 1. Assuming we have a feature vector with given parameters, what are the possibilities of this sample belongs to different classes and which class shows the highest interest in taking in this sample.

Table 4.42 A flowchart of KNN and LogR

	KNN	LogR
Inputs	Training set: m_1 samples * n features Training label: m_1 samples * 1 Testing set: m_2 samples * n features Testing label: m_2 samples * 1 *k*: the number of nearest neighbors *w*: weight adjustment for imbalance data *c*: the number of classes *D*: a $m_1 * m_2$ storage matrix for distance between training and testing *A*: a $m_2 * c$ storage matrix for voting	Training set: m_1 samples * n features Training label: m_1 samples * 1 Testing set: m_2 samples * n features Testing label: m_2 samples * 1 *I*: the maximum iteration time Adding a bias term for each train sample Initialize thetas for all features plus a bias term Initialize a vector of thetas n+1 Test sample + a bias term
Core	Loop 1: for i = 1 to m_2 Replicate each testing sample m_1 times by row then compare to the training set matrix to calculate the Euclidean Distance end Sort the *D* matrix from the closest to the farthest by row Take first *k* indices and match them with the actual label from the training set Loop 2: for i = 1 to m_2 Estimate testing label based on all nearest neighbors' labels by the majority voting end	Loop: Minimize the Cost Function $J(\theta_0,\theta_1,...,\theta_n)=\frac{1}{m}\sum_{i=1}^{m}Cost(h_\theta(x^{(i)}),y^{(i)})$ $Cost(h_\theta(x),y)=$ $\begin{cases}-\log(h_\theta(x)) & if\ y=1\\ -\log(1-h_\theta(x)) & if\ y=0\end{cases}$ Update Gradient Descent $\theta_j := \theta_j - \alpha\sum_{i=1}^{m}(h_\theta(x^{(i)})-y^{(i)})x_j^{(i)}$ Prediction $g(z)=g(\theta^T x)=\frac{1}{1+e^{-\theta^T x}}$ $\theta^T x \geq 0, h_\theta(x)\geq 0.5, y=1$ $\theta^T x < 0, h_\theta(x) < 0.5, y=0$ end
Output	Estimated label for each testing sample	Estimated label for each testing sample

Table 4.43 A selective comparison table of KNN, LogR and NN with/without feature selection

With/Without Feature Selection	Category	KNN	LogR	NN
No Feature Selection or Regularization	-	63.75%	58.75%	60.75%
Regularization Lambda = 100	Embedded	-	58.75%	65.50%
Mutual Information (Top 45)	Filter	69.50%	67.75%	79.50%
Statistical Dependency (Top 45)	Filter	61.75%	58.75%	78.00%
Partial Correlation (Top 45)	Filter	60.25%	59.00%	60.25%
Distance Based Discriminant (Top 45)	Filter	68.75%	59.00%	75.25%
Random Subset Feature Selection (Top 50)	Wrapper	71.05%	59.50%	68.25%
Mutual Information Based Random Subset Feature Selection (Top 50)	Filter + Wrapper	62.00%	60.00%	61.25%
Statistical Dependency Based Random Subset Feature Selection (Top 50)	Filter + Wrapper	57.75%	58.75%	59.00%

From the above table, we can clearly see that with any feature selection, all classification algorithms achieve an equal or higher classification accuracy rate than without it. In any case, NN outperforms LogR. In most cases, NN yields better results than KNN. However, NN requires greater computational power and has more parameters to adjust, while LogR only needs to worry about the length of the iteration and KNN only needs to consider different choices of neighbors. In turns of the overall processing time, LogR is the easiest and fastest approach among three classifiers. NN is the slowest due to the inside and outside loops. Surprisingly, using a filter MI/SD then the wrapper RSFS does not improve the classification results under a limited number of iterations. A Possible explanation might be narrowing down the scope with MI/SD

before using a wrapper also shrinks the inner loop subset size to start with. Results from previous tables indicate that feature selection is necessary since a large set of features does not always enhance the performance especially when many of them are not that critical. Last but not the least, NN with Mutual Information is highly recommended in dealing with similar decision-making problem.

4.10 Conclusions

NN is a powerful ML algorithm but its cognition ability largely depends on the quality of the inputs and the existence of a pattern. Very often, lack of critical information or fail to set up a clear and tangible goal undercuts NN's performance in practice. For instance, many researchers utilize NN to forecast stock prices, which is contradict to the reality that stock prices tend to be unpredictable and are snapshots of the market evaluation on companies. In addition, inputs are often predetermined and limited based on expert's experience. On the contrary, we focus on the stability in the learning process by following the fundamental investment principles. Although we cannot avoid the issue of insufficient information, the stability embedded in companies' long-term records and more strict labeling process help us to construct a better classification process. Furthermore, referring to the components of NN, we use the traditional Epsilon initiation process and sigmoid function, but there are many ways to initiate parameters and transform inputs. Therefore, tests on other parameters initialization process and activation functions might yield interesting results. In sum, our main purpose here is steering away from the traditional price forecasting route but focusing on the stability in investment analysis.

CHAPTER 5 CONCLUSIONS AND EXTENSION

This dissertation focuses on the ML algorithms application on real-world financial datasets. Main research contributions are as follows: First, by changing the structure of the objective cost function and combing with a points filling process, we successfully turn the linear regression into a classification method. Second, by setting and varying weights in front of different classes' cost of errors, we demonstrate that the classification accuracy can be improved if there is an imbalance tendency in the dataset. Thirdly, we integrate and compare three classification algorithms with various feature selection methods based on the concept of long-term stability. Test results from the last chapter also indicate that information abundance sometimes might not be a bless but a curse since basic algorithms are set in a way to fit for the training set by taking consideration of all available information. Therefore, they might easily lose focus on the key attributes. Future extension can be carried out as follows: reducing the dimensionality with other feature selection techniques, using different classifiers, collecting more information to construct new critical features (e.g. the industry sector), classifying companies according to different standards, categorizing companies into multiple classes, assessing data imbalance influence, and exploring advanced deep learning methods.

BIBLIOGRAPHY

Abdelghany, K., Abdelghany, A., & Raina, S. (2005). A model for the airlines' fuel management strategies. *Journal of Air Transport Management, 11(4), 199-206* .

Acharya, V. V., Afonso, G., & Kovner, A. (2016). How do global banks scramble for liquidity? Evidence from the asset-backed commercial paper freeze of 2007. *Journal of Financial Intermediation* .

Aggarwal, C. C. (2015). Outlier analysis. *In Data Mining, Springer International Publishing* , 237-263.

Air France Flight 4590. (n.d.). Retrieved from https://en.wikipedia.org/wiki/Air_France_Flight_4590

Air travel disruption after the 2010 Eyjafjallajökull eruption. (2015). Retrieved from https://en.wikipedia.org/wiki/Air_travel_disruption_after_the_2010_Eyjafjallaj%C3%B6kull_eruption

Alpaydin, E. (2014). *Introduction to machine learning.* MIT press.

Alvarez-Ramirez, J., Alvarez, J., & Rodriguez, E. (2008). Short-term predictability of crude oil markets: a detrended fluctuation analysis approach. *Energy Economics, 30(5)* , 2645-2656.

American Trucking Associations. (2008). *American Trucking Associations Report*. Retrieved from http://www.trucking.org/News_and_Information_Reports.aspx

Andrew-Ng-Machine-Learning-Programming-solutions. (n.d.). Retrieved from coursera.org Stanford University: https://github.com/LilianYe/Andrew-Ng-Machine-Learning-Programming-solutions-

Angelini, E., Tollo, G., & Roli, A. (2008). A neural network approach for credit risk evaluation. *The quarterly review of economics and finance, 48(4), 733-755* .

Angiulli, F., & Pizzuti, C. (2002). Fast outlier detection in high dimensional spaces. *In European Conference on Principles of Data Mining and Knowledge Discovery* , (pp. 15-27). Springer Berlin Heidelberg.

Arora, M., Singh, H., & Kaur, A. (2015). Distance based verification techniques for online signature verification system. *In 2015 2nd International Conference on Recent Advances in Engineering & Computational Sciences (RAECS)* (pp. 1-5). IEEE.

Ashcraft, A. B., & Schuermann, T. (2008). *Understanding the securitization of subprime mortgage credit (No. 3).* Now Publishers Inc.

Asif, M., & Muneer, T. (2007). Energy supply, its demand and security issues for developed and emerging economies. *Renewable and Sustainable Energy Reviews,11(7)* , 1388-1413.

Ayliffe, G. A., Green, W. E., Livingston, R., & Lowbury, E. J. (1977). Antibiotic-resistant Staphylococcus aureus in dermatology and burn wards. *Journal of clinical pathology, 30(1), 40-44* .

Baba, N., & Kozaki, M. (1992). An intelligent forecasting system of stock price using neural networks. *In Neural Networks, 1992. IJCNN., International Joint Conference on (Vol. 1)* (pp. 371-377). IEEE.

Bailey, D. L., & Thompson, D. (1990). Developing neural-network applications. *AI expert, 5(9), 34-41* .

Bartlett, P. L., Mendelson, S., & Neeman, J. (2012). ℓ 1-regularized linear regression: persistence and oracle inequalities. *Probability theory and related fields, 154(1-2), 193-224* .

Baum, E. B., & Haussler, D. (1989). What size net gives valid generalization? *Neural computation, 1(1), 151-160* .

Becken, S., & Lennox, J. (2012). Implications of a long-term increase in oil prices for tourism. *Tourism Management, 33(1)* , 133-142.

Bergerson, K., & Wunsch, D. C. (1991). A commodity trading model based on a neural network-expert system hybrid. *Neural Networks* .

Berkshire Hathaway . (1997-2000). *Berkshire Hathaway Annual and Interim Reports.* Berkshire Hathaway.

Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000). LOF: identifying density-based local outliers. *In ACM sigmod record (Vol. 29, No. 2)* , 93-104.

Brunnermeier, M. K. (2009). Deciphering the liquidity and credit crunch 2007–2008. *The Journal of economic perspectives, 23(1), 77-100* .

Burman, P. (1989). A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika, 76(3), 503-514* .

Calomiris, C. W. (2009). *The debasement of ratings: What's wrong and how we can fix it.* Unpublished working paper, Columbia Business School.

Cambini, C., & Jiang, Y. (2009). Broadband investment and regulation: A literature review. *Telecommunications Policy,33(10)* , 559-574.

Carson, R. T., Mitchell, R. C., Hanemann, M., Kopp, R. J., Presser, S., & Ruud, P. A. (2003). Contingent valuation and lost passive use: damages from the Exxon Valdez oil spill. *Environmental and resource economics, 25(3)* , 257-286.

Carter, D. A., Rogers, D. A., & Simkins, B. J. (2006). Does hedging affect firm value? Evidence from the US airline industry. *Financial management, 35(1)* , 53-86.

Case Study: The Collapse of Lehman Brothers. (2008). Retrieved from Investopedia: http://www.investopedia.com/articles/economics/09/lehman-brothers-collapse.asp

Chai, X., Shan, S., Chen, X., & Gao, W. (2007). Locally linear regression for pose-invariant face recognition. *IEEE Transactions on Image Processing, 16(7), 1716-1725* .

Cheh, J. J., Weinberg, R. S., & Yook, K. C. (2013). An application of an artificial neural network investment system to predict takeover targets. *Journal of Applied Business Research (JABR), 15(4), 33-46* .

Chen, S., & Gopalakrishnan, P. (1998). Speaker, environment and channel change detection and clustering via the bayesian information criterion. *DARPA Broadcast News Transcription and Understanding Workshop (Vol. 8),* (pp. 127-132).

Chen, Y., Miao, D., & Zhang, H. (2010). Neighborhood outlier detection. *Expert Systems with Applications 37(12)* , 8745-8749.

Chow, G. C., & Lin, A. L. (1971). Best linear unbiased interpolation, distribution, and extrapolation of time series by related series. *The review of Economics and Statistics, 372-375* .

Cleland, A. C., Earle, M. D., & Boag, I. F. (1981). Application of multiple linear regression to analysis of data from factory energy surveys. *International Journal of Food Science & Technology, 16(5), 481-492* .

CROX: summary for Crocs, Inc (2009). (n.d.). Retrieved from Yahoo Finance: http://finance.yahoo.com/q?s=CROX

Davies, H., & Green, D. (2013). *Global Financial Regulation: The Essential Guide (Now with a Revised Introduction).* John Wiley & Sons.

Davis, P. J. (1975). *Interpolation and approximation.* Courier Corporation.

Deichmann, J., Eshghi, A., Haughton, D., Sayek, S., & Teebagy, N. (2002). Application of multiple adaptive regression splines (MARS) in direct response modeling. *Journal of Interactive Marketing, 16(4), 15-27* .

Diamond, D. W., & Rajan, R. (2009). *The credit crisis: Conjectures about causes and remedies* . National Bureau of Economic Research.

Ding, C., & Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. *Journal of bioinformatics and computational biology, 3(02), 185-205* .

Ellerman, A. D., & Harrison Jr, D. (2003). Emissions trading in the US: Experience, lessons, and considerations for greenhouse gases.

Enders, C. K. (2010). *Applied missing data analysis.* Guilford Press.

Ersoy, O. (1990). *Tutorial at Hawaii International Conference on Systems Sciences.*

Examining The Delphi Bankruptcy's Impact on Workers and Retirees. (2009). Retrieved from https://www.gpo.gov/fdsys/pkg/CHRG-111hhrg53719/html/CHRG-111hhrg53719.htm

Faraway, J. J. (2016). *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models.* CRC press.

Fayyad, U. M., Reina, C., & Bradley, P. S. (1998). Initialization of Iterative Refinement Clustering Algorithms. *In KDD* , 194-198.

Feature selection code. (2015). Retrieved from http://users.spa.aalto.fi/jpohjala/featureselection/

Financial Crisis Inquiry Commission. (2011). *The financial crisis inquiry report: Final report of the national commission on the causes of the financial and economic crisis in the United States.* PublicAffairs.

FiskerAutomotive's road to ruin: How a "Billion-Dollar Startup Became a Billion-Dollar Disaster". (2013). Retrieved from Private Company Financial Intelligence: http://www.privco.com/fisker-automotives-road-to-ruin/

FJ., A. (1973). Graphs in statistical analysis. *American Statistician 27: 17–21* .

Florida, R. (2009). How the crash will reshape America. *The Atlantic, 303(2)* , 44-56.

Forbes, J. D. (1857). Further Experiments and Remarks on the Measurement of Heights by the Boiling Point of Water. *Transactions of the Royal Society of Edinburgh, 21(02), 235-243* .

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The annals of statistics* , 1-67.

Friedman, J., & Popescu, B. E. (2003). *Gradient directed regularization for linear regression and classification.* Technical Report, Statistics Department, Stanford University.

Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. *Statistics in medicine, 27(15), 2865-2873* .

Gentle, J. E. (2006). *Random number generation and Monte Carlo methods.* Springer Science & Business Media.

Gilbert, R., & Perl, A. (2013). *Transport revolutions: moving people and freight without oil.* New Society Publishers.

Gillingham, K., Newell, R., & Palmer, K. (2006). Energy efficiency policies: a retrospective examination. *Annu. Rev. Environ. Resour., 31* , 161-192.

Gittell, J. H., Nordenflycht, V., & Kochan, T. A. (2004). Mutual gains or zero sum? Labor relations and firm performance in the airline industry. *Industrial & Labor Relations Review, 57(2)* , 163-180.

Goonatilake, P. C. (1981). Postoperative wound infection: Application of multiple regression analysis to patients parameters. *International journal of bio-medical computing, 12(5), 393-399* .

Goonatilake, P. C. (1981). Postoperative wound infection: Application of multiple regression analysis to patients parameters. *International journal of bio-medical computing, 12(5), 393-399* .

Graham, B. (2009). *The Intelligent Investor.* Harper Collins .

Graham, B., & Dodd, D. L. (1934). *Security analysis: principles and technique.* McGraw-Hill.

Gramacy, R. B., & Lee, H. K. (2012). Bayesian treed Gaussian process models with an application to computer modeling. *Journal of the American Statistical Association* .

Guresen, E., Kayakutlu, G., & Daim, T. U. (2011). Using artificial neural network models in stock market index prediction. *Expert Systems with Applications, 38(8), 10389-10397* .

Guyon, I. W., J., B., S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning, 46(1-3), 389-422* .

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research, 3(Mar), 1157-1182* .

Hadavandi, E., Shavandi, H., & Ghanbari, A. (2010). Integration of genetic fuzzy systems and artificial neural networks for stock price forecasting. *Knowledge-Based Systems, 23(8), 800-808* .

Hall, M. A. (1999). *Correlation-based feature selection for machine learning.* Doctoral dissertation, The University of Waikato.

Harris, G. (2009). *Pfizer pays $2.3 billion to settle marketing case.* Retrieved from http://www.nytimes.com/2009/09/03/business/03health.html

Hazell, L., & Shakir, S. A. (2006). Under-reporting of adverse drug reactions. *Drug Safety, 29(5), 385-396* .

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering, 21(9), 1263-1284* .

Headey, D., & Fan, S. (2008). Anatomy of a crisis: the causes and consequences of surging food prices. *Agricultural Economics,39(s1)* , 375-391.

Hebb, D. O. (1949). *The organization of behavior: A neuropsychological theory.* Psychology Press.

Hidalgo, B., & Goodman, M. (2013). Hidalgo and Goodman Respond. *American journal of public health, 103(6)* .

Hilts, P. J. (2003). *Protecting America's health: The FDA, business, and one hundred years of regulation.* New York: Alfred A. Knopf: 23.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science, 313(5786), 504-507* .

History of Sycamore Networks, Inc. (2001). Retrieved from http://www.fundinguniverse.com/company-histories/sycamore-networks-inc-history/

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural networks, 4(2), 251-257* .

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks, 2(5), 359-366* .

Huang, G. B., Zhu, Q. Y., & Siew, C. K. (2004). Extreme learning machine: a new learning scheme of feedforward neural networks. *2004 IEEE International Joint Conference on (Vol. 2)* (pp. 985-990). In Neural Networks, 2004. Proceedings.

Jacobs, R. A. (1988). Increased rates of convergence through learning rate adaptation. *Neural networks, 1(4), 295-307* .

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR), 31(3)* , 264-323.

Janić, M. (2005). Modeling the large scale disruptions of an airline network. *Journal of transportation engineering, 131(4), 249-260* .

Jaramillo, J., Borgemeister, C., & Baker, P. (2006). Coffee berry borer Hypothenemus hampei (Coleoptera: Curculionidae): searching for sustainable control strategies. *Bulletin of entomological research, 96(03), 223-233* .

Jeffers, A. E. (2010). How Lehman Brothers used Repo 105 to manipulate their financial statements. *Journal of Leadership, Accountability and Ethics, 8(5)* , 44.

Johnes, G. (1999). Forecasting unemployment. *Applied Economics Letters, 6(9), 605-607* .

Kaastra, I., & Boyd, M. (1996). Designing a neural network for forecasting financial and economic time series. *Neurocomputing, 10(3), 215-236* .

Kaufmann, R. K., Dees, S. G., A., & Mann, M. (2008). Oil prices: the role of refinery utilization, futures markets and non-linearities. *Energy Economics, 30(5)* , 2609-2622.

Kelo, S. M., & Dudul, S. V. (2011). Short-term Maharashtra state electrical power load prediction with special emphasis on seasonal changes using a novel focused time lagged recurrent neural network based on time delay neural network model. *Expert Systems with Applications, 38(3), 1554-1564* .

Kelsey, J., Schneier, B., & Ferguson, N. (1999). Yarrow-160: Notes on the design and analysis of the yarrow cryptographic pseudorandom number generator. *In International Workshop on Selected Areas in Cryptography* (pp. 13-33). Springer Berlin Heide.

Kilian, L. (2006). *Not all oil price shocks are alike: Disentangling demand and supply shocks in the crude oil market.*

Kilian, L., & Murphy, D. P. (2014). The role of inventories and speculative trading in the global market for crude oil. *Journal of Applied Econometrics, 29(3)* , 454-478.

Kilian, L., & Park, C. (2009). The impact of oil price shocks on the US stock market. *International Economic Review, 50(4)* , 1267-1287.

Klier, T., & Linn, J. (2010). The price of gasoline and new vehicle fuel economy: evidence from monthly sales data. *American Economic Journal: Economic Policy, 2(3)* , 134-153.

Klimasauskas, C. C. (1993). Applying neural networks. *Neural networks in finance and investing* .

Knorr, E. M., Ng, R. T., & Tucakov, V. (2000). Distance-based outliers: algorithms and applications. *The VLDB Journal—The International Journal on Very Large Data Bases, 8(3-4)* , 237-253.

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial intelligence, 97(1), 273-324* .

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). *Supervised machine learning: A review of classification techniques.*

Krogh, A., & Vedelsby, J. (1995). Neural network ensembles, cross validation, and active learning. *Advances in neural information processing systems, 7, 231-238* .

Kumar, A., & Sawant, K. K. (2014). Application of multiple regression analysis in optimization of anastrozole-loaded PLGA nanoparticles. *Journal of microencapsulation, 31(2), 105-114* .

Kurt, I., Ture, M., & Kurum, A. T. (2008). Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. *Expert Systems with Applications, 34(1), 366-374* .

Lam, M. (2004). Neural network techniques for financial performance prediction: integrating fundamental and technical analysis. *Decision support systems, 37(4), 567-581* .

Lawrence, R. (1997). *Using neural networks to forecast stock market prices.* University of Manitoba.

Lebl, D. R., Bono, C. M., Velmahos, G., Metkar, U., Nguyen, J., & Harris, M. B. (2013). Vertebral artery injury associated with blunt cervical spine trauma: a multivariate regression analysis. *Spine, 38(16), 1352-1361* .

Lee, M. C. (1959). Multiple Regression vs. Pattern Analysis: A Comparative Application of Linear and Non-Linear Multivariate Methods of Prediction. *In XVth International Congress of Psychology, Brussels, Belgium (Vol. 28).*

Leroy, A. M., & Rousseeuw, P. J. (1987). *Robust regression and outlier detection.* New York: Wiley Series in Probability and Mathematical Statistics.

Levenberg, K. (1994). A method for the solution of certain non-linear problems in least squares. *Quarterly of applied mathematics, 2(2), 164-168* .

Liang, J., Yang, S., & Winstanley, A. (2008). Invariant optimal feature selection: A distance discriminant and feature ranking based solution. *Pattern Recognition, 41(5), 1429-1439* .

Lippmann, R. P. (1988). An introduction to computing with neural nets. *ACM SIGARCH Computer Architecture News, 16(1), 7-25* .

Liu, H., & Setiono, R. (1995). Chi2: Feature selection and discretization of numeric attributes. *ICTAI* , 388-391.

Liu, Z., Liu, J., Shi, X., Wang, L., Yang, Y., Tao, M., et al. (2016). Comparing calculated free testosterone with total testosterone for screening and diagnosing late-onset hypogonadism in aged males: A cross-sectional study. *Journal of Clinical Laboratory Analysis* .

Lorenzetti, L. (2015, September 21). *Here's why Turing Pharmaceuticals says 5,000% price bump is necessary.* Retrieved from http://fortune.com/2015/09/21/turing-pharmaceuticals-martin-shkreli-response/

Maloney, C. B. (2009). *The Ripple Effect: Why Failure of The Big Three is Not An Option.*

Mason, J. R. (2010). *The economic cost of a moratorium on offshore oil and gas exploration to the gulf region.* Louisiana State University.

Masters, T. (1993). *Practical neural network recipes in C++*.

Matsumoto, M., & Nishimura, T. (1998). Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation (TOMACS), 8(1), 3-30* .

McDonald, R., & Paulson, A. (2015). AIG in Hindsight. *The Journal of Economic Perspectives, 29(2), 81-105* .

McManus, W. S. (2005). *In the tank: how oil prices threaten automakers' profits and jobs.*

Mena, J. (2011). *Machine learning forensics for law enforcement, security, and intelligence.* CRC Press.

Michie, D., Spiegelhalter, J., D., & Taylor, C. C. (1994). *Machine learning, neural and statistical classification.*

Milesi-Ferretti, G. M., & Tille, C. (2011). The great retrenchment: international capital flows during the global financial crisis. *Economic Policy, 26(66)* , 289-346.

Morrell, P., & Swan, W. (2006). Airline jet fuel hedging: Theory and practice. *Transport Reviews, 26(6), 713-730* .

Moustris, K. P., Nastos, P. T., Larissi, I. K., & Paliatsos, A. G. (2012). *Application of multiple linear regression models and artificial neural networks on the surface ozone forecast in the greater Athens area, Greece.* Advances in Meteorology.

Musshoff, O., Odening, M., & Xu, W. (2011). Management of climate risks in agriculture–will weather derivatives permeate? *Applied Economics, 43(9), 1067-1077* .

Naseem, I., Togneri, R., & Bennamoun, M. (2010). Linear regression for face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(11), 2106-2112* .

National Philanthropic Trust. (2015). *Charitable Giving Statistics*. Retrieved from https://www.nptrust.org/philanthropic-resources/charitable-giving-statistics/

Ng, A. (2013). *http://www.andrewng.org/courses/*. Retrieved from Machine Learning Courses.

Ng, A. Y. (2004). Feature selection, L 1 vs. L 2 regularization, and rotational invariance. *In Proceedings of the twenty-first international conference on Machine learning* (p. 78). ACM.

Ng, A. Y. (2004). Feature selection, L 1 vs. L 2 regularization, and rotational invariance. *In Proceedings of the twenty-first international conference on Machine learning* (p. 78). ACM.

Ng, A. Y. (1998). On feature selection: learning with exponentially many irrevelant features as training examples.

Nguyen, A. P., & Enomoto, C. E. (2009). The Troubled Asset Relief Program (TARP) and the financial crisis of 2007-2008. *Journal of Business & Economics Research, 7(12), 91* .

Nigam, K., McCallum, A., Thrun, S., & & Mitchell, T. (1998). Learning to classify text from labeled and unlabeled documents. *AAAI/IAAI, 792* .

Nocedal, J. (1992). Theory of algorithms for unconstrained optimization. *Acta numerica, 1, 199-242* .

Ofek, E., & Richardson, M. (2003). Dotcom mania: The rise and fall of internet stock prices. *The Journal of Finance, 58(3), 1113-1137* .

Ogutu, J. O., Schulz-Streeck, T., & Piepho, H. P. (2012). Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. *In BMC proceedings* (p. 1). BioMed Central.

Park, S. K., & Miller, K. W. (1988). Random number generators: good ones are hard to find. *Communications of the ACM, 31(10), 1192-1201* .

Pearson, K., Yule, G. U., Blanchard, N., & Lee, A. (1903). The law of ancestral heredity. *Biometrika, 2(2), 211-236* .

Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence, 27(8), 1226-1238* .

Pengelly, M. (2010). Flash cordons. *Risk, 23(11)* , 58.

Pohjalainen, J., Räsänen, O., & Kadioglu, S. (2015). Feature selection methods and their combinations in high-dimensional classification of speaker likability, intelligibility and personality traits. *Computer Speech & Language, 29(1), 145-171* .

Porto, I. D., Cardoso, F. L., & Sacomori, C. (2016). Sports practice, resilience, body and sexual esteem, and higher educational level are associated with better sexual adjustment in men with acquired paraplegia. *Journal of rehabilitation medicine* .

Pyle, D., & Jose, C. S. (2015). *An executive's guide to machine learning.* McKinsey Quarterly.

Quackenbush, J. (2002). Microarray data normalization and transformation. *Nature genetics, 32, 496-501* .

Quah, T. S., & Srinivasan, B. (1999). Improving returns on stock investment through neural network selection. *Expert Systems with Applications, 17(4), 295-301* .

Quinlan, J. R. (1987). Simplifying decision trees. *International journal of man-machine studies, 27(3), 221-234* .

Ramaswamy, S., Rastogi, R., & Shim, K. (2000). Efficient algorithms for mining outliers from large data sets. *ACM SIGMOD Record (Vol. 29, No. 2)* , 427-438.

Ramsey, J. B. (1999). The contribution of wavelets to the analysis of economic and financial data. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, 357(1760), 2593-2606* .

Rao, K. R., & Lakshminarayanan, S. (. (2007). Partial correlation based variable selection approach for multivariate data classification methods. *Chemometrics and intelligent laboratory systems, 86(1), 68-81* .

Räsänen, O., & Pohjalainen, J. (2013). Random subset feature selection in automatic recognition of developmental disorders, affective states, and level of conflict from speech. *In INTERSPEECH* , 210-214.

Ray-Mukherjee, J., Nimon, K., Mukherjee, S., Morris, D. W., Slotow, R., & Hamer, M. (2014). Using commonality analysis in multiple regressions: a tool to decompose regression effects in the face of multicollinearity. *Methods in Ecology and Evolution, 5(4)* .

Reed, R., Marks, R. J., & Oh, S. (1995). Similarities of error regularization, sigmoid gain scaling, target smoothing, and training with jitter. *IEEE Transactions on Neural Networks, 6(3), 529-538* .

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-validation. *In Encyclopedia of database systems. Springer US* , 532-538.

Rencher, A. C., & Schaalje, G. B. (2008). *Linear models in statistics.* John Wiley & Sons.

Ritter, J. (2014). *Initial Public Offerings: Updated Statistics.*

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review, 65(6), 386* .

Saad, E. W., Prokhorov, D. V., & Wunsch, D. C. (1998). Comparative study of stock trend prediction using time delay, recurrent and probabilistic neural networks. *IEEE Transactions on neural networks, 9(6), 1456-1470* .

Samuel, A. L. (1959). *Some studies in machine learning using the game of checkers.* IBM Journal of research and development, 3(3), 210-229.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data.* CRC press.

Schimek, M. G. (2013). *Smoothing and regression: approaches, computation, and application.* John Wiley & Sons.

Schloegl, A. (2010). *NaN-Tb: A statistics toolbox*. Retrieved from http://lasp.colorado.edu/cism/CISM_DX/code/CISM_DX-0.50/required_packages/octave-forge/extra/NaN/README.TXT

Schloemer, E., Li, W., Ernst, K., & Keest, K. (2006). *Foreclosures in the subprime market and their cost to homeowners.* Center for Responsible Lending.

Schwarcz, S. L. (2008). Disclosure's failure in the subprime mortgage crisis. *Utah Law Review, 1109* .

Scrosati, B., & Garche, J. (2010). Lithium batteries: Status, prospects and future. *Journal of Power Sources,195(9)* , 2419-2430.

Seo, S. N. (2013). An essay on the impact of climate change on US agriculture: weather fluctuations, climatic shifts, and adaptation strategies. *Climatic Change, 121(2), 115-124* .

Setiono, R. (2001). Feedforward neural network construction using cross validation. *Neural Computation, 13(12), 2865-2877* .

Setiono, R., & Liu, H. (1997). Neural-network feature selector. *IEEE transactions on neural networks, 8(3), 654-662* .

Shanker, M., Hu, M. Y., & Hung, M. S. (1996). Effect of data standardization on neural network training. *Omega, 24(4), 385-397* .

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American statistical Association, 88(422), 486-494* .

Shiller, R. J. (2015). *Irrational exuberance.* Princeton university press.

Smith, F. (2010). Madoff Ponzi Scheme Exposes the Myth of the Sophisticated Investor. *U. Balt. L. Rev., 40, 215* .

Smith, T. (2009). Going to Seed: Using Monsanto as a Case Study to Examine the Patent and Antitrust Implications of the Sale and Use of Genetically Modified Seed. *Ala. L. Rev., 61, 629* .

Sparrow, M. K. (2000). *License to steal: how fraud bleeds America's health care system.* Basic Books.

Sun, W., & Yuan, Y. X. (2006). *Optimization theory and methods: nonlinear programming.* Springer Science & Business Media.

Swales, G. S. (1992). Applying artificial neural networks to investment analysis. *Financial Analysts Journal, 48(5)* .

Syla, S. (2013). Application of Multiple Linear Regression Analysis of Employment through ALMP. *International Journal of Academic Research in Business and Social Sciences, 3(12), 252* .

Symonds, M. R., & Moussalli, A. (2011). A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike's information criterion. *Behavioral Ecology and Sociobiology, 65(1), 13-21* .

Tadele, H., & Rao, P. M. (2014). Corporate Governance and Ethical issues in Microfinance Institutions (MFIs)-A study of Microfinance Crises in Andhra Pradesh, India. *Journal of Business Management & Social Sciences Research, 3(1)* , 21-26.

Teräsvirta, T., Lin, C. F., & Granger, C. W. (1993). Power of the neural network linearity test. *Journal of Time Series Analysis, 14(2), 209-220* .

Terrin, N., Schmid, C. H., Griffith, J. L., D'Agostino, R. B., & Selker, H. P. (2003). External validity of predictive models: a comparison of logistic regression, classification trees, and neural networks. *Journal of clinical epidemiology, 56(8), 721-729* .

The Associated Press. (2008, April 19). *Caterpillar's Profit Climbs on Strength of Foreign Sales.* Retrieved from http://www.nytimes.com/2008/04/19/business/19caterpillar.html

Thomas, D. M., Amburgey, J., & Ellis, L. (2016). Anti-transgender prejudice mediates the association of just world beliefs and victim blame attribution. *International Journal of Transgenderism, 1-9* .

Ticknor, J. L. (2013). A Bayesian regularized artificial neural network for stock market forecasting. *Expert Systems with Applications, 40(14), 5501-5506* .

Timilsina, G. R., Mevel, S., & Shrestha, A. (. (2011). Oil price, biofuels and food supply. *Energy Policy, 39(12)* , 8098-8105.

Tkacz, G. (2001). Neural network forecasting of Canadian GDP growth. *International Journal of Forecasting, 17(1), 57-69* .

Toh, B. L., Yeoh, H. K., Teoh, W. H., & Chin, L. C. (2016). Surface mount adhesive: in search of a perfect dot. *The International Journal of Advanced Manufacturing Technology, 1-12* .

Tolosana-Delgado, R., & von Eynatten, H. (2009). Grain-size control on petrographic composition of sediments: compositional regression and rounded zeros. *Mathematical geosciences, 41(8), 869-886* .

Trippi, R. R., & Turban, E. (1992). *Neural networks in finance and investing: Using artificial intelligence to improve real world performance.* McGraw-Hill.

Tyner, W. E. (2008). The US ethanol and biofuels boom: Its origins, current status, and future prospects. *BioScience, 58(7)* , 646-653.

Umar, M., & Sun, G. (2016). Determinants of different types of bank liquidity: evidence from BRICS countries. *China Finance Review International, 6(4)* .

United States Department of Labor. (2009). *Bureau of Labor Statistics.*

USPS. (2014). *A decade of facts and figures*. Retrieved from https://about.usps.com/who-we-are/postal-facts/decade-of-facts-and-figures.htm

Wang, Y. M., & Traore, S. (2009). Time-lagged recurrent network for forecasting episodic event suspended sediment load in typhoon prone area. *International Journal of Physical Sciences, 4(9), 519-528* .

Weber, J. G. (2012). The effects of a natural gas boom on employment and income in Colorado, Texas, and Wyoming. *Energy Economics, 34(5)* , 1580-1588.

Weisberg, S. (2005). *Applied linear regression.* John Wiley & Sons.

Weistroffer, C., Speyer, B. K., S., & Mayer, T. (2009). *Credit default swaps.* Deutsche bank research.

Wilson, D. R., & Martinez, T. R. (2003). The general inefficiency of batch training for gradient descent learning. *Neural Networks, 16(10), 1429-1451* .

Wirasingha, S. G., Schofield, N., & Emadi, A. (2008). Plug-in hybrid electric vehicle developments in the US: Trends, barriers, and economic feasibility. *2008 IEEE Vehicle Power and Propulsion Conference* (pp. 1-8). IEEE .

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann.

Xu, Y. L., & Chen, D. R. (2009). Partially-linear least-squares regularized regression for system identification. *IEEE Transactions on Automatic Control, 54(11), 2637-2641* .

Yan, X. (2009). *Linear regression analysis: theory and computing.* World Scientific.

Ye, M., Zyren, J., & Shore, J. (2002). Forecasting crude oil spot price using OECD petroleum inventory levels. *International Advances in Economic Research, 8(4)* , 324-333.

Yeung, N., Botvinick, M. M., & Cohen, J. D. (2004). The neural basis of error detection: conflict monitoring and the error-related negativity. *Psychological review, 111(4), 931* .

Yoon, Y., & Swales, G. (1991). Predicting stock price performance: A neural network approach. *System Sciences. Proceedings of the Twenty-Fourth Annual Hawaii International Conference on (Vol. 4, pp. 156-162).* Hawaii: IEEE.

Zhang, Z., & Trafalis, B. (2016). A Two-stage Feature Selection Neural Network (NN) Analysis in Investment Analysis. Working paper in progress.

Appendix A: DATA PRE-PROCESSING

Combine Workbooks

```
' Open the first workbook in the folder
' Pick the folder
' Save as Combined.xlsm

Sub comb()

On Error Resume Next

Dim wb As Workbook, sh As Worksheet

Dim fn As String, pt As String

startT = Timer

' Picking a folder
With Application.FileDialog(msoFileDialogFolderPicker)

  .Show

  .AllowMultiSelect = False

  If .SelectedItems.Count = 0 Then

    Exit Sub

  Else

    pt = .SelectedItems(1)

  End If

End With

' Return a list of filenames with .csv extension
fn = Dir(pt & "\*.csv")

Do While fn <> ""

  If fn <> ThisWorkbook.Name Then 'Skip the first one

  k = k + 1
```

```
  Application.ScreenUpdating = False

  Application.DisplayAlerts = False

  Set wb = Workbooks.Open(pt & "\" & fn, , True)

  Set sh =
ThisWorkbook.Worksheets.Add(after:=ThisWorkbook.Worksheets(ThisWorkbook.Wo
rksheets.Count))

  wb.Worksheets(1).Rows.Copy sh.Rows

  sh.Name = Left(fn, Len(fn) - IIf(Right(fn, 1) = "x", 30, 29)) '
"A_quarterly_financial_data.csv"

  wb.Close

  Application.DisplayAlerts = True

  Application.ScreenUpdating = True

  End If

  fn = Dir

Loop

  endT = Timer

  ThisWorkbook.Worksheets(1).Name = Left(Thisworkbook.Name,
Len(Thisworkbook.Name) - IIf(Right(Thisworkbook.Name, 1) = "x", 30, 29))

  ThisWorkbook.Save

  MsgBox "Combined " & k+1 & " Companies " & " @ " & endT - startT & " Seconds"

End Sub
```

Copy Qualified Sheets

```
' Run in the combined.xlsx
' Need three new workbooks

Sub copyQualifiedSheets()

On Error Resume Next
```

```
Dim wb As Workbook, sh As Worksheet

Set wb = Workbooks.Open("H:\Big Data\Test\2000-2014.xlsx")

'Clear storage workbook
For Each ws In wb.Worksheets
   If Not ws.Name = "Sheet1" Then ws.Delete
Next ws

sheetsCount = ThisWorkbook.Worksheets.Count
MsgBox "# of companies from the original dataset: " & sheetsCount

For i = 1 To sheetsCount

   If ThisWorkbook.Worksheets(i).Range("A2") >= DateValue("December 31,2014") _
   And ThisWorkbook.Worksheets(i).Range("A:A").End(xlDown) <=
DateValue("December 31,1999") _
   Then
   'MsgBox ThisWorkbook.Worksheets(i).Name & ": Copy"

   ThisWorkbook.Worksheets(i).Copy After:=wb.Sheets(wb.Sheets.Count)
   wb.Sheets(wb.Sheets.Count).Name = ThisWorkbook.Worksheets(i).Name

   Else
   'MsgBox Worksheets(i).Name & ": Don't Copy"

End If

Next i

wb.Sheets(1).Delete

MsgBox wb.Worksheets.Count & " companies qualified."

End Sub
```

Identify and Copy Selected Ranges

```
'Identify and copy data from 2010 - 2014 to wb1, 2000-2009 to wb2
'Run in 2000-2015.xlsx

Sub findRangeCopy()

Application.ScreenUpdating = False
```

```
Dim j, n, k As Long
Dim wb1, wb2 As Workbook
Dim ws As Worksheet

n = 19
m = 39
countSh = ThisWorkbook.Worksheets.Count
colN = ThisWorkbook.Worksheets(1).UsedRange.Columns.Count

MsgBox countSh & " qualified companies " & " with " & colN - 1 & " attributes"

'Open & clean subStorage workbooks
Set wb1 = Workbooks.Open("H:\Big Data\Test\2010-2014.xlsx")
Set wb2 = Workbooks.Open("H:\Big Data\Test\2000-2009.xlsx")


For Each ws In wb1.Worksheets
   If Not ws.Name = "Sheet1" Then ws.Delete
Next ws

For Each ws In wb2.Worksheets
   If Not ws.Name = "Sheet1" Then ws.Delete
Next ws


For j = 1 To countSh

   For i = 1 To Rows.Count

      If ThisWorkbook.Sheets(j).Cells(i, 1).Value <= DateValue("December 31, 2014")
And ThisWorkbook.Sheets(j).Cells(i, 1).Value >= DateValue("September 1, 2014")
Then

      UpperRange = ThisWorkbook.Sheets(j).Cells(i, 1)
      LowerRange = ThisWorkbook.Sheets(j).Cells(i + n, 1)
      'MsgBox ThisWorkbook.Sheets(j).Name & " : " & UpperRange & "  -  " &
LowerRange
      UpperRange2 = ThisWorkbook.Sheets(j).Cells(i + n + 1, 1)
      LowerRange2 = ThisWorkbook.Sheets(j).Cells(i + n + 1 + m, 1)

     '2010-2014.xlsx
      wb1.Sheets.Add after:=wb1.Sheets(wb1.Sheets.Count)
      wb1.Worksheets(wb1.Sheets.Count).Rows(1).Value =
ThisWorkbook.Sheets(j).Rows(1).Value
      wb1.Worksheets(wb1.Sheets.Count).Cells(2, 1).Resize(n + 1, 31).Value =
ThisWorkbook.Sheets(j).Cells(i, 1).Resize(n + 1, 31).Value
```

```
        wb1.Worksheets(wb1.Sheets.Count).Name = ThisWorkbook.Sheets(j).Name
        'MsgBox "Workbook: " & wb1.Name & vbNewLine & "Worksheet " &
wb1.Worksheets(wb1.Sheets.Count).Name _
        & ": " & LowerRange & "  -  " & UpperRange


       '2000-2009.xlsx
        wb2.Sheets.Add after:=wb2.Sheets(wb2.Sheets.Count)
        wb2.Worksheets(wb2.Sheets.Count).Rows(1).Value =
ThisWorkbook.Sheets(j).Rows(1).Value
        wb2.Worksheets(wb2.Sheets.Count).Cells(2, 1).Resize(m + 1, 31).Value =
ThisWorkbook.Sheets(j).Cells(i + n + 1, 1).Resize(m + 1, 31).Value
        wb2.Worksheets(wb2.Sheets.Count).Name = ThisWorkbook.Sheets(j).Name
        'MsgBox "Workbook: " & wb2.Name & vbNewLine & "Worksheet " &
wb2.Worksheets(wb2.Sheets.Count).Name _
        & ": " & LowerRange2 & "  -  " & UpperRange2

        Exit For

       End If

    Next i

Next j

wb1.Save
wb2.Save

verifyNo = wb2.Worksheets.Count - 1
MsgBox "Done with " & verifyNo & " companies!"

End Sub
```

Clean & Attributes Selection 2000-2008/2010-2014

```
' Run in 2000-2009 & 2010-2014 seperately
' Find "None", replace it with 0
' Use 'Sheet1' as evaluation sheet, paste and calcuate data
' Shares split adjusted EPS, Div, P!

Sub repModifyCal()

Dim countWS, I, J, k As Integer
Dim lastRow, lastCol As Integer
Dim fid As Variant
Dim rpl As Double
Dim ws1, wsE As Worksheet
```

```
countWS = ThisWorkbook.Worksheets.Count
Application.ReplaceFormat.Font.Color = 255
fid = "None"
rpl = 0

On Error Resume Next

'Build up an evaluation sheet
Set ws1 = ThisWorkbook.Worksheets(1)
With ws1
   .Cells.Clear
   .Cells.HorizontalAlignment = xlCenter
   .Cells.WrapText = True
   .Rows(1).Font.Bold = True
   .Columns(1).Font.Bold = True
   .Cells(1, 1).Value = "COMPANY"
   .Cells(1, 1).Font.Color = 255
End With

noYear = 9  'Ignore 2009

For J = 2 To countWS
   Set wsE = Worksheets(J)
   lastRow = wsE.Cells(wsE.Rows.Count, "A").End(xlUp).Row

   'Replace NONE to 0
   wsE.Cells.Replace What:=fid, Replacement:=rpl, MatchCase:=False,
SearchFormat:=True, ReplaceFormat:=True
   Application.ScreenUpdating = Flase

   'List company name
   ws1.Cells(J, 1) = wsE.Name

   'No of increased shares
   ws1.Cells(1, 2) = "Averaged Increased Shares Split Adjusted 2000 - 2008"
   ws1.Cells(J, 2) =
(Application.WorksheetFunction.Average(wsE.Range("C6").Resize(4, 1)) -
Application.WorksheetFunction.Average(wsE.Range("C38").Resize(4, 1))) / noYear

   'Assets/Liabilities
   ws1.Cells(1, 3) = "Averaged Assets/Liabilities 2000 - 2008"
   Set RngT = wsE.Range(wsE.Cells(6, 33), wsE.Cells(41, 33))
   RngT.Formula = "=E6/F6"
   ws1.Cells(J, 3) = Application.WorksheetFunction.Average(RngT)
   ws1.Cells(J, 3).NumberFormat = "0.0"
   RngT.Clear
```

```
    'Changes of Equity
    ws1.Cells(1, 4) = "Averaged Changes of Equity 2000 - 2008"
    ws1.Cells(J, 4) =
(Application.WorksheetFunction.Average(wsE.Range("G6").Resize(4, 1)) -
Application.WorksheetFunction.Average(wsE.Range("G38").Resize(4, 1))) / noYear

    'Changes of Long Term Debt
    ws1.Cells(1, 5) = "Averaged Changes of LT Debt 2000 - 2008"
    ws1.Cells(J, 5) =
(Application.WorksheetFunction.Average(wsE.Range("K6").Resize(4, 1)) -
Application.WorksheetFunction.Average(wsE.Range("K38").Resize(4, 1))) / noYear

    'Changes of Revenue
    ws1.Cells(1, 6) = "Averaged Changes of Revenue 2000 - 2008"
    ws1.Cells(J, 6) = (Application.WorksheetFunction.Sum(wsE.Range("L6").Resize(4,
1)) - Application.WorksheetFunction.Sum(wsE.Range("L38").Resize(4, 1))) / noYear

    'Average net margin
    ws1.Cells(1, 7) = "Averaged Net Margin 2000 - 2008"
    Set RngT = wsE.Range(wsE.Cells(6, 33), wsE.Cells(41, 33))
    RngT.Formula = "=M6/L6"
    ws1.Cells(J, 7) = Application.WorksheetFunction.Average(RngT)
    ws1.Cells(J, 7).NumberFormat = "0.0%"
    RngT.Clear

    'Shares split adjusted EPS 2000 - 2008
    ws1.Cells(1, 8) = "Averaged Adjusted EPS 2000 - 2008"
    Set RngT = wsE.Range(wsE.Cells(6, 33), wsE.Cells(41, 33))
    RngT.Formula = "=N6/C6"
    ws1.Cells(J, 8) = Application.WorksheetFunction.Average(RngT) * 4
    ws1.Cells(J, 8).NumberFormat = "0.0"
    RngT.Clear

    'Shares split adjusted Div 2000 - 2008
    ws1.Cells(1, 9) = "Averaged Adjusted Div 2000 - 2008"
    Set RngT = wsE.Range(wsE.Cells(6, 33), wsE.Cells(41, 33))
    RngT.Formula = "=Q6/D6"
    ws1.Cells(J, 9) = Application.WorksheetFunction.Average(RngT) * 4
    ws1.Cells(J, 9).NumberFormat = "0.0"
    RngT.Clear

    'Shares split adjusted Average Price Difference 2000 - 2008
    ws1.Cells(1, 10) = "Shares split adjusted Average Price Difference 2000 - 2008"
    Set RngT = wsE.Range(wsE.Cells(6, 33), wsE.Cells(41, 33))
    RngT.Formula = "=R6/D6"
```

```
   ws1.Cells(J, 10) = (Application.WorksheetFunction.Average(wsE.Cells(6,
33).Resize(4, 1)) - Application.WorksheetFunction.Average(wsE.Cells(38, 33).Resize(4,
1))) / noYear
   ws1.Cells(J, 10).NumberFormat = "0.0"
   RngT.Clear

   'Shares split adjusted average(P_High - P_Low) 2000 - 2008
   ws1.Cells(1, 11) = "SUM (P_High - P_Low) 2000 - 2008"
   Set RngT1 = wsE.Range(wsE.Cells(6, 33), wsE.Cells(41, 33))
   Set RngT2 = wsE.Range(wsE.Cells(6, 34), wsE.Cells(41, 34))
   RngT1.Formula = "=S6/D6"
   RngT2.Formula = "=T6/D6"
   ws1.Cells(J, 11) = (Application.WorksheetFunction.Sum(RngT1) -
Application.WorksheetFunction.Sum(RngT2)) / noYear
   ws1.Cells(J, 11).NumberFormat = "0.0"
   RngT1.Clear
   RngT2.Clear

Next J

MsgBox "Selected Features of  " & ThisWorkbook.Name & " " & countWS - 1 & "
companies."

ws1.Cells(1, 12) = "CLASS"
For k = 2 To countWS
   If ws1.Cells(k, 8) <= 0 Or ws1.Cells(k, 9) <= 0 Or ws1.Cells(k, 10) <= 0 Then
      ws1.Cells(k, 12) = 0
      Else
      ws1.Cells(k, 12) = 1
   End If
Next k

class1 = Application.WorksheetFunction.Sum(Columns(12))
MsgBox class1

End Sub
```

Classification 1: EPS, Div, Price Appreciation

```
' Define standards : Average price, earnings and dividends appreciation, Assets inc >
Liabilities inc
' run in Classification.xlsx

Sub classify2Methods()

Dim wb, wb1, wb2 As Workbook
Dim ws As Worksheet
```

```
ScreenUpdating = False

Set wb = ThisWorkbook
Set wb1 = Workbooks.Open("H:\Big Data\Test\2010-2014.xlsx")
Set wb2 = Workbooks.Open("H:\Big Data\Test\2000-2009.xlsx")
Set ws1 = wb1.Worksheets(1)
Set ws2 = wb2.Worksheets(1)

For Each ws In wb.Worksheets
   If Not ws.Name = "Sheet1" Then ws.Delete
Next ws

wb.Worksheets.Add After:=Worksheets(wb.Worksheets.Count)
wb.Worksheets(wb.Worksheets.Count).Name = "Final"
Set wsFin = wb.Worksheets("Final")

n = ws1.Cells(ws1.Rows.Count, "A").End(xlUp).Row ' # of companies
'Copy Shares split adjusted results
wsFin.Cells(1, 1).Resize(n, 1).Value = ws1.Cells(1, 1).Resize(n, 1).Value
wsFin.Cells(1, 14).Resize(n, 5).Value = ws2.Cells(1, 8).Resize(n, 5).Value
wsFin.Cells(1, 19).Resize(n, 5).Value = ws1.Cells(1, 8).Resize(n, 5).Value
wsFin.Cells(1, 24) = "Misclassification"

mis = 0
investWrong = 0
specWrong = 0

For h = 2 To n
   If wsFin.Cells(h, 18).Value <> wsFin.Cells(h, 23).Value Then
      wsFin.Cells(h, 24) = "Y"
      mis = mis + 1

      If wsFin.Cells(h, 18).Value > wsFin.Cells(h, 23).Value Then investWrong =
investWrong + 1
      If wsFin.Cells(h, 18).Value < wsFin.Cells(h, 23).Value Then specWrong =
specWrong + 1

   End If
Next h

totalMisRate = mis / (n - 1)

Set wbT1 = wb2
Set wbT2 = wb1
```

```
rowNo1 = wbT1.Worksheets(2).Cells(wbT1.Worksheets(2).Rows.Count,
"A").End(xlUp).Row '41-1 records
rowNo2 = wbT2.Worksheets(2).Cells(wbT2.Worksheets(2).Rows.Count,
"A").End(xlUp).Row '21-1 recrods

class1 = Application.WorksheetFunction.Sum(wsFin.Columns(18))
class1_1 = Application.WorksheetFunction.Sum(wsFin.Columns(23))

MsgBox "Companies: " & wbT1.Worksheets.Count - 1 & vbLf _
& "2000 - 2009 Records: " & rowNo1 - 1 & vbLf _
& "2010 - 2014 Records: " & rowNo2 - 1 & vbLf _
& "2000 - 2009 Class 1: " & class1 & vbLf _
& "2010 - 2014 Class 1: " & class1_1 & vbLf _
& "Overall Classification accuracy: " & Format(1 - totalMisRate, "Percent") & vbLf _
& "Mistaken investment: " & investWrong & vbLf _
& "Mistaken specualtion: " & specWrong

' Methods 2

With wsFin
    .Cells.WrapText = True
    .Columns(1).Font.Bold = True
    .Rows(1).Font.Bold = True
    .Cells.HorizontalAlignment = xlCenter
    .Cells(1, 2) = "Industry"
    .Cells(1, 4) = "Years of Deficits 2000-2007"
    .Cells(1, 5) = "Years of 0 Div 2000-2007"
    .Cells(1, 6) = "P_2007 - P_2000"
    .Cells(1, 7) = "Years of Deficits 2010-2014"
    .Cells(1, 8) = "Years of 0 Div 2010-2014"
    .Cells(1, 9) = "P_2014 - P_2010"
    .Cells(1, 10) = "Class 2000 - 2007"
    .Cells(1, 11) = "Class 2010 - 2014"
    .Cells(1, 12) = "Misclassification"
End With

' Check 2000 - 2007
For comp = 2 To wbT1.Worksheets.Count

   def1 = 0
   noDiv = 0
   y = 0

   For a1 = 10 To rowNo1 Step 4

      Set rng1 = wbT1.Worksheets(comp).Cells(a1, 15).Resize(4, 1)
```

```
        Set rng2 = wbT1.Worksheets(comp).Cells(a1, 17).Resize(4, 1)
        Set rng3 = wbT1.Worksheets(comp).Cells(a1, 18).Resize(4, 1)


        If Application.WorksheetFunction.Sum(rng1) <= 0 Then def1 = def1 + 1
        If Application.WorksheetFunction.Sum(rng2) <= 0 Then noDiv = noDiv + 1
        aveP = Application.WorksheetFunction.Average(rng3)

        If a1 = 10 Then P2007 = aveP
        If a1 = 38 Then P2000 = aveP

    Next a1

    ThisWorkbook.Worksheets("Final").Cells(comp, 4) = def1
    ThisWorkbook.Worksheets("Final").Cells(comp, 5) = noDiv
    ThisWorkbook.Worksheets("Final").Cells(comp, 6) = P2007 - P2000

    If def1 > 0 Or noDiv > 0 Or P2007 - P2000 < 0 Then
ThisWorkbook.Worksheets("Final").Cells(comp, 10) = 0 _
    Else ThisWorkbook.Worksheets("Final").Cells(comp, 10) = 1



Next comp

' Check 2010 - 2014
For comp = 2 To wbT2.Worksheets.Count

    def1 = 0
    noDiv = 0
    y = 0

    For a1 = 2 To rowNo2 Step 4

        Set rng1 = wbT2.Worksheets(comp).Cells(a1, 15).Resize(4, 1)
        Set rng2 = wbT2.Worksheets(comp).Cells(a1, 17).Resize(4, 1)
        Set rng3 = wbT2.Worksheets(comp).Cells(a1, 18).Resize(4, 1)

        If Application.WorksheetFunction.Sum(rng1) <= 0 Then def1 = def1 + 1
        If Application.WorksheetFunction.Sum(rng2) <= 0 Then noDiv = noDiv + 1
        aveP = Application.WorksheetFunction.Average(rng3)
        If a1 = 2 Then P2014 = aveP
        If a1 = 18 Then P2010 = aveP

    Next a1
```

```
   ThisWorkbook.Worksheets("Final").Cells(comp, 7) = def1
   ThisWorkbook.Worksheets("Final").Cells(comp, 8) = noDiv
   ThisWorkbook.Worksheets("Final").Cells(comp, 9) = P2014 - P2010

   If def1 > 0 Or noDiv > 0 Or P2014 - P2010 < 0 Then
ThisWorkbook.Worksheets("Final").Cells(comp, 11) = 0 _
   Else ThisWorkbook.Worksheets("Final").Cells(comp, 11) = 1

Next comp

mis = 0
investWrong = 0
specWrong = 0

For h = 2 To wsFin.UsedRange.Rows(wsFin.UsedRange.Rows.Count).Row
   If wsFin.Cells(h, 10).Value <> wsFin.Cells(h, 11).Value Then
      wsFin.Cells(h, 12) = "Y"
      mis = mis + 1

      If wsFin.Cells(h, 10).Value > wsFin.Cells(h, 11).Value Then investWrong =
investWrong + 1
      If wsFin.Cells(h, 10).Value < wsFin.Cells(h, 11).Value Then specWrong =
specWrong + 1

   End If
Next h

totalMisRate = mis / (wbT1.Worksheets.Count - 1)

class1 = Application.WorksheetFunction.Sum(wsFin.Columns(10))
class1_1 = Application.WorksheetFunction.Sum(wsFin.Columns(11))

MsgBox "Companies: " & wbT1.Worksheets.Count - 1 & vbLf _
& "2000 - 2007 Records: " & rowNo1 - 1 & vbLf _
& "2010 - 2014 Records: " & rowNo2 - 1 & vbLf _
& "2000 - 2009 Class 1: " & class1 & vbLf _
& "2010 - 2014 Class 1: " & class1_1 & vbLf _
& "Overall Classification accuracy: " & Format(1 - totalMisRate, "Percent") & vbLf _
& "Mistaken investment: " & investWrong & vbLf _
& "Mistaken specualtion: " & specWrong

End Sub
```

Classification 2: Average EPS, Div, Price

```
' Define standards : Average price, earnings and dividends appreciation, Assets inc > Liabilities inc
' run in Classification.xlsx

Sub classificationRel()

Dim wb, wb1, wb2 As Workbook
Dim ws As Worksheet

Set wb = ThisWorkbook
Set wb1 = Workbooks.Open("C:\Users\zhan2383\Downloads\161 Companies\Test\2010-2014.xlsx")
Set wb2 = Workbooks.Open("C:\Users\zhan2383\Downloads\161 Companies\Test\2000-2009.xlsx")
Set ws1 = wb1.Worksheets(1)
Set ws2 = wb2.Worksheets(1)

For Each ws In wb.Worksheets
   If Not ws.Name = "Sheet1" Then ws.Delete
Next ws

ws1.Copy After:=wb.Worksheets(wb.Worksheets.Count)
wb.Worksheets(wb.Worksheets.Count).Name = "Evaluation 2010-2014"

ws2.Copy After:=wb.Worksheets(wb.Worksheets.Count)
wb.Worksheets(wb.Worksheets.Count).Name = "Evaluation 2000-2009"

wb.Worksheets.Add After:=Worksheets(Worksheets.Count)
wb.Worksheets(Worksheets.Count).Name = "Calculation"

Set wsCla = wb.Worksheets("Calculation")
Set ws1014 = wb.Worksheets("Evaluation 2010-2014")
Set ws0009 = wb.Worksheets("Evaluation 2000-2009")

wsCla.Rows(1).Value = wb.Worksheets(2).Rows(1).Value
wsCla.Columns(1).Value = wb.Worksheets(2).Columns(1).Value

lastRow = wsCla.UsedRange.Rows(wsCla.UsedRange.Rows.Count).Row
lastCol = wsCla.UsedRange.Columns(wsCla.UsedRange.Columns.Count).Column


For k = 2 To lastRow
   For g = 2 To lastCol
      wsCla.Cells(k, g) = ws1014.Cells(k, g) - ws0009.Cells(k, g)
   Next g
Next k
```

```
'Classification I: quarterly averaged comparision
wb.Worksheets.Add After:=Worksheets(Worksheets.Count)
wb.Worksheets(Worksheets.Count).Name = "Final"
Set wsFin = wb.Worksheets("Final")

wsFin.Columns(1).Value = wb.Worksheets(2).Columns(1).Value
wsFin.Columns(1).Font.Bold = True
wsFin.Cells.HorizontalAlignment = xlCenter
wsFin.Cells(1, 1).Value = " Company "
wsFin.Cells(1, 2).Value = " Class by averaged comparison"

'Pick averaged attributes records to set standards
a = 15 'EPS
b = 16 'EPS diluted
c = 17 'Div
d = 18 'Average price
' e = 23 'BVPS

For k = 2 To lastRow
      If wsCla.Cells(k, 15) >= 0 And _
         wsCla.Cells(k, 16) >= 0 And _
         wsCla.Cells(k, 17) >= 0 And _
         wsCla.Cells(k, 18) >= 0 Then
'And _ wsCla.Cells(k, 23) >= 0 Then

         wsFin.Cells(k, 2) = 1
         Else
         wsFin.Cells(k, 2) = 0

      End If

Next k

class1 = Application.WorksheetFunction.Sum(wsFin.Columns(2))

MsgBox "Classified " & " " & lastRow - 1 & " companies." & vbLf _
& "Business improved : " & class1 & vbLf _
& "No improvement: " & lastRow - 1 - class1

'Classification II: specific standards
With wsFin
    .Cells.WrapText = True
    .Cells.HorizontalAlignment = xlCenter
    .Cells.Font.Bold = True
```

```
    .Cells(1, 3) = "Industry"
    .Cells(1, 4) = "Years of Deficits 2000-2009"
    .Cells(1, 5) = "Years of 0 Div 2000-2009"
    .Cells(1, 6) = "P_2009 - P_2000"
    .Cells(1, 7) = "Years of Deficits 2010-2014"
    .Cells(1, 8) = "Years of 0 Div 2010-2014"
    .Cells(1, 9) = "P_2014 - P_2010"
    .Cells(1, 10) = "Class 2000 - 2009"
    .Cells(1, 11) = "Class 2010 - 2014"
    .Cells(1, 12) = "Misclassification"

End With

Set wbT1 = wb2
Set wbT2 = wb1

rowNo1 =
wbT1.Worksheets(2).UsedRange.Rows(wbT1.Worksheets(2).UsedRange.Rows.Count).
Row
rowNo2 =
wbT2.Worksheets(2).UsedRange.Rows(wbT2.Worksheets(2).UsedRange.Rows.Count).
Row

' Ignore 2008 - 2009 record
For comp = 2 To wbT1.Worksheets.Count

   def1 = 0
   noDiv = 0
   y = 0

   For a1 = 6 To rowNo1 Step 4

      Set rng1 = wbT1.Worksheets(comp).Cells(a1, 15).Resize(4, 1) ' EPS
      Set rng2 = wbT1.Worksheets(comp).Cells(a1, 17).Resize(4, 1) ' Div
      Set rng3 = wbT1.Worksheets(comp).Cells(a1, 18).Resize(4, 1) ' Price

      If Application.WorksheetFunction.Sum(rng1) <= 0 Then def1 = def1 + 1
      If Application.WorksheetFunction.Sum(rng2) <= 0 Then noDiv = noDiv + 1
      aveP = Application.WorksheetFunction.Average(rng3)

      'If a1 = 2 Then P2009 = aveP
      'If a1 = 38 Then P2000 = aveP
      If a1 = 6 Then P2008 = aveP
      If a1 = 38 Then P2000 = aveP
```

```
    Next a1

    ThisWorkbook.Worksheets("Final").Cells(comp, 4) = def1
    ThisWorkbook.Worksheets("Final").Cells(comp, 5) = noDiv
    ThisWorkbook.Worksheets("Final").Cells(comp, 6) = P2008 - P2000

    If def1 > 0 Or noDiv > 0 Or P2008 - P2000 < 0 Then
ThisWorkbook.Worksheets("Final").Cells(comp, 10) = 0 _
    Else ThisWorkbook.Worksheets("Final").Cells(comp, 10) = 1

Next comp


' Check 2010 - 2014
For comp = 2 To wbT2.Worksheets.Count

    def1 = 0
    noDiv = 0
    y = 0

    For a1 = 2 To rowNo2 Step 4

        Set rng1 = wbT2.Worksheets(comp).Cells(a1, 15).Resize(4, 1)
        Set rng2 = wbT2.Worksheets(comp).Cells(a1, 17).Resize(4, 1)
        Set rng3 = wbT2.Worksheets(comp).Cells(a1, 18).Resize(4, 1)

        If Application.WorksheetFunction.Sum(rng1) <= 0 Then def1 = def1 + 1
        If Application.WorksheetFunction.Sum(rng2) <= 0 Then noDiv = noDiv + 1
        aveP = Application.WorksheetFunction.Average(rng3)
        If a1 = 2 Then P2014 = aveP
        If a1 = 18 Then P2010 = aveP


    Next a1

    ThisWorkbook.Worksheets("Final").Cells(comp, 7) = def1
    ThisWorkbook.Worksheets("Final").Cells(comp, 8) = noDiv
    ThisWorkbook.Worksheets("Final").Cells(comp, 9) = P2014 - P2010

    If def1 > 0 Or noDiv > 0 Or P2014 - P2010 < 0 Then
ThisWorkbook.Worksheets("Final").Cells(comp, 11) = 0 _
    Else ThisWorkbook.Worksheets("Final").Cells(comp, 11) = 1

Next comp

mis = 0
```

```
investWrong = 0
specWrong = 0

For h = 2 To wsFin.UsedRange.Rows(wsFin.UsedRange.Rows.Count).Row
   If wsFin.Cells(h, 10).Value <> wsFin.Cells(h, 11).Value Then
      wsFin.Cells(h, 12) = "Y"
      mis = mis + 1

      If wsFin.Cells(h, 10).Value > wsFin.Cells(h, 11).Value Then investWrong =
investWrong + 1
      If wsFin.Cells(h, 10).Value < wsFin.Cells(h, 11).Value Then specWrong =
specWrong + 1

   End If
Next h

totalMisRate = mis / (wbT1.Worksheets.Count - 1)

MsgBox "Companies: " & wbT1.Worksheets.Count - 1 & vbLf _
& "2000 - 2008 Records: " & rowNo1 - 5 & vbLf _
& "2010 - 2014 Records: " & rowNo2 - 1 & vbLf _
& "Overall Classification accuracy: " & Format(1 - totalMisRate, "Percent") & vbLf _
& "Mistaken investment: " & investWrong & vbLf _
& "Mistaken specualtion: " & specWrong

End Sub
```

Annualize Quarterly Financial Statistics

```
' Run in a CLEAN 2000YEAR workbook seperately

Sub yearRec()

Application.ScreenUpdating = False

Set wb2 = ThisWorkbook
Set wb1 = Workbooks.Open("H:\Big Data\Test\2000-2009.xlsx")

For comp = 2 To wb1.Worksheets.Count

wb1.Worksheets(comp).Range("AF1") = "Shares split adjusted EPS"
wb1.Worksheets(comp).Range("AG1") = "Shares split adjusted Div"
wb1.Worksheets(comp).Range("AH1") = "Shares split adjusted P"
wb1.Worksheets(comp).Range("AI1") = "Shares split adjusted P_High"
wb1.Worksheets(comp).Range("AJ1") = "Share split adjusted P_low"
```

```
wb1.Worksheets(comp).Range("AF2: AF41").Formula = "=N2/C2" 'Shares split
adjusted EPS
wb1.Worksheets(comp).Range("AG2: AG41").Formula = "=Q2/D2" 'Shares split
adjusted Div
wb1.Worksheets(comp).Range("AH2: AH41").Formula = "=R2/D2" 'Shares split
adjusted P
wb1.Worksheets(comp).Range("AI2: AI41").Formula = "=S2/D2" 'Shares split adjusted
P_High
wb1.Worksheets(comp).Range("AJ2: AJ41").Formula = "=T2/D2" 'Share split adjusted
P_low

wb2.Sheets.Add after:=wb2.Sheets(wb2.Sheets.Count)

With wb2.Worksheets(wb2.Sheets.Count)
   .Rows(1).Value = wb1.Worksheets(comp).Rows(1).Value
   .Name = wb1.Worksheets(comp).Name
   .Cells.HorizontalAlignment = xlCenter
   .Cells.WrapText = True
   .Range("O2:AJ20").NumberFormat = "0.0"
End With

'rowNo1 =
wb1.Worksheets(2).UsedRange.Rows(wb1.Worksheets(2).UsedRange.Rows.Count).Ro
w
'colNo1 =
wb1.Worksheets(2).UsedRange.Columns(wb1.Worksheets(2).UsedRange.Columns.Co
unt).Column

Set sht = wb1.Worksheets(2)
rowNo1 = sht.Cells(sht.Rows.Count, "A").End(xlUp).Row
colNo1 = sht.Cells(7, sht.Columns.Count).End(xlToLeft).Column

y = 0
   For r1 = 2 To rowNo1 Step 4
      For c1 = 2 To colNo1
         Set rng1 = wb1.Worksheets(comp).Cells(r1, c1).Resize(4, 1)
         yr = 2009 - y

         If c1 = 12 Or c1 = 13 Or c1 = 14 Or c1 = 15 Or c1 = 16 Or c1 = 17 Then
            val1 = Application.WorksheetFunction.Sum(rng1)
         Else
            val1 = Application.WorksheetFunction.Average(rng1)
         End If

        wb2.Worksheets(comp).Cells(y + 2, 1) = yr
        wb2.Worksheets(comp).Cells(y + 2, c1) = val1
```

```
      Next c1
         y = y + 1
   Next r1

Next comp


MsgBox "Annualized " & wb2.Worksheets.Count - 1 & " companies " & " @ " &
wb1.Name

End Sub

************************************************
'Combine companies row by row
Sub rowByRow()

On Error Resume Next
Dim wb As Workbook, sh As Worksheet

sheetsCount = ThisWorkbook.Worksheets.Count
Set ws1 = ThisWorkbook.Worksheets(1)
Set ws0 = ThisWorkbook.Worksheets(2)

' List Company's Name
For j = 2 To sheetsCount
   ws1.Cells(j + 1, 1).Value = ThisWorkbook.Worksheets(j).Name
Next j

' Label year
For A = 1 To 10
   ws1.Range(ws1.Cells(1, 35 * (A - 1) + 2), ws1.Cells(1, 35 * A + 1)).Value =
ws0.Cells(A + 1, 1).Value
Next A

For A = 1 To 10
   ws0.Range(ws0.Cells(1, 2), ws0.Cells(1, 36)).Copy
Destination:=ws1.Range(ws1.Cells(2, 35 * (A - 1) + 2), ws1.Cells(2, 35 * A + 1))
Next A


k = 3
For i = 2 To sheetsCount
   For R = 2 To 11
   Set ws2 = ThisWorkbook.Worksheets(i)
```

```
    ws2.Range(ws2.Cells(R, 2), ws2.Cells(R, 36)).Copy
Destination:=ws1.Range(ws1.Cells(k, 35 * (R - 2) + 2), ws1.Cells(k, 35 * (R - 1) + 1))
    Next R
k = k + 1
Next i
End Sub
```

Appendix B: PART OF THE NEURAL NETWORK, FEATURE SELECTION AND CROSS VALIDATION CODE

%% ================ Stage 1 Neural Network ================
%% Modified based on Andrew Ng' NN code (Andrew-Ng-Machine-Learning-Programming-solutions)**; NaN-Tb: A statistics toolbox** (Schloegl, 2010)**; J. Pohjalainen, O. Räsänen and S. Kadioglu's feature selection repository** (Feature selection code, 2015)**.**

```
% ================= Neural Network_Feature Selection_Cross Validation
% Adding Packages
addpath ('C:\Users\zhen\Desktop\Dissertation\Neural Network II');
clear; close all; clc

% Hand-Written Data, 500 iter 87%
%load('C:\Users\zhen\Desktop\Machine Learning\Neural Network II\ex4data1.mat');
%features = X;
%labels = y;
%N = size(features,1);

% Company Data
%dataset = load('C:\Users\zhen\Desktop\Dissertation\Logistic
Regression\350features\2000-2009 label 2000-2007.txt');
 dataset = load('C:\Users\zhen\Desktop\Dissertation\Logistic
Regression\350features\2000-2009 label 2010-2014.txt');
 dataset = dataset(randperm(400),:); % 400 samples 350 features 1 output
 features = dataset(:,1:350);% 400*350
 labels = dataset(:,351)+1;% 400*1
 N = size(features,1);% 400 samples

%% Parameters
Top = 45; % Top features
Q = 15; % Number of blocks for MI/SD
nodes = 5; % No. nodes
iter = 1000; % NN Iteration
lambda = 0; % NN Regularization


%% Stage 1: Known Key 18 Features: 2000-2007 EPS, Div; 2000 & 2007 P
%     sf1 = 85; sf2= 86;
%     sel_fea = [87,332,sf1,sf2];
%     while sf2 <= 330
%      sf1 = sf1 + 35;
%      sf2 = sf2 + 35;
%      sel_fea = [sel_fea sf1 sf2];
%      end;
```

```
%     (sel_fea);


% Cross-Validation
ncv = 5; %
cvblocksize = N/5; % 5 parts
dataorder = randperm(N); % Randomlize the sample index

% Estimated output vector
hypos_orig = zeros(N,1);
hypos_MI = zeros(N,1);
hypos_SD = zeros(N,1);
hypos_FSDD = zeros(N,1);
hypos_PPC = zeros(N,1);
hypos_RSFS = zeros(N,1);

for cvi=1:ncv
   fprintf('Cross validation partition %d/%d\n',cvi,ncv);

   % Testing Indices
   testidx = dataorder(((cvi-1)*cvblocksize+1):min(N,cvi*cvblocksize));
   % 1+step*(i-1) to min(total,step*i)
   % cvi = 1, dataorder(1: min(400,80))
   % cvi = 2, dataorder(81: min(400,160))
   ...
   % cvi = 5, dataorder(321: min(400,400))


   % Training indices
   trainidx = setdiff(1:N,testidx); % A has B does not, A > B
   trainidx = trainidx(randperm(length(trainidx))); % randomlize traing index

   % For RSFS, divide training data into two halves ("train + dev")
   Ntrain = length(trainidx); % 320 train
   trainidx1 = trainidx(1:round(Ntrain/2));
   trainidx2 = trainidx((round(Ntrain/2)+1):end);

   %% Mutual Information
%    [F_MI,W_MI] = MI(features(trainidx,:),labels(trainidx),Q);


   %% Statsical Dependence
%    [F_SD,W_SD] = SD(features(trainidx,:),labels(trainidx),Q);


   %% FSDD
```

```
    %[score,ind] = FS_NaN_FSDD (features(trainidx,:),labels(trainidx));
    % F_FSDD = ind';

     %% Pearson Partial Correlation
    %[score ind] = FS_NaN_Pearson(features, labels);
    % F_PPC = ind';


     %% Pure RSFS_NN
     %[F_RSFS,W_RSFS] =
RSFS_NN(features(trainidx1,:),features(trainidx2,:),labels(trainidx1),labels(trainidx2),'
max_iters',100,'max_delta',0.01);

     %% RSFS_NN_MI/SD, select the top 50 features and then use RSFS
     %[F_MI,W_MI] = MI(features(trainidx,:),labels(trainidx),Q);
     %[F_RSFS,W_RSFS] =
RSFS_NN(features(trainidx1,F_MI(1:Top)),features(trainidx2,F_MI(1:Top)),labels(trai
nidx1),labels(trainidx2),'max_iters',100,'max_delta',0.01);

     %[F_SD,W_SD] = SD(features(trainidx,:),labels(trainidx),Q);
     %[F_RSFS,W_RSFS] =
RSFS_NN(features(trainidx1,F_SD(1:Top)),features(trainidx2,F_SD(1:Top)),labels(trai
nidx1),labels(trainidx2), 'max_iters',100,'max_delta',0.01);


     %% Sequential Forward Selection (SFS)
%    k_sfs = 5:5:20; % Values of KNN k parameter over which Sequential Forward
Selection (SFS) is performed
%    t_sfs = 3;      % How many iterations is SFS run beyond the first detected
performance maximum?
%    [F_SFS,W_SFS] =
SFS(features(trainidx1,:),features(trainidx2,:),labels(trainidx1),labels(trainidx2),k_sfs,t_
sfs);


%    %% Sequential Floating Forward Selection (SFS)
%    k_sffs = 5:5:20; % Values of KNN k parameter over which Sequential Floating
Forward Selection (SFS) is performed
%    t_sffs = 3;      % How many iterations is SFFS run beyond the first detected
performance maximum?
%    [F_SFFS,W_SFFS] =
SFFS(features(trainidx1,:),features(trainidx2,:),labels(trainidx1),labels(trainidx2),k_sffs
,t_sffs);


    % Stage 1: Feature Scaling
```

```
    % X = features(trainidx,:); % Non-Scaling
    % X = (X - min(X))./(max(X)- min(X)); % Normalization [0,1]
    % X = X - mean(X)./std(X); % Standardization [-1,1]
    % X = sigmoid(X); %
    % hypos_orig(testidx) = NN(X,features(testidx,:),labels(trainidx),nodes,iter,lambda);


    % Stage 1: 18 Key Features (2000-2007 EPS, Div, 2000 P, 2007 P)
    % hypos_orig(testidx) =
NN(features(trainidx,sel_fea),features(testidx,sel_fea),labels(trainidx),nodes,iter,lambda
);


    % Stage 2
    % hypos_orig(testidx) =
NN(features(trainidx,:),features(testidx,:),labels(trainidx),nodes,iter,lambda);
    % hypos_MI(testidx) =
NN(features(trainidx,F_MI(1:Top)),features(testidx,F_MI(1:Top)),labels(trainidx),node
s,iter,lambda);
    % hypos_SD(testidx) =
NN(features(trainidx,F_SD(1:Top)),features(testidx,F_SD(1:Top)),labels(trainidx),node
s,iter,lambda);
    % hypos_FSDD(testidx) =
NN(features(trainidx,F_FSDD(1:Top)),features(testidx,F_FSDD(1:Top)),labels(trainidx
),nodes,iter,lambda);
    % hypos_PPC(testidx) =
NN(features(trainidx,F_PPC(1:Top)),features(testidx,F_PPC(1:Top)),labels(trainidx),no
des,iter,lambda);
    % hypos_RSFS(testidx) =
NN(features(trainidx,F_RSFS),features(testidx,F_RSFS),labels(trainidx),nodes,iter,lam
bda);


% Results
fprintf('Original %d features: %0.2f%% correct.\n',size(features,2),sum(hypos_orig ==
labels)/length(labels)*100);
fprintf('Best %d features from MI: %0.2f%% correct.\n',Top,sum(hypos_MI ==
labels)/length(labels)*100);
fprintf('Best %d features from SD: %0.2f%% correct.\n',Top,sum(hypos_SD ==
labels)/length(labels)*100);
fprintf('Best %d features from FSDD: %0.2f%% correct.\n',Top,sum(hypos_FSDD ==
labels)/length(labels)*100);
fprintf('Best %d features from FSDD: %0.2f%% correct.\n',Top,sum(hypos_PPC==
labels)/length(labels)*100);
fprintf('RSFS feature sets (%0.1f features on average): %0.2f%%
correct.\n',nfeat_RSFS,sum(hypos_RSFS == labels)/length(labels)*100);
```