

ตัวแบบประมาณการความสามารถในการสร้างรายได้ของหุ้นด้วยการจัดหมวดหมู่ ภูมิศึกษาตลาด
หุ้นไทย (SET)



บทคัดย่อและแฟ้มข้อมูลฉบับเต็มของวิทยานิพนธ์ตั้งแต่ปีการศึกษา 2554 ที่ให้บริการในคลังปัญญาจุฬาฯ (CUIR)
เป็นแฟ้มข้อมูลของนิสิตเจ้าของวิทยานิพนธ์ ที่ส่งผ่านทางบัณฑิตวิทยาลัย

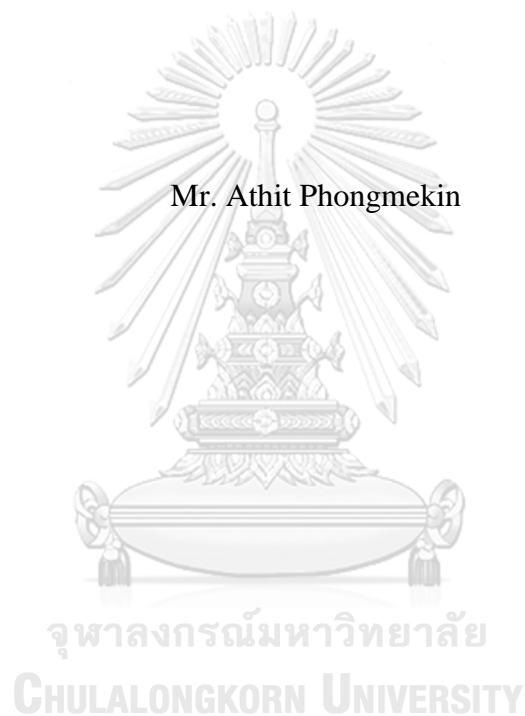
The abstract and full text of theses from the academic year 2011 in Chulalongkorn University Intellectual Repository (CUIR)
are the thesis authors' files submitted through the University Graduate School.

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมอุตสาหการ ภาควิชาวิศวกรรมอุตสาหการ
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2560
ลิขสิทธิ์ของจุฬาลงกรณ์มหาวิทยาลัย

Classification

Models for Stocks' Performance Prediction: A Case Study in Stock Exchange of Thailand
and (SET)

Mr. Athit Phongmekin



A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Engineering Program in Industrial Engineering
Department of Industrial Engineering
Faculty of Engineering
Chulalongkorn University
Academic Year 2017
Copyright of Chulalongkorn University

Thesis Title	Classification Models for Stocks' Performance Prediction: A Case Study in Stock Exchange of Thailand (SET)
By	Mr. Athit Phongmekin
Field of Study	Industrial Engineering
Thesis Advisor	Assistant Professor Pisit Jarumaneeroj, Ph.D.

Accepted by the Faculty of Engineering, Chulalongkorn University in
Partial Fulfillment of the Requirements for the Master's Degree

..... Dean of the Faculty of Engineering
(Associate Professor Supot Teachavorasinskun, D.Eng.)

THESIS COMMITTEE

..... Chairman
(Associate Professor Wipawee Tharmmaphornphilas, Ph.D.)

..... Thesis Advisor
(Assistant Professor Pisit Jarumaneeroj, Ph.D.)

..... Examiner
(Assistant Professor Naragain Phumchusri, Ph.D.)

..... External Examiner
(Siravit Swangnop, Ph.D.)



จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

อริศ ผ่องเมฆินทร์ : ตัวแบบประมาณการความสามารถในการสร้างรายได้ของหุ้นด้วยการจัดหมวดหมู่ กรณีศึกษาตลาดหุ้นไทย (SET) (Classification Models for Stocks' Performance Prediction: A Case Study in Stock Exchange of Thailand (SET)) อ.ที่ปริกษาวิทยานิพนธ์หลัก: ผศ. ดร. พิศิษฐ์ จารุมณีโรจน์, 93 หน้า.

ตลาดหลักทรัพย์ คือ ศูนย์รวมการซื้อขายหุ้นส่วนบริษัทที่อยู่ในหลักทรัพย์เมื่อผู้ซื้อและผู้ขายได้มีการตกลงราคาซื้อขายที่ทั้งสองฝ่ายยอมรับ โดยวัตถุประสงค์หลักของผู้ซื้อและผู้ขายคือการทำกำไรจากราคาส่วนต่างของหุ้นจากความคาดหวังในมูลค่าของหุ้นในอนาคตที่นักลงทุนแต่ละคนประมาณไว้ ปัจจุบันไม่มีกลยุทธ์การลงทุนหุ้นที่เป็นเอกฉันท์ว่าเป็นกลยุทธ์ที่ดีที่สุด ทั้งนี้การสร้างเกณฑ์ หรือกลยุทธ์การลงทุนขึ้นอยู่กับประสบการณ์ส่วนตัวของนักลงทุนคนเป็นส่วนใหญ่ เพื่อลดปัญหาความจำเป็นของการใช้ประสบการณ์ส่วนตัว งานวิจัยนี้ได้ใช้ดัชนีทางการเงิน และการแยกแยะอุตสาหกรรมในการสร้างตัวแบบประมาณการที่ใช้ในการอธิบายความสามารถในการสร้างรายได้ของหุ้นในตลาดหลักทรัพย์ประเทศไทย (SET) 2 ชนิด ได้แก่ 1. ตัวแบบที่ประมาณการว่าราคาหุ้นภายใน 1 ปีจะมีกำไรมากกว่าตลาดหรือไม่ 2. ตัวแบบที่ประมาณการว่าหุ้นภายใน 1 ปีจะมีกำไรเป็นบวกหรือลบ ด้วยการจัดหมวดหมู่ (classification model) แบบต่างๆ โดยความสามารถของตัวแบบจะถูกวัดผ่าน Area Under the Curve (AUC)

ข้อมูลหุ้นที่นำมาใช้ในการสร้างตัวแบบนั้นถูกแยกเป็น 6 หมวดตามชนิดของอุตสาหกรรม โดยแต่ละหมวดจะมีตัวแบบการจัดหมวดหมู่ (classification model) 4 แบบย่อย ซึ่งรวมเป็นตัวแบบทั้งสิ้น 48 แบบ ทั้งนี้ผู้วิจัยได้เลือกแต่ตัวแบบที่มีคะแนน AUC มากที่สุดของหุ้นในแต่ละหมวดมาใช้เป็นเครื่องมือในการเลือกหุ้นที่น่าสนใจจาก 582 บริษัทใน SET และเมื่ออ้างอิงมาตรฐาน AUC ของ Deloitte เห็นว่าตัวแบบแต่ละตัวนั้นมีประโยชน์อยู่ในระดับที่ยอมรับได้ หรือดี จากค่าของ AUC ระหว่าง 0.7 ถึง 0.8

ภาควิชา วิศวกรรมอุตสาหการ

ลายมือชื่อนิสิต

สาขาวิชา วิศวกรรมอุตสาหการ

ลายมือชื่อ อ.ที่ปริกษาหลัก

ปีการศึกษา 2560

5970351021 : MAJOR INDUSTRIAL ENGINEERING

KEYWORDS: DATA MINING / CLASSIFICATION MODEL / LOGISTIC REGRESSION / DECISION TREE / LINEAR DISCRIMINANT ANALYSIS (LDA) / K NEAREST NEIGHBORS / STOCK MARKET / THAI STOCK MARKET

ATHIT PHONGMEKIN: Classification Models for Stocks' Performance Prediction: A Case Study in Stock Exchange of Thailand (SET). ADVISOR: ASST. PROF. PISIT JARUMANEEROJ, Ph.D., 93 pp.

Within stock market, the objective of both stock buyers and sellers is to make profit on price difference based on their expectation on a company's current and future value. There is no investing strategy that is considered to be the best by experts, and investing decision criteria remain contingent upon an individual investor's experiences and bias. To address the challenges, this paper uses financial ratios and company's industry data to construct forecasting models that quantitatively describe the return on stock investment. Various classification models, including Logistic Regression (LR), Decision Tree (DT), Linear Discriminant Analysis (LDA) and K-nearest neighbor are used in the current study to find the best model with high predictive power. Two types of classification models for predicting whether a stock's one-year return in SET will outperform or underperform the SET Index and whether the return will be positive or negative are constructed in this study with Area Under the Curve (AUC) of Receiver Operating Characteristic (ROC) curve as measurement for models' performance. This study primarily focuses on the Stock Exchange of Thailand. The resulting AUCs demonstrates that the usefulness of these models can be rated as "Acceptable" to "Good" with AUC range from 0.7 to above 0.8 using Deloitte's standard.

Department: Industrial Engineering Student's Signature

Field of Study: Industrial Engineering Advisor's Signature

Academic Year: 2017

ACKNOWLEDGEMENTS

I would like to thank and offer my advisor, Assistant Professor Pisit Jurumaneeroj, a sincere gratitude for his encouragement and guidance. I really appreciate the commitment of my advisor who always responded promptly to my work with constructive comments. Also, I am really thankful for his contribution in helping me summarizing my paper for conference despite a tight deadline. Attending the ICEAST conference wouldn't be possible without my advisor's help.



CONTENTS

	Page
THAI ABSTRACT	iv
ENGLISH ABSTRACT.....	v
ACKNOWLEDGEMENTS	vi
CONTENTS.....	vii
TABLE OF FIGURES	1
LIST OF TABLES	3
Chapter I: Introduction.....	5
1.1 Thesis Background.....	5
1.2 Stock Market Investment Strategies.....	6
1.3 Challenges in Investment.....	7
1.4 Proposed Methodology.....	8
1.5 Objectives of the Study.....	9
Chapter II: Literature Review	10
2.1 Financial Ratios in Investing.....	10
2.2 Forecasting Models	10
2.2.1 Forecasting Stock Price	11
2.2.2 Forecasting Qualitative Stock Performance	11
2.2.3 Other Uses of the Classification Model in Finance	12
2.3.4 Models Used in the Study	13
A) Parametric models	13
B) Non-parametric models	15
Chapter III: Methodology	17
3.1 Variables.....	17
3.1.1 Independent Variables.....	17
3.1.2 Dependent Variable.....	21
3.2 Classification Models	22
3.2.1 CART Decision Tree.....	22
3.2.2 Logistic Regression (LR).....	27

	Page
3.2.3 Linear Discriminant Analysis.....	29
3.2.4 <i>K-Nearest Neighbor</i>	30
3.3 <i>Performance Measurement</i>	33
3.4 Data Preprocessing	39
3.5 Model Training Process	44
3.5.1 <i>Segmentation by Industry Sector</i>	45
3.5.2 <i>Qualitative Independent Variable Conversion</i>	47
3.5.3 <i>Optimization of Model Parameters</i>	47
3.5.4 <i>Independent Variable Selection</i>	47
Chapter IV: Results and Discussion	49
4.1 The Finance Sector	49
4.1.1 <i>Performance Relative to the SET Index Model</i>	49
4.1.2 <i>The Price Movement Model</i>	54
4.2 The Consumer Cyclical Sector	58
4.2.1 <i>Performance Relative to the SET Index Model</i>	58
4.2.2 <i>Price Movement Model</i>	62
4.3 The Consumer Non-Cyclical Sector.....	64
4.3.1 <i>Performance Relative to the SET Index Model</i>	64
4.3.2 <i>The Price Movement Model</i>	66
4.4 The Industrial Sector.....	69
4.4.1 <i>Performance Relative to the SET Index Model</i>	69
4.4.2 <i>The Price Movement Model</i>	71
4.5 Communication + Technology + Diversified Sectors	75
4.5.1 <i>Performance Relative to the SET Index Model</i>	75
4.5.2 <i>The Price Movement Model</i>	78
4.6 Basic Materials + Energy + Utilities Sectors.....	80
4.6.1 <i>Performance Relative to the SET Index Model</i>	80
4.6.2 <i>The Price Movement Model</i>	82
4.7 Remark.....	86

	Page
Chapter V: Conclusion.....	87
REFERENCES	90
VITA.....	93



TABLE OF FIGURES

Figure 1: ADVANC Price, Demand and Supply from www.settrade.com	5
Figure 2: ADVANC price from www.settrade.com	6
Figure 3: Graph of LDA with binary outcomes (x_i vs. Z_i)	14
Figure 4: Graph of Logistic Regression with binary outcomes (p_i vs. Z_i)	14
Figure 5: Classification Tree Example	16
Figure 6: KNN Example	16
Figure 7: Data Process Flow	17
Figure 8: Graph of Logistic Regression with binary outcomes (p_i vs. Z_i)	28
Figure 9: Graph of LDA with binary outcomes (x_i vs. Z_i)	29
Figure 10: KNN Example	31
Figure 11: Perfect Classification Model's ROC Curve	35
Figure 12: The ROC Curve for Random Guessing.....	35
Figure 13: ROC Curve Plotted from Table 7.....	38
Figure 14: ROC curve of the 3-year dataset (AUC = 0.552).....	42
Figure 15: ROC curve of the 4-year dataset from LR (AUC = 0.521)	42
Figure 16: ROC curve of the 5-year dataset from LR (AUC = 0.522)	43
Figure 17: Model Training/Testing Flow	44
Figure 18: Frequency of data by Industry Sector.....	46
Figure 19: Distribution of DIVIDEND_YEAR 3 in the Finance Sector	51
Figure 20: Distribution of PB_YEAR_2 in the Finance Sector.....	51
Figure 21: The Finance Sector by Industry Group	53
Figure 22: ROA_YEAR 2 of the Finance Sector.....	54
Figure 23: The GICS Subindustry in the Finance Sector	56
Figure 24: Interaction of the GIC Subindustry & NET_DEBT_TO_EQUITY_YEAR 3	57
Figure 25: LDA Regression Equation for Relative Performance Model in Consumer Cyclical.....	59
Figure 26: The GICS Industry of the Consumer Cyclical Sector	60

Figure 27: The GICS Subindustry of the Consumer Cyclical Sector	61
Figure 28: The GICS Subindustry for the Consumer Cyclical Sector (Price Movement).....	63
Figure 29: DT of the Relative Performance Model for the Consumer Non-Cyclical Sector	64
Figure 30: The ICB Industry in the Consumer Non-Cyclical Sector.....	68
Figure 31: Scatter Plot of Revenue Growth Year 1 and Profit Margin Year 2.....	70
Figure 32: The BICS Level 2 Industry Group (Industrial Sector).....	73
Figure 33: Profit Margin vs. Asset Turnover Scatterplot (Price Movement Model for the Industrial Sector)	74
Figure 34: ASSET_TURNOVER_YEAR 1 vs. ROIC_YEAR 2 (Relative Performance LR Model)	76
Figure 35: Industry Index Categories as Unique Integers	76
Figure 36: Asset Turnover by Industry Sector.....	77
Figure 37: Independent Variable for KNN Price Movement Models.....	78
Figure 38: ICB Industry vs. INCOME_GROWTH_YEAR 3	79
Figure 39: LDA Regression Equation for the Relative Performance Model.....	81
Figure 40: The GICS Subindustry for the Relative Performance Model.....	81
Figure 41: The GICS Subindustry for the Price Movement Model.....	84
Figure 42: ROE_YEAR 3 vs. DIV_YIELD_YEAR 2.....	85

LIST OF TABLES

Table 1: Example for Gini index and Entropy Calculation	23
Table 2: Example for Gini and Entropy Calculation with Independent Variable.....	24
Table 3: Example for finding the best split point m	26
Table 4: KNN Example based on Training Data	32
Table 5: KNN Example	33
Table 6: Confusion Matrix.....	33
Table 7: ROC Curve Construction Example	37
Table 8: 3-year Observation Period ($N = 3$)	39
Table 9: 4-year Observation Period ($N = 4$)	40
Table 10: 5-year Observation Period ($N = 5$)	40
Table 11: Relative Performance Model in Finance	49
Table 12: Independent Variables of LR for Relative Performance Model in Finance.....	50
Table 13: The Price Movement Model in the Finance Sector	54
Table 14: Independent Variables of LDA for Relative Price Movement Model in Finance.....	55
Table 15: The Relative Performance Model for the Consumer Cyclical Sector	58
Table 16: The Price Movement Model for the Consumer Cyclical Sector.....	62
Table 17: LDA Regression Equation for Price Movement Model in Consumer Cyclical	62
Table 18: The Relative Performance Model's AUC for the Consumer Non-Cyclical Sector	64
Table 19: Regression Coefficient for Relative Performance LDA Model (Consumer Non-Cyclical Sector).....	66
Table 20: The Price Movement Model for the Consumer Non-Cyclical Sector	66
Table 21: Regression Coefficient for Price Movement LDA Model (Consumer Non-Cyclical Sector)	67
Table 22: Relative Performance Model AUC for the Industrial Sector.....	69

Table 23: Independent Variables for the LDA Relative Performance Model (Industrial Sector)	69
Table 24: Price Movement Model for the Industrial Sector	71
Table 25: Independent Variables for the LDA Price Movement Model (Industrial Sector)	72
Table 26: Relative Performance Model AUC for Communication + Technology + Diversified Sectors	75
Table 27: Independent Variables for the LR Relative Performance Model (Communication + Technology + Diversified Sectors)	75
Table 28: Price Movement Models Communication + Technology + Diversified Sectors	78
Table 29: Relative Performance Model AUC for Basic Materials + Energy + Utilities Sectors	80
Table 30: Price Movement Models for Basic Materials + Energy + Utilities Sectors	82
Table 31: Independent Variables for the Price Movement LDA Model	83
Table 32: AUC Benchmark	86
Table 33: Summary of Top-Performing Models	87
Table 34: Summary of Important Observations by Sector	88

Chapter I: Introduction

1.1 Thesis Background

The stock market is a place where public company ownership is traded on an agreed upon price and volume between stock sellers and buyers. The objective of both buyers and sellers is to make profit on price difference based on their expectation on a company's current and future value. Typically, these investors (buyers and sellers) can create their expectation on the company's future price using a historical trend and the company's financial performances.

Investors in Thailand engage in stock trading through the Stock Exchange of Thailand. Figure 1 illustrates an example of real-time stock information of a telecommunication company ADVANC from www.settrade.com including the last price the transaction occurred labeled as "Last Trade"; "Bid/Volume Bid" represents the price and volume of the stock demand, and "Offer / Volume Offer" represents the price and amount of the stock supply.

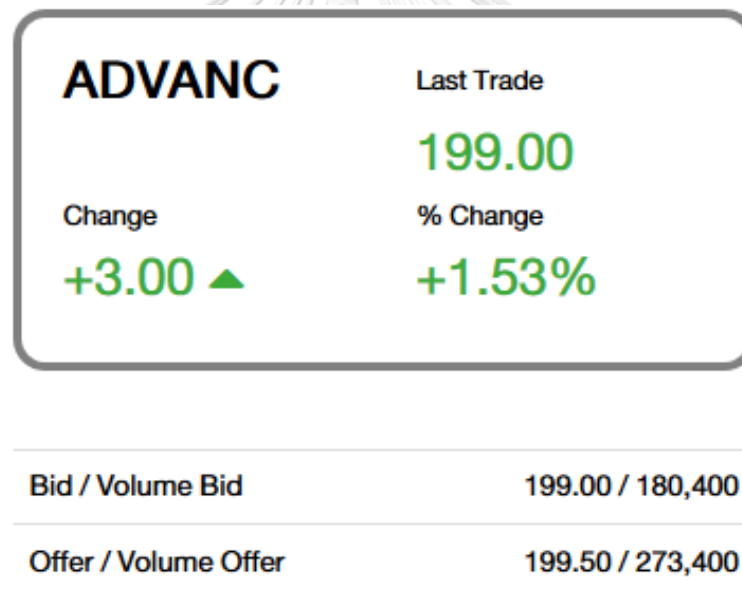


Figure 1: ADVANC Price, Demand and Supply from www.settrade.com

Investors engaged in trading can set their own volume and bid price for buying and selling stock. The transaction occurs when there is a match in prices; otherwise, investors have to wait until an offer–bid price match occurs. In the case of Figure 1, an investor can buy the stock immediately at the offer price of 199.50 THB per share, while seller can do the same by selling immediately at 199.00 THB per share as indicated by the bid price.

Prior	196.00	Volume (Shares)	9,088,581
Open	197.00	Value ('000 Baht)	1,798,591.97
High	200.00	Par Value (Baht)	1.00
Low	195.00	Ceiling	254.00
Average Price	197.90	Floor	137.50

Figure 2: ADVANC price from www.settrade.com

Other than providing the current price, demand and supply, Settrade's website also provides historical prices during a day for 10.00 am to 12.30 pm and 2.00 pm to 4.30 pm, including the total volume traded on that day, labeled as "Volume (Shares)"; the opening trading price; as well as upper and lower boundaries between which the price can move, labeled as "Ceiling" and "Floor", respectively.

Besides, investors also have access to a company's annual financial performance, financial ratios and an annual report that provide information of the company from many angles. Risk-averse investors typically perform extensive study on their company of interest, including financial performance, the nature of businesses and the competitive landscape to project how much the company will be worth in the future before making a decision to invest in the company's stock.

The challenge arises in that the investment decision remains subjective, and there is no clear-cut methodology in valuing a company's worth in the future. There are many valuation methods used by financial analysts to map the future stock price; however, such valuation models require various assumptions based on an analyst's knowledge and experience. Thus, the estimated price and recommendation vary from analyst to analyst. As for beginner investors who have no experience, they either study the market themselves or review financial analysts' recommendations before making an investment decision. Despite a large amount of available information, predicting the future price of a stock with accuracy to ensure profit remains a difficult task, especially for beginner investors.

1.2 Stock Market Investment Strategies

There are various investment strategies used by investors: one is technical analysis, which evaluates stocks and forecasts their movement from trading activity (supply and demand). The second method is called fundamental analysis. This second method takes a closer look at a company's financial performance and nature of business to find a stock with a price lower than its intrinsic value. Fundamental analysis is the root of value and growth investing.

Value investing is a popular investing strategy introduced by Benjamin Graham and adopted by many successful investors. Analysis of fundamental metrics is a critical component of value investing. Value investors analyze fundamental metrics and

formulate heuristic investing criteria for choosing a stock that has a price lower than its intrinsic value [1]. Although the proponent of this strategy asserts that value investing gives higher return, there is little evidence to support this claim [2], and predicting stock performance remains a complicated task because of unforeseen events such as insider trading and the participation of large players such as Foreign Institute Investors (FIIs) and Domestic Institute Investors (DIIs) [3]. Despite such complication, fundamental ratios are an objective measurement for company performance and have forecasting power on a stock's performance.

Another common investing strategy is called growth investment, in which an investor focuses on buying the stock of a company with high growth. Although value investing focuses on buying cheap stock with low P/E, growth investors buy a stock with high P/E [4] with expectation that the company will further grow in value. There is no consensus on which investment strategy is superior [4]. However, there are many studies that indicate that a value stock outperforms a growth stock [4].

The basic fundamental parameters for investment criteria include price-to-book value (P/B), price-to-earnings ratio (P/E), price-to-cashflow ratio, dividend yield, return on equity (ROE), return on asset (ROA) and return on capital [1]. For value investing, these ratios are typically benchmarked with the industry average to determine whether the stock price is affordable or expensive compared to those of other companies in the same industry. Additionally, these parameters can also be used for growth investing as ROA and ROE are measurements of profitability used to project the growth rate of a growth stock.

Evidently, fundamental analysis is a crucial part in both value and growth investing as financial ratios in the analysis give investors some idea about the future value of the companies they invest in.

1.3 Challenges in Investment

As mentioned in the previous section, there is no investing strategy that is considered to be the best by experts, and investing decision criteria remain contingent upon an individual investor's experiences and bias. Investors with experiences are more likely to predict the future stock's price more accurately than the inexperienced investors, as they experienced investors formulate better investing criteria for choosing a stock.

Currently, SET consists of stocks of 582 companies, and each of these companies have different financial performance and nature of business. To make an investment decision, investors have to review massive information regarding each company's financial performance and business model. To help investors make an investment decision easier, an objective tool is needed to screen out unattractive companies for investors, so that they can spend time reviewing companies with a higher chance of generating a return and acceptable investment risk. This tool is especially helpful to inexperienced investors when choosing stocks to review without formulating criteria that rely on subjective experience. Since the constructed models in this study only indicate the relationship between independent and dependent variables, it is recommended that investors should use these models only to screen stocks from 582 companies in SET for further analysis before investing, as investors have different

diversification requirements, degrees of risk they are willing to take, and fund availability.

1.4 Proposed Methodology

To address the challenges, financial ratios and company's sector/industry can be used to construct models that quantitatively describe the return on stock investment. The forecasting model analyzes how much each metric affects those returns that are likely to allow investors to make better and more objective investment decisions.

To do this, various studies have been reviewed to find models for constructing a predictive model on stock performance using fundamental financial ratios. The models are used to avoid subjective rules of thumb created by investors' personal experiences in choosing a stock with a profitable performance. There are many studies in non-Thai stock markets in which financial ratios are used in a classification model to predict price movement and individual stocks' performance relative to the market. Thus, various classification models, including Logistic Regression (LR), Decision Tree (DT), Linear Discriminant Analysis (LDA) and K-nearest neighbor were used in the current study to find the best model with high predictive power using fundamental financial ratios and company's sector/industry classification as input and performance as output.

The purpose of the current study is to perform a quantitative analysis of fundamental ratios and company's sector/industry using parametric and non-parametric statistical methods to construct various forecasting models for stock performance in Stock Exchange of Thailand (SET). There are two types of the classification model, with the first type being the classification model to predict whether the stock "outperformed" or "underperformed" the SET market, and the second type being whether the stock generates "positive" or "negative" return. These models are compared for accuracy to allow for better investing decision.

In the current study, the concept of the Receiver Operating Characteristic (ROC) curve is introduced as a measurement for models' performance. As its Cartesian coordinates, the ROC curve is a plot of sensitivity (true positive rate) on the y-axis and fall-out (false positive rate) on the x-axis, where sensitivity and fall-out are measurements from applying testing data set to a binomial classification models mentioned above. A numerical measurement acquired from the ROC curve for measuring the classification model's performance is called Area Under the Curve (AUC) and ranges from 0 to 1; a higher number means a better classification model. The ROC curve is commonly used as performance measurement for the classification model applied for credit scoring. Banks use a credit score converted from a classification model to decide creditworthiness of loan applicants, and the ROC curve is a measurement of usefulness of the classification model. The ROC curve is useful for the two types of the classification model presented above because what investors expect from these two models is the stocks classified as both outperformed and positive; thus, investors need to primarily focus on how many instances the stock predicted as outperformed is actually outperformed (true positive rate) and how many instances the stock predicted as outperformed is actually underperformed (false positive rate) to understand the risk they are taking when using the models. By contrast, examining the stocks predicted in

the classification model as an underperformed and negative return is not productive as investors do not invest in these stocks.

1.5 Objectives of the Study

1. To construct various predictive classification models using financial ratios and industry segmentations as input to predict whether a stock's one-year return will outperform or underperform the SET index.
2. To construct various predictive classification models using financial ratios and industry segmentations as input to predict whether a stock's one-year return will be positive or negative.
3. To find the most useful models from objectives 1 and 2 to be used for choosing stocks for investment using AUC of ROC curves from all models as a measurement for the models' usefulness.



Chapter II: Literature Review

This chapter reviews some studies regarding uses of financial ratios and forecasting models in forecasting stock movement.

2.1 Financial Ratios in Investing

One of value investing strategies was introduced by Sareewiwatthana for screening underpriced stock using fundamental parameters such as P/E, P/B, dividend yield and ROE [1]. With a 15 years' test period, Sareewiwatthana's study concluded that the proposed screening method produced significantly higher annual and total portfolio return than the SET index [1]. Despite great performance, the stock selection rules do not include any quantitative analysis. The study demonstrated an opportunity to further investigations into the financial ratios to validate whether the ratios used as stock selection criteria have a positive or negative relationship with a stock's return using parametric and non-parametric statistical methods.

In another paper, a study was conducted on the Japanese stock market. Chan et al. observed the return of Japanese stock through the behaviors of four fundamental variables including earning yield, size, book-to-market ratio and cashflow yield [5]. Through statistical analysis, the study concluded that these fundamental variables have a significant relationship with expected return in the Japanese stock market, with the book-to-market ratio and cashflow yield having the most effect [5]. Unlike Paiboon [1], Chan et al. analyzed the fundamental variables with a statistical method called Seemingly Unrelated Regression (SUR). The limitation of the study is that the results only describe the Japanese stock market, and there are many other financial ratios that can be included for higher-dimension analysis.

Although Paiboon [1] and Chan et al [5] used only few fundamental financial ratios in stock selection, there is still a need for converting various financial indicators into fundamental financial strength to help investors understand the investment-worthiness of a company. To do so, Edirisinghe and Zhang performed data envelopment analysis and developed a new metric called the Relative Financial Strength Indicator (RFSI), which has a high correlation with the stock price return [6]. The RFSI includes financial ratios that measure a company's profitability, asset utilization efficiency, current value, growth and liquidity. The limitation of the study is that although RFSI is useful for selecting a stock with high return, the measurement does not directly help in predicting a return relative to a market that represents the opportunity cost for investing in the stock market.

2.2 Forecasting Models

The studies discussed in this chapter can be categorized into three types: forecasting of continuous variables such as a stock's price, forecasting of binary variables such as a stock's performance with two qualitative outcomes or a classification model, and other use of the same classification model in the financial world.

2.2.1 Forecasting Stock Price

To forecast a closing price of one particular stock, Grigoryan [7] introduced principle component analysis (PCA) with artificial neural network (ANN) and technical parameters as independent variables [7]. Using the mean square error (MSE) as a performance measurement, this study demonstrated that a combination of the PCA-ANN model is worth exploring and that the proposed model can be used for financial time series forecasting. Hakob's study focused on predicting the price of one company's stock using only technical parameters to forecast the price movement. It also highlighted the use of a statistical method on technical analysis, which is another investment strategy centered around the demand and supply of a stock; fundamental analysis focuses on financial ratios that reflect a company's performance.

An example of the price forecasting model on companies in the Stock Exchange of Thailand (SET) was done by Wanrapee Banchuenvijit [8], who used multiple linear regression (MLR) with the current ratio, debt-to-equity ratio, net profit margin, asset turnover and agriculture production index as independent variables to forecast the price of four different agriculture firms [8]. In contrast to Hakob's paper [7], a study by Wanrapee [8] constructed four different MLR models for each of the four companies using parameters only in fundamental analysis. The purpose of the current study is to examine fundamental financial ratios that have a statistically significant relationship with stock prices of companies in the agriculture sector using the p-value and significant level in MLR as measurements. Wanrapee [8] concluded that financial ratios had consistent effects on a stock's price, as stated in another paper on fundamental analysis with current ratio, net profit margin and asset turnover, which had a positive correlation with the stock price, while the debt-to-equity ratio had a negative correlation with the price as it represented a company's risk.

Unlike in the above-mentioned studies, stock price is affected by news events that affect investors' expectations and thus demand, supply and price; Hiral et al. [9] incorporated news into MLR and ANN to forecast a stock's open price [9]. Using semantic analysis to incorporate news impact and technical parameters, the study attained accuracy of 82% and 70% for MLR and ANN models, respectively.

2.2.2 Forecasting Qualitative Stock Performance

Complementary to the work of Hiral et al. [9], Alostad and Davulcu [10] incorporated news from social media platforms such as Twitter into support vector machine (SVM) and logistic regression (LR) classification models to predict the movement of stock price direction [10]. The results indicated that using LR and incorporating news from social media provides high accuracy above 70%.

Dutta et al. [11] used LR to predict whether a selected stock outperformed or underperformed the Indian Stock Market index (NIFTY) using 12 months' financial ratios [11]. In contrast to Hiral et al., who incorporated news impact and technical parameters as the classification model input, the eight financial ratios used in this study are primarily fundamental ratios, which indicate a company's performance. Moreover, the prediction model in a study by Dutta et al. indicates whether a company is likely to generate return above or below the market. Unlike other studies so far, the model is practical to investors because it benchmarks individual stock's return to the overall

market return, which represents the opportunity cost of investing in the market. The paper demonstrated that the LR prediction model can achieve 74.6% prediction accuracy while using only fundamental financial ratios.

Tsai and Wang [12] utilized ANN and decision tree (ANN-DT) classification methods to build a predictive model on the rise and fall of stocks in Taiwan's electronics industry [12]. The paper demonstrated that the hybrid model of ANN-DT using fundamental, technical and macroeconomic parameters provides superior accuracy than either of the single models ANN and DT. The study is distinct from previous works mentioned because the authors use hybrid classification models that incorporate fundamental, technical and macroeconomic ratios to predict the price movement (up or down) of stocks. Unlike Dutta et al. [11], who predicted whether a stock underperformed or outperformed the market, the current study provides a model that predicts a positive-return stock for investors. This prediction is important because a stock can outperform a market and yet generate a negative return if the market also has a negative return. However, the limitation of the model is that the prediction only applies to stocks of companies in one industry.

In Tufekci [13] used the classification models for forecasting up and down movement of the Istanbul Stock Exchange National 100 (ISEN 100) index using macroeconomic indicators, gold price, oil price, exchange rate, stock price index in other countries and the technical ISEN 100 data [13]. The researchers used three different methods including LR, bagging logistic regression (BLR) and ANN. Unlike other studies, the study provided models for predicting the movement of the ISEN 100 index, not individual stocks, thus not using any technical or fundamental parameters. This demonstrates another use of the classification model in the stock market as the ability to predict the ISEN 100 index's movement that can help investors formulate investment decisions.

2.2.3 Other Uses of the Classification Model in Finance

Other than predicting a stock's price and return, the above-mentioned methodologies (ANN, SVM, LR, and DT) can also be utilized to assess the financial condition of a company. Shie et al. found that the incorporated particle swarm optimization algorithm with SVM yields higher prediction accuracy for predicting whether a banking company is in financial distress [14]. In this study, the training set consisted of 54 banking companies' fundamental ratios and a label for whether the banks are in financial distress. The authors labeled each company as in financial distress based on an auditor's report, and the absence of such report classified the bank as not in financial distress. Shie et al. [14] also demonstrated the use of fundamental financial ratios instead of the stock price in the classification model to predict a company's financial position.

In a more extensive modeling study, Cheng and Wang [15] used various attribute selection and classification methods to predict financial distress of over-the-counter electronic firms. In contrast to Shie et al., the researchers defined financial distress differently as they used *earnings before tax/total asset interest ratio* as an indicator instead of an auditor's report. Moreover, the study was performed in a different sector (not the banking industry). Cheng and Wang used companies' financial ratios as input for the prediction model, and the result indicates that using LR for variable selection

followed by a rough set theory provides the best prediction accuracy over other combinations [15]. This proves once again that financial ratios are useful for predicting various outcomes (stock price, return and financial distress) regardless of a company's sector.

Other than predicting financial distress, Sinthupundaja et al. adopted LR, ANN and SVM to predict whether the return on asset (ROA) of various companies in SET is above average by including fundamental ratios and external factors (GDP, Economy, Social Behaviors and Technology) as independent parameters [16]. The result of the study demonstrates that using machine learning methods (ANN and SVM) provided better accuracy than LR [16].

From literature reviews, various observations can be made:

- There are very few prediction models employed in SET, and none of them help investors to choose stocks that outperform the SET index and focus on investment return.
- Machine learning methodologies (SVM and ANN) are used more frequently than non-machine learning technologies (LR and DT).
- No literature uses simpler classification methods such as K-nearest neighbor or linear discriminant.
- Most of the literature uses accuracy as a performance measurement for modeling and the accuracy range around 70–80%.
- Companies' financial ratios are useful as predictors for various outcomes such as stock price, stock price movement, a stock's return relative to the stock market or financial distress regardless of industry in which the sample companies are.

In the current study, the researcher proposes prediction models to classify stock performance in SET. The model uses companies' 3–5-year financial ratios as independent variables and a stock one-year performance as outcome variables.

2.3.4 Models Used in the Study

The models utilized in the study consist of two parametric and two non-parametric statistical models.

A) Parametric models

The parametric models used in the current study are logistic regression and linear discriminant analysis. These two methods are regression analysis in which a hyperplane is created from a regression equation to best separate the two dichotomous classes (dependent variables). For linear discriminant analysis, the hyperplane takes the form of a linear plane.

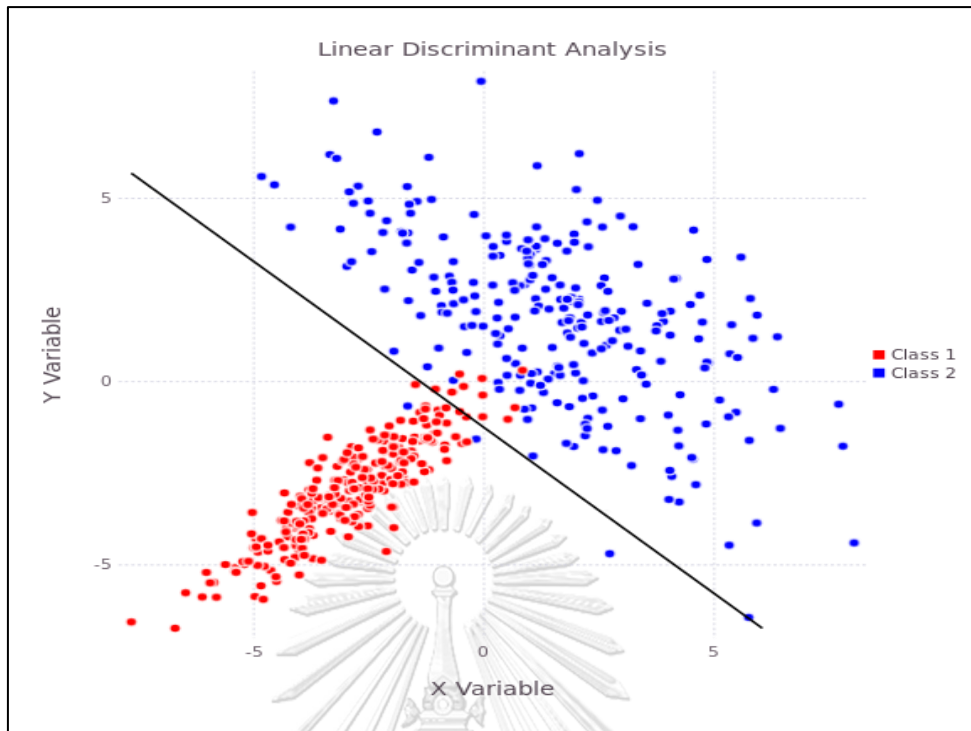


Figure 3: Graph of LDA with binary outcomes (x_i vs. Z_i)

The regression equation (Z-score) takes the form

$$Z = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k \quad (1)$$

By contrast, the hyperplane of the logistic regression takes the logarithmic shape as the regression equation takes the form of a log odd ratio.

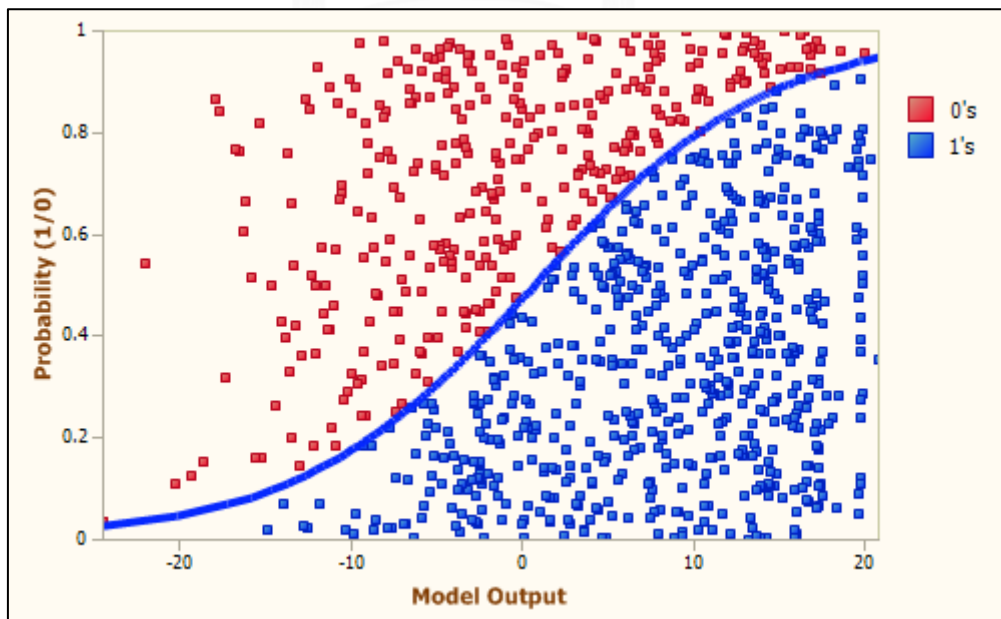


Figure 4: Graph of Logistic Regression with binary outcomes (p_i vs. Z_i)

It has the following regression equation:

$$z_i = \ln \left(\frac{p_i}{1-p_i} \right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k \quad (2)$$

where p_i is the probability of dichotomous events to occur describe in the dependent variable, $\beta(s)$ are coefficients calculated using a maximum likelihood function, k is the number of independent variables, and Z_i is called the odd ratio.

B) Non-parametric models

A decision tree is a standard classification method in which independent variables are selected to provide the most significant split into two categories (outperformed or underperformed). The tree consists of a root node, branches as well as parent, child and leaf nodes. The parent node represents an independent variable (Node 0 in Figure 5), and its branches split into child nodes that represent different categories (Node 1 and Node 2 in Figure 5). The root node is the first or topmost decision node (Node 0 in Figure 5). The leaf node represents the classification decision or the last child node. The decision tree node grows by measuring information gain, gain ratio or Gini gain to create the best possible categorization. As more nodes grow, the information regarding classification becomes purer. Every leaf node consists of a training sample categorized into dichotomous groups according to their dependent variable labels (“good” or “bad” in Figure 5), and the leaf is classified into the category of the majority class. Taking Figure 5 as an example, Node 1 is classified as “bad”, with 82.09% of the sample categorized as “bad”. By contrast, Node 2 is classified as “good”, with 70.38% of the sample categorized as “good”. If the proportion of the sample in the leaf node in either category is zero, then the node is defined as a pure node and the information is zero [17].

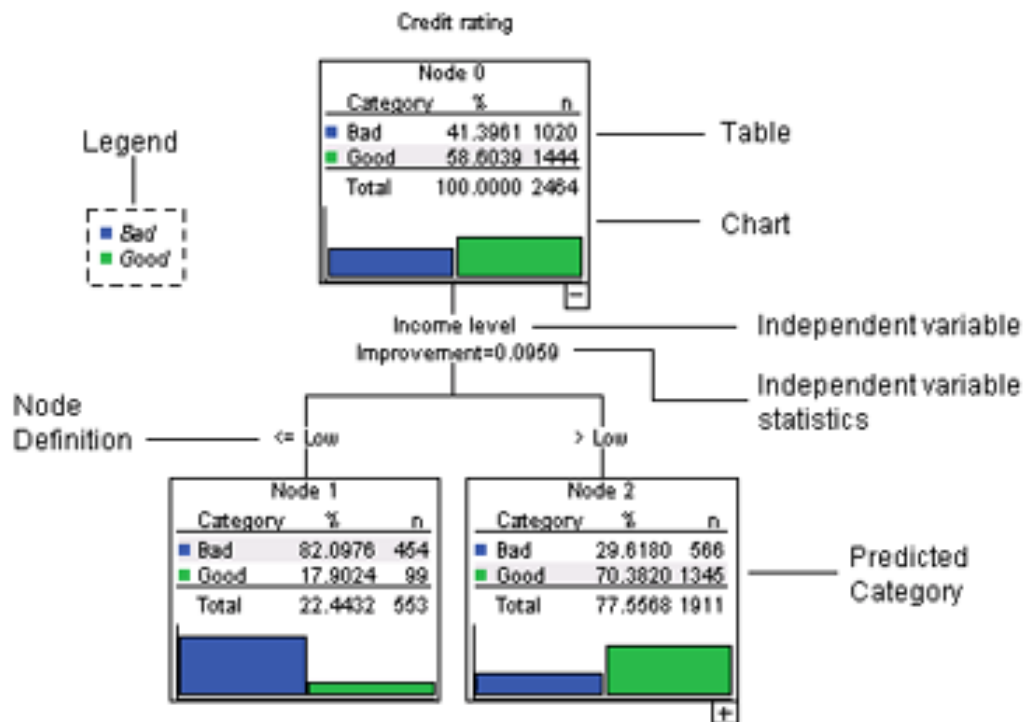


Figure 5: Classification Tree Example

The second non-parametric model is called the K-nearest neighbor, in which the classification of a testing instance is determined by the majority class of its K nearest neighbors. Nearest neighbors can be found from calculating distances between the testing instance and all training instances by using the distance function and ranking those distances in ascending order.

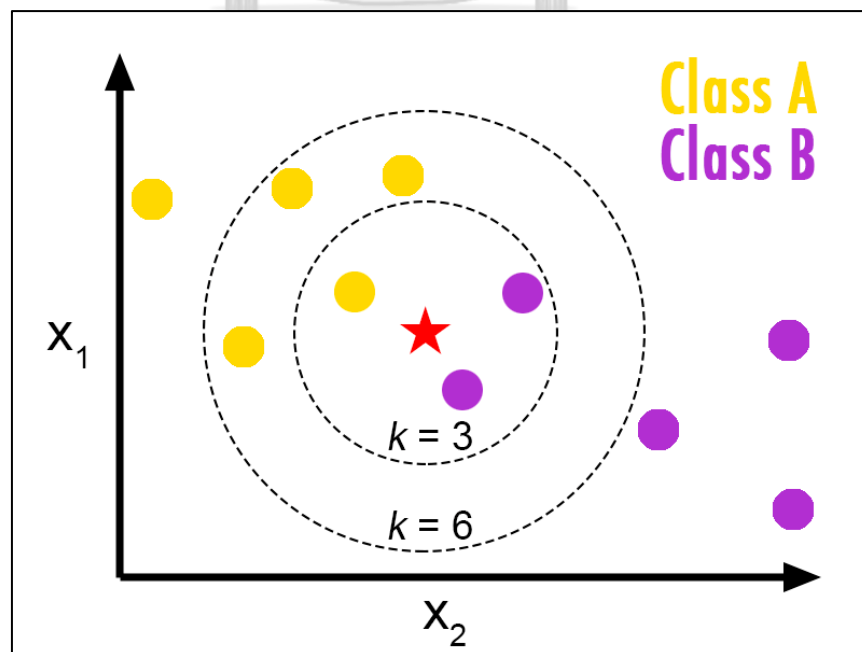


Figure 6: KNN Example

Chapter III: Methodology

The big picture of methodology could be divided into four major stages. First, the data of companies' stocks and financial ratios in SET are retrieved from a Bloomberg terminal into an Excel spreadsheet. Within Excel, the data are arranged into a row of instances in which one instance (one row) consists of a company's name, 33–55 financial ratios of that company (depending on observation period) and two prices for calculating the 1-year return (see Table 8, Table 9 and Table 10). Second, these data are preprocessed to remove all instances without a value, and both dependent variables are then computed for all instances in Excel. After the removal of empty instances and computation of dependent variables, the data are imported into Rapidminer for further processing.

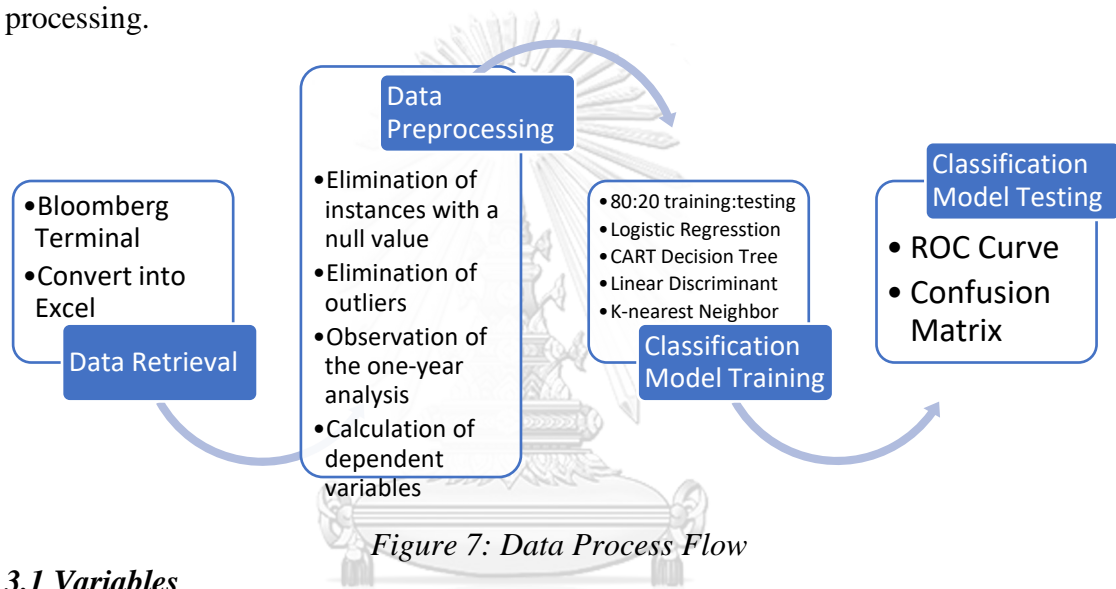


Figure 7: Data Process Flow

3.1 Variables

3.1.1 Independent Variables

For training the models, 3–5 years of 11 different financial ratios are used as independent variables; thus, the number of independent variables ranges from $3 \times 11 = 33$ to $5 \times 11 = 55$. The financial ratios are chosen based on the ratios used in other studies and their availability. SET includes all types of company in many sectors, some of which do not have a certain ratio because of the nature of business and accounting methods. For example, Enterprise Value/Earnings-before-interest-tax-depreciation-amortization (EV/EBITDA) is not available in the banking sector since earnings are primarily from interest payment, and the banking sector is not capital intensive to have depreciation or amortization cost. To be inclusive, the chosen ratios are common in all industries and are aimed at avoiding omission of any particular sector in SET. These independent variables include the following:

1. The price-to-Earnings ratio (PE), which is calculated by dividing the current price of a stock by earnings per share:

$$PE = \frac{\text{Current Price}}{\text{Earning per Share (EPS)}}$$

This ratio indicates how much investors are willing to pay for each unit of earnings by the company. Assuming the earnings are constant annually, this ratio would represent the number of years it would take for a company to earn enough to satisfy the current price. PE also represents a prospective intrinsic value of the company by investors, as low PE compare to industrial average can be seen as the company having a high intrinsic value or the current price of company's ownership being low [1].

2. Price-to-book (PB) is calculated by dividing the current price by book value per share. Book value of a company comes from total equity in the company's balance sheet:

$$PB = \frac{\text{Current Price}}{(\text{Asset} - \text{Liability})/\text{no. of share}}$$

Like PE, PB also represents how much investors perceive a company's intrinsic value [1]. A low PB ratio means that the company is undervalued and has a high potential of gaining return on investment as it grows. Similarly, PB is typically compared with the average PB of other companies in the same industry to decide whether the value is low or high.

3. Return on Equity (ROE) is the amount of earnings per unit shareholder equity invested in the company:

$$ROE = \frac{\text{Net Income}}{\text{Asset} - \text{Liability}} \times 100$$

The ratio measures a company's profitability, and higher ROE means that the company is more profitable and thus worthy of investing in.

4. Return on Invested Capital (ROIC) is calculated as

$$ROIC = \frac{\text{Operating Income}}{\text{Average Invested Capital}} \times 100$$

where Average Invested Capital = (Invested Capital at the Beginning of the Year + Invested Capital at the End of the Year) / 2.

ROIC is a measurement for a company's efficiency at allocating capital to generate a return. This capital includes property, plant and equipment or land in which a company invested to conduct business. In a sense, the ratio conveys how the company uses its money to generate profit. This ratio is compared with Weighted Average Cost of Capital (WACC), which represents the opportunity cost of holding capital. If $ROIC > WACC$, then the company uses the capital effectively as it generates a return higher than the opportunity cost.

5. Return on Asset (ROA) is a profitability indicator that measures a company's earnings relative to assets. Investors can gauge how efficient the company uses its assets to generate earnings. ROA is calculated as

$$ROA = \frac{\text{Net Income}}{\text{Asset}} \times 100$$

This ratio is typically compared with ROAs of other companies in the same industry.

6. Asset Turnover is an efficiency ratio that measures the amount of revenue generated per monetary unit of asset or how efficient the company deploys assets:

$$\text{Asset Turnover} = \frac{\text{Revenue}}{\text{Average Total Asset}}$$

where Average Total Asset = (Asset at the Beginning of the Year + Asset at the End of the Year) / 2.

Asset Turnover typically varies depending on the industry. Therefore, its comparison has to be made with companies in the same industry.

7. Revenue Growth is the percentage increase in revenue generated annually. This measurement indicates how the company's revenue grows over time. An investor looks for consistent revenue growth when choosing to invest in a company. The measurement also tells the investor and manager how well the company is doing over a year.

$$\text{Revenue Growth} = \frac{\text{Revenue}[t] - \text{Revenue}[t-1]}{\text{Revenue}[t-1]}$$

In most cases, a consistent decline in revenue is a bad sign for investors.

8. Observing Net Income Growth with Revenue Growth gives an investor a sense of a company's ability to control cost. For example, an increase in revenue but a decrease in the bottom line (Net Income) demonstrates that the company spends too much in generating revenue. Net Income Growth can be computed as follows:

$$\text{Net Income Growth} = \frac{\text{Net Income}[t] - \text{Net Income}[t-1]}{\text{Net Income}[t-1]}$$

9. Net Debt to Equity ratio is the measurement of a company's interest bearing debt relative to the value of stock. Risk-taking investors perceive high Net Debt to Equity as a sign of borrowing money for aggressive expansion to increase a company's value, whereas risk-averse investors perceive it as increased risk in liquidation in case an expansion plan fails. By contrast, too low Net Debt to Equity can also mean that the company is not financing its growth properly. Net Debt to Equity is calculated as

$$\text{Net Debt to Equity} = \frac{\text{Total Interest Bearing Debt} - \text{Cash}}{\text{Total Equity}} \times 100$$

Observing the ratio over time can also demonstrate a company's ability to finance growth. For example, a huge increase in Net Debt to Equity followed by a gradual decrease over time can indicate that the company is growing and able to pay back borrowed money.

10. Profit Margin is another profitability ratio that measures how much a company earns for every unit of revenue:

$$\text{Profit Margin} = \frac{\text{Net Income}}{\text{Revenue}} \times 100$$

This ratio is important because even though the revenue increases over time, the company may not earn more money. Profit Margin considers a company's cost structure and how much it pays to achieve revenue growth. A decrease of the profit margin with growing revenue could mean that the company is paying too much to generate additional revenue.

Profit Margin also measures a company's performance relative to other companies in the same industry or business model. Compared to that of a company with the same business model, lower profit margin could mean that the company does not use its expense as efficiently as the company with higher profit margin.

11. Dividend Yield indicates how much a company pays out as dividend each year relative to share price:

$$\text{Dividend Yield} = \frac{\text{Annual Dividend per Share}}{\text{Current Price per Share}}$$

This ratio measures how much cash goes to investors as they bought and hold ownership of a company. Generally, if a company has a quarter with good financial performance, the board of directors might choose to pay higher dividends to investors. Risk-averse investors can choose to buy shares that pay out dividends consistently to ensure that they get some money back from investing in a company.

12. Market and industry segmentation. There are various ways each company can be classified into a segment that better describes its business characteristics base on various standards. These various ways of carving out companies' segments and sub-segments is an important factor in the classification model. In the current study, 11 different variables for classifying companies into industry and subindustry are chosen as qualitative independent variables; these include the following:

- a. Bloomberg Industry Classification System (BICS) Level 1 Sector
- b. BICS Level 2 Industry Group

- c. Global Industry Classification Standard (GICS) Sector
- d. GICS Industry Group
- e. GICS Subindustry
- f. Industry Classification Benchmark (ICB) Industry
- g. ICB Subindustry
- h. Industry Sector
- i. Industry Group
- j. Industry Issuer (Subindustry)
- k. Industry Index Name

“Sector” is the broadest classification,; “industry group” is more detailed classification within each sector; and subindustry is the most in-depth classification within each industry group. *Industry Sector* and *Industry Group* variables are BICS level 1 and BICS level 2 classifications, respectively, but they incorporate the business or economic function and characteristic of a company, which make the classification slightly different from *BICS Level 1 Sector* and *BICS Level 2 Industry Group* variables. Industry Issuer is the classification from the issuer of the stock. Finally, Industry Index Name represents the industry index to which a company belongs.

3.1.2 Dependent Variable

There are two types of model: one being the model that predicts whether a stock will outperform or underperform the SET index and another determining whether the return is positive or negative, i.e. there are two dependent variables.

The first dependent variable measures the stock performance relative to the entire SET market. In this case, the *SET Return* can be perceived as an opportunity cost for holding any stock in the SET market, and the goal is to generate return above the opportunity cost. Thus, the percentage return of SET and a sample stock are compared to tabulate the binary dependent variable. The SET return and the sample stock return are calculated as follows:

$$\text{SET Return} = \frac{\text{SET Index}_t - \text{SET Index}_{t-1}}{\text{SET Index}_{t-1}} \quad (3)$$

and

$$\text{Stock A Return} = \frac{\text{Price of Stock A}_t - \text{Price of Stock A}_{t-1}}{\text{Price of Stock A}_{t-1}}, \quad (4)$$

where A represents any company, and t represents year. These two are compared to determine whether the stock of company A outperformed or underperformed the SET market.

$$\begin{aligned} y_1 &= \text{"Outperform"} \text{ if Stock A Return} > \text{SET Return} \\ &= \text{"Underperform"} \text{ Otherwise} \end{aligned} \quad (5)$$

Another dependent variable measures only the movement of the stock price. The model to predict this second dependent variable is built to ensure investors' return is positive even if the SET market is falling and the stock outperforms the market.

$$\begin{aligned} y_2 &= \text{"Positive"} \text{ if Stock A Return} > 0 \\ &= \text{"Negative"} \text{ Otherwise} \end{aligned} \quad (6)$$

3.2 Classification Models

To build a model classifying whether a stock will outperform or underperform the SET index, various classification methodologies with categorical dependent variables were used in the current study.

3.2.1 CART Decision Tree

The classification and regression tree (CART) is one of many types of the decision tree used in classification. Unlike the Chi-square Automatic Interaction Detector (CHAID) decision tree that uses iterative Pearson's Chi-squared test for independence to grow the decision tree [18], CART uses entropy or the Gini index as a measurement for reducing uncertainty and growing the decision tree.

The Gini index is a measurement of statistical dispersion developed by statistician Corrado Gini in 1912. In the macroeconomic field, the Gini index is a measurement for economic inequality whose range is from 0 to 1. A Gini index of 0 represents perfect equality, and that of 1 perfect inequality. In CART, a low Gini index means better classification as the proportion of frequencies becomes more asymmetrical.

Given a categorized training sample, the Gini index can be calculated as follows:

$$Gini(y = i) = 1 - \sum_{i=1}^c p_i^2, \quad (7)$$

where y is a dependent qualitative variable, p_i is the proportion of data categorized as type i , and c is the total number of classification types.

Information entropy is defined as the summation of negative logarithm probability mass functions. It is a measurement of uncertainty introduced by Claude Shannon in 1948 with a unit called the bit. Higher-uncertainty events give higher information entropy measurements because more information can still be obtained through more random sampling. By contrast, an event with no uncertainty contains 0 bits of information entropy since no new information can be acquired through more sampling. For CART, lower entropy means purer classification as proportion of frequency is mostly skewed to one class. The formula for computing entropy is as follows:

$$Entropy(y = i) = -\sum_{i=1}^c p_i \log p_i . \quad (8)$$

An example of the Gini index and entropy calculation for data in Table 1 is provided below the table.

Table 1: Example for Gini index and Entropy Calculation

Classification Type	Frequency	Proportion
Outperformed	80	40%
Underperformed	120	60%
Total	200	100%

The Gini index and entropy can be computed as follows:

$$\begin{aligned} Gini(y = i) &= 1 - p_{\text{outperform}}^2 - p_{\text{underperform}}^2 \\ &= 1 - 0.4^2 - 0.6^2 = 0.48 \end{aligned}$$

and

$$\begin{aligned} Entropy(y = i) &= -p_{\text{outperform}} \log p_{\text{outperform}} - p_{\text{underperform}} \log p_{\text{underperform}} \\ &= -0.4 \log (0.4) - 0.6 \log (0.6) = 0.2923. \end{aligned}$$

However, these calculations are based on only the labeled classification (dependent variable). To incorporate the effect of the independent variable, detailed equations are needed to compute entropy and the Gini index with the independent variable's effect. Let x be independent variable and y be a binary classification label. The Gini index and entropy after the split can be computed as the expected value from each node:

$$Gini(x, y) = P_{x < m} Gini(y = i, x < m) + P_{x > m} Gini(y = i, x > m) \quad (9)$$

and

$$Entropy(x, y) = p_{x < m} Entropy(y = i, x < m) + p_{x > m} Entropy(y = i, x > m) \quad (10)$$

where $Gini(y = i, x < m)$ and $Entropy(y = i, x < m)$ are the Gini index and entropy, respectively, given that $x < m$, $P_{x < m}$ is the proportion of data in the $x < m$ class, and m is the cutoff value for the best split point. To compute these parameters, all data instances must be discretized by frequency using x as a binary class.

An example is presented in Table 2 below, where x is price-to-earnings (PE) ratio of a stock, and y is a stock's performance relative to the market.

Table 2: Example for Gini and Entropy Calculation with Independent Variable

Frequency Table x	y		Total
	Outperformed	Underperformed	
PE > 10	50	60	110
PE < 10	30	60	90
Total	80	120	200

The Gini index for each node using the two PE classes as a split can be computed as

$$\begin{aligned}
 Gini(y = i) &= 1 - p_{\text{outperform}}^2 - p_{\text{underperform}}^2 \\
 &= 1 - (80/200)^2 - (120/200)^2 = 0.48 \\
 Gini(y = i, PE > 10) &= 1 - p_{\text{outperform, PE > 10}}^2 - p_{\text{underperform, PE > 10}}^2 \\
 &= 1 - (50/110)^2 - (60/110)^2 = 0.4959 \\
 Gini(y = i, PE < 10) &= 1 - p_{\text{outperform, PE < 10}}^2 - p_{\text{underperform, PE < 10}}^2 \\
 &= 1 - (30/90)^2 - (60/90)^2 = 0.4444.
 \end{aligned}$$

Based on equation (9), the Gini index after the split is

$$\begin{aligned}
 Gini(x, y) &= p_{PE > 10} Gini(y = i, PE > 10) + p_{PE < 10} Gini(y = i, PE < 10) \\
 &= (110/200)0.4959 + (90/200)0.4444 = 0.4727.
 \end{aligned}$$

From equation (10), the entropy of the split can be calculated as the expected value of the entropy in the node after the split. The first calculation is for the entropy of each node:

$$\begin{aligned} Entropy(y = i, PE > 10) &= -p_{out, PE > 10} \log(p_{out, PE > 10}) - p_{under, PE > 10} \log(p_{under, PE > 10}) \\ &= -(50/110)\log(50/110) - (60/110)\log(60/110) = 0.2992. \end{aligned}$$

and

$$\begin{aligned} Entropy(y = i, PE < 10) &= -p_{out, PE < 10} \log(p_{out, PE < 10}) - p_{under, PE < 10} \log(p_{under, PE < 10}) \\ &= -(30/90)\log(30/90) - (60/90)\log(60/90) = 0.2764. \end{aligned}$$

Therefore, the expected value of the split can be computed by applying equation (10):

$$\begin{aligned} Entropy(PE, y) &= p_{PE < 10} Entropy(y = i, PE < 10) + p_{PE > 10} Entropy(y = i, PE > 10) \\ &= (90/200) \times 0.2764 + (110/200) \times 0.2992 = 0.2890. \end{aligned}$$

The entropy of 0.2890 is information after the split. As mentioned earlier, the criteria for creating the split are either information gain, the gain ratio or Gini gain. These numbers can be specified as a minimum gain to conduct the split and a termination criterion for growing a decision tree.

In the case of the above example, the information and Gini gain before and after a split can be computed as follows:

$$\begin{aligned} \text{Information Gain} &= Entropy(y = i) - Entropy(PE, y) \\ &= 0.2923 - 0.2890 = 0.0033 \end{aligned} \tag{11}$$

and

$$\begin{aligned} \text{Gini Gain} &= Gini(y = i) - Gini(x, y) \\ &= 0.48 - 0.4727 = 0.0073. \end{aligned} \tag{12}$$

It is important to note that the Gini index and entropy before the split (of a parent node) is always higher than those after the split (child nodes). This is because as more independent variables are incorporated into the splits, the classification becomes purer.

Information gain has a bias for independent variables with many distinct features. For example, if there is a data ID attribute that has a distinct value for all data points, the information gain before and after the split will be maximal or equal to the information

contained in the parent node before the split. The gain ratio is used instead to account for this by normalizing the information gain with intrinsic information. This can be computed as follows:

$$\begin{aligned} \text{Intrinsic Information} &= -p_{x<m} \log(p_{x<m}) - p_{x>m} \log(p_{x>m}) \quad (13) \\ &= -(90/200)\log(90/200) - (110/200)\log(110/200) = 0.2989. \end{aligned}$$

Based on this information, the gain ratio is

$$\begin{aligned} \text{Gain Ratio} &= \frac{\text{Information Gain}}{\text{Intrinsic Information}} \quad (14) \\ &= 0.0033/0.2989 = 0.0110. \end{aligned}$$

In the previous example of calculation of information and Gini bases on a continuous independent variable PE, it is important to note that a calculation for the best binary split point m is important in building a decision tree for a continuous variable with many distinct values. The best binary split point in the previous example is assumed to be at $m = 10$ ($PE = 10$). One of the common approaches in finding the best split point is to use the middle point of two distinct values [19]. Assume there are eight data points sorted in ascending order; the possible cutoff m can be computed for all two adjacent distinct values.

Table 3: Example for finding the best split point m

Data point	PE	m	y
1	5.7		Outperform
2	8	6.85	Outperform
3	12	10	Underperform
4	15	13.5	Underperform
5	16.5	15.75	Outperform
6	18	17.25	Outperform
7	18.25	18.125	Outperform
8	20	19.125	Underperform

Then, for all m in Table 3, the Gini index and information gain are computed iteratively to find the highest gain, which represents the best split.

3.2.2 Logistic Regression (LR)

Regression is one of many tools used for analyzing a relationship between dependent and independent variables. In regression analysis, independent variables are assumed to have a linear relationship with dependent variables. However, when a dependent variable is binary, non-linear regression such as logistic regression is often used to find a relationship between, either categorical or continuous, independent variables with a discrete binary dependent variable (outperformed or underperformed). Logistic regression is a parametric analysis using “logit” or natural log of an odd ratio. Despite being a parametric method, logistic regression is not restricted by normality or equal variance (homoscedasticity) assumptions. The model, however, still requires

1. independent variables that have little to no multicollinearity
- and
2. assumed linearity between independent variables and a log odd ratio.

The logistic regression equation takes the form

$$z_i = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k \quad (15)$$

where p_i is the probability that dichotomous events that occur describe in dependent variable, $\beta(s)$ are coefficients calculated using the maximum likelihood function, k is the number of independent variables, and Z_i is called the odd ratio.

From equation (15), the probability that a binary event occurs given a set of independent variables $x(s)$ is given by

$$P(Y = y / x_1, x_2, \dots, x_k) = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k}} \quad (16)$$

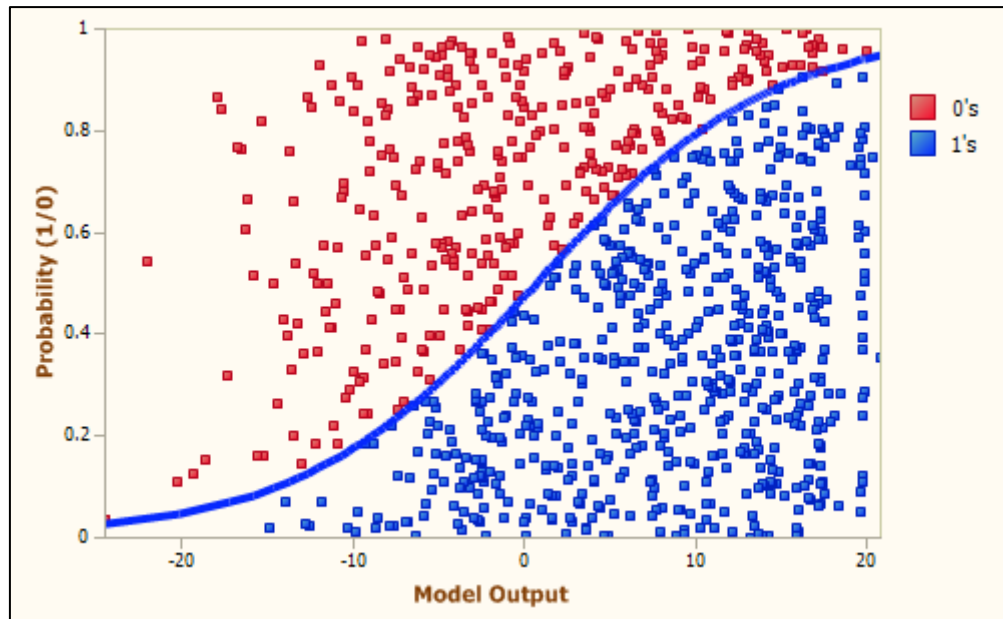


Figure 8: Graph of Logistic Regression with binary outcomes (π_i vs. Z_i)

Since logistic regression has a binary dependence output, the likelihood function for deriving $\beta(s)$ takes the form of the Bernoulli probability mass function:

$$L(\beta) = \prod_{i=1}^n p(x_i)^{y_i} (1 - p(x_i))^{1 - y_i} \quad (17)$$

where n is the number of data points, and y_i is 0 or 1 for binary outcomes.

After taking log and substituting equations (15) and (16) into equation (17), the likelihood function for logistic regression becomes

$$L(\beta) = \sum_{i=1}^n -\log\left(1 + e^{\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k}\right) + \sum_{i=1}^n y_i (\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k) \quad (18)$$

$\beta(s)$ can be solved by taking the derivative of equation (18) with respect to $\beta(s)$ and setting it equal to zero or maximizing the likelihood function.

Compared to the normal linear regression, the benefit of using the logistic regression is that the technique does not require normality or homoscedasticity assumptions for independent variables that make it easier to use [11]. Other than building a forecasting model, logistic regression can also be used for independent variable selection through the p-value and significance level. This was done by Cheng and Wang, who are mentioned in the Related Work section [15].

3.2.3 Linear Discriminant Analysis

Linear discriminant analysis (LDA) is a parametric classification method developed by Ronald Fischer in which dependent variables are categorical and independent variables are continuous. The basic concept of LDA is to obtain a linear regression equation that best separates the two classifications.

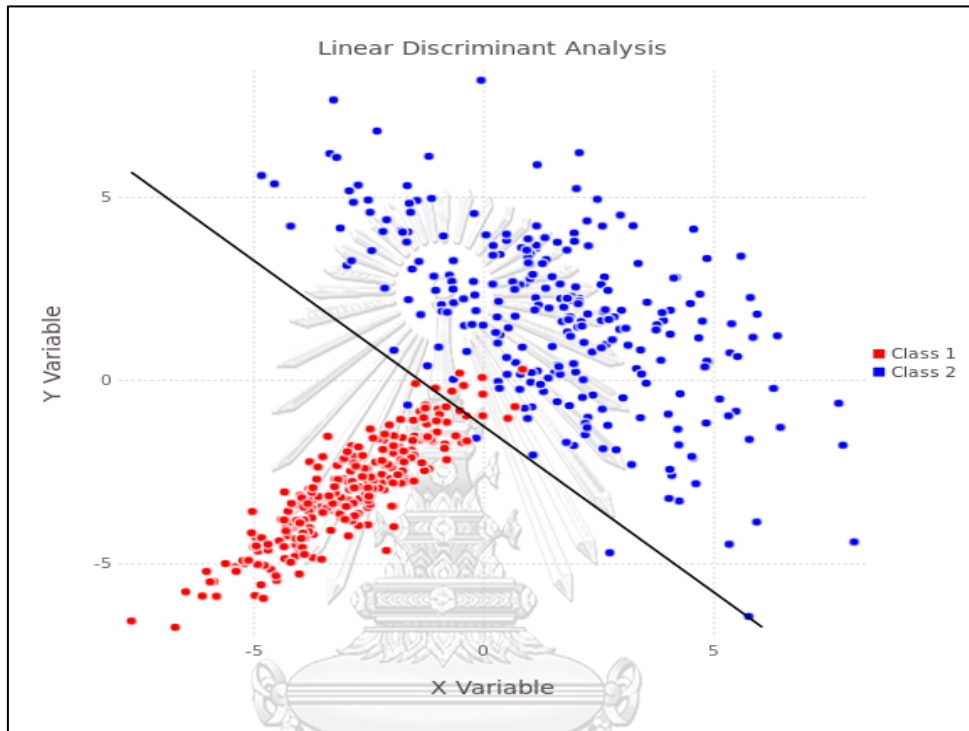


Figure 9: Graph of LDA with binary outcomes (x_i vs. Z_i)

The regression equation (Z-score) takes the form

$$Z = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k \quad (19)$$

The coefficients β (s) are calculated from a linear combination of independent variables that best separate the two classes. To capture such a separation, Fischer defined the following score function:

$$s(\beta) = \frac{\beta^T \mu_1 - \beta^T \mu_2}{\beta^T C \beta} \quad (20)$$

where β is the vector of coefficients, μ_i is mean vector for independent variables classified into dependent variable categories, and C is covariance matrix. After solving equation (20), the vector of coefficients (β), the solution for maximizing the score function is obtained as follows:

$$\beta = C^{-1}(\mu_1 - \mu_2) \quad (21)$$

and

$$C = \frac{1}{n_1 + n_2}(n_1 C_1 + n_2 C_2) \quad (22)$$

where C is the pooled covariance matrix, C_i are covariance matrix vectors for independent variables classified into dependent variable categories, n_i is the sample size for each of the dependent variable categories. Like logistic regression, LDA can produce an estimate probability of binary events occurring through a set of independent variables with the following equation for calculating the probability of a negative outcome:

$$p(Y = 2 / x_1, x_2, \dots, x_k) = \frac{1}{1 + \frac{1 - \pi_2}{\pi_2} e^{z - 0.5\beta(\mu_1 - \mu_2)}} \quad (23)$$

From (19), an instance can be classified into two categories (based on dependent variables) if the Z-score is either more or less than the cutoff score. The drawbacks of using this methodology is that, unlike the logistic regression, the model requires restrictive normality and homoscedasticity assumptions. However, the violation of the normality assumption is not typically fatal, and the significance test is still trustworthy [20].

3.2.4 K-Nearest Neighbor

K-Nearest Neighbor (KNN) is a non-parametric statistical approach classifier. For given testing data, KNN classifies the test data using the distance function. After the distances are calculated between the testing data and all training data, the ranks are given in ascending order of the distance. Finally, the K nearest distances (K closest distance or K lowest rank) are chosen, and the mentioned testing data point are classified based on the majority results of the K nearest distances of training data points. This is illustrated in Figure 11.

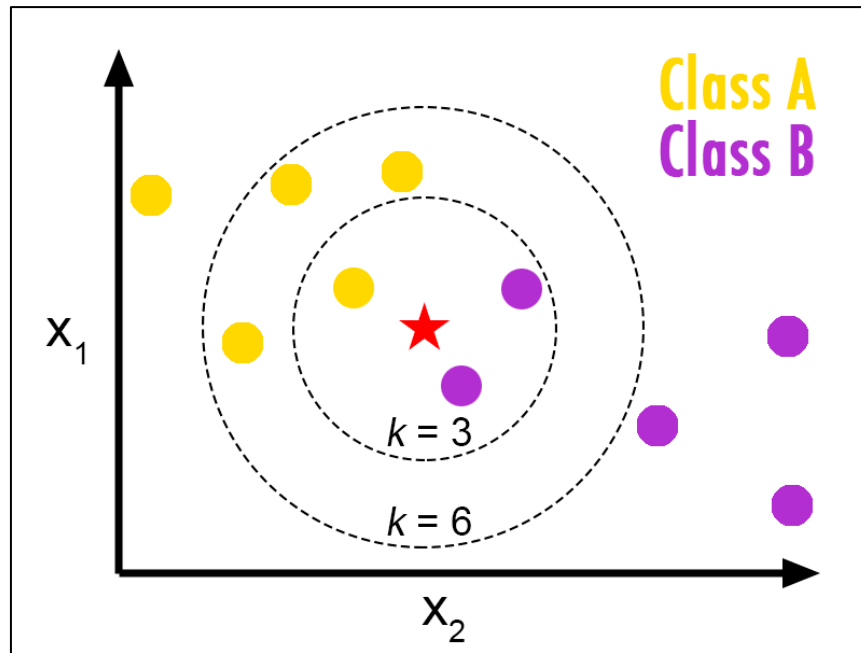


Figure 10: KNN Example

In Figure 10, the new testing data, denoted by the star, are classified into class B since the majority of nearest neighbors (distance) are class B training data, considering $K = 3$ nearest neighbors. By contrast, if $K = 6$, the testing data are classified into class A since the majority of nearest neighbors are in class A. From this observation, selecting the right K value is crucial in building an accurate model. Generally, the K value is selected as an odd number to avoid a tie in both classes.

Various distance functions can be used to calculate distances and determine the nearest neighbors. The three basic distance functions are Minkowski, Manhattan and Euclidean and represented by equations (24), (25) and (26), respectively:

$$D_{\text{Minkowski}} = \sqrt[q]{\sum_{i=1}^k |x_{i,\text{training}} - x_{i,\text{testing}}|^q} \quad (24)$$

$$D_{\text{Manhattan}} = \sum_{i=1}^k |x_{i,\text{training}} - x_{i,\text{testing}}| \quad (25)$$

and

$$D_{\text{Euclidean}} = \sqrt{\sum_{i=1}^k |x_{i,\text{training}} - x_{i,\text{testing}}|^2} \quad (26)$$

where $x_{i,\text{training}}$ is the independent variable from training data, $x_{i,\text{testing}}$ is the independent variable from testing data, and k is the number of independent variables. The major drawback of using the distance function is that independent variables are very likely to have different measurement scales. To nullify the effect of different measurement

scales, the numerical value of these variables needs to be normalized using any of the range transformations (interquartile transformation and Z-transformation) with equation (27), (28)) or (29), respectively:

$$x_{i,\text{normalized}} = \frac{x_i - x_{i,\text{min}}}{x_{i,\text{max}} - x_{i,\text{min}}} \quad (27)$$

where $x_{i,\text{min}}$ and $x_{i,\text{max}}$ are minimum and maximum values of x_i from training data;

$$x_{i,\text{normalized}} = \frac{x_i - \text{median}}{IQR} \quad (28)$$

where IQR is the interquartile range and median is the 50th quartile; and

$$x_{i,\text{normalized}} = \frac{x_i - \text{mean}}{\sigma} \quad (29)$$

where σ is the standard deviation of that independent variable data.

An example calculation based on training data is presented in Table 4 with PE and price-to-book ratio (PB) as independent variables.

Table 4: KNN Example based on Training Data

Data point	PE	PB	y
1	5	0.9	Outperform
2	8	1.2	Outperform
3	12	3	Underperform
4	15	0.8	Underperform
5	16.5	1.0	Outperform
6	18	0.85	Outperform
7	18.25	4.2	Outperform
8	20	2.5	Underperform

Applying equation (27), PE and PB are normalized; then, the Manhattan distance is calculated for testing data with PE = 10 and PB = 1.36. The normalized PE and PB, from equation (27), for the testing data are $PE_{\text{normalize}} = (10 - 5) / (20 - 5) = 0.33$ and $PB_{\text{normalize}} = (1.36 - 0.8) / (4.2 - 0.8) = 0.1647$. After calculating the Manhattan distance, ranks are assigned in ascending order of distance.

Table 5: KNN Example

Data point	PE	PB	y	Manhattan Distance	Rank
1	0.000	0.029	Outperformed	$Abs(0-0.33) + Abs(0.029-0.1647) = 0.47$	2
2	0.200	0.118	Outperformed	$Abs(0.2-0.33) + Abs(0.118-0.1647) = 0.18$	1
3	0.467	0.647	Underperformed	0.62	5
4	0.667	0.000	Underperformed	0.50	3
5	0.767	0.059	Outperformed	0.54	4
6	0.867	0.015	Outperformed	0.68	6
7	0.883	1.000	Outperformed	1.39	8
8	1.000	0.500	Underperformed	1.00	7

For KNN with $K = 3$, the data points with ranks 1 to 3 are considered. In Table 5, data points with ranks 1 to 3 have two “Outperform” and 1 “Underperform” classifications; therefore, the testing data point is classified as “Outperform” by majority results. If $K = 5$, the data points with ranks 1 to 5 are considered, which consist of three “Outperform” and two “Underperform”; therefore, the testing instance is classified as “Outperform”.

3.3 Performance Measurement

The performance of a classification model can be measured by counting the number of correct and incorrect classifications. This can be done by using the confusion matrix presented in Table 6, which accounts for the predicted class vs. the actual class.

Table 6: Confusion Matrix

Confusion Matrix		Actual Outcome	
		Outperformed	Underperformed
Predicted Outcome	Outperformed	True Positive (TP)	False Positive (FP)
	Underperformed	False Negative (FN)	True Negative (TN)

From the confusion matrix, the accuracy of a prediction model can be computed as

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (30)$$

Accuracy measurement considers how many instances are correctly predicted by a model. However, this measurement has a drawback: accuracy measurement does not consider the cost and benefit of correctly predicting an outperforming or underperforming class. For example, if the model predicted all testing instances to be

100% “Underperform”, then TP and FP would be 0 but the accuracy can still be high if TN is high. In this case, the measurement is not very useful to investors, since they are interested in choosing outperforming stocks.

Sensitivity calculation considers the benefit of correct prediction. Sensitivity can be computed as

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (31)$$

The ratio only considers how many instances have been predicted as Outperform out of all the actual Outperform outcomes. This is also called the true positive rate (TPR).

Another ratio that accounts for the cost of investing is called the Fall-Out rate.

$$\text{Fall-Out} = \frac{FP}{FP + TN} \quad (32)$$

This ratio measures the risk of using the model. Since investors are likely to choose a stock that has been predicted to outperform, it is vital to know how many Outperform predictions are incorrect. This ratio is also called the false positive rate (FPR).

Another technique for measuring classification performance is the receiver operating characteristic (ROC) curve. The ROC is a graph used to visualize and choose a classification model by plotting sensitivity and fall-out rate [21].

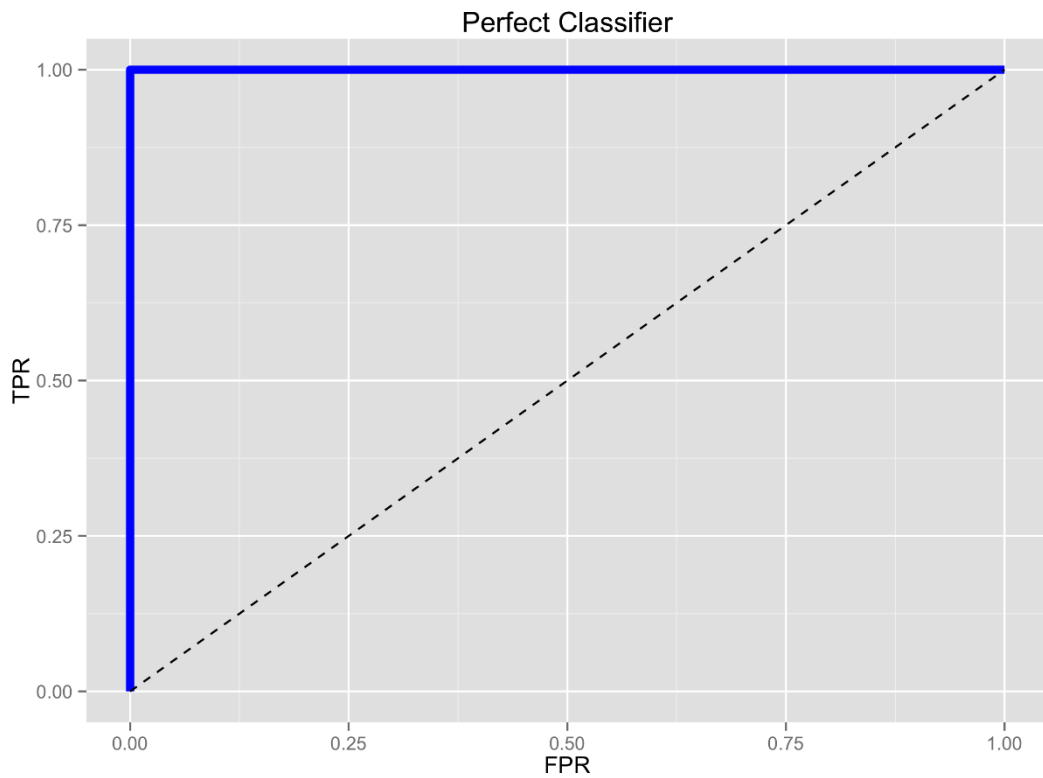


Figure 11: Perfect Classification Model's ROC Curve

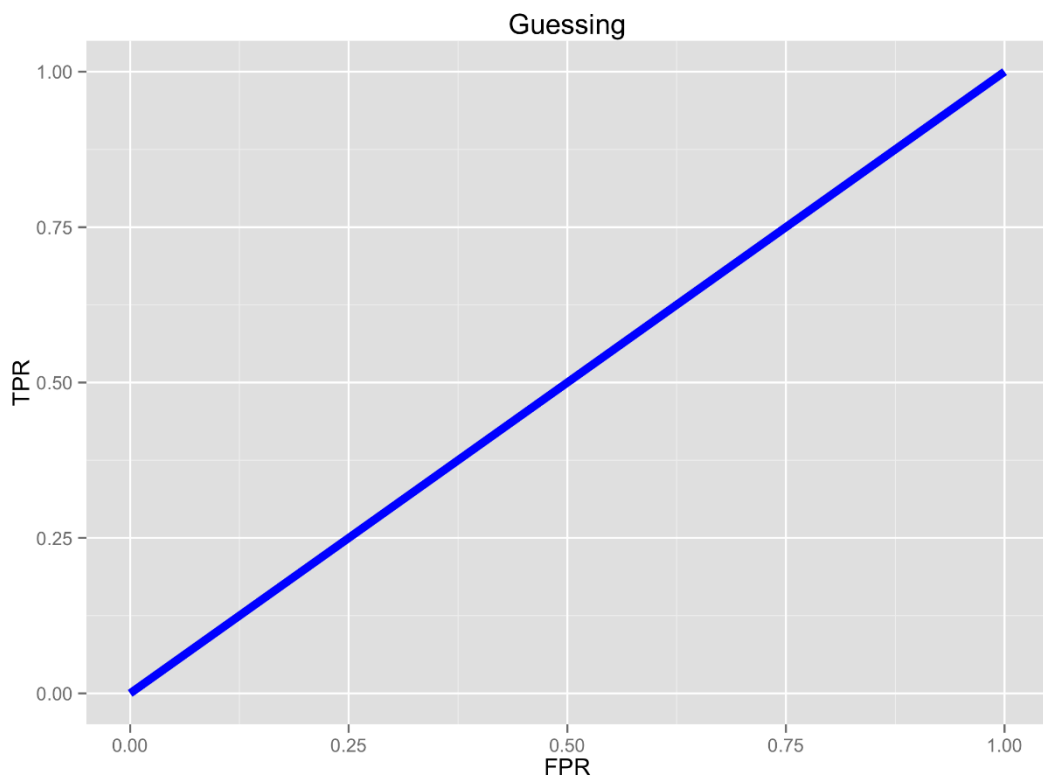


Figure 12: The ROC Curve for Random Guessing

There are many important points to note in ROC space. At point (0, 0), the classification model issues no positive (“Outperform”) prediction but also gains no FPR. Point (0, 1) represents perfect classification, in which all positive outcomes are predicted correctly and there is no FPR. By contrast, point (1, 0) represents the worst classification, in which all positive predictions are FP.

The ROC curve for a classification model is considered better when the curve bends to the left or is in the $TPR > FPR$ region. This is because if FPR is higher than TPR, then the model can be considered to be not useful in making decisions since most of its positive predictions will be inaccurate.

The ROC curve can be constructed by adjusting a threshold in a classification model. Classification models such as LR and LDA have equations (16) and (23), respectively, which translate prediction into estimated probability. The estimated probability is then compared with the threshold value to determine the classification in which a testing instance is. If the calculated probability is higher than or equal to the threshold value, the testing instance is categorized as positive or “Outperform”; otherwise, the instance is negative or “Underperform”. The confusion matrix (Table 6) can be tabulated for every threshold value for which FPR and TPR differ. ROC is plotted by sorting the calculated probability of the testing data in descending order, moving the threshold according to the calculated probability and plotting FPR and TPR for every threshold. Using 20 instances of testing data in Table 7, in which the estimated probabilities by a model are sorted in descending order, ROC is constructed as follows:



Table 7: ROC Curve Construction Example

Instance	Actual Outcome/Class	Estimated Probability/Threshold	FPR (x-axis) FP/(FP+TN)	TPR (y-axis) TP/(TP+FN)
1	Outperform	0.9	$0/(0+10) = 0$	$1/(1+9) = 0.1$
2	Outperform	0.8	$0/(0+10) = 0$	$2/(2+8) = 0.2$
3	Underperform	0.7	$1/(1+9) = 0.1$	$2/(2+8) = 0.2$
4	Outperform	0.6	$1/(1+9) = 0.1$	$3/(3+7) = 0.3$
5	Outperform	0.55	$1/(1+9) = 0.1$	$4/(4+6) = 0.4$
6	Outperform	0.54	$1/(1+9) = 0.1$	$5/(5+5) = 0.5$
7	Underperform	0.53	$2/(2+8) = 0.2$	$5/(5+5) = 0.5$
8	Underperform	0.52	$3/(3+7) = 0.3$	$5/(5+5) = 0.5$
9	Outperform	0.51	$3/(3+7) = 0.3$	$6/(6+4) = 0.6$
10	Underperform	0.505	$4/(4+6) = 0.4$	$6/(6+4) = 0.6$
11	Outperform	0.4	$4/(4+6) = 0.4$	$7/(7+3) = 0.6$
12	Underperform	0.39	$5/(5+5) = 0.5$	$7/(7+3) = 0.7$
13	Outperform	0.38	$5/(5+5) = 0.5$	$8/(8+2) = 0.8$
14	Underperform	0.37	$6/(6+4) = 0.6$	$8/(8+2) = 0.8$
15	Underperform	0.36	$7/(7+3) = 0.7$	$8/(8+2) = 0.8$
16	Underperform	0.35	$8/(8+2) = 0.8$	$8/(8+2) = 0.8$
17	Outperform	0.34	$8/(8+2) = 0.8$	$9/(9+1) = 0.9$
18	Underperform	0.33	$9/(9+1) = 0.9$	$9/(9+1) = 0.9$
19	Outperform	0.30	$9/(9+1) = 0.9$	$10/(10+0) = 1$
20	Underperform	0.1	$10/(10+0) = 1$	$10/(10+0) = 1$

According to Table 7, if the threshold is set at 0.53 (instance 7), then the model predicts that training instances 1 to 7 outperform since their probability exceeds the threshold. However, there are two data in instances 1 to 7 for which the actual class is Underperform, and the rest is correctly categorized as Outperform; therefore, FP = 2, and TP = 5. At the same threshold, instances 8 to 20 are classified by the model as Underperform, but only eight instances actually underperform, and five instances outperform; thus, TN = 8, and FN = 5. The thresholds are adjusted to the same value as estimated probability in every instance and (FPR, TPR) values are plotted as illustrated in Figure 13.

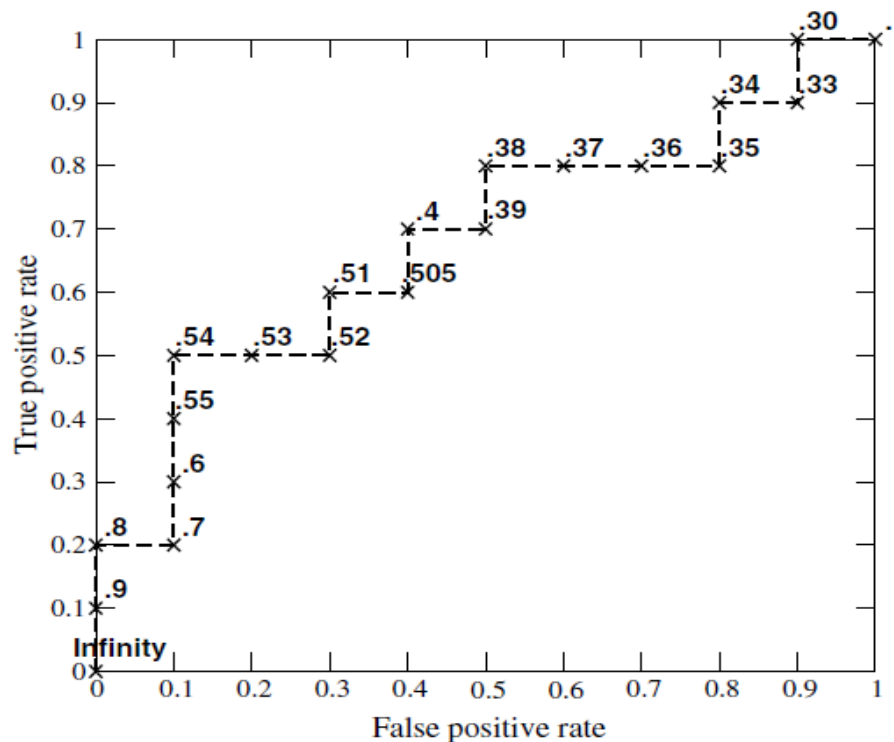


Figure 13: ROC Curve Plotted from Table 7

The ROC curve starts at a threshold of infinity, where $FPR = TPR = 0$, and ends at a threshold of negative infinity, where $FPR = TPR = 1$. The closer the ROC curve to the top-left corner, the better the classification model.

As for the CART Decision Tree and KNN, unlike LDA and LR, which have an estimated probability function, the estimated probability becomes a score or proportion range from 0 to 1 instead. The score of the decision tree is the proportion of the majority class in a leaf node. For KNN, the score is also the proportion of the majority class in K nearest neighbors.

In the ROC curve, the model's performance is measured visually by the curvature of the ROC curve or numerically by the Area Under the Curve (AUC). Visually, the model is considered more useful when the curve leans more toward point (1, 1). Numerically, AUC, which ranges from 0 to 1, is calculated where 1 represented the perfect classification. The model with a higher AUC value is considered more useful, as on average, it has the smallest number of false positive instances and the largest number of true positive instances. The AUC represent how well the model can separate the two class.

Since ROC curve is plotted by probability scores and adjusted by classification thresholds, the AUC represents the probability of random positive instance being ranked higher than a random negative instance. To explain in terms of investing, given a pair of positive and negative stock, AUC is the probability that the model selects positive stock with higher confidence.

3.4 Data Preprocessing

The data of 582 companies in SET were pulled from the Bloomberg terminal, which is a common platform of financial and securities data utilized by investors. The data consist of three dimensions, including company name, financial ratios and year. In the current study, the relationship between a company's end-of-year financial performance and 1-year capital gain was observed (see Table 8, Table 9 and Table 10).

The financial ratios are taken at the end of a year, whereas the stock prices for computing a 1-year return are taken at the end of quarter 1 of next year because there is a time gap before the public release of end-of-year financial performance. All the independent variables (financial ratios and industry/subindustry) are arranged in the observation period from year 1 to year N regardless of their actual year, and the prices are arranged in performance period from year N+1 to year N+2 for calculating the 1-year return of a stock, where $N = 3, 4, \text{ and } 5$. The return is translated into categorical dependent variables to build a model describing the relationship of financial ratios for year 1 to year N and the dependent variable. Observation periods of 3, 4 and 5 years are analyzed to find the best observation period for the training model. This is performed by measuring the AUC of LR, and the number of observation periods with the highest AUC are selected for further analysis.

Table 8: 3-year Observation Period ($N = 3$)

End of Year																	
2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	
Observation			Performance														
	Observation		Performance														
		Observation		Performance													
			Observation		Performance												
				Observation		Performance											
					Observation		Performance										
						Observation		Performance									
							Observation		Performance								
								Observation		Performance							
									Observation		Performance						
										Observation		Performance					
											Observation		Performance				
												Observation		Performance			
													Observation		Performance		

Table 9: 4-year Observation Period ($N = 4$)

End of Year																	
2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	
Observation				Performance													
	Observation			Performance													
		Observation		Performance													
			Observation		Performance												
				Observation			Performance										
					Observation				Performance								
						Observation				Performance							
							Observation				Performance						
								Observation				Performance					
									Observation				Performance				
										Observation				Performance			
											Observation				Performance		
												Observation				Performance	

Table 10: 5-year Observation Period ($N = 5$)

End of Year																	
2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	
Observation					Performance												
	Observation				Performance												
		Observation			Performance												
			Observation		Performance												
				Observation			Performance										
					Observation				Performance								
						Observation				Performance							
							Observation				Performance						
								Observation				Performance					
									Observation				Performance				
										Observation				Performance			
											Observation				Performance		

CHULALONGKORN UNIVERSITY

Every data instance consists of an observation period that contains columns of independent variables $X_1, X_2, \dots, X_{11N+11}$ (where $11N$ is 11 financial ratios for N years, and the other 11 is for industry/subindustry classification variables) and a performance period that contains a stock's price and the SET index for calculating dependent variables Y_1 and Y_2 . It is important to notice from Table 8, Table 9, and Table 10 that a longer observation period means less data instances. For example, for the 3-year Observation period, there are 13 instances (13 rows in Table 8) for a company stock, each with an instance consisting of 11 financial ratios over 3 years (observation period) plus 11 industry segmentation variables or 44 independent variables $X_1, \dots, \text{and } X_{44}$ and 2 dependent variables from 1-year performance periods Y_1 and Y_2 .

By contrast, for the 5-year observation period, there are 11 instances (11 rows in Table 10) for a stock, each with an instance consisting of 11 financial ratios over 5 years (observation period) plus 11 industry segmentation variables or independent variables $X_1, \dots, \text{and } X_{66}$ and 2 dependent variables from 1-year performance periods Y_1 and Y_2 .

After arranging the data and eliminating instances results in a null value: there are a total of 3,030, 2,524 and 2,102 instances for 3-year, 4-year and 5-year observation periods, respectively.

We use observation period of 3-year because investors typically need to see the consistency of company's growth and profitability over the long term. While there is no rule on the minimum number of observation period, 1-year period is too short as investors may not see any trend or unable to evaluate consistency of company's growth. While 2-year period allows investors to see linear trend of financial ratios, such a trend can be misleading as non-linear trends could not be captured by such a setting. Therefore, 3-year period seems to be the minimal observation period.

The elimination of outliers is performed using KNN outlier detection. This process entails calculating the distance between any instance and its k-th nearest neighbor (see KNN in the Classification Model section) and then declaring the top 30 instances with the highest distance as outliers for the 5-year observation period's data set. For the 5-year observation period's data set, the top 30 instances are chosen because, as mentioned earlier, after eliminating instances with a null value, the sample size becomes 2,102 instances. According to Hosmer and Lemeshow [22], the sample size for performing logistic regression is 10 times the number of independent variables. However, LeBlanc and Fitzgerald [23] suggested that 30 times the number of independent variable should be used. Since the ration of training to testing is 70:30 and there are 55 independent variables, the safest sample size required for model training is $55 \times 30 = 1,650$ instances and training account for 80% of all instances; therefore, $1650 \div 0.8 = 2,063$ instances are required. The filter data contain 2,102 instances; thus, the maximum number of outliers that can be eliminated is $2,102 - 2,063 = 39$ instances. These 30 instances account for 1.43% of 2,102 instances, and this percentage is used for eliminating outliers in the 3-year and 4-year observation period's data sets; thus, 44 and 36 outliers are screened out from 3,030 instances in the 3-year observation period's data and 2,524 instances in the 4-year observation period's data, respectively.

The elimination of outliers is essential because outliers later affect data normalization and parametric analysis. For detecting outliers with many continuous variables, the popular distance function is Euclidean distance [24].

After identifying outliers from the three datasets (3-year, 4-year and 5-year), an observation year analysis is performed using logistic regression and comparing AUC from the ROC curve to find the best number of observation years to be used in training the model.

Logistic regression is used for this analysis because it is the simplest model that can be applied without highly complicated settings such as termination criterions for the decision tree, the number of neighbors to consider in KNN. More importantly, LR can be used without the risk of violating normality and homoscedasticity assumptions required in LDA. The ratio of training to testing data in the analysis was 70:30. The ROC curves of these three data sets are illustrated in Figures 15–17.

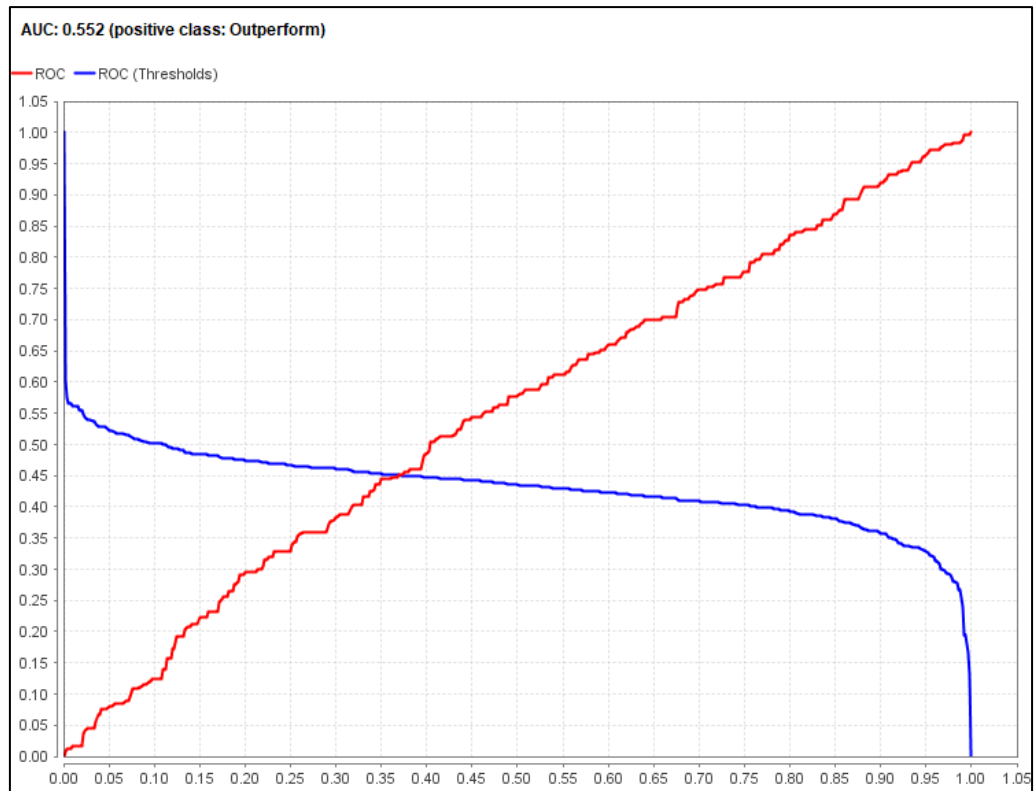


Figure 14: ROC curve of the 3-year dataset (AUC = 0.552)

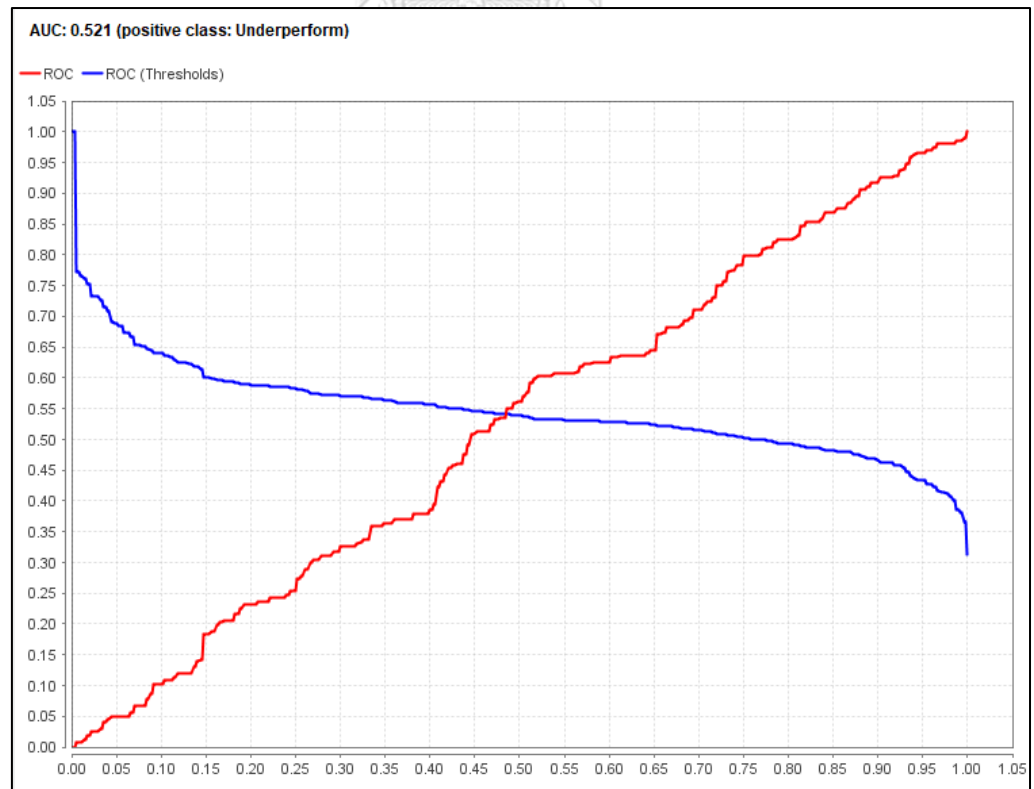


Figure 15: ROC curve of the 4-year dataset from LR (AUC = 0.521)

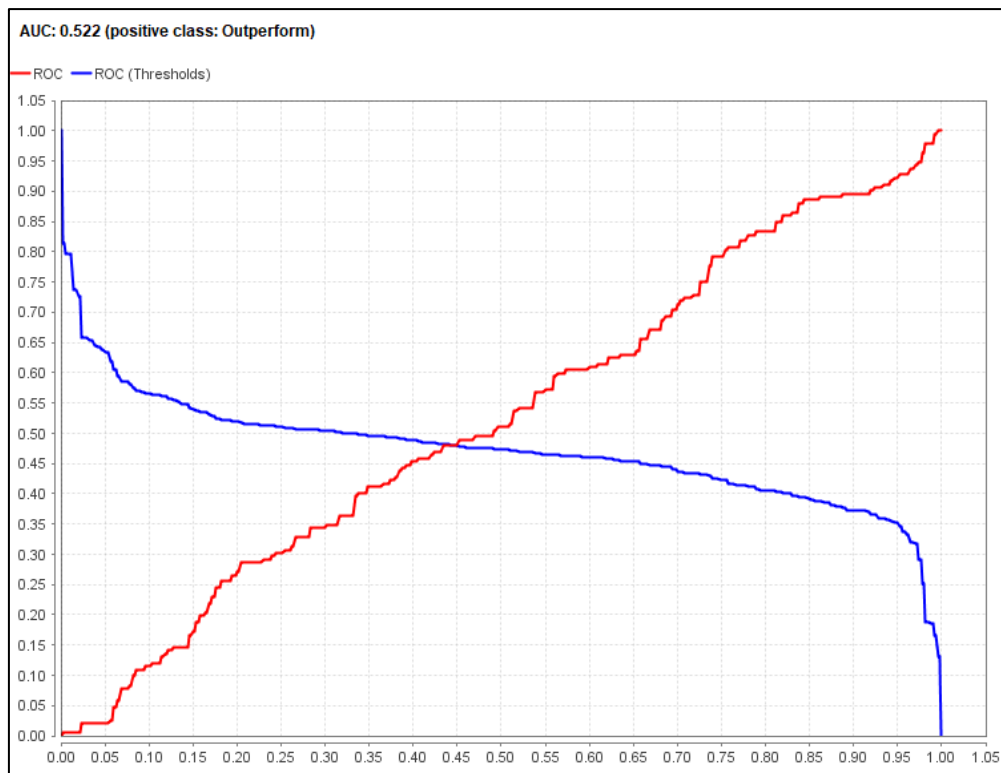


Figure 16: ROC curve of the 5-year dataset from LR (AUC = 0.522)

From the analysis, the 3-year data set has the highest AUC (0.552). As mentioned earlier, the 3-year data set has the highest number of instances (3,030) and the least number of independent variables (11 financial ratio variables \times 3 years = 33 variables) from the three data sets. These results demonstrate that model building becomes slightly more robust with more instances vs. more independent variables, especially in the 5-year data set, which has only 2,102 instances. Only 1,682 instances (80% of 2102) are used for training; this number is slightly higher than that recommended by LeBlanc and Fitzgerald [23], which is 30 times the number of independent variables. With a total of 55 independent variables (11 financial ratio \times 5 years = 55 variables) in the 5-year dataset, the number of instances recommended by LeBlanc and Fitzgerald [23] would be $30 \times 55 = 1,650$.

With the highest AUC of 0.552, the 3-year data set will be used for training the classification model. The advantage of using the 3-year data set is that the data have a smaller number of independent variables; this can make a model like the classification tree much simpler compared to using 44 or 55 variables. Moreover, the data have a larger number of instances, thus making the models more accurate.

The detection of outliers with K-nearest neighbors as well as classification model training and testing was performed in data mining software called Rapidminer.

3.5 Model Training Process

In the 3-year data set, there were a total of 3,030 instances, in which 500 are eliminated as outliers through multivariate outlier detection or the KNN method. During data preprocessing, only 44 outliers were eliminated to ensure the same proportion as that of the 5-year data set, which set the limit of how many outliers can be eliminated. However, using the 3-year data set allows for more outliers to be eliminated as there is a larger number of instances. The purpose of eliminating outlier is to ensure the smoothness of data distribution and removal of extreme values. Then, the model training process consists of four steps as illustrated in the flowchart below. These four steps are performed to optimize AUC and get the most useful model out of all classifiers.

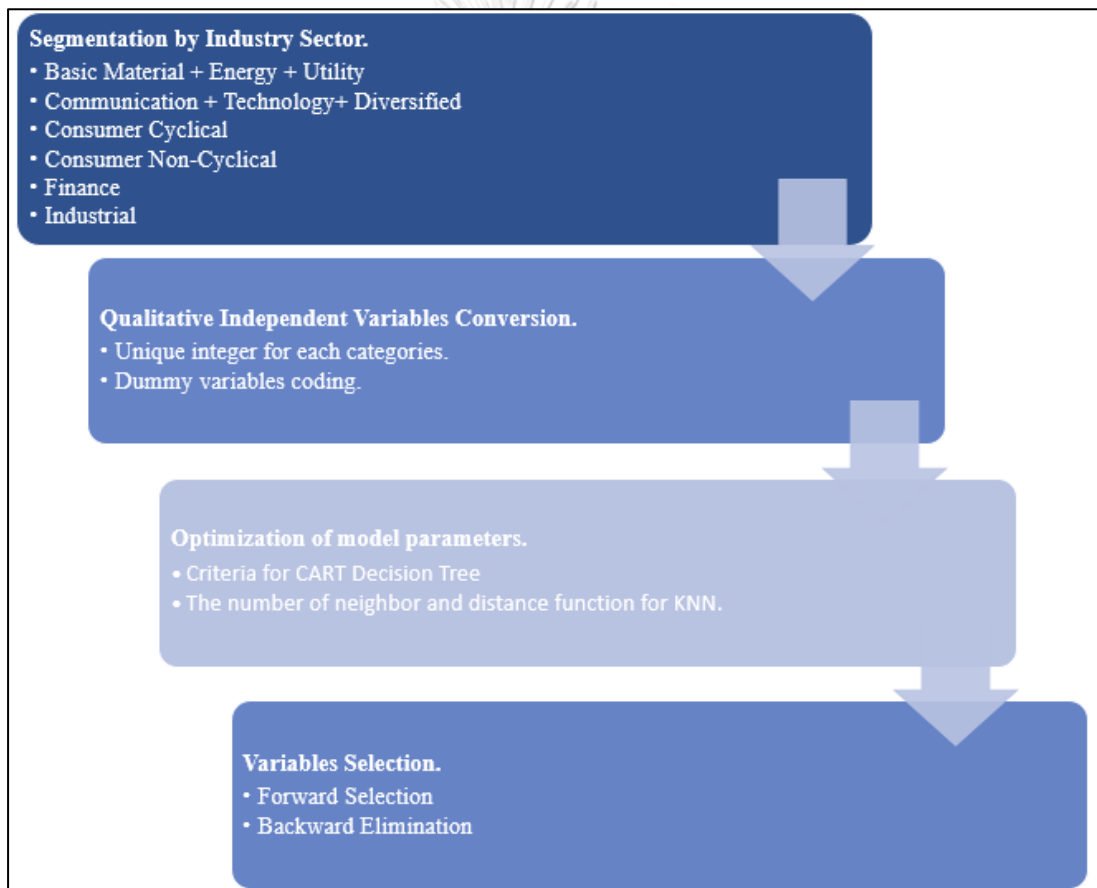


Figure 17: Model Training/Testing Flow

First, the data instances were separated into 10 industry sectors: Basic Material, Energy, Utility, Communication, Technology, Diversified, Consumer Cyclical, Consumer Non-Cyclical, Finance, and Industrial. Some of these sectors are grouped together to increase the number of instances, which are too low in some sectors. This is done primarily to improve model AUC and improve prediction as the generalized model without dividing sectors has too many characteristics that each classifier cannot completely identify and

results in lower AUC. This is proven from the AUC(s) that range around 0.50–0.55 in the data preprocessing section where data are not divided into sectors.

Second, the qualitative independent variables, i.e. the industry/market segmentation variables, are converted using either unique integer conversion or dummy variables' coding, whichever leads to a more efficient model building process. This needs to be done because some classifiers such as KNN and LDA cannot use categorical variables.

Third, the best parameter setting for the CART Decision Tree and KNN can be obtained through an optimization process in which the decision variable for DT would be termination criteria such as the maximal tree depth, minimal number of instances in a leaf node, minimal number of instances for splitting and type of a decision tree (information gain, the gain ratio, or the Gini index). For the KNN model, the optimization determines the best K (number of neighbors), the distance function (Euclidean or Manhattan), and normalization method.

Fourth, the variable selection is another optimization in which independent variables are chosen to best optimize the model's AUC performance. This is done by forward selection or backward elimination, whichever gives higher AUC.

The third and fourth steps are both optimization and were taken primarily to get the best AUC results.

3.5.1 Segmentation by Industry Sector

As mentioned earlier, the data set is segmented by industry sector, which includes 10 different sectors. Some of the industry sectors are compounded to make up for a small number of data instances for training and testing. Figure 19 illustrates the Pareto chart of all industry segments in descending order.

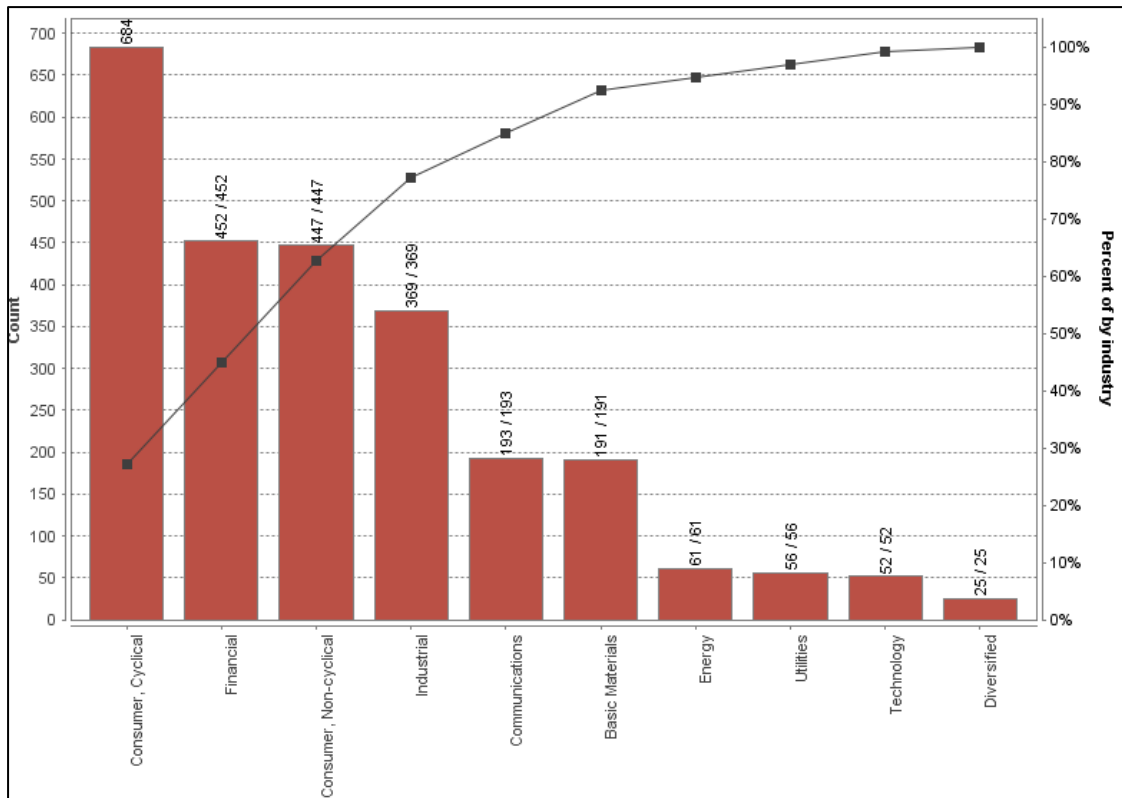


Figure 18: Frequency of data by Industry Sector

From the chart, more than 70% of all data instances are Consumer Cyclical, Finance, Consumer Non-Cyclical and Industrial sectors. Each of the sectors has less than 200 data instances, with Energy, Utilities, Technology and Diversified having less than 70 instances. To make up for the lack, the six sectors with the fewest instances were compounded into two sectors. The communication sector were merged with Technology and Diversified (the two lowest), and Basic Material was combined with Energy and Utilities to make up at least 10% of all instances. The logic behind the grouping is from Hosmer and Lemeshow [22], who suggested that the number of instances should be at least 10 times the number of independent variables. After optimization through the variable selection process, the model should be able to contain around 25–30 variables, which lead to approximately 300 instances in the two compounded sectors. Additionally, communication and technology sectors are closely related to one another, as communication companies need to invest in technologies such as cellphones and internet frequency. Thus, there are a total of six segments for model building: Consumer Cyclical, Finance, Consumer Non-cyclical, Industrial, Communication + Technology + Diversified and Basic Material + Energy + Utilities. There are a total of four classifier methods with six segments and two types of models (stock performance compared with the SET index and price movement models); therefore, the current study had a total of $4 \times 6 \times 2 = 48$ models.

3.5.2 Qualitative Independent Variable Conversion

The 11 independent variables that identify a market/industry can be converted into numerical measurements. This is vital, especially for KNN and the LDA operator in Rapidminer, which cannot handle a qualitative variable. There are two ways in which these qualitative variables can be converted:

a) Dummy Coding

In dummy coding, for each category of a qualitative variable, a new variable is created that can only take a value of 0 or 1. For example, there are a total of 10 categories for the Industry Sector variable; then, 10 dummy variables are created, with each representing a sector. If a data instance is a company in the finance sector, then the dummy variable representing the finance sector takes the value of 1 for this data instance while the rest of the nine dummy variables take the value of 0. The advantage of dummy coding is that it increases the number of variables that provide more degrees of freedom in the regression analysis. With more degrees of freedom, the optimization process of variable selection (backward elimination and forward selection) could lead to a higher AUC with more variable combinations to pick from.

b) Unique Integer

For this conversion, each category of a qualitative variable can be seen as equally ranked; thus, each category is simply converted to a real value that represents that category. The new real values are equidistant.

3.5.3 Optimization of Model Parameters

This step only applies in the CART decision tree and KNN classifier. There are many parameters in CART decision tree that can be set in Rapidminer. The optimizer used for DT helps in choosing the minimum number of instances in the leaf node, the minimum number of instances required for splitting more nodes, the maximum tree depth that limits the size of the decision tree. These are called termination conditions for building a decision tree. The splitting criterion (information gain, the gain ratio, and the Gini index) mentioned in Section 3.2.1 is another parameter to be determined by the optimizer. For the KNN classifier, the parameters that can be optimized are the normalization method (range, interquartile or Z transformation), the number of K neighbors and the distance function. These parameters were chosen to maximize AUC.

3.5.4 Independent Variable Selection

The variable screening is another optimization in the model-building process. This is done for the purpose of maximizing AUC by removing non-explanatory independent variables from the models. There are two operators in Rapidminer: Backward Elimination and Forward Selection.

In backward elimination, the model starts with all independent variables and the operator removes the independent variables one by one as long as it increases the AUC measurement. The elimination stops when the AUC can no longer increase. The

operators can go through many trials of the backward elimination process specified to avoid getting stuck in the local optimum.

For forward selection, the model starts with no variables, and the operator keeps adding more variables one by one as long as the variables increase AUC. Like backward elimination, the process stops when AUC can no longer increase by adding variables, and the process can repeat many trials to avoid the local optimum.



Chapter IV: Results and Discussion

As mentioned earlier, there are two types of a model: one predicts whether a stock's one-year return will outperform or underperform the SET Index, and another predicts whether the stock's one-year return will be positive or negative. The performances of models are measured in AUC of the ROC curve. Since the models are segmented into six sectors, using the industry sector variable as a basis, with one sector consisting of four classifiers (KNN, LDA, DT and LR) and two types of a model, there are a total of 48 models. In all models, the ratio of training to testing is 70:30. However, according to the objective of the current study, only the best model with the highest AUC from the four classifiers was chosen for real application. Thus, there are a total of 12 models (one for each segment for both types of a model) that are deemed useful for application. Due to the large number of models, we only discuss the following:

- the model with the best AUC and less uncertainty,
- the independent variables acquired after optimizing AUC with backward elimination and forward selection, and
- the explanation of why these variables optimize AUC.

4.1 The Finance Sector

4.1.1 Performance Relative to the SET Index Model

The summary of AUCs for stock's performance relative to the SET index for the finance sector are presented in Table 11.

Table 11: Relative Performance Model in Finance

Classifier	AUC (Neutral)	AUC (Optimistic)	AUC (Pessimistic)
KNN	0.759	0.816	0.702
LR	0.744	-	-
LDA	0.743	-	-
DT	0.713	0.794	0.631

In Table 11, the neutral AUC is the average of pessimistic and optimistic AUCs. The optimistic and pessimistic measurements of AUC for KNN and DT classifiers are different from neutral AUCs. This is because, as mentioned in Section 3.3, since KNN and DT do not have an estimated probability function such as a parametric classifier, the ROC curve is plotted by adjusting the threshold based on the proportion score, which comes from the majority proportion in the leaf node for DT and the majority proportion from KNN. Therefore, some of the testing data end up with the same proportion score (if the testing instances end up at the same leaf node or the same KNN),

and at this same score/threshold, some of the testing instances are true positive and some are false positive. For DT, more testing instances will land on the same score/threshold if there are many pure leaf nodes (leaf nodes that have only one instance; thus, the score/threshold of the testing instance becomes 1). The pessimistic AUC is constructed by counting the false positive instances before true positive instances, at the same score/threshold; thus results in the ROC curve leaning more toward the right direction and vice versa for optimistic AUC. By contrast, LR and LDA have an estimated probability function that makes the score for threshold adjustment continuous. Thus, LR and LDA have no testing instances that land on the same score/threshold. Consequently, the pessimistic and optimistic AUCs for these two classifiers are always the same.

From these results, a risk-averse investor would choose LR as the best model with stable AUC at 0.744. Although KNN has the highest neutral AUC, the classifier still suffers uncertainty from optimistic and pessimistic measurements.

There are slight differences in the AUCs of LDA and LR. The optimized independent variables in both LDA and LR are exactly the same (see Table 12). The small difference could arise from the fact that LDA requires independent variables to be normally distributed, but LR does not.

Table 12: Independent Variables of LR for Relative Performance Model in Finance

Attribute	Coefficient ↑
ROA_YEAR 2	-0.038
Intercept	-0.029
GICs Sector	-0.024
PROFIT MARGIN_YEAR 3	-0.023
PE_YEAR 1	-0.005
Industry Group	-0.005
REVENUE GROWTH_YEAR 3	-0.004
PE_YEAR 3	-0.001
INCOME GROWTH_YEAR 3	0.003
PROFIT MARGIN_YEAR 2	0.017
DIV YIELD_YEAR 3	0.023
PB_YEAR 2	0.195

Taking a look at some of these variables, their distribution is far from normal (Figure 19 and Figure 20).

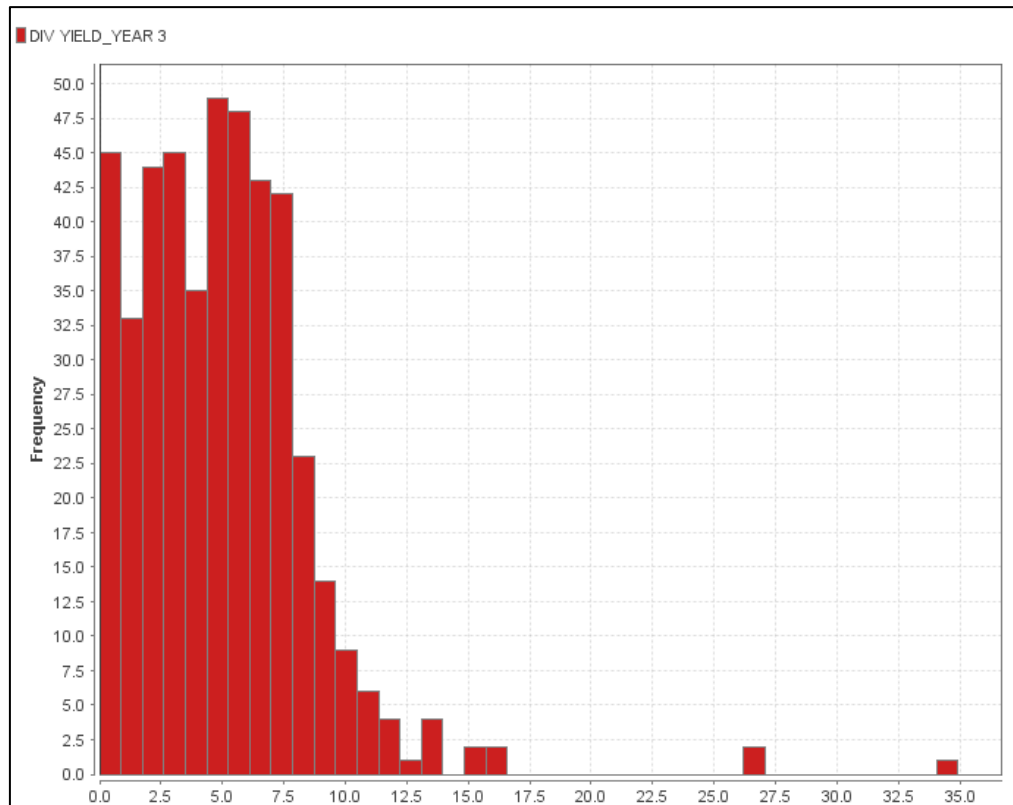


Figure 19: Distribution of DIVIDEND_YEAR 3 in the Finance Sector

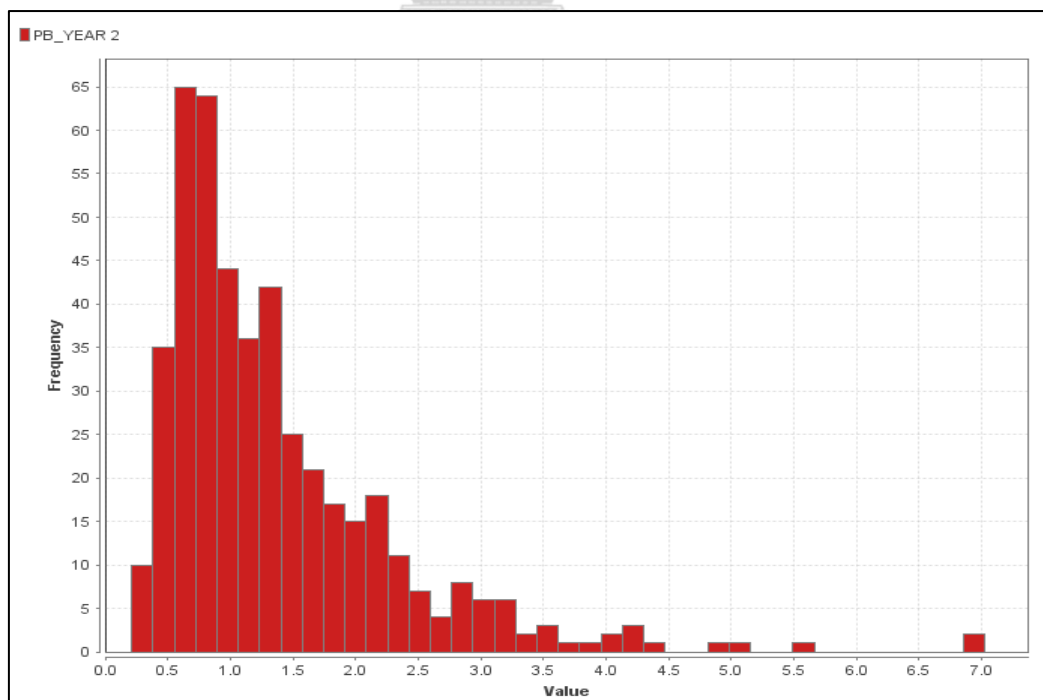


Figure 20: Distribution of PB_YEAR_2 in the Finance Sector

As the distribution of some of these optimized variables is not normal, the LR classifier becomes slightly more robust in predicting. Contrastingly, it has also been proved that the violation of assumption does not fatally hinder the LDA model's performance [20].

The optimized independent variables for LR and LDA models include Industry Group, GICS Sector, PE, PB, ROA, Profit Margin, Dividend Yield, Income Growth, and Revenue Growth as shown in Table 12.

An interesting observation can be found when looking at Industry Group, which consists of industries within the finance sector. Companies within the finance sector are divided into five different industries, including Real Estate, Diversified Financial Services, Insurance, Banking and Real Estate Investment Trusts (REITs). According to the overall trend, there are more data instances classified as "Underperform", but this is different from the Real Estate group as illustrated in Figure 21. There are more instances in which Real Estate outperforms when the performance of this sector should be close to that of REITS as both industries are real estate investment. However, in these data, a significant proportion of instances in the REITS group are classified as underperform. This is due to the fact that companies investing in REIT and Real Estate have different risks. Due to its nature, an REIT is less risky as a company invests in a secondary market and can invest any amount in the trust fund. However, direct real estate financing requires much cash and offers a higher return because of its higher entry cost and risk. This scenario demonstrates that although REITS companies can generate returns, these returns are not high enough to outperform the market. In addition, investors view real estate firms as more risky; thus, the companies' stocks are more risky and generate higher returns.



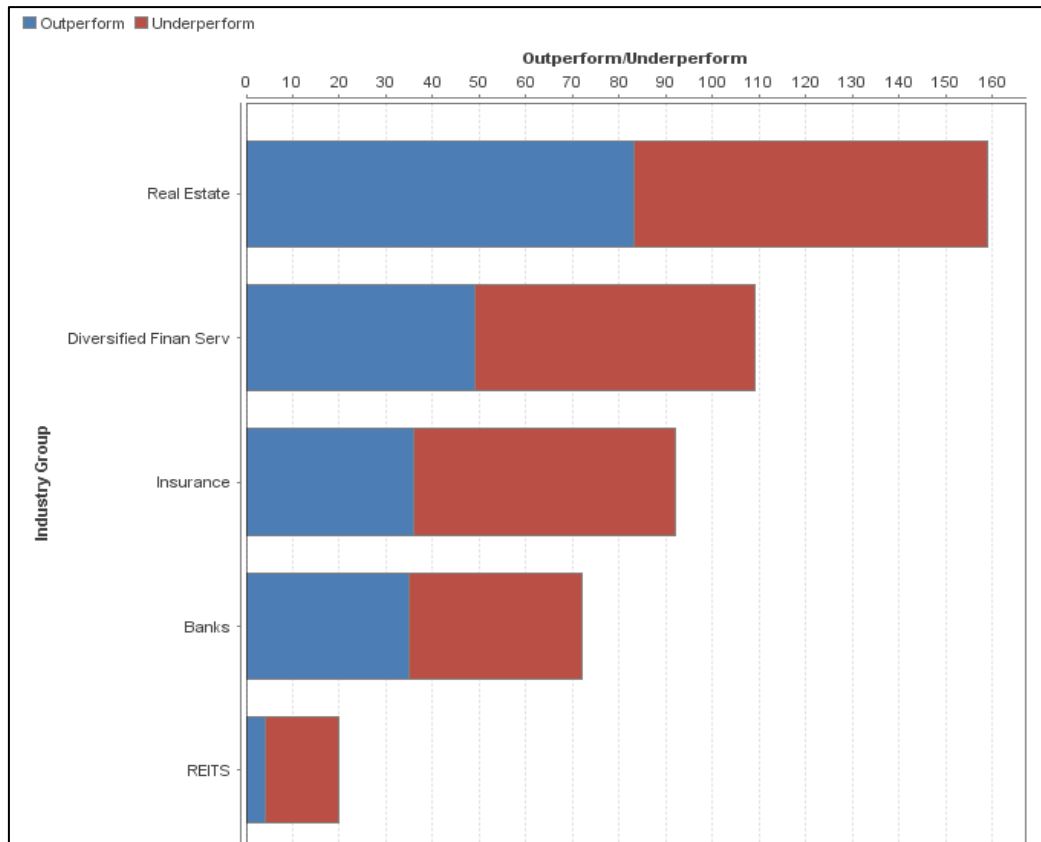


Figure 21: The Finance Sector by Industry Group

For LR and LDA, the PB ratio indicates a positive correlation with likelihood to outperform the market. As the finance sector is dominated by real estate firms with a high proportion of tangible assets (building and property), these companies have a high depreciation cost, which lowers the book value and increases the PB ratio. This is probably the reason why higher PB leads to outperformance as the ratio indicates real estate firms that are more likely to outperform.

Another observation to note is that PE coefficients in LR and LDA are negative; thus, higher PE leads to decreased probability of outperformance. This is common in how investors view stocks in the market as higher PE within the same sector means the stock is more expensive than it should be. Income growth and dividend yield have a positive correlation with likelihood to outperform as higher income growth is expected of a good company, and dividend yield reflects the reward to an investor for investing in the company.

Many of the regression coefficients are counterintuitive. For example, higher ROA leads to a decrease in outperformance likelihood. A case such as this can be demonstrated in the distribution of ROA_YEAR 2, in which higher values (19–22) are dominated by the Underperform class, as illustrated in Figure 22.

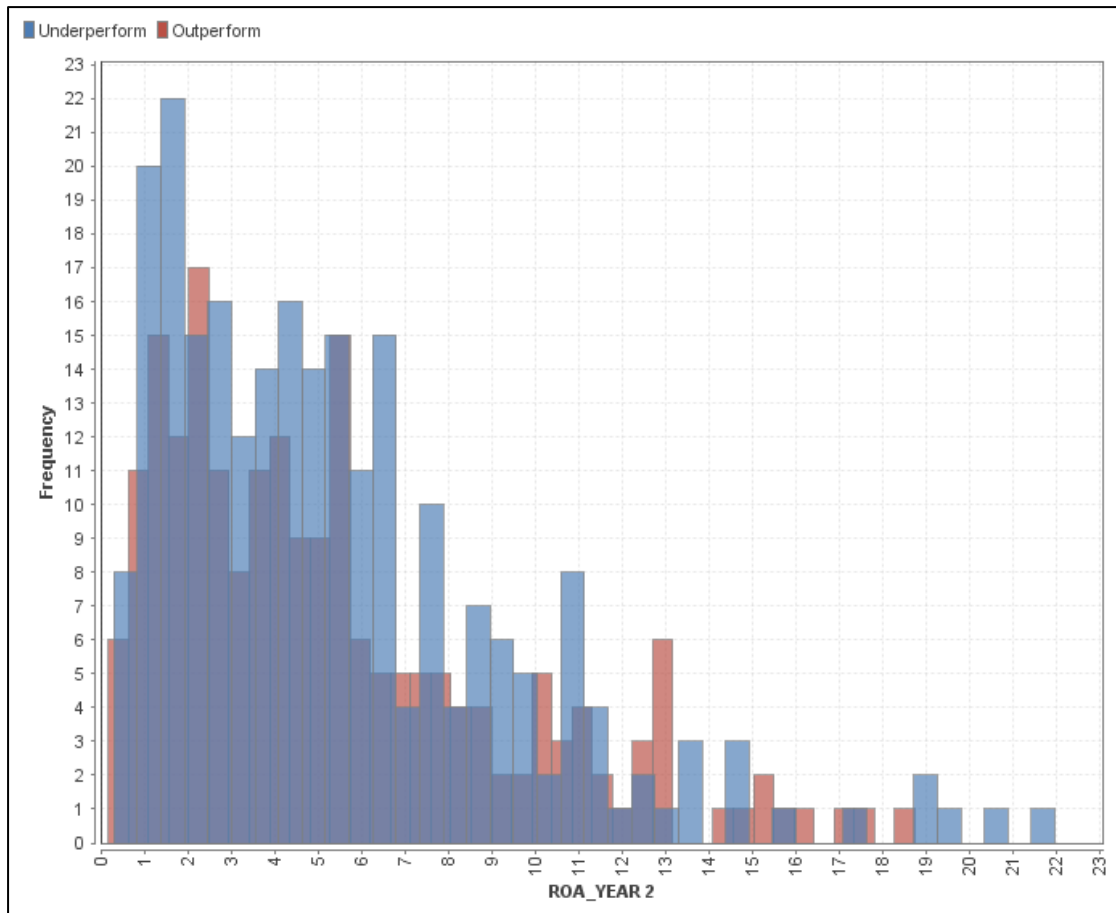


Figure 22: ROA_YEAR 2 of the Finance Sector

Despite some counterintuitive correlation, the LR classifier accurately classified the testing data with AUC of 0.744.

4.1.2 The Price Movement Model

The resultant AUCs for one-year price movement models are presented in Table 13.

Table 13: The Price Movement Model in the Finance Sector

Classifiers	AUC (Neutral)	AUC (Optimistic)	AUC (Pessimistic)
KNN	0.724	0.760	0.687
LR	0.711	-	-
LDA	0.712	-	-
DT	0.661	0.872	0.460

In Table 13, KNN provides slightly higher neutral AUC than the rest of the model in the finance sector. At the same time, using KNN comes with uncertainty of optimistic and pessimistic measurements. The trend also demonstrates that LR and LDA are much similar when used in the finance sector for prediction, regardless of the model type as the AUCs are very close to one another in both cases. The lower uncertainty between

pessimistic and optimistic AUCs for the KNN classifier (± 0.036 from neutral AUC for this model) indicates that a smaller number of testing instances fall into the same proportion score despite the same amount of optimized neighbors ($K = 183$) having been used. This means KNN is much more precise in predicting price movement than predicting whether an instance outperforms the SET index (± 0.057 from neutral AUC for performance relative to the SET Index model).

PB_YEAR 3 (third year of the observation period) for the price movement model in LDA has a negative correlation with the probability of positive price movement. This is common for investors when they expect the price of a stock to rise, as lower PB means the stock is cheap and likely to have a positive return.

Table 14: Independent Variables of LDA for Relative Price Movement Model in Finance

Attribute	Coefficient
GICS Sub Industry	0.000
PB_YEAR 1	0.048
PB_YEAR 3	-0.104
ROIC_YEAR 1	0.008
ROA_YEAR 1	-0.003
ASSET TURNOVER_YEAR 1	-0.199
ASSET TURNOVER_YEAR 3	0.096
REVENUE GROWTH_YEAR 3	-0.000
INCOME GROWTH_YEAR 3	0.000
NET DEBT TO EQUITY_YEAR 3	0.000
PROFIT MARGIN_YEAR 1	-0.000
(Intercept)	0.649

Another interesting note is that the GICS Subindustry plays a role in improving a classifier. In Figure 24, the GICS Subindustry divides the industries within the finance sector more thoroughly (Figure 23).

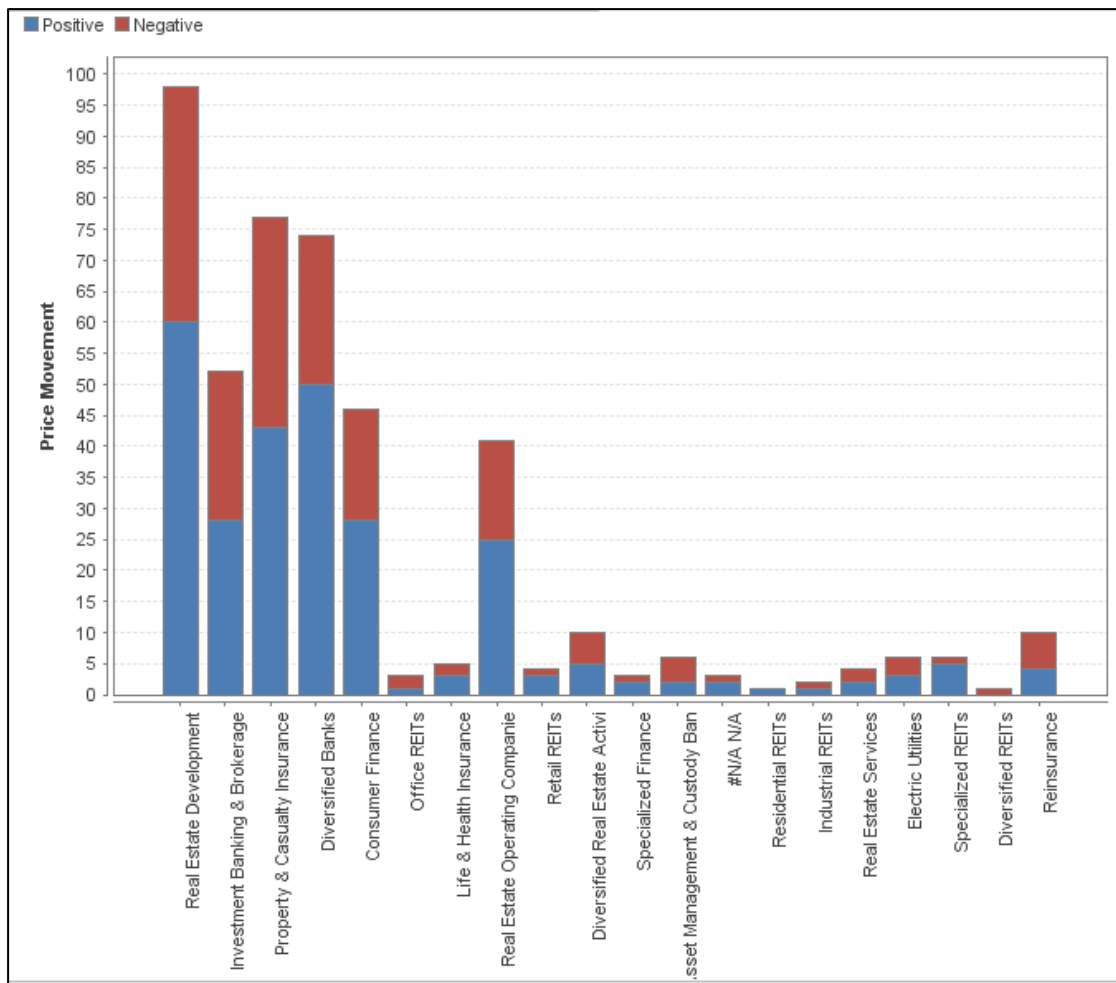


Figure 23: The GICS Subindustry in the Finance Sector

Although most subindustries are dominated by the “Positive” classification, the interaction of these variables with others should provide a better insight into improving the classification model as illustrated in Figure 24: Interaction of the GIC Sub, in which the *GIC Subindustry* is plotted with *NET_DEBT_TO_EQUITY_YEAR 3*.

According to the regression coefficient, higher *NET_DEBT_TO_EQUITY_YEAR 3* provides a higher tendency for a positive return. This can be observed in a scatterplot, especially for the real estate subindustry.

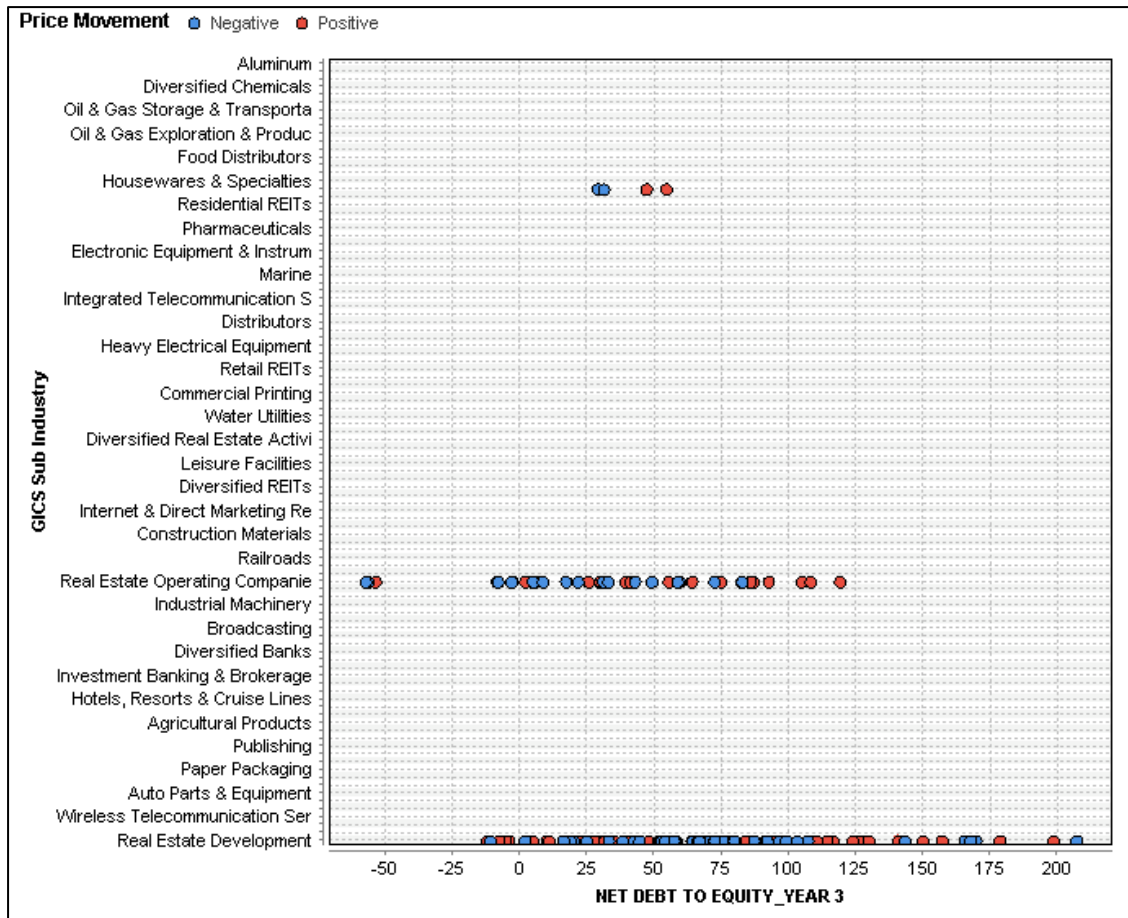


Figure 24: Interaction of the GIC Subindustry & NET_DEBT_TO_EQUITY_YEAR 3

The reason behind the positive return for the real estate subindustry when Net Debt to Equity is higher is that the companies investing in real estate need to borrow money for large upfront investment for their businesses; therefore, it not uncommon for these companies to have a large debt ratio. Thus, investors found it acceptable for such companies to have high debt. Net Debt to Equity has an positive effect on the LDA classifier's performance probably because Net Debt to Equity improves the model's performance in this subindustry.

In addition to high net debt to equity, ROIC also has a positive correlation with the likelihood of a positive return. ROIC measures the profitability of a company based on its invested capital, which includes the company's interest-bearing debt. Investors view ROIC as the measurement of effectiveness in utilizing debt. Thus, higher ROIC means a company invests more effectively and thus has increased likelihood of a positive return.

4.2 The Consumer Cyclical Sector

4.2.1 Performance Relative to the SET Index Model

The summary of AUCs for a stock's performance relative to the SET index for the cyclical consumer sector is presented in Table 15.

Table 15: The Relative Performance Model for the Consumer Cyclical Sector

Classifiers	AUC (Neutral)	AUC (Optimistic)	AUC (Pessimistic)
KNN	0.715	0.745	0.685
LR	0.723	-	-
LDA	0.773	-	-
DT	0.643	0.680	0.606

In Table 15, it is quite clear that the best model with the highest AUC is LDA. To train this LDA model, unlike in the finance sector, the qualitative independent variables are converted into dummy variables that represent all categories of the variables. This results in a total of above 500 independent variables since there are many categories to account for when dummy coding creates variables for all sectors, industries and subindustries. With significantly more variables to process, the optimization using backward elimination and forward selection takes more runtime to go through more possible combinations. With more variables during the optimization, the LDA algorithm has more degrees of freedom to choose variables and thus provides more possibilities to select a combination that gives higher AUC performance. Despite more possibilities in dummy coding, the method does not guarantee a higher AUC as illustrated in the finance sector, which uses unique integer conversion instead of dummy coding.

```

0.088 * BICS Level 2 Industry Group = Industrial Services
+ 0.151 * BICS Level 2 Industry Group = Chemicals
- 0.087 * GICS Industry = Materials
- 0.404 * GICS Industry = Energy
+ 0.199 * GICS Industry = Commercial & Professional Serv
- 0.082 * GICS Sub Industry = Specialty Stores
- 0.460 * GICS Sub Industry = Internet & Direct Marketing Re
- 0.290 * GICS Sub Industry = Office Services & Supplies
- 0.054 * GICS Sub Industry = Home Furnishings
- 0.162 * GICS Sub Industry = Housewares & Specialties
+ 0.010 * GICS Sub Industry = Department Stores
+ 0.026 * ICB Sector = Household Goods & Home Construction
+ 0.091 * ICB Sector = General Retailers
+ 0.387 * ICB Sector = Food & Drug Retailers
- 0.032 * Industry Group = Leisure Time
- 0.084 * Industry Index Name = SETENTER
- 0.001 * Industry Issuer = Home Furnishings
- 0.241 * Industry Issuer = Intimate Apparel
+ 0.003 * PE_YEAR 1
- 0.001 * PE_YEAR 3
- 0.001 * ROE_YEAR 1
- 0.000 * REVENUE GROWTH_YEAR 3
+ 0.000 * INCOME GROWTH_YEAR 2
+ 0.402

```

Figure 25: LDA Regression Equation for Relative Performance Model in Consumer Cyclical

According to the result of optimized variables (see Figure 25), a stock's performance relative to the SET Index in the consumer cyclical sector is primarily described by market segmentation variables or dummy variables. Notable dummy variables are the GICS industry and GICS subindustries. The LDA indicates that dummy variables representing Energy and Materials in the GICS Industry have a negative correlation with likelihood to outperform the market. This can be accounted by the majority proportion of underperformance instances in both industries. By contrast, the dummy variable representing commercial and professional services in the GICS industry has a positive correlation with likelihood of outperformance, as the majority of data instances in this industry outperformed the market.

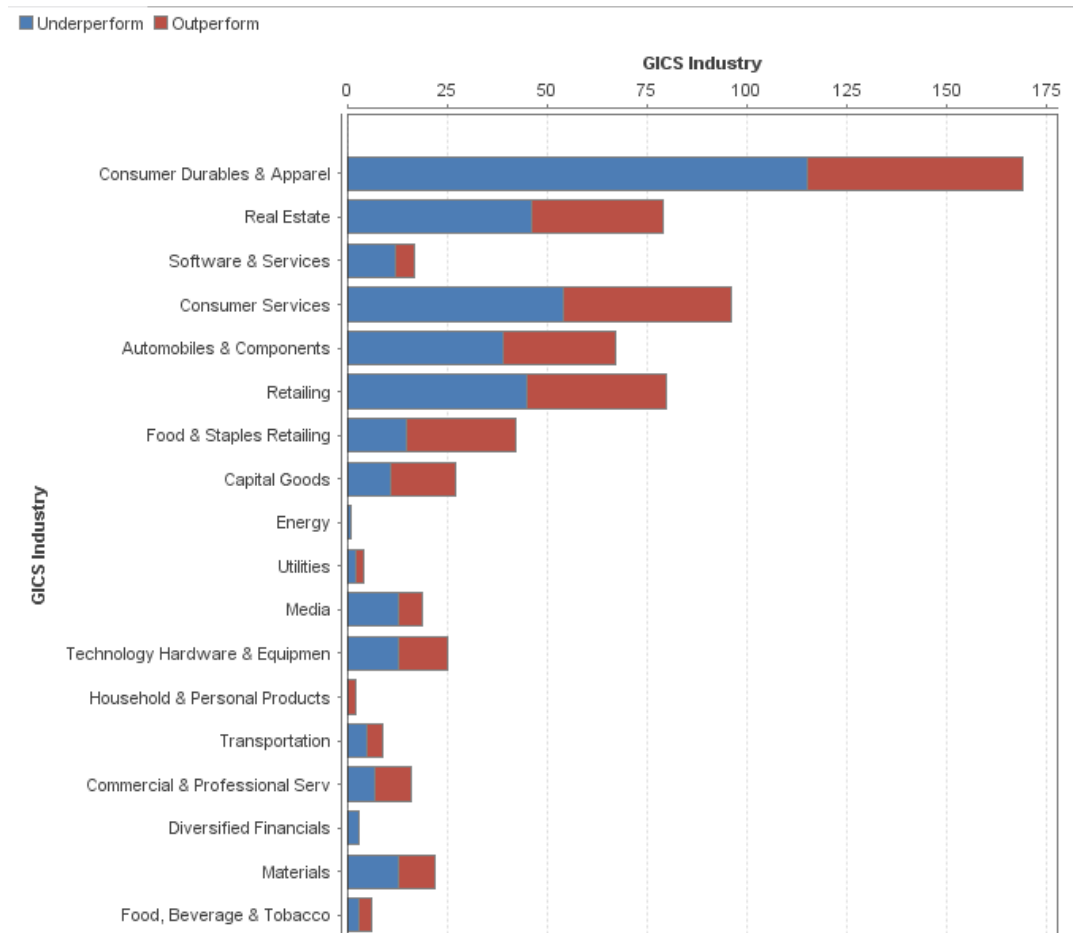


Figure 26: The GICS Industry of the Consumer Cyclical Sector

The LDA also delves deeper into market segmentation as many of GICS subindustry dummy variables increase the model's AUC. The dummy variables of the GICS subindustry that have a negative correlation with the probability of outperforming the market are Home Furnishing, Houseware and Specialty, Internet and Direct Marketing, Office Service and Supply and Specialty Stores. These subindustries have underperformance instances as a major classification, thus giving negative regression coefficients. However, the dummy variable representing Department Stores in the GICS subindustry has outperformance instances as the majority class, thus giving a positive correlation and coefficient of the LDA regression equation (see Figure 27).

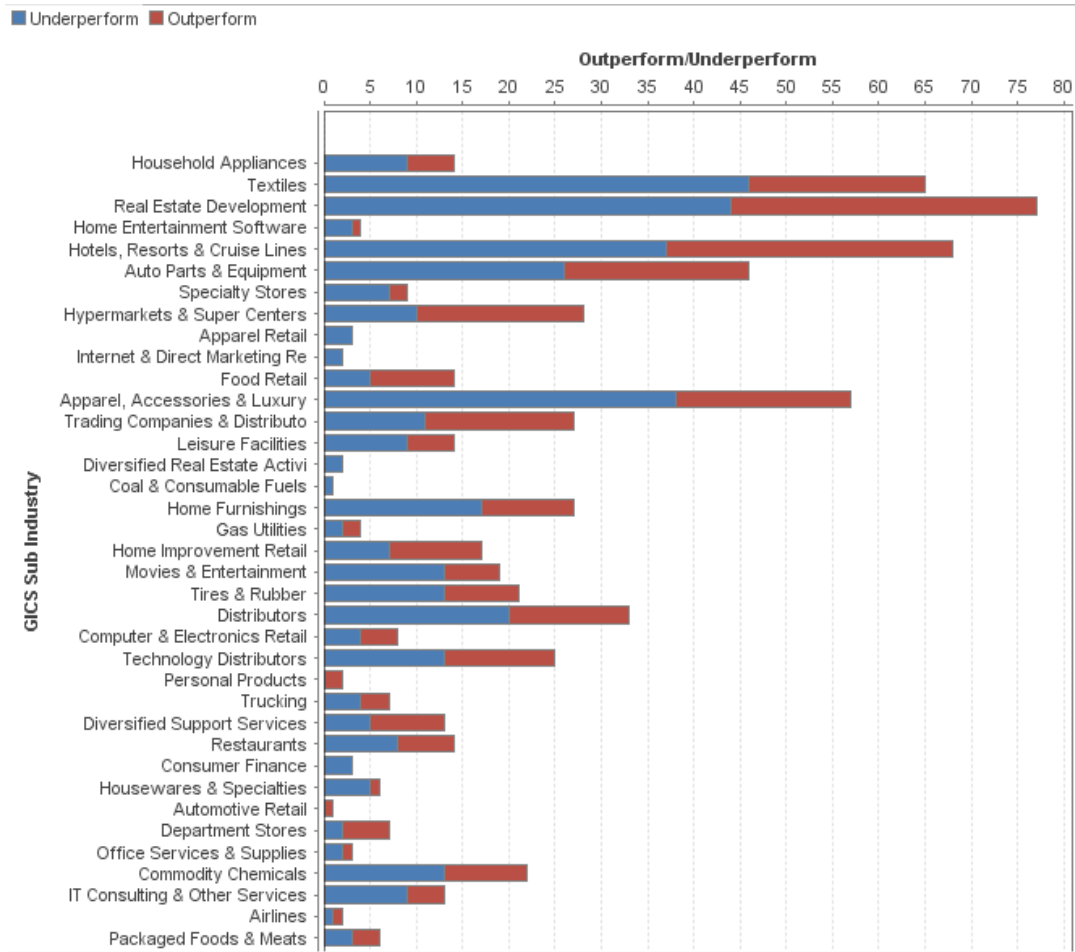


Figure 27: The GICS Subindustry of the Consumer Cyclical Sector

Other than the GIC Industry and GIC Subindustry dummy variables, other types of segmentation variables such as BICS Level 2 Industry Group and ICB Sector also appear to play a role in improving AUC of this model. There are only 4 out of 33 financial ratios presented as independent variables for optimized AUC. Observation can be drawn that the consumer cyclical sector is a largely generalized sector consisting of many subindustries with their own unique characteristic. This is the reason why identifying a company's industry and subindustry plays a greater role in predicting the performance of stocks relative to the SET Index than identifying financial ratios does. It also demonstrates that financial ratios in this sector are too vague to predict the performance as the sector contains too many characteristics that cannot be described by this mix of financial ratios. For investment application, formulating an investment strategy in this sector relies heavily on identifying a proper industry or subindustry in which to invest.

4.2.2 Price Movement Model

The resultant AUCs for one-year price movement models are presented in Table 16.

Table 16: The Price Movement Model for the Consumer Cyclical Sector

Classifier	AUC (Neutral)	AUC (Optimistic)	AUC (Pessimistic)
KNN	0.675	0.712	0.638
LR	0.658	-	-
LDA	0.693	-	-
DT	0.629	0.665	0.602

Although LDA remains the best model with the highest AUC, it is important to note that overall, the AUCs are significantly lower compare to performance relative to market models. The results of independent variables that optimized AUC are also the same as those of the previous model, in which sector/industry segmentation dummy variables were the main source of the prediction score in the regression equation. Therefore, the same conclusion can be drawn: the sector consists of too many characteristics that cannot be described by financial ratios.

Table 17: LDA Regression Equation for Price Movement Model in Consumer Cyclical

BICS Level 2 Industry Group = Manufactured Goods	-0.053
BICS Level 2 Industry Group = Recreation Facilities & Svcs	0.180
GICS Industry = Commercial & Professional Serv	0.051
GICS Sub Industry = Household Appliances	0.007
GICS Sub Industry = Food Retail	0.004
GICS Sub Industry = Diversified Real Estate Activi	-0.482
GICS Sub Industry = Housewares & Specialties	-0.539
GICS Sub Industry = Automotive Retail	0.466
GICS Sub Industry = Department Stores	0.061
GICs Sector = Consumer Staples	0.004
Industry Group = Retail	0.013
Industry Group = Apparel	-0.055
Industry Index Name = SETENTER	-0.064
Industry Index Name = SETCONMT	-0.102
PE_YEAR 3	-0.001
(Intercept)	0.544

The dummy variable with the highest absolute weight in the LDA equation is the variable that represents Houseware and Specialty for the GICS subindustry in which

the subindustry is completely dominated by negative price movement instances, thus resulting in a strong negative correlation to likelihood for positive price movement. The next highest absolute weight in line is the dummy variable that represents Diversified Real Estate Activities in the GICS subindustry. This subindustry is also completely occupied by negative price movement instances, thus also giving a strong negative correlation.

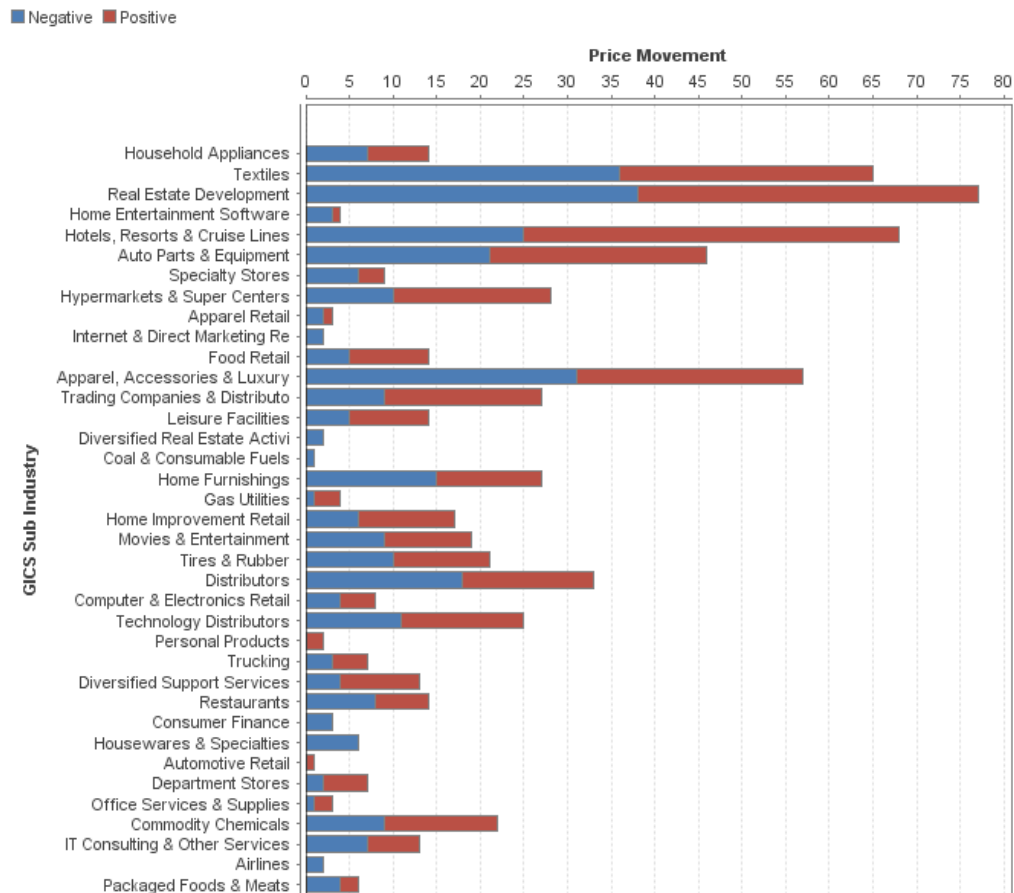


Figure 28: The GICS Subindustry for the Consumer Cyclical Sector (Price Movement)

Since the performance relative to market models have higher AUC that is primarily characterized by industry and subindustry variables, the industry segmentation is better at predicting whether a stock will win the market. In other words, it is easier to predict price movements when the movements are benchmarked by the SET market. The lower AUC also proves that the price movement of a stock in the consumer cyclical sector is much more erratic than the performance relative to the market. This reasoning is further reinforced by the fact that this sector has a large spread of industries within one sector, thus making the prediction model less robust as the classifiers cannot capture all characteristics.

4.3 The Consumer Non-Cyclical Sector

4.3.1 Performance Relative to the SET Index Model

The summary of AUCs for stocks' performance relative to the SET index for the non-cyclical consumer sector is presented in Table 18.

Table 18: The Relative Performance Model's AUC for the Consumer Non-Cyclical Sector

Classifier	AUC (Neutral)	AUC (Optimistic)	AUC (Pessimistic)
KNN	0.732	0.760	0.705
LR	0.723	-	-
LDA	0.740	-	-
DT	0.769	0.825	0.712

In this sector, the top models seem to be LDA and DT with pessimistic AUC of DT very close to LR. However, the independent variables that optimized AUC for these two models are very different. LDA variables contain almost an equal number of industry segmentation dummy variables and financial ratios, whereas DT is mostly explained by financial ratios. The difference in the types of variables on these two models can be explained by the manner in which each model works.

DT is constructed by finding the best combination within each variable to best separate the two classes and using information gained as a measurement. Thus, DT picks out anything that separates the two classes without considering the importance or weight of each variable.

```

PE_YEAR 2 > 6.240
| REVENUE GROWTH_YEAR 2 > -7.278
| | GICs Sector = Consumer Discretionary: Underperform {Underperform=7, Outperform=3}
| | GICs Sector = Consumer Staples
| | | ROIC_YEAR 2 > 10.696
| | | | PE_YEAR 1 > 1.185: Outperform {Underperform=23, Outperform=28}
| | | | PE_YEAR 1 ≤ 1.185: Underperform {Underperform=17, Outperform=3}
| | | | ROIC_YEAR 2 ≤ 10.696
| | | | REVENUE GROWTH_YEAR 2 > 15.750: Underperform {Underperform=12, Outperform=7}
| | | | REVENUE GROWTH_YEAR 2 ≤ 15.750: Outperform {Underperform=19, Outperform=36}
| | GICs Sector = Health Care
| | | PE_YEAR 2 > 26.375: Underperform {Underperform=8, Outperform=2}
| | | PE_YEAR 2 ≤ 26.375
| | | | REVENUE GROWTH_YEAR 2 > 2.436
| | | | | ROIC_YEAR 2 > 9.047: Outperform {Underperform=21, Outperform=41}
| | | | | ROIC_YEAR 2 ≤ 9.047: Underperform {Underperform=5, Outperform=3}
| | | | REVENUE GROWTH_YEAR 2 ≤ 2.436: Outperform {Underperform=0, Outperform=6}
| | GICs Sector = Industrials: Outperform {Underperform=12, Outperform=19}
| REVENUE GROWTH_YEAR 2 ≤ -7.278: Underperform {Underperform=17, Outperform=6}
PE_YEAR 2 ≤ 6.240: Underperform {Underperform=17, Outperform=1}

```

Figure 29: DT of the Relative Performance Model for the Consumer Non-Cyclical Sector

From the resultant DT model in Figure 29, DT could have picked a sector that is likely to outperform the market naturally and classify any testing instance that falls into the sector as outperforming without considering financial ratios. In the above result, anything that falls into Industrials GICS Sector is classified as outperforming because most of the training instances in this leaf node actually outperform the market. Conversely, any instance that falls into the Consumer Discretionary GICS Sector is classified as underperforming. These results of DT demonstrated that the model easily identifies a sector that is likely to grow faster or slower than the market. This could be useful for further research on why specific sectors tend to outperform or underperform the SET Index.

The leaf nodes in *GICS Sector = Consumer Staple*, in which $ROIC_YEAR\ 2 > 10.696$ indicate that companies with a high PB ratio (>10.696) tend to outperform the market, whereas a low PB ratio tends to underperform the market. This is not a common assumption amongst value investors as a lower-PB stock can be viewed as an undervalued stock. By contrast, a low PB value can reflect that investors see no growth potential in a company and thus do not buy its stock to drive the price up. In the other branch in which $ROIC_YEAR\ 2 \leq 10.696$, the only reasonable explanation for why high revenue growth correlates with underperformance and low revenue growth correlate to outperformance is that investors who invest in this sector also look at other financial ratios other than only ROIC and revenue growth, but this DT is unable to distinguish those characteristics.

For the *GICS Sector = the Health Care* node, the pattern is quite common as lower PE as well as higher revenue growth and ROIC lead to outperformance, and vice versa.

As for the LDA model, despite having various variables, the magnitude of weight in the regression equation seems to be heavy for dummy variables only (see Table 19). As these weights directly affect the probability of outperformance, identifying what industry or subindustry a company is in plays a major role in proper investment in this sector. From coefficient, a company in the Household and Personal Product GICS Industry is an attractive company to invest in because of the industry's high likelihood to outperform.

Another weight to notice in the equation is PB_YEAR 1, which is 0.071, because the average for this variable is around 2, which could significantly affect the score. PB_YEAR 1 in this regression function has the same characteristic as in DT, in which higher PB_YEAR 1 lead to more likelihood of outperformance (positive correlation). Therefore, this parameter could represent the *GICS Sector = the Consumer Staple*, which is the dominant GICS sector, and the same intuition can be derived as low PB value in this sector means that a company has no growth potential.

Table 19: Regression Coefficient for Relative Performance LDA Model (Consumer Non-Cyclical Sector)

Attribute	Coefficient
BICS Level 2 Industry Group = Health Care Facilities & Svcs	0.026
BICS Level 2 Industry Group = Biotech & Pharma	-0.237
GICS Industry = Household & Personal Products	0.264
ICB Industry Name = Consumer Goods	-0.117
Industry Index Name = SETAGRI	-0.013
Industry Index Name = SETPERS	-0.125
Industry Issuer = Food-Meat Products	0.002
PE_YEAR 3	-0.002
PB_YEAR 1	0.071
PB_YEAR 2	-0.042
ROIC_YEAR 1	-0.002
INCOME GROWTH_YEAR 2	0.000
PROFIT MARGIN_YEAR 2	-0.003
(Intercept)	0.595

4.3.2 The Price Movement Model

The summary of AUC for price movement models for the non-cyclical consumer sector is illustrated in Table 20.

Table 20: The Price Movement Model for the Consumer Non-Cyclical Sector

Classifier	AUC (Neutral)	AUC (Optimistic)	AUC (Pessimistic)
KNN	0.7	-	-
LR	0.784	-	-
LDA	0.786	-	-
DT	0.663	0.8	0.557

In this case, the most useful model with the highest AUC is clearly LDA. The independent variables that optimized AUC mostly include financials ratios as presented in Table 21.

Table 21: Regression Coefficient for Price Movement LDA Model (Consumer Non-Cyclical Sector)

Attribute	Coefficient ↑
PB_YEAR 3	-0.062
ROA_YEAR 2	-0.020
PE_YEAR 3	-0.003
ICB Sector	-0.003
INCOME GROWTH_YEAR 3	-0.001
NET DEBT TO EQUITY_YEAR 1	0.000
ROIC_YEAR 3	0.006
PROFIT MARGIN_YEAR 1	0.011
PB_YEAR 1	0.020
ICB Industry Name	0.024
PB_YEAR 2	0.044
(Intercept)	0.629

All the PBs for 3 years are included in this model, in which PB_YEAR 3 has a negative correlation while PB_YEAR 1 and PB_YEAR 2 have positive correlation with likelihood of getting a positive price movement. Normally, investors would look for low PB as it represents an undervalued stock that will later have a positive price movement; however, PB_YEAR 3 correlation indicates that higher PB leads to lower probability of positive price movement. Observing it for one year, PB might not make logical sense; however, observing all PBs together can give a different interpretation. In this case, investors might want to see a lower PB in year 3 as the model assumes that investors will invest after observing financial ratios of year 3. Thus, a lower PB at the end of year 3 means the stock is cheap when investors are ready to invest; thus, lower PE_YEAR 3 leads to a higher probability of positive price movement. Additionally, if investors see a high-PB record in years 1 and 2 but lower PB in year 3, they would expect the PB to rise again to the same level as that of the previous year, thus expecting positive price movement. To sum this up, a stock has a higher probability of positive price movement when PB_YEAR 3 is lower and investors are ready to buy them (as the model assumed); at the same time, the stock should have a historical record of high PB (in years 1 and 2) to prove that the company can grow to that level.

Another interesting part of this model is the ICB Industry, which also gives weights as high as PBs. For this model, the ICB Industry variable uses a unique integer as a coding method for categorical variable conversion. The drawback of this method is that it does not guarantee better AUC performance than dummy coding, and it assumes that some industries are “more” than others by assigning a unique real integer to them.

For this LDA model, there are four categories in the ICB Industry, including Healthcare, Consumer Goods, as well as Industrial and Consumer Services, and their unique integers are 4, 2, 7 and 1, respectively. The values of 7 and 4 are assigned to Industrial and Healthcare, respectively, which have higher proportions of the positive price movement class (Figure 30). Since the ICB industry has a positive correlation with the positive price movement probability, the unique integer coding's drawback inadvertently becomes the strength of the model instead, because the coding algorithm assumes that Industrial and Healthcare have more value than Consumer Goods and Consumer Service, thus allowing the model to distinguish industries that have a higher tendency to give positive price movement and AUC, which rely on true positive measurements.

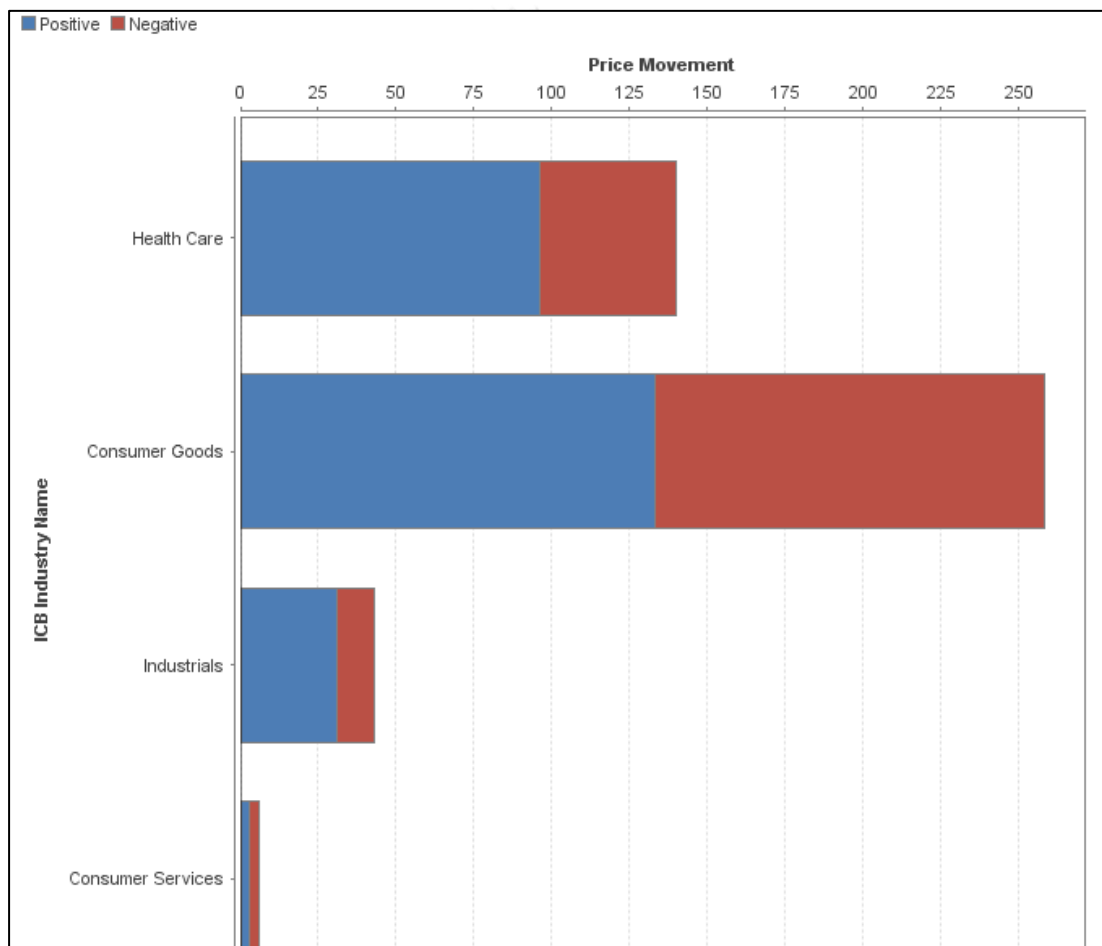


Figure 30: The ICB Industry in the Consumer Non-Cyclical Sector

4.4 The Industrial Sector

4.4.1 Performance Relative to the SET Index Model

The summary of AUCs for stocks' performance relative to the SET index for the industrial sector is presented in Table 22. The results indicate that LDA has the highest AUC with the least uncertainty from optimistic and pessimistic measurements.

Table 22: Relative Performance Model AUC for the Industrial Sector

Classifier	AUC (Neutral)	AUC (Optimistic)	AUC (Pessimistic)
KNN	0.743	0.707	0.672
LR	0.733	-	-
LDA	0.789	-	-
DT	0.610	0.790	0.430

According to independent variables for which AUC for LDA model was optimized, the industry and subindustry with a high proportion of outperformed instances were chosen. These variables include *BICS Level 2 Industry Group = Waste and Environmental Services and Equipment*, *ICB Industry Name = Consumer Goods*, and *Industry Issuer = Container-Metal/Glass* which have the highest positive weight in the LDA equation. From the investment perspective, the model helps investors to identify attractive industries/subindustries in which to invest. Historical records indicate that stocks in these industries/subindustries are likely to outperform the SET Index.

Table 23: Independent Variables for the LDA Relative Performance Model (Industrial Sector)

Attribute	Coefficient ↓
BICS Level 2 Industry Group = Waste & Environ Svcs & Equip	0.265
ICB Industry Name = Consumer Goods	0.131
Industry Issuer = Containers-Metal/Glass	0.086
Industry Index Name = SETTRANS	0.045
BICS Level 1 Sector Name = Consumer Discretionary	0.032
ICB Sector = Chemicals	0.026
Industry Index Name = SETIMM	0.018
Industry Issuer = Bldg Prod-Cement/Aggreg	0.009
ROIC_YEAR 3	0.003
DIV YIELD_YEAR 1	0.000
REVENUE GROWTH_YEAR 3	-0.000
ROE_YEAR 2	-0.001
REVENUE GROWTH_YEAR 1	-0.002
PROFIT MARGIN_YEAR 2	-0.005
Industry Issuer = Building&Construct-Misc	-0.089
GICs Sector = Consumer Discretionary	-0.094

In addition, the model indicates that companies in which the *GICS Sector = Consumer Discretionary* or *Industry Issuer = Building ND Construct-Misc* are not attractive to invest in as they have a high negative coefficient.

Another important aspect to note is the number of data instances representing *BICS Level 2 Industry Group = Waste and Environmental Services & Equipment*, *ICB Industry Name = Consumer Goods*, or *Industry Issuer = Container-Metal/Glass* is very small, with the number of instances for each of them lower than 40 out of 369 instances. The lower number of instances could be an indication that these industries/subindustries are small with a very small number of companies. From the macroeconomic perspective, it is normal for a small segment to outperform the market as a large competitive industry is the key driver of the SET index and is unlikely to outperform the market.

For independent variables that represent financial ratios, many of coefficients are counterintuitive such as negative coefficient for revenue growth and profit margin. The negative coefficients mean that as these variables increase the likelihood of outperformance decreases. From investors' perspective, higher profit margin and revenue growth mean companies perform better, and this should thus increase the likelihood of their outperformance. Despite the coefficients such as revenue growth and profit margin being counterintuitive from the investment perspective, there are explanations for why the LDA classifier gives these variables negative coefficients.

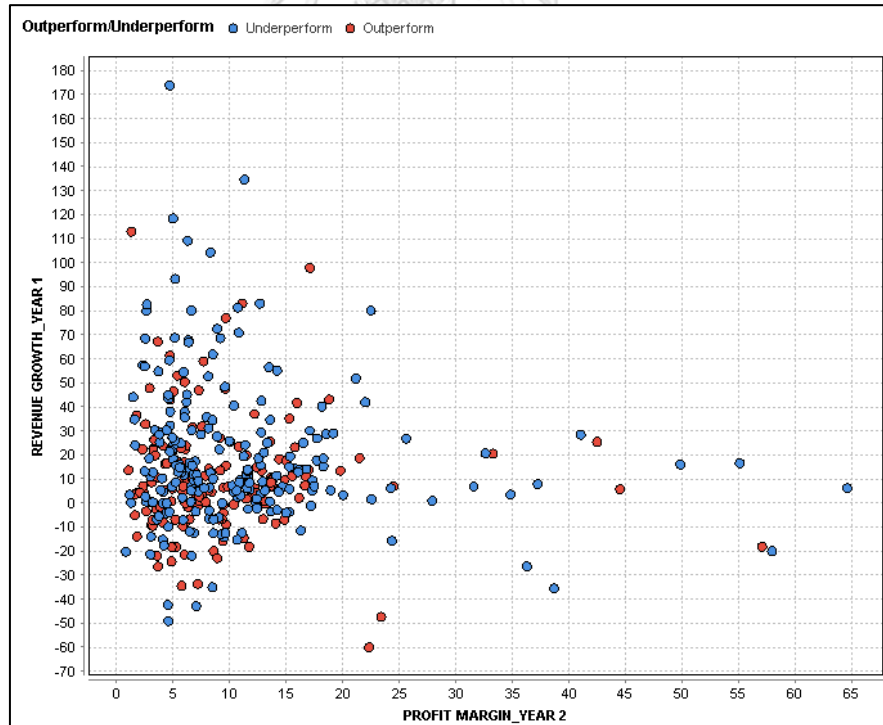


Figure 31: Scatter Plot of Revenue Growth Year 1 and Profit Margin Year 2

The scatter plot in Figure 31 demonstrates that instances with high PROFIT_MARGIN_YEAR 2 or REVENUE_GROWTH YEAR 1 are dominated by the underperformance classification, thus giving negative coefficients.

Additionally, the weights of these financial ratio variables are small. When the mean each variable is calculated, the positive and negative scores of these variables almost cancel out, thus making a small contribution to overall score in regression equation.

4.4.2 The Price Movement Model

The summary of AUC for price movement models of the industrial sector is presented in Table 24. The results indicate that LDA has the highest AUC with the least uncertainty from optimistic and pessimistic measurements.

Table 24: Price Movement Model for the Industrial Sector

Classifier	AUC (Neutral)	AUC (Optimistic)	AUC (Pessimistic)
KNN	0.712	0.752	0.671
LR	0.716	0.716	0.716
LDA	0.723	0.723	0.723
DT	0.711	0.951	0.485

Half of the independent variables that optimized AUC is industry segmentation, and the other half is financial ratios.

Table 25: Independent Variables for the LDA Price Movement Model (Industrial Sector)

Attribute	Coefficient ↓
(Intercept)	0.710
BICS Level 2 Industry Group = Automotive	0.133
BICS Level 2 Industry Group = Hardware	0.110
GICS Industry = Technology Hardware & Equipmen	0.096
BICS Level 2 Industry Group = Iron & Steel	0.053
ROA_YEAR 3	0.004
NET DEBT TO EQUITY_YEAR 3	-0.000
REVENUE GROWTH_YEAR 3	-0.001
REVENUE GROWTH_YEAR 1	-0.002
PE_YEAR 2	-0.003
BICS Level 2 Industry Group = Construction Materials	-0.008
PROFIT MARGIN_YEAR 2	-0.008
GICs Sector = Consumer Discretionary	-0.054
ASSET TURNOVER_YEAR 2	-0.109

This result indicates that the LDA classifier attempts to pick out an industry and subindustry with a positive one-year price movement. In Figure 36, the number of instances in these industries (Automotive, Hardware, Iron and Steel) is lower than that in the whole data but is dominated by positive price movement classification, thus giving positive regression coefficient and positive correlation to the probability of positive price movement. The same evidence can be observed for instances with *GICS Industry = Technology Hardware and Equipment*. Moreover, the automotive, hardware and iron and steel industries from the *BICS Level 2 Industry Group* are closely related to *GICS Industry = Technology Hardware and Equipment*; thus, all of them having positive coefficients is not surprising.

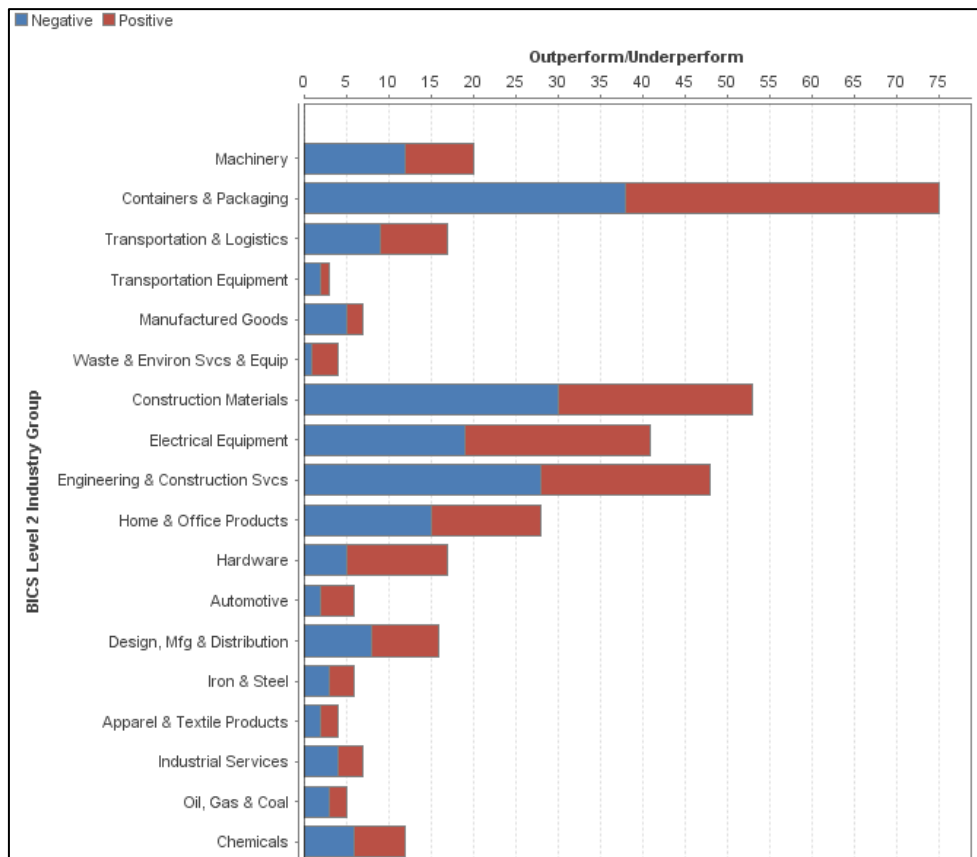


Figure 32: The BICS Level 2 Industry Group (Industrial Sector)

The top two independent variables that contribute the most to the regression score are ASSET_TURNOVER_YEAR 2 and PROFIT_MARGIN_YEAR 2. Logically, higher asset turnover and profit margin should lead to higher probability of positive price movement as asset turnover measures a company's efficiency in utilizing assets, and profit margin represents profitability. However, evidence for this sector indicates the contrary. From the scatterplot below, high ASSET_TURNOVER_YEAR 2 (top-left corner) and PROFIT_MARGIN_YEAR 2 (bottom-right corner) are mostly covered by data instances of the negative-price-movement class. Thus, it is not surprising that LDA gives a negative coefficient or higher values of these two variables lead to a lower probability of positive price movement.

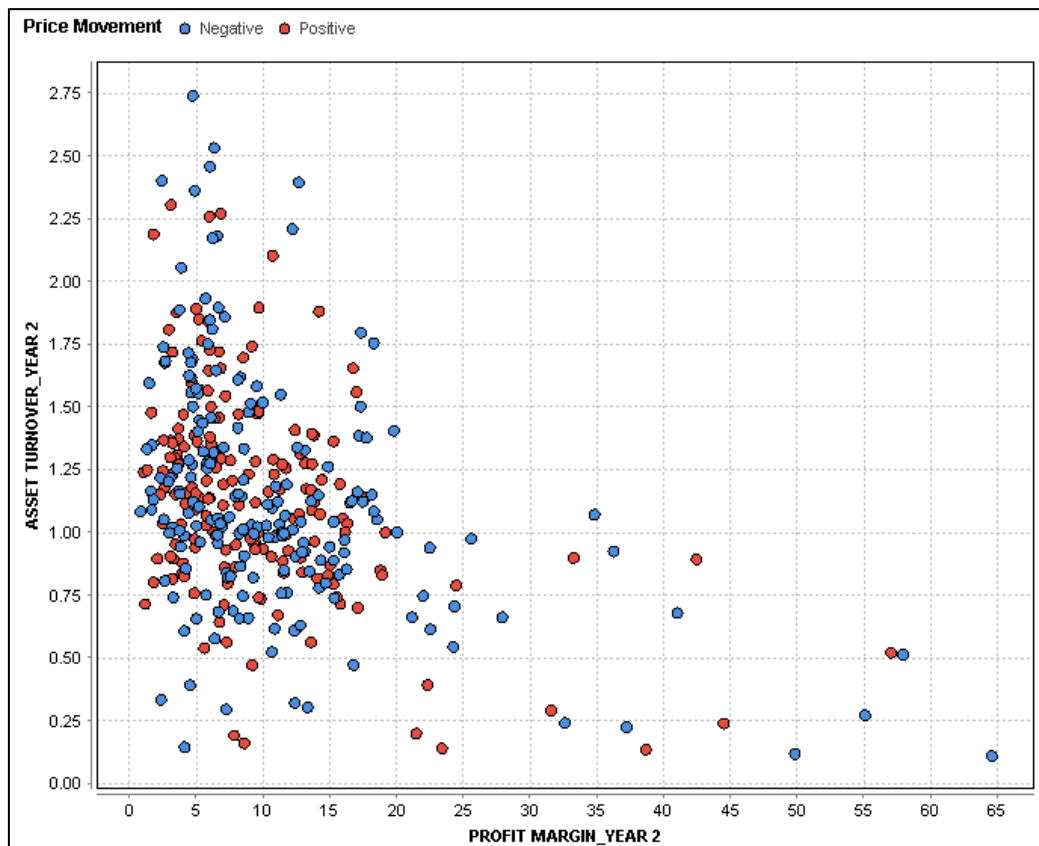


Figure 33: Profit Margin vs. Asset Turnover Scatterplot (Price Movement Model for the Industrial Sector)

An argument can be made that looking at one-year financial ratios such as asset turnover and profit margin is not enough to draw a logical conclusion for price movement or investment pattern. Thus, interactions of many financial ratios and many years should be observed instead.

4.5 Communication + Technology + Diversified Sectors

4.5.1 Performance Relative to the SET Index Model

The summary of AUCs for stocks' performance relative to the SET index models for these three sectors is presented in Table 26. The results indicate that LR has the highest AUC with the least uncertainty from optimistic and pessimistic measurements.

Table 26: Relative Performance Model AUC for Communication + Technology + Diversified Sectors

Classifier	AUC (Neutral)	AUC (Optimistic)	AUC (Pessimistic)
KNN	0.738	0.804	0.672
LR	0.823	0.823	0.823
LDA	0.814	0.814	0.814
DT	0.630	0.837	0.444

The variables that optimized AUC for the LR model include ROIC_YEAR 2, INCOME GROWTH YEAR 3, Industry Group, Industry Index and ASSET TURNOVER_YEAR 1, as illustrated in Table 27.

Table 27: Independent Variables for the LR Relative Performance Model (Communication + Technology + Diversified Sectors)

Attribute	Coefficient ↑
Intercept	-0.457
ROIC_YEAR 2	-0.003
INCOME GROWTH_YEAR 3	0.001
Industry Group	0.001
Industry Index Name	0.007
ASSET TURNOVER_YEAR 1	0.067

Normally, ROIC should have a positive coefficient as higher ROIC means better profitability and thus increased likelihood of outperformance. However, ROIC_YEAR 2 in this case demonstrated the contrary. In Figure 39, most of the high-ROIC_YEAR 2 instances (ROIC_YEAR 2 > 35) are classified as underperforming; thus, the model captures a negative correlation.

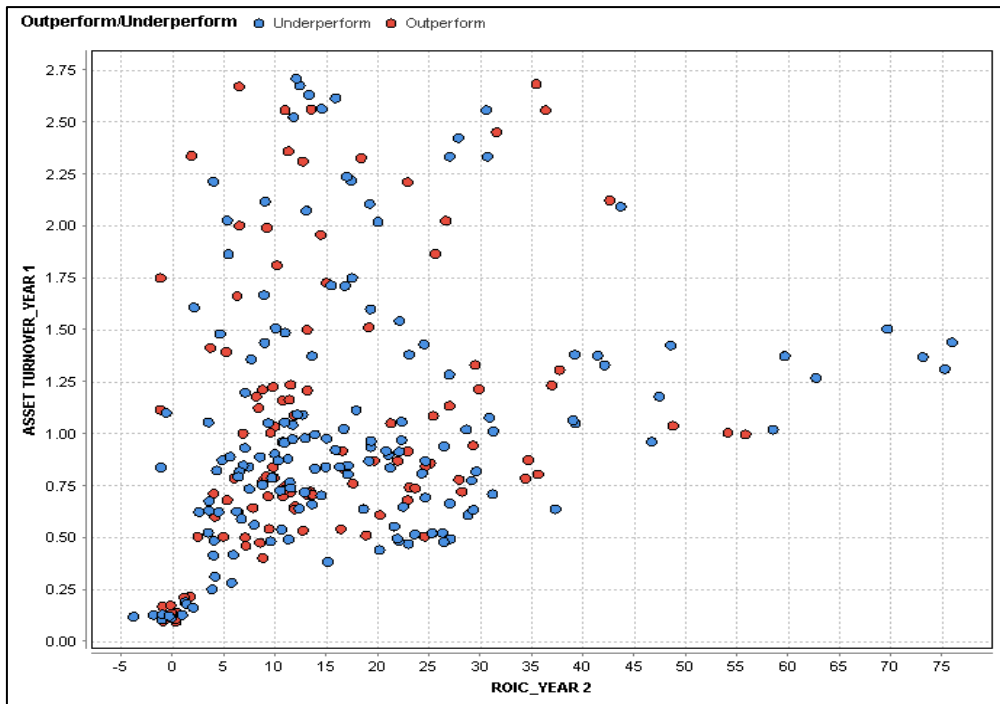


Figure 34: ASSET TURNOVER_YEAR 1 vs. ROIC_YEAR 2 (Relative Performance LR Model)

For categorical independent variables, unique integer coding helps provide the LR classifier with a positive regression coefficient for the industry index, as the categories with high proportion of outperformance instances are assigned with high unique integers than categories with high proportion of underperformance instances. This is illustrated in the industry indexes converted into unique integers in Figure 35.

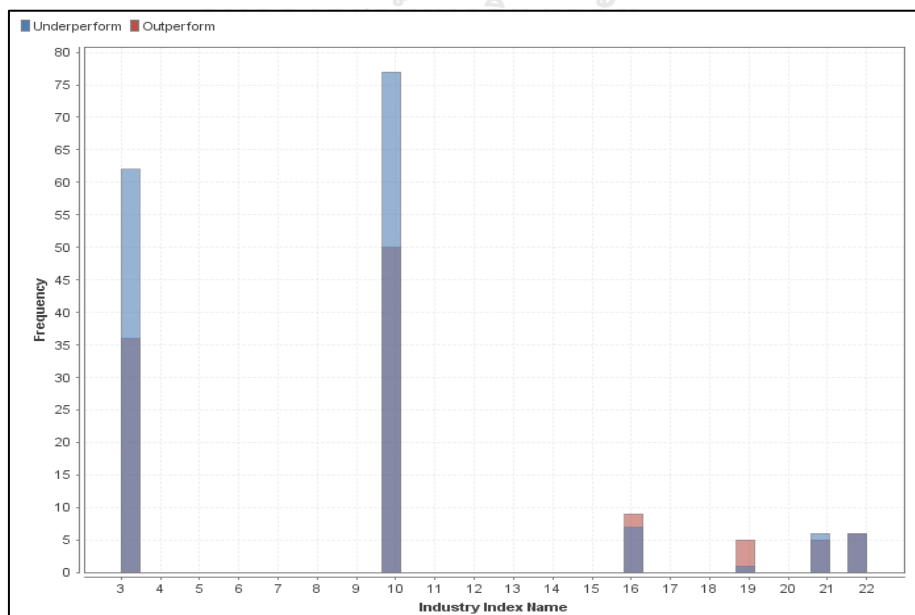


Figure 35: Industry Index Categories as Unique Integers

There is no clear evidence that the instances with high ASSET_TURNOVER_YEAR 1 are classified as outperforming as illustrated in Figure 34. However, if the variable is separated by the Industry Sector, there are three Industry Sectors, and a higher ASSET_TURNOVER_YEAR 1 in the technology sector is associated with outperformance instances, thus giving a positive coefficient in the LR model. A conclusion can be drawn that the range of asset turnover can help investors identify an attractive industry in this sector.

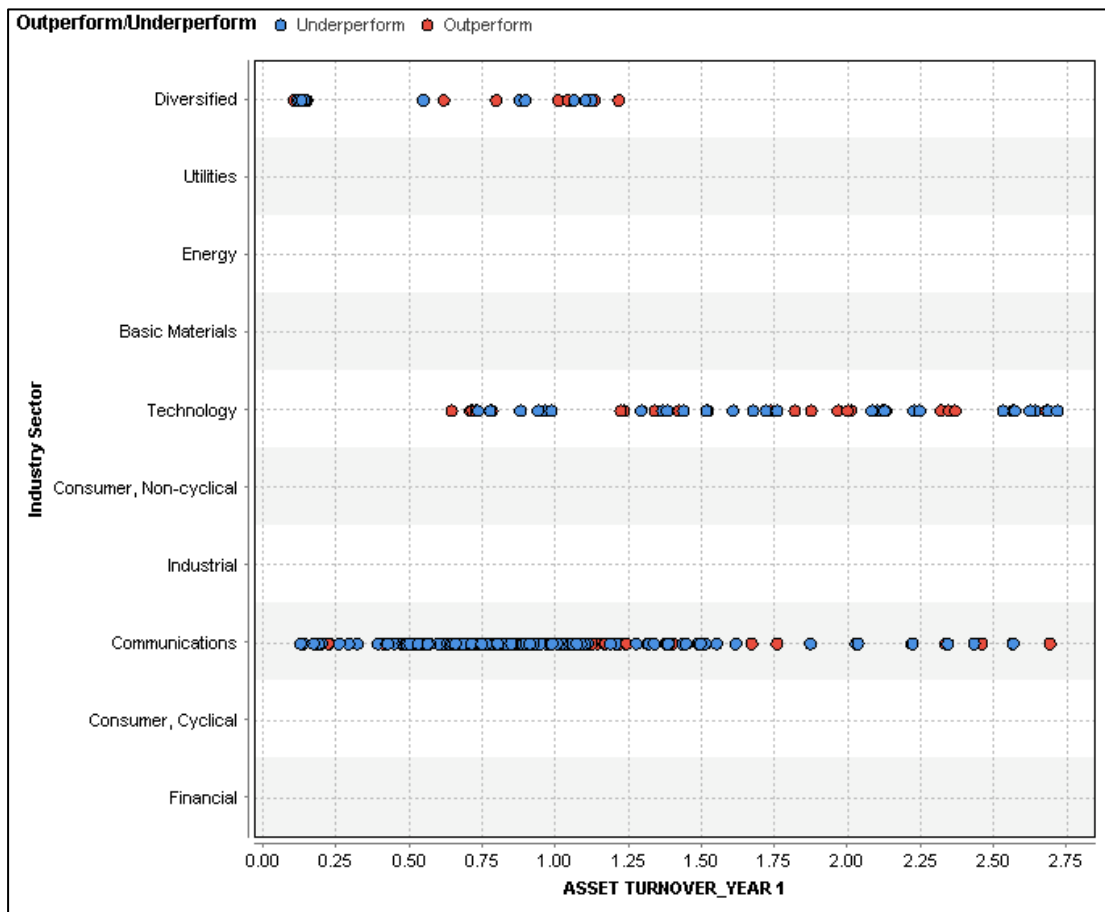


Figure 36: Asset Turnover by Industry Sector

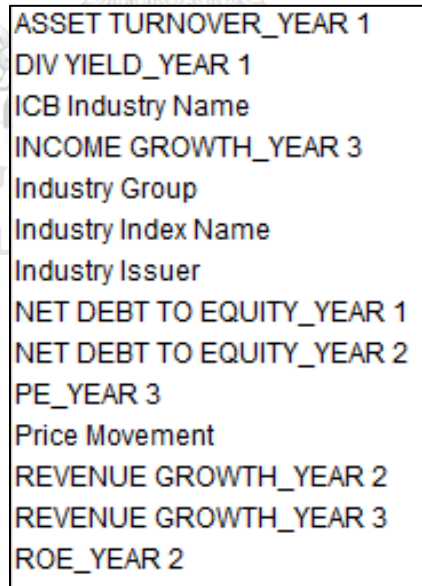
4.5.2 The Price Movement Model

The summary of AUC for price movement models of Communication + Technology + Diversified sectors is presented in Table 28. The results indicated that KNN has the highest AUC with the least uncertainty from optimistic and pessimistic measurements.

Table 28: Price Movement Models Communication + Technology + Diversified Sectors

Classifier	AUC (Neutral)	AUC (Optimistic)	AUC (Pessimistic)
KNN	0.818	0.818	0.818
LR	0.780	0.780	0.780
LDA	0.782	0.782	0.782
DT	0.780	0.864	0.696

For this model, the AUCs of the KNN model are equal in all cases because the concept of weight vote is used for the model. Although the weighted vote option does not guarantee higher neutral AUC, the option eliminates the chance of testing instances to get the same proportion score as the model weighs the contribution of nearest neighbors by assigning a higher weight to neighbors with smaller distance. Thus, the majority vote rule of the KNN model views nearer neighbors as more valuable than farther neighbors.



```

ASSET TURNOVER_YEAR 1
DIV YIELD_YEAR 1
ICB Industry Name
INCOME GROWTH_YEAR 3
Industry Group
Industry Index Name
Industry Issuer
NET DEBT TO EQUITY_YEAR 1
NET DEBT TO EQUITY_YEAR 2
PE_YEAR 3
Price Movement
REVENUE GROWTH_YEAR 2
REVENUE GROWTH_YEAR 3
ROE_YEAR 2

```

Figure 37: Independent Variable for KNN Price Movement Models

Since the categorical independent variables are converted using unique integer coding and there are four categorical independent variables in the KNN model, if a testing instance is in the same industry or subindustry as a training instance, then the distance

value reduces significantly as the difference between the instances goes to zero with the same unique integer. Hence, predicting price movement in these sectors highly depends on what industries or subindustries the testing instances are in.

The advantage of using KNN is that for a testing instance, the distance values are calculated with all training instances. Therefore, the model can capture characteristics that are likely to make a testing instance positive (or negative) price movement regardless of the variable distribution. For example, for testing instances in the ICB Industry, in which the representative unique integer is 2, the instances with normalized INCOME_GROWTH_YEAR 3 between 0.5 and 1.5 are likely to have positive price movement as prediction, and anything outside this range is likely to have negative price movement as prediction (as illustrated in Figure 43). As for the neighbor that is not even in the same ICB Industry, the distance value is too large and the neighbor be thus considered unimportant under the weight vote system.

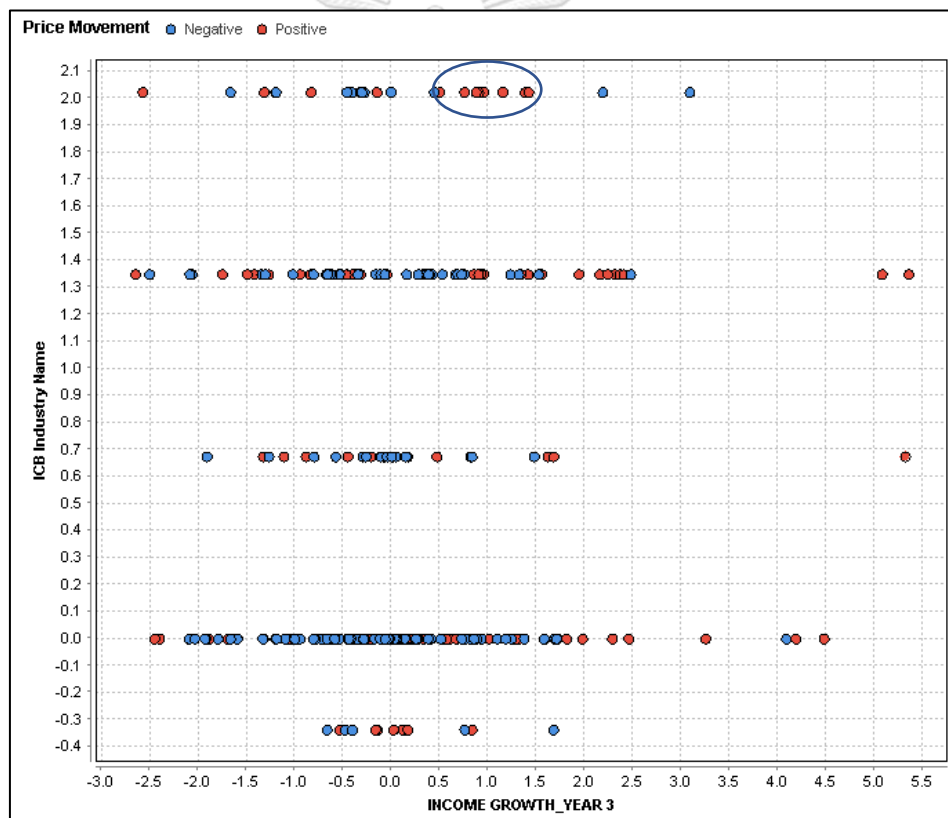


Figure 38: ICB Industry vs. INCOME_GROWTH_YEAR 3

From the example mentioned above, we can deduce that each industry or subindustry will have financial ratios within a certain range in which training instances are likely to have a polarized classification that characterizes that industry or subindustry.

4.6 Basic Materials + Energy + Utilities Sectors

4.6.1 Performance Relative to the SET Index Model

The summary of AUC for stocks' performance relative to the SET index for these three sectors is illustrated in Table 29. The results indicate that LDA has the highest AUC with the least uncertainty from optimistic and pessimistic measurements.

Table 29: Relative Performance Model AUC for Basic Materials + Energy + Utilities Sectors

Classifier	AUC (Neutral)	AUC (Optimistic)	AUC (Pessimistic)
KNN	0.752	0.807	0.697
LR	0.770	0.770	0.770
LDA	0.809	0.809	0.809
DT	0.631	0.900	0.367

By observing the variables and coefficients from the regression equation of LDA, attractive industries can be identified. Clearly, the *GICS Subindustry = Renewable Energy* is a highly profitable industry to look into as the weight of its coefficient contributes greatly to the regression score. On the contrary, the *GICS Subindustry = Oil and Gas Refining and Marketing* is not a profitable industry and has a negative coefficient. These facts are also reflected globally as oil price decreases and renewable energy is gaining popularity as the world moves toward clean and renewable energy. Despite being a small industry with few data instances (see Figure 45), in the Stock Exchange of Thailand, the renewable energy industry has potential and, this model is another evidence as it is based on historical performance.

Another industry to notice is the *GICS Industry = Automobiles and Component* with positive correlation to probability of outperformance. Thailand's automotive industry is the largest in Southeast Asia and is still reported to have a positive outlook.

```

0.033 * BICS Level 1 Sector Name = Utilities
+ 0.064 * GICS Industry = Automobiles & Components
- 0.282 * GICS Industry = Capital Goods
- 0.281 * GICS Sub Industry = Oil & Gas Refining & Marketing
+ 0.665 * GICS Sub Industry = Renewable Electricity
+ 0.147 * ICB Sector = Forestry & Paper
+ 0.222 * Industry Issuer = Metal-Copper
+ 0.203 * Industry Issuer = Diversified Minerals
- 0.003 * PE_YEAR 3
+ 0.082 * PB_YEAR 1
- 0.038 * PB_YEAR 3
- 0.013 * ROIC_YEAR 2
+ 0.009 * ROIC_YEAR 3
- 0.002 * INCOME GROWTH_YEAR 3
- 0.001 * NET DEBT TO EQUITY_YEAR 3
+ 0.005 * PROFIT MARGIN_YEAR 1
- 0.005 * PROFIT MARGIN_YEAR 3
+ 0.468
    
```

Figure 39: LDA Regression Equation for the Relative Performance Model

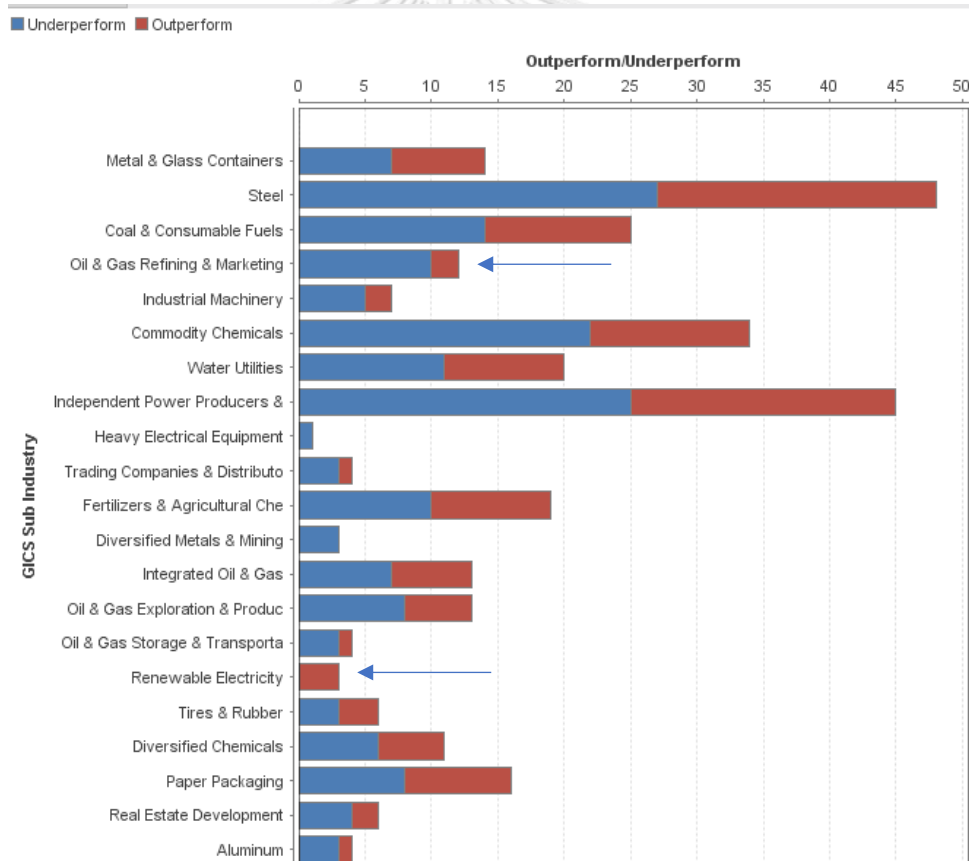


Figure 40: The GICS Subindustry for the Relative Performance Model

Other attractive and unattractive industries and subindustries are also being identified in this model through their coefficients. However, financial ratios also play a significant role in the regression equation.

PE_YEAR 3 and PB_YEAR 3 have negative coefficients; thus, higher values of these ratios lead to lower probability of outperformance. In this case, it is not illogical or counterintuitive to an investor for year-3 ratios to have negative coefficients as the model builds on the assumption that investments will be made shortly after the end of year 3. Thus, a stock's low ratios mean that the stock is cheap, thus giving higher probability of its outperformance. By contrast, PB_YEAR 1 has an opposite relationship. If an investor observes a stock with high PB_YEAR 1 but low PB_YEAR 3, such a stock looks attractive as investors might expect the price to increase to the same level as PB_YEAR 1 when they choose to invest at the end of year 3.

There are many financial ratios such as profit margin, ROIC and income growth that are counterintuitive to an investment strategy. These ratios should give positive outcomes when they are higher, but the coefficient indicates the opposite. An explanation for this would be that these ratios are still not enough to accurately distinguish an investment pattern in the SET market. However, it is also important to understand that the regression coefficients are a product of historical records in the stock market that can be useful for prediction through multivariate analysis.

4.6.2 The Price Movement Model

The summary of AUCs for price movement models for Basic Materials + Energy + Utilities sectors is presented in Table 30. The results indicate that KNN has the highest AUC with the least uncertainty from optimistic and pessimistic measurements.

Table 30: Price Movement Models for Basic Materials + Energy + Utilities Sectors

Classifiers	AUC (Neutral)	AUC (Optimistic)	AUC (Pessimistic)
KNN	0.725	0.787	0.662
LR	0.705	0.705	0.705
LDA	0.783	0.783	0.783
DT	0.750	0.791	0.715

After investigating the independent variables that optimized AUC in the LDA model, an observation can be made that earning a one-year profit from these three sectors highly depends on choosing the right industry or subindustry, as the variables are mostly dummy variables that represent various segments.

Table 31: Independent Variables for the Price Movement LDA Model

Attribute	Coef... ↑
GICS Sub Industry = Oil & Gas Refining & Marketing	-0.186
Industry Issuer = Petrochemicals	-0.166
Industry Index Name = SETPETRO	-0.159
ROE_YEAR 3	-0.007
PE_YEAR 2	-0.005
BICS Level 2 Industry Group = Containers & Packaging	0.001
DIV YIELD_YEAR 2	0.013
Industry Issuer = Chemicals-Plastics	0.024
BICS Level 2 Industry Group = Chemicals	0.085
BICS Level 1 Sector Name = Energy	0.100
BICS Level 1 Sector Name = Utilities	0.126
GICS Sub Industry = Fertilizers & Agricultural Che	0.170
GICS Sub Industry = Metal & Glass Containers	0.367
(Intercept)	0.558

Once again, the dummy variables with negative regression coefficients prove the decline of the petroleum and oil industry. As mentioned in relative performance model section 4.6.1, this decline is the result of a decrease in oil's price. Moreover, the trend of increasing investment in clean energy slowly renders the oil sector obsolete. The negative coefficient for the *GICS Subindustry = Oil and Gas Refining and Marketing*, *Industry Index = SETPETRO*, and *Industry Issuer = Petrochemical* indicated that companies in these segments are not doing well, and a company in this industry has a reduced chance of getting positive price movement.

As for other subindustries, after investigation, we found that all the data instances with the *GICS Subindustry = Metal and Glass Container* were within the *BICS Level 2 Industry Group = Container and Packaging*, which is a larger subindustry. Therefore, the positive correlation of these two dummy variables indicates that they complement each other. As the *BICS Level 2 Industry Group = Container and Packaging* is a larger subindustry that contains the *GICS Subindustry = Metal and Glass Container* instances, the coefficient of the *BICS Level 2 Industry Group = Container and Packaging* is lower, as the proportion of positive instances is lower.

Another relationship to notice within the dummy variables is the *BICS Level 2 Industry Group = Chemicals*, which is a broadly defined industry that contains all data in *Industry Issuer = Chemicals-Plastics* and *GICS Subindustry = Fertilizer and Agriculture Chemicals*. By far, the *GICS Subindustry = Fertilizer and Agriculture Chemicals* contribute in increasing the chance of a positive price movement. This can be explained by the fact that rice production in Thailand represents a significant portion of the Thai economy, and the production of rice relies on chemical fertilizer. The higher

proportion of positive price movement in the *GICS Subindustry = Fertilizer and Agriculture Chemicals* is attributable to increasing demand in chemical fertilizer and thus attracts investors to this subindustry.

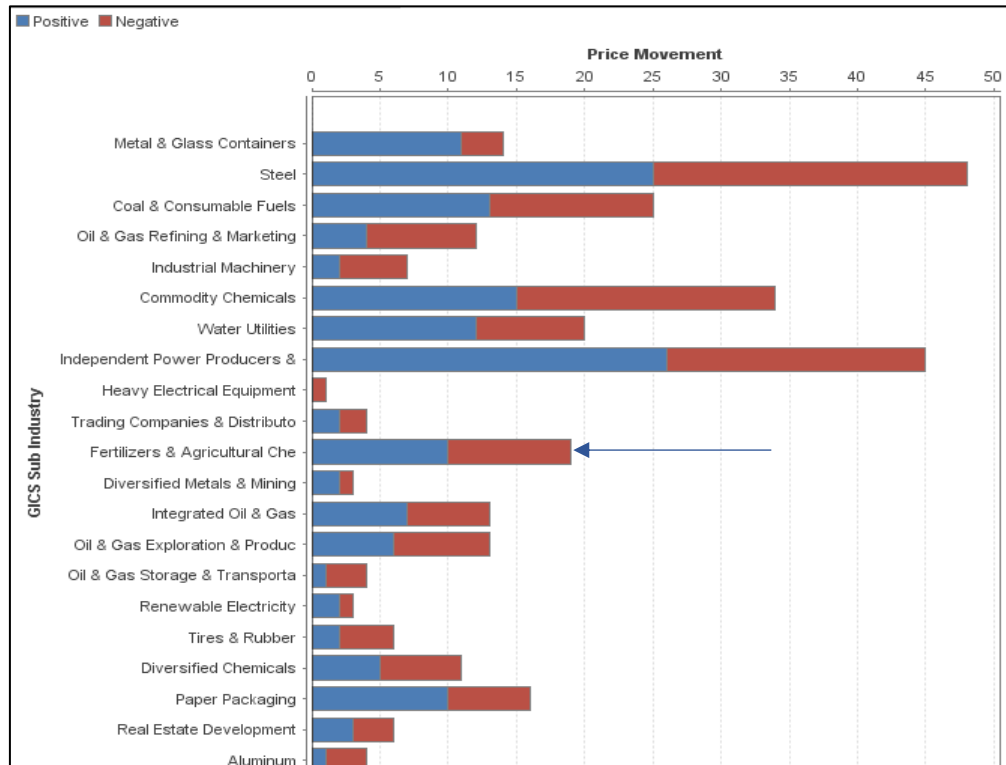


Figure 41: The GICS Subindustry for the Price Movement Model

The coefficients of financial ratios can be explained by Figure 48. Clearly, there is a higher proportion of positive instances when $DIV_YEAR\ 2$ is above 12.5, which causes the model to indicate that higher $DIV_YEAR\ 2$ increases the chance of positive price movement. There are no clear patterns for $ROE_YEAR\ 3$ that are attributable to smaller absolute weights in the regression coefficient.

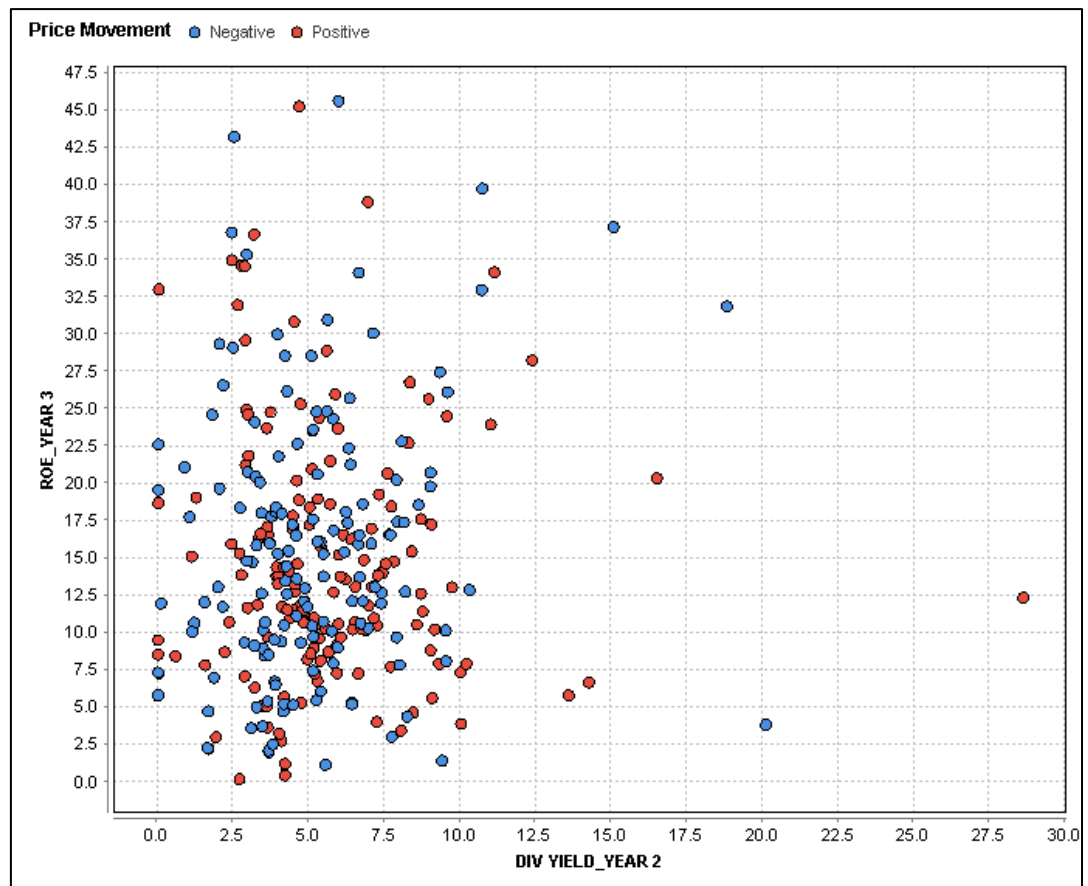


Figure 42: ROE_YEAR 3 vs. DIV_YIELD_YEAR 2

4.7 Remark

After investigating various classifiers, parametric models (LDA and LR) are more frequently robust in predicting stocks' relative performance to market and price movement than non-parametric models do, as they only have one AUC value (no uncertainty from optimistic and pessimistic AUCs). Normally, risk-averse investors likely choose a model with lower uncertainty and rely more on the price movement model, which better guarantees profit, as the relative performance model predicts only the price movement relative to the SET Index but not profit.

Although financial ratios play an important role in company analysis, many regression coefficients indicate relationships that are not common among investors. However, statistical models rely on historical records, and counterintuitive relationships demonstrate that not all investors buy stock based on fundamental ratio analysis alone. There are some aspects that the models fail to cover such as insider trading, i.e. investment based on hearsay, which is not uncommon.

From all the results, we can see the effect of identifying sector/industry/subindustry is critical in predicting a stock's performance. Generally, there are too many industries to look into and investors have to spend time going through the details of industries/subindustries that are attractive to invest in. The models, however, can be a tool that helps inexperienced investors in shortlisting attractive segments with potential and profitable records.

Finally, benchmarking the model's performance measurement (AUC) and what it means is important for investors to understand the appropriateness of these models. To use the model, the user should know at what AUC the model is considered acceptable. One example that can be used for benchmarking AUC is found in study from Deloitte's credit-scoring case [25].

Table 32: AUC Benchmark

Predictive Power	Area Under ROC
Acceptable	>70%
Good	>80%
Very Good	>85%

According to the study [25], financial institutions benchmark the usefulness of a credit-scoring model using Figure 49. Based on the figure, our models' AUC ranges between "acceptable" and "good" in the credit-scoring perspective, which should be enough when applied in stock investment.

Chapter V: Conclusion

The goal of the current study was to build two types of classification models for predicting whether a stock's one-year return in SET will outperform or underperform the SET Index and whether the return will be positive or negative. In order to do this, four different classifiers are applied on training data consisting of 3-year financial ratios and industry/subindustry segments as independent variables as well as binary outputs as dependent variables. The data were separated into six different sectors with each sector having the mentioned four classifiers and two types of a model, resulting in 48 models in total. For each sector and model type, the top-performing classifiers with the highest AUC and least uncertainty were chosen for prediction application. The results of top-performing classifiers for each sector and model type are summarized in Table 33.

Table 33: Summary of Top-Performing Models

Sector	Classifier for Performance Relative to the SET Index	AUC	Classifier for One-year Price Movement	AUC
Finance	LR	0.744	LR	0.712
Consumer Cyclical	LDA	0.773	LDA	0.693
Consumer Non-Cyclical	DT	0.769	LDA	0.786
Industrial	LDA	0.789	LDA	0.723
Communication + Technology + Diversified	LR	0.823	KNN	0.818
Basic Materials + Energy + Utilities	LDA	0.809	LDA	0.783

The purpose of the models in Table 33 is to serve as a tool for shortlisting attractive stocks from 582 companies in SET for further research and investment. After benchmarking the AUCs, the usefulness of these models can be rated as “Acceptable” to “Good” using Deloitte’s credit-scoring standard [25].

An important observation in Table 33 is that the LDA classifier is the best model in many cases. From the literature review, the most popular model many researchers use is the Artificial Neural Network (ANN). Therefore, the current study proves that using a classifier such as LDA is also acceptable. One of the reasons for why LDA performs better than other classifiers is that the LDA model can handle dummy variables. When dummy variable coding are used for categorical independent variables, the number of variables increases significantly as dummy variables are created to represent each category in all categorical variables. Thus, the AUC optimization with variable

selection (forward selection/backward elimination) has more degrees of freedom or more combinations of variables to choose from to optimize AUC. Another reason is that the violation of normality and homoscedasticity assumptions does not have a significant effect on AUC or prediction power, as Mircea et al. mentioned that the violations of assumptions are not fatal [20].

A summary of important observations from investigating independent variables that optimized AUC is presented in

Table 34: Summary of Important Observations by Sector

Sector	One-year Relative Performance Model	One-year Price Movement Model
Finance	Real estate is an attractive industry in the finance sector with a better historical record of outperformance than other industries.	Higher net debt is correlated with a higher chance of positive return, as it is not uncommon in the real estate industry to have high debt to support large upfront investment.
Consumer Cyclical	Identifying a company's industry and subindustry plays a greater role in predicting the performance of a stock relative to the SET Index than financial ratios do as the sector contain too many characteristics to be identified by financial ratios.	With lower AUC, this sector has a large spread of industries within one sector, thus making price movement erratic and the prediction model less robust than the relative performance model.
Consumer Non-Cyclical		PB ratios play an important role in investment decision in this sector. Attractive companies should have a good record of high PB.
Industrial	In this sector, small industries are more likely to outperform the market as a large competitive industry is the key driver of the SET index and is unlikely to outperform the market.	The automotive, hardware as well as iron and steel industries in this sector are attractive and worth further investigation for future investment.
Communication + Technology + Diversified	The range of asset turnover can help identify the attractive industry in this sector, which is technology industry.	Predicting price movement in these sectors highly depends on the type of industry or subindustry.

Table 34: Summary of Important Observations by Sector (Continue)

Basic Materials + Energy + Utilities	Oil and gas and petrochemical industries are no longer attractive due to lower oil price. This observation can be seen in both models.
--	--

In practice, these models can be applied with the latest input data, i.e. 3-year financial observations from 2013 to 2015, to predict the stock's performance and price movement between 2016 and 2017. We only need to pull out financial data of 582 stocks in the SET from Bloomberg Terminal, together with 10 industry/subindustry classifications. As these data can be further separated to sectors, stock preselection could be done so that we can focus only those in the classes of “outperform” and “positive”.

As for limitation of these models, the application of these models still required some degree of data mining skills. Retraining or updating these models as new data become available will be complicated as the data are separated into six sectors and optimization of classifiers has to be performed all over again. Further research direction should be looking into a more generalized model that does not separate data into six sectors and require training and optimizing 48 models. Such a generalized model will reduce runtime, data handling and complication in applying and updating models. Although the generalized model is likely to have lower AUC, the AUC can be improved by various methods such as using the popular Artificial Neural Network (ANN) as mentioned in the literature review. Additionally, incorporating more independent variables such as macroeconomic or news analysis using text mining has been found to improve model accuracy.

Another aspect to improve is the AUC itself. Some studies used an optimization algorithm such particle swarm optimization (PSO) and saw an improvement of accuracy. Applying ensemble learning methods such as stacking and boosting should also be studied as these methods could improve accuracy.

REFERENCES

- [1] P. Sareewiwatthana, "Value investing in Thailand: the test of basic screening rules," *International Review of Business Research Papers*, vol. 7, no. 4, pp. 1-13, 2011.
- [2] U-Wen Kok, Jason Ribando, and R. Sloan, "Facts about Formulaic Value Investing," *Financial Analysts Journal*, vol. 73, no. 2, pp. 81-98, 2017.
- [3] Manminder Singh Saluja and Y. Shaikh, "Decoding Investment Pattern of FIIs and DIIs in Indian Stock Market using," *International Journal of Advanced Research in Computer Science*, vol. 8, no. 3, pp. 911-916, 2017.
- [4] G. Alfonso Perez, "Value Investing in the Stock Market of Thailand," *International Journal of Financial Studies*, vol. 5, no. 4, 2017.
- [5] Louis K.C. Chan, Yasushi Hamao, and J. Lakonishok, "Fundamentals and stock returns in Japan,"
- [6] N. C. P. Edirisinghe and X. Zhang, "Generalized DEA model of fundamental analysis and its application to portfolio optimization," *Journal of Banking & Finance*, vol. 31, no. 11, pp. 3311-3335, 2007.
- [7] H. Grigoryan, "Stock Market Prediction using Artificial Neural Networks. Case Study of TALIT, Nasdaq OMX Baltic Stock," *Database Systems Journal*, vol. 6, no. 2, pp. 14-23, 2015.
- [8] W. Banchuenvijit, "Financial Ratios and Stock Prices: Evidence from the Agriculture Firms Listed on the Stock Exchange of Thailand," *UTCC International Journal of Business & Economics*, vol. 8, no. 2, pp. 23-29, 2016.
- [9] Hiral R. Patel, Satyen M. Parikh, and D. N. Darji, "Prediction model for stock market using news based different Classification, Regression and Statistical Techniques: (PMSMN)," presented at the 2016 International Conference on ICT in Business Industry & Government (ICTBIG) Indore, India, 2016.
- [10] H. Alostad and H. Davulcu, "Directional prediction of stock prices using breaking news on Twitter," *Web Intelligence*, vol. 15, no. 1, pp. 1-17, 2017.
- [11] Avijan Dutta, Gautam Bandopadhyay, and S. Sengupta, "Prediction of Stock Performance in the Indian Stock Market Using Logistic Regression," *International Journal of Business and Information*, vol. 7, no. 1, pp. 105-136, 2012.
- [12] C.-F. Tsai and S.-P. Wang, "Stock Price Forecasting by Hybrid Machine Learning Techniques," in *International MultiConference of Engineers and Computer Scientists*, Hong Kong, 2009, vol. 1.
- [13] P. Tüfekci, "Classification-based prediction models for stock price index movement," *Intelligent Data Analysis*, vol. 20, no. 2, pp. 357-376, 2016.
- [14] F. S. Shie, M.-Y. Chen, and Y.-S. Liu, "Prediction of corporate financial distress: an application of the America banking industry," *Neural Computing and Applications*, vol. 21, no. 7, pp. 1687-1696, 2012.
- [15] C.-H. Cheng and S.-H. Wang, "A quarterly time-series classifier based on a reduced-dimension generated rules method for identifying financial distress," *Quantitative Finance*, vol. 15, no. 12, pp. 1979-1994, 2015.
- [16] Janthorn Sinthupundaja, Navee Chiadamrong, and N. Suppakitjarak, "Financial Prediction Models from Internal and External Firm Factors Based

- on Companies Listed on the Stock Exchange of Thailand," *Suranaree Journal of Science & Technology*, vol. 24, no. 1, pp. 83-98, 2017.
- [17] M. Santini. (2015). *Lecture 4 Decision Trees (2): Entropy, Information Gain, Gain Ratio*. Available: <https://www.slideshare.net/marinasantini1/lecture-4-decision-trees-2-entropy-information-gain-gain-ratio-55241087>
- [18] G. Ritschard, "CHAID and Earlier Supervised Tree Methods," ed. University of Geneva, 2010.
- [19] (2017). *CART Algorithm for Decision Tree*. Available: <http://dni-institute.in/blogs/cart-algorithm-for-decision-tree/>
- [20] Gabriela Mircea, Marilen Pirtea, Mihaela Neamtu, and S. Bazavan, "Discriminant analysis in a credit scoring model," presented at the Recent Advances in Applied and Biomedical Informatics and Computational Engineering in Systems Applications, Florence, 2011.
- [21] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861-874, 2006.
- [22] David W. Hosmer and S. Lemeshow, *Applied Logistic Regression*, 2 ed. John Wiley & Sons, Inc., 2000.
- [23] Michael LeBlanc and S. Fitzgerald, "Logistic Regression for School Psychologists," *School Psychology Quarterly*, vol. 15, no. 3, pp. 344-358, 2000.
- [24] Shuchita Upadhyaya and K. Singh, "Nearest Neighbour Based Outlier Detection Techniques," *International Journal of Computer Trends and Technology*, vol. 3, no. 2, pp. 299-303, 2012.
- [25] Nikos Skantzos and N. Castelein, "Credit Scoring - Case study in data analytic," 2016.

APPENDIX



จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

VITA

Athit Phongmekin was born in 1991. He received Bachelor of Science degree from Purdue University in the field of chemical engineering. He received both of my master degree from a joint dual degree program from Sasin Graduate Institute of Business Administration and Chulalongkorn University. He earned Master of Engineering in industrial engineering and Master of Business Administration.

His previous undergraduate research experience in chemical engineering had to do with developing mathematical model for simulating white blood cells dynamic during chemotherapy. His primary role was to assist my advisor in data collection and model testing.

Before entering master degree program, he was employed at Professional Environmental Management (PEM) Company Limited where he worked as sales engineering. Due to the nature of a new and small business, he was responsible for many roles including from finding target customer, make sale pitch and designing waste water recycling system. He stayed in PEM for 2 years before joining master degree program.

His interest in this thesis topic primary root from my interest in finance and further reinforce by my study in MBA degree at Sasin.