Customer Churn Prediction for a Software-as-a-Service Inventory Management Company

Mr. Phongsatorn Amornvetchayakul

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Engineering in Industrial Engineering
Department of Industrial Engineering
FACULTY OF ENGINEERING
Chulalongkorn University
Academic Year 2019

การพยากรณ์การหยุดใช้บริการของลูกค้าสำหรับบริษัทซอฟต์แวร์ให้บริการจัดการคลังสินค้า

นายพงศธร อมรเวชยกุล

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมอุตสาหการ ภาควิชาวิศวกรรมอุตสาหการ
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2562

Thesis Title	Customer Churn Prediction for a Software-as-a-Service Inventory Management Company
By	Mr. Phongsatorn Amornvetchayakul
Field of Study	Industrial Engineering
Thesis Advisor	Associate Professor NARAGAIN PHUMCHUSRI, Ph.D.

Accepted by the FACULTY OF ENGINEERING, Chulalongkorn University in Partial Fulfillment of the Requirement for the Master of Engineering

..... Dean of the FACULTY OF ENGINEERING
(Professor SUPOT TEACHAVORASINSKUN, D.Eng.)

THESIS COMMITTEE

..... Chairman
(Associate Professor WIPAWEE THARMMAPHORNPHILAS, Ph.D.)

..... Thesis Advisor
(Associate Professor NARAGAIN PHUMCHUSRI, Ph.D.)

..... Examiner
(Nantachai Kantanantha, Ph.D.)

..... External Examiner
(Assistant Professor Manop Reodecha, Ph.D.)

พงศธร อมรเวชยกุล : การพยากรณ์การหยุดใช้บริการของลูกค้าสำหรับบริษัทซอฟต์แวร์ให้บริการจัดการคลังสินค้า. (Customer Churn Prediction for a Software-as-a-Service Inventory Management Company) อ.ที่ปรึกษาหลัก : รศ. ดร.นระเกณฑ์ พุ่มชูศรี

วิทยานิพนธ์นี้แสดงถึงการพยากรณ์การหยุดใช้บริการของลูกค้าสำหรับบริษัทซอฟต์แวร์ที่ให้บริการจัดการคลังสินค้าในประเทศไทย โดยกลุ่มธุรกิจบริการซอฟต์แวร์ในลักษณะ Software-as-a-Service เป็นหนึ่งในธุรกิจที่มีการเติบโตอย่างรวดเร็ว และมีมูลค่าทางการตลาดสูง เป็นผลสืบเนื่องมาจากการเกิดใหม่ของธุรกิจออนไลน์ในปัจจุบัน โดยในธุรกิจนี้ การหยุดใช้บริการของลูกค้านั้นถือเป็นหนึ่งในปัจจัยสำคัญที่มีผลกระทบกับธุรกิจโดยตรง ดังนั้นการศึกษานี้จะมุ่งเน้นไปที่การหาแบบจำลองการพยากรณ์การหยุดใช้บริการของลูกค้าสำหรับบริษัทซอฟต์แวร์ที่ให้บริการจัดการคลังสินค้าในประเทศไทยที่กำลังประสบปัญหากับอัตราการหยุดใช้บริการของลูกค้าที่อยู่ในระดับสูง ในวิทยานิพนธ์นี้นำเสนอรูปแบบการพยากรณ์ที่หลากหลายโดยประยุกต์ใช้แบบจำลองการเรียนรู้ของเครื่อง หรือ Machine Learning ทั้งหมด 4 แบบได้แก่ Logistic Regression, Support Vector Machine, Decision Tree และ Random Forest จากผลการศึกษาพบว่าแบบจำลอง Decision Tree ที่ได้รับการปรับค่าตัวแปรที่เกี่ยวข้องต่าง ๆ อย่างเหมาะสมนั้นสามารถทำได้ดีกว่าแบบจำลองอื่น ๆ ในกรณีที่ใช้การประเมินผลด้วย Recall เป็นหลัก โดยสามารถทำผลการทดสอบแบบ Recall ได้ถึง 94.4% และการทดสอบแบบ F1-score ได้ 88.2% นอกจากนี้แบบจำลองยังสามารถบ่งชี้ถึงตัวแปรสำคัญที่มีผลต่อการหยุดใช้บริการของลูกค้าซึ่งเป็นข้อมูลเชิงลึกที่เป็นประโยชน์ต่อบริษัทที่เป็นกรณีศึกษา ด้วยเหตุนี้วิทยานิพนธ์นี้สามารถช่วยให้บริษัทที่เป็นกรณีศึกษานั้นสามารถบ่งชี้ลูกค้าที่กำลังหยุดใช้บริการได้อย่างถูกต้อง และช่วยให้บริษัทสามารถวางแผนในด้านการตลาดและการบริหารงานได้อย่างมีประสิทธิภาพมากยิ่งขึ้น

สาขาวิชา	วิศวกรรมอุตสาหการ	ลายมือชื่อนิสิต .....
ปีการศึกษา	2562	ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

6170218821 : MAJOR INDUSTRIAL ENGINEERING
KEYWORD: churn prediction, Machine learning, Software-as-a-Service (SaaS)

Phongsatorn Amornvetchayakul : Customer Churn Prediction for a Software-as-a-Service Inventory Management Company. Advisor: Assoc. Prof. NARAGAIN PHUMCHUSRI, Ph.D.

This thesis proposes customer churn prediction model for a Software-as-a-Service inventory management company in Thailand. Software-as-a-Service is the fast growing and high market values industry as a new emerging online business. Customer churn is a critical measure for this business. Thus, this thesis focuses on seeking a customer churn prediction model specified for a Software-as-a-Service inventory management company in Thailand which is facing a high churn rate issue. This thesis executes the prediction models with four machine learning algorithms: logistic regression, support vector machine, decision tree and random forest. The results show that the optimized decision tree model is capable to outperform other classification models toward recall scorer with validated testing scores of 94.4% of recall and 88.2% of F1-score. Moreover, feature importance scores can highlight useful insights of case-study that business metrics are significantly related to churn behavior. As a result, this paper is beneficial to the case-study company to help indicate real churn customer correctly and enhance the effectiveness in making executive decision and marketing campaign.

Field of Study:	Industrial Engineering	Student's Signature .....
Academic Year:	2019	Advisor's Signature .....

ACKNOWLEDGEMENTS

Phongsatorn Amornvetchayakul

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

Chapter 1 Introduction

1.1. Cloud Computing and Software-as-a-Service

Software-as-a-Service, or SaaS, is a model where software is licensed over the internet on a subscription basis by a third-party provider. Software-as-a-Service (SaaS) is one of three major types of cloud computing, together with Infrastructure-as-a-Service (IaaS) and Platform-as-a-Service (PaaS).

Cloud Computing Services

The core business of those different services, aforementioned, are when typical hardware and data centers – such as storage and servers – are on the cloud. When compared to on-premises solutions, cloud computing provides plenty of major benefits, thus quickly becoming highly preferential for businesses today. Scalability, cost effectiveness, immediate availability, performance, and security are just some benefits of cloud computing services, i.e., Software-as-a-Service (SaaS), Infrastructure-as-a-Service (IaaS), and Platform-as-a-Service (PaaS).

For better understanding of the different services, the cloud computing services can be analogized to transportation (refer to Figure 1) On-premise infrastructure is similar to owning a car – the user has its full responsibility and upgrades often mean buying a whole new vehicle. Infrastructure-as-a-Service (IaaS) is like leasing a car – the user has full control on which car they want to lease, where they want to drive it to, and upgrades as easier done by leasing a different car. Platform-as-a-Service (PaaS) is similar to taking a taxi – users don't "drive" the taxi, they simply dictate the direction. Software-as-a-Service (SaaS) is like taking a bus – not much is in the user's control, neither the vehicle nor the direction.

Software-as-a-Service business model

For traditional, on-premise services, the upfront sale held the highest value towards a business. All sales energy and resources were focused on the making the deal and

acquiring the customer. For as-a-service models, efforts need to be directed not only towards acquiring the new customer, but also to retaining customer usage is high and their retention is secure. For recurring revenue models such as these, operational and service costs are realized over time, right alongside the recurring revenue. This model, thus, entails a decrease in top-line revenue and bottom-line profitability amid the transaction.

Software-as-a-Service (SaaS)'s leading benefit as a business model is revenue recurrence. As advancing towards a world of as-a-Service cloud computing, revenue opportunities switch from one-sale transactions to subscription based, recurring revenues.

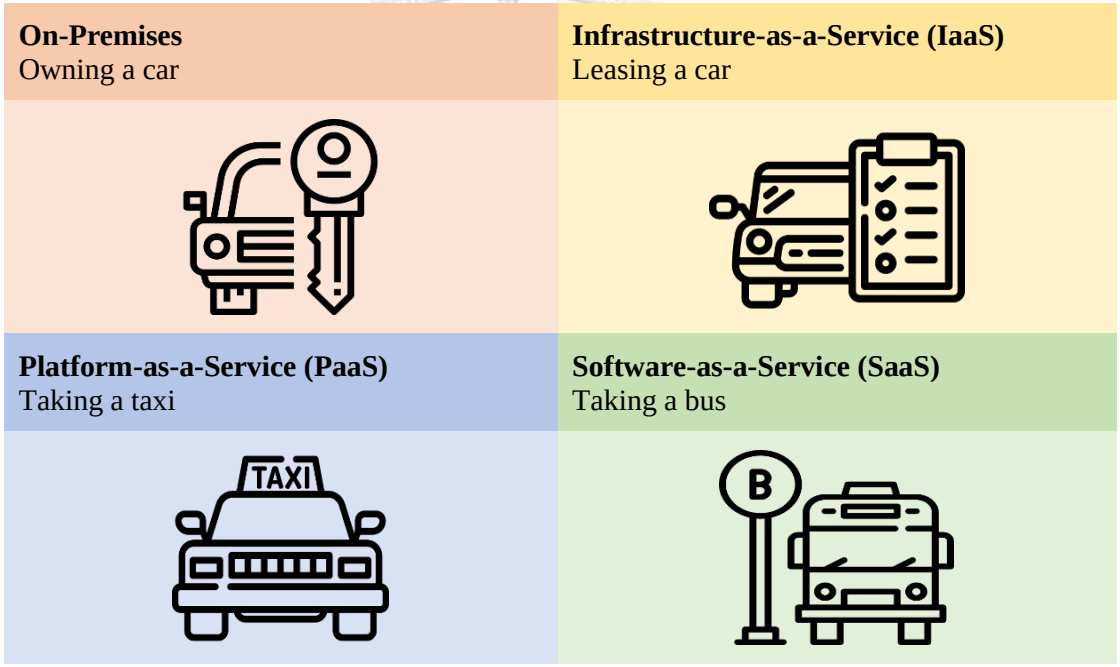


Figure 1 Types of cloud services in analogies[1]

For such Software-as-a-Service (SaaS) models, revenue is generally recognized over a prolonged time frame. Thus, the higher the customer retention, the longer the subscription model, and the more profitable the customer becomes for the business. If, for whatever reason, the customer decides to leave the business prematurely, the business makes realizes the cost of bringing in a new, replacement customer (also known as acquisition cost). For such cases, the acquisition cost is generally higher than the profitability from the initial customer – leaving the business at a loss. Hence,

for as-a-service businesses, customer retention and low customer churn is crucial towards profitability.

Due to its flexible and scalable nature, the use of cloud computing is exponentially rising, especially in this digital transformation landscape. The use has accelerated the growth of clouding service. According to IDC's Worldwide Semiannual Public Cloud Services Spending Guide, published in second half of 2018[2], cloud service is growing more than 4.5 times the rate of IT growth as a whole with a five-year compound annual growth rate (CAGR) of 22.3%. Software-as-a-Service (SaaS), especially, is the largest market segment and is expected to maintain this top position. Due to the scalability of subscription-based software, the Software-as-a-Service (SaaS) sector is forecasted to grow to $116 Billion by 2020 and expected to reach $150 billion in 2022 in terms of public cloud service revenue (refer to Table 1).

Table 1 Worldwide Public Cloud Service Revenue Forecast (Billions of U.S. Dollars)[3]

Types of Cloud Service	**2018**	**2019**	**2020**	**2021**	**2022**
Cloud Business Process as a Services (BPaaS)	41.7	43.7	46.9	50.2	53.8
Cloud Application Infrastructure Services (PaaS)	26.4	32.2	39.7	48.3	58
Cloud Application Services (SaaS)	85.7	99.5	116	133	151.1
Cloud Management and Security Services	10.5	12	13.8	15.7	17.6
Cloud System Infrastructure Services (IaaS)	32.4	40.3	50	61.3	74.1
Total Market	196.7	227.8	266.4	308.5	354.6

Note: Totals may not add up due to rounding.

As a result of that, the Software-as-a-Service (SaaS) will remain the dominant cloud computing type in next few years, capturing over one-thirds of all public cloud service revenue.

Professional services, discrete manufacturing, and banking, together, account for more than one-third of all cloud service spending. According to International Data Corporation (IDC)'s research press release[4], professional services industry is expected to expand in public cloud service, with a 5-year CAGR of 25.6% which is the fastest growing industry in public cloud services (refer to Figure 2).

In addition, Software-as-a-Service (SaaS) spending, containing applications and system infrastructure software (SIS), is seen to be governed by application purchases. Customer Relationship Management (CRM) and Enterprise Resource Management (ERM) are the prominent Software-as-a-Service (SaaS) applications, followed by content workflow, and collaborative applications. It notes that Inventory Management is considered as a part of Enterprise Resource Management (ERM).

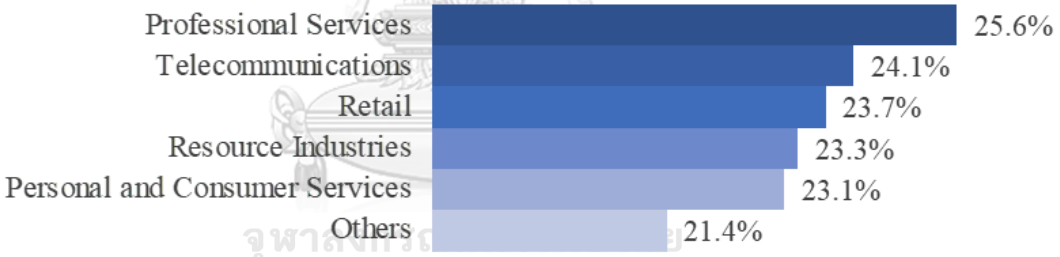


Figure 2. Top industry based on 5-year CAGR (2018-2023)[4]

Considering Software-as-a-Service (SaaS) market in the application of Enterprise Resource Planning (ERP) based on enterprise resource planning trends survey in the United Kingdom by the Accenture SAP Business Group in 2018[5], it found that even though SMEs (Small and medium-sized enterprises) or subsidiaries of larger companies are the main adopted users of Software-as-a-Service Enterprise Resource Planning software market, the larger enterprise tends to significantly shift the deployment model from On-Premises to Software-as-a-Service (SaaS) urging for large-scale development and expansion.

1.2.Inventory management

Inventory management software is used for tracking inventory levels, purchase orders, sales, delivery, and much more. The ease-of-use and high visibility of inventory tracking tools are just some of the factors that are driving the use of such software worldwide. A huge variety of inventory management, tracking, control, and visibility tools are available on a multitude of platforms, such as iOS and Android, just to name a few. Such a software helps users view, control, and track inventory level, no matter the size of the plant or the location of the user. Large manufacturing plants are usually interconnected through a wireless network and can be operated from any remote location.

As result of the benefits of inventory tracking solution and the increment of connected device adoption rate along the global upcoming trend, the demand in inventory management has been increased extensively. However, there is a large initial investment towards the installation of such an intertwined system. The cost of an inventory management software typically depends on the features it offers, its capabilities, and the number of users it accommodates. With Software-as-a-Service (SaaS) business model, the inventory management software can reduce the friction of this issue. Additionally, data security is a major concern when it comes to such systems. With such a software carrying numerous and integral information on customers, products, internal processes, and more, the risk of exposure is high and critical.

The inventory management software market can be divided based on the different 5 main categories: product segment, application, deployment model, organization size, and vertical (Industry) (refer to Figure 3.). By product segment, it comprises a manually managed inventory system, an advanced radio frequency system, and a barcode scanning system. The application category of inventory management software market includes order management, asset tracking, service management, product differentiation, and inventory optimization. By deployment model, the market comprises on-premises and cloud-based on which this research focuses. By organization size, it comprises SMEs (Small and medium-sized enterprises) and large

enterprises. And by vertical category or by industry, the market includes retail, manufacturing, healthcare, automotive, oil and gas, and others. Since inventory management software provides several benefits in terms of handling stock effectively and efficiently, tracking, tracing, and analyzing the value chain from upstream to downstream, this type of software has been used on a large scale in various industries.

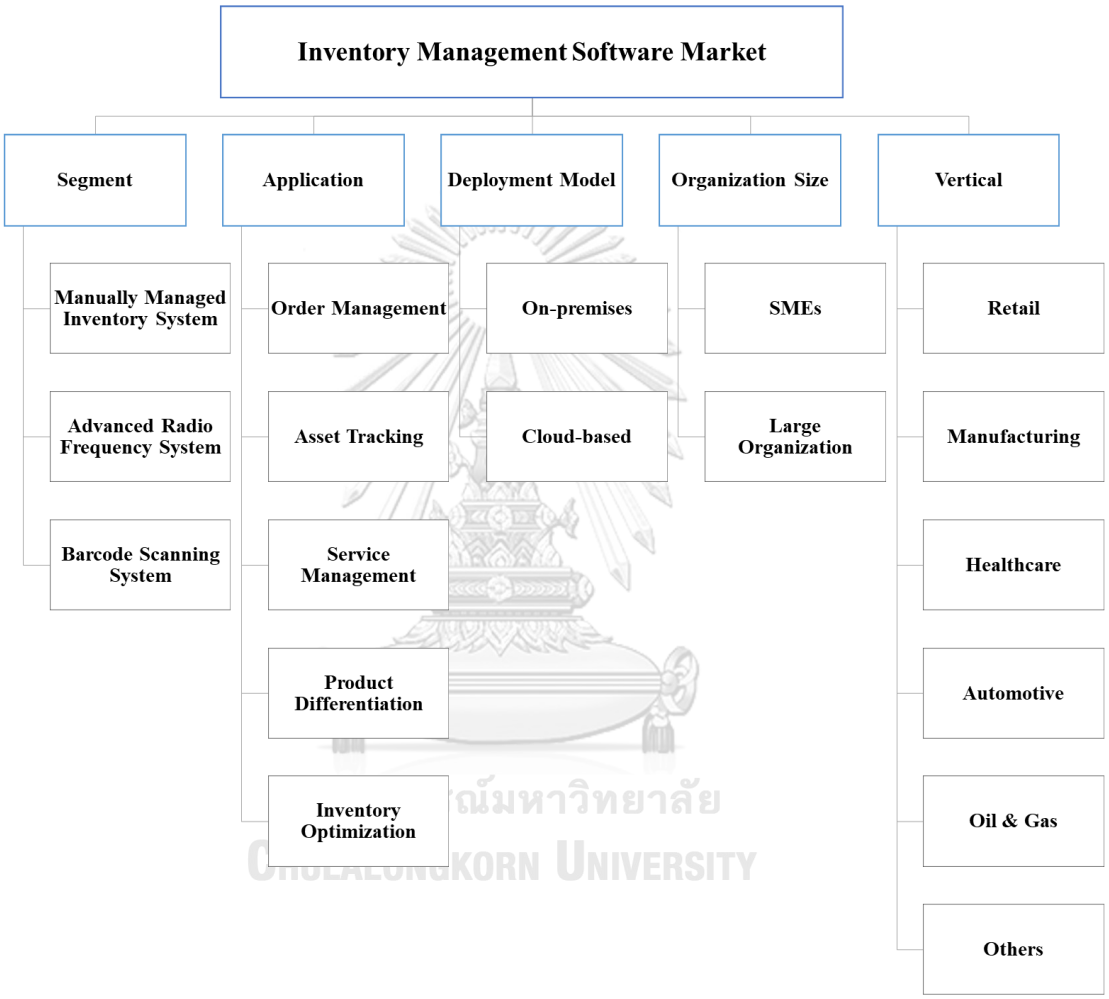


Figure 3. Inventory management software market segmentation

According to the report from GlobeNewswire[6] , the global market of inventory management software is expected to expand with a CAGR of over 6.1% between 2019 and 2025. For the sake of that, the global market of inventory management software is project to be approximately USD 3.2 billion in 2025 regarding the previous market size in 2018 was account for USD 2.1 billion.

Considering cloud base deployment model, MarketWatch released that cloud-base deployment model of inventory management software assist SMEs enhancing the business competitiveness in terms of scalability, versatility, and advanced security[7]. The cloud-base model, especially Software-as-a-Service (SaaS) model which is the idea of pay-as-you-go, can provide the inventory management software in affordable prices which reduce SMEs' financial burdens in terms of initial investment or capital expenditure on infrastructure in early stage; therefore, the cloud deployment sector has been adopted extensively. And its market share is expected to be obtained over 35% of the global inventory management software market by 2024.

Trends in Asia

In Asia Pacific, the emerging Asian economies have propelled the trend of growing in an inventory management software market. The concern in productivity improvement by using the technological tools or any other automation system is a part of the growing factor of this market. Furthermore, the growth of e-commerce and rising number of SMEs in this region have also driven the expansion of the inventory management software market along the trends, especially in retail industry. Focusing on South East Asia (SEA), a market study on the global inventory management software market[8], Future Market Insights (FMI) found that South East Asia become a new target market of the inventory management software providers which particularly expected to penetrate in the retail industry.

1.3.Customer churn rate and its effect

Customer churn rate, sometimes known as attrition rate or cancellation rate, is the percent of customers who have cancelled their subscription or use of a product or service. Online business considers their customer "churned" when a predefined amount of time as gone by since the customer's last contact with the site.

Generally, churn rate is calculated by dividing the number of customers lost during a period of time by the total number of customers at the start of the time period. The

total cost related to customer churn comprises of lost revenue from the absent customer, as well as the replacement of the customer with new ones.

It is established that there are three main strategies of revenue growth dependent on customers: customer expansion, product or service price inflation, and reduction of customer churn. Price inflation is a tough strategy, often almost impossible to accomplish without disappointing existing customers. Customer expansion is an expensive tactic, with research stating that “acquiring a new customer is anywhere from five to 25 times more expensive than retaining an existing one”[9].Customer retention, on the other hand, is not quite as expensive. The strategy saves a lot of time and resources to find and acquire new customers. To further support this point, research done by Frederick Reichheld of Bain & Company states that increasing customer retention rates by 5% increases profits by 25% to 95%[10].

The Importance of Predicting Customer Churn

The accurate prediction of customer churn is one of the most valuable information businesses are currently seeking. The ability to identify customers who are high-risk or very likely to terminate their relationship with a service or discontinue their use of a product is a key in strategic decision making.

As a result of better strategic moving to handle the risky customer with better predicting customer churn, the customer retention rate will be improved. It means that the company can attain more repeat customers which are profitable. Regarding the amount of purchases, when a customer once purchased, that customer tends to repurchase at the same store with only 27% chance. While if the customer return to buy for the second times, the opportunities for the third purchase will be increased up to 45%[11]. This can be indicated that the repeat customers are more likely to retain relationship with a particular company in long term. According to a research in retail sector by Bain & Company[12], the repeat customers further spend 23 percent more per orders after purchasing with the same company for 30 months **(**refer to Figure 4.**)** that the power of repeat customer significantly impacts on the spending growth.

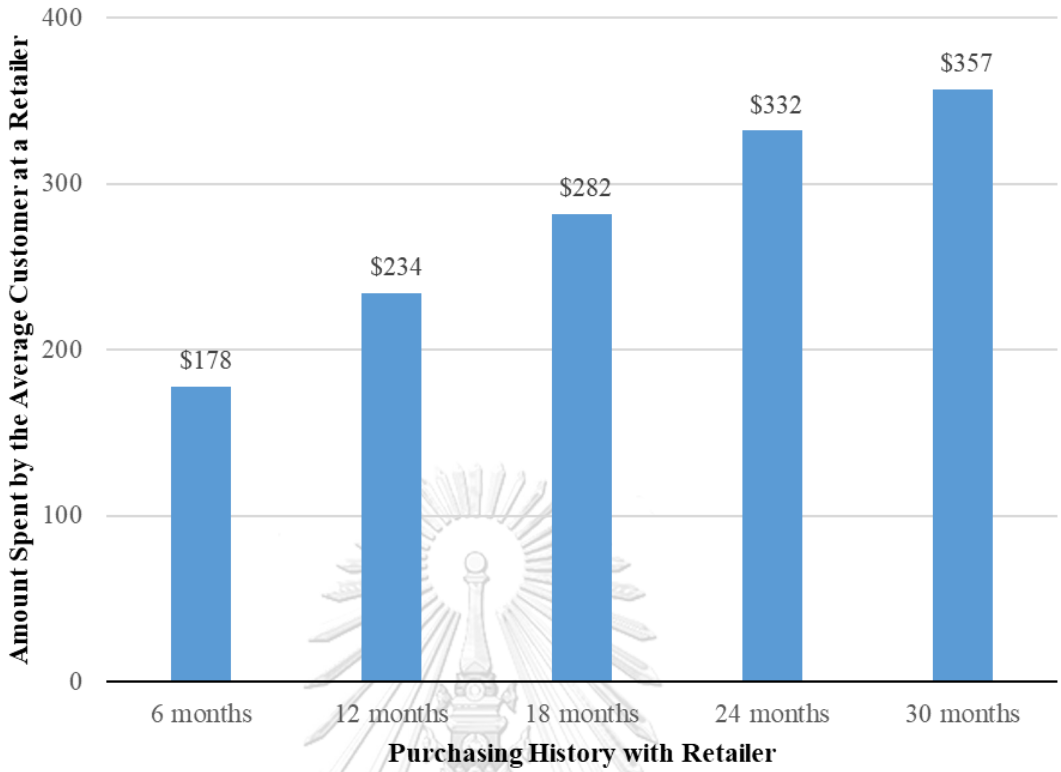


Figure 4. Spending Growth Impact[12]

Moreover, Bain & Company's study found that the number of people referred to retailer since first purchase increases as the number of purchases (refer to Figure 5.). Considering that, the repeat customers also potentially provide a massive marketing advertisement as a part of loyalty which is another benefit of repeat customers.

With the current economic situation moving towards a highly disruptive and rapidly saturating nature, low customer churn is now becoming a crucial metric towards success. The prediction of customer churn plays a significant role in evaluating not only customers, but also business and industries.

Furthermore, return on investment – another essential metric for business success – is highly affected by customer churn. Businesses that have higher customer retention than customer acquisition (a much more expensive process) tend to have a higher return on investment, and therefore higher business success.

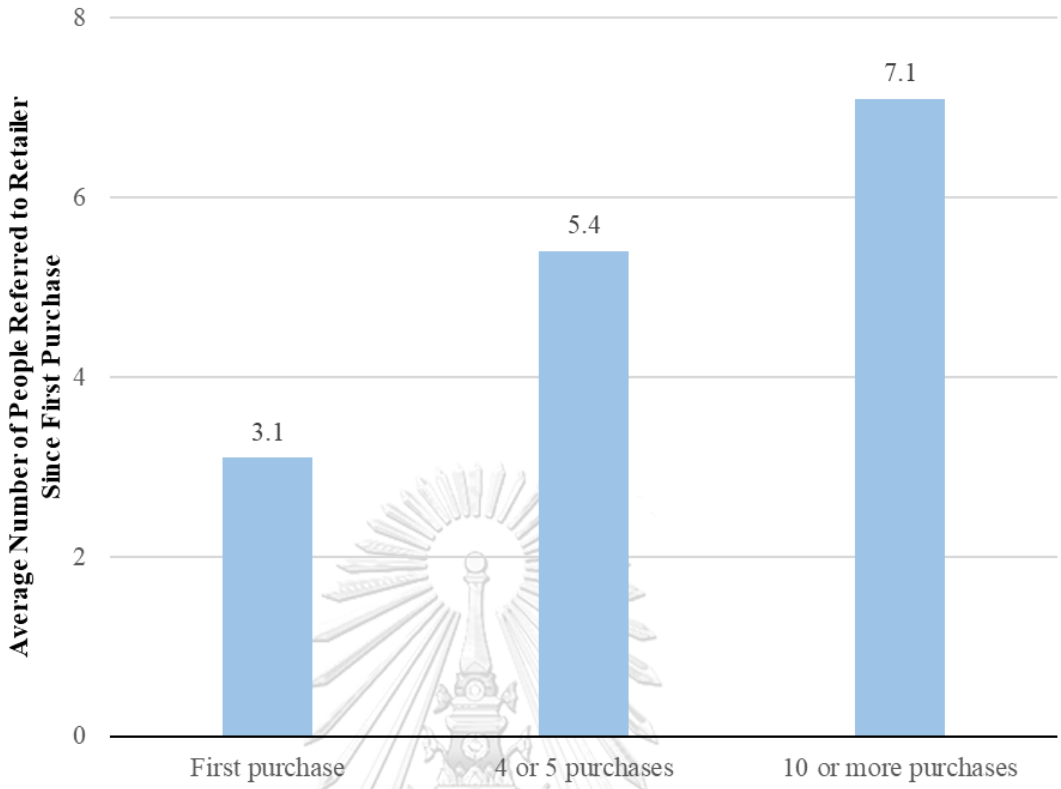


Figure 5. Referral impact[12]

Churn rate is typically calculated annually, quarterly, or monthly depending on the industry and the nature of business. Annual churn rate is generally what most businesses use as a metric, however, certain companies such as phone or software service providers that price products on a monthly basis prefer measuring monthly churn rate.

1.4. Case study general company information

This paper studies Customer Churn Prediction for the company which deployed a Software-as-a-Service business model provided Inventory Management software and solution in Thailand. As an inventory software on cloud, the case study company also functionally integrate with renowned services along with its customers' business value chain, i.e., website, the marketplace, and logistic providers. These combinations enhance the ability of a company to help its customers automatically controls stock and records orders across multiple e-commerce platforms which have been

extensively used in the region e.g., Lazada and Shopee. Moreover, the company acts as a central information source to work across their customer team members and locations, and to see a big picture of their customer business at glance.

Due to the Software-as-a-Service inventory management software nature considered as a Business-to-Business (B2B) model, this case study company offers Integrated Inventory Management Platform and inventory software on cloud tools for SMEs in many types of packages provided the various different options and functional services in the different price tiers. Since the company is a funded startup in which its organization is still under 50 employees, this company needs to quickly capture the market share by acquiring new customers and create the reliability for its customers to keep the company stays in a positive customer retention rate. Therefore, under this circumstance, the company interests in this topic in order to retain its customers with a long-term relationship which brought the company to investigate deeply the factors affect to customer churn rate. And the company also needs the model to be able to predict their risky customers who likely be churned so as to identify the risky customers and directly engage them to resolve the problem before they decide to end the relationship with the company.

Regarding the data provided by the case-study company, the company is facing the severe situation in a high customer churn rate with over 50 percent. Under this circumstance, it is noticeable that the case-study company needs to focus on customer relationship management through customer churn rate prediction. Moreover, the company is also willing to find a model to solve the issue as soon as possible. Since the number of customers has been increased significantly as a result of that the company has already spend a large amount of investment in marketing for a new customer acquisition. On the contrary, the percentage of customer churn has not reduced causing the problem in the case-study company's cash flow owning to the unpredictable revenue stream.

1.5.Problem statement

In Software-as-a-Service (SaaS), customer relation significantly affects a particular company's planning and strategizing which associates with customer loyalty. For an entrepreneurial firm in Software-as-a-Service business, management style and strategic actions that secure loyal customer relations become a tough challenge of the company, especially a rapidly growing company which has limited resources and budgets. As a result of that, the company must carefully concern the factors that sustain its business in long-term. customers' action is a key element to be observed in order to comprehend the factors affecting customer behavior. Moreover, it is also indicated that there is a relationship between market competition and retention strategies. Since those retention strategies can create the company's competitiveness in cruel market situation[13].

Expansion in e-commerce exacerbates the sophistication of the retail industry by changing the retail environment and forcing the digitalization of inventory management. Increment of the frequency of orders and return rates directly affects retailers in inventory management which increase demand for advanced system. Not only the retail industry but every industry along supply chains are also facing the difficulty of complicated inventory management and digital technology revolution.

Due to those sophistication of supply chains and the industry revolution, cloud-based inventory management software which might also referred to Customer Relationship Management (CRM) and Enterprise Resource Planning (ERP) can provide the advantages in many different aspects, e.g., customer tracking, global real-time data, financial management, customer behavior analytics, accessed on-demand and as-needed services using the cloud. Moreover, some big companies tend to integrate vertically through supply chains which can minimize the costs and enhance competitiveness in a specific industry.

Digital technologies enable e-commerce across Business-to-Business. Data-driven, cloud computing for planning, operating and sales, these transformations create opportunities in how service and retail industries track their customers' behavior and

together with transactions, save costs and forecast budgets in order to serve their customers.

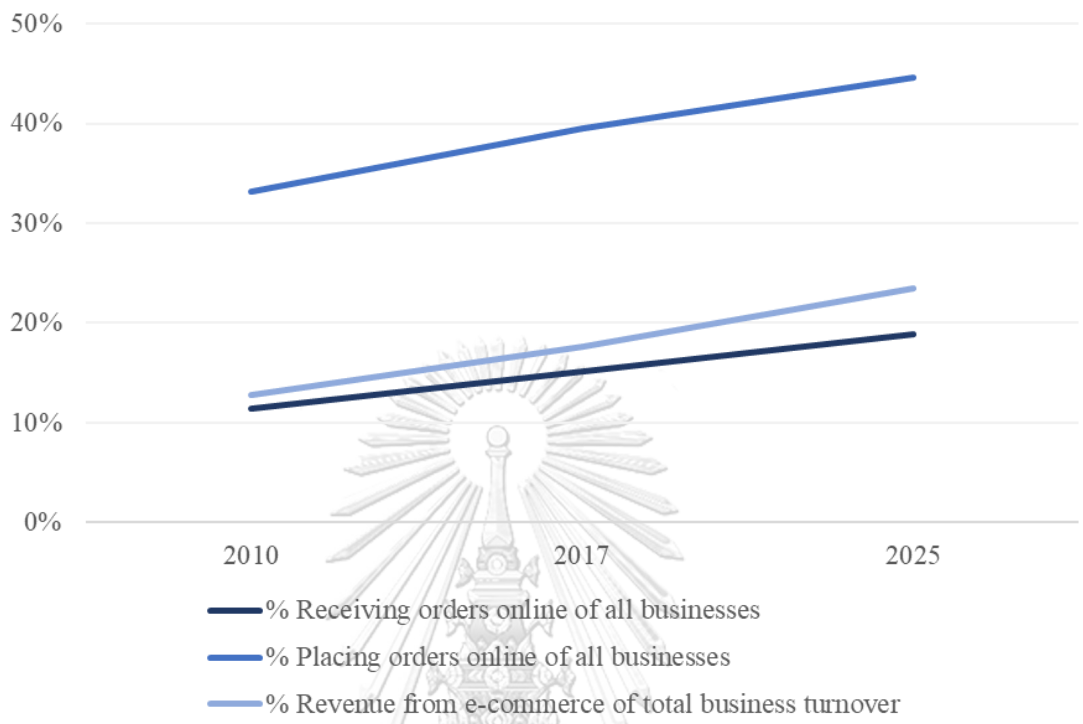
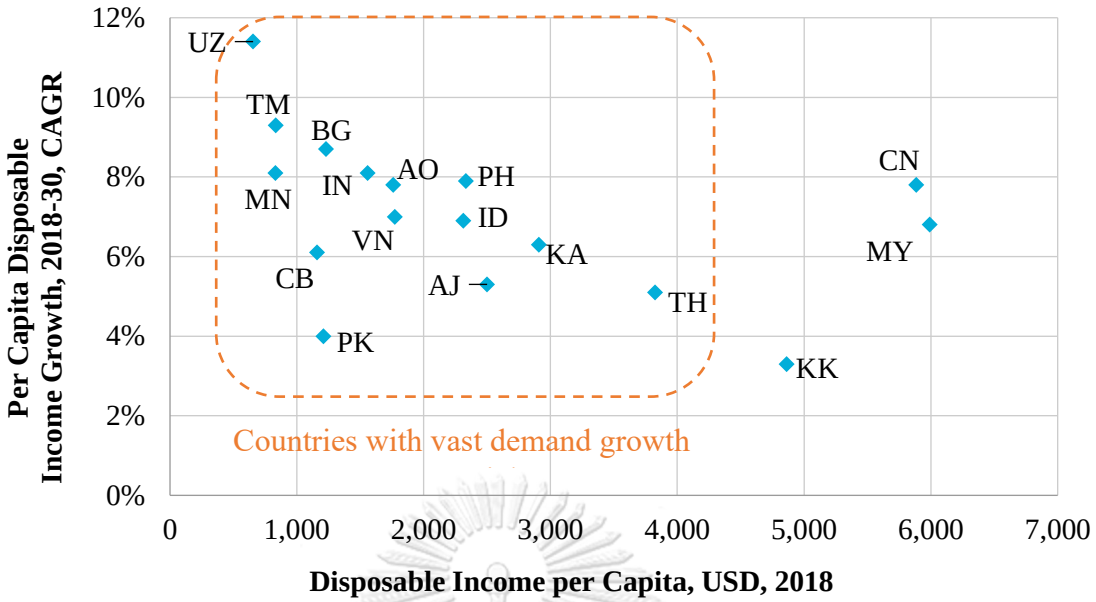


Figure 6. Digital business statistics[14]

Business-to Business sector tends to continue focusing on the technological improvement and efficiency of supply chains. The trend has been recognized for almost a decade and potentially continue accelerate over the next few years. Over the long term, companies will continue digitalized their business in terms of information and transaction streams, which will also increase the growth of e-commerce in Business-to-Business sector. According to statistical information over the past 10 years (refer to Figure 6.), the percentage of market activities of all businesses significantly shifts to online and also project to be increased in the future. The percentage of receiving orders online tend to reach a near 10 percent of all receiving orders in 2025. In case of placing orders, online activity is clearly settled over 40 percent at the present. And Revenue from e-commerce of total business turnover is also increased over 10 percent for the past 10 years. With this information, it can be indicated that the trend businesses seem inevitably moving to online activities and increasing the intense in market competition around the world.



Note: UZ-Uzbekistan, TM- Turkmenistan, BG- Bangladesh, MN- Myanmar, IN- India, AO- Laos, PH- Philippines, ID- Indonesia, VN- Vietnam, CB- Cambodia, AJ- Azerbaijan, PK- Pakistan KA- Sri Lanka, TH- Thailand, MY- Malaysia, CN- China and KK- Kazakhstan

Figure 7. Disposable income growth in Asia Pacific (2018-2030)[15]

Asia is expected to be the dominant economic power by 2030 due to increase in the number of populations, middle class consumers growth, and economic expansion. Asian countries could be divided into 2 groups that support the emerging markets in the region. The first group is the countries that already emerged and continued growing along the previous trend. The second group is the countries that is indicated as new emerging frontier markets. The group on which this study focuses is the new emerging frontiers in Asia pacific that provide an opportunity factors in terms of population and economic growth. Regarding Euromonitor[15], Asia is expected to lead the world's economy by holding the major production capacity which the product output in the market is projected to be originated in this region over 50% of world's production by 2025.

The countries with vast demand growth potential including Thailand have most new growth opportunities, especially for multinational companies (refer to Figure 7.). To support that statement, the total number of middle-class consumers of the countries

with vast demand growth potential mentioned hereinbefore are projected to be over 14.3 million by 2030. In addition, a relatively low number of players in the market empower the new growth opportunities since the competition in consumer markets are weakened by the lower current income levels

In Thailand, considering a recent telecommunication growth over decade and a high mobile device adoption rate, these factors have played a significant role in a further boost to e-commerce activity. While the government also concerns in rapidly developing the digital infrastructure, ICT policy framework 2020 had been introduced that have been expected to initiate to provide over 95 percent of Thailand population with broadband internet at the end of the year 2020 as well as improving the capabilities in terms of e-commerce activity.

Businesses using the internet in Thailand have been increased from 20.5% of businesses in 2013 to 30.1% in 2018 (refer to Figure 8.). This can be implied that the private sector tends to adopt the new venture of internet in terms of improving efficiency and creating new opportunities for their businesses.

Nowadays the businesses in every sector are exposed to the disruption of new innovative startup companies which lead by a scalable business model with entrepreneurial personnel and effective development plan. When the business is growing to a larger scale by rapidly expanding their customer base, the customer churn rate also tends to be ramped up severely caused by unknown factors which the traditional common resolution cannot handle the issues in case of large-scale company. These behind reasons and factors mentioned hereinbefore can be considered that the customer churn rate prediction becomes an important part of solutions to achieve the positive development and high growth in order to capture the market extensively.

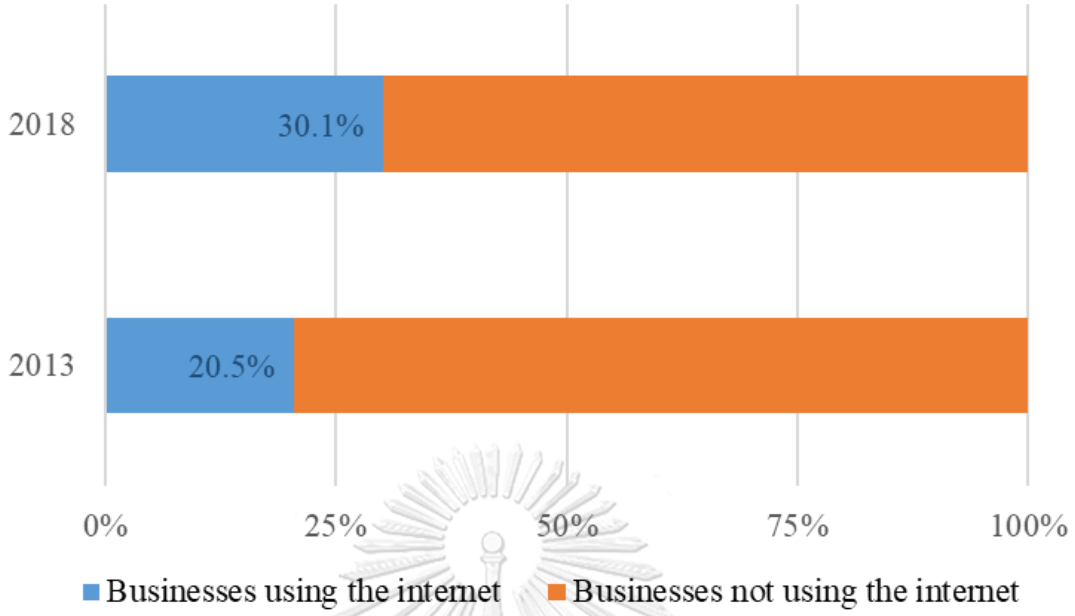


Figure 8. Proportion of Businesses using the internet in Thailand (2013-2018)[16]

The case study company is currently facing the same problem of high customer churn rate due to a larger business and highly competitive environment of software companies in Thailand market. While the case-study company has never analyzed their data in term of customer churn prediction, the company has been trying to find a solution for expansion their revenue growth in order to retain their existing customers by a traditional solution from scratch which is not effective enough to identified enigma. The case-study company needs to indicate the risky customer who can potentially be churned. Therefore, the customer churn prediction for a Software-as-a-Service inventory management company becomes the most important issue for the case-study company to achieve in order to generate continuously reoccurring revenue and save the cost by retain the current customer instead of the new customer.

Regarding the case-study company situation, the company is facing a high customer churn rate with over 50 percent according to the provided data. Under this circumstance, it is noticeable that the case-study company needs to focus on customer relationship management through customer churn rate prediction. Moreover, the company is also willing to find a model to solve the issue as soon as possible. Since the number of customers has been increased significantly as a result of that the

company has already spend a large amount of investment in marketing for a new customer acquisition. On the contrary, the percentage of customer churn has not reduced causing the problem in the case-study company's cash flow owning to the unpredictable revenue stream. As a result of that aforementioned, customer churn prediction become an urgent issue for the company to obtain the first step of resolution as the company aims to reduce the high customer churn rate.

1.6. Objectives

The objective of this thesis is to search for suitable customer churn prediction models for a Software-as-a-Service inventory management company in Thailand.

1.7. Scopes

1. In this study, customer churn is defined as the customer who have not been active consecutively for more than 14 days regarding the case-study company marketers' opinion derived from the marketers' experience in difficulty of customer retention. It means that a particular customer will be indicated as "churn" when that customer has been inactive consecutively for over 14 days. And if that customer has returned to be active within 14 days after being inactive, it will be identified as "Not Churn". It is noticeable that period of churn is based on October 1 to October 14, 2019.

2. This thesis focuses on the inventory management software company adopting Software-as-a-Service business model. Regarding the case-study company being well-known in the industry, especially retailing, the company granted a permission to use their informative expectation and insights and raw data. the data relate to the business and customers, this paper extracted data from the case-study company database based on the data collecting from October 2015 to October 2019. The whole raw data consists of 1788 observations with contributions of 23 feature variables.

3. To model the solution, machine learning classification models using in customer churn predication are Logistic Regression, Support vector machines, Decision Tree, and Random Forest.

4. To evaluate models, evaluation measures used to evaluate the performance of the models are the major terms of evaluation matrixes in regard to order of importance, Recall, F1-score (also known as F-measure), Accuracy and Precision respectively which based on confusion matrix. It is noticeable that this paper considers Recall as an essential metric because, in customer churn detection, if a real churn customer is predicted as a non-churn customer, the consequence will severely affect the balance sheet of the company as the cost of false negative. However, this research considers F1-score as another important metric which also predominates over Accuracy on account of hereinabove mentioned. Moreover, the results of final customer churn prediction models are required to be at least 0.8 or 80% in every dimension (Recall, F1-score, Accuracy, and Precision) regarding the case-study company.

1.8. Thesis outcomes

The expected outcomes of this thesis consist of:

1. Prediction model as a significant tool for Customer Churn Prediction for a case-study Software-as-a-Service inventory management company in Thailand.

2. Factors contributing the customer churn rate in a case-study Software-as-a-Service inventory management company in Thailand.

1.9. Thesis benefits

The expected benefits of this research are as follow:

1. The result provides a reliable customer churn prediction model, this prediction model can assist the company to analysis the company's information, identify customers who potentially be churned. Moreover, it can prevent these potential churn customers who predicted to be churned by prediction model from becoming a real new churn customer by advance the customer relationship management in company which is intentionally keen on adapting to constantly changing market and consumer sentiment to be more effective. It means that the company can reduce customer churn rate by using the customer churn prediction model.
2. The features or factors importance being informative insights can assist the case-study management team in their future strategic planning with comprehension of the behind reasons in churn with significant features and actively encounter with those risky churners who likely to be churned by using proactively engagement or rigorous promotions beforehand.

1.10. Research timeline

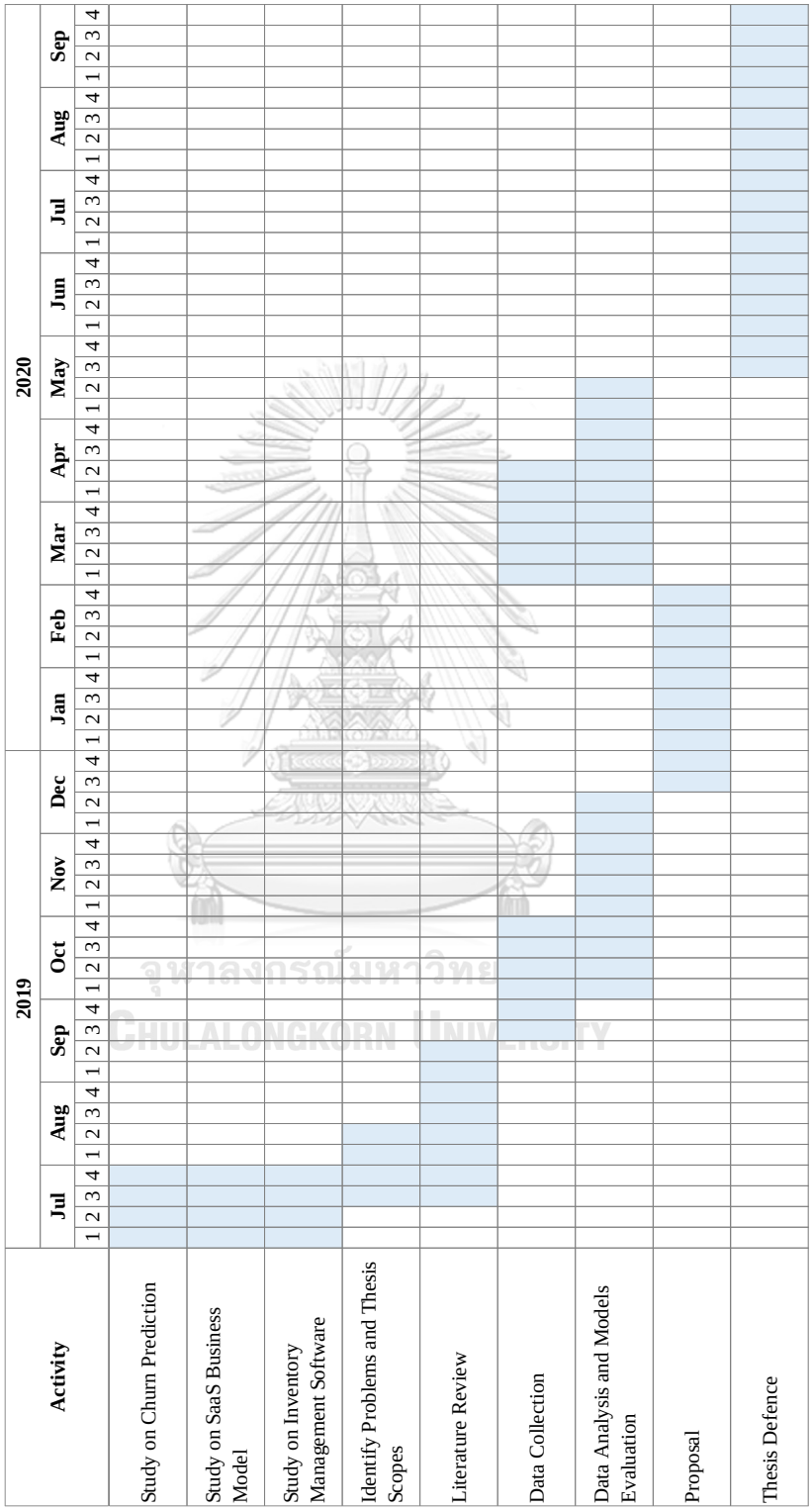


Figure 9. Research timeline

Chapter 2 Literature Review

Over the past few decades, many papers have tried to overcome a customer churn problem in several aspects. Some papers have applied machine learning techniques to execute the problematic customer churn behavior. Each research also extent their interesting to many different industries which define the customer churn in the different ways.

In term of **machine learning techniques**, since 2004, [17]proposed churn management in telecommunication applying data mining including Decision Tree and Neural Network. This paper is nevertheless limited due to changing regulation in the study country. [18]studied on churn analysis for non-contractual FMCG (Fast-moving consumer goods) retail. the paper found that RFM (Recency, frequency, monetary value) variables are the best predictors of partial customer defection. [19]studied the impact of preprocessing on data mining for publishing industry showed that data preparation methods had a positive impact on the prediction results.

Then, [20]indicated that the application of Support Vector Machine (SVM) model is able to maximize the effectiveness for churn management on telecommunication industry. [21]also concentrated on using an extended Support Vector Machine (SVM) forecasting framework in online Retailing comparing the other techniques, Neural Network and Decision Tree. [22]focused on Customer lifetime value (CLV) as a major attribution of the model in Banking sector. [23]studied the application of data mining techniques, e.g., Decision Tree and Neural Network classifier, in predicting risky-churned customers. The results show that efficiency of these techniques, especially Decision Tree are capable to predict and select the related variables.

Interestingly, [24]investigated the customer churn prediction in online gambling industry. Its industry causes the attributes tend to highly relate to RFM (Recency, frequency, monetary value), length of relationship and demographic information. This paper indicates that the ensembles model can provide better performance and more robust than the single models. [25]applied boosting algorithm to Logistic Regression as a basic learner model. the results show that boosting models provide a better

clusters of churn data compared to single Logistic Regression model. [26]assessed the business impact in gaming industry derived from various models; Neural Network, Support Vector Machine, Decision Tree, Logistic Regression, and Hidden Markov Model (HMM).

Table 2. The Studies on Customer Churn Prediction in telecommunication industry

References	Industry	Methodology	Churn Definition
Hung and Wang (2004)[17]	Telecommunication (Wireless)	Decision Tree (C5.0), and Artificial Neural Networks (BPN)	Subscriber switches to a competitor over a period of time
Guo-en and Wei-dong (2008)[20]	Telecommunication	Support Vector Machine (linear), Artificial Neural Networks, Decision Tree (C4.5), Logistic Regression, and Naive Bayes	"Customer does not enjoy all services of
Umayaparvathi and Iyakutti (2012)[23]	Telecommunication	Decision Tree, and Artificial Neural Networks	"Customer shifts to other service provider
Lu et al. (2014)[25]	Telecommunication (Mobile)	Logistic Regression, and ensemble learner (boosting)	"Subscriber switches to another service during a period of time
Wanchai, P. (2017)[27]	Telecommunication (Mobile)	Decision Tree (C4.5), Logistic Regression, and Artificial Neural Networks	Consumer switch or choose to cancel their subscription
Amin et al. (2018)[28]	Telecommunication	Naive Bayes	Customer leave the service or company

After that, [29]investigated managing Business-to-Business (B2B) customer churn with FMCG (Fast-moving consumer goods) retailers' data. the results confirm that the

model provide the performance in churn management surpass managerial heuristics in traditional way. [30]investigated a churn prediction model for large-scale enterprise subscription enhancing Decision Tree techniques with SMOTE.[28] conducted classification model based on data uncertainty for customer churn prediction with Naïve Bayes technique. [31]used Logistic Regression, Gradient Boosted Tree and Neural Networks together with preprocessing and feature selection methods to predict customer churn in an Online Streaming.

Table 3. The Studies on Customer Churn Prediction in banking industry

References	Industry	Methodology	Churn Definition
Glady et al. (2009)[22]	Banking	Decision Tree (Cost-sensitive), Artificial Neural Networks (MLP), Logistic Regression, and ensemble learner (AdaCost)	Customer lifetime value (CLV) decreases over a period of time
Ali and Arıtürk (2014)[32]	Banking	Independently trained binary, Multinomial and Ordinal Logistic Regression, and Survival analysis	Customer's portfolio size is below a specific threshold value during a period of time
Hemalatha and Amalanathan (2019)[33]	Banking	Support Vector Machine, KSVM and Artificial Neural Networks	Not mentioned

Recently, [34]proposed the novel augmented model with existing models of Support Vector Machine to evaluate on online retailers' data which outperform the baseline models. [33]used Support Vector Machine (SVM), kernel function (KSVM) and Artificial Neural Network techniques in prediction models. The results show that the prediction models integrated with those techniques can outperform the original single models.

Table 4. The Studies on Customer Churn Prediction in retail industry

References	Industry	Methodology	Churn Definition
Buckinx and Van den Poel (2005)[18]	Retailing (FMCG)	Artificial Neural Networks (ANN), Logistic Regression, and ensemble learner (Random forest)	Customer changes purchasing behavior during a particular period of time
Yu et al. (2011)[21]	Retailing (Online)	Artificial Neural Networks (BP), Decision Tree (C4.5), and Support Vector Machine (Linear polynomial, RBF)	Customer shifts to a competitor
Tamaddoni Jahromi et al. (2014)[29]	Retailing (FMCG)	Decision Tree (simple, cost-sensitive, CART), Logistic Regression, and ensemble learner (boosting)	Customer has no purchase in prediction period
Vadakattu et al. (2015)[30]	Retailing (Online)	Decision Tree (C4.5) and SMOTE	Customer deactivate in next 3 months
Chen (2016)[35]	Retailing	gamma cumulative sum (CUSUM) chart	Customer not login during a particular period of time
Berger and Kompan (2019)[34]	Retailing (Online)	Support Vector Machine	User considers this session is the last session

Overall, as mentioned hereinbefore, it may be concluded that the studies deployed various different models, e.g., Logistic Regression, Support Vector Machine (SVM), Neural Network, Naïve Bayes, Decision Tree and Random Forest. Additionally, each paper had not used an only single machine learning method to solve the problem, but

many methodologies had been adopted simultaneously for the benefit of performance comparison.

Table 5. The Studies on Customer Churn Prediction in other industries

References	Industry	Methodology	Churn Definition
Crone et al. (2006)[19]	Publishing	Support Vector Machine (RBF), Artificial Neural Networks (MLP), and Decision Tree (C4.5)	Not mentioned
Coussement and De Bock (2013)[24]	Gambling (Online)	Decision Tree (CART), generalized additive model (GAM), and ensemble learners (random forests, GAMens)	Customer not having played during a period of time
Runge et al. (2014)[26]	Gaming (Online)	Artificial Neural Networks, Support Vector Machine (RBF), Decision Tree, Logistic Regression, and Hidden Markov Model (HMM)	Player permanently leaves game

Regarding **the type of business**, customer churn prediction has been studied in various industries. Most of researches commonly focus on Telecommunication industry[17, 20, 23, 25, 27, 28], as shown in Table 2, that customer churn prediction has been remarkably deployed and widely applied. Due to the importance of telecommunication as a main infrastructure of any countries' development, customer churn prediction in Telecommunication has been developed over the time along with changing in technology, i.e., landline, mobile and wireless, respectively.

The common case of customer churn in the telecommunication industry is substitution on account of aggressive competition and new technological disruption. While another famous industry in which studies concentrated is Banking[22, 32, 33], as

illustrated in Table 3, that churn rate directly causes the industry in many aspects, e.g., customer portfolio's size or frequency of transactions. As shown in Table 4, retailing industry[18, 21, 29, 30, 34, 35] also had been investigated by many papers. Those studies are exposed in the different sectors considering the customer churn prediction in the different angles. [18]focused on retailing in fast-moving consumer goods (FMCG). The paper considers the change of customer purchasing behavior as a focal point while recent papers tend to concentrate in online retailing according to market trends[21, 30, 34].

However, the limited number of papers studies in online application, shown in Table 5, such as gaming[26], streaming [31] and online gambling[24] despite emerging of technology. This might be caused by commercial confidentiality that the data of a particular company in this field are considered as an invaluable asset. As well as Software-as-a-Service industry[36-38], there are only few studies compared to the market size of this industry due to limited resources and data access authorization. Most of studies in this Software-as-a-Service industry processed with open data sources that could not catch up the changing in customer behavior in the present time.

Considering **customer churn definition**, there is the difference in each study that is defined to suitable in a particular situation or industry, specific factors and each case-study company's needs while some studies did not mention. Customer churn definition is not only mentioned in the dimension of Customer terminating[26-28, 31, 34, 38], switching to competitors [17, 21, 23, 25, 27] or deactivating any activities [24, 29, 30, 35, 37] but it is also defined in the other aspects of interest, e.g., Customer's portfolio size being lower than a specific threshold value[32], customer changing purchasing behavior[18], decrease in customer's lifetime value (CLV)[22].

Based on these papers' conclusions as shown in Table 2 to Table 6, it can be found that the studies have been conducted in the different angles by each reference, i.e., methodology, type of industries, attributes and customer churn definition.

In Thailand, the study related to customer churn prediction are limited on Telecommunication industry since Telecommunication industry has settled in the country for over a century while Software-as-a-Service industry has prevailed recently

in early stage along with the emergence of e-commerce disruption trend around the world. [27]studying the customer churn analysis for a case study on the telecommunication industry of Thailand found that C4.5 **Decision Tree** algorithm performs well with the optimal result among other techniques; **Logistic Regression** and Neural Network. Additionally, in term of feature category, most of selected features are likely related to usage matric.

However, the difference in the interesting industry can significantly affect the model outcomes. Therefore, the studies related to Software-as-a-Service (SaaS), as described in Table 6, become important references to consider. In early-stage of emergence of Software-as-a-Service (SaaS), [36]applied churn analysis in telecommunications as a baseline with K-means and **Decision Tree** model to Software-as-a-Service (SaaS). Despite of the fact that this paper provides the benefit of features comparison between Software-as-a-Service (SaaS) and Telecommunication, there is still a gap in term of the number or variety of attributes used in analysis. Then [37] applied **Logistic Regression, Random Forest** and also ensemble learner of XGBoost to the model under the churn definition; Customer suspended payment for at least 2 months. The result of this paper shows that XGBoost model can optimize performance effectively assessed by Receiver Operating Characteristic (ROC) and Area under the curve (AUC). For a recent paper, [38]studies the customer churn prediction in Software-as-a-Service (SaaS) with various machine learning algorithm, the result is nevertheless shown that **Support Vector Machine** is the most precise model in this case. It is noticeable that the case study company of this paper is digital marketing campaign management offering marketing solution in domestic which the quality of service highly depends on other online platform policies.

The literature shows that various machine learning techniques have been executed for customer churn prediction in many different industries with unique churn definition in each research. As a result of that different definition, condition or involved industry as aforementioned, the conflicts between the conclusions of these published papers arise. Therefore, this paper studies on prediction of customer churn for a Software-as-a-Service inventory management company based on case study in Thailand.

Table 6. The Studies on Customer Churn Prediction in Software-as-a-Service (SaaS) industries

References	Industry	Methodology	Churn Definition
Frank, B., & Pittges, J. (2009)[36]	Software-as-a-Service	K-Means and Decision Tree	Not mentioned
Ge Yizhe et al. (2017)[37]	Software-as-a-Service	Logistic Regression, and ensemble learner (Random forest and XGboost)	Customer suspended payment for at least 2 months
Rautio, Anton (2019)[38]	Software-as-a-Service	Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Support Vector Machine, and RF	Customer quits using a company's services completely
This thesis	Software-as-a-Service	Logistic Regression, Support Vector Machine, Decision Tree and Random forest	Customer has not been active for more than 14 days

In term of churn definition, it is still relatively constructed with two main bodies; one is the number of interesting objects or individuals, and the another is a specific period of time. In this case-study as mentioned hereinbefore, the customer churn is defined as the customer who have not been active consecutively for more than 14 days regarding the case-study company marketers' opinion derived from the marketers' experience

Instead of using single models' technique, this thesis also applies various machine learning classification models to solve the problematic customer churn including with Logistic Regression, Support vector machines, Decision Tree, and Random Forest.

In conclusion, the existing researches mostly concentrated on major industries, i.e., telecommunication, retailing and banking, while only few papers focused on SaaS industry. Thus, this paper studies on prediction of customer churn for a SaaS inventory management software company based on case-study in Thailand. Moreover, this paper defines churn according to the case-study company marketers' requirement as a customer who have been inactive consecutively for more than 14 days which is different from previous papers

.

Chapter 3 Methodology

In previous chapter, the literature review shows that various machine learning techniques have been executed for customer churn prediction in many different industries with unique churn definition in each research. As a result of that different definition, condition or involved industry as aforementioned, the conflicts between the conclusions of these published papers arise. Therefore, this paper study Customer Churn Prediction for the company which deployed a Software-as-a-Service business model provided Inventory Management software in Thailand.

This paper relies on the machine learning process which is illustrated in Figure 10.

This process begins with collecting raw data that all features are derived from customer usage and business metric regarding the insight of case-study company marketers. Then that data is transformed in order to prepare for execution with machine learning algorithms. Data transformation is including with 3 processes; 1. Data Cleaning, 2. Data Processing and 3. Feature Selection.

After the data is already prepared by these transformation process, the data is split into training dataset and testing dataset by holdout method. For training dataset, this set is used for training model in each machine learning techniques with hyper-parameters tuning and validating the model with K-fold cross validation in this paper. Moreover, evaluating the model is executed on the split testing dataset in order to confirm the final testing results of each machine learning.

Eventually, these final testing results of every machine learning models (Logistic Regression, Support Vector Machine, Decision Tree, and Random Forest) are compared by evaluation metrics to receive the best-performed model.

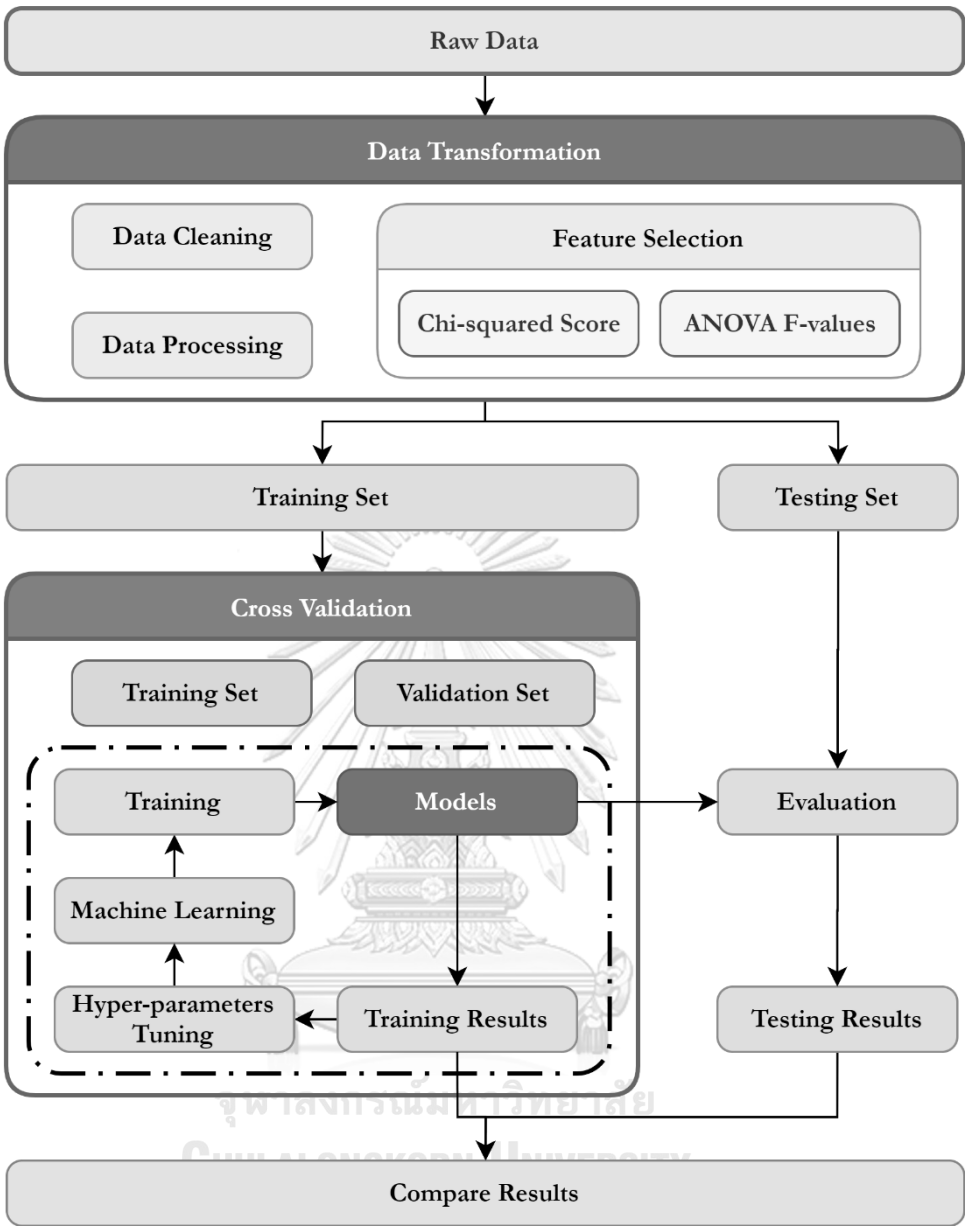


Figure 10. Machine learning process.

3.1.Data collection

This paper is encouraged by the case-study company's intention in improvement the customer churn issues which directly affect the company balance sheet. Regarding In early stage of data exploration, thanks to the company permission, researcher was granted to use their informative expectation and insights and intelligent raw data. researcher had to consider the possibility of data collection in each dimension. Due to

experience and competency, the insights of business provided by the case-study company marketers are derived into various data matric which mostly are related to customer usage and business metric. Since the company concerns the privacy of their customer, some specific data that can be used to trace back or indicate the customer identification are not allowed to access. And the data are also sorted randomly for the same reason. Therefore, the data collection in the first step must have been checked and confirmed by the company authorities. Nevertheless, the provided data are enough to generate and execute the models.

Then, this paper extracted data from the case-study company database based on the data collecting from October 2015 to October 2019. The whole raw data consists of 1788 observations. Regarding data extraction, the related features or attributes are listed in Table 7 which are also described the detail in each attribute. the features contribute of 23 variables in terms of customer usage behavior and business matric such as transactions. It is noticeable that this paper defines churn as a customer who have not been active consecutively for more than 14 days while period of churn is based on October 1 to October 14, 2019.

Table 7 List of attributes

Attribute Name	Description
daysToAct	Days from registration to first action
daysToExpire	Days until an expected expiration date.
totalTrans	Number of all transactions
currPeriodTrans	Number of transactions in current period of 14 days
prevPeriodTrans	Number of transactions in previous period of 14 days
amountSpend	Amount of customer spending
numAct	Number of actions per customer

Attribute Name	Description
Act_Day	Average number of actions per day
numCargo	Number of cargoes
numUser	Number of users per customer
UserAct	Average number of actions per user
lastContactDays	Days since the last contact with salesperson
everContact	Represent if salesperson ever contacted to a customer (Ever Contact, Never Contact)
numContact	Number of contacts
amountTrans	Amount of transaction values via platform
hasPhone	Represent if contact is provided (has Contact, has no Contact)
amountSpend_log	Amount of customer spending per login
numAct_log	Average number of actions per login
numUser_log	Average number of users per login
amountTrans_log	Average amount of transaction values via platform per login
usernumAct_log	Average number of user actions per login
Act_Day_log	Average number of actions per day per login
numCargo_log	Average number of cargoes per login
churn	Represent if a customer has not been active for more than 14 days (Churn, Not Churn)

3.2.Data transformation

3.2.1. Data cleaning

Regarding the raw data, missing data, meaning that there is no data value provided for the observed variable in a sample, are under consideration. This paper applied the listwise deletion technique to handle this problem by excluding an observation in case of missing any single data value[39]. This method is also applied in several customer churn prediction papers[27, 40]. In this case, Listwise deletion reduce the missing data from the 1788 observations of the whole raw data to 1718 observations. The data also includes both churn and non-churn customer records; 927 churn samples and 791 non-churn samples. This original dataset seems to be likely balanced distribution with 53.96% of churn samples and 46.04% of non-churn samples.

3.2.2. Data processing

The collected raw dataset is typically provided in the two different types of variables, i.e., quantitative variables and categorical variables. Preprocessing data transform raw data to be more appropriate for machine learning and fitting in the prediction models. In case of categorical data, in general, this type of variables was encoded by various techniques. Due to the nature of categorical data provided by case-study company, this paper uses one hot encoding method to convert the data. Furthermore, standardization of dataset is applied in this case. Regarding the benefits in implementation, machine learning classification algorithms commonly requires function scale in standardization. This methodology neglects the shape of distribution by scaling to the feature's standard deviation after removing the mean of a certain feature which is simple and effective.

3.2.3. Feature Selection

In general, feature selection is an important process which enhances the performance of machine learning algorithms and decrease computational complexity or time complexity by reducing the number of independent variables or explanatory variables. The feature selection is sophisticated part since it is not provided in a specific

procedures or rules, meaning that both basic methods like statistical filter or searching algorithm and advance methods like dimensionality reduction can be applied without any guarantee of its performance. It can be inferred that each feature selection technique is suitable to apply in a certain problem.

This paper focuses on two types of univariate feature selection or statistical filter method for classification, i.e., Chi-squared score and ANOVA F-values. Both Chi-squared and ANOVA filter methods statistically measure each features' importance separately and rank these features variables in an order of importance.

Then, the top ranks of highest-score features are selected to be used in any machine learning in next step while the number of selected features is a parameter needed to be assigned. As a result of that, this paper executes all possible numbers of selected features.

3.3.Machine Learning

Machine learning classification is a supervised learning method in which algorithm learns from data input and predict the class of given observed data. Since there is various of classification functions applied in many different researches with single or multiple models as mentioned in Literature Review, the appropriate techniques for this paper, studying customer churn prediction for a Software-as-a-Service inventory management company in Thailand, are deliberately selected regarding the functionality, reliability and performance of those algorithms in previous researches. In this paper, these algorithms are functioned using Scikit-learn focusing on machine learning in Python[41].

Moreover, in this paper, grid search technique is used to hyper-parameters tuning for each machine learning classification methods instead of running with only default parameters. The following machine learning techniques is essentially generated by various values or type of hyper-parameters. These models with different constraints or values of hyper-parameters provides unique results. Therefore, choosing relevant hyper-parameters for each model are necessary[38]. And hyper-parameters tuning is

important to optimally solve the problem by each machine learning classification methods.

As was pointed out earlier, the machine learning process includes transforming data, training model, testing model, and comparing results. These steps are demonstrated in Figure 11

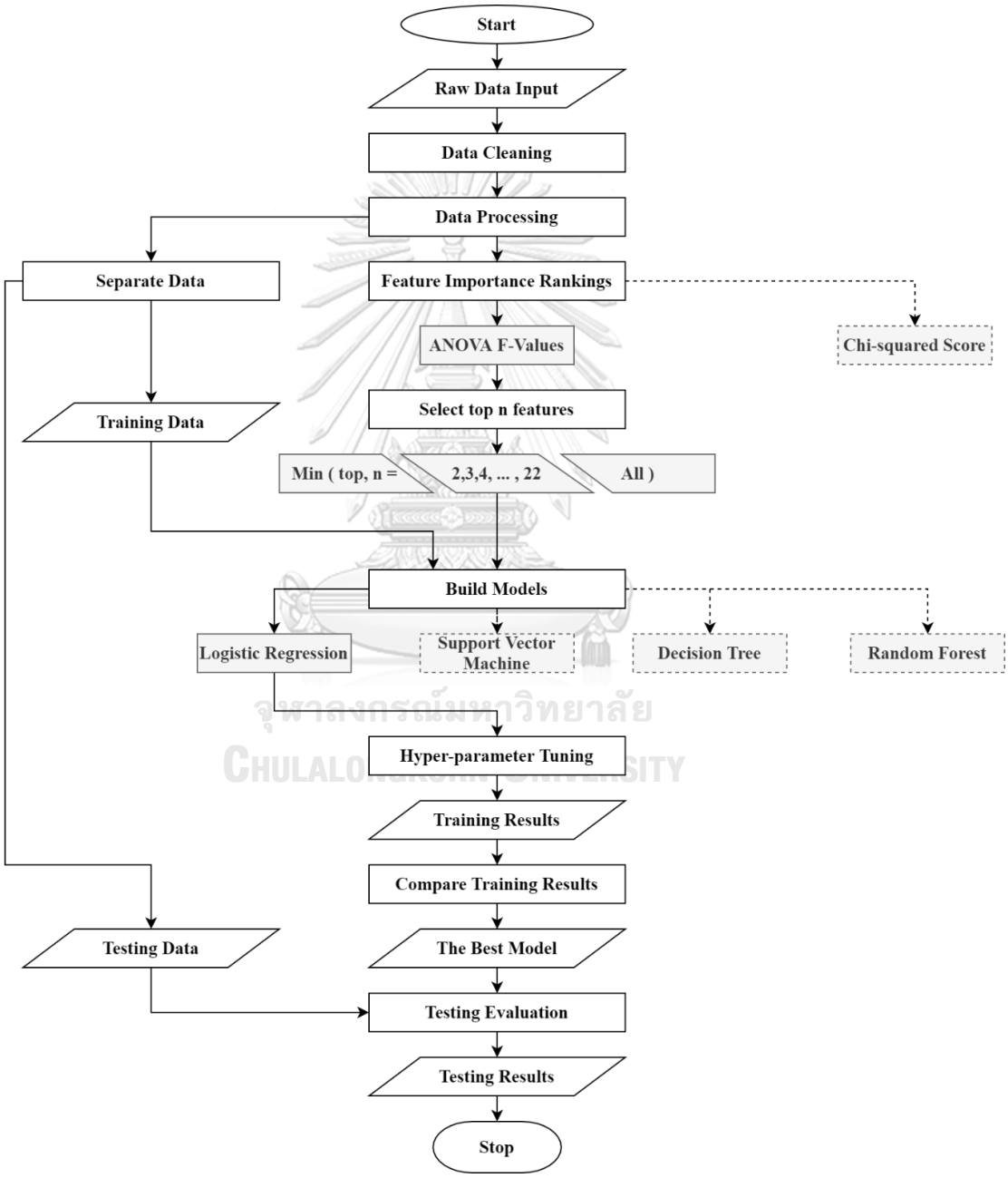


Figure 11. Machine learning flowchart

The overall steps of machine learning are explained step by step in detail as follows.

1. Collect the raw data from case-study company.
2. Clean data by removing any missing value data.
3. Transform data to be able to properly use in the models
4. Rank important features with two different methods: ANOVA F-values and Chi-squared method.
5. Separate data in 2 portions: 80% training data and 20% Testing data
6. Train each classification model with all different set of selected top n features, for example, training dataset which selected only top 10 important features by ANOVA method are brought to train the logistic regression model. This investigation has to train each model by varying the number of selected top important features from minimum features to maximum features for both ranking feature importance methods: ANOVA F-values and Chi-squared method. The machine learning classification models herein includes logistic regression, support vector machine, decision tree, and random forest. This training model step is shown in Figure 12.
7. For each training, hyper-parameter tuning is necessary to be achieved because a particular machine learning model with different input data can provide the best result by a certain set of hyper-parameters. Not only hyper-parameter tuning, the training also includes the k-fold cross validation.
8. Compare the results of every trainings (different feature selection methods, different number of selected top important features and different machine learning algorithms) are compared to each other to receiving the best model.
9. Validate the best model with testing data which separated previously.

A more detailed account of machine learning algorithms is given in the following parts.

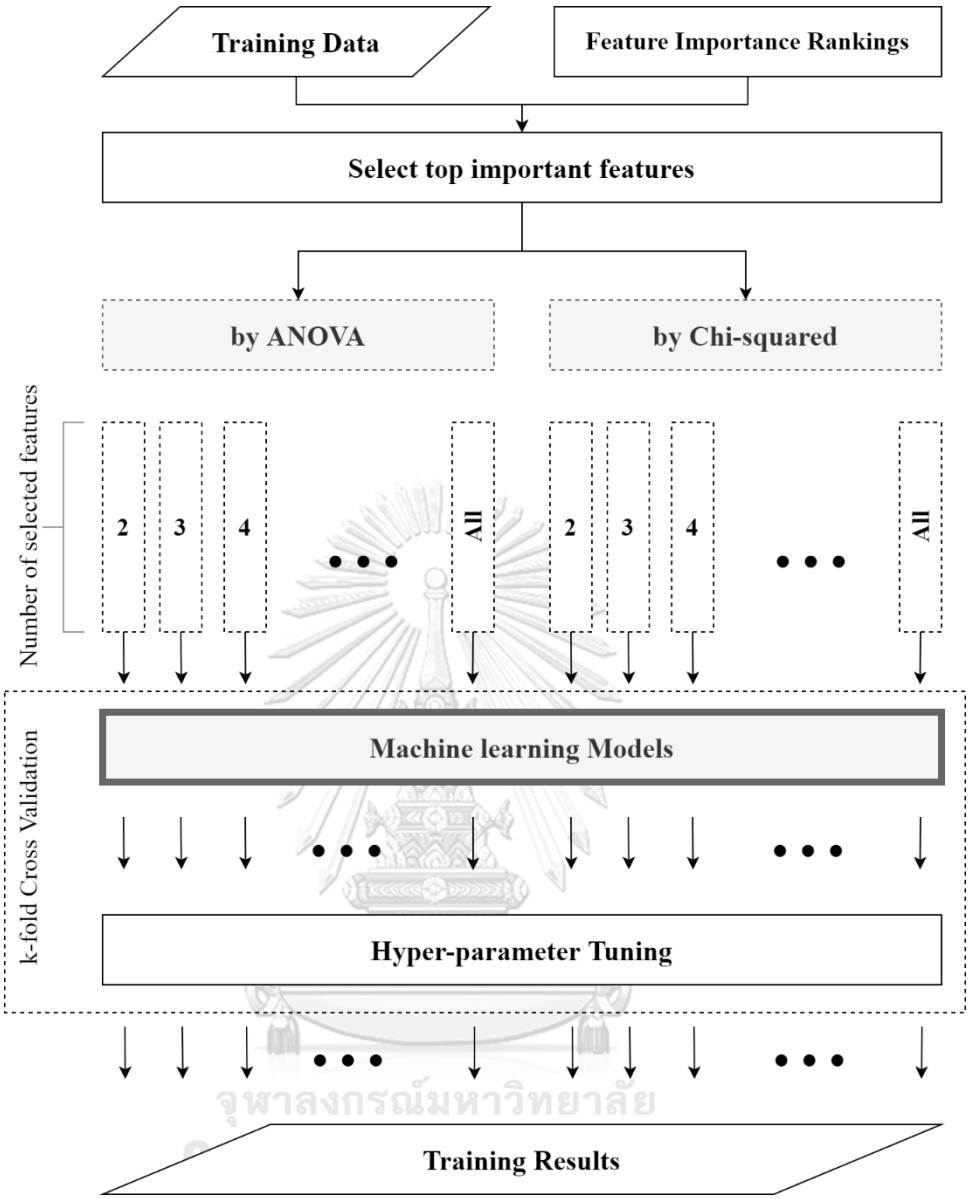


Figure 12. Training model process

3.3.1. Logistic Regression

Logistic regression is one of the well-known classification methods in machine learning. The Logistic regression model represent the relation between independent variables and dependent variable in probabilities while dependent variable has two possible outcomes, i.e., churn and non-churn.

Theoretically, logistic regression is implemented to be able to fit binary, one-vs-rest or multinomial where this thesis focuses on binary method. the core term of logistic regression is penalty term which expressed as cost function or loss function.

Considering an optimization problem, L1 is defined as squared magnitude of coefficient in penalization on the loss function in regression model in order to optimize the following optimization problem as equation (1):

$$cost\ function = \min_{w,c} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \log\left(\exp\left(-y_i(X_i^T w + c)\right) + 1\right) \quad (1)$$

Similarly, L2 is defined as absolute value of magnitude of coefficient in penalization on the loss function in regression model. L2 is expressed in solving the following optimization problem as equation (2):

$$cost\ function = \min_{w,c} \| w \|_1 + C \sum_{i=1}^{n} \log\left(\exp\left(-y_i(X_i^T w + c)\right) + 1\right) \quad (2)$$

While a combination of L1 and L2 is elastic-net regularization and penalizes on the loss function in regression model in order to optimize the following optimization problem as equation (3):

$$\min_{w,c} \frac{1-\rho}{2} w^T w + \rho \| w \|_1 + C \sum_{i=1}^{n} \log\left(\exp\left(-y_i(X_i^T w + c)\right) + 1\right) \quad (3)$$

This method performed effectively in customer churn problem for Software-as-a-Service business[37].

Considering hyper-parameters tuning for logistic regression classifier, the parameters which this paper selected are essential parts to control the model's complexity or detail in application of different techniques in some steps of classification. In this

case, the hyper-parameters for logistic regression in which this paper is interested consists of 3 parameters; 1. penalty, 2. C, and 3. tol. (see Table 8)

Table 8. Hyper-parameters list for logistic regression.

Hyper-parameter	Definition
penalty	the types of norm used as penalty term. • L1 is defined as squared magnitude of coefficient in penalization on the loss function in regression model. • L2 is defined as absolute value of magnitude of coefficient in penalization on the loss function in regression model.
C	the values of model simplicity or irregularization strength. • The higher value of C indicates simpler model. The lower value of C creates more complex model or stronger regularization.
tol	the values of tolerance criteria. • The model stops searching toward the objective or optimization when the indicated tolerance is reached.

3.3.3. Support vector machines

Support vector machines is one of the classifiers that distinguish the class of the output with hyperplane. This technique is able to handle the linear and non-linear classification problems.

Support vector machine function consists of subset of training dataset which is the support vectors. A related problem is under the criteria of optimization problem and mathematically related to Lagrange multiplier method. Support vector machine arranges the data which is non-linear separable into a higher dimension of space in order to define the hyperplane that be able to partition the sample data in groups.

In support vector machine function, kernel function is used to identify the mapping model of hyperplane to separate the data and weaken the complexity of mapping function. It can be indicated as the following equation(4):

$$K(x_i, x_i) = \emptyset(x_i) \cdot \emptyset(x_j) \tag{4}$$

The kernel function can be formed with various terms of function. The kernel functions are indicated herein, i.e., Polynomial kernel, RBF (Radial basis function) kernel and Sigmoid kernel as the following.

Polynomial kernel shown in equation:

$$K(x_i, x_i) = (x_i \cdot x_j + 1)^p \tag{5}$$

RBF (Radial basis function) kernel shown in equation:

$$K(x_i, x_i) = e^{-\gamma(x_i - x_j)^2} \tag{6}$$

Sigmoid kernel shown in equation:

$$K(x_i, x_i) = \tanh(\eta x_i \cdot x_j + \upsilon) \tag{7}$$

Many studies were successful in solving classification problems.[21], focused on retailing industry, is one of those papers that getting a favorable outcome for Support vector machine as well as [20] who studied in telecommunication industry. Moreover,

support vector machine brought a better performance in the results of algorithm comparison for churn prediction in Software-as-a-Service.

Considering hyper-parameters tuning for support vector machine classifier, the parameters which this paper selected are essential parts to control the model's complexity or detail in application of different techniques in some steps of classification. In this case, the hyper-parameters for support vector machine in which this paper is interested consists of 3 parameters; 1. kernel, 2. C, and 3. tol. (see Table 9)

Table 9. Hyper-parameters list for support vector machine.

Hyper-parameter	**Definition**
kernel	the types of kernel used in algorithms. • Kernel type indicate the mapping model of hyperplane to separate the data.
C	the values of model simplicity or irregularization strength. • The higher value of C indicates simpler model. The lower value of C creates more complex model or stronger regularization.
tol	the values of tolerance criteria. • The model stops searching toward the objective or optimization when the indicated tolerance is reached.

3.3.5. Decision Tree

Decision tree is generally a common prediction model method and widely used in various fields, e.g., field statistics, data mining and machine learning. of supervised machine learning problems A decision tree is a tree where each node represents a feature or an interested item, each branch represents a decision dividing a node into two or more sections and each leaf or terminal node represents a target value being discrete in case of classification tree. An example diagram of decision tree is shown in Figure 13.

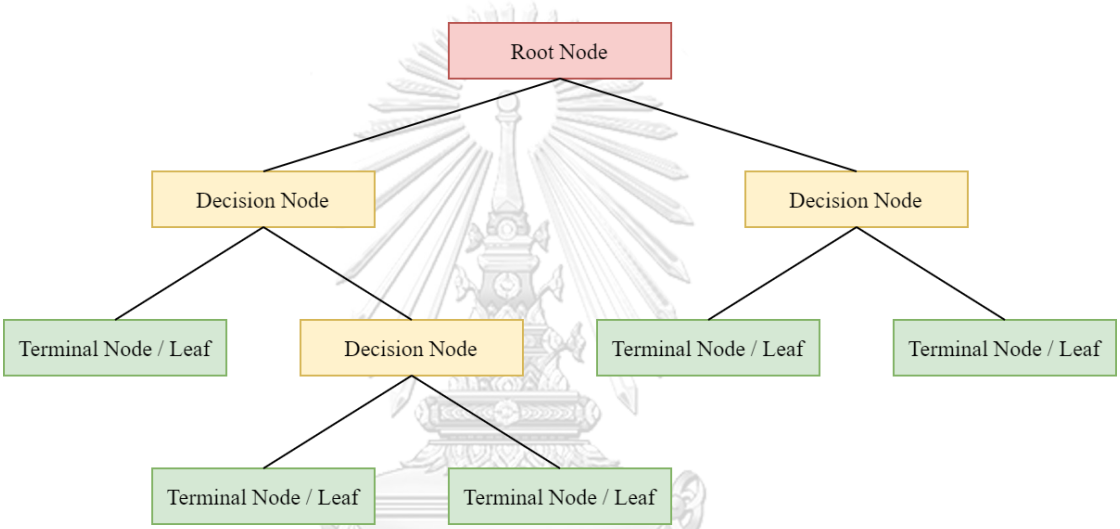


Figure 13 Decision tree diagram

Function that related to decision tree mainly focuses on the techniques used to measure the impurity quality of split. The types of functions are herein including with Gini impurity or Gini index and information gain.

Gini impurity or Gini index is used to evaluate the probability occurrences of incorrectly splits in dataset as a cost function. It works with only binary on the categorical target variable. This Gini index can calculate by subtracting sum of the squared probability of a particular labeled variable for every target variable as equation (8):

$$\text{Gini} = 1 - \sum_{i=1}^{c} (p_i)^2 \qquad (8)$$

Information gain is statistically described in term of entropy which is an index of randomness of information being used in the process. Entropy for a particular feature can be represented as the following equation (9):

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i \quad (9)$$

Information gain is measured by the difference between entropy before split and average entropy after split using the data of a particular feature or attribute values as indicated in equation (10):

$$\text{Information gain} = \text{Entropy(before)} - \sum_{j=1}^{K} \text{Entropy(j, after)} \quad (10)$$

This technique can be applied very simple and straightforward with breakdown in smaller portions[29].

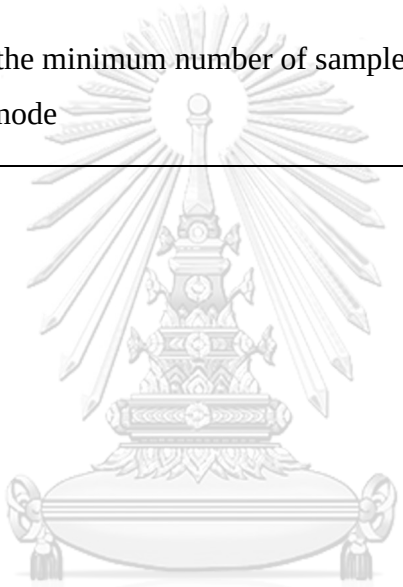
Decision tree was performed accurately for churn prediction problem studied by [17] while, in Thailand, [27]received a good result for churn prediction in telecommunication industry. [36]also applied decision tree classification to predicting customer churn in Software-as-a-Service industry.

Considering hyper-parameters tuning for decision tree classifier, the parameters which this paper selected are essential parts to control the model's complexity or detail in application of different techniques in some steps of classification. In this case, the hyper-parameters for decision tree on which this paper focused consists of 5 parameters; 1. criterion, 2. splitter, 3. max_depth, 4. min_samples_split, and 5. min_samples_leaf. (see Table 10)

Table 10. Hyper-parameters list for decision tree.

Hyper-parameter	**Definition**
criterion	the types of function to measure the impurity quality of split at each node. • 'gini' is defined as Gini impurity or Gini index.

	• 'entropy' is defined as information gain.
splitter	the techniques used to split samples at each node. • 'best' is indicated choosing the best split. • 'random' is indicated choosing the best random split.
max_depth	the maximum depth of decision nodes
min_samples_split	the minimum number of samples required to split an internal or decision node
min_samples_leaf	the minimum number of samples required to generate a leaf node



3.3.7. Random Forest

Random forest is generated by combining with a numerous decision trees in process of machine learning as one of ensemble learning classifier techniques. This technique is also stable in term of diverse data and better reduction in risk of overfitting effect.

The background concept of random forest is based on the idea of bootstrap aggregating or bagging to decision tree learners. A training dataset X with responses Y bagging constantly B times fits to train a classification. After training samples, the prediction function can be express as the following equation(11):

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(x') \tag{11}$$

The number of samples or trees is a particular parameter that generally assigned since a few numbers to several thousand trees. And the optimal size of samples can be obtained by applying cross validation. While the random forests use modified decision tree learning algorithm that chooses the split random of features set. By processing this feature bagging, the predictors can be selected with only few strong features regarding the correlation of the trees. However the number of features will be specified in each classification problems in order to obtain the optimal value by tuning parameters.

[38], who studied in churn prediction in Software-as-a-Service, found that this random forest technique provides a great performance as similar as support vector machine as hereinbefore mentioned.

Considering hyper-parameters tuning for random forest classifier, the parameters which this paper selected are essential parts to control the model's complexity or detail in application of different techniques in some steps of classification. In this case, the hyper-parameters for random forest on which this paper focused consists of 5 parameters; 1. n_estimaters, 2. criterion, 3. max_depth, 4. min_samples_split, and 5. min_samples_leaf. (see Table 11)

Table 11. Hyper-parameters list for random forest.

Hyper-parameter	Definition
n_estimators	the number of decision trees
criterion	the types of function to measure the impurity quality of split at each node. • 'gini' is defined as Gini impurity or Gini index. • 'entropy' is defined as information gain.
max_depth	the maximum depth of decision nodes
min_samples_split	the minimum number of samples required to split an internal or decision node
min_samples_leaf	the minimum number of samples required to generate a leaf node

In conclusion, this paper applies these 4 different techniques of machine learning classification, i.e., Logistic Regression, Support Vector Machine, Decision Tree and Random Forest. The behind reason of this selection is that this paper gathers the machine learning algorithms which performs successfully as hereinbefore mentioned in each classification techniques, especially related to Software-as-a-Service business model.

3.4.Cross Validation

Cross validation is a common technique to assess and compare the results from different prediction machine learning models. This paper applies two general methods including with holdout method and k-fold cross validation (refer to Figure 14). The dataset is partitioned in to 80% of testing data and 20% of training data under holdout method. This holdout testing set is used to evaluate the final model while the training set is proceeded k-fold cross validation.

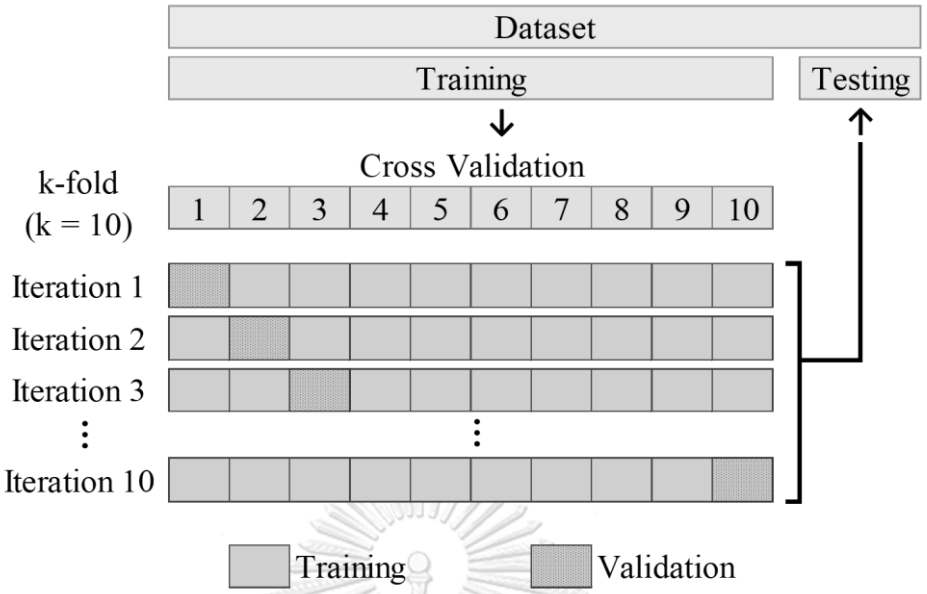


Figure 14. Cross validation

K-fold cross validation is considered as important procedure. This method sticks to k parameter which is the number splitting sample dataset in cross validation. Mostly, k value is popularly assigned to be equaled to 10. [42]mentioned that there is not any specific rule for valuing k while a value of k is commonly equal to 10. It is noticeable that this value of k = 10 is valid to high variance data which create the bias of k-fold cross validation technique. As a result of that, in this case, k is defined as equal to 10, meaning that the cross-validation dataset is randomly split into 10 portions, processing 10 iterations with different validation datasets. In this step, k-fold cross validation is considered as the cross validation in training model process.

Since holdout method and k-fold cross validation is the process of data randomly splitting, it is concerned that the randomly split might generate different results for each execution of the cross validation. Thus, this paper has set a specific splitter or random state to ensure that every model is executed on the identical split of training and testing dataset including split samples in k-fold cross validation.

Overall, this paper tries to negate the effect of bias and overfitting by applying k-fold cross validation. Therefore, all results of evaluation metrics, mentioned hereinafter, from k-fold cross validation are used to compare the performance of prediction models. Then, the top performance prediction models are chosen. In addition to that,

the final customer churn prediction model using handout testing dataset which imitated the real data is evaluated to confirm the capability of the model performance on unknown dataset.

3.5.Evaluation metrics

In order to evaluate and compare the performance or accuracy of prediction models, confusion matrix is a suitable tool to describe the outcome between predication and actuals in a binary classification, churn and non-churn, given in Table 12. it is indicated in 4 terms: true positive (TP), false positive (FP), false negative (FN) and true negative (TN). Considering the confusion matrix, it results can be used to calculate in 4 different dimensions including with Accuracy, Precision, Recall and F1-score. Each metrics' dimension can satisfy the problem depending on the purpose of each business issue.

Table 12. Confusion matrix for customer churn prediction

		Predicted	
		Negative (Non-churn)	**Positive (Churn)**
Actual	**Negative (Non-churn)**	TN	FP
	Positive (Churn)	FN	TP

Accuracy is the ratio of total number of correct predictions to all observations which generally indicated the performance of predication in terms of frequency. And its calculation is shown in eq. (12):

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{12}$$

Precision is the ratio of correctly predicted churns to all predicted churns. precision is normally used in the problem focusing on false positive results or incorrectly churn prediction. And its calculation is shown in the following eq. (13):

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{13}$$

Recall is the ratio of correctly predicted churns to all churns meaning that the performance in prediction of how much real churns are correctly indicated. And its calculation is shown in eq. (14):

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{14}$$

F1 score is the weighted average of precision and recall which also indicate the overall prediction accuracy. And its calculation is shown in eq. (15):

$$\text{F1 score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{15}$$

Regarding the case-study business objectives, the various measures of confusion matrix are considered in the order of importance. Due to the definition of churn, the paper aims to mitigate the cost of a new customer acquisition which is significantly high compared to the cost of a customer retentions caused by churn customers. It can be concluded that it is necessary to correctly identify a churned customer.

This paper considers Recall as an essential metric because, in customer churn detection, if a real churn customer is predicted as a non-churn customer, the consequence will severely affect the balance sheet of the company as the cost of false negative. As a result, the scoring parameter in grid search mentioned in machine learning section is evaluated by Recall. Precision which values false positive as the most significant is considered as the last order of importance among other measures. F1-score and Accuracy are likely the same in term of overall performance indicator. The disadvantage of Accuracy is the dependence on data classification balance while F1-score can handle the issue of imbalanced data. Even though accuracy can be affected by imbalanced data, this case-study dataset, combining with 53.96% of churn samples and 46.04% of non-churn samples, can be implied that the data are balanced.

Therefore, Accuracy is still applicable. However, this research considers F1-score as another important metric which also predominates over Accuracy on account of hereinabove mentioned. Moreover, the results of final customer churn prediction models are required to be at least 0.8 or 80% in every dimension (Recall, F1-score, Accuracy, and Precision) regarding the case-study company.

Chapter 4 Results and Discussion

In this section, the whole processes are based on details in 30Methodology section, especially illustrated in Figure 10, Figure 11, and Figure 12. The particularized results in each step are given in the following parts.

4.1.Data Collection

This research applies four machine learning algorithms on the data granted by the case-study company. The raw data was extracted from the case-study company database while the whole raw data consists of 1788 observations with 23 features in both customer usage behavior and business matric.

4.2.Data Transformation

4.2.1. Data cleaning

Regarding data transformation as explained in the methodology, the dataset had been firstly removed every missing-value observation by listwise method that is data cleaning step. As result, the 1788 observations of the whole raw data had been cleaned the missing data to become 1718 observations. The data includes both churn and non-churn customer records; 927 churn samples and 791 non-churn samples. This original dataset tends to have likely balanced distribution with 53.96% of churn samples and 46.04% of non-churn samples.

4.2.2. Data processing

Then, the data processing had been secondly proceeded in order to prepare data to be readily executed in other steps. the data processing includes one hot encoding method used to convert the categorial data and standardization used to regularize the dataset without effect of different means.

4.2.3. Feature selection

The last step in data transformation is feature selection or feature importance ranking, the dataset was evaluated the feature importance by two types of univariate feature selection or statistical filter method for classification, i.e., Chi-squared score and ANOVA F-values. The order of feature importance calculated by ANOVA F-values is shown in Figure 15 while the order of feature importance determined by Chi-squared score is shown in Figure 16 the higher score feature means more related to output or churn in this case considering the feature selection methods. The raw scores of feature importance calculated by both ANOVA F-values and Chi-squared are demonstrated in Appendix 1. It is notable that this thesis used every data to analyst the feature importance.

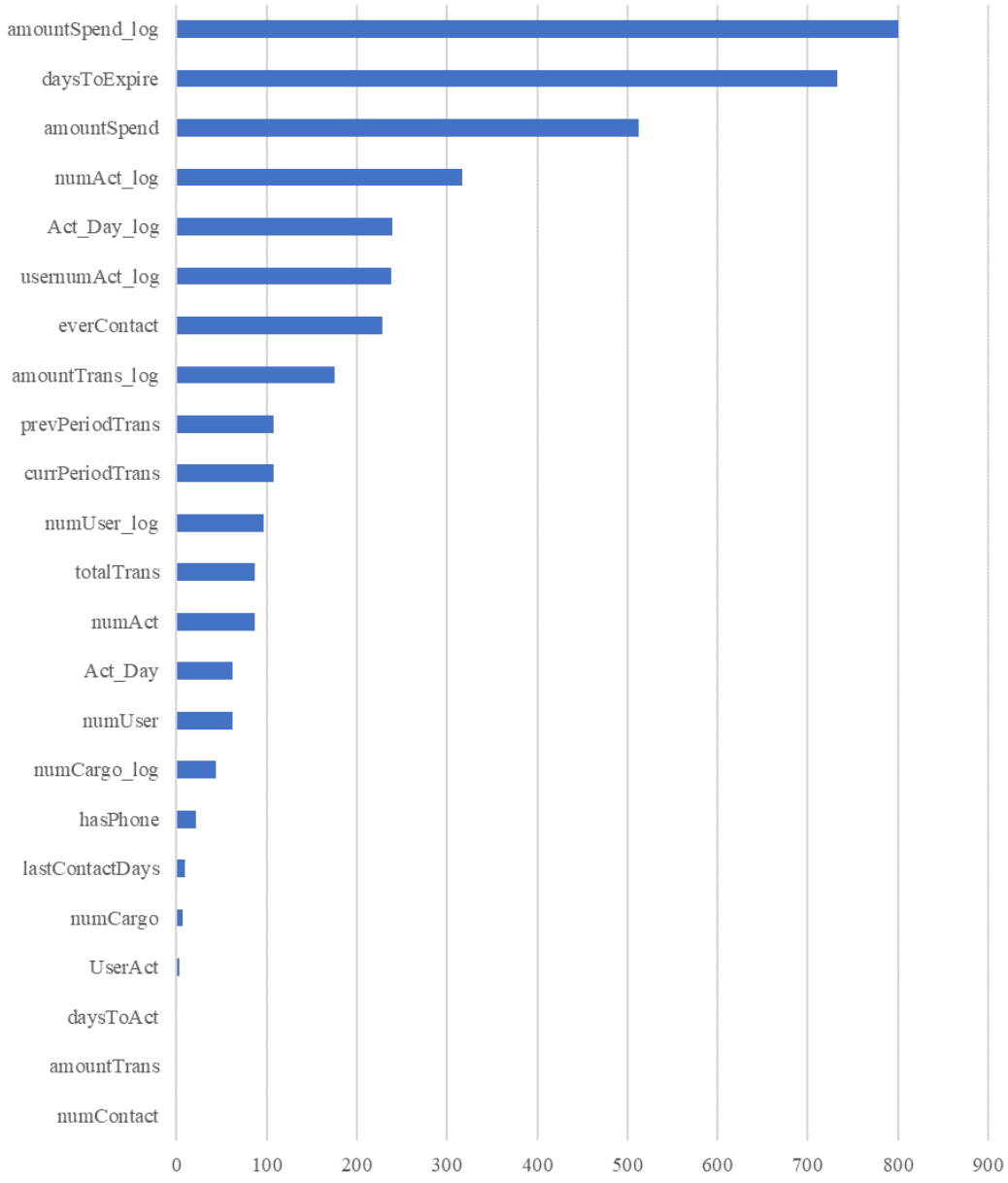


Figure 15. Feature importance by ANOVA.

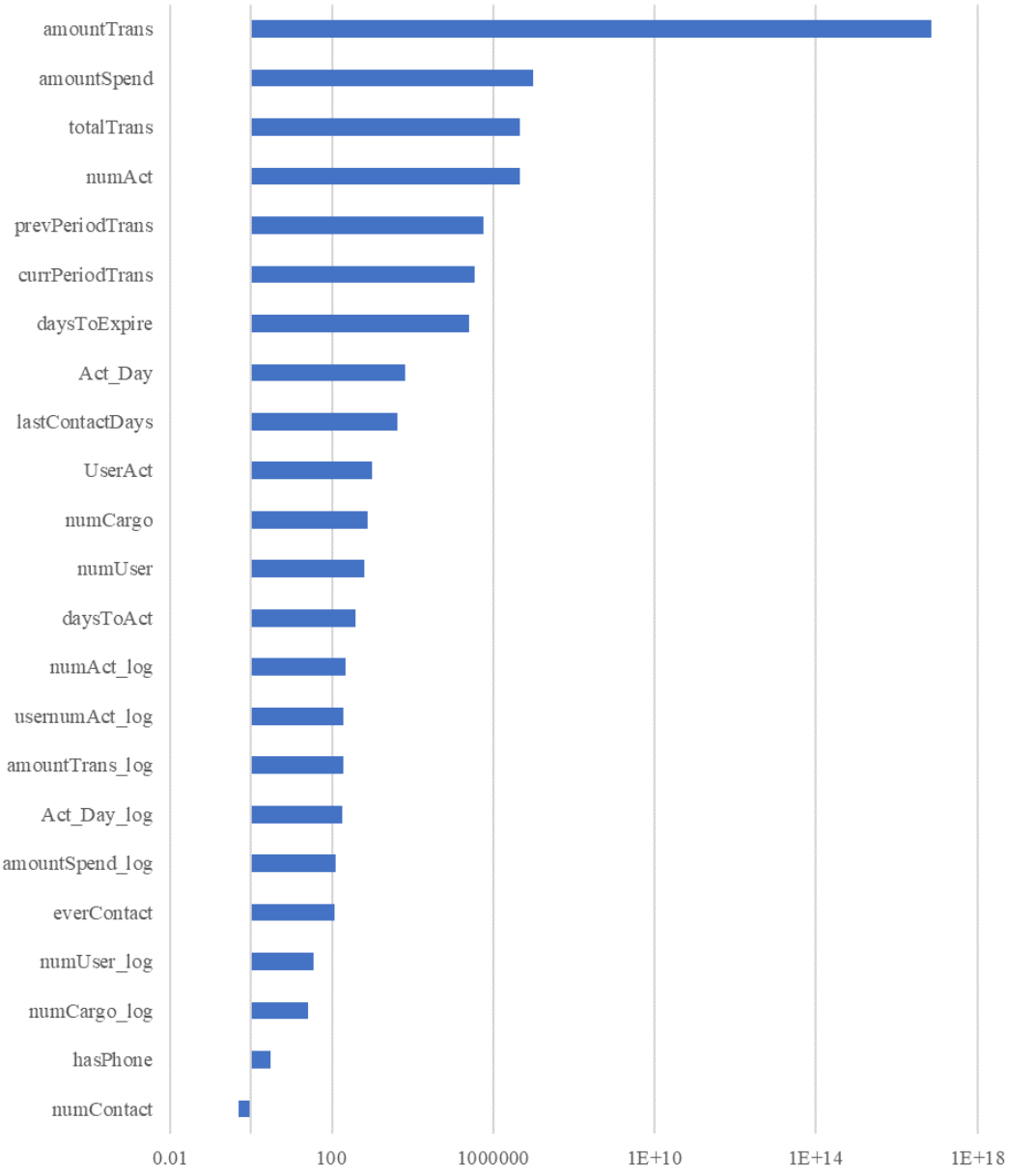


Figure 16. Feature importance by Chi-squared

4.3.Model Training

After data transformation had been executed, the transformed data which consisted of 1718 observations of 23 explanatory variables was separated into 2 portions: 80% of training dataset and 20% of testing dataset. It is noticeable that the split data was kept the ratio of churn data to be the same or as close as original dataset which are

approximately 54% of churn and 46% of non-churn. It is important to imitate the proportion of churn to training dataset to preserve the condition of distribution.

Next, this investigation had trained each model by varying the number of selected top important features from minimum features to maximum features for both ranking feature importance methods: ANOVA F-values and Chi-squared method. The machine learning classification models herein includes logistic regression, support vector machine, decision tree, and random forest. For each training, hyper-parameter tuning was achieved for a particular machine learning model by using grid search. Not only hyper-parameter tuning, the training also includes the k-fold cross validation (k=10) in order to validate the results of model training. The model training for each classification algorithm are given in follows:

4.3.1. Logistic Regression

Logistic regression algorithm is performed as the classifier. The logistic regression was trained with training dataset which had already done the data transformation. The logistic regression model was generated using Scikit-learn as mentioned in methodology section. Training model process for logistic regression is shown in Figure 17.

After training data with varying the number of selected features (n) as a certain parameter on logistic regression models, each logistic regression model with n-number selected features had been optimized toward recall with its own optimal value set of hyper-parameters whilst, as previously described, the hyper-parameters for logistic regression in which this paper is interested consists of 3 parameters; 1. penalty, 2. C, and 3. tol. the evaluation results applied grid search used for hyper-parameters tuning and k-fold cross validation (k =10) are shown in Figure 18 and Figure 19 And it is noted that the results of model are required to be at least 0.8 or 80% as a baseline in every dimension (Recall, F1-score, Accuracy, and Precision) regarding the case-study company.

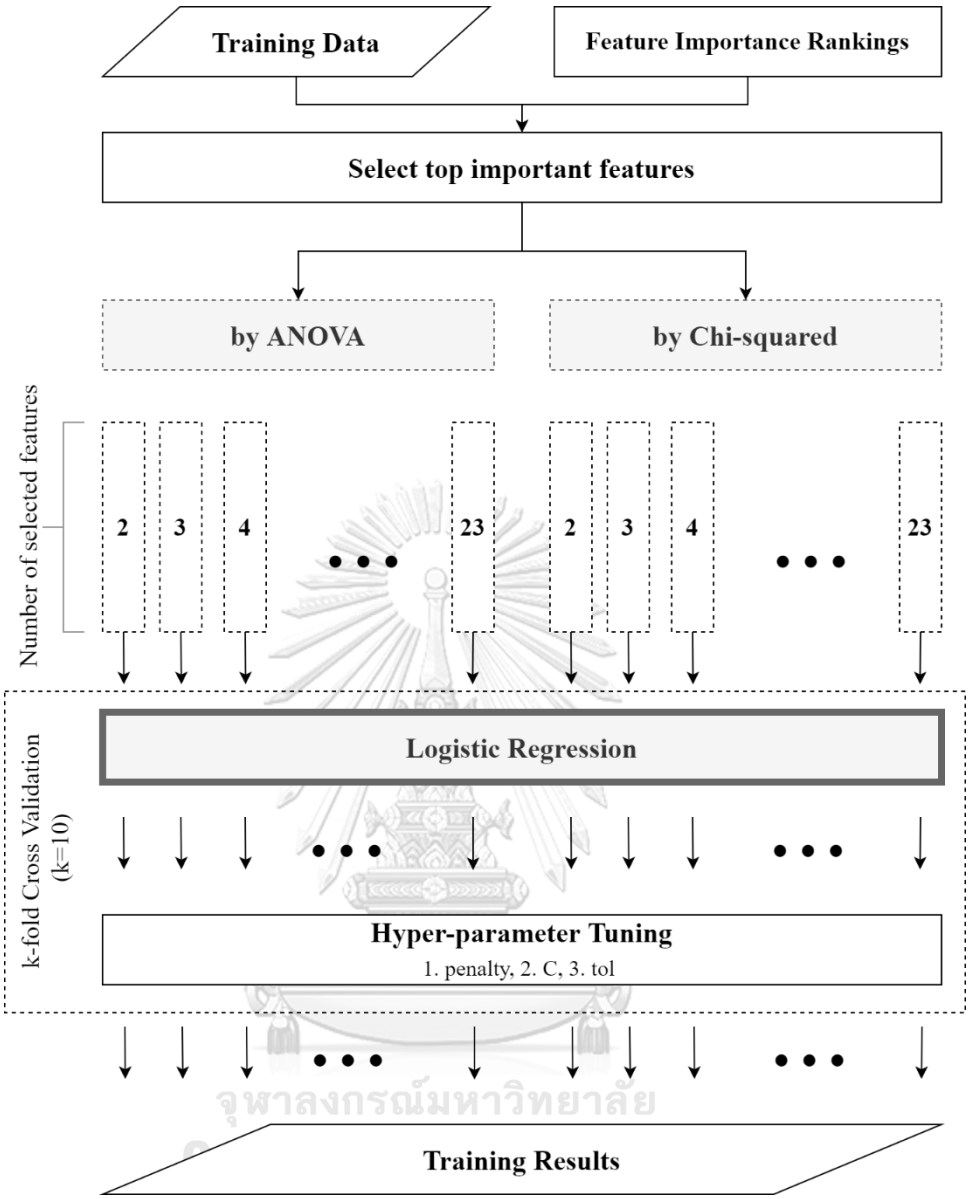


Figure 17. Training model process for logistic regression

The results of logistic regression optimized hyper-parameters and varied the number of feature importance (n) by ANOVA filter method in Figure 18 illustrate that every n-number selected feature can provide above the baseline requirement of 0.8 or 80% in every dimension (Recall, F1-score, Accuracy, and Precision). However, recall scores of every model which had not passed over 0.9 or 90% is slightly low compared to other dimension despite of that recall is the most preferable metric in this case. Regarding to the experimental results, the top-performance model of logistic regression with ANOVA filter method is the optimized logistic regression with top 21

important features (n=21) selected by ANOVA method. This model can provide recall of 86.8%, F1-score of 90.2%, accuracy of 89.6% and precision of 93.9%.

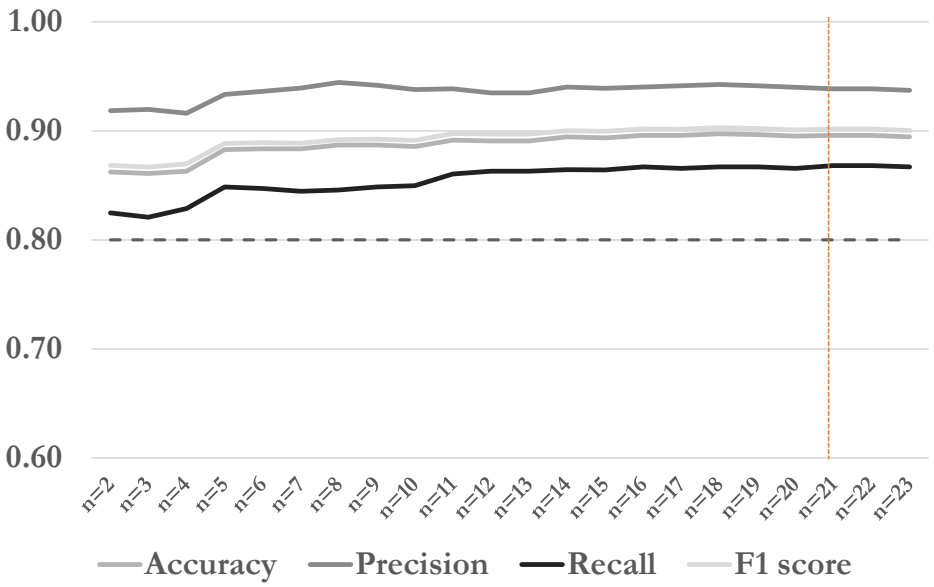


Figure 18. Results of Logistic Regression with ANOVA filter method.

Considering this optimized model of logistic regression with top 21 important features (n=21) selected by ANOVA method, grid search values for each hyper-parameter are presented in Table 13. And it also shows the results of the optimal values for each hyper-parameter using in the optimized model.

Table 13. Hyper-parameters of Logistic Regression with ANOVA filter method.

Hyper-parameter	**Grid Search Values**	**Optimal Value**
penalty	l1, l2	l2
C	1, 10, 100	10
tol	0.01,0.001,0.0001,0.00001	0.001

The results of logistic regression optimized hyper-parameters and varied the number of feature importance (n) by Chi-squared filter method in Figure 19 illustrate that

although, if the n-number selected features is lower than 7 features, recall score is steeply increased, F1-score, accuracy and precision is reversely plunged down and lower than the baseline requirement of 0.8 or 80%. This can be implied that those logistic regression models of n-number selected feature by Chi-squared method is not suitable to be applied in the case-study. While the top-performance model of logistic regression with Chi-squared filter method is the optimized logistic regression with top 22 important features (n=22) selected by Chi-squared method. This model can provide recall of 86.8%, F1-score of 90.2%, accuracy of 89.6% and precision of 93.9%, similar to the logistic regression model with ANOVA filter method. However, it still has the same issue with low recall score compare to other dimensions.

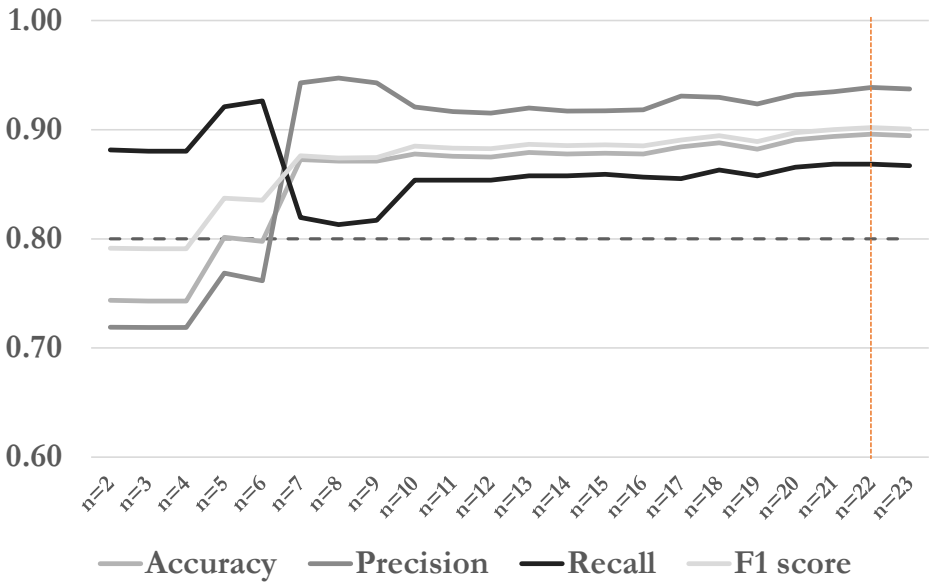


Figure 19. Results of Logistic Regression with Chi-squared filter method.

Focusing on the optimized model of logistic regression with top 22 important features (n=22) selected by Chi-squared method, grid search values for each hyper-parameter are presented in

Table 14. And it also shows the results of the optimal values for each hyper-parameter using in the optimized model.

Table 14. Hyper-parameters of Logistic Regression with Chi-squared filter method.

Hyper-parameter	Grid Search Values	Optimal Value
penalty	l1, l2	l2
C	1, 10, 100	10
tol	0.01,0.001,0.0001,0.00001	0.001

4.3.2. Support Vector Machine

Using the same training data as for logistic regression, classifier is set to be support vector machine algorithm. The support vector machine was trained with transformed training dataset completely cleaned and processed with encoding and standardization. The support vector machine model was created using Scikit-learn as described in methodology section. Training model process for support vector machine is shown in Figure 20.

After training data with varying the number of selected features (n) as a certain parameter on support vector machine models, each support vector machine model with n-number selected features had been optimized in accordance with recall score with its own optimal value set of hyper-parameters whilst, as was pointed out earlier, the hyper-parameters for support vector machine in which this paper is interested consists of 3 parameters; 1. kernel, 2. C, and 3. tol. the evaluation results applied grid search cross validation used for hyper-parameters tuning and k-fold cross validation (k =10) are shown in Figure 21 and

Figure 22. And it is mentioned that the results of model are required to be at least 0.8 or 80% as a baseline in every dimension (Recall, F1-score, Accuracy, and Precision) regarding the case-study company.

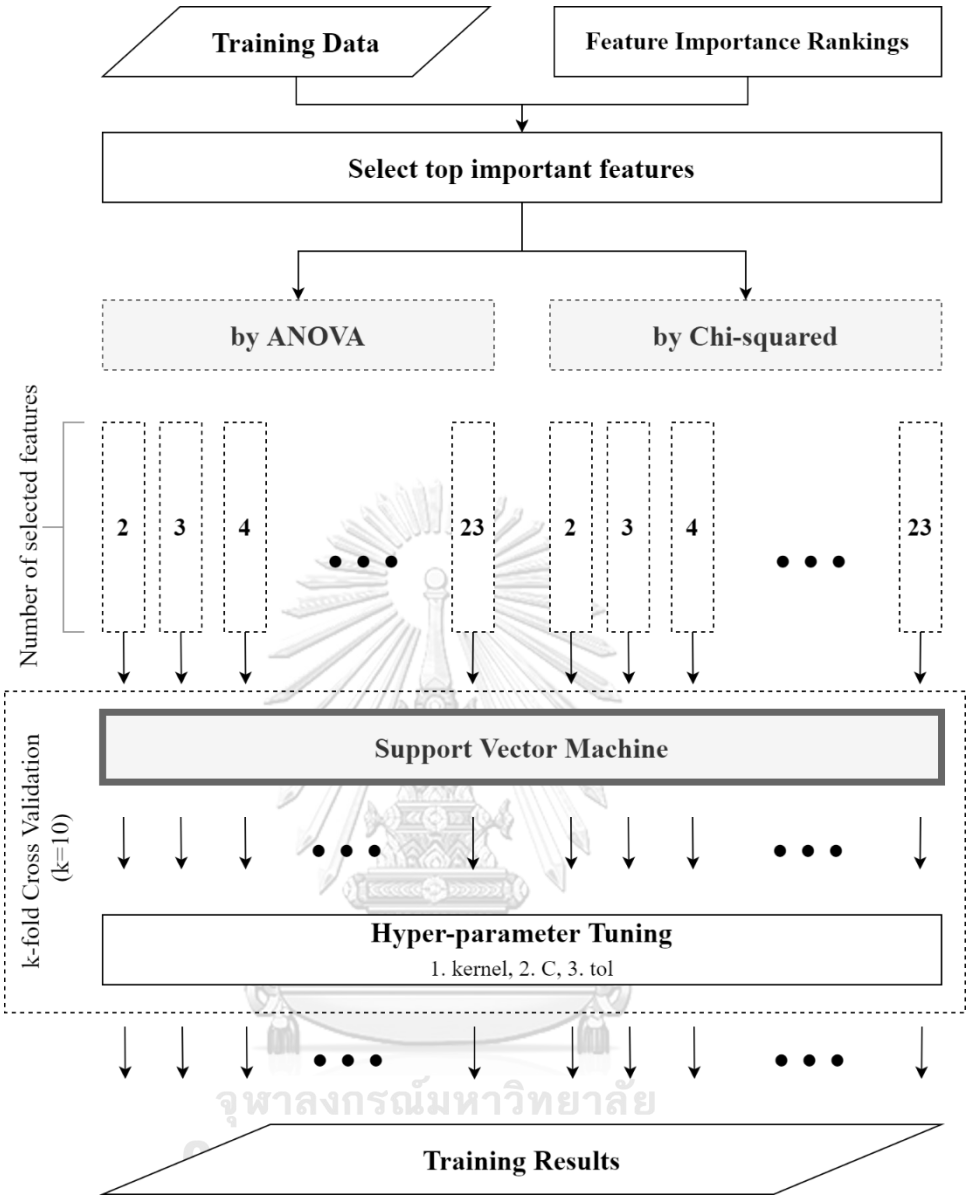


Figure 20. Training model process for support vector machine

The results of support vector machine optimized hyper-parameters and varied the number of feature importance (n) by ANOVA filter method in Figure 21 illustrate that every n-number selected feature can provide recall score above a 0.8 or 80% baseline. In spite of that, other dimensions (F1-score, accuracy and precision) are dramatically low. In a contrary, when F1-score, accuracy and precision get improved, recall score seems to be declined. Regarding to the experimental results, the top-performance model of support vector machine with ANOVA filter method is the optimized support vector machine with top 18 important features (n=18) selected by

ANOVA method. This model can provide recall of 88.1%, F1-score of 83.9%, accuracy of 81.4% and precision of 80.3% that meets the criteria of 0.8 or 80% in overall scores.

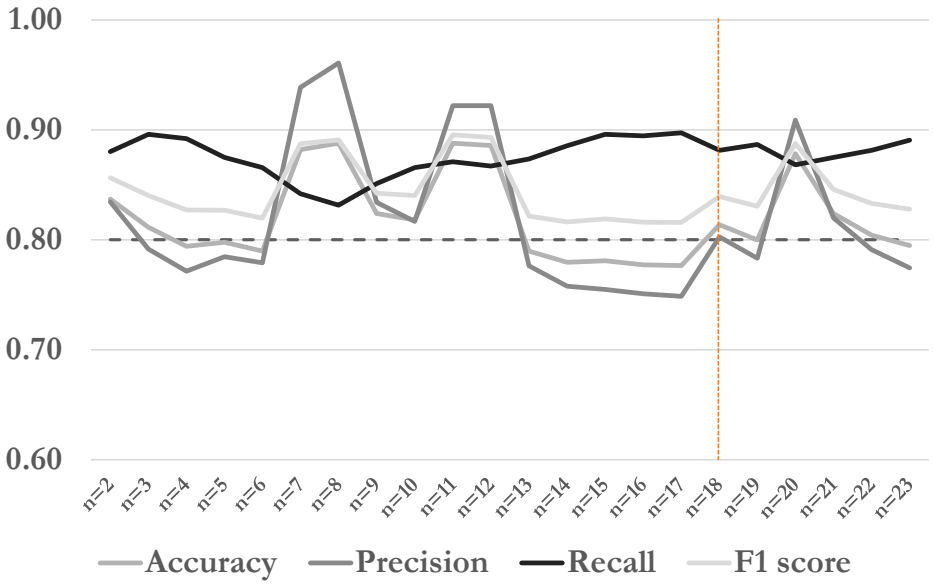


Figure 21. Results of Support Vector Machine with ANOVA filter method.

Considering this optimized model of support vector machine with top 18 important features (n=18) selected by ANOVA method, grid search values for each hyper-parameter are presented in Table 15. And it also shows the results of the optimal values for each hyper-parameter using in the optimized model.

Table 15. Hyper-parameters of Support Vector Machine with ANOVA filter method.

Hyper-parameter	**Grid Search Values**	**Optimal Value**
Kernel	linear, poly, rbf	poly
C	1, 10, 100	1
tol	0.01,0.001,0.0001,0.00001	0.001

The results of support vector machine optimized hyper-parameters and varied the number of feature importance (n) by Chi-squared filter method in

Figure 22 illustrate that although many n-number of selected feature by Chi-squared method can improve recall of the support vector machine model over 0.9 or 90% score, other dimensions (F1-score, accuracy and precision) are dropped sharply, especially lower than 10 selected features (n) that accuracy and precision scores dive to nearly 0.6 or 60% score. While the top-performance model of support vector machine with Chi-squared filter method is the optimized support vector machine with top 19 important features (n=19) selected by Chi-squared method. This model can provide recall of 88.0%, F1-score of 90.0%, accuracy of 89.2% and precision of 92.3%. This recall score is almost equivalent to the recall of the support vector machine model with ANOVA filter method. Nevertheless, other dimensions of support vector machine model with Chi-squared filter method are significantly higher than support vector machine with ANOVA method.

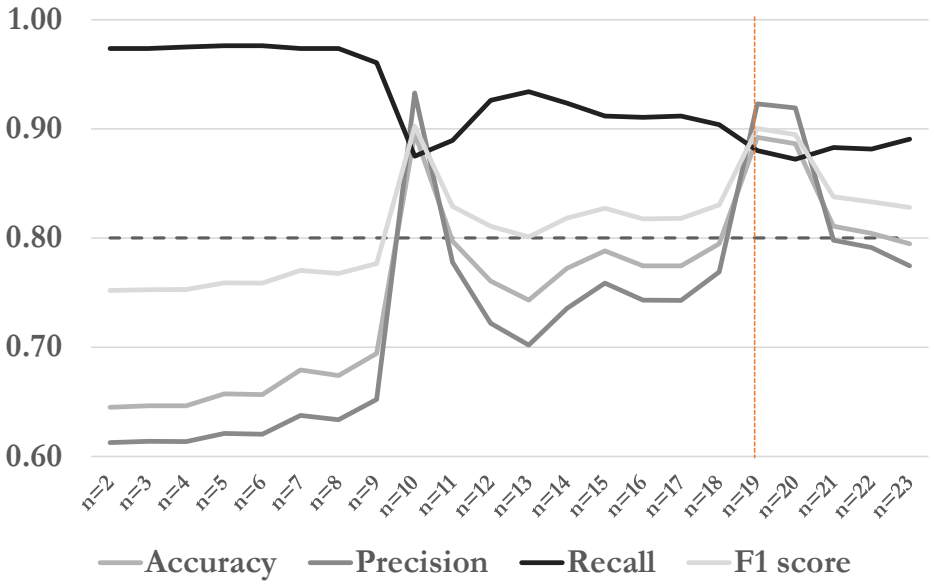


Figure 22. Results of Support Vector Machine with Chi-squared filter method

Focusing on the optimized model of support vector machine with top 19 important features (n=19) selected by Chi-squared method, grid search values for each hyper-parameter are presented in Table 16. And it also shows the results of the optimal values for each hyper-parameter using in the optimized model.

Table 16. Hyper-parameters of Support Vector Machine with Chi-squared filter method.

Hyper-parameter	Grid Search Values	Optimal Value
Kernel	linear, poly, rbf	poly
C	1, 10, 100	100
tol	0.01,0.001,0.0001,0.00001	0.001

4.3.3. Decision Tree

Similar to support vector machine, decision tree algorithm is performed as the classifier. The decision tree was trained with training dataset which had already done the data transformation. The decision tree model was built using Scikit-learn as explained earlier in methodology section. Training model process for decision tree is shown in Figure 23.

After training data with varying the number of selected features (n) as a certain parameter on decision tree models, each decision tree model with n-number selected features had been optimized considering recall as objective scorer with its own optimal value set of hyper-parameters whilst, as was mentioned previously, the hyper-parameters for decision tree on which this paper focused consists of 5 parameters; 1. criterion, 2. splitter, 3. max_depth, 4. min_samples_split, and 5. min_samples_leaf. the evaluation results applied grid search cross validation used for hyper-parameters tuning and k-fold cross validation (k =10) are shown in Figure 24 and Figure 25. And it is noted that the results of model are required to be at least 0.8 or 80% as a baseline in every dimension (Recall, F1-score, Accuracy, and Precision) regarding the case-study company.

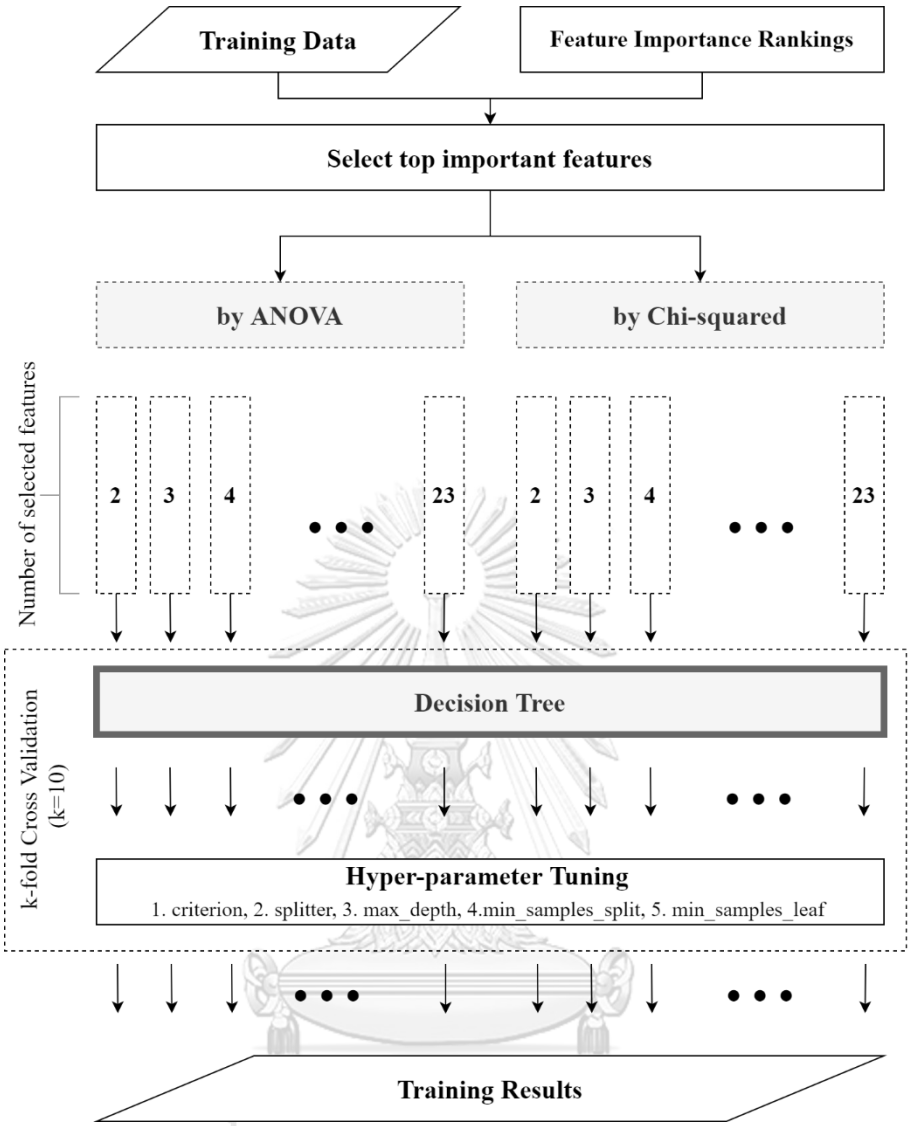


Figure 23. Training model process for decision tree

The results of decision tree optimized hyper-parameters and varied the number of feature importance (n) by ANOVA filter method in Figure 24 illustrate that this decision tree model with every n-number selected feature by ANOVA filter method can provide evaluation scores above the 0.8 or 80% baseline requirement in every dimension (Recall, F1-score, Accuracy, and Precision). Every n-number selected feature can provide above the baseline requirement of 0.8 or 80% in every dimension (Recall, F1-score, Accuracy, and Precision). However, recall scores of every model which had not passed over 0.9 or 90% is slightly low compared to other dimension

despite of that recall is the most preferable metric in this case. The recall scores seem to be slightly increased when more n-number selected features.

However, the decision tree model with 15 selected features (n=15) by ANOVA method is outperformed the others in recall score. As a result, the top-performance model of decision tree with ANOVA filter method is the optimized decision tree with top 15 important features (n=15) selected by ANOVA method. This model can provide recall of 93.2%, F1-score of 87.6%, accuracy of 85.4% and precision of 82.7%.

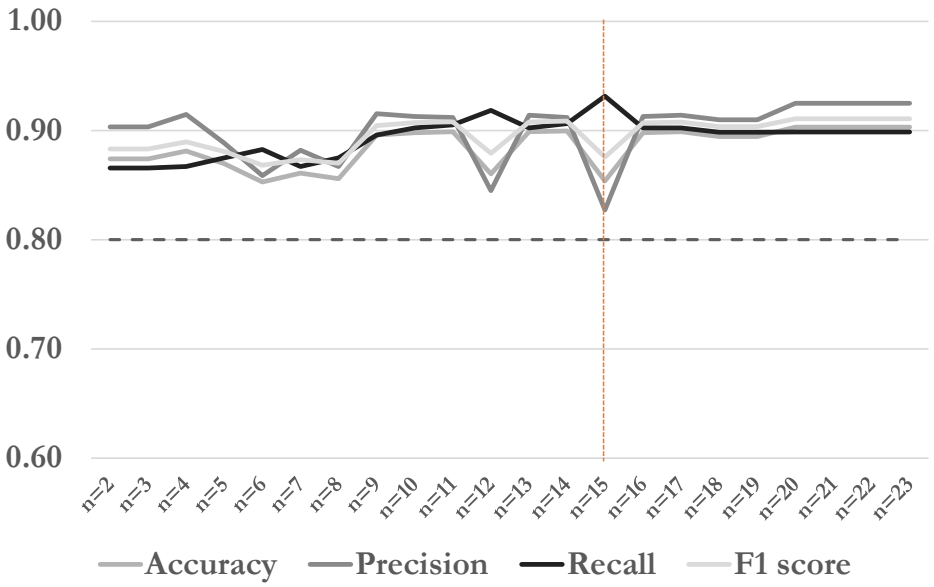


Figure 24. Results of Decision Tree with ANOVA filter method.

Considering this optimized model of decision tree with top 15 important features (n=15) selected by ANOVA method, grid search values for each hyper-parameter are presented in

The results of decision tree optimized hyper-parameters and varied the number of feature importance (n) by Chi-squared filter method in Figure 25 illustrate that although the decision tree models with 2 to 4 selected features (n<5) by Chi-squared method cannot reach the criteria of a baseline requirement of 0.8 or 80% score in F1-score, accuracy and precision, the models with above 5 n-number selected features

(n>4) by Chi-squared method are able to satisfy the baseline criteria in every evaluation dimension (Recall, F1-score, Accuracy, and Precision).

Table 17. And it also shows the results of the optimal values for each hyper-parameter using in the optimized model.

The results of decision tree optimized hyper-parameters and varied the number of feature importance (n) by Chi-squared filter method in Figure 25 illustrate that although the decision tree models with 2 to 4 selected features (n<5) by Chi-squared method cannot reach the criteria of a baseline requirement of 0.8 or 80% score in F1-score, accuracy and precision, the models with above 5 n-number selected features (n>4) by Chi-squared method are able to satisfy the baseline criteria in every evaluation dimension (Recall, F1-score, Accuracy, and Precision).

Table 17. Hyper-parameters of Decision Tree with ANOVA filter method.

Hyper-parameter	Grid Search Values	Optimal Value
criterion	gini, entropy	entropy
splitter	best, random	random
max_depth	[1-10], None	3
min_samples_split	2, 4, 6, 8, 10	2
min_samples_leaf	[1-10]	1

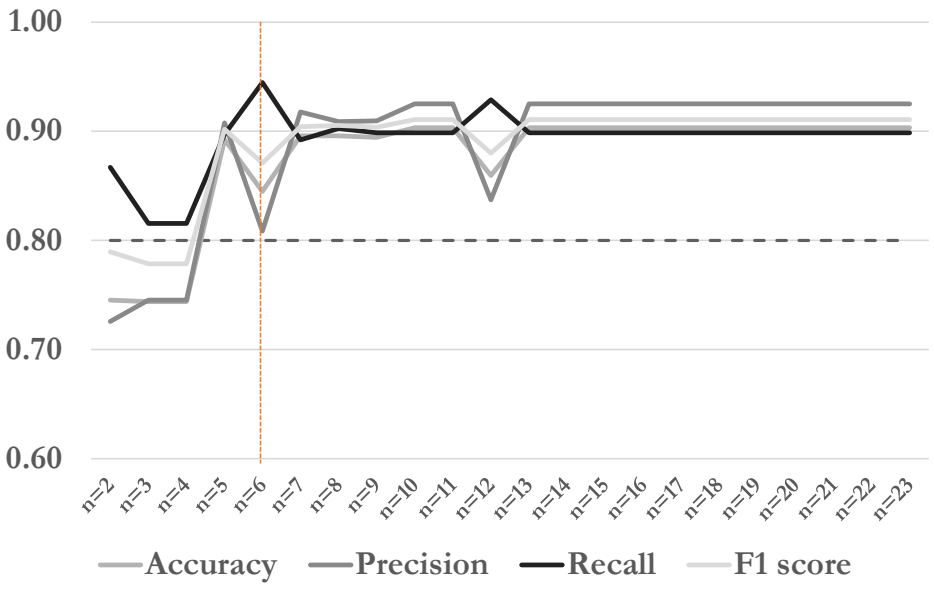


Figure 25. Results of Decision Tree with Chi-squared filter method

Regarding recall score, the top-performance model of decision tree with Chi-squared filter method is the optimized decision tree with top 6 important features (n=6) selected by Chi-squared method. This model can provide recall of 94.5%, F1-score of 87.1%, accuracy of 84.5% and precision of 80.9%.

Focusing on the optimized model of decision tree with top 6 important features (n=6) selected by Chi-squared method, grid search values for each hyper-parameter are presented in Table 18. And it also shows the results of the optimal values for each hyper-parameter using in the optimized model.

Table 18. Hyper-parameters of Decision Tree with Chi-squared filter method.

Hyper-parameter	Grid Search Values	Optimal Value
criterion	gini, entropy	gini
splitter	best, random	random
max_depth	[1-10], None	3
min_samples_split	2, 4, 6, 8, 10	2

min_samples_leaf	[1-10]	1

4.3.4. Random Forest

Same as decision tree, random forest algorithm is treated as the classifier. The random forest model was trained with the same training dataset which had already done the data transformation (data cleaning and data processing). The random forest model was built using Scikit-learn as was previously pointed out in methodology section. Training model process for random forest is shown in Figure 26.

After training data with varying the number of selected features (n) as a particular parameter on logistic regression models, each logistic regression model with n-number selected features had been optimized toward the purpose of recall scorer with its own optimal value set of hyper-parameters whilst, as stated earlier, the hyper-parameters for random forest on which this paper focused consists of 5 parameters; 1. n_estimaters, 2. criterion, 3. max_depth, 4. min_samples_split, and 5. min_samples_leaf. the evaluation results applied grid search cross validation used for hyper-parameters tuning and k-fold cross validation (k =10) are shown in Figure 27 and Figure 28. And it is noted that the results of model are required to be at least 0.8 or 80% as a baseline in every dimension (Recall, F1-score, Accuracy, and Precision) regarding the case-study company.

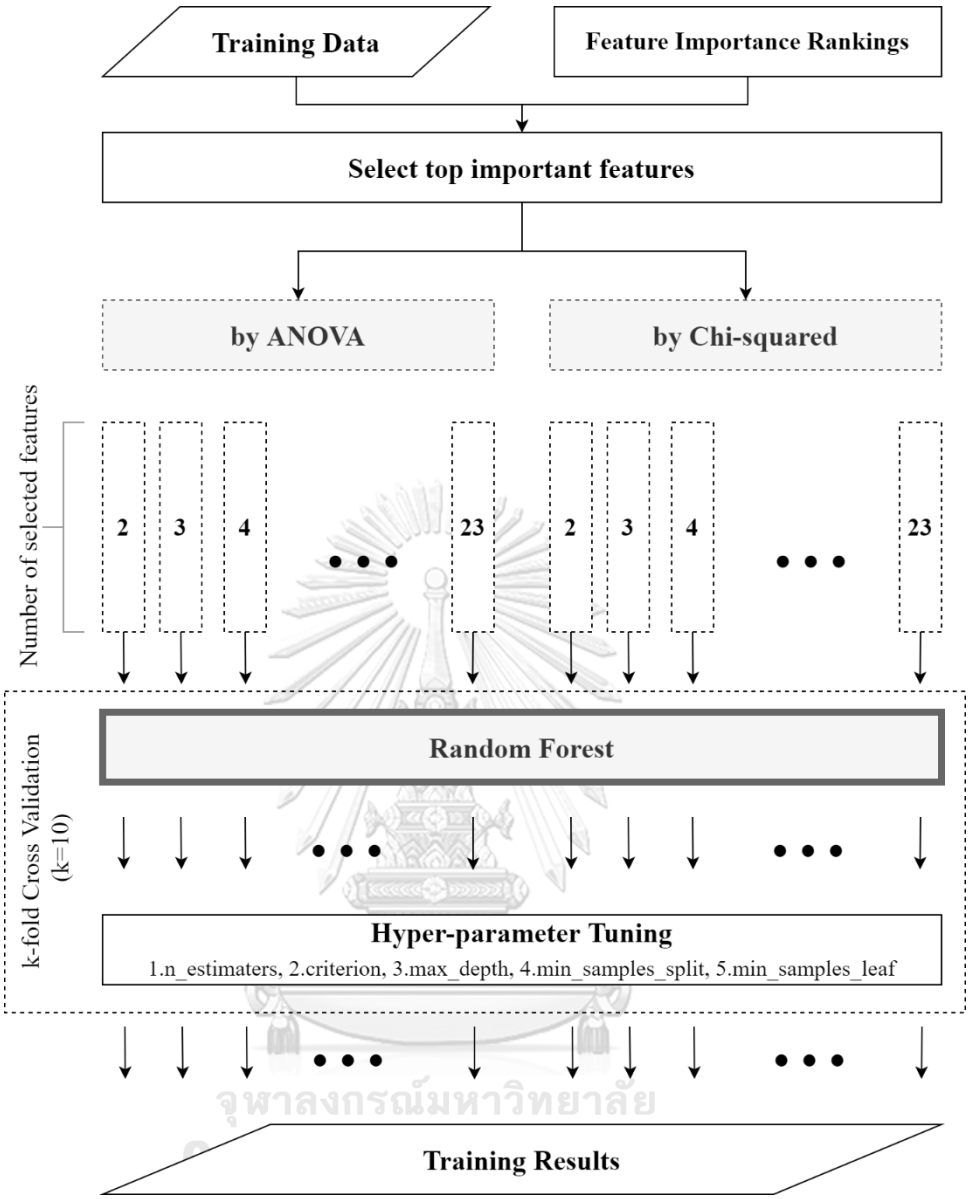


Figure 26. Training model process for random forest

The results of random forest optimized hyper-parameters and varied the number of feature importance (n) by ANOVA filter method in Figure 27 illustrate that every n-number selected feature can satisfy the requirement of 0.8 or 80% score in every evaluation dimension (Recall, F1-score, Accuracy, and Precision). Even though the random forest model with more than 9 number selected features (n>9) by ANOVA method tends to have steadily outcomes and the values of scores close to each other, the top recall score model of random forest with ANOVA filter method is the optimized random forest with top 21 important features (n=21) selected by ANOVA

method. This model can provide recall of 91.4%, F1-score of 92.6%, accuracy of 91.9% and precision of 93.9% that is outstanding with over 0.9 or 90% of scores close to each other. Noticeably, the models with more than 4-number selected features (n>4) by ANOVA method do not only meet the criteria but these models also perform over 0.9 or 90% score in every evaluation dimension.

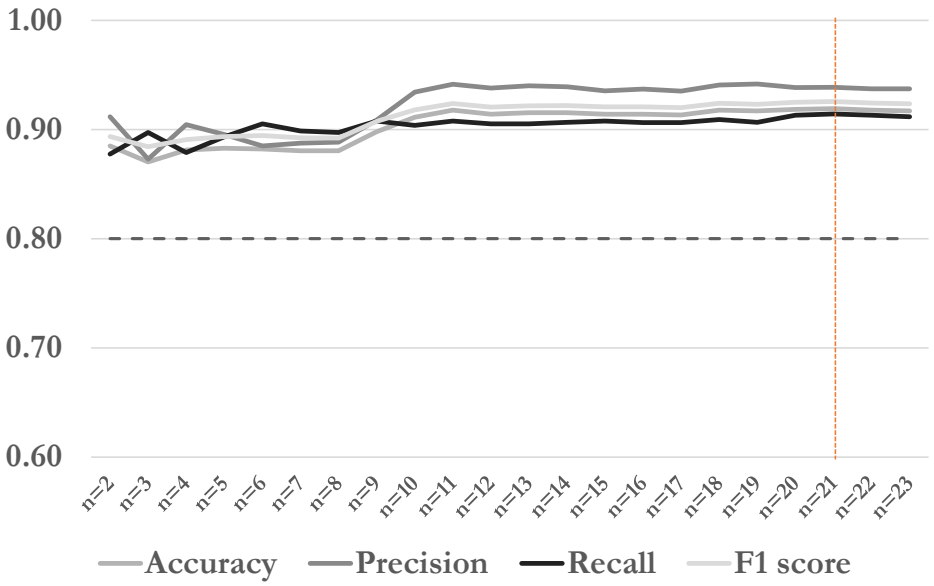


Figure 27. Results of Random Forest with ANOVA filter method.

Considering this optimized model of random forest with top 21 important features (n=21) selected by ANOVA method, grid search values for each hyper-parameter are presented in

Table 19. And it also shows the results of the optimal values for each hyper-parameter using in the optimized model.

The results of logistic regression optimized hyper-parameters and varied the number of feature importance (n) by Chi-squared filter method in Figure 28 illustrate that every n-number selected feature can provide above the baseline requirement of 0.8 or 80% in every dimension (Recall, F1-score, Accuracy, and Precision) except 2, 3 and 4 number selected features (n<5). The random forest models with more than 4-number selected features (n>4) by ANOVA method seem to provide steadily outcomes and the values of scores close to each other.

Table 19. Hyper-parameters of Random Forest with ANOVA filter method.

Hyper-parameter	Grid Search Values	Optimal Value
n_estimators	10, 20, 40, 60, 80, 100, 200	200
criterion	gini, entropy	gini
max_depth	[1-10], None	None
min_samples_split	2, 4, 6, 8, 10	2
min_samples_leaf	[1-10]	1

Noticeably, the models with more than 4-number selected features (n>4) by ANOVA method do not only meet the criteria but these models also perform over 0.9 or 90% score in every evaluation dimension.

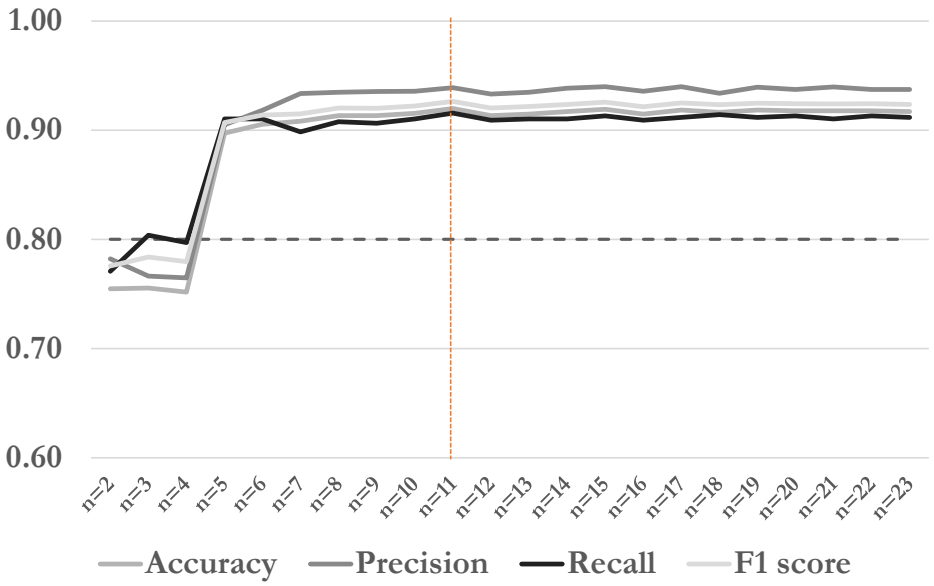


Figure 28. Results of Random Forest with Chi-squared filter method

Considering the best recall score, the top-performance model of random forest with Chi-squared filter method is the optimized random forest with top 11 important features (n=11) selected by Chi-squared method. This model can provide recall of

91.6%, F1-score of 92.6%, accuracy of 92.0% and precision of 93.9%, be slightly better scores comparing to the random forest model with ANOVA filter method.

Focusing on the optimized model of random forest with top 11 important features (n=11) selected by Chi-squared method, grid search values for each hyper-parameter are presented in Table 20. And it also shows the results of the optimal values for each hyper-parameter using in the optimized model.

Table 20. Hyper-parameters of Random Forest with Chi-squared filter method.

Hyper-parameter	Grid Search Values	Optimal Value
n_estimators	10, 20, 40, 60, 80, 100, 200	100
criterion	gini, entropy	gini
max_depth	[1-10], None	None
min_samples_split	2, 4, 6, 8, 10	2
min_samples_leaf	[1-10]	1

4.3.5. Result Comparison

Every training result in details for each prediction model are shown in Appendix 2. Considering the top-performance of each model classification with different filter methods, these top- performance models results are presented in Table 21. top-performance models are listed herein including 8 models as the following:

1. The optimized logistic regression with top 21 important features selected by ANOVA method (LR ANOVA n=21)
2. The optimized model of logistic regression with top 22 important features selected by Chi-squared method (LR Chi2 n=22)
3. The optimized model of support vector machine with top 18 important features selected by ANOVA method (SVM ANOVA n=18)

4. The optimized model of support vector machine with top 19 important features selected by Chi-squared method (SVM Chi2 n=19)
5. The optimized model of decision tree with top 15 important features selected by ANOVA method (DT ANOVA n=15)
6. The optimized model of decision tree with top 6 important features selected by Chi-squared method (DT Chi2 n=6)
7. The optimized model of random forest with top 21 important features selected by ANOVA method (RF ANOVA n=21)
8. The optimized model of random forest with top 11 important features selected by Chi-squared method (RF Chi2 n=11).

Table 21. Result comparison

Models	**Evaluation metrics**			
	Accuracy	**Precision**	**Recall**	**F1-score**
LR ANOVA n=21	0.896	0.939	0.868	0.902
LR Chi2 n=22	0.896	0.939	0.868	0.902
SVM ANOVA n=18	0.814	0.803	0.881	0.839
SVM Chi2 n=19	0.892	0.923	0.880	0.900
DT ANOVA n=15	0.854	0.827	0.932	0.876
DT Chi2 n=6	0.845	0.809	0.945	0.871

Models	Evaluation metrics			
	Accuracy	**Precision**	**Recall**	**F1-score**
RF ANOVA n=21	0.919	0.939	0.914	0.926
RF Chi2 n=11	0.920	0.939	0.916	0.926

From result comparison, the optimized decision tree model with top 6 important features selected by Chi-squared method outperform other models regarding the objective towards recall scorer with recall of 94.5%, F1-score of 87.1%, accuracy of 84.5% and precision of 80.9%. Even though the optimized decision tree model with top 6 important features selected by Chi-squared method is the best models in term of recall score, it is noticeable that in term of overall evaluation results, the optimized random forest models both with top 21 important features selected by ANOVA method and with top 11 important features selected by Chi-squared method are capable to provide over 0.9 or 90% scores in every dimension instead of focusing on only recall.

4.5.Model Testing

As was mentioned earlier, the transformed data which consisted of 1718 observations of 23 explanatory variables was separated into 2 portions: 80% of training dataset and 20% of testing dataset. The first 80% portion was used in model training in previous section. Another portion of 20% was separated to be used in model testing to confirm the model validation. It is the same as training data that the data was kept the ratio of churn data to be equivalent to original data which approximately consists of 54% churn and 46% non-churn.

As a result of classifiers comparison in previous step, the model that performed outstandingly apart from other algorithms is the optimized decision tree model with top 6 important features selected by Chi-squared method. Therefore, to confirm the performance of the final model, this model is necessary to be investigated holdout cross validation with firstly separated testing dataset. The testing results are shown in Table 22. In testing result, the optimized decision tree model with top 6 important features selected by Chi-squared method is capable to perform similarly to the result of cross validation training dataset with 94.4% of recall and 88.2% of F1-score, 85.8% of accuracy and 82.9% of precision that can also satisfy baseline criteria of 0.8 or 80% score in every evaluation metrics (Recall, F1-score, Accuracy, and Precision). It noted that every testing result in details for every prediction model are shown in Appendix 3.

Table 22. Testing result of the optimized decision tree model with top 6 important features selected by Chi-squared method.

DT Chi2 n=6	**Evaluation metrics**			
	Accuracy	**Precision**	**Recall**	**F1-score**
Training Result	0.845	0.809	0.945	0.871
Testing Result	0.858	0.829	0.944	0.882

Since these evaluation metrics are generally originated from the calculation of confusion matrix which describes the outcome between predications and actuals in a binary classification as was mentioned previously in methodology, Table 23 shows details in confusion matrix of the testing result of final model, the optimized decision tree model with top 6 important features selected by Chi-squared method, which indicated in 4 terms: true positive (TP), false positive (FP), false negative (FN) and true negative (TN) as mentioned earlier in evaluation metric.

Table 23. Confusion matrix of the holdout testing result of the optimized decision tree model with top 6 important features selected by Chi-squared method.

		Predicted	
		Negative (Non-churn)	**Positive (Churn)**
Actual	**Negative (Non-churn)**	111	38
	Positive (Churn)	11	184

In addition to evaluate the classification model, the final model can also describe the related features. Table 24 shows the summary of the top 6 important features sorted by feature importance scores which derived from the final classification model. Overall, the important features impacting to the prediction model are mainly related to business features like transaction and usage frequency.

Table 24. Top 6 important features in the final model.

Rank	Attribute Name	Description
1	amountTrans	Amount of transaction values via platform
2	amountSpend	Amount of customer spending
3	numAct	Number of actions per customer

4	totalTrans	Number of all transactions
5	currPeriodTrans	Number of transactions in current period of 14 days
6	prevPeriodTrans	Number of transactions in previous period of 14 days

Chapter 5 Conclusion and Future Work

5.1. Conclusions

In SaaS industry, the market is continuously expanding due to the global economic growth as well as inventory management software market. This case-study, SaaS inventory management software company in Thailand, is also expected to capture the trends and scale up their market. However, the company is facing a high customer churn rate problem. Developing the solution models which improves the customer churn rate becomes an important issue.

This paper found the most effective customer churn prediction model in term of recall as scorer is the optimized decision tree model using top 6 important features selected by Chi-squared filter method. It shows the capability in prediction with the highest average training recall score and being above case-study baseline criteria of 0.8 or 80% score in every evaluation metrics. The training result can score up to 94.5% recall, 87.1% F1-score, 84.5% accuracy and 80.9% precision. Additionally, the holdout testing result of this model is also validated its performance in customer churn prediction with 94.4% of recall and 88.2% of F1-score, 85.8% of accuracy and 82.9% of precision. Since this final prediction model of decision tree classifier was optimized by hyper-parameters tuning, the optimal values for each hyper-parameter using grid search cross validation including with criterion, splitter, max_dept, min_samples_split and min_samples_leaf, are provided 1. gini criterion, 2. Random splitter, 3. 3 of max_dept, 4. 2 of min_samples_split and 5. 1 of min_samples_leaf, respectively. Hence this final model is enough to satisfy the case-study objective indicating real churn customer from all churn customer correctly and perform better than others prediction models. It noted that these optimal values of hyper-parameters are considered among specific sets of hyper-parameters which are only subset of every possible hyper-parameters. Moreover, the classification model provides details in attributes or important futures that related to customer churn. these contributions of feature importance scores provide beneficial insights to the case-study company. It highlights that business metrics such as transactions are the top feature importance. It

can also imply that if customer is online more frequent and creates more amount of transactions via the platform, that customer is less likely to churn. The final model, the optimized decision tree model using top 6 important features selected by Chi-squared filter method, shows that churn prediction can concentrate on only these 6 important figures.

However, it is noticeable that the final model proposed in this paper is under the objective of case-study to receive the best recall score while others are needed to only be above 0.8 or 80% baseline. In case of overall performance, it is recommended that the optimized random forest models both with top 21 important features selected by ANOVA method and with top 11 important features selected by Chi-squared method are more suitable to be properly used as the churn classification prediction model with over 0.9 or 90% in all dimensions according to the results. Thus, if the case-study company found that the prediction model needs to be improved in other evaluation dimensions in order to meet new criteria in the future, these optimized random forest models would be able to perform properly as well.

In conclusion, this paper provides the final model of customer churn prediction and features importance list that is able to satisfy the objective of this paper and also the case-study company requirements. As result, the case-study company will be able to indicate the right risky churn customers by using the final customer churn prediction model and handle them with effective marketing campaigns. Also, this paper can enhance the efficiency and effectiveness of managerial decision for case-study company.

5.2. Limitations for this research

Due to limitation of case-study company's information management system, most of available features had been exposed in business-related feature usage area, e.g., transactions and number of various types of usages while these data still lack of quality metrics like service rating or customer satisfaction in many dimensions and also internal information like outage incidents, software errors. These all kind of data can enhance the prediction model to understand further dimensions of case-study

related factors. Considering the current situation, The COVID-19 pandemic has completely changed customer related factors, e.g., customer behavior, financial conditions. The prediction models from this thesis have not applied with the current updated information yet. Therefore, the results might be different which should be significantly concerned.

In case of feature selection process, this thesis applied feature selection method with whole data after data processing while these features should be generally ranked by using only training data. The feature importance scores by ANOVA filter method and Chi-squared method using training data (shown in Appendix 4) are not as exactly same as the ranking using the whole data in calculation. As a result, this difference changes the ranking position of some features that could cause performance of prediction models. Therefore, this specification of the feature ranking step in this thesis should take in consideration before applying the prediction model from this thesis to the real data.

In addition to cross validation, although this paper applies both holdout method and k-fold cross validation to confirm the performance of prediction models, the final prediction model should also be tested with the updated actual data from the case-study company in order to satisfy practically the target of the company in practice.

5.3. Future work

Although this paper provides the customer churn prediction model satisfied the target in churn customer identification, it can be extended to forecast the number of customers, the future transaction and also the revenue of the company[43]. Due to the nature of SaaS business model as subscription base, customer churn is potentially variated to the projected future revenues. As a result of that extensions, the company's executives can practically apply this more concrete information like financial data in a managerial decision.

Appendix 1

Table 25. Feature importance score by ANOVA filter method using whole data

Ranking	Attribute Name	Score
1	amountSpend_log	800.23
2	daysToExpire	732.69
3	amountSpend	513.00
4	numAct_log	316.47
5	Act_Day_log	238.86
6	usernumAct_log	238.08
7	everContact	228.00
8	amountTrans_log	175.33
9	prevPeriodTrans	108.24
10	currPeriodTrans	107.69
11	numUser_log	96.12
12	totalTrans	86.99
13	numAct	86.97
14	Act_Day	62.26
15	numUser	61.66
16	numCargo_log	43.45
17	hasPhone	22.15
18	lastContactDays	9.27
19	numCargo	7.05
20	UserAct	3.63
21	daysToAct	0.85
22	amountTrans	0.80
23	numContact	0.49

Table 26. Feature importance score by Chi-squared filter method using whole data

Ranking	Attribute Name	Score
1	amountTrans	73837300000000000.00
2	amountSpend	9864920.00
3	totalTrans	4537420.00
4	numAct	4536180.00
5	prevPeriodTrans	570715.00
6	currPeriodTrans	357923.00
7	daysToExpire	253610.00
8	Act_Day	6760.09
9	lastContactDays	4195.99
10	UserAct	991.10
11	numCargo	757.35
12	numUser	642.46
13	daysToAct	384.33
14	numAct_log	220.34
15	usernumAct_log	195.38
16	amountTrans_log	195.17
17	Act_Day_log	186.97
18	amountSpend_log	121.85
19	everContact	119.63
20	numUser_log	36.38
21	numCargo_log	26.56
22	hasPhone	2.98
23	numContact	0.49

Appendix 2

Table 27. Training results in each evaluation metrics for Logistic Regression with ANOVA filter method.

Number of Selected Features	Evaluation metrics			
	Accuracy	**Precision**	**Recall**	**F1-score**
n=2	0.862	0.919	0.825	0.869
n=3	0.861	0.920	0.821	0.867
n=4	0.863	0.916	0.829	0.870
n=5	0.883	0.934	0.849	0.889
n=6	0.884	0.936	0.847	0.889
n=7	0.884	0.939	0.845	0.889
n=8	0.887	0.944	0.846	0.892
n=9	0.887	0.942	0.849	0.892
n=10	0.886	0.938	0.850	0.891
n=11	0.892	0.939	0.860	0.898
n=12	0.891	0.935	0.863	0.897
n=13	0.891	0.935	0.863	0.897
n=14	0.894	0.940	0.864	0.900
n=15	0.894	0.939	0.864	0.900
n=16	0.896	0.940	0.867	0.902
n=17	0.896	0.942	0.866	0.902
n=18	0.897	0.943	0.867	0.903
n=19	0.897	0.941	0.867	0.902
n=20	0.895	0.940	0.866	0.901
n=21	0.896	0.939	0.868	0.902
n=22	0.896	0.939	0.868	0.902
n=23	0.894	0.937	0.867	0.900

Table 28. Standard Deviation of training results in each evaluation metric for Logistic Regression with ANOVA filter method.

Number of Selected Features	Standard Deviation			
	Accuracy	**Precision**	**Recall**	**F1-score**
n=2	0.028	0.028	0.041	0.028
n=3	0.027	0.028	0.042	0.028
n=4	0.021	0.023	0.034	0.021
n=5	0.022	0.017	0.039	0.023
n=6	0.021	0.020	0.034	0.021
n=7	0.019	0.019	0.041	0.020
n=8	0.023	0.019	0.038	0.023
n=9	0.023	0.024	0.034	0.022
n=10	0.021	0.021	0.035	0.021
n=11	0.022	0.024	0.029	0.021
n=12	0.021	0.022	0.030	0.020
n=13	0.021	0.022	0.030	0.020
n=14	0.023	0.021	0.037	0.023
n=15	0.022	0.024	0.034	0.021
n=16	0.022	0.019	0.035	0.022
n=17	0.022	0.018	0.036	0.022
n=18	0.025	0.019	0.034	0.024
n=19	0.024	0.018	0.036	0.024
n=20	0.023	0.016	0.036	0.023
n=21	0.026	0.020	0.036	0.026
n=22	0.026	0.020	0.036	0.026
n=23	0.029	0.022	0.038	0.028

Table 29. Training results in each evaluation metrics for Logistic Regression with Chi-squared filter method.

Number of Selected Features	Evaluation metrics			
	Accuracy	**Precision**	**Recall**	**F1-score**
n=2	0.744	0.719	0.881	0.791
n=3	0.743	0.719	0.880	0.791
n=4	0.743	0.719	0.880	0.791
n=5	0.801	0.768	0.921	0.837
n=6	0.798	0.762	0.926	0.835
n=7	0.873	0.943	0.820	0.876
n=8	0.871	0.947	0.813	0.874
n=9	0.871	0.943	0.817	0.874
n=10	0.878	0.921	0.854	0.885
n=11	0.876	0.917	0.854	0.883
n=12	0.875	0.915	0.854	0.883
n=13	0.879	0.920	0.858	0.887
n=14	0.878	0.917	0.858	0.885
n=15	0.878	0.917	0.859	0.886
n=16	0.878	0.918	0.856	0.885
n=17	0.884	0.931	0.855	0.891
n=18	0.888	0.930	0.863	0.895
n=19	0.882	0.924	0.858	0.889
n=20	0.891	0.932	0.866	0.897
n=21	0.894	0.935	0.868	0.900
n=22	0.896	0.939	0.868	0.902
n=23	0.894	0.937	0.867	0.900

Table 30. Standard Deviation of training results in each evaluation metric for Logistic Regression with Chi-squared filter method.

Number of Selected Features	Standard Deviation			
	Accuracy	**Precision**	**Recall**	**F1-score**
n=2	0.033	0.027	0.048	0.028
n=3	0.036	0.029	0.047	0.030
n=4	0.036	0.029	0.047	0.030
n=5	0.038	0.040	0.026	0.027
n=6	0.031	0.034	0.026	0.022
n=7	0.030	0.024	0.047	0.031
n=8	0.028	0.030	0.047	0.029
n=9	0.032	0.026	0.052	0.033
n=10	0.021	0.030	0.041	0.021
n=11	0.020	0.028	0.041	0.021
n=12	0.022	0.027	0.043	0.022
n=13	0.024	0.031	0.044	0.024
n=14	0.025	0.030	0.045	0.025
n=15	0.026	0.033	0.045	0.025
n=16	0.026	0.031	0.046	0.026
n=17	0.024	0.026	0.042	0.024
n=18	0.021	0.020	0.036	0.021
n=19	0.025	0.020	0.039	0.024
n=20	0.022	0.017	0.032	0.022
n=21	0.028	0.022	0.036	0.027
n=22	0.026	0.020	0.036	0.026
n=23	0.029	0.022	0.038	0.028

Table 31. Training results in each evaluation metrics for Support Vector Machine with ANOVA filter method.

Number of Selected Features	Evaluation metrics			
	Accuracy	**Precision**	**Recall**	**F1-score**
n=2	0.837	0.835	0.880	0.856
n=3	0.811	0.792	0.896	0.840
n=4	0.794	0.771	0.892	0.827
n=5	0.798	0.785	0.875	0.827
n=6	0.790	0.779	0.866	0.820
n=7	0.882	0.939	0.842	0.887
n=8	0.888	0.961	0.831	0.891
n=9	0.824	0.834	0.851	0.842
n=10	0.818	0.817	0.866	0.840
n=11	0.888	0.922	0.871	0.896
n=12	0.886	0.922	0.867	0.893
n=13	0.790	0.776	0.874	0.821
n=14	0.779	0.758	0.885	0.816
n=15	0.781	0.755	0.896	0.819
n=16	0.777	0.751	0.895	0.816
n=17	0.777	0.749	0.897	0.816
n=18	0.814	0.803	0.881	0.839
n=19	0.800	0.783	0.887	0.831
n=20	0.878	0.909	0.868	0.888
n=21	0.824	0.820	0.875	0.846
n=22	0.804	0.791	0.881	0.833
n=23	0.795	0.774	0.891	0.828

Table 32. Standard Deviation of training results in each evaluation metric for Support Vector Machine with ANOVA filter method.

Number of Selected Features	Standard Deviation			
	Accuracy	**Precision**	**Recall**	**F1-score**
n=2	0.030	0.029	0.035	0.026
n=3	0.029	0.029	0.028	0.023
n=4	0.032	0.026	0.036	0.027
n=5	0.033	0.030	0.040	0.028
n=6	0.039	0.034	0.044	0.034
n=7	0.022	0.022	0.033	0.022
n=8	0.021	0.017	0.038	0.022
n=9	0.028	0.028	0.030	0.025
n=10	0.031	0.030	0.029	0.026
n=11	0.036	0.032	0.041	0.034
n=12	0.033	0.032	0.041	0.031
n=13	0.038	0.041	0.031	0.030
n=14	0.030	0.030	0.028	0.023
n=15	0.030	0.029	0.030	0.024
n=16	0.030	0.027	0.037	0.025
n=17	0.024	0.018	0.035	0.021
n=18	0.030	0.033	0.038	0.025
n=19	0.034	0.040	0.037	0.026
n=20	0.019	0.029	0.029	0.017
n=21	0.023	0.028	0.032	0.019
n=22	0.037	0.043	0.032	0.029
n=23	0.039	0.041	0.029	0.030

Table 33. Training results in each evaluation metrics for Support Vector Machine with Chi-squared filter method.

Number of Selected Features	Evaluation metrics			
	Accuracy	**Precision**	**Recall**	**F1-score**
n=2	0.645	0.613	0.974	0.752
n=3	0.646	0.614	0.974	0.753
n=4	0.646	0.613	0.975	0.753
n=5	0.657	0.621	0.976	0.759
n=6	0.656	0.620	0.976	0.759
n=7	0.679	0.637	0.974	0.770
n=8	0.674	0.633	0.974	0.767
n=9	0.694	0.652	0.960	0.777
n=10	0.896	0.933	0.875	0.902
n=11	0.797	0.778	0.889	0.829
n=12	0.761	0.722	0.926	0.811
n=13	0.743	0.702	0.934	0.801
n=14	0.772	0.735	0.924	0.818
n=15	0.788	0.759	0.912	0.827
n=16	0.774	0.743	0.910	0.818
n=17	0.774	0.743	0.912	0.818
n=18	0.795	0.769	0.904	0.830
n=19	0.892	0.923	0.880	0.900
n=20	0.886	0.919	0.872	0.895
n=21	0.811	0.798	0.883	0.838
n=22	0.804	0.791	0.881	0.833
n=23	0.795	0.774	0.891	0.828

Table 34. Standard Deviation of training results in each evaluation metric for Support Vector Machine with Chi-squared filter method.

Number of Selected Features	Standard Deviation			
	Accuracy	**Precision**	**Recall**	**F1-score**
n=2	0.023	0.016	0.012	0.013
n=3	0.021	0.015	0.012	0.012
n=4	0.022	0.015	0.009	0.012
n=5	0.020	0.014	0.010	0.012
n=6	0.020	0.014	0.010	0.012
n=7	0.022	0.016	0.014	0.014
n=8	0.023	0.016	0.014	0.015
n=9	0.023	0.018	0.016	0.014
n=10	0.020	0.016	0.038	0.020
n=11	0.026	0.034	0.026	0.019
n=12	0.033	0.033	0.022	0.022
n=13	0.033	0.031	0.019	0.022
n=14	0.041	0.040	0.019	0.027
n=15	0.046	0.048	0.021	0.031
n=16	0.044	0.043	0.021	0.030
n=17	0.052	0.049	0.027	0.036
n=18	0.041	0.044	0.020	0.029
n=19	0.013	0.030	0.022	0.011
n=20	0.014	0.025	0.020	0.012
n=21	0.029	0.033	0.029	0.024
n=22	0.037	0.043	0.032	0.029
n=23	0.039	0.041	0.029	0.030

Table 35. Training results in each evaluation metrics for Decision Tree with ANOVA filter method.

Number of Selected Features	Evaluation metrics			
	Accuracy	**Precision**	**Recall**	**F1-score**
n=2	0.874	0.903	0.866	0.883
n=3	0.874	0.903	0.866	0.883
n=4	0.881	0.915	0.867	0.890
n=5	0.870	0.888	0.875	0.881
n=6	0.853	0.859	0.883	0.868
n=7	0.861	0.882	0.867	0.873
n=8	0.856	0.867	0.875	0.870
n=9	0.896	0.915	0.896	0.905
n=10	0.898	0.913	0.903	0.907
n=11	0.899	0.912	0.905	0.908
n=12	0.860	0.845	0.918	0.879
n=13	0.899	0.914	0.903	0.908
n=14	0.900	0.912	0.906	0.909
n=15	0.854	0.827	0.932	0.876
n=16	0.898	0.913	0.903	0.907
n=17	0.899	0.914	0.903	0.908
n=18	0.894	0.910	0.899	0.904
n=19	0.894	0.910	0.899	0.904
n=20	0.903	0.925	0.899	0.911
n=21	0.903	0.925	0.899	0.911
n=22	0.903	0.925	0.899	0.911
n=23	0.903	0.925	0.899	0.911

Table 36. Standard Deviation of training results in each evaluation metric for Decision Tree with ANOVA filter method.

Number of Selected Features	Standard Deviation			
	Accuracy	**Precision**	**Recall**	**F1-score**
n=2	0.034	0.031	0.052	0.033
n=3	0.034	0.031	0.052	0.033
n=4	0.031	0.033	0.043	0.029
n=5	0.023	0.021	0.044	0.023
n=6	0.030	0.039	0.063	0.029
n=7	0.025	0.038	0.037	0.021
n=8	0.033	0.036	0.044	0.030
n=9	0.033	0.035	0.050	0.030
n=10	0.023	0.027	0.030	0.021
n=11	0.023	0.028	0.032	0.022
n=12	0.032	0.041	0.033	0.026
n=13	0.023	0.027	0.030	0.021
n=14	0.023	0.028	0.032	0.022
n=15	0.023	0.030	0.032	0.018
n=16	0.023	0.027	0.034	0.022
n=17	0.024	0.027	0.034	0.022
n=18	0.022	0.021	0.037	0.021
n=19	0.022	0.021	0.037	0.021
n=20	0.018	0.019	0.043	0.018
n=21	0.018	0.019	0.043	0.018
n=22	0.018	0.019	0.043	0.018
n=23	0.018	0.019	0.043	0.018

Table 37. Training results in each evaluation metrics for Decision Tree with Chi-squared filter method.

Number of Selected Features	Evaluation metrics			
	Accuracy	**Precision**	**Recall**	**F1-score**
n=2	0.745	0.726	0.867	0.790
n=3	0.744	0.745	0.816	0.779
n=4	0.744	0.745	0.816	0.779
n=5	0.892	0.908	0.897	0.902
n=6	0.845	0.809	0.945	0.871
n=7	0.896	0.918	0.892	0.904
n=8	0.896	0.909	0.903	0.905
n=9	0.894	0.910	0.899	0.904
n=10	0.903	0.925	0.899	0.911
n=11	0.903	0.925	0.899	0.911
n=12	0.860	0.837	0.929	0.880
n=13	0.903	0.925	0.899	0.911
n=14	0.903	0.925	0.899	0.911
n=15	0.903	0.925	0.899	0.911
n=16	0.903	0.925	0.899	0.911
n=17	0.903	0.925	0.899	0.911
n=18	0.903	0.925	0.899	0.911
n=19	0.903	0.925	0.899	0.911
n=20	0.903	0.925	0.899	0.911
n=21	0.903	0.925	0.899	0.911
n=22	0.903	0.925	0.899	0.911
n=23	0.903	0.925	0.899	0.911

Table 38. Standard Deviation of training results in each evaluation metric for Decision Tree with Chi-squared filter method.

Number of Selected Features	Standard Deviation			
	Accuracy	**Precision**	**Recall**	**F1-score**
n=2	0.043	0.032	0.054	0.037
n=3	0.048	0.041	0.049	0.041
n=4	0.048	0.041	0.049	0.041
n=5	0.026	0.028	0.042	0.024
n=6	0.029	0.032	0.030	0.023
n=7	0.028	0.022	0.041	0.027
n=8	0.026	0.023	0.036	0.024
n=9	0.024	0.021	0.036	0.023
n=10	0.018	0.019	0.043	0.018
n=11	0.018	0.019	0.043	0.018
n=12	0.030	0.039	0.026	0.024
n=13	0.018	0.019	0.043	0.018
n=14	0.018	0.019	0.043	0.018
n=15	0.018	0.019	0.043	0.018
n=16	0.018	0.019	0.043	0.018
n=17	0.018	0.019	0.043	0.018
n=18	0.018	0.019	0.043	0.018
n=19	0.018	0.019	0.043	0.018
n=20	0.018	0.019	0.043	0.018
n=21	0.018	0.019	0.043	0.018
n=22	0.018	0.019	0.043	0.018
n=23	0.018	0.019	0.043	0.018

Table 39. Training results in each evaluation metrics for Random Forest with ANOVA filter method.

Number of Selected Features	Evaluation metrics			
	Accuracy	**Precision**	**Recall**	**F1-score**
n=2	0.885	0.912	0.878	0.894
n=3	0.870	0.873	0.897	0.884
n=4	0.881	0.905	0.879	0.891
n=5	0.883	0.895	0.893	0.894
n=6	0.882	0.885	0.905	0.894
n=7	0.881	0.888	0.899	0.892
n=8	0.881	0.888	0.897	0.892
n=9	0.898	0.909	0.908	0.907
n=10	0.911	0.934	0.904	0.918
n=11	0.918	0.942	0.908	0.924
n=12	0.914	0.938	0.905	0.921
n=13	0.916	0.940	0.905	0.922
n=14	0.916	0.939	0.907	0.922
n=15	0.914	0.935	0.908	0.921
n=16	0.914	0.937	0.907	0.921
n=17	0.913	0.935	0.907	0.920
n=18	0.918	0.941	0.909	0.924
n=19	0.917	0.942	0.907	0.923
n=20	0.918	0.939	0.913	0.925
n=21	0.919	0.939	0.914	0.926
n=22	0.918	0.937	0.913	0.924
n=23	0.917	0.937	0.912	0.924

Table 40. Standard Deviation of training results in each evaluation metric for Random Forest with ANOVA filter method.

Number of Selected Features	Standard Deviation			
	Accuracy	**Precision**	**Recall**	**F1-score**
n=2	0.030	0.028	0.046	0.029
n=3	0.019	0.023	0.037	0.018
n=4	0.025	0.025	0.043	0.024
n=5	0.022	0.026	0.038	0.020
n=6	0.023	0.025	0.035	0.021
n=7	0.018	0.020	0.039	0.018
n=8	0.020	0.020	0.038	0.019
n=9	0.031	0.030	0.046	0.029
n=10	0.025	0.027	0.042	0.024
n=11	0.027	0.021	0.042	0.026
n=12	0.024	0.024	0.043	0.023
n=13	0.028	0.024	0.042	0.026
n=14	0.029	0.026	0.044	0.027
n=15	0.023	0.023	0.039	0.022
n=16	0.023	0.026	0.043	0.022
n=17	0.025	0.024	0.038	0.024
n=18	0.022	0.025	0.036	0.021
n=19	0.024	0.024	0.039	0.023
n=20	0.022	0.024	0.038	0.021
n=21	0.021	0.024	0.039	0.020
n=22	0.026	0.027	0.041	0.024
n=23	0.021	0.025	0.038	0.020

Table 41. Training results in each evaluation metrics for Random Forest with Chi-squared filter method.

Number of Selected Features	Evaluation metrics			
	Accuracy	**Precision**	**Recall**	**F1-score**
n=2	0.755	0.782	0.771	0.776
n=3	0.755	0.766	0.804	0.784
n=4	0.752	0.765	0.797	0.780
n=5	0.897	0.905	0.910	0.907
n=6	0.905	0.918	0.910	0.914
n=7	0.908	0.934	0.899	0.915
n=8	0.913	0.935	0.908	0.920
n=9	0.913	0.935	0.907	0.920
n=10	0.916	0.936	0.910	0.922
n=11	0.920	0.939	0.916	0.926
n=12	0.913	0.933	0.909	0.920
n=13	0.915	0.935	0.910	0.922
n=14	0.917	0.939	0.910	0.924
n=15	0.919	0.940	0.913	0.926
n=16	0.915	0.936	0.909	0.922
n=17	0.918	0.940	0.912	0.925
n=18	0.916	0.934	0.914	0.923
n=19	0.918	0.940	0.912	0.925
n=20	0.918	0.937	0.913	0.924
n=21	0.918	0.940	0.910	0.924
n=22	0.918	0.937	0.913	0.924
n=23	0.917	0.937	0.912	0.924

Table 42. Standard Deviation of training results in each evaluation metric for Random Forest with Chi-squared filter method.

Number of Selected Features	Standard Deviation			
	Accuracy	**Precision**	**Recall**	**F1-score**
n=2	0.040	0.034	0.057	0.040
n=3	0.034	0.033	0.048	0.031
n=4	0.029	0.027	0.048	0.028
n=5	0.022	0.020	0.039	0.021
n=6	0.026	0.027	0.039	0.024
n=7	0.028	0.028	0.043	0.027
n=8	0.025	0.031	0.039	0.023
n=9	0.024	0.023	0.043	0.023
n=10	0.027	0.027	0.043	0.025
n=11	0.026	0.028	0.044	0.025
n=12	0.020	0.021	0.039	0.020
n=13	0.025	0.029	0.041	0.024
n=14	0.024	0.028	0.038	0.022
n=15	0.021	0.025	0.038	0.020
n=16	0.027	0.029	0.041	0.025
n=17	0.021	0.026	0.039	0.020
n=18	0.022	0.029	0.037	0.020
n=19	0.025	0.023	0.040	0.023
n=20	0.022	0.025	0.038	0.021
n=21	0.023	0.026	0.040	0.022
n=22	0.026	0.027	0.041	0.024
n=23	0.021	0.025	0.038	0.020

Appendix 3

Table 43. Testing results in each evaluation metrics for Logistic Regression with ANOVA filter method.

Number of Selected Features	Evaluation metrics			
	Accuracy	**Precision**	**Recall**	**F1-score**
n=2	0.858	0.920	0.821	0.867
n=3	0.858	0.920	0.821	0.867
n=4	0.869	0.936	0.826	0.877
n=5	0.904	0.955	0.872	0.912
n=6	0.904	0.955	0.872	0.912
n=7	0.895	0.954	0.856	0.903
n=8	0.884	0.938	0.851	0.892
n=9	0.887	0.938	0.856	0.895
n=10	0.890	0.939	0.862	0.898
n=11	0.890	0.929	0.872	0.899
n=12	0.890	0.929	0.872	0.899
n=13	0.890	0.929	0.872	0.899
n=14	0.890	0.934	0.867	0.899
n=15	0.890	0.929	0.872	0.899
n=16	0.890	0.934	0.867	0.899
n=17	0.890	0.934	0.867	0.899
n=18	0.895	0.934	0.877	0.905
n=19	0.892	0.934	0.872	0.902
n=20	0.884	0.914	0.877	0.895
n=21	0.881	0.918	0.867	0.892
n=22	0.881	0.918	0.867	0.892
n=23	0.881	0.918	0.867	0.892

Table 44. Confusion matrix of testing results for Logistic Regression with ANOVA filter method.

Number of Selected Features	Confusion matrix			
	TN	**FP**	**FN**	**TP**
n=2	135	14	35	160
n=3	135	14	35	160
n=4	138	11	34	161
n=5	141	8	25	170
n=6	141	8	25	170
n=7	141	8	28	167
n=8	138	11	29	166
n=9	138	11	28	167
n=10	138	11	27	168
n=11	136	13	25	170
n=12	136	13	25	170
n=13	136	13	25	170
n=14	137	12	26	169
n=15	136	13	25	170
n=16	137	12	26	169
n=17	137	12	26	169
n=18	137	12	24	171
n=19	137	12	25	170
n=20	133	16	24	171
n=21	134	15	26	169
n=22	134	15	26	169
n=23	134	15	26	169

Table 45. Testing results in each evaluation metrics for Logistic Regression with Chi-squared filter method.

Number of Selected Features	Evaluation metrics			
	Accuracy	**Precision**	**Recall**	**F1-score**
n=2	0.744	0.720	0.897	0.799
n=3	0.747	0.723	0.897	0.801
n=4	0.747	0.723	0.897	0.801
n=5	0.808	0.772	0.938	0.847
n=6	0.814	0.779	0.938	0.851
n=7	0.869	0.936	0.826	0.877
n=8	0.863	0.940	0.810	0.871
n=9	0.858	0.935	0.805	0.865
n=10	0.875	0.918	0.856	0.886
n=11	0.875	0.918	0.856	0.886
n=12	0.878	0.927	0.851	0.888
n=13	0.872	0.913	0.856	0.884
n=14	0.869	0.908	0.856	0.881
n=15	0.875	0.913	0.862	0.887
n=16	0.878	0.918	0.862	0.889
n=17	0.881	0.938	0.846	0.889
n=18	0.887	0.938	0.856	0.895
n=19	0.895	0.934	0.877	0.905
n=20	0.881	0.918	0.867	0.892
n=21	0.878	0.918	0.862	0.889
n=22	0.881	0.918	0.867	0.892
n=23	0.881	0.918	0.867	0.892

Table 46. Confusion matrix of testing results for Logistic Regression with Chi-squared filter method.

Number of Selected Features	Confusion matrix			
	TN	**FP**	**FN**	**TP**
n=2	81	68	20	175
n=3	82	67	20	175
n=4	82	67	20	175
n=5	95	54	12	183
n=6	97	52	12	183
n=7	138	11	34	161
n=8	139	10	37	158
n=9	138	11	38	157
n=10	134	15	28	167
n=11	134	15	28	167
n=12	136	13	29	166
n=13	133	16	28	167
n=14	132	17	28	167
n=15	133	16	27	168
n=16	134	15	27	168
n=17	138	11	30	165
n=18	138	11	28	167
n=19	137	12	24	171
n=20	134	15	26	169
n=21	134	15	27	168
n=22	134	15	26	169
n=23	134	15	26	169

Table 47. Testing results in each evaluation metrics for Support Vector Machine with ANOVA filter method.

Number of Selected Features	Evaluation metrics			
	Accuracy	**Precision**	**Recall**	**F1-score**
n=2	0.831	0.822	0.897	0.858
n=3	0.802	0.775	0.918	0.840
n=4	0.808	0.779	0.923	0.845
n=5	0.834	0.811	0.923	0.863
n=6	0.814	0.799	0.897	0.845
n=7	0.892	0.944	0.862	0.901
n=8	0.884	0.948	0.841	0.891
n=9	0.843	0.847	0.882	0.864
n=10	0.840	0.833	0.897	0.864
n=11	0.901	0.930	0.892	0.911
n=12	0.895	0.930	0.882	0.905
n=13	0.826	0.820	0.887	0.852
n=14	0.823	0.805	0.908	0.853
n=15	0.817	0.792	0.918	0.850
n=16	0.817	0.789	0.923	0.851
n=17	0.814	0.784	0.928	0.850
n=18	0.846	0.832	0.913	0.870
n=19	0.840	0.818	0.923	0.867
n=20	0.895	0.908	0.908	0.908
n=21	0.852	0.846	0.903	0.873
n=22	0.852	0.840	0.913	0.875
n=23	0.843	0.828	0.913	0.868

Table 48. Confusion matrix of testing results for Support Vector Machine with ANOVA filter method.

Number of Selected Features	Confusion matrix			
	TN	**FP**	**FN**	**TP**
n=2	111	38	20	175
n=3	97	52	16	179
n=4	98	51	15	180
n=5	107	42	15	180
n=6	105	44	20	175
n=7	139	10	27	168
n=8	140	9	31	164
n=9	118	31	23	172
n=10	114	35	20	175
n=11	136	13	21	174
n=12	136	13	23	172
n=13	111	38	22	173
n=14	106	43	18	177
n=15	102	47	16	179
n=16	101	48	15	180
n=17	99	50	14	181
n=18	113	36	17	178
n=19	109	40	15	180
n=20	131	18	18	177
n=21	117	32	19	176
n=22	115	34	17	178
n=23	112	37	17	178

Table 49. Testing results in each evaluation metrics for Support Vector Machine with Chi-squared filter method.

Number of Selected Features	Evaluation metrics			
	Accuracy	**Precision**	**Recall**	**F1-score**
n=2	0.642	0.617	0.974	0.755
n=3	0.640	0.616	0.969	0.753
n=4	0.640	0.616	0.969	0.753
n=5	0.663	0.631	0.974	0.766
n=6	0.657	0.627	0.974	0.763
n=7	0.680	0.643	0.979	0.776
n=8	0.680	0.643	0.979	0.776
n=9	0.712	0.668	0.979	0.794
n=10	0.907	0.931	0.903	0.917
n=11	0.823	0.825	0.872	0.848
n=12	0.799	0.765	0.933	0.841
n=13	0.788	0.748	0.944	0.834
n=14	0.808	0.777	0.928	0.846
n=15	0.826	0.803	0.918	0.856
n=16	0.811	0.780	0.928	0.848
n=17	0.823	0.794	0.928	0.856
n=18	0.849	0.824	0.933	0.875
n=19	0.890	0.911	0.892	0.902
n=20	0.898	0.930	0.887	0.908
n=21	0.863	0.856	0.913	0.883
n=22	0.852	0.840	0.913	0.875
n=23	0.843	0.828	0.913	0.868

Table 50. Confusion matrix of testing results for Support Vector Machine with Chi-squared filter method.

Number of Selected Features	Confusion matrix			
	TN	**FP**	**FN**	**TP**
n=2	31	118	5	190
n=3	31	118	6	189
n=4	31	118	6	189
n=5	38	111	5	190
n=6	36	113	5	190
n=7	43	106	4	191
n=8	43	106	4	191
n=9	54	95	4	191
n=10	136	13	19	176
n=11	113	36	25	170
n=12	93	56	13	182
n=13	87	62	11	184
n=14	97	52	14	181
n=15	105	44	16	179
n=16	98	51	14	181
n=17	102	47	14	181
n=18	110	39	13	182
n=19	132	17	21	174
n=20	136	13	22	173
n=21	119	30	17	178
n=22	115	34	17	178
n=23	112	37	17	178

Table 51. Testing results in each evaluation metrics for Decision Tree with ANOVA filter method.

Number of Selected Features	Evaluation metrics			
	Accuracy	**Precision**	**Recall**	**F1-score**
n=2	0.887	0.906	0.892	0.899
n=3	0.887	0.906	0.892	0.899
n=4	0.910	0.946	0.892	0.918
n=5	0.892	0.916	0.892	0.904
n=6	0.869	0.854	0.928	0.889
n=7	0.878	0.901	0.882	0.891
n=8	0.872	0.895	0.877	0.886
n=9	0.887	0.906	0.892	0.899
n=10	0.913	0.915	0.933	0.924
n=11	0.913	0.915	0.933	0.924
n=12	0.881	0.867	0.933	0.899
n=13	0.913	0.910	0.938	0.924
n=14	0.913	0.910	0.938	0.924
n=15	0.852	0.824	0.938	0.878
n=16	0.913	0.910	0.938	0.924
n=17	0.913	0.910	0.938	0.924
n=18	0.913	0.919	0.928	0.923
n=19	0.913	0.919	0.928	0.923
n=20	0.919	0.937	0.918	0.927
n=21	0.919	0.937	0.918	0.927
n=22	0.919	0.937	0.918	0.927
n=23	0.919	0.937	0.918	0.927

Table 52. Confusion matrix of testing results for Decision Tree with ANOVA filter method.

Number of Selected Features	Confusion matrix			
	TN	**FP**	**FN**	**TP**
n=2	131	18	21	174
n=3	131	18	21	174
n=4	139	10	21	174
n=5	133	16	21	174
n=6	118	31	14	181
n=7	130	19	23	172
n=8	129	20	24	171
n=9	131	18	21	174
n=10	132	17	13	182
n=11	132	17	13	182
n=12	121	28	13	182
n=13	131	18	12	183
n=14	131	18	12	183
n=15	110	39	12	183
n=16	131	18	12	183
n=17	131	18	12	183
n=18	133	16	14	181
n=19	133	16	14	181
n=20	137	12	16	179
n=21	137	12	16	179
n=22	137	12	16	179
n=23	137	12	16	179

Table 53. Testing results in each evaluation metrics for Decision Tree with Chi-squared filter method.

Number of Selected Features	Evaluation metrics			
	Accuracy	**Precision**	**Recall**	**F1-score**
n=2	0.756	0.738	0.882	0.804
n=3	0.767	0.773	0.836	0.803
n=4	0.767	0.773	0.836	0.803
n=5	0.913	0.923	0.923	0.923
n=6	0.858	0.829	0.944	0.882
n=7	0.895	0.912	0.903	0.907
n=8	0.878	0.901	0.882	0.891
n=9	0.913	0.919	0.928	0.923
n=10	0.919	0.937	0.918	0.927
n=11	0.919	0.937	0.918	0.927
n=12	0.881	0.867	0.933	0.899
n=13	0.919	0.937	0.918	0.927
n=14	0.919	0.937	0.918	0.927
n=15	0.919	0.937	0.918	0.927
n=16	0.919	0.937	0.918	0.927
n=17	0.919	0.937	0.918	0.927
n=18	0.919	0.937	0.918	0.927
n=19	0.919	0.937	0.918	0.927
n=20	0.919	0.937	0.918	0.927
n=21	0.919	0.937	0.918	0.927
n=22	0.919	0.937	0.918	0.927
n=23	0.919	0.937	0.918	0.927

Table 54. Confusion matrix of testing results for Decision Tree with Chi-squared filter method.

Number of Selected Features	Confusion matrix			
	TN	**FP**	**FN**	**TP**
n=2	88	61	23	172
n=3	101	48	32	163
n=4	101	48	32	163
n=5	134	15	15	180
n=6	111	38	11	184
n=7	132	17	19	176
n=8	130	19	23	172
n=9	133	16	14	181
n=10	137	12	16	179
n=11	137	12	16	179
n=12	121	28	13	182
n=13	137	12	16	179
n=14	137	12	16	179
n=15	137	12	16	179
n=16	137	12	16	179
n=17	137	12	16	179
n=18	137	12	16	179
n=19	137	12	16	179
n=20	137	12	16	179
n=21	137	12	16	179
n=22	137	12	16	179
n=23	137	12	16	179

Table 55. Testing results in each evaluation metrics for Random Forest with ANOVA filter method.

Number of Selected Features	Evaluation metrics			
	Accuracy	**Precision**	**Recall**	**F1-score**
n=2	0.887	0.906	0.892	0.899
n=3	0.875	0.869	0.918	0.893
n=4	0.890	0.907	0.897	0.902
n=5	0.892	0.895	0.918	0.906
n=6	0.887	0.882	0.923	0.902
n=7	0.892	0.895	0.918	0.906
n=8	0.890	0.891	0.918	0.904
n=9	0.895	0.904	0.913	0.908
n=10	0.924	0.938	0.928	0.933
n=11	0.930	0.943	0.933	0.938
n=12	0.939	0.939	0.954	0.947
n=13	0.930	0.934	0.944	0.939
n=14	0.930	0.938	0.938	0.938
n=15	0.933	0.939	0.944	0.941
n=16	0.927	0.929	0.944	0.936
n=17	0.930	0.938	0.938	0.938
n=18	0.930	0.938	0.938	0.938
n=19	0.927	0.938	0.933	0.936
n=20	0.924	0.929	0.938	0.934
n=21	0.933	0.939	0.944	0.941
n=22	0.924	0.933	0.933	0.933
n=23	0.933	0.939	0.944	0.941

Table 56. Confusion matrix of testing results for Random Forest with ANOVA filter method.

Number of Selected Features	Confusion matrix			
	TN	**FP**	**FN**	**TP**
n=2	131	18	21	174
n=3	122	27	16	179
n=4	131	18	20	175
n=5	128	21	16	179
n=6	125	24	15	180
n=7	128	21	16	179
n=8	127	22	16	179
n=9	130	19	17	178
n=10	137	12	14	181
n=11	138	11	13	182
n=12	137	12	9	186
n=13	136	13	11	184
n=14	137	12	12	183
n=15	137	12	11	184
n=16	135	14	11	184
n=17	137	12	12	183
n=18	137	12	12	183
n=19	137	12	13	182
n=20	135	14	12	183
n=21	137	12	11	184
n=22	136	13	13	182
n=23	137	12	11	184

Table 57. Testing results in each evaluation metrics for Random Forest with Chi-squared filter method.

Number of Selected Features	Evaluation metrics			
	Accuracy	**Precision**	**Recall**	**F1-score**
n=2	0.759	0.772	0.815	0.793
n=3	0.765	0.766	0.841	0.802
n=4	0.756	0.768	0.815	0.791
n=5	0.884	0.889	0.908	0.898
n=6	0.913	0.919	0.928	0.923
n=7	0.919	0.933	0.923	0.928
n=8	0.927	0.947	0.923	0.935
n=9	0.936	0.944	0.944	0.944
n=10	0.933	0.957	0.923	0.940
n=11	0.939	0.948	0.944	0.946
n=12	0.942	0.958	0.938	0.948
n=13	0.936	0.953	0.933	0.943
n=14	0.933	0.943	0.938	0.941
n=15	0.936	0.948	0.938	0.943
n=16	0.930	0.952	0.923	0.938
n=17	0.936	0.944	0.944	0.944
n=18	0.936	0.935	0.954	0.944
n=19	0.927	0.934	0.938	0.936
n=20	0.930	0.938	0.938	0.938
n=21	0.936	0.944	0.944	0.944
n=22	0.924	0.933	0.933	0.933
n=23	0.933	0.939	0.944	0.941

Table 58. Confusion matrix of testing results for Random Forest with Chi-squared filter method.

Number of Selected Features	Confusion matrix			
	TN	**FP**	**FN**	**TP**
n=2	102	47	36	159
n=3	99	50	31	164
n=4	101	48	36	159
n=5	127	22	18	177
n=6	133	16	14	181
n=7	136	13	15	180
n=8	139	10	15	180
n=9	138	11	11	184
n=10	141	8	15	180
n=11	139	10	11	184
n=12	141	8	12	183
n=13	140	9	13	182
n=14	138	11	12	183
n=15	139	10	12	183
n=16	140	9	15	180
n=17	138	11	11	184
n=18	136	13	9	186
n=19	136	13	12	183
n=20	137	12	12	183
n=21	138	11	11	184
n=22	136	13	13	182
n=23	137	12	11	184

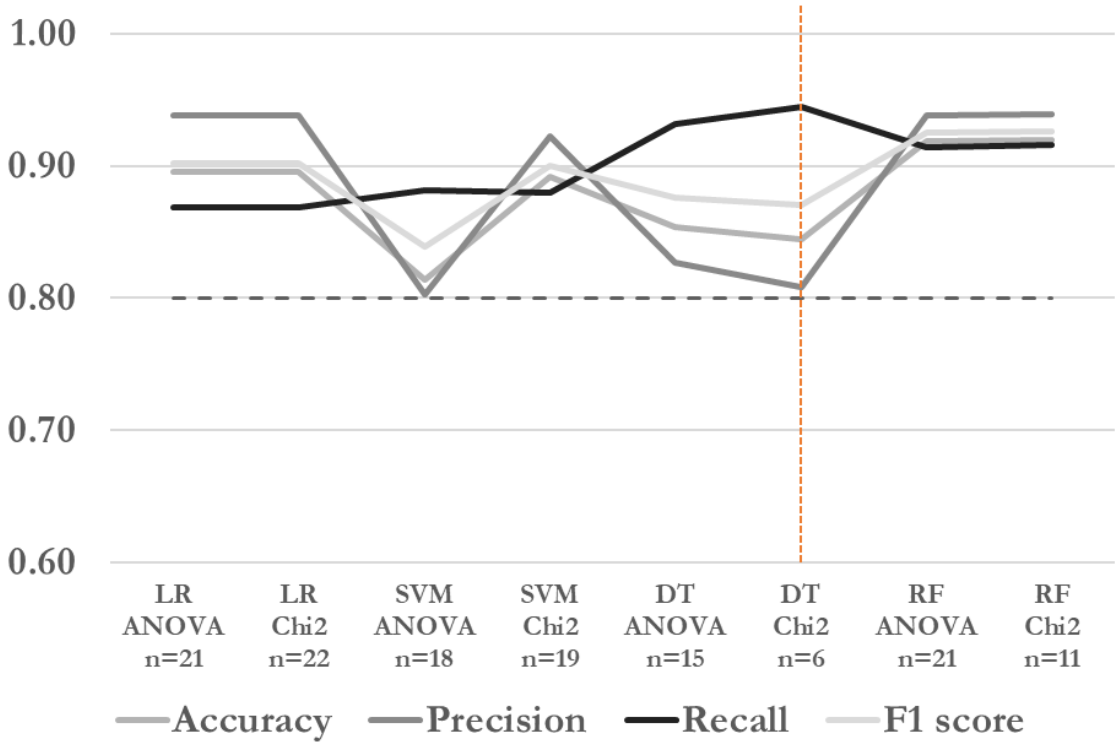


Figure 29. Results comparison

Appendix 4

Table 59 Feature importance score by ANOVA filter method using training data

Ranking	Attribute Name	Score
1	amountSpend_log	628.70
2	daysToExpire	578.39
3	amountSpend	413.50
4	numAct_log	251.23
5	everContact	185.67
6	usernumAct_log	184.60
7	Act_Day_log	179.36
8	amountTrans_log	140.70
9	prevPeriodTrans	82.28
10	currPeriodTrans	79.70
11	numUser_log	77.95
12	totalTrans	71.03
13	numAct	71.01
14	numUser	47.34
15	Act_Day	43.51
16	numCargo_log	28.65
17	hasPhone	15.65
18	lastContactDays	10.62
19	numCargo	6.79
20	UserAct	3.07
21	daysToAct	0.85
22	amountTrans	0.81
23	numContact	0.16

Table 60 Feature importance score by Chi-squared filter method using training data

Ranking	Attribute Name	Score
1	amountTrans	74707500000000000.00
2	amountSpend	7732070.00
3	totalTrans	3399550.00
4	numAct	3398520.00
5	prevPeriodTrans	450694.00
6	currPeriodTrans	281145.00
7	daysToExpire	202702.00
8	Act_Day	4945.87
9	lastContactDays	4752.73
10	UserAct	838.15
11	numCargo	706.77
12	numUser	528.38
13	daysToAct	390.74
14	numAct_log	175.10
15	amountTrans_log	158.10
16	usernumAct_log	153.08
17	Act_Day_log	141.97
18	everContact	96.67
19	amountSpend_log	95.80
20	numUser_log	29.51
21	numCargo_log	17.73
22	hasPhone	2.14
23	numContact	0.16

REFERENCES

[1] G. B, "Rubygarage." [Online]. Available: https://rubygarage.org/blog/iaas-vs-paas-vs-saas

[2] M. Framingham, "Worldwide Public Cloud Services Revenue Grows to Nearly $183 Billion in 2018, Led by the Top 5 Service Providers and Accelerating Public Cloud Services Spending in China." [Online]. Available: https://www.idc.com/getdoc.jsp?containerId=prUS45411519

[3] "Gartner Forecasts Worldwide Public Cloud Revenue to Grow 17% in 2020." [Online]. Available: https://www.gartner.com/en/newsroom/press-releases/2019-11-13-gartner-forecasts-worldwide-public-cloud-revenue-to-grow-17-percent-in-2020

[4] M. Framingham, "Worldwide Public Cloud Services Spending Will More Than Double by 2023." [Online]. Available: https://www.idc.com/getdoc.jsp?containerId=prUS45340719

[5] "Unleashing exponential evolution — 2019 ERP trends." [Online]. Available: https://www.accenture.com/_acnmedia/pdf-90/accenture-unleashing-exponential-evolution-pdf.pdf

[6] "Global Inventory Management Software Market Will Reach USD 3.2 Billion By 2025: Zion Market Research." [Online]. Available: https://www.globenewswire.com/news-release/2019/08/30/1909046/0/en/Global-Inventory-Management-Software-Market-Will-Reach-USD-3-2-Billion-By-2025-Zion-Market-Research.html

[7] "Global Inventory Management Software Market to Hit USD 3 billion by 2024." [Online]. Available: https://www.marketwatch.com/press-release/global-inventory-management-software-market-to-hit-usd-3-billion-by-2024-2019-11-26

[8] "Market Study On The Global Inventory Management Software Market." [Online]. Available: https://www.globalbankingandfinance.com/market-study-on-the-global-inventory-management-software-market-the-inventory-management-software-market-is-expected-to-grow-significantly-in-the-coming-years-owing-to-an-increase-in-emphasis-on-omni/

[9] A. Gallo, "The Value of Keeping the Right Customers," ed: harvard business school publishing corporation, 2014.

[10] F. Reichheld, "BAIN & COMPANY, INC.." [Online]. Available: http://www2.bain.com/Images/BB_Prescription_cutting_costs.pdf

[11] "Luxury Client Experience Board Reveals How Successful Sales Teams Turn First-Time Shoppers into Long-Term Clients." [Online]. Available: https://www.globenewswire.com/news-release/2016/10/19/1394364/0/en/Luxury-Client-Experience-Board-Reveals-How-Successful-Sales-Teams-Turn-First-Time-Shoppers-into-Long-Term-Clients.html

[12] "the value of online customer loyalty and how you can capture it." [Online]. Available: https://www.bain.com/insights/the-value-of-online-customer-loyalty-and-how-you-can-capture-it/

[13] W. Bern and J. Hermansson, "Software-as-a-Service: strategizing for customer loyalty (Master's thesis)," Gothenburg University,, Gothenburg, Sweden,

Master's thesis 2017.
[14] "How E-Commerce is Transforming B2B Sectors." [Online]. Available: https://www.portal.euromonitor.com/
[15] "Shifting Market Frontiers State of Play 2019." [Online]. Available: https://www.portal.euromonitor.com/
[16] "Business Dynamics Thailand." [Online]. Available: https://www.portal.euromonitor.com/
[17] S. Y. Hung and H. Y. Wang, "Applying data mining to telecom churn management," in *Pacific Asia Conference on Information Systems (PACIS)*, 2004.
[18] W. Buckinx and D. Van den Poel, "Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting," *European Journal of Operational*, vol. 164, no. 1, pp. 252-268, 2005.
[19] S. F. Crone, S. Lessmann, and R. Stahlbock, "The impact of preprocessing on data mining: an evaluation of classifier sensitivity in direct marketing," *European Journal of Operational Research*, vol. 173, no. 3, pp. 781-800, 2006.
[20] X. Guo-en and J. Wei-dong, "Model of customer churn prediction on support vector machine," *Systems Engineering–Theory and Practice*, vol. 28, no. 1, pp. 71-77, 2008.
[21] X. Yu, S. Guo, J. Guo, and X. Huang, "An extended support vector machine forecasting framework for customer churn in e-commerce," *Expert Systems with Applications*, vol. 38, no. 3, pp. 1425-1430, 2011.
[22] N. Glady, B. Baesens, and Croux, "Modeling churn using customer lifetime value," *European Journal of Operational Research*, vol. 197, no. 1, pp. 402-411, 2009.
[23] V. Umayaparvathi and K. Iyakutti, "Applications of data mining techniques in telecom churn prediction," *International Journal of Computer Applications*, vol. 42, no. 20, pp. 5-9, 2012.
[24] K. Coussement and K. W. De Bock, "Customer churn prediction in the online gambling industry: the beneficial effect of ensemble learning," *Journal of Business Research*, vol. 66, no. 9, pp. 1629-1636, 2013.
[25] N. Lu, H. Lin, J. Lu, and G. Zhang, "A customer churn prediction model in telecom industry using boosting," *IEEE Transactions on Industrial Informatics*, vol. 10, no. 1, pp. 1659-1665, 2014.
[26] J. Runge, P. Gao, F. Garcin, and B. Faltings, "Chum prediction for high-value players in casual social games," in *2014 IEEE Conference on Computational Intelligence and Games, IEEE Computer Society Press,*, Washington, DC., 2014.
[27] P. Wanchai, "Customer churn analysis : A case study on the telecommunication industry of Thailand," in *2017 12th International Conference for Internet Technology and Secured Transactions (ICITST)*, Cambridge, 2017.
[28] A. Amin, F. Al-Obeidat, B. Shah, A. Adnan, J. Loo, and S. Anwar, "Customer churn prediction in telecommunication industry using data certainty," *Journal of Business Research*, vol. 94, pp. 290-301, 2019.
[29] A. Tamaddoni Jahromi, S. Stakhovych, and M. Ewing, "Managing B2B customer churn, retention and profitability," *Industrial Marketing Management*, vol. 43, no. 7, pp. 1258-1268, 2014.
[30] R. Vadakattu, B. Panda, S. Narayan, and H. Godhia, "Enterprise subscription

churn prediction," in *2015 IEEE International Conference on Big Data (Big Data)*, Santa Clara, California, 2015.

[31] A. Schill, "Customer Churn Prediction in an Online Streaming Service," Lund University, Lund, Sweden, 2018.

[32] O. G. Ali and U. Arıtürk, "Dynamic churn prediction framework with more effective use of rare event data: the case of private banking," *Expert Systems with Applications,* vol. 41, no. 17, pp. 7889-7903, 2014.

[33] P. Hemalatha and G. M. Amalanathan, "A Hybrid Classification Approach for Customer Churn Prediction using Supervised Learning Methods: Banking Sector," in *2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN)*, Vellore, India, 2019.

[34] P. Berger and M. Kompan, "User Modeling for Churn Prediction in E-Commerce," (in English), *Ieee Intelligent Systems,* vol. 34, no. 2, pp. 44-52, Mar-Apr 2019, doi: 10.1109/Mis.2019.2895788.

[35] S. H. Chen, "The gamma CUSUM chart method for online customer churn prediction," *Electronic Commerce Research and Applications,* vol. 17, no. 1, pp. 99-111, 2016.

[36] B. Frank and J. Pittges, "Analyzing Customer Churn in the Software as a Service (SaaS) Industry," Radford university, Virginia, 2009.

[37] Y. Ge, S. He, J. Xiong, and D. E. Brown, "Customer churn analysis for a software-as-a-service company," in *2017 Systems and Information Engineering Design Symposium (SIEDS)*, Charlottesville, VA, 2017.

[38] A. Rautio, "Churn Prediction in SaaS using Machine Learning," Tampere University, Tampere, Finland, 2019.

[39] P. D. Allison, "Missing data, Quantitative applications in the social sciences," SAGE Publications, Inc., Thousand Oaks, California, 2002.

[40] S. Khodabandehlou and M. Zivari Rahman, "Comparison of supervised machine learning techniques for customer churn prediction based on analysis of customer behavior," *Journal of Systems and Information Technology,* vol. 19, no. 1/2, pp. 65-93, 2017.

[41] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *The Journal of Machine Learning Research (JMLR)*, vol. 12, pp. 2825-2830, 2011.

[42] M. Kuhn and K. Johnson, "Over-Fitting and Model Tuning," in *Applied Predictive Modeling*, M. Kuhn and K. Johnson Eds. New York, NY: Springer New York, 2013, pp. 61-92.

[43] A. Sukow and R. Grant, "Forecasting and the Role of Churn in Software-as-a-Service Business Models," *iBusiness,* vol. 05, no. 01, pp. 49-57, 2013.

VITA

NAME	Phongsatorn Amornvetchayakul
DATE OF BIRTH	1 June 1992
PLACE OF BIRTH	Bangkok
INSTITUTIONS ATTENDED	Mechanical Engineering, Bachelor of Engineering degree, Chulalongkorn University
HOME ADDRESS	199 M.3 T. Nong-Hoi A.Muang Chiang Mai
PUBLICATION	-
AWARD RECEIVED	-