Cover Page – MSc Business Analytics Consultancy Project/Dissertation 2021-22

Title of Project : **Predicting Customer Retention in the E-Commerce Industry in Indonesia: A Machine Learning Approach**

Candidate Number : **RSCB1**

Date : **30 July 2022**

Word Count : **10,326**

Disclaimer:

I hereby declare that this dissertation is my individual work and to the best of my knowledge and confidence, it has not already been accepted in substance for the award of any other degree and is not concurrently submitted in candidature for any degree. It is the end product of my own independent study except where other acknowledgement has been stated in the text.

Marking Sheet – MSc Business Analytics Consultancy Project/Dissertation 2021-22

Criteria/Weight	Supervisor's comments
Topic, theoretical framework, literature, and methodology (35%): Topic is clearly identified and boundaries are asserted. Knowledge of relevant theories and their limitations. Current and relevant literature coming from reliable sources. Appropriate and adequate methodology for topics. Detailed methodology facilitating replication of project and reproducibility of results.	
Analysis and conclusions /recommendations (35%): Use of primary and/or secondary data. Rigorous analysis and interpretations. Alternative interpretations/arguments are considered. Limitations are identified and justified by reasonable arguments. Conclusions/recommendations are fully consistent with evidence presented.	
Structure, originality and presentation (10%): Provides a concise summary. Demonstrates an understanding of business context. Coherent and appropriate structure. Adequate presentation, language, style, graphs, tables, and referencing. Appropriate use of visualisation. Presents business recommendations.	
Complexity of project scope and progress made towards business goals (10%): Progress made towards overcoming technical and operational challenges encountered during the project. Progress made in overcoming problem framing and theoretical and data related problems encountered during the project.	
Project Management (10%): Good use of project management and communication tools. Use of Kanban board for structuring project work. Evidence of objectives being broken down in appropriate tasks and timely engagement with primary supervisor.	

Abstract

For the past decade, managing complex customer retention has presented a constant challenge for the retail industry. As a result, companies invest in customer relationship management (CRM) tools and leverage their data to understand customers' behaviour, identify profitable segments, predict the non-recurring (churning) customers, and formulate their retention strategy.

Previous studies mainly highlighted this problem in an arguably more traditional industry, such as telecommunications or airlines. Moreover, the focus often falls on finding the best model by implementing and comparing state-of-the-art algorithms without emphasising how it applies in modern business settings.

This research used e-commerce data from Indonesia in 2021, one of the fastest-growing countries for online retail, to predict whether a customer will make another transaction in the following 30 days. Several algorithms are compared to ensure a robust analysis: decision tree, random forest, XGBoost, neural network, and voting classifier. Overall, the random forest has the best accuracy (0.751243). In contrast, XGBoost has the best precision (0.816262), and the neural network has the best recall (0.715383).

Finally, two business application approaches, applying both precision and recall metrics, are examined in detail to encourage implementation.

Table of Contents

Chapter 1: Introduction

1.1 Background

E-commerce, generally defined as purchase and sale of goods or services through the internet (Lawrence et al., 1998), has experienced rapid development since the internet has become accessible to the general population.

Following this trend, for the past decade e-commerce in Indonesia has experienced phenomenal growth due to several factors such as strong digital-savvy customer base, robust middle-income segment, solid digital merchant growth, and healthy investment appetite from venture capitals in the region (Google, 2019). Furthermore, recent studies forecast Indonesia's e-commerce Gross Merchandise Value (GMV) to reach US$ 104 billion in 2025 (Google, 2021).

However, this growth comes with its own unique challenges. Google (2019) identified six key barriers to unlock growth in South-East Asia: access to the internet, funding for e-commerce players, consumer trust, payment, logistics, and talent.

SIRCLO (www.sirclo.com) is an e-commerce enabler company based in Indonesia that tries to solve these hurdles. Founded in 2013, the company focuses on helping brands sell online by solving technology integration, brand management, marketing, and fulfilment challenges.

To further strengthen its offering as a one-stop solution company in the e-commerce industry, the company merged with Orami, a parenting-focused e-commerce platform, and Warung Pintar, a new-retail enabler, in the last couple of years. Currently, SIRCLO has 2,000 employees, serves more than 3,000 brands and manages more than US$ 300 million in annual GMV.

1.2 Motivation

For the past decade, customer retention has become a central topic for management and marketing decisions (Lariviere and Van Den Poel, 2005). Alongside with customer acquisition, it presents a constant challenge for the retail industry; thus, improvement in both sectors are a significant concern (Manivannan, R. et al., 2021). Furthermore, given

budget constraints, companies are pressured to optimise their resource allocation into the right marketing strategy to compete effectively in the market (Angeloni and Rossi, 2021).

As the Indonesian e-commerce market gets intensified, customer acquisition costs steadily increase (Google, 2021). Research shows that with internet penetration reaching 75% in Indonesia, e-commerce players need to unlock new customers beyond their tier-1 cities, defined as top 15 cities based on economic output, to sustain their astronomical growth (Google, 2019; Google, 2021). Unfortunately, these digital laggards customers are three to five years behind urban areas in digital adoption and present considerable unlocking obstacles (Alpha JWC & Kearney, 2020).

As a result, players in the industry are exploring ways to allocate their investment to solidify their existing customer base instead of finding a new one. In line with that, studies found that retaining and keeping existing customers is more profitable than acquiring highly churning new customers (Reinartz, W. J. & Kumar, V. (2003)). In 2021, for example, pre-pandemic consumers used four more digital services than before, and 60 million new pandemic-era Indonesian consumers are here to stay (Google, 2021). Chapter 2 will elaborate the e-commerce industry trend in Indonesia further.

Companies realise that their customers' data is one of their most valuable strategic assets to support retention strategy in this technology-driven era (Popovic, D. and Basic, B.D., 2009). An exemplary implementation of customer relationship management (CRM) tools makes companies perform better in managing their customers throughout their life cycle (Reinartz, Krafft, and Hoyer, 2004). In short, formulating a strategy to improve customer retention means understanding their behaviour, identifying profitable customers, and predicting the churning individuals by leveraging the limited resources and data availability.

The focus company, SIRCLO, is an e-commerce company that collects valuable transactional data from approximately 500,000 monthly orders from hundreds of brands, containing detailed customer demographics and behaviour. Recently, it has aspired to improve its customer retention and profitability.

1.3 Project Objective

As of 2022, SIRCLO focuses on expanding user growth and industry market share instead of optimising a retention strategy. However, with recent industry development, the management is interested in lowering its customers' attrition rate and prolonging its customers' lifetime value.

To unlock this opportunity, companies must discover the reasons behind customers' churn decisions and design an effective churn model for managing customer retention (Coussement, K. & Van den Poel, D., 2008). In addition to that, a robust retention model enables companies to allocate proper promotions directly to the right customer.

This paper thereby aims to find relevant patterns and predict whether SIRCLO's customers will make a one-off or repetitive purchase behaviour and thus, become high-value and profitable customers. Developing an accurate model also enables the company to target specific customer bases with the highest probability of churn, improving the efficient use of limited marketing resources (Verbeke et al., 2011).

Hence, this project phrased the objective as follows:

"Given thirty days of historical transaction data, how accurately can we predict whether customers' will make another purchase in the next thirty days?"

Figure 1 below shows the data input and prediction workflow. The primary data point consists of detailed transactions from multiple e-commerce sites in Indonesia. Since there is hardly any common customer identification (ID) across platforms, phone numbers are used as unique customer identification. Finally, user features using thirty days historical data are built on top of that ID.

The classification model then consumes the data to predict which IDs are likely to make purchases in the next thirty days. This output becomes the primary data input to the Company's CRM tools.

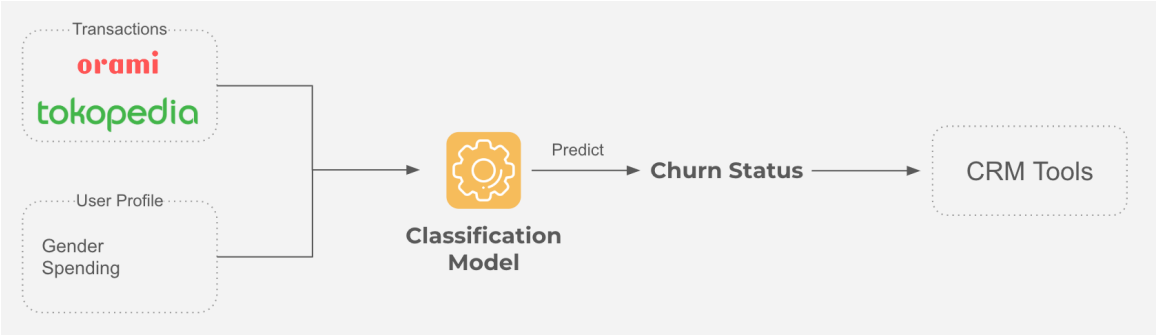


Figure 1 - Overview of project scope

1.4 Report Structure

This report will be divided into six chapters and begins with an introduction of the company, project scope, and motivation. The subsequent chapter explains the relevant models and methodologies used in this paper. Relevant research is then summarised and examined to provide a thorough literature review. Chapter 3 focuses on the available dataset and relevant insights through visualisation and descriptive analytics. Detailed methodologies are explained in Chapter 4, and chapter 5 describes findings and the best-performing model. Finally, Chapter 6 concludes the project results, including its implementation strategy and limitations.

Chapter 2: Literature Review

In the first section, this chapter presents recent studies on digital customer behaviour in Indonesia, our target studies. Then, the following section focuses on detailed studies about customer churn and retention. Lastly, the third section will explain relevant technical models for predicting customers' behaviour and the contribution of this study.

2.1 Indonesia Digital Economy Landscape

Avgerou (2003) and Datta (2010) argue that e-commerce is an essential and critical factor to advance social and economic progress in developing countries.

With the launch of the fast-growing rise of e-commerce platforms in 2016, Indonesia has experienced a significant shift in customer behaviour from an offline-first mindset to digital shoppers. The behaviour shift has been further accelerated by the Covid-19 pandemic (SIRCLO, 2021). From the start of the pandemic alone, Indonesia has seen 21 million new digital consumers, 72% from non-metro areas (Google, 2021). As of 2022, out of its 270 million population (BPS, 2021), 79.2% are digital consumers who intend to continue using digital services in the future (Google, 2021).

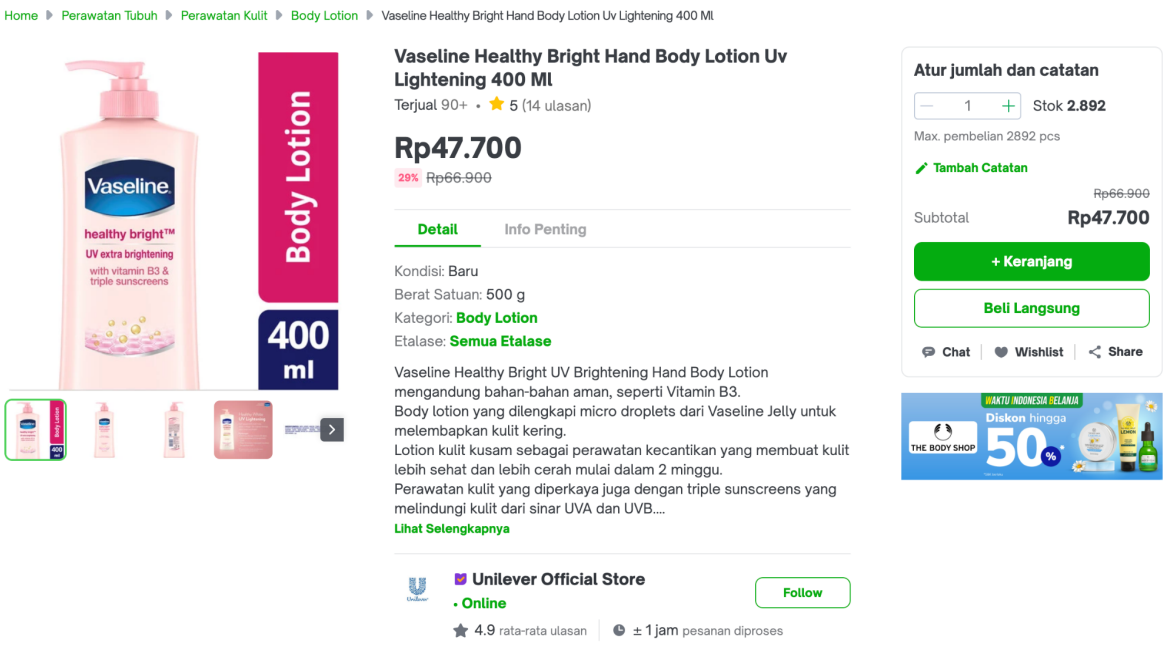


Figure 2 - Product page example from Tokopedia, the biggest e-commerce marketplace in Indonesia with a total monthly visit of 157,233,300 in Q1 2022 (iPrice, 2022).

Tokopedia (IDX: GOTO) and Bukalapak (IDX: BUKA) are known as pioneers in Indonesia's e-commerce landscape. Having a similar business model to the Amazon marketplace but focusing on the Indonesian market, they enable Indonesian consumers to transact safely and securely through the platform. This movement was further strengthened by solid global investors and venture capital funding. Tokopedia raised $147 million and $1.1 billion in 2016 and 2017, respectively. This investment was followed up by another $1 billion in 2018, putting their estimated valuation to $7 billion as of 2019.

Instead of having one dominating marketplace platform, the market heated up when Shopee (NYSE: SE), a regional player from Singapore, entered the Indonesian market in 2016.

A rapidly growing digital-savvy middle class further strengthens this digitalisation pattern (Jurriens & Tapsell, 2017). A recent study shows that 36.8% of consumers shop online at least once a month with an average basket size of Rp250,000, approximately GBP 13, or 20% of average national monthly spending per capita (SIRCLO, 2021). Moreover, pre-pandemic digital consumers have spent 3.6 more services since the pandemic began, with 86% of users satisfied with the services (Google, 2021).

A study of impact shows that the digital economy benefits stakeholders in several ways: increasing discoverability and match between seller and buyer, streamlining price discovery and modifying costs associated with enforcing contractors' contracts (Pangestu & Dewi, 2017).

Moreover, a recent survey shows that digital merchants are increasingly becoming tech-savvy and highly reliant on e-commerce. For example, Google (2021) shows that 28% of merchants believe they would not have survived Covid-19 pandemics without e-commerce. In addition, payment, commonly considered a barrier in a high bankless population like Indonesia, shifted, with 98% of digital merchants accepting payments and 59% accepting digital lending.

With these strong fundamentals, Indonesia's digital economy is predicted to reach $146 billion, maintaining a 20% compound annual growth rate (CAGR) for the next five years (Google, 2021). Its most substantial contribution comes from e-commerce, with $104 billion, reflecting almost 72% of the digital economy. Nevertheless, big as it is, many consider it relatively small and holds more room for growth. Indonesia's e-commerce contributes only

3% of total retail (Ariansyah et al., 2021), while China, the fastest growing e-commerce market, contributes 24.5% of total retail in 2021 (Ma, Y., 2022).

2.2 CRM, Customer Acquisition, Retention, and Churn

Customer acquisition refers to finding new customers by persuading them to buy or utilise the company's products and services. On the other hand, customer churn is the percentage of customers that stopped using a company's products and services, and the percentage of customers that stay is called the retention rate. Customer Relationship Management (CRM) is software to track, record, and manage customer life cycles.

Based on Kumar and Petersen (2021), there are four main primary objectives of investing in CRM: to acquire profitable customers, retain profitable customers, prevent migration of profitable customers, and win back profitable customers from competitors.

Manivannan (2021) shows that customer acquisition and retention are significant concerns in the retail industry. Of these two, companies have recognised the greater importance of identifying churning customers as the market gets increasingly saturated (Hadden, Tiwari, et al., 2005). Another study also finds that cost associated with acquiring new customers is substantially higher than retaining existing ones; thus, churn customer management is critical to survive in the industry (Berger and Kompan, 2019).

Research from significant telecommunications players in Korea shows that the customer churn rate might reach as high as annually (Ahn, Han, & Lee, 2006). In order to manage this attrition, the research finds that companies need to understand contributing factors that make customers churn. Berger and Kompan (2019) classifies these into six main factors: web session of customers, purchase behaviour of customers, behaviour changes between sessions, user interaction with the website, actions made during interactions, and ratings given from customers from the last purchase. Out of those attributes, research shows that attributes related to user interaction is the best indicator of churn intent, while purchasing becomes the second best (Berger and Kompan, 2019).

Churning customers itself can be categorised into two categories: non-voluntary and voluntary churns. The first group, non-voluntary churners, are customers whose access has been revoked or banned by the company because of violating terms and conditions or other

reasons. This part is easier to predict because the company has total control and is able to specifically define the reason. On the contrary, the voluntary churners are hard to understand due to the limitless possibilities and motivation behind the decision. These conscious decisions of not making another purchase can be affected by service quality, moving to competitors, financial struggles, and many more (Hadden, Tiwari, et al., 2005).

The complexity of understanding the intention and reason of a customer to make a purchase or churn makes the need to develop a machine learning model of customer behaviour in high demand.

2.3 Data-driven Customer Management

CRM has become critical for retail customer companies that implement online and social media strategies in this information era. More often than not, retail companies serve abundant users with their unique attributes and behaviour. Therefore, modern online CRM should be designed with customer-centric and data-driven analytics to understand business and customers' information (Chiang, 2019).

Research shows that enabling data-driven decisions across organisations may improve productivity by 5% - 6% while considering further investments and the leverage of information technology (Brynjolfsson, Hitt and Kim, 2011).

Popularly known as big data analytics (BDA), this analysis needs more sophisticated tools to process information effectively and efficiently. For example, due to human limitations, discovering patterns hidden in a million rows of data cannot be easily obtained. To overcome the computational barrier in analysing enormous data, the discovery process must be enhanced by machine learning (ML) and artificial intelligence (AI) (Ranjan & Foropon, 2021).

One of the essential elements of data-driven decision-making comes from predictive analytics using the ML model, which may help companies prepare for appropriate future action (Kraus, Feuerriegel and Oztekin, 2020). However, there are still some barriers for companies to adopt this novel and fast-growing method. A recent survey shows that the question is "how", rather than "if", to maximise the use of ML and predictive analytics in the business and improve companies' performance (Kiron & Schrage, 2019). Furthermore, if business users are not empowered to understand the result, BDA will only provide little

value. AI and ML have great potential, but only if decision-makers can interpret, use, and evaluate it properly (Kozak et al., 2021).

Based on the complexity of BDA and the importance of predictive analytics, this paper tries to solve CRM and customer retention problems using ML techniques. The subsequent chapter explores methodologies and approaches taken to answer this objective.

2.4 Classification Model

Classification is one of the most popular types of supervised machine learning models. Unlike unsupervised machine learning models, where models learn the pattern of unlabeled data, supervised machine learning depends on a targeted label defined beforehand. Specifically, classification models are trained using labelled data with specific and defined classes, and it learns how to segregate other data into its relevant classes (Geron, 2017).

2.3.1 Decision Tree

The decision tree is a supervised learning method for regression, numerical processing variables, and classification, processing categorical variables. The fundamental idea of the decision tree is to predict an output by learning to create certain boundaries from the training data.

One of its main advantages is its simplicity. This algorithm is robust because it does not need complex data preparation. Furthermore, the model's output can be visualised to make it easier to interpret and fine-tune.

On the other hand, it may create an overfitting model due to over-complex trees. A decision tree in its simplest form is also susceptible as a simple movement in the data may change the model. However, this can be mitigated by using ensemble methods. Unbalanced learning data may also create a significant problem for this algorithm, as it learns to predict the false majority instead of the true minority.

2.3.2 Random Forest

The random forest is an improvement from the decision tree method and part of an ensemble algorithm, which combines several prediction results to give a more robust result over a single estimator. This algorithm builds several trees from several bootstrap samples, a

set of samples from the dataset drawn with replacement, and a decision tree algorithm as its base algorithm. Then, it combines the prediction results from all the trees and predicts the class with the most votes.

This method overcomes several limitations of the decision tree. First, it has built-in randomness to cancel out errors and reduce overfitting. Second, it can grow into a complex tree without losing much accuracy due to unforeseen data, improving prediction quality (Tin Kam Ho, 1995). Moreover, it improves model stability and reduces sensitivity due to data variation while maintaining its easy-to-use and processing speed.

2.3.3 Support Vector Machines (SVM)

SVM is a versatile supervised machine learning method for regression, classification, and outlier detection. This method works well in high-dimensional data, even if the number of features exceeds the number of samples. The algorithm may implement a regularisation method to avoid overfitting when the data has many features. SVM uses a subset of training data to have efficient memory usage. However, its output does not directly show probability estimates and is very sensitive to feature scales. Therefore, scaling the data is highly recommended while using this method.

2.3.4 K-Nearest Neighbours

Commonly used for regression and classification problems, k-nearest neighbours is a supervised machine learning method whose primary idea is using proximity to classify data into a certain number of groups, denominated by K.

In its basic form, it is computed from a simple majority vote: a single data point is assigned to a group with the highest representatives within the nearest neighbour of the data. However, this algorithm can accommodate weighted neighbours, which may contribute to a better fit.

As K is a user-defined variable, it may be challenging to define the K value. Therefore, one possible solution is to run KNN using various K-values and select one with the lowest error rate.

2.3.5 XGBoost

XGBoost is one gradient boosting method, a classification and regression machine learning method that gives predictions from an ensemble of several weak models, with a decision tree being the most common base model. This method usually outperforms the Random Forest algorithm.

Gradient boosting optimises an arbitrary differentiable loss function, enabling a better-performing prediction. On the other hand, gradient boosting generally sacrifices interpretability for performance, as it becomes increasingly harder to interpret multiple combinations of trees from the same dataset.

2.3.6 Neural Network

A neural network is part of deep learning that utilises a multi-layer network of neurons to make predictions. Like other machine learning algorithms, the neural network can be used for supervised and unsupervised learnings, classification or clustering, and even time series prediction.

In general, it has three primary types of layers. The first is the input layer, where neurons take input from the training data. Meanwhile, an output layer is the end layer where neurons output the prediction results. Between that layer can be multiple intermediate hidden layers with multiple combinations of neurons. These hidden layers are the core of the neural network. Artificial neurons in hidden layers will process the input based on predetermined activation and output the weighted result to be passed on to the next layer.

Further explanations of neural network parameters and models are described in Chapter 4: Methodology.

2.3.7 Voting Classifier

The voting classifier is a part of ensemble learning that trains and predicts several model classifiers and outputs the aggregate predictions of all base models. This method combines weak predictions and expects a more robust model.

There are two types of ensemble learning:

Hard voting: aggregate prediction based on the prediction of base models

Soft voting: aggregate prediction based on the prediction probability of base models

2.5 Feature Importance

Understanding and interpreting models is one of the most challenging parts of Machine Learning. Accuracy scores and other metrics do not necessarily mean the model can find a suitable business use case. Therefore, understanding features and how they affect output is critical in many applications. Unfortunately, the development of machine learning algorithms makes this even harder. Some models, like ensemble learning, become more and more like a black box and make accuracy and interpretability a trade-off.

This project uses SHapley Additive exPlanations (SHAP) to interpret feature importance. Based on a game-theory unified approach by Lundberg and Lee (2017), SHAP assigns a value to each feature based on its importance in predicting output value. This relatively new method aligns better with human intuition and enables easier understanding than the older features importance method.

2.6 Gap and Contribution

While the importance of churn management and the use of ML in BDA is highly discussed, relevant studies such as Kozak et al. (2021) focus more on evaluating the algorithm's performance in data-driven churn management problems. In addition, more traditional research about churn management spotlighted other industries, such as telecommunication (Ahn, Han, & Lee, 2006) or airlines (Chiang, 2019).

Compared to previous studies, this project aims to fill the gap by predicting customer retention using transaction data to support decision-making. The data comes from a fast-paced e-commerce industry in a growing Indonesian market that still lacks relevant research. The output will then be implemented in the internal CRM tools, opening more possibilities for further research about its effectiveness.

Most importantly, this paper tries to answer the "how" by giving a realistic example of ML application in a business setting by leveraging limited data commonly available to e-commerce players and relevant algorithms.

Chapter 3: Data Exploration and Pre-processing

3.1 Dataset

3.1.1 Dataset Description

SIRCLO provided all transaction data for this project and analysis. They classify their data into three parts: order management system, operational, and customer data platform. Order management system handles all day-to-day transactional data, designed for speed and efficiency to handle millions of updates daily. These data are then sent via messaging middleware into Netsuite, SIRCLO internal enterprise resource planning (ERP) software, to be processed and fulfilled by the operations team. Concurrently, customer data is updated into the customer data platform as a single source of truth of all customers' journeys across SIRCLO Group.

The company has established company-wide data pipelines from several sources using Apache Airflow (airflow.apache.org) as orchestrator, extract load and transfer (ETL) process using DBT (www.getdbt.com), and then collects everything in Google Bigquery as their data warehouse (DWH). This DWH is a core component of the group's business intelligence and analytics service. This project utilises the company's data mart from the DWH, a stable data layer commonly used as the primary table for internal analysis.

The table covers 21 months of transactions from July 2020 to March 2022. It consists of 17,153,806 rows and 28 columns, consisting of three main variables: order details, customer details, and internal fulfilment details. Order details show the store where customers buy their products, the e-commerce platform where the transactions happened, product name, product category, quantity, value, and shipping type, among others. The customer details section shows delivery address, cities, provinces, phone numbers, and other buyer details. Finally, internal fulfilment details present warehouse origin, internal sales person, and order status.

Each row records the order transaction line details. For example, if a customer buys five different products, it is recorded as five rows with the same order number. However, the product name and SKU are unique in every order; no matter how much the order quantity for that product, it is only created once. Table 1 below is a transposed sample of SIRCLO

transaction data along with a brief description for every column. All values given in the table are artificial to mask buyer's personally identifiable information.

Table 1 - Descriptive Table of Raw Dataset (values are for illustrative purposes only)

Column Name	Value	Description
Order_Tstamp	2021-09-30 16:47	Order timestamp
Order_Number	PULI1000332558	Internal unique order number
Customer_Reference	INV/20210930	Reference number from external parties
Effective_Date	2021-09-30	Internal process date
Status	Sales Order	Internal order status
Marketplace	Tokopedia	E-Commerce platform as sales channel
Official_Store	AAAA Official Store	Brand official store name in sales channel
SKAM	Budi	Internal sales code
Warehouse	Depok	Internal warehouse origin
Buyer	Tono	Buyer name
Gender	L	Buyer gender: L = Male and P = Female
Phone	08188188188	Buyer phone number
Address	Jalan Jakarta	Buyer detail shipping address
City	Jakarta	Buyer city
Province	Jakarta	Buyer province
Postcode	16954	Buyer postcode
Shipping	TIKI - Reguler	Shipping type
AWB	0001234555	Airwaybill or shipping tracking number
Principal	AAAA	Brand principal name
Category	Body Wash	Product category
SKU	LYA1G16002	Product SKU
Product	Sabun Cair	Product Name
Replacement_For	NaN	Internal SKU replacement code
Barcode	8999999515577	Product Barcode
Qty_Order	1.0	Quantity order
Qty_Delivered	1.0	Quantity delivered
Value_Untaxed	67181.82	Value without tax
Value_Taxed	73900.0	Value with tax

3.1.2 Dataset Cleaning and Aggregation

As this project intends to analyse customers' purchasing behaviour based on their orders, the raw data need to be aggregated by Order Number to build an aggregated summary of

customers' shopping patterns. Therefore, this table has unique order numbers and aggregated order quantities and values.

In order to prepare suitable and proper data required for further analysis, data cleaning was performed. First of all, null values were dropped, and cancelled orders were removed. The business-to-business transaction was also cleaned out as it has a different pattern and business strategy, therefore out of this project scope.

Our final order table consists of 8,942,963 eligible orders from 3,084,072 unique customers. The figure below shows 17 columns from the order table.

Table 2 - Descriptive Table of Final Dataset (values are for illustrative purposes only)

Column Name	Value	Description
Order_Tstamp	2021-09-30 16:47	Order timestamp
Order_Number	PULI1000332558	Internal unqiue order number
Status	Sales Order	Internal order status
Marketplace	Tokopedia	E-Commerce platform as sales channel
Official_Store	AAAA	Brand official store name in sales channel
SKAM	Budi	Internal sales code
Warehouse	Depok	Internal warehouse origin
Buyer	Tono	Buyer name
Gender	L	Buyer gender: L = Male and P = Female
Phone	08188188188	Buyer phone number
Address	Jalan Jakarta	Buyer detail shipping address
City	Jakarta	Buyer city
Province	Jakarta	Buyer province
Postcode	16954	Buyer postcode
Shipping	TIKI - Reguler	Shipping type
Principal	AAAA	Brand principal name
Qty_Order	1.0	Quantity order
Value_Taxed	73900.0	Value with tax

Moving forward, this order table is used as the main data and table for analysis.

3.2 Exploratory Data Analysis

Exploratory data analysis (EDA) is an analytical framework that provides conceptual and computational tools to discover patterns in the data to build and sharpen hypotheses. In

addition, EDA helps interpret the analysis results and may show identified, unexpected, or misleading patterns (Behrens, 1997).

The exploration in this chapter was performed to understand the customer profile and discover interesting patterns that are useful for feature engineering in the next step. In order to provide a comprehensive overview of the order table, this chapter is divided into three sections. Brief industry and geographical context are included in each analysis to provide a more thorough background for readers.

3.2.1 Gross Merchandise Volume and Order

The company considers GMV as one of the most critical metrics. In the e-commerce and online retail industry, GMV means the total value of merchandise sold across time through the exchange platform. This topline metric is commonly considered the performance benchmark for growth and measures the overall health of e-commerce players, as also shown in Yan et al. (2017) studies.

Figure 4 shows that SIRCLO's weekly GMV grew from 6 to 20 billion Rupiah, Indonesian local currency, while orders grew from 50,000 to 120,000 in the past six quarters. Moreover, there was a significant increase in both metrics between weeks 45 and 49 likely due to Indonesia's online platforms and e-commerce marketplaces doing an annual shopping festival during this period.

Figure 3 - Example of online shopping festival banners from Tokopedia and Shopee, specifically for 12.12 or 12th December shopping day. This promo date is similar to 11.11 in China and Cyber Monday in the USA.

On the other hand, the lowest sales happened in the first week of January, probably due to the long year-end holiday and the significant buyer going out of town. Small spikes that happened every four weeks, though not significant, were likely caused by payday promo by marketplaces. This marketing campaign specifically targets working people, a significant portion of Indonesia's population.

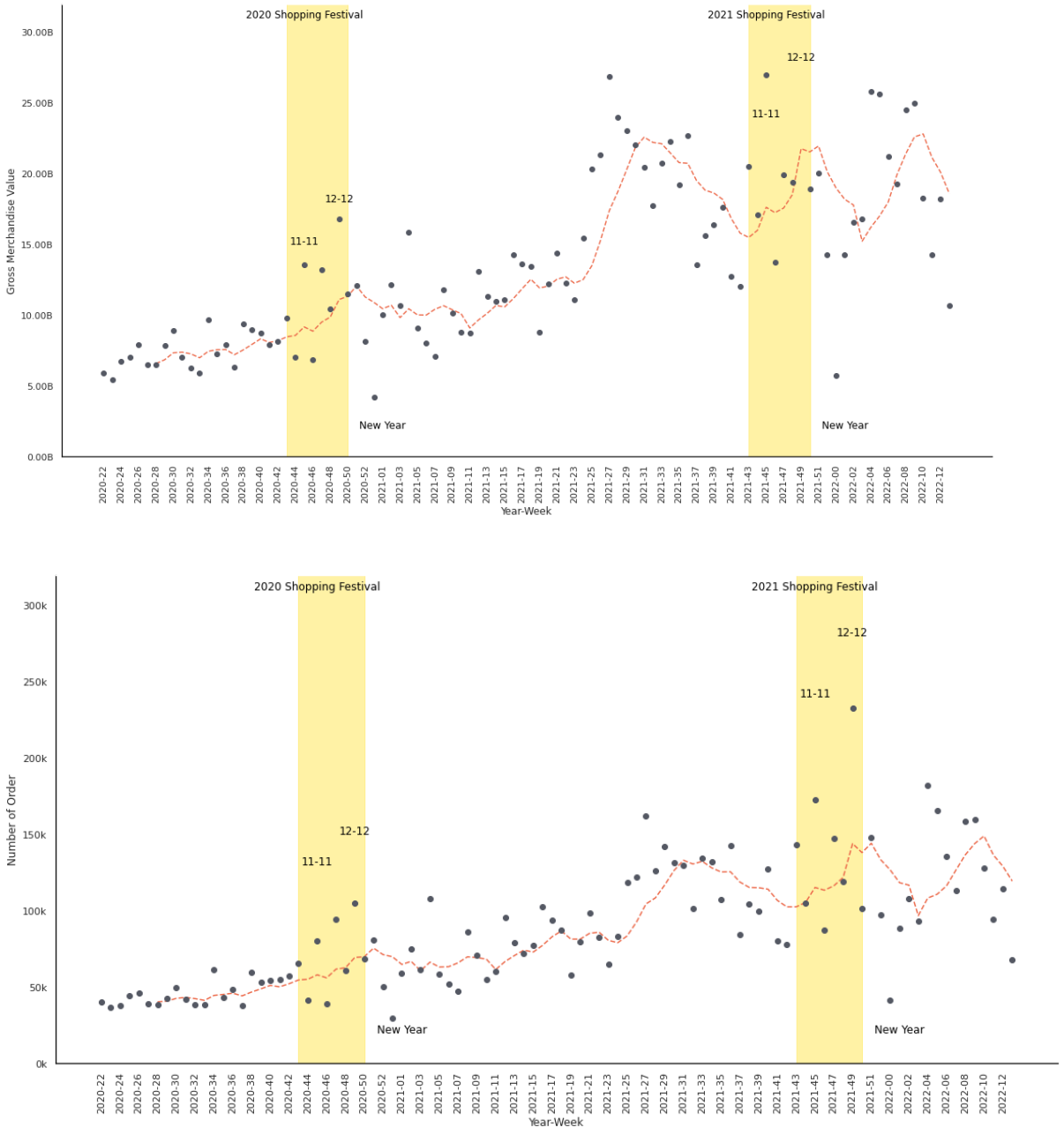


Figure 4 - SIRCLO's weekly number of orders and GMV in Rupiah (local currency of Indonesia). It shows an increasing trend from 2020 to 2022. The area highlighted in yellow emphasises the online shopping festival, the most significant e-commerce promotion period in Indonesia. Dash line shows 7-weeks moving average.

More interesting patterns in these metrics, such as gender behaviour and urban vs rural differences, are explained later in this chapter.

3.2.2 Gender Behaviour Analysis

Previous section shows a general increasing trend of SIRCLO's GMV. In a deep-dive analysis to find the contributing factor, different shopping behaviour between males and females emerges.

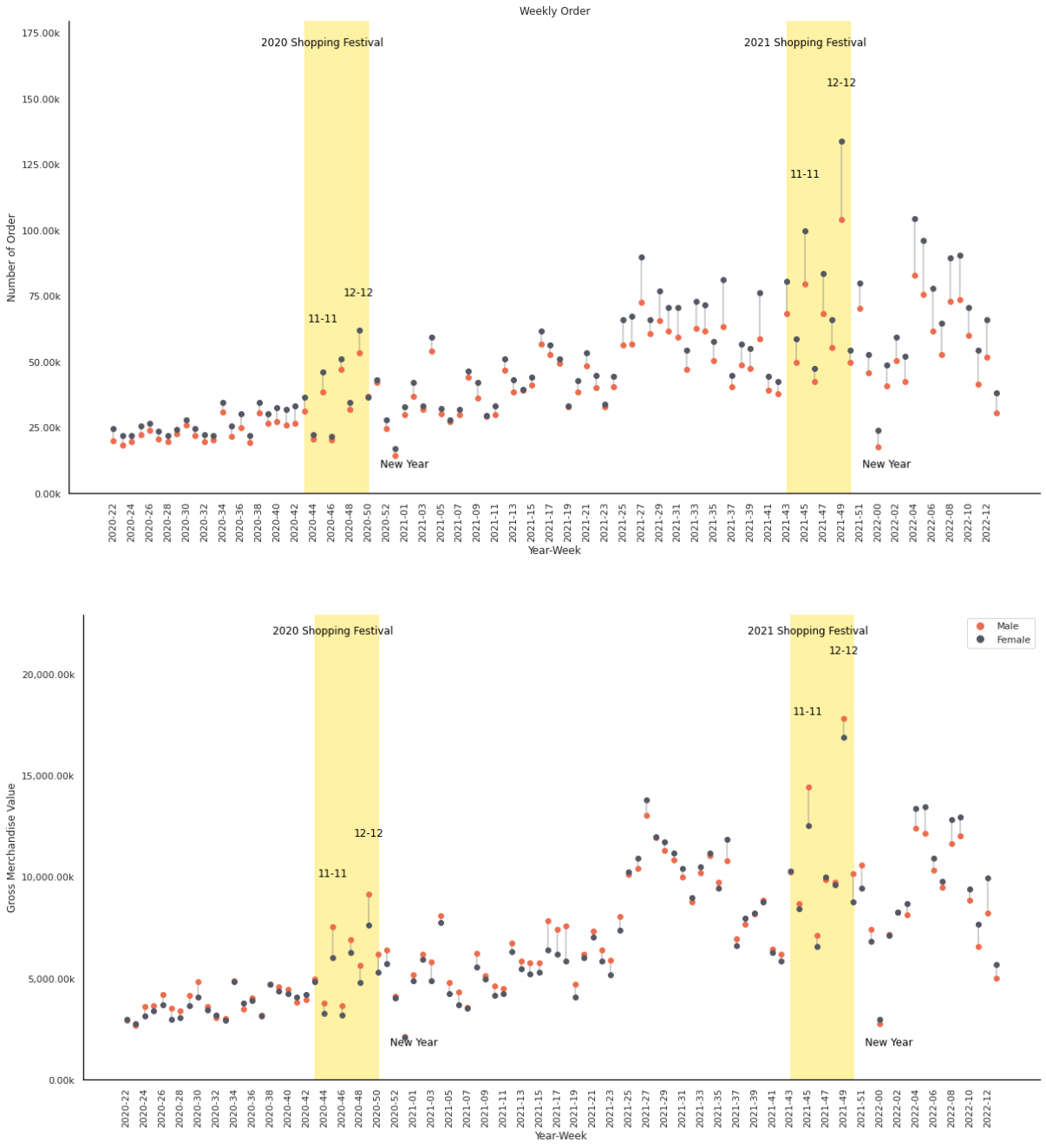


Figure 5 - SIRCLO's weekly number of orders and GMV in Rupiah (local currency of Indonesia), grouped by

gender. In general, males contribute higher GMV in SIRCLO before weakening and overlapping with their female counterparts in 2021.

Figure 5 shows that male buyers generally contributed a higher portion of weekly GMV, especially during the first half of 2021. However, this trend weakens in the second half as more female buyers outspend their male counterparts.

Figure 6 - SIRCLO's weekly number of orders, grouped by gender

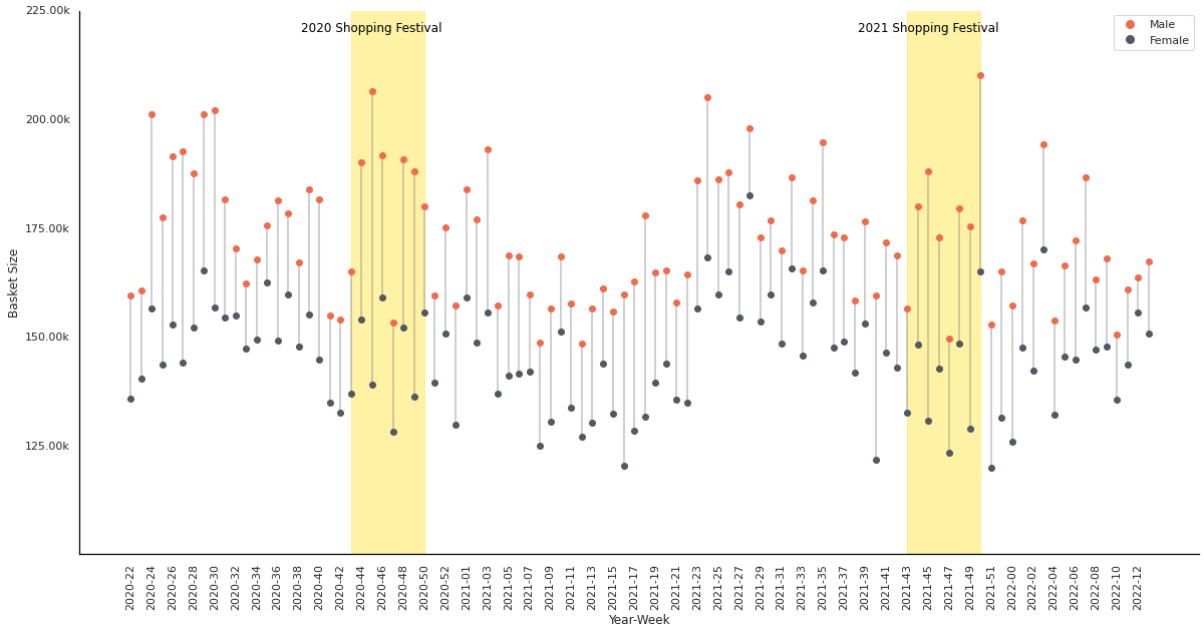


Figure 6 - SIRCLO's weekly average basket size, grouped by gender. This chart shows that males have higher order value per transaction than females.

Clearly, GMV is derived from price and quantity, further analysis on Figure 5 shows an interesting trend: female buyers transact more orders per week than males. On the other hand, male buyers tend to buy higher value in every order, as shown in Figure 6, approximately 20% greater than female buyers.

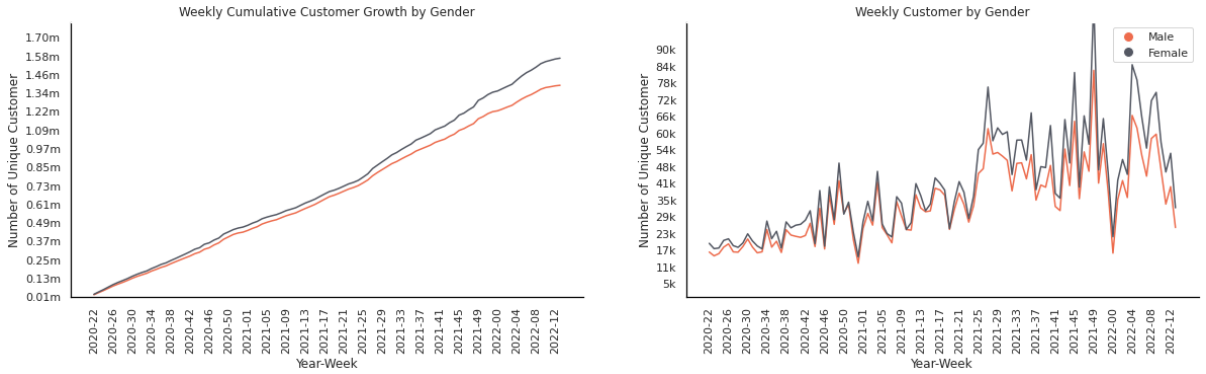


Figure 7 - SIRCLO accumulative customer growth and weekly customer by gender. Left chart shows SIRCLO has a larger female customer base. Right chart shows weekly customer acquisition by gender.

On average, male and female customers buy 2.7 orders during their lifetime; however, Figure 7 shows that SIRCLO has a higher female customer base. In conclusion, while males have higher order value per transaction, females buy more occasionally; the latter contributes more GMV in SIRCLO because the company has a higher female customer base. The following chart, Figure 8, also shows that gender has significantly different basket sizes and orders and behaves differently based on brands and regions.

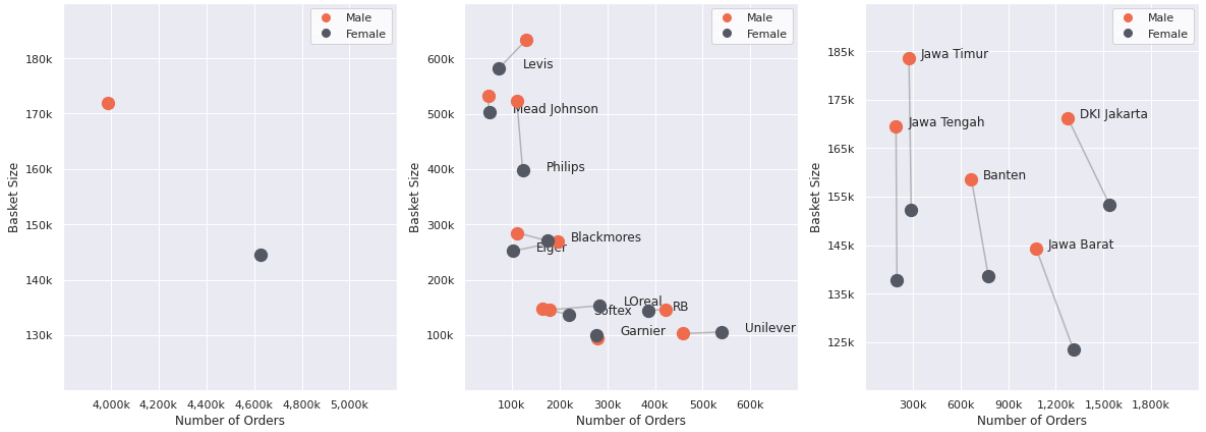


Figure 8 - SIRCLO weekly basket size, by gender, by principals, and by region.

To conclude, there is an interesting contrasting buying behaviour between males and females; therefore, this report will include this as a feature in the model.

3.2.3 Regional Analysis

Like gender, provinces where customers live affect GMV distribution and show exciting patterns of customer shopping behaviour.

Indonesia has a diverging range of spending and economic power per capita between provinces. For example, DKI Jakarta, the capital and largest metropolitan in the country, has 13x higher Gross Regional Product (GRP) per capita than the lowest province, East Nusa Tenggara (BPS, 2020).

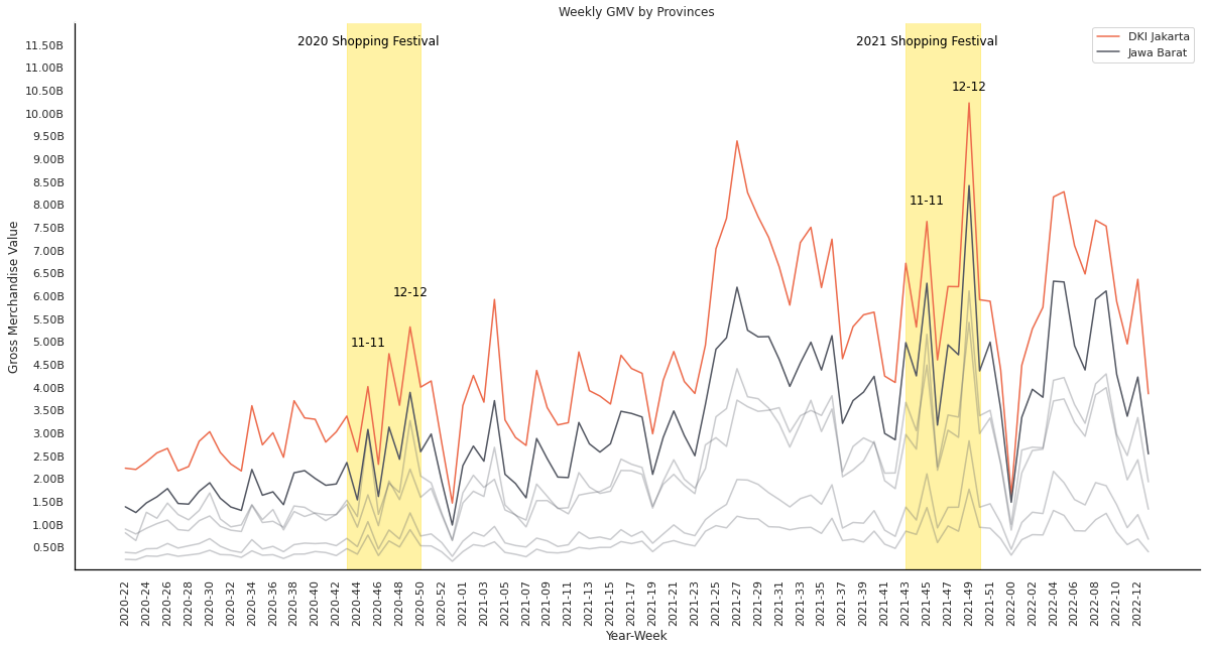


Figure 9 - SIRCLO weekly top 5 region GMV sorted by total value. DKI Jakarta, the capital and biggest metro, outperforms all other provinces.

Furthermore, Indonesia has more than 17,000 islands and spans 5,100km from east to west that makes shipping and logistics, which are critical factors for the growth of online commerce (Google, 2016), difficult and expensive. For instance, shipping one kilogram of a package differs from Rp9,000 within DKI Jakarta to Rp105,000 from DKI Jakarta to Jayapura in the east, almost twelve times higher. A full map of Indonesia can be seen in Appendix A.

This economic factor and shipping cost affects customers' might affect online spending significantly. SIRCLO's internal analysis finds a positive correlation between shipping cost and average order value (AOV). On the other hand, shipping cost negatively correlates with the number of orders. In short, customers far from DKI Jakarta, where most merchants ship, tend to buy less frequently with a more considerable AOV.

Figure 11 below observes that the top three provinces with the most significant number of orders have the lowest AOV. On the other hand, provinces with smaller orders tend to have

greater AOV. Furthermore, to eliminate bias due to differing populations, the chart on the right shows the average number of orders per customer, grouped by their provinces.

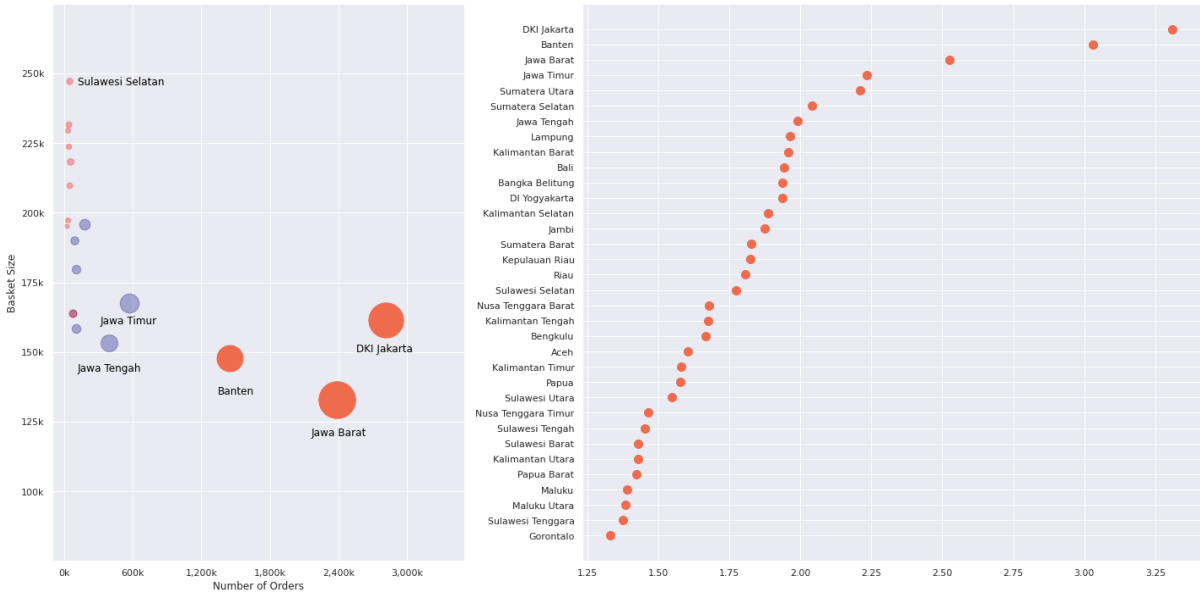


Figure 10 - SIRCLO basket size and number of orders by region. DKI Jakarta and neighbouring provinces, Banten and Jawa Barat, have the highest number of orders but lower average basket size. This might indicate negative correlation between shipping cost, basket size, and region.

3.2.4 Dependent Variable Formulation

Before generating the dependent variable, clearly defining recurring and churning customers is fundamental.

In the Software-as-a-Service (SaaS) industry, like Spotify, churning customers are usually identified when customers' subscription ends or there is an absence of interaction between customers and the service. However, defining recurring and churning customers in online retail is delicate and has no generally accepted definition. Unlike SaaS, shoppers may make platform transactions without any commitment or subscription beforehand.

This project defines churning customers as those without subsequent purchases in the next thirty days. There are several reasons for this. First, domain experts from the company argue that most shoppers are susceptible to promotion and marketing. This behaviour happens due to the relatively fast-paced nature of e-commerce, where promotion may span from a few minutes or even seconds. Impulse buying behaviour, therefore, may be formed.

Subsequently, this project aims to improve marketing and promotion effectiveness, typically formulated every month to align with the e-commerce pace. Finally, monthly analysis aligns with SIRCLO's cycle of the external report to the brand owners.

This project denotes the binary dependent variable for every transaction by looking at the next thirty days based on the order timestamp for that specific customer. A value of 1 means the customer makes a subsequent purchase within 30 days and is defined as a non-churning customer. Vice versa, when the customer does not purchase anything within the period, the value of the dependent variable equals 0 and is called a churning customer.

For instance, Customer 1 made a transaction on the platform on 20 January 2022. Then, fourteen days later, the customer made another purchase and therefore was defined as non-churning for that transaction.

However, Customer 1 did not make another purchase in February and March 2022. Hence, the customer is designated as churning, and the dependent variable is 0.

Thirty days prior data, called the "Feature Extraction Period", is collected to generate customers' features. For example, features for transactions that happened on 30 March are generated from 1-30 March data. The figure below demonstrates this process.

3.2.5 Dependent Variable Exploration

This section explores the dependent variables generated in the previous section. The figure below explores patterns between recurring and non-recurring customers, with the former having a lower value in average basket size and number of orders. The middle chart shows that there is more churn than retaining transactions. On the other hand, on average retaining customers buy three times more orders than churning customers.

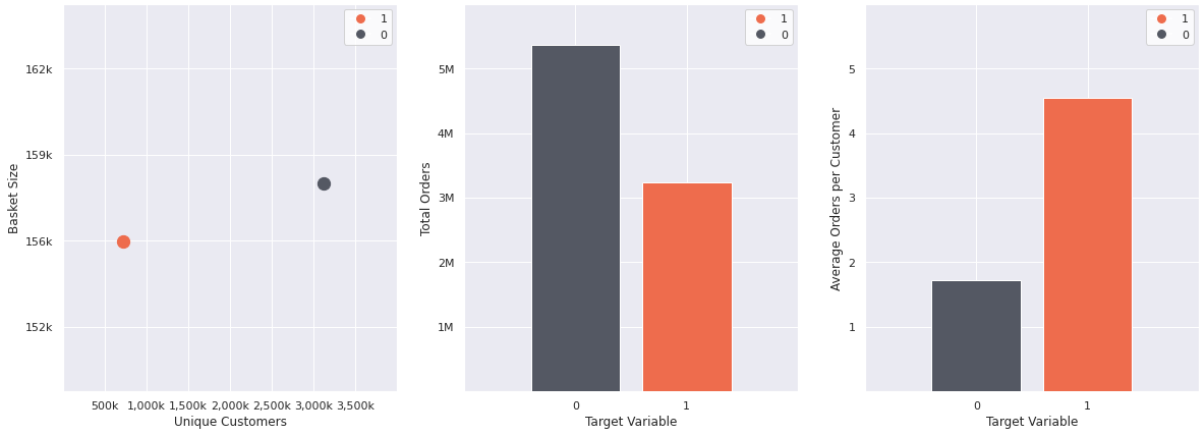


Figure 11 - SIRCLO basket size and number of orders by region, grouped by target variable. The leftmost figure shows target variable 0 has a larger basket size and unique customers. The centre and right figures show that target variable 1 has a lower total order. However, it has a higher average order per customer.

From a geographical perspective, there is no observable difference between retaining and churning customers. DKI Jakarta and Banten held the highest non-churning customers by percentage, while Sulawesi and Papua Barat, the easternmost and farthest island from the capital, had the lowest non-churning customers.

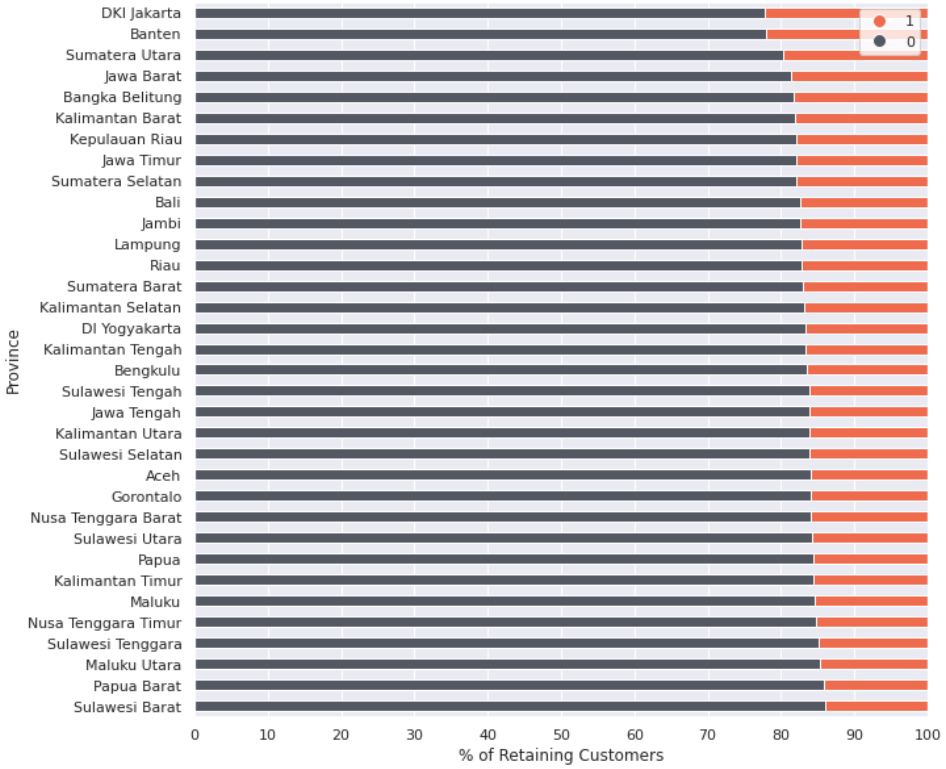


Figure 12 - SIRCLO basket size and number of orders by region. DKI Jakarta is the province with the most returning customers, while Sulawesi Barat is the lowest. Map of Indonesia can be seen in Appendix A.

3.2.6 EDA Summary

In summary, EDA shows that customers behave differently depending on their region, gender, and several other factors. Non-churning customers generally have a higher basket size, more quantity, and explore more stores while shopping. They are also more likely to use express and same day shipping, which are typically more expensive than regular shipping. These factors will be useful for feature engineering in the next chapter.

Chapter 4: Methodology

This study aims to predict customer retention using a new, more straightforward approach, leveraging companies' internal transaction data exclusively in the model. Feasibility studies of this method are critical due to limited public data availability in the e-commerce industry, especially in South-East Asia, which makes the conventional model challenging to implement. The following section discusses features engineering processes based on the dataset discussed in Chapter 3. Furthermore, the subsequent section explains the model selection and parameter tuning thoroughly. Finally, the metrics evaluation method and importance are discussed.

4.1 Feature Engineering and Selection

Table below shows final features used in the model:

Table 3 - Final features for training and test datasets

Category	Features Code and Name		Description
Order Detail	OMYE	Order Month Year	Month and year of the order
Order Detail	OQTY	Order Quantity	Order Quantity
Order Detail	OVAL	Order Value	Order Value
Historical Data	PRMO	Prior Month	Last previous order month from the customer
Historical Data	TVAL	Total Value	Historical order value
Historical Data	TQTY	Total Quantity	Historical order quantity
Historical Data	CNOD	Count Order	Historical count of order
Historical Data	CNOS	Count OS	Historical count of OS
Historical Data	CNPR	Count Principal	Historical count of principals
Historical Data	PRNC	Principal	List of all principals ever bought
Order Detail	CRMO	Current Month	Current order month
Other Feature	MOR3	More Than 3	1 = sum of CNOD, CNOS, and CNPR > 3
Other Feature	JAVA	In Java	1 = province is located in Java island

4.1.1 Current and Prior Order Month

Chapter 3 demonstrates how the number of orders and GMV fluctuates heavily based on months. Heavy discounts and promotions during certain months may push customers to make impulse buys and affect churning behaviour. Therefore, this study creates these

features to capture seasonality movement. Current Month (CRMO) is generated from Order Timestamp. Likewise, Prior Month (PRMO) is derived by finding the nearest prior order from that transaction for each customer, then using its Order Timestamp.

4.1.2 Count of Principals, Official Stores, and Order

Principals are brand owners, usually multinational corporations such as Unilever, Levi's, L'Oreal, Reckitt Benckiser, and Nivea. Having the right to manufacture and distribute their products globally, they are given Official Store status for their brand in the marketplace. This status helps increase sellers' credibility and customers' trust to transact within the C2C platform, which usually has a high rate of fake products. The relationship is one to many: each principal can have more than one official store; however, one official store can only be managed by one principal.

Making purchases from diverse principals and official stores indicates customers' proficiency and familiarity with e-commerce platforms. It also shows trust and habits have begun to form, which may affect churning probabilities.

Similar to the count of principals (CNPR) and count of official stores (CNOS), this study aggregates the number of orders (CNOD) from each specific customer. All these features are generated using the last thirty days of data.

4.1.3 Combination of Three Indicators

To strengthen the signal further, this study creates a new feature called MOR3, derived from the total of CNPR, CNOS, and CNOD. When the sum of those features is more than three, MOR3 is denoted as 1; and vice versa, MOR3 equals 0 if the total is below 3.

The intuition behind this feature is that customers who purchase more than one from various official stores and principals are more likely to make another purchase. This pattern shows consumer trust on the platform, which has been identified as a challenge for the e-commerce industry (Google, 2019).

4.1.4 Total GMV and Quantity

These features aggregate total transaction value (TVAL) and product quantity (TQTY) customers buy for the last thirty days. Total transaction value is affected by how often

customers make a transaction and what category they usually buy. Generally, groceries have lower transaction value than fashion and beauty products while having higher quantity.

4.1.4 Province Located in Java Island

Java is the most extensive and most populous island in Indonesia, contributing almost 60% of Indonesia's GDP (BPS, 2021). Previous analysis in Chapter 3 shows that provinces where customers live correlate with different shopping behaviour. For example, DKI Jakarta, Banten, and Jawa Barat hold the top three highest orders and values; all are provinces in Java. Therefore, this feature (JAVA) is generated to capture that pattern. Transactions from provinces in Java island are denoted as 1, and otherwise provinces outside Java as 0.

4.1.5 Omitted Features

Several features have been explored and omitted to improve model performance. For example, though exploratory data analysis in Chapter 3 shows a different trend between male and female purchasing patterns, the model does not consider that an important signal; instead, it decreases accuracy by around 5%. In addition, the company's internal operation data, such as the name of the key account manager (sales) or warehouse name, also become noise for the model. Therefore, these features are omitted from the final model. The complete table and results can be seen in Appendix B.

4.2 Data Splitting

Data splitting is an essential step in the machine learning process. By taking this measure, we divide the data into training and test data. Training data is used as an input to the machine to learn the pattern. On the other hand, test data is reserved and not made available until the machine finishes its learning. Afterwards, it checks and evaluates the models based on relevant metrics using unseen data. This guarantees the model performance is unbiased and more reliable. Faraway (2016) shows that data splitting is superior than using an entire dataset, unless for niche circumstances. Using complete data for model training may make it unable to generalise well and leads to overfitting, where training performance is high but does not perform well in unseen data.

This project splits the data into 70:30 for training and test data. Stratified sampling is also used to ensure both datasets preserve the same distribution of classes.

4.3 Feature Scaling

Scaling means transforming the data into a specific scale. Some models are sensitive to values; therefore, in general, scaling may improve predictions by giving all the features the same importance and proportion. Roy (2020) also proves that models may run much faster using scaled features.

In general, there are two methods of data scaling: standardisation and normalisation. Standardisation means transforming the data to have a 0 mean and variance of 1. Meanwhile, normalisation is bounding the data into a specific scale, usually between 0 and 1.

This project utilises the standard scaling method from the SKLearn library, which standardised features by removing the mean and scaling it to unit variance.

4.4 Model Algorithm and Parameter

4.4.1 Baseline Model

Baseline models act as the benchmark for machine learning projects, and all advanced models' performance is measured against the baseline model performance. This model is by nature not complex and has little or no hyperparameter tuning options.

This project chooses a decision tree classifier by the SKLearn library as a baseline model. Cross-validation with five folds validation is performed to avoid overfitting that may lead to lower performance.

4.4.2 Advanced Model

The first advanced model used in this project is Random Forest, using the SKLearn library. In addition, cross-validation with five fold validation and hyperparameter tuning using GridSearchCV is performed.

There are three hyperparameters to explore: maximum depth, minimum sample leaf, and minimum sample split. Maximum depth shows the maximum depth of each tree in the forest. A deeper tree can capture more data but demands a longer training time. It may also be unable to generalise well and leads to overfitting. The second parameter, the minimum sample leaf, is the required minimum number of nodes for each leaf. Moreover finally,

minimum sample splits tune the number of samples required before splitting the internal node.

The table below shows the parameters used in GridSearchCV for Random Forest and brief descriptions for each parameter.

Table 4 - Random Forest GridSearchCV - Parameter Tuning

Parameters	Parameters Value	Description
max_depth	25, 50, 100	Maximum depth of each tree in the forest
min_samples_leaf	25, 50, 100	Minimum number of nodes for each leaf
min_samples_split	25, 50,100	Minimum sample splits before node splitting

Consequently, XGBoost is chosen for the gradient boosting algorithm using precision as evaluation metrics. This model has various parameter tuning and is relatively fast compared to other advanced models. However, the training time and power consumption correlate with the data size. Due to the limitation of computing power, this project selects several critical hyperparameter tuning based on their importance.

Minimum child weight and maximum depth are chosen to find the minimum weight required in a child and the maximum depth of the trees. Subsequently, gamma is tested to find the optimum minimal loss reduction required to split a node. Furthermore, the subsample and colsample bytree are selected to find the minimum loss reduction required for node splitting and the number of random samples needed for each tree. Finally, the learning rate is assessed to ensure the model is robust.

GridSearchCV are performed with tested parameters listed below.

Table 5 - XGBoost GridSearchCV - Parameter Tuning

Parameters	Parameters Value	Description
min_child_weight	1, 3, 5, 10	Minimum sub of weights required in a child
max_depth	1, 3, 5, 10	Maximum depth of a tree
gamma	0.0, 0.1, 0.2, 0.3	Minimum loss reduction requires to split a node
subsample	0.0 - 1.0	Random samples of observations for each tree
colsample_bytree	0.0 - 1.0	Random samples of columns for each tree
learning_rate	0.01, 0.1, 0.5	Learning rate of the model

Best parameters for each advanced model are summarised in Chapter 5 section 1: Model Evaluation.

4.4.3 Neural Network

This project uses Adam as an optimisation algorithm for gradient descent. Kingma & Ba (2017) introduces this method as optimisation of stochastic gradient descent. It has several advantages, such as being more efficient, faster, and well suited for data with big datasets and many parameters.

The total layers are six, including four hidden layers, with a learning rate of 0.0001. Multiple combinations of layers, epochs, and batch sizes are explored. However, the final model is summarised in the table below.

Table 6 - Neural Network - Model: Sequential

Layer	Output Shape	Parameters #
Input Layer	500	62,500
Hidden Layer 1	250	125,250
Hidden Layer 2	100	25,100
Hidden Layer 3	75	7,575
Hidden Layer 4	25	1,900
Output Layer	1	26

4.4.4 Voting Classifier

The main goal of ensemble learning is to combine some weak models and enhance predictive results. The previous section has explained the implementation of Random Forest and XGBoost, which are examples of bagging and boosting ensemble learning. In the bagging method, multiple decision trees are combined and averaged. Meanwhile, the boosting method adds an ensemble variable that corrects and improves previous predictions and gives weighted average results as an output.

This section discusses voting classifiers, another form of ensemble learning which combines various models to give a more robust one. It may be a base model such as decision trees and logistic regression or another ensemble method such as random forest and XGBoost.

There are two main types of voting classifiers: soft and hard voting. Soft voting calculates predicted probability from each estimator and generates the class with the highest average predicted probability. Meanwhile, hard voting is calculated from the predicted class and takes the highest vote.

This project explores both soft and hard voting with several combinations of models.

The table below summarises this process.

Table 7 - Ensemble Learning

Model Name	Voting Type	DT	RF	XGB	NN
ECLF	Soft	Yes	Yes	Yes	No
ECLF2	Hard	Yes	Yes	Yes	No
ECLF3	Soft	No	Yes	Yes	No
ECLF4	Hard	No	Yes	Yes	No

4.5 Model Evaluation

There are several evaluation methods to decide which model performs the best prediction. Though usually several metrics are considered collectively, each has its purpose and shows greater importance than others depending on the purpose of prediction. This section elaborates the definition and concludes why this project selects specific metrics to evaluate mode performance.

4.5.1 Confusion Matrix

The confusion matrix is a popular measure for classification problems. It shows and compares predicted against true values of data. Grouping it into a matrix shows the prediction's true false positives and negatives. True positives (TP) indicate a positive predictive value classified accurately as positive. Similarly, true negatives (TN) indicate a negative prediction label for the actual negative label data. False positive (FP) shows actual negative values predicted as positives, and false negative (FN) shows actual positive values predicted as negatives.

4.5.2 Accuracy, Recall, and Precision

Accuracy measures how accurate the true predictions are against total predictions made.

$$Accuracy = \frac{True\ Predictions}{Total\ Predictions}$$

Recall is the number of true predictions against the total of samples that should be in that class.

$$Recall = \frac{True\ Predictions}{True\ Positives + False\ Negatives}$$

Precision is the ratio of true predictions against the total of predictions in that class.

$$Precision = \frac{True\ Predictions}{True\ Positives + False\ Positives}$$

The use of these metrics highly depends on the use case of predictions. For example, while predicting the cancer patients for treatment, having a low recall is much more important than having a high precision. It is preferable to over treat patients (false positives) than excluding the wrongly predicted cancerous patients (false negatives).

4.5.3 Receiver Operating Characteristic (ROC) Curve and Area Under the Curve (AUC)

ROC curve is widely used to evaluate the performance of model classifiers in machine learning (Kulkarni, Chong and Batarseh, 2020). This curve is made by plotting true negative and false positive on the y and x-axis, respectively. The results of the threshold between 0 and 1 are plotted in the chart, thus producing a comparison between true negative and false positive rates. A diagonal line in the middle shows random guessing, and a curve below that line is considered a low-performing model, worse than guessing. On the contrary, the highest top-left corner is considered the best performing model and desirable.

Performance evaluation usually also calculates the area under the curve or AUC, which represents a number between 0 and 1. A more significant value of AUC is desirable, and a curve below the diagonal line will have AUC below 0.5.

Both ROC and AUC are usually considered good performance metrics for imbalanced datasets (Kulkarni, Chong and Batarseh, 2020).

Chapter 5: Result and Findings

5.1 Model Evaluation

Models explained in chapter 4 are trained using 3,986,606 rows of the training dataset. First, the decision tree is performed as a baseline model. Then, the random forest and XGBoost are trained using cross-validation and tuned using grid search. Finally, the neural network and voting classifier are executed. Table 8 shows the best hyperparameters tuning and metrics improvement after GridSearch for each model classifier:

Table 8 - GridSearchCV - Best Parameters

			Accuracy Score	
Classifier	Parameters	Value	Before Tuning	After Tuning
Random Forest	max_depth	100	0.715405	0.755122
	min_samples_leaf	50		
	min_samples_split	50		
XGBoost	min_child_weight	1	0.750700	0.816554
	max_depth	1		
	gamma	0.0		
	subsample	0.875		
	colsample_bytree	0.525		
	learning_rate	0.01		

The models are trained using those parameters to predict the value of the test dataset. The final results are shown in the table below.

Table 9 - Model Performance using Test Data

Model Name	Accuracy	Precision	Recall	F1 Score
Decision Tree	0.684436	0.609241	0.558467	0.582750
Random Forest	**0.751243**	0.754243	0.548202	0.634925
XGBoost	0.742682	**0.816262**	0.448941	0.579280
Neural Network	0.736047	0.650527	**0.715383**	**0.681415**
ECLF	0.712430	0.650564	0.585946	0.616567
ECLF2	0.745407	0.715060	0.589828	0.646434
ECLF3	0.750022	0.729877	0.581811	0.647487
ECLF4	0.750967	0.765824	0.531364	0.627405

As expected, the Decision Tree has the lowest accuracy score, with only 0.684436. This score suggests that a simple tree could not sufficiently capture the pattern in the dataset. It has a precision score of 0.609241 and a recall score of 0.558467.

The second model, or the first advanced model, is Random Forest, which has an accuracy score of 0.751243. This score is produced after tuning some hyperparameters, explained in Table 8, using GridSearchCV and is the highest accuracy score from all the models. Tuned for precision, this model also outperforms the Decision Tree on precision metric with a score of 0.754243. However, it performs slightly lower on recall, 0.548202. These results are expected because precision and recall optimisation are trade-offs.

While not having the highest accuracy score (0.742682), XGBoost precision score is 0.816262, the highest precision score among all models. It also has the lowest recall of 0.448941.

The Neural Network is tuned for recall and has the highest recall score of 0.715383 while maintaining its precision at 0.650527. Thus, it also has the highest f1 score of 0.681415.

All ensemble learning performs similarly, with the highest accuracy held by ECLF4 with a score of 0.750967, the hard voting classifier between RF and XGB. On the other hand, the soft voting classifier between DT, RF, and XGB (ECLF) has the highest recall score of 0.585946 and the lowest 0.650564 precision. Overall, ensemble learning does not have the highest score on all metrics among all models.

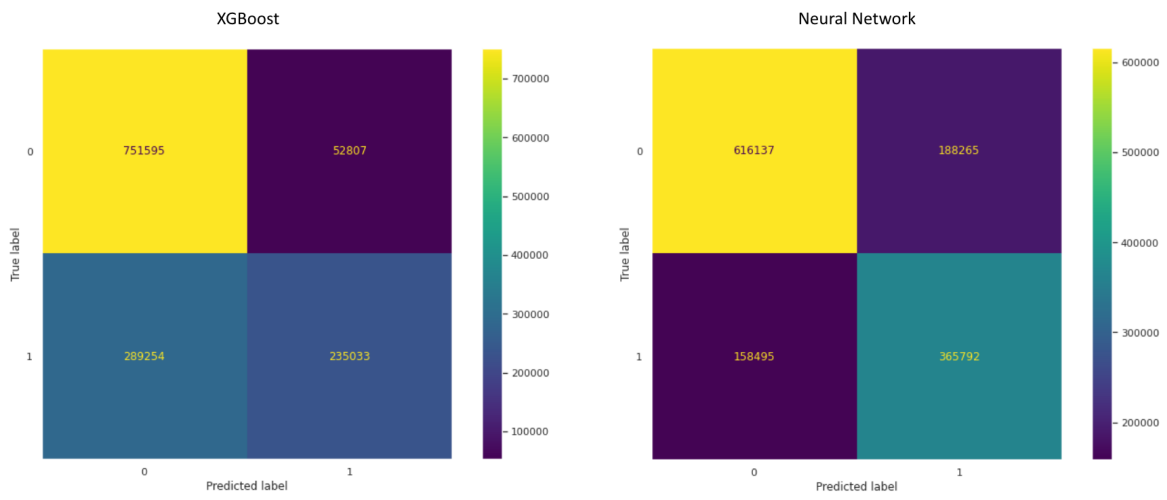


Figure 13 - Confusion matrix for XGBoost and neural network model. Top right corner shows False Positives while the bottom left corner shows False Negatives value.

Next, a closer look is taken at the confusion matrix of XGBoost and Neural Network, the model with the highest score of precision and recall, respectively. Figure 15 shows that XGBoost has correctly predicted customer retention from 235,033 data with only 52,807 false negatives. Meanwhile, the Neural Network model has predicted 365,792 with 188,265 false negatives, leading to a lower precision score.

On the other hand, the Neural Network model has 158,495 false negatives, while XGBoost has significantly higher values of 289,254. This score shows that the Neural Network model has a better recall. The following section will explain why this is important and how to implement these models for the Company.

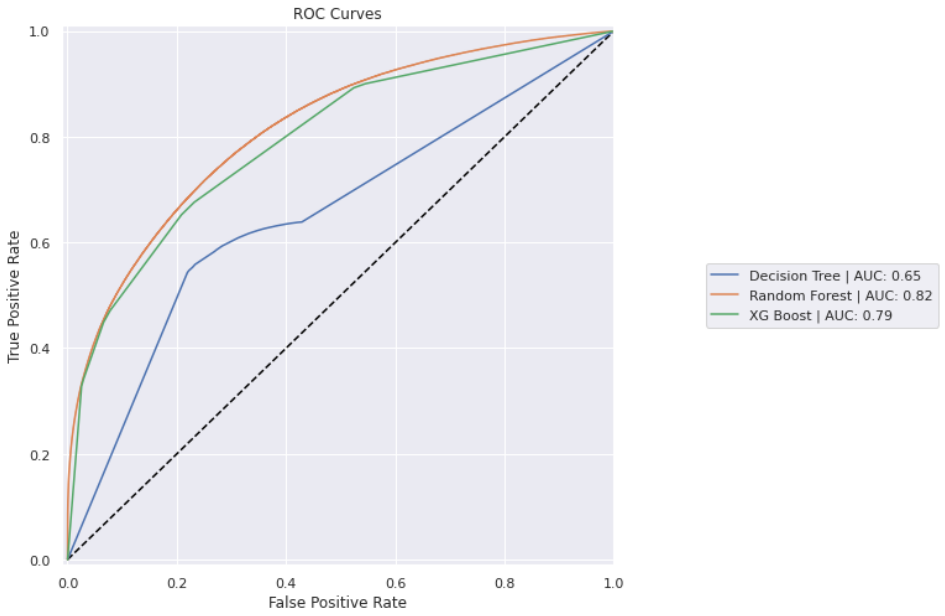


Figure 14 - ROC curves and AUC score for Decision Tree, Random Forest, and XGBoost.

Figure 15 shows that the random forest has the highest AUC with a score of 0.82, meaning it is the most robust model to differentiate the classes. Not long behind is XGBoost, with an AUC score of 0.79. Meanwhile, the decision tree, the baseline model, has a low AUC score of 0.65. This improvement shows that advanced models outperform the baseline model. Similarly, the random forest and XGBoost perform better than the decision tree in ROC curves. Though both overlap at specific points, this research chose XGBoost due to its better performance at precision against the test set.

The models' accuracy for this project ranges between 0.684436 to 0.751243. The only metric that passed 0.8 is XGBoost's precision with a score of 0.816262. Performance of these

metrics are reasonable in predicting customers' churn and retention. For example, Dror et al. (2012) best model while predicting churn of Yahoo! Answers is 0.758 for AUC. Althoff & Leskovec (2015) model performance to predict donor retention in online crowdfunding communities achieves 74.2% for ROC. Lastly, Dave et al. (2013) predicted user retention using the time spent model of web surfing has 71% prediction accuracy.

5.2 Feature Importance

Based on the explanation from the previous section, we use XGBoost as a basis for our feature importance analysis. This section focuses on explaining the top five most important features based on the SHAP values. An important thing to note is the SHAP does not imply causality; instead, it shows a feature's impact magnitude to the probability of predicted output.

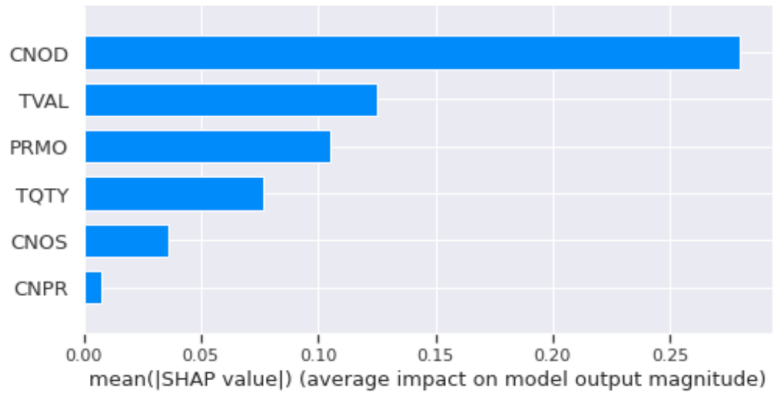


Figure 15 - SHAP summary of features with the most significant impact. CNOD (count of order) has the highest value, followed by TVAL (total value) and PRMO (prior month). Complete table of variables' description can be seen in section 4.1 Feature Engineering and Selection.

Based on SHAP in Figure 15, CNOD has the most significant impact, with a mean score of more than 2.5. The second and third are value_taxed (OVAL) and CNOS. We examined these figures further using a summary plot from SHAP libraries.

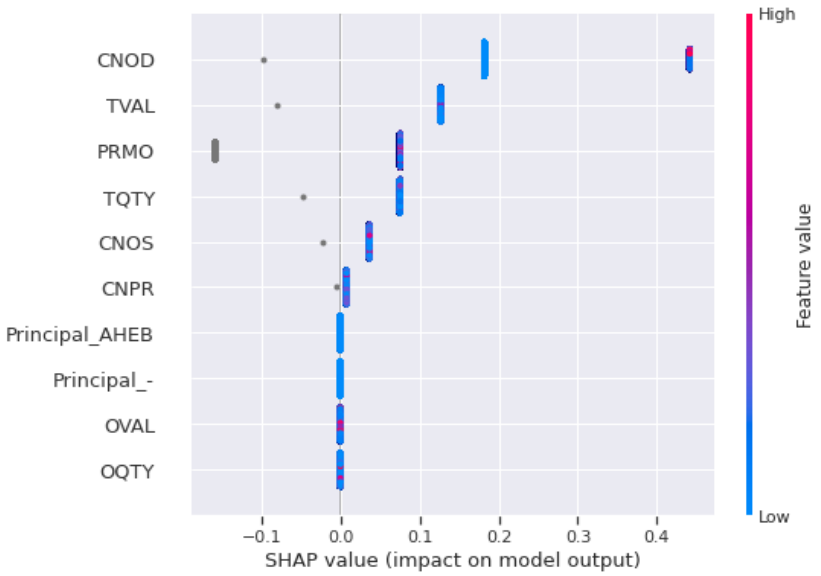


Figure 16 - SHAP summary of features with the most significant impact. CNOD (count of order) has the highest value, followed by TVAL (total value) and PRMO (prior month). Complete table of variables' description can be seen in section 4.1 Feature Engineering and Selection.

The figure above may explain more about the features. Each dot represents one data sample, x-axis showing SHAP Values. The features are sorted by their importance, with the highest as the most important by the sum of SHAP value magnitudes. The colour represents the feature value, where red is the highest.

The chart shows that the count of order (CNOD) positively impacts the probability of a predicted value of 1. A higher value of orders also improves this probability. On the other hand, a total of the historical transaction value (OVAL) positively impacts the probability of a predicted value of 1, but a higher feature value does not necessarily have a higher SHAP value.

The number of official stores (CNOS) shows an exciting distribution on the chart. Low feature values, meaning the customers do not buy from various official stores in their prior purchases, lower the probability of them making another purchase within 30 days. Meanwhile, the high value of CNOS increases the probability of them returning and making another transaction. This impact supports SIRCLO leaders' intuition about the importance of having customers buy from various official stores, which may increase their trust in the platform.

Second purchase also plays an essential role in customers' churn. SHAP shows that customers with no historical purchase lower the probability of returning. Meanwhile, customers on their second or more purchases are more likely to return.

Lastly, the chart shows that high order quantity decreases customers' probability of making another purchase. Although this seems counterintuitive, this shows that some customers are hoarder who occasionally buys products in high quantities. These customers may have different factors and behaviour than retail customers; for example, they might only buy when product discounts reach a certain threshold. Further studies are needed to understand this specific type of customer.

5.3 Business Application

Finding a good business fit is often challenging for machine learning implementation. Though many businesses want to exploit the benefit, most are having difficulty in finding specific use cases. Thus, this section discusses how to apply the model and predictive results from previous sections in SIRCLO. Moreover, it elaborates on two general approaches to maximise the use of different models with different scoring metrics.

Approach 1

Assume the company wants to create a specific marketing campaign targeting returning customers. This campaign may be for a new product launching or new principals with a similar existing category. For this project, having a model with good precision metrics is critical.

In this scenario, the campaign acts as a start of business workflows and expects a user to make a specific action while involving high direct variable costs for each customer; for example, by giving them a free product sample. By having high precision, the business can maximise its investment by focusing on its target market: its loyal and returning customers.

A lower score on recall is less critical because the cost of having a widespread false positive promotion is higher than missing specific false negative customers. Therefore, the business should use the predicted model with the highest recall: XGBoost with a score of 0.820469.

Approach 2

The second scenario emphasises the opportunity costs of excluding the returning customers (false negatives). Assume that the company wants to predict as many returning customers as possible while able to maintain low variable retention costs. The business is less worried about targeting the churning customers, as the marketing campaign cost is not substantial.

This approach benefits more from having a high recall score. Therefore the company should implement Neural Network prediction, which has the highest recall score of 0.709944.

This method might be suitable for sending marketing emails or having an under-capacity customer service team, which can call more users without incurring high additional costs.

Chapter 6: Conclusion and Remarks

6.1 Conclusion

After implementing various models and algorithms, this project successfully predicts customer retention and churn. Furthermore, feature importance analysis shows that the count of historical orders and transaction values are essential and may indicate a higher probability of customers making another transaction.

However, having a good prediction model is only one step of many. Therefore, two business applications are presented in Chapter 5 to provide more prominent examples of how the company might implement this. The XGBoost model has the best performance for precision, while the neural network is optimised for recall.

These results indicate that adopting internal transaction data for churn prediction shows promising results, and businesses should explore this approach further.

Limitations and future improvement is elaborated in the next section.

6.2 Limitations and Future Improvements

Based on the scope and objectives, this study predicts whether customers will make another purchase within 30 days while omitting any industry-specific behaviour. Thus, this study may be less relevant in specific industries with niche customer behaviour. Instead, this project provides a high-level overview that transaction data is adequate to make a robust customer retention prediction model, regardless of the industry and segment.

The company should expand its model training using multi-year transaction data. More complex and diverse feature engineering may also be explored to capture more data patterns. Although these processes are resource intensives, improvement in prediction accuracy is expected.

While promotion and marketing are massive drivers in the e-commerce industry, this project does not utilise discount and marketing value in its feature engineering due to unavailable data. The company should keep track of their promotion and leverage those to improve the model.

Moreover, as customers' behaviour may change depending on competitors' prices and discounts, competitor price analysis may also benefit the prediction. Having detailed competitors' product pricing and promotion may also enable the company to expand this model to measure promotion effectiveness and therefore allocate their marketing budget efficiently.

In short, this project shows the company that its transaction data is an invaluable strategic asset. It opens up possibilities and opportunities to shift from descriptive to more advanced predictive analysis. Furthermore, the rapid growth of machine learning lowers the barrier to business implementation and gives management essential insights to enhance their data-driven decision-making process.

References

Ahn, J.-H. et al. (2006) Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry. Telecommunications policy. [Online] 30 (10), 552–568.

Althoff, T. and Leskovec, J. (2015). Donor Retention in Online Crowdfunding Communities: A Case Study of DonorsChoose.org. Proceedings of the ... International World-Wide Web Conference. International WWW Conference, [online] 2015, pp.34–44. doi:10.1145/2736277.2741120.

Alpha JWC, Kearney. (2020). Read @Kearney: Unlocking the next wave of digital growth: beyond metropolitan Indonesia. [online] Available at: https://www.kearney.com/digital/article/-/insights/unlocking-the-next-wave-of-digital-growth-beyond-metropolitan-indonesia [Accessed 14 Jun. 2022].

Angeloni, S. and Rossi, C. (2021). An analytical model for comparing the profitability of competing online marketing channels: Search engine marketing versus e-commerce marketplace. Journal of Marketing Theory and Practice, pp.1–16. doi:10.1080/10696679.2021.1879656.

Ariansyah, K., Sirait, E.R.E., Nugroho, B.A. and Suryanegara, M. (2021). Drivers of and barriers to e-commerce adoption in Indonesia: Individuals' perspectives and the implications. Telecommunications Policy, p.102219. doi:10.1016/j.telpol.2021.102219.

Avgerou, C. (2003). Information Systems and Global Diversity. [online] Oxford University Press. Oxford, New York: Oxford University Press. Available at: https://global.oup.com/academic/product/information-systems-and-global-diversity-9780199263424?cc=gb&lang=en& [Accessed 25 Jul. 2022].

Behrens, J. T. (1997). Principles and Procedures of Exploratory Data Analysis. Psychological Methods, 2 (2), 131-160.

Berger, P. and Kompan, M. (2019). User Modeling for Churn Prediction in E-Commerce. IEEE Intelligent Systems, [online] 34(2), pp.44–52. doi:10.1109/MIS.2019.2895788.

Brynjolfsson, E., Hitt, L.M. and Kim, H.H. (2011). Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance? SSRN Electronic Journal. doi:10.2139/ssrn.1819486.

Chiang, W.-Y. (2019) Establishing high value markets for data-driven customer relationship management systems: An empirical case study. Kybernetes. [Online] 48 (3), 650–662.

Coussement, K. & Van den Poel, D. (2008) Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. Expert systems with applications. [Online] 34 (1), 313–327.

Datta, P. (2010). A preliminary study of ecommerce adoption in developing countries. Information Systems Journal, 21(1), pp.3–32. doi:10.1111/j.1365-2575.2009.00344.x.

Dave, K., Vaingankar, V., Kolar, S., Varma, V., Stumbleupon, S. and Francisco (2013). Timespent Based Models for Predicting User Retention. [online] Available at: http://www2013.w3c.br/proceedings/p331.pdf.

Dror, G., Pelleg, D., Rokhlenko, O. and Szpektor, I. (2012). Churn prediction in new users of Yahoo! answers. Proceedings of the 21st international conference companion on World Wide Web - WWW '12 Companion. [online] doi:10.1145/2187980.2188207.

Faraway, J.J. (2014). Does data splitting improve prediction? Statistics and Computing, 26(1-2), pp.49–60. doi:10.1007/s11222-014-9522-9.

Geron, Aurelien (2017). Hands-on machine learning with Scikit-Learn and TensorFlow : concepts, tools, and techniques to build intelligent systems. Sebastopol, Ca: O'reilly Media.

Hadden, J., Tiwari, A., Roy, R. and Ruta, D. (2007). Computer assisted customer churn management: State-of-the-art and future trends. Computers & Operations Research, 34(10), pp.2902–2917. doi:10.1016/j.cor.2005.11.007.

Iprice (2022). Top 50 E-Commerce Sites & Apps in Indonesia 2022. [online] Available at: https://iprice.co.id/insights/mapofecommerce/en/.

Jurriens, E. & Tapsell, R. (2017) Digital Indonesia: Connectivity and Divergence. 1st ed. SG: ISEAS-Yusof Ishak Institute.

Kingma, D.P. and Ba, J. (2014). Adam: A Method for Stochastic Optimization. [online] arXiv.org. Available at: https://arxiv.org/abs/1412.6980.

Kozak, J., Kania, K., Juszczuk, P. and Mitręga, M. (2021). Swarm intelligence goal-oriented approach to data-driven innovation in customer churn management. International Journal of Information Management, p.102357. doi:10.1016/j.ijinfomgt.2021.102357.

Kraus, M., Feuerriegel, S. and Oztekin, A. (2020). Deep learning in business analytics and operations research: Models, applications and managerial implications. European Journal of Operational Research, 281(3), pp.628–641. doi:10.1016/j.ejor.2019.09.018.

Kulkarni, A., Chong, D. and Batarseh, F.A. (2020). 5 - Foundations of data imbalance and solutions for a data democracy. [online] ScienceDirect. Available at: https://www.sciencedirect.com/science/article/pii/B9780128183663000058.

Kumar, V. & Petersen, J. (2012) Statistical Methods in Customer Relationship Management / Kumar, V.. 1st edition. Wiley.

Lariviere, B. and Vandenpoel, D. (2005). Predicting customer retention and profitability by using random forests and regression forests techniques. Expert Systems with Applications, 29(2), pp.472–484. doi:10.1016/j.eswa.2005.04.043.

Lawrence, Elaine & Newton, Stephen & Corbitt, Brian & Lawrence, John & Dann, Stephen & Thanasankit, Theerasak. (2003). Internet commerce : digital models for business.

Ma, Y. (2022). China: e-commerce share of retail sales 2019. [online] Statista. Available at: https://www.statista.com/statistics/1129915/china-ecommerce-share-of-retail-sales/.

Manivannan, R., Saminathan, R. and Saravanan, S. (2019). An improved analytical approach for customer churn prediction using Grey Wolf Optimization approach based on stochastic customer profiling over a retail shopping analysis: CUPGO. Evolutionary Intelligence. doi:10.1007/s12065-019-00282-x.

Popovic, D. & Basic, B. D. (2009) Churn prediction model in retail banking using fuzzy C-means algorithm. Informatica (Ljubljana). 33 (2), 243–248.

Pangestu, Mari and Dewi, Grace. (2017) "13. Indonesia and the digital economy: creative destruction, opportunities and challenges". Digital Indonesia: Connectivity and Divergence, edited by Edwin Jurriens, Singapore: ISEAS Publishing, 2017, pp. 227-255.

Ranjan, J. and Foropon, C. (2021). Big Data Analytics in Building the Competitive Intelligence of Organisations. International Journal of Information Management, 56, p.102231. doi:10.1016/j.ijinfomgt.2020.102231.

Reinartz, W., Krafft, M. and Hoyer, W.D. (2004). The Customer Relationship Management Process: Its Measurement and Impact on Performance. Journal of Marketing Research, 41(3), pp.293–305. doi:10.1509/jmkr.41.3.293.35991.

Reinartz, W. J. & Kumar, V. (2003) The Impact of Customer Relationship Characteristics on Profitable Lifetime Duration. Journal of marketing. [Online] 67 (1), 77–99.

Scikit-learn.org. (2019). 1.4. Support Vector Machines — scikit-learn 0.22 documentation. [online] Available at: https://scikit-learn.org/stable/modules/svm.html#svm-classification.

Scikit-learn.org. (2012). 1.11. Ensemble methods — scikit-learn 0.22 documentation. [online] Available at: https://scikit-learn.org/stable/modules/ensemble.html#forests-of-randomized-trees.

Scikit-learn. (n.d.). 1.11. Ensemble methods. [online] Available at: https://scikit-learn.org/stable/modules/ensemble.html# [Accessed 13 Jun. 2022].

Siber, Richard (1997) Combating the churn phenomenon. Telecommunications international. 31 (10), 77–.

SIRCLO (2021). E-Commerce Insights by SIRCLO: Navigating Market Opportunities. [online] Available at: https://insights.sirclo.com/ [Accessed 14 Jun. 2022].

Think with Google. (2019). e-Conomy SEA 2019: Swipe up and to the right: Southeast Asia's $100 billion internet economy. [Online].

Think with Google. (2020). e-Conomy SEA 2020: At Full Velocity: Resilient and Racing Ahead. [Online].

Think with Google. (2021). e-Conomy SEA 2021: Roaring the 20s: The SEA Digital Decade. [Online].

Tin Kam Ho (1995). Random decision forests. [online] IEEE Xplore. doi:10.1109/ICDAR.1995.598994.

Verbeke, Wouter, David Martens, Christophe Mues, and Bart Baesens. 2011. “Building Comprehensible Customer Churn Prediction Models with Advanced Rule Induction Techniques.” *Expert Systems with Applications* 38 (3): 2354–64.

www.bps.go.id. (2021). Badan Pusat Statistik. [online] Available at: https://www.bps.go.id/pressrelease/2021/01/21/1854/hasil-sensus-penduduk-2020.html.

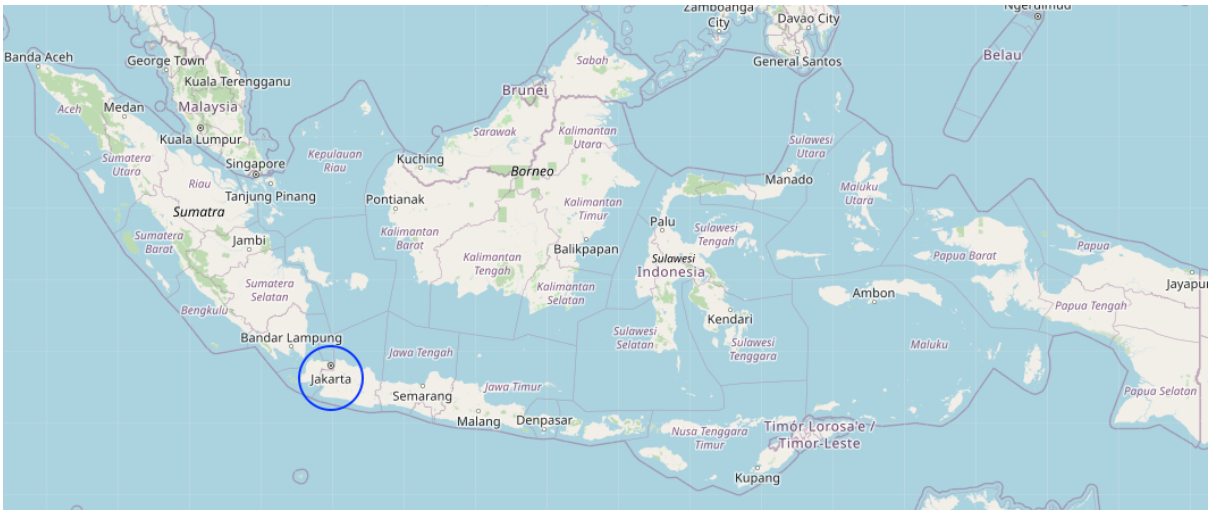
www.bps.go.id. (2021). Badan Pusat Statistik. [online] Available at: https://www.bps.go.id/indicator/52/286/1/-seri-2021-produk-domestik-regional-bruto-.html.

Yan, Y., Guo, W., Zhao, M., Hu, J. and Yan, W.P. (2017). Optimizing Gross Merchandise Volume via DNN-MAB Dynamic Ranking Paradigm. arXiv:1708.03993 [cs]. [online] Available at: https://arxiv.org/abs/1708.03993 [Accessed 25 Jul. 2022].

Appendix

Appendix A - Map of Indonesia

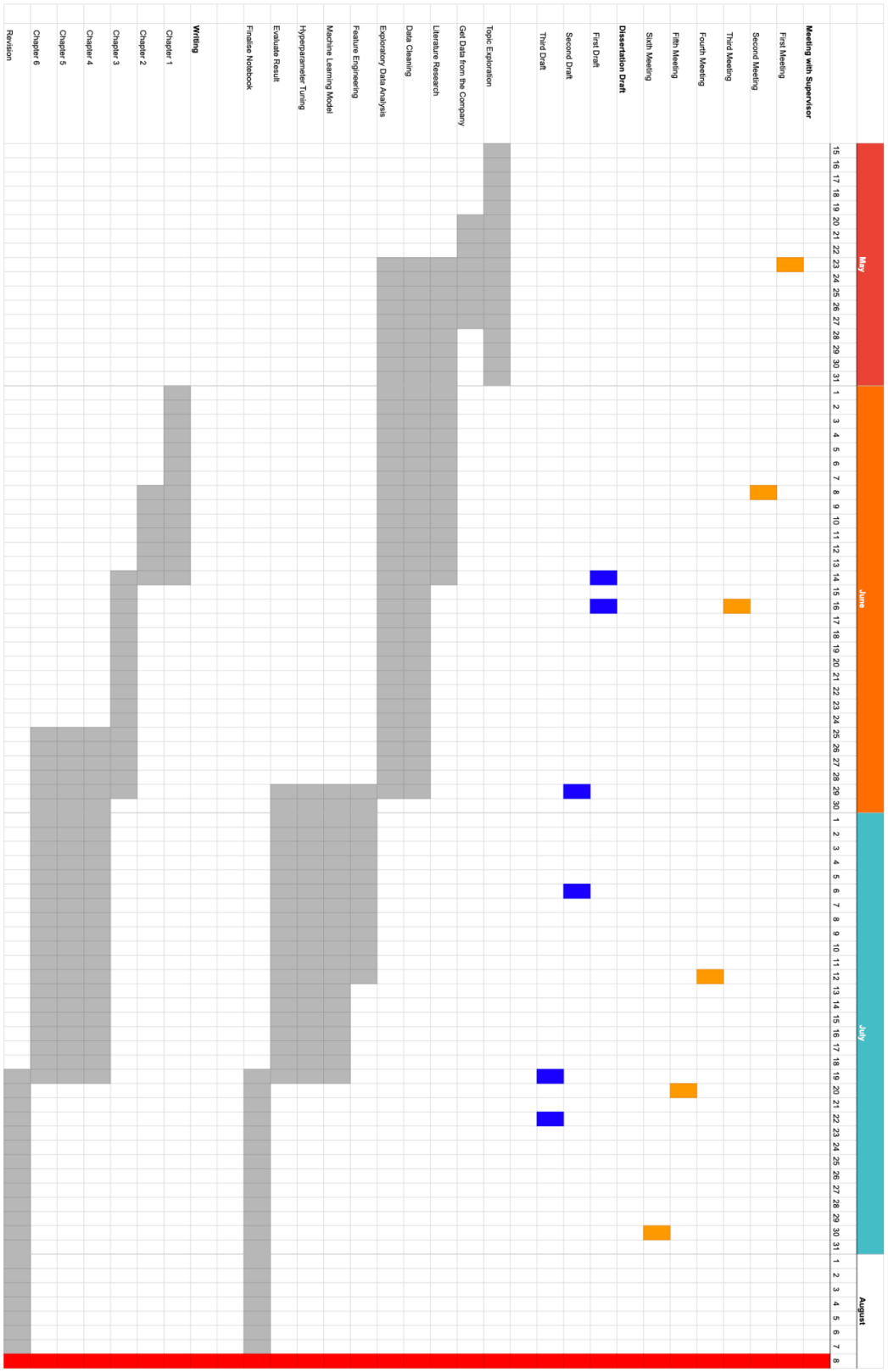
Jakarta, the capital and the biggest metropolitan city, is highlighted in blue.



Appendix B - Features Exploration and Selection

No	Omitted Variables	DT	RF	XGB
Exploration 1	Order Status, SKAM, Warehouse, City, Postcode, Shipping, Marketplace, Official Store, Shipping Category, Province Category, Province	0.674737	0.701542	0.745958
Exploration 2	Order Status, SKAM, Warehouse, City, Postcode, Shipping, Shipping Category, Province Category, Province	0.683115	0.713421	0.749333
Exploration 3	Order Status, SKAM, Warehouse, City, Postcode, Shipping, Gender, Shipping Category, Province Category, Province	0.681069	0.709398	0.748504
Exploration 4	Order Status, SKAM, Warehouse, City, Postcode, Shipping, Gender, Shipping Category, Province Category	0.678019	0.714185	0.750037
Exploration 5	Order Status, SKAM, Warehouse, City, Postcode, Shipping, Gender, Shipping Category, Province	0.679973	0.713060	0.749995
Exploration 6 (chosen)	**Order Status, SKAM, Warehouse, City, Postcode, Shipping, Marketplace, Official Store, Gender, Shipping Category, Province Category, Province**	**0.695487**	**0.715405**	**0.750700**

Appendix C - Project Timeline



Appendix D - Repository

Repository for this project can be accessed through github:

https://github.com/louisbernardus/MSIN0114