E-Commerce Customer Churn Analysis and Prediction

Yeo Jie Hui[*], Associate Prof. Dr. Azah Anir Binti Norman[+]

[#]*Faculty of Computer Science and Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia*
u2005330@siswa.um.edu.my
[+]*Faculty of Computer Science and Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia*
azahnorman@um.edu.my

***Abstract*—Customer churn is a big concern for e-commerce industry due to its associated costs. Client retention and preventing customer churn have become crucial issues for e-commerce business operations and growth. This short paper briefly explains how to work on customer churn analysis and prediction for e-commerce and the results of the work. Different machine learning approach is compared to choose the best model in accurately predicting customers who will change and turn to another e-commerce platform for the same or similar service or products. Dataset used in this project is retrieved from a well-known online e-commerce company data in Kaggle. This paper aims to build an e-commerce customer churn prediction model and show the importance of analysis and prediction of e-commerce customer churn in helping decision making for choosing the proper decisions on behalf of the e-commerce company. The Cross Industry Standard Process for Data Mining (CRISP-DM) is a process model that serves as the base for a data science process. Thus, CRISP-DM is applied as the methodology throughout the paper and the paper is concluded with future directions for the study.**

Keywords—e-commerce; customer churn problem; prediction and analysis model; data science; CRISP-DM; machine learning

I. INTRODUCTION

Firstly, we will need to know what means customer churn, basically it is the customers who are going to change their usage of a certain service to another competing service company.[1] Normally it is due to some reasons for churning, and this will be one of the focus that needed to be further analyse in this project.

Customers are one of an organization's most valuable assets and are crucial to increasing the company's performance and competitiveness in the market.[2] Customers can choose from a wide range of goods or service providers despite strong market competition.[3] It has become crucial for businesses to figure out how to utilise current consumer resources and prevent the loss of existing customers in order to keep market advantages.[4]

According to the statistic by Statista, it is estimated that e-commerce sales will reach $6.3 trillion by 2024. In addition, a study by Wunderman Thompson claims that 65% of consumers believe they will increase their use of online shopping channels in the future.[5] Thus, we can see that the market of e-commerce is growing bigger and bigger, so definitely it will be vital to predict customer churn and provide suggestions or promote based on customers' preferences to prevent they leave.

In this project, the key reasons for customer churn in e-commerce will be analysed and machine learning approach is used to build an e-commerce customer churn prediction model based on those reasons.

II. LITERATURE REVIEW / PREVIOUS RESEARCH

The term "customer churn" describes the phenomenon wherein a company's early customers stop using its products or services and start using those of its rivals.[6] In the case of a non-contractual relationship, one type of customer churn is seen in e-commerce. Even if this type of business-customer relationship were to end in a non-contractual partnership, it would be challenging for the company to anticipate it.[7] The goal of e-commerce customer churn prediction is to combine customer data collected over time and create models for predicting customer churn by looking at consumer purchase patterns.[8] Then, offer e-commerce customer churn retention strategies to lessen customer churn, find high-value non-churn e-commerce customers, and perform admirably in terms of retaining customers.

Based on the research of Shao,[7] indicates that although new customers desire to increase e-commerce profits, the current customers do not require high costs. Both new and established customers have different purchasing habits, yet it is still important to pinpoint the factors that contributed to the customer's loss. In support, Lu et al.,[8] said that in the e-commerce industry, it is crucial to analyse customer loss, predict which customers may be lost, and then take appropriate action to keep these customers and prevent their loss. Currently, the majority of e-commerce businesses have thoroughly analysed customer basic characteristic data and transaction behaviour data. They then utilise a variety of techniques and technologies to create and research customer churn prediction models.[9]

According to Agrawal et al.,[10] customers become a crucial component in how profitable businesses are as well as a crucial resource for gaining a competitive edge. Customers that shop online often leave after a short while. Customers must be developed into stable, ongoing connections if businesses wish to form long-term agreements with them. The consumer base for e-commerce is vast, intricate, and valued differently. A popular topic in the world of e-commerce has emerged as a result of the Saghir et al.,[11] study, which highlighted the question of how to precisely identify high-value consumers, predict churn, and keep them in advance. In response, Wu et

al.,[12] conducted research showing that measuring customer value can help businesses identify valuable customers among a wide range of consumers and implement various customer management strategies in line with various customer values, maximising the impact of the company's limited resources.

Customer churn has always been a source of interest. According to a paper titled Prediction of Customer Churn in eRetailing that was released in 2020,[13] ensemble techniques work best for identifying churners. Sonali Agarwal, Divya Tomar, and Pretam Jayswal demonstrated that gradient boosted trees produced the highest results in customer churn analysis using bagging and boosting ensemble learning approaches to predict customer churn with classifiers like Decision Trees, Random forests, and trees.[14] They also employed optimizations to improve their findings. Irfan Ullah studied the use of random forests in churn prediction in 2019. In this study, K-Means Clustering and RF Classifiers were employed to profile the customers.[15] Thus, based on the research of previous works, the decision is made where Logistic Regression, Linear Discriminant Analysis (LDA), Random Forest, XGBoost, four different machine learning algorithms will be used in this project.

III. Problem Statement

Based on Zendesk, it is found that an average of 66% of consumers had terminated their relationship with a company due to poor customer service. Although good customer service and merchandise can help e-commerce retain customers but getting to know a client well is the best approach for a business to stop customer attrition. However, e-commerce businesses frequently put a great deal of work and money into recruiting new clients rather than concentrating on keeping their current ones.[16]

It turns out that in the world of e-commerce, acquiring new clients is significantly more expensive for a business than keeping its existing clientele,[11] especially when e-commerce acting as a popular field in gaining revenue with various competitors, for example, Shopee, Lazada and more.

A research paper also explored that keeping consumers is more profitable than finding new ones, according to multiple studies, such as Customer Loyalty: Toward an Integrated Conceptual Framework,[17] Customer Loyalty in eCommerce,[18] and Customer base analysis: Partial defection of behaviourally loyal clients in a non-contractual FMCG retail.[19] It mentioned high customer retention rates are always a common challenge for businesses.[20]

Therefore, to solve the issue of lack attention from e-commerce industry in customer churn, I decide on doing this project to show the importance of analysis and prediction of e-commerce customer churn in helping decision making for choosing the proper decisions on behalf of the e-commerce company.

IV. Methodology

For this project, I would employ the CRISP-DM methodology as it is one of the most widely used methodology, which is composed of cyclical stages starting with a knowledge of the business (business understanding) where we identify the issue and look at its commercial opportunity. The next step is data understanding, which is the data acquisition in this project, it is the step where we collect data from various sources. Following that, I will do data preparation which entails going through several rounds of cleaning the data by removing any missing values, outliers, and extraneous columns. Then, I will do Exploratory Data Analysis (EDA) to visualize the content of our dataset more clearly. Last but not least, I will conclude with modelling, evaluation, and deployment, during which we will put our models into practise and evaluate their accuracy and performance.[21] Fig.1 shows the graphical view of CRISP-DM methodology stages.

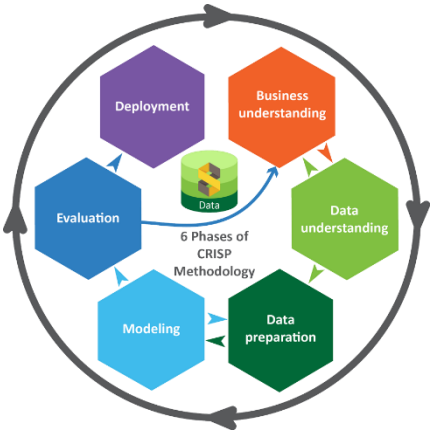

Fig.1 Phases of CRISP-DM Methodology

A. *Business Understanding*

This is the first step in the methodology. As the e-commerce nowadays will need to know whether a customer will churn and why they churn, therefore I decided to build a model by testing different machine learning algorithms or approach to do an accurate e-commerce customer churn analysis and prediction project.

B. *Data Acquisition*

This step can be also known as data understanding and data collection step. For this step is to gather the dataset needed for this project. In this project, the dataset being used is obtained from Kaggle, this dataset is the property of a well-known online e-commerce company, where it wants to determine which consumers are likely to leave so that the company can contact them with special offers or promotions. It is historical data with customer information and experience, and its result is a flag for customer churn (churn=1, no churn=0). The dataset displays interactions and preferences among more than 5000 customers using the platform. This data's usefulness comes from the fact that it includes some precise, granular attributes that will aid in customer segmentation, like preferred login device, satisfaction shore, and other characteristics. These characteristics will support my project into the factors contributing to customer churn in each customer segment. I had retrieved the dataset from Kaggle.[22] This dataset consists of 5630 rows and 20 columns and below Table I is the data dictionary which explain the meaning of each column attributes:

TABLE I
DATA DICTIONARY

Variable	Description
CustomerID	Unique customer ID
Churn	Churn Flag (0-Customer Not Churn, 1-Customer Churn)
Tenure	Tenure of customer in the e-commerce platform
PreferredLoginDevice	Preferred login device of customer
CityTier	City tier
WarehouseToHome	Distance in between warehouse to home of customer
PreferredPaymentMode	Preferred payment method of customer
Gender	Gender of customer
HourSpendOnApp	Number of hours spend on mobile application or website
NumberOfDeviceRegistered	Total number of devices is registered on particular customer
PreferedOrderCat	Preferred order category of customer in last month
SatisfactionScore	Satisfactory score of customers on service

In this step, since we need to understand data, we can actually do exploratory data analysis (EDA) to access and visualize our data. As for me, I will prefer to do EDA before and after data preparation as the data visualize after data preparation is already cleaned and more organized. I will import libraries in Google Colab to do visualization such as bar chart to show churn column so that it is clear to view the number of customers churn versus number of customers do not churn.

C. *Data Preparation*

Data preparation, which also referred to data preprocessing, is a very important step to ensure the quality of data by transforming and cleaning raw data for extracting useful insights and make accurate predictions when running it through machine learning algorithms. Data preprocessing is known to be one of the most meaningful issues within the famous Knowledge Discovery from Data process. Since inconsistent and redundant data will almost certainly exist, commencing a data analysis procedure with such data is not appropriate. Data preprocessing enables the processing of data that would otherwise be impractical by enabling the data to be adjusted to the needs posed by each data mining technique.[23]

Basically, the key process of data preparation is data selection to choose required data for modelling phase; data preprocessing or data cleaning to fix inconsistencies in data, handle missing and duplicate values in the dataset and also data tidying so that the data is organized in certain format to ease the process of modelling; data transformation to reconstruct the data and derive necessary attributes, data integration to merge different data sources (in this project will not be apply as the dataset is only from one source) and data formatting such as naming conventions.

For data cleaning of this project is impute missing values with mean values. By using the isnull.sum() function, it is noticed that the columns “Tenure”, “WarehouseToHome”, “HourSpendOnApp”, “OrderAmountHikeFromlastYear”, “CouponUsed”, “OrderCount”, “ DaySinceLastOrder” consist of missing values. Total number of rows of missing values: 1856 and there are no duplicate values in the data. To make sure the model is accurate, these missing values should be addressed. Information loss may occur if all rows with missing values are removed. Therefore, the missing values is substituted by the median values. Although replacing missing values with the median may reduce the variance of the data, it is simple to execute and can keep the judgement of the model from being influenced by skewness in the data.[24]

While for data transformation in this project is done where in PreferedOrderCat column, there were two variables indicating the same meaning “Mobile” and “Mobile Phone” which were grouped in one level labelled as “Mobile”. Then, In PreferredPaymentMode column “Cash on Delivery” and “COD” levels that have been grouped in one level labelled as “Cash on Delivery” in addition to “Credit Card” and “CC” levels have been grouped in in group called “Credit Card”.

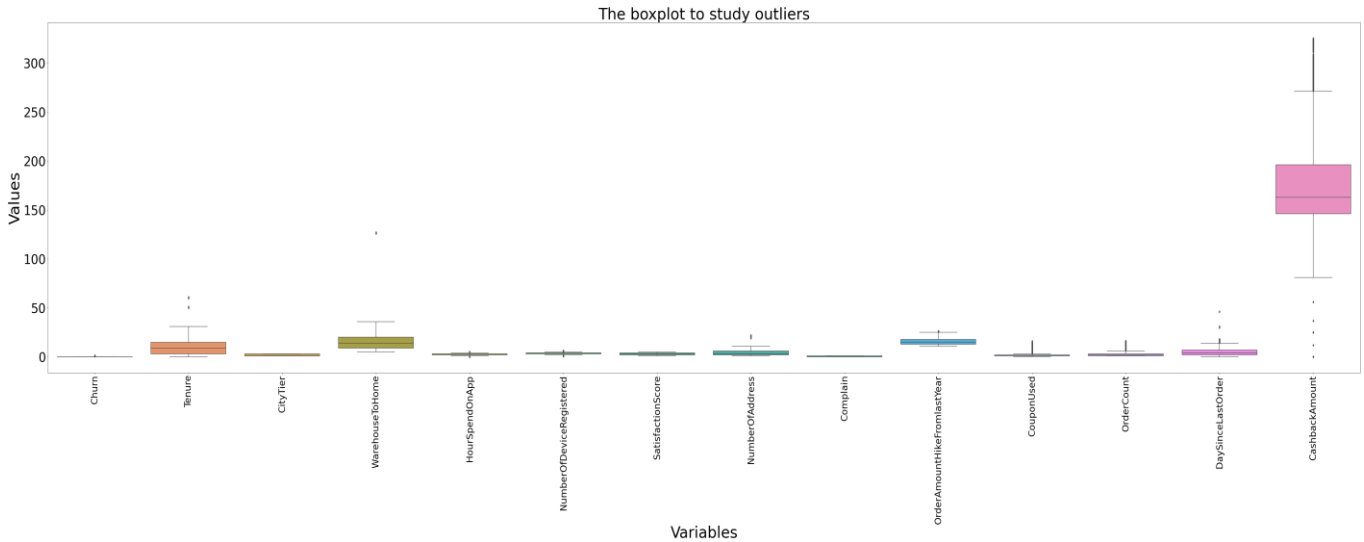

Fig. 2 Boxplot to Study Outliers of Each Variable

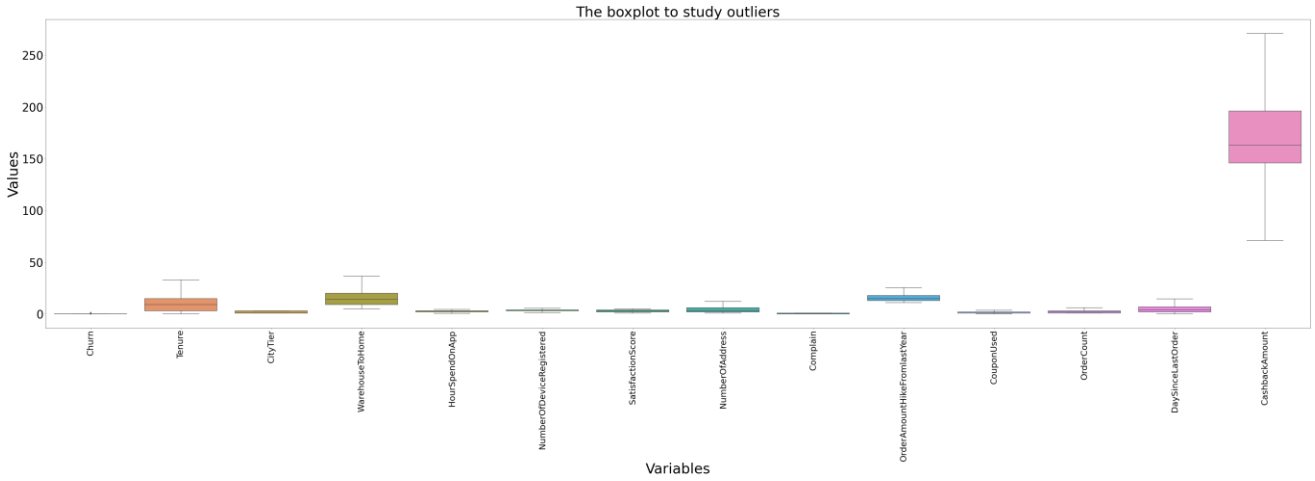

Fig. 3 Boxplot Free from Outliers After Outlier Treatment

Lastly, outlier treatment is done for data preparation as outliers are observations that deviate from the statistical distribution of the vast majority of the data, which could result in inaccurate results from statistical analysis.[25] There are total 2287 outliers for our data. To treat it, we define the lower range and upper range which is going to be at a distance of 1.5 times the Interquartile range from the respective whiskers. After treatment, the outliers are now replaced with their corresponding upper range or lower range values and the data is free from outliers. Fig. 2 shows the outlier of all variables in boxplot while Fig. 3 shows the result after outlier treatment.

D. Exploratory Data Analysis (EDA)

Actually, exploratory data analysis can be done before and after the data preparation process. Since there is no strict order of execution in this CRISP-DM methodology, I will also do EDA after the data preparation so that the data is cleaned and easier to visualize. For EDA, it is to summarize the data main characteristics and also plot the data to see the patterns or anomalies. So, in this project, Python is used for exploratory data analysis in the Google Colab for small data. In an effort to find trends in the data, Python libraries - "matplotlib" and "seaborn" are used to create engaging representations of the dataset. For example, we plot bar chart to study the distribution of churn and not churn rate in our dataset, some histograms to distinguish between difference of churned and retained customers when react with certain columns and also use heatmap to show relationship between churn or not churn with each variable.

E. Modelling

This is the model building step which is use different machine learning approach based on the nature of problem to be tackled. There are a variety of algorithms to choose from, each with its own set of advantages and disadvantages. Exploration of the various algorithms is essential to determine the best algorithms to be used to build the model for web application. As to compare the model accuracy, four machine learning models is chosen, which is Logistic Regression, Linear Discriminant Analysis, Random Forest and XGBoost. Further description of the selected algorithms and reasons for using them are as follow:

1) Logistic Regression: Logistic regression is a process where probability of discrete outcome is modelled which is usually used for binary outcomes. Logistic regression is widely used on classification problems especially when the aim of the study is to determine if a sample appropriately fitting into a class. and considered as one of the main analytical algorithms.[26] The reason of choosing it is because it is a classification technique used in machine learning, which suites to our dataset that only having class 0 (not churn) and class 1(churn).

2) Linear Discriminant Analysis: Linear Discriminant Analysis (LDA) is a dimensionality reduction technique. In order to avoid the dimensionality curse, save resources, and cut dimensional costs, it is used to project features from higher dimensions space onto a lower dimensional space.[27] The reason of choosing this algorithm is because it is a supervised classification technique machine learning model, same as Logistic Regression, suitable for this classification problem project.

3) Random Forest: Random forest is machine learning algorithm which combines the output of various decision aiming for a single output, it overcomes some overfitting and bias issues associated with decision trees and gives accurate predictions especially when individual trees are uncorrelated with each other.[28] The reason of choosing it is due to its characteristic of generally considered very accurate predictive model and can handle categorical data just like some categorical type data which exist in this project dataset.

4) XGBoost: XGBoost is a distributed gradient boosting library that has been developed to be very effective, adaptable, and portable. It uses the Gradient Boosting framework to implement machine learning algorithms. A parallel tree boosting method called XGBoost (also known as GBDT or GBM) is available to address a variety of data science issues

quickly and accurately. The same algorithm can answer problems with more than a trillion examples and runs on key distributed environments (Hadoop, SGE, MPI).[29] This algorithm is chosen as one because it is very flexible open-source library capable of binary classification.

F. Evaluation and Interpretation

In this stage, evaluation is required to confirm the effectiveness of the best model selected in producing results, insights, and findings in order to make sure it is correct for e-commerce customer churn analysis and prediction. ROC curve and F1-score are a few examples of evaluation measures. The construction of the confusion matrix served as the major way of evaluation, and the accuracy, recall rate, and accuracy were computed on the basis of this matrix. The prediction impact of the model was evaluated using these three indicators. The detailed explanation of the evaluation matrices are as follows:

1) Accuracy: Accuracy is the proportion between the positive and negative sample size and the total sample size that is accurate in the overall prediction. It mainly measures the accuracy of the overall predictions of the model.

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

2) Recall: Recall is the proportion of correctly predicted positive sample size to the actual number of positive samples, and it reflects the coverage of the model.

$$Recall = \frac{TP}{TP + FN}$$

3) Precision: Precision is the proportion of the number of positive samples correctly predicted to the number of positive samples, mainly indicating the accuracy of the positive samples.

$$Precision = \frac{TP}{TP + FP}$$

TP (true positive) and TN (true negative) indicate how many predictions were correct. FP (false positive) and FN (false negative) indicate how many predictions were wrong.

The results of the model are then interpreted to determine whether it should be used in practise.

G. Deployment

Finally, it will be deployed into a data product which is a web application. A prototype for the web application is created to demonstrate how does the system work based on the model built. This prototype helps to identify about-to-withdraw customers and act accordingly to ensure that the e-commerce company can take the best-possible actions to retain those customers. Also, the prototype predicts the possible reasons of leaving for a customer which may give a better picture of customer thoughts. In the web application, it will have the basic features such as, a user-friendly interface, show all details of the dataset feed in and visualize it using different graphs then predict probability of customer churn. The web application will be done through Streamlit which turn data scripts into shareable web apps in minutes, only applicable to Python language and also an open source plus free of charge platform.

V. ANALYSIS/MODELS

A. EDA/Data Visualization

To clearly understand and analyse our data, charts or graphs are plotted as follows:

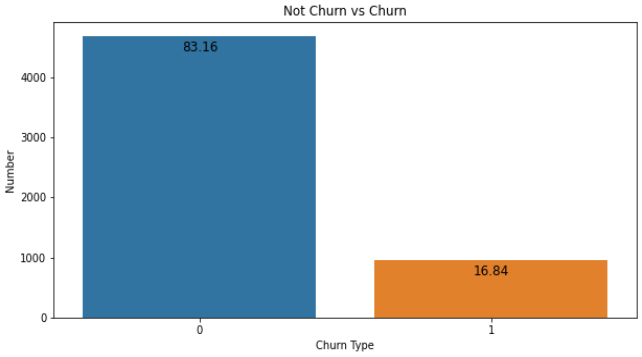

Fig.4 Bar chart of Not Churn (0) vs Churn (1) With Percentage

Based on Fig.4, the x-axis of the chart represents the different groups being compared (churned customers vs. non-churned customers), while the y-axis represents the percentage of customers in each group. It shows that more not churn customer data than churn customer data exist in the dataset used.

This indirectly indicates that the target of the dataset is imbalanced, so data resampling is needed before train test data splitting and modelling.

Fig. 5 presents different histograms of the distribution of certain columns in the dataset to gain valuable insights. To understand the histogram of recency of customer, recency is gotten from the day since last order of each customer. Customers with low number of days since last order are recent customers and vice versa.

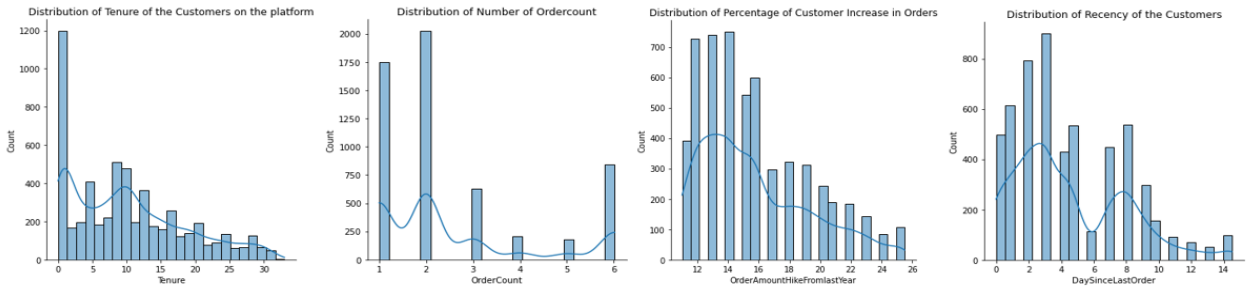

Fig. 5 Histograms for Different Distribution of Customer Behaviour and Pattern

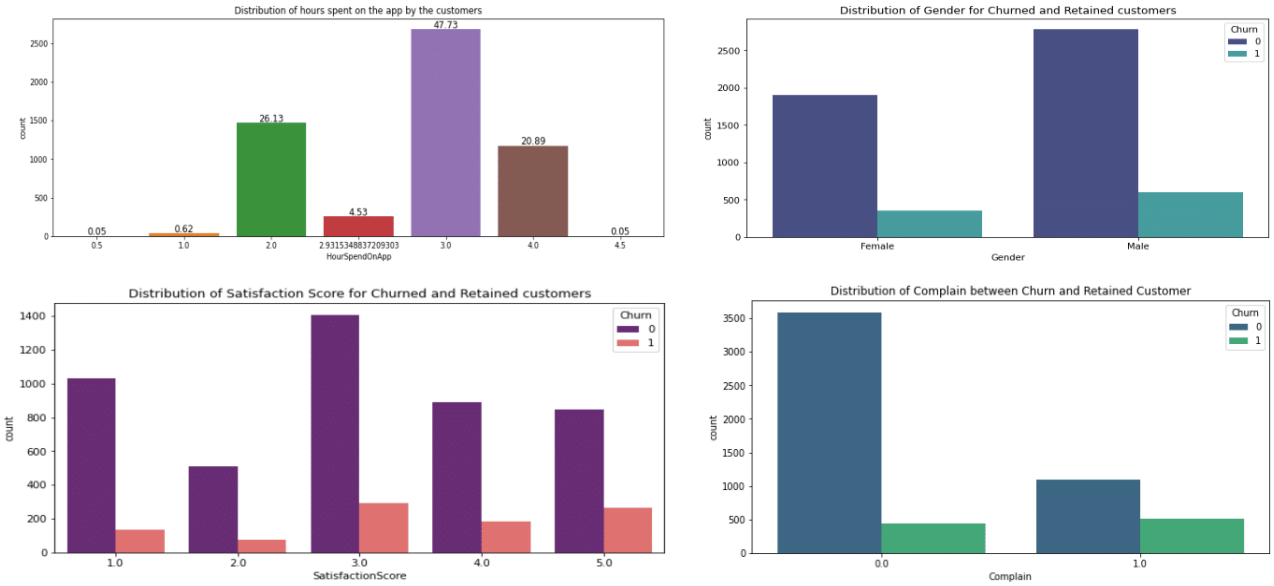

Fig. 6 Bar Charts Showing Difference in Characteristics of Churned and Retained Customer

Based on Fig. 6 above, some information can be concluded. For instance, most of the customers spend 3 hours on the e-commerce app, and 99.28% of the customers spend between 2 and 4 hours on the app. Other than that, customers with complain have higher churn rate than customers without complain.

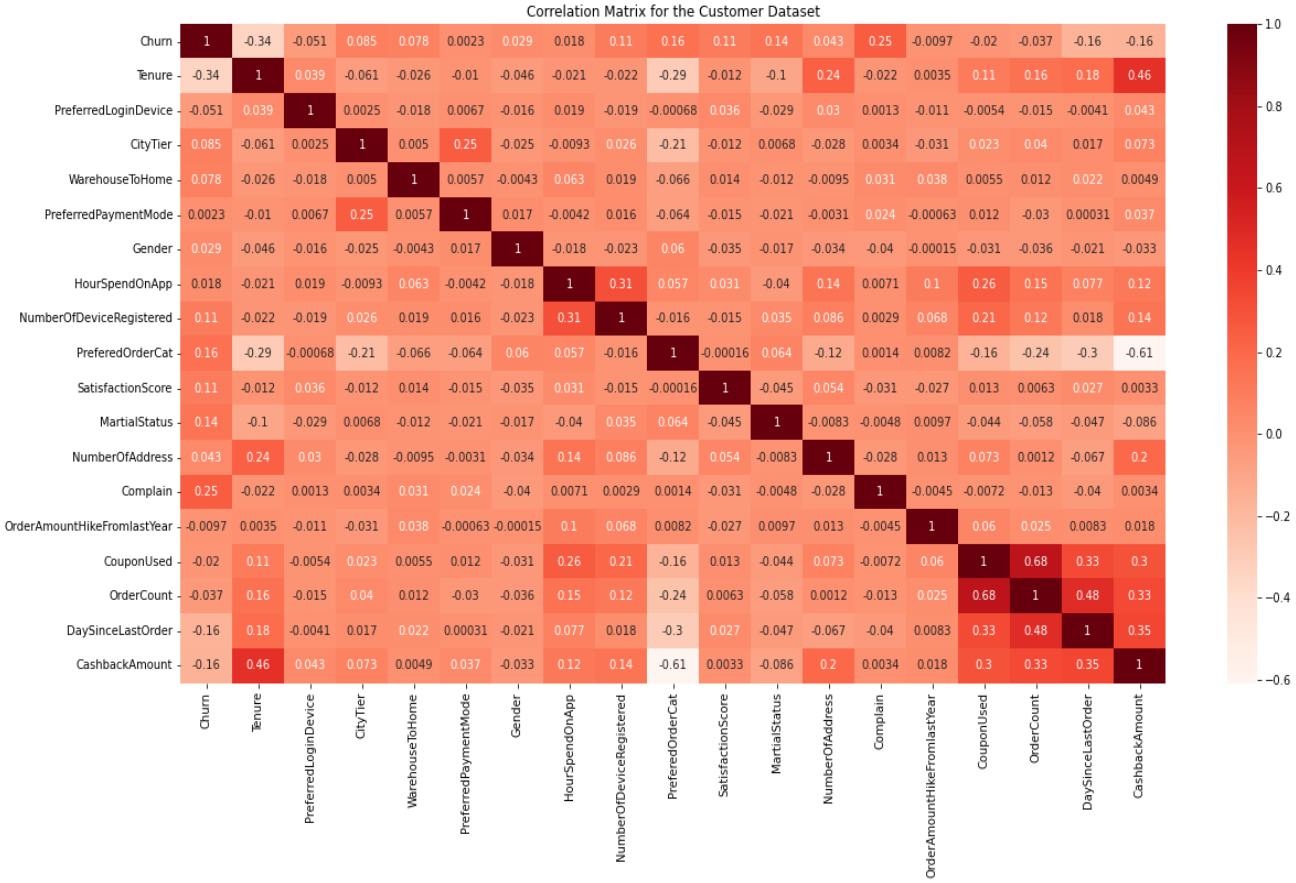

Fig. 7 Heatmap to Show Relationship Between Churn or Not Churn with Each Variable

By looking at Fig. 7, each square shows the correlation between the variables on each axis. Correlation ranges from -1 to +1. Values closer to zero means there is no linear trend between the two variables. The close to 1 the correlation is the more positively correlated they are, that is as one increases so does the other and the closer to 1 the stronger this relationship is. A correlation closer to -1 is similar, but instead of both increasing one variable will decrease as the other increases. The diagonals are all 1 because those squares are correlating each variable to itself (so it's a perfect correlation). For the rest the larger the number and darker the colour the higher the correlation between the two variables. So, this heatmap can be used to quickly identify which variables are most strongly related to customer churn where it can be observed that “Complain” is the most correlated variable with “Churn” (+0.25) whereas “Tenure” is the least correlated variable with “Churn” (-0.34).

B. *Model Development*

1) Data Splitting: To ensure that models are developed to provide the highest performance, a number of procedures must be taken before moving deep into the modelling phase. Data must first be divided into training and test sets. 80% of the overall dataset will be made up of training data, and 20% will be test data. The training set is used for model training while the testing set is used for model evaluation.

2) Data Resampling: Just like what had mentioned in the bar chart showing churn and not churn percentage of data, the target data is imbalanced, so data resampling is needed. Normally, SMOTE (Synthetic Minority Oversampling Technique) is used but in this project, SMOTEENN is applied, this resampling method combine SMOTE with Edited Nearest Neighbour (ENN) using Python to balance the dataset. Integrating this technique with oversampled data done by SMOTE helps in doing extensive data cleaning. Here on misclassification by NN’s samples from both the classes are removed. This results in a more clear and concise class separation. Table II shows the counts of target data before and after resampling.

TABLE II
BEFORE AND AFTER DATA RESAMPLING USING SMOTEENN

State	Retained Customer	Churned Customer
Before Resample	4682	948
After Resample	4660	3776

3) Model Building: In the modelling part, GridSearchCV is used in choosing the best model among Logistic Regression, LDA, Random Forest and XGBoost. The hyperparameters of a model are one essential aspect in determining how well it performs; if the correct values for the hyperparameters are selected, a model's performance can be considerably enhanced. Therefore, that is the reason for using GridSearchCV to determine the ideal values for a model's hyperparameters. This procedure is also known as hyperparameter tuning.

VI. RESULTS

A. *Feature Importance*

Fig. 8 is a chart that represents the importance of different features that contribute to the model, meaning data in the ‘tenure’ attribute is the most influential column towards our e-commerce customer churn prediction. We can observe the descending order of level of importance for the features in this graph and analyse the top 5 factors that cause customer churn in e-commerce, which are, tenure, complaint, cashback amount, number of days since last order and preferred order category by customers. There is, however, a discrepancy between the prior Heatmap results presented in this study, where "Complaint" is the value, most strongly connected with "Churn," and the Feature Importance graph, where "Tenure" is the most significant feature in determining churn results. To explain that, compared to "Tenure" (0 to 13), the variable "Complaint" has a significantly smaller range of possible values (0 and 1). This means that each random forest tree can only employ this binary characteristic once, but "Tenure" may appear much more frequently on different levels of the random forest trees. As a result, such a binary characteristic will be given relatively little weight or frequency, but a great deal of gain and coverage relevance.

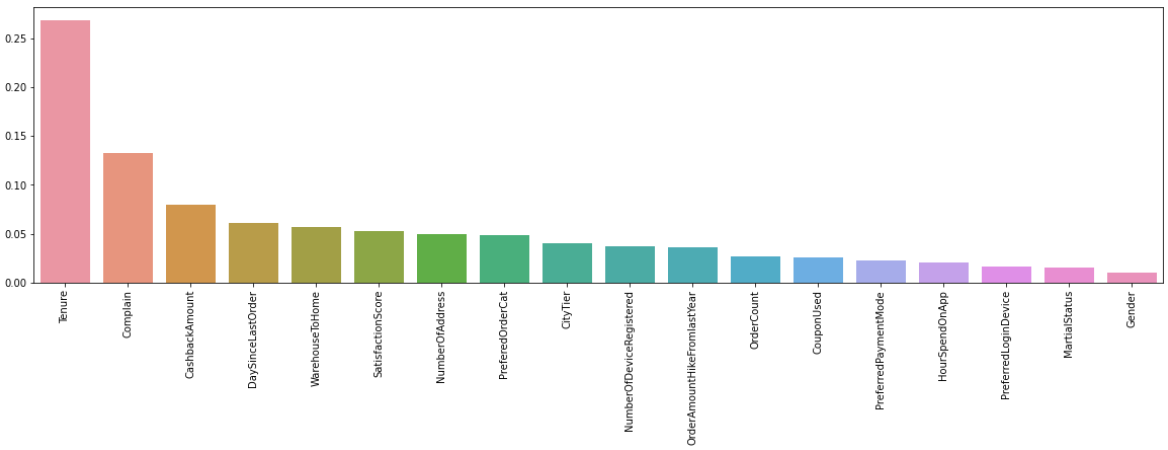

Fig. 9 Feature Importance Graph Showing Variable Importance Level in Descending Order

TABLE III
ACCURACY SCORE OF MODELS

Model	Parameter	Rank	Accuracy
RandomForestClassifier(criterion='entropy')_entropy	{'clf': RandomForestClassifier(criterion='entropy'), 'clf__criterion': 'entropy'}	1	0.987700
RandomForestClassifier(criterion='entropy')_gini	{'clf': RandomForestClassifier(criterion='entropy'), 'clf__criterion': 'gini'}	2	0.986515
XGBClassifier()_0.3_0.1_0.01	{'clf': XGBClassifier(), 'clf__learning_rate': 0.3, 'clf__reg_alpha': 0.1, 'clf__reg_lambda': 0.01}	3	0.968287
XGBClassifier()_0.3_0_0.01	{'clf': XGBClassifier(), 'clf__learning_rate': 0.3, 'clf__reg_alpha': 0, 'clf__reg_lambda': 0.01}	4	0.967991
XGBClassifier()_0.3_0.01_0.1	{'clf': XGBClassifier(), 'clf__learning_rate': 0.3, 'clf__reg_alpha': 0.01, 'clf__reg_lambda': 0.1}	5	0.967843

B. *Model Evaluation and Best Predictive Model*

The performance of the model is evaluated using some evaluation metrics. Before showing the model performance of the best model, Table III above is the scoring of accuracy for the models with different parameters feed in GridSearchCV. Based on the table, it is obvious that Random Forest model ranked number one and has the highest accuracy score among other models, which is 0.9877. Thus, Random Forest is the best predictive model chosen. By using this model to do prediction on test set, the confusion matrix is visualized as the Fig. 10:

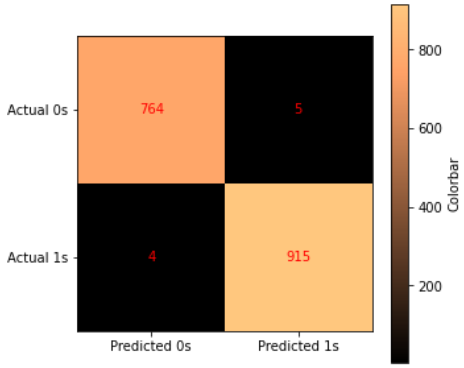

Fig. 10 Confusion Matrix

To interpret the confusion matrix, it means this model correctly predicted 764 not churn data and 915 churn data, and only wrongly predicted 5 not churn data and 4 churn data. Only the confusion metric report and ROC curves of the best performing classifier which is Random Forest will be displayed. So, from Table IV, the classification report below, it can be analysed where the model has high rate of 99% in all evaluation metrics, such as accuracy, precision, recall and F1-score. Fig. 11 also shows high value of area under ROC Curve, which is 0.99.

TABLE IV
CLASSIFICATION REPORT OF BEST MODEL: RANDOM FOREST

	Precision	Recall	F1-score	Support
0	0.99	0.99	0.99	769
1	0.99	1.00	1.00	919
accuracy			0.99	1688
macro avg	0.99	0.99	0.99	1688
weighted avg	0.99	0.99	0.99	1688

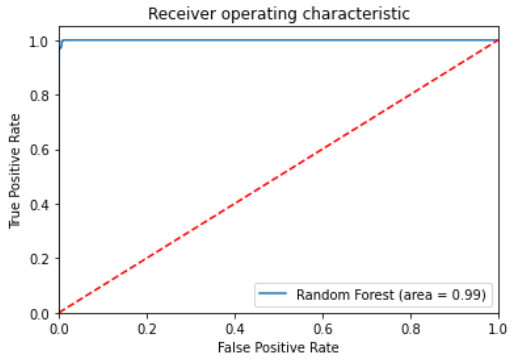

Fig. 11 ROC Curve of Best Model: Random Forest

C. *Prototype-Web Application*

After the predictive model is built successfully, it is deployed into a web application that enables public users to access it. For the application, which is the end product of this project is deployed using Streamlit with Github. The link to the E-Commerce Customer Churn Analysis and Prediction Web Application is https://jiehui0827-ecom-customerchurn-churn-app-1hdb4m.streamlit.app/. This application consists of different pages for user to navigate through:

1) Home Page: This is the interface of the main page of the application when once user clicks into the link to the website. This is also the default page for the web application. It is just showing the brief overview on what this application is for. Refer to Fig. 12 below for this page view.

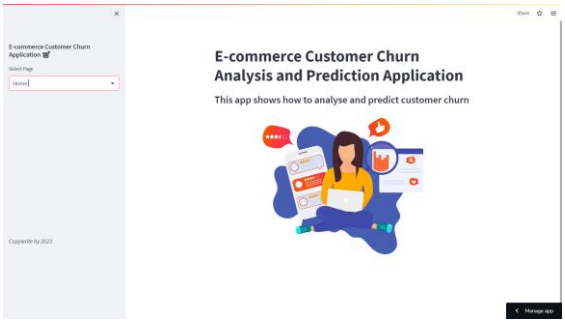

Fig. 12 Home Page

2) User Manual Page: This is where the users can read the information and know the ways to use the application, either

predict customer churn through online key in function or upload a set of customers data in file for prediction. So, if the user prefer choice 1 which is manual insert data, he or she can follow the instruction for choice 1, while for choice 2 is the same concept. Refer to the website for this page view.

3) Analysis and Prediction Page: This is the most important page where the user can interact with the web application to fill in customer details then predict whether the customer will churn or not churn. User can firstly choose manual insert customer data or insert csv file with customer data, if the user chooses manual, he or she can just key in for number input boxes or select value in boxes, then the user can view the final input data as a table below "Customer Details". When the user wants to predict that customer will churn or not, just simply press the button "Predict Customer Churn & Probability", it will load the result and show either churn or not churn with churn probability and the pie chart of the probability output (Fig. 14). If the user choose file, he or she need to upload a file with the required features and data, then the dataset details will be shown, and then it takes some time to clean and process the file. After that, it will show the missing values cleaned, and the user can choose which kind of graph to explore, for example, the user can select order count to view the total number of orders made by customer. Then, since the user file have the list of customer ID, he or she can select the customer that wish to know whether will churn or not using the customer ID, the app will predict and show the result same as the manual insert one. Refer to Fig. 15 for this page view.

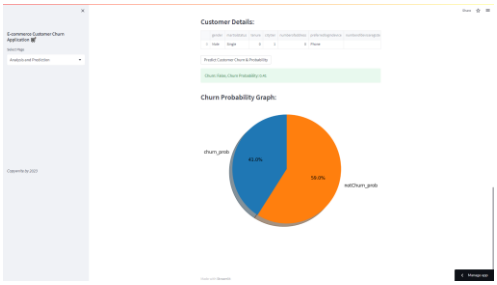

Fig. 14 Example of Churn Probability Pie Chart

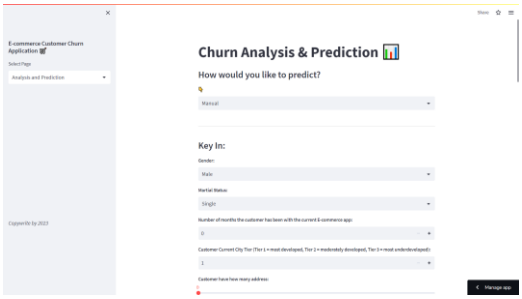

Fig. 15 Analysis and Prediction Page

4) About Page: This page basically just the introduction of author for this app, and credits to the dataset source and also the code for this app in GitHub link. Refer to the website for this page view.

5) Sidebar Menu: There are four menus for this application where it can be seen from the sidebar, and user can select it to browse to various pages. Firstly, the main menu, which is the Home page, secondly the User Manual page, then the Analysis and Prediction page, lastly the About page. Refer to the website for this sidebar view.

VII. DISCUSSION

To discuss on the best model, Random Forest is sensitive to noise and outliers, so it is important to pre-process the data before training the model and remove any irrelevant data or outliers. Only that the accuracy of the predictive model can be high. However, for this project, the model accuracy is considered very high of 99%, which normally be interpreted due to overfitting. The truth is nothing to do with overfitting. To simplify, Random Forest consists of 1) fully grown trees, 2) built on bootstrapped data, 3) and the majority vote rule to make predictions. Think about a straightforward binary classification issue just like this project. More than half of all trees in the final ensemble "know" the correct class for each observation if each tree is fully developed and each leaf is a pure leaf, and if each observation has a 62.5% chance to be sampled during bootstrapping. With the majority vote rule in place, that leads to 99%, or nearly 100% accuracy on the training set.

While for the E-commerce Customer Churn Analysis and Prediction Web Application, based on the churn probability pie chart shown in the 'Analysis and Prediction' page, by comparing the percentage of customers who have churned to those who have not, a company can get a sense of how effective their retention efforts have been. If the percentage of churned customers is high, it may indicate that the company needs to take steps to improve customer satisfaction or address issues that are causing customers to leave. On the other hand, if the percentage of churned customers is low, it may indicate that the company is effectively retaining its customers. Thus, this will be useful especially to e-commerce company as it is important to note that churn rate is a key metric for businesses, and it is a good indicator of how well a company is keeping its customers. The higher the churn rate the less stable the customer base is and the more difficult it will be to grow.

Lastly, since the Feature Importance graph already shown the importance ratios of factors affecting customer churn from high to low, so e-commerce platform should take note to retain longer tenure customer, cautious on customer complaints, provide more cashback and do more promotion once detect any customer churn based on the prediction application.

VIII. CONCLUSION, LIMITATION, AND FUTURE DIRECTIONS

Predicting customer churn often aims to pinpoint at-risk customers so that proactive measures can be taken to keep them, including individualized incentives or targeted marketing campaigns. It is crucial to remember that the ability to estimate customer churn properly can assist e-commerce businesses in identifying and resolving the root causes of customer churn. Additionally, by identifying the customers who are most likely to remain loyal, churn prediction can assist e-commerce platforms in maximizing their efforts to acquire new customers. Four different machine algorithms were applied to predict customer churn which are Logistic regression, Linear

Discriminant Analysis, Random Forest and XGBoost. It was found that Random Forest model has the best accuracy at 99%.

However, actually attempting to address the issue that it is challenging to accomplish high-precision customer churn prediction with a single model, my Random Forest prediction model and the E-commerce Customer Churn Analysis and Prediction App in Streamlit still have some limitations. One limitation of using a random forest prediction model for e-commerce customer churn prediction is that it may not be able to handle a large number of categorical variables. Random forest models are well-suited to handling a large number of input variables, but they are not as effective when the input variables are categorical. This is because random forests use decision trees as the building blocks, and decision trees are less effective at handling categorical variables. Another limitation is that random forest models can be complex and interpreting the results can be difficult. Since it is an ensemble model, it is hard to understand the feature importance and specific rule it uses for the prediction. This can make it difficult for business users to understand and use the model's predictions.

Therefore, to ensure a better model is developed in future for customer churn prediction in e-commerce, building combined prediction model will be a consideration for future work. This is because based on recent research done, it shows that the combined prediction model can not only have a better interpretation ability like a decision tree model, but also a higher prediction accuracy rate of a neural network model, which can better make up for the shortcomings of a single prediction model and can also get more stable and accurate prediction results.

ACKNOWLEDGMENT

On this great occasion of successful completion of my assignment, a very sincere appreciation to my supervisor, Dr. Azah for her guidance and support throughout the project. I also would like to thank all the organizations, for example IEEE Xplore and Kaggle, which provided the data source and information retrieval for my current project. Last but not least, I would like to thank everyone who motivated me to work on this assignment.

REFERENCES

[1] Mathai, Paul. (2020). Customer Churn Prediction:A Survey.

[2] Bi, Q.Q. Cultivating loyal customers through online customer communities: A psychological contract perspective. *J. Bus. Res.* **2019**, *103*, 34–44.

[3] Maria, O.; Bravo, C.; Verbeke, W.; Sarraute, C.; Baesens, B.; Vanthienen, J. Social network analytics for churn prediction in telco: Model building, evaluation and network architecture. *Expert. Syst. Appl.* **2017**, *85*, 204–220.

[4] Gordini, N.; Veglio, V. Customers churn prediction and marketing retention strategies: An application of support vector machines based on the AUC parameter-selection technique in B2B e-commerce industry. *Ind. Market. Manag.* **2017**, *62*, 100–107.

[5] Benitez, C. (2022, October 26). *25+ top ecommerce statistics and trends for 2023*. Startup Bonsai. Retrieved January 26, 2023, from https://startupbonsai.com/ecommerce-statistics/

[6] Wu, X. J., & Meng, S. S. (2017). *Research on e-commerce customer churn prediction based on customer segmentation and Ada-Boost.* Industrial Engineering, 20(02), 99- 107.

[7] Shao, D. (2016). *Analysis and prediction of insurance company's customer loss based on BP neural network.* Lanzhou University

[8] Lu, N., Liu, X. W., & Lee, L. (2018). *Research on customer value segmentation of online shop based on RFM.* Computer Knowledge and Technology, 14(18), 275-276, 284.

[9] Huang, J. (2018). *A Comparative Study of Social E-Commerce and Traditional Ecommerce.* Economic and Trade Practice, (23), 188-189.

[10] Agrawal, S., Das, A., Gaikwad, A., & Dhage, S. (2018, July). *Customer churn prediction modelling based on behavioural patterns analysis using deep learning.* In 33 2018 International conference on smart computing and electronic enterprise (ICSCEE) (pp. 1-6). IEEE.

[11] Saghir, M., Bibi, Z., Bashir, S., & Khan, F. H. (2019, January). *Churn prediction using neural network-based individual and ensemble models.* In 2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST) (pp. 634-639). IEEE.

[12] Wu, S., Yau, W. C., Ong, T. S., & Chong, S. C. (2021). *Integrated Churn Prediction and Customer Segmentation Framework for Telco Business.* IEEE Access.

[13] M. Jaeyalakshmi, S. Gnanavel, K. S. Guhapriya, H. Phriyaa, and K. Sree, "Prediction of Customer Churn on eRetailing," *International Journal of Recent Technology and Engineering*, no. 6, pp. 2277–3878, 2020, doi: 10.35940/ijrte.F9550.038620.

[14] P. Jayaswal, B. R. Prasad, and S. Agarwal, "An Ensemble Approach for Efficient Churn Prediction in Telecom Industry," *International Journal of Database Theory and Application*, vol. 9, no. 8, pp. 211–232, 2016, doi: 10.14257/ijdta.2016.9.8.21.

[15] I. Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam, and S. W. Kim, "A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector," *IEEE Access*, vol. 7, pp. 60134–60149, 2019, doi: 10.1109/ACCESS.2019.2914999.

[16] Alshamsi, A. (2022). *Customer churn prediction in ECommerce sector*. RIT Scholar Works. Retrieved November 9, 2022, from https://scholarworks.rit.edu/theses/11183/

[17] Dick, A.S.; Basu, K. *Customer Loyalty: Toward an Integrated Conceptual Framework.* J. Acad. Mark. Sci. 1994, 22, 99–113.

[18] Gefen, D. *Customer Loyalty in e-Commerce.* J. Assoc. Inf. Syst. 2002, 3, 2.

[19] Buckinx, W.; Poel, D.V.D. Customer base analysis: *Partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting.* Eur. J. Oper. Res. 2005, 164, 252–268.

[20] Matuszelański, Kamil & Kopczewska, Katarzyna. (2022). *Customer Churn in Retail E*-Commerce Business: Spatial and Machine Learning Approach. Journal of Theoretical and Applied Electronic Commerce Research. 17. 165-198. 10.3390/jtaer17010009.

[21] Huber, S., Wiemer, H., Schneider, D., & Ihlenfeldt, S. (2019). *DMME: Data mining methodology for engineering applications–a holistic extension to the CRISP-DM model.* Procedia Cirp, 79, 403-408.

[22] Kaggle. (2021, March 8). *E-commercecustomerchurn*. Kaggle. Retrieved January 26, 2023, from https://www.kaggle.com/code/ankitverma2010/e-commercecustomerchurn/data

[23] García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J. M., & Herrera, F. (2016). Big data preprocessing: methods and prospects. *Big Data Analytics*, *1*(1), 1-22.

[24] Eric. (2021). *Introduction to Handling Missing Values.* Aptech. Retrieved from https://www.aptech.com/blog/introduction-to-handling-missing-values/

[25] Liu, H., Shah, S., & Jiang, W. (2004). On-line outlier detection and data cleaning. *Computers & chemical engineering*, *28*(9), 1635-1647.

[26] Thomas W. Edgarm, D. O. (2017). *Research Methods of Cyber Security.*

[27] Sarkar, P. (2019, September 30). *What is linear discriminant analysis(lda)?* What is Linear Discriminant Analysis (LDA)? Retrieved November 9, 2022, from https://www.knowledgehut.com/blog/data-science/linear-discriminant-analysis-for-machine-learning

[28] Random Forest. (2020, December 7). Retrieved from IBM: https://www.ibm.com/cloud/learn/random-forest#:~:text=%20What%20is%20random%20forest%3F%20%201%20Decision,ba

[29] *XGBoost documentation*. XGBoost Documentation - xgboost 1.7.1 documentation. (2022). Retrieved November 9, 2022, from https://xgboost.readthedocs.io/en/stable/