Customer Churn Prediction with RFM Analysis

Prepared By: Yew Jin-Nee Supervisor: Dr. Hoo Wai Lam

Abstract—The evolution of e-commerce has heavily influence how businesses manage their relationship with customers. To become a customer-centric organization, businesses are taking advantage of the massive amount of customer data readily available to plan their strategies and make data-driven decisions. Therefore, businesses are constantly trying to understand their customers from various perspectives to identify the suitable market audience that can maximize revenue. A common application is Recency, Frequency, and Monetary (RFM) analysis that is derived from customers' purchasing behaviour. Furthermore, customer churn prediction is also a crucial application in ensuring a business survival. In this project, RFM analysis is incorporated in the customer churn prediction model by analysing transactional data provided by Olist Store. Three different algorithms, namely Weighted Logistic Regression, Isolation Forest, and One-Class Support Vector Machine have been used in model training that addresses the issue of imbalanced data. The main metrics used to evaluate the models were Recall and Balanced Accuracy. Besides that, the effects of the dataset size on the model performance was also investigated. Lastly, a dashboard was developed that summarizes the findings in this project such as the model performance and characteristics of churning and non-churning customers. The results shown found that One-Class Support Vector Machine is the overall most suitable algorithm with a recall score of 0.59 and balanced accuracy score of 0.54.

Index Terms—Churn Prediction, e-Commerce, RFM Analysis, Imbalanced Data, One-class Classification.

1 Introduction

Customer relationship management (CRM) consists of a set of processes and enabling systems that support a business strategy with the purpose of building long term, profitable relationships with specific customers according to Ling and Yen (2001). The exponential growth of the Internet and big data along with the rapid development of information technology (IT) tools have changed the marketing landscape and transformed the ways customers' relationships with businesses are managed (Ngai et al., 2009). Hence, the need for organizations to implement analytical CRM where customer characteristics and behaviours are analysed have greatly increase. Moreover, a successful CRM strategy requires customer data and IT tools as the fundamental building blocks. According to Monalisa and Kurnia (2019), customers are the organization's asset that is essential to maintain, handle, and understand their needs. However, this raises the questions on knowing how and which customers to retain. The most apparent answer would be to preserve loyal customers with high profitability. Based on research done by Bahari and Elayidom (2015), organizations with the ability to communicate meaningfully with customers to understand and influence their behaviour have shown improvement in terms of customer acquisition, loyalty, retention, and profitability. Therefore, it is evident that analysing customer behaviours leads to an increase of old and new potential customers so as to maximizing customer value (Ngai et al., 2009).

In light of the above mentioned, maintaining customers should be one of the top priorities of commercial entities. In order to achieve a high customer retention rate, businesses need to be capable of predicting customers churn. According to Huang et al. (2012), customer churning occurs when commercial organizations suffer lost in terms of customers to their competitors. The term customer churning ,also known as customer attrition is a marketing term that depicts a customer gaining interest in another organization or product (Hung et al., 2006). After a customer churns, the organization will be faced challenges such as reduction in sales, revenues and an increase in customer acquisition cost. This grave issue poses a huge threat to various industries, such as e-commerce, telecommunication, and finance. The research by Van den Poel and Lariviere (2004) highlights the importance of customer churn management in ensuring the survival of an organization. Due to the significance of this issue, there are many studies conducted to predict customer churn and its function, for example researches by Ali and Arıtürk (2014), Hung et al. (2006), and Vafeiadis et al. (2015).

Furthermore, e-commerce has come a long way since the world's first e-commerce company was introduced, which was Boston Computer Exchange, an online marketplace for people to sell and buy used computers. In addition, retail e-commerce sales in the United States alone were $375 billion by the end of 2020 and experts forecast sales to exceed $476 billion by the end of 2024 (BigCommerce, 2021). While e-commerce has been continuously developed for decades, the remarkable surge in online sales have indicated that customers, products' quality and services have also evolved. On the other hand, competition among businesses also grew more intense as businesses race against each other to retain and attract more customers. Consequently, many organizations have been adopting consumer-centric business strategies such as predicting customer activities to maximize profit. For instance, Chen et al. (2012) have stated online retailers are trying to implement data mining practices coupled with business metrics to increase revenue, while research by He and Li

(2016) said that "enterprise business strategy should be also be "customer centric", because only to find the target consumer group, can the merchants know that to who the value recommended and what value to give customer."

It is crucial that organizations are able to distinguish between churning and non-churning customers because this clear understanding can give the organizations their competitive edge in the market. Conversely, organizations who treat all their customers in a similar manner will end up losing revenue and wasting a large portion of the organizational resources. For example, if an organization is constantly reaching out to a non-churning customer rather than a churning customer, the probability of the latter actually churning will drastically increase and resources used on the former are squandered. Therefore, an organization's success can be achieved by implementing models to accurately identify customers who are at risk of churning in the near future (Jahromi et al., 2014). Many customer churn prediction models have utilized machine learning techniques such as artificial neural networks (ANN), decision trees (DT), support vector machine (SVM), and logistic regression (LR). Besides that, there are models that also use behavioral variables to predict customer churn. One common example is recency, frequency, and monetary (RFM) analysis, a data mining method used to predict customer behavior that is widely adopted in the field of direct marketing (Wei et al., 2010).

In this study, customer churn prediction would be performed based on transactional data provided by Olist Store, a Brazilian online marketplace. The framework proposed includes expanding RFM by integrating suitable variables such as the customer average review score and demographic information, comparing the performance of machine learning algorithms in predicting customer churn, and the algorithms' performance on different customer churn time periods. The objectives of this study is to predict customer churn based on their purchasing behaviour, develop a dashboard that summarizes the characteristics of the churned customers, and determine the effects of observation period length on model performance. The insights obtained should be able to aid businesses better understand their customers and make data-drive decisions.

2 Literature Review

Customer churn prediction is often implemented in the business domain. Based on the studies by marketing experts, many companies annually lose about 25 per cent of their customers on average (Chiang et al., 2003). There also have been reports that some companies suffer customer churn up to 36 percent (Dursun and Caber, 2016). While companies that successfully reduce customer churn by 5 percent, reported that there was a 25 (Marcus, 1998; Reichheld, 1993) to 85 percent (Ivanovic et al., 2011; Reichheld and Sasser, 1990) increase in revenues. Furthermore, companies can plan more targeted and personalized marketing strategies by analyzing customer information and identifying the customers' value (Panuš et al., 2016).

According to Osisanwo et al. (2017), Machine Learning is defined as "the automated detection of meaningful patterns in data with tools endowed with the ability to learn and adapt". The application of Machine Learning such as data mining has also become one of the fundamental building blocks of Information Technology today especially with the growing trend of Big Data analytics and technologies. Moreover, there are two types of Machine Learning algorithms that are differentiated by the expected outcome. For example, inputs that are mapped to desired outputs based on a generated function is known as supervised learning. Classification problems are a common instance of supervised learning where the aim is to train a computer to learn the classification model built. This study is considered as a classification problem because a learner will learn a function which maps a vector into one of the two classes, customer churned or customer retained, by looking at the input and output examples in the training set.

There are several related works that have involved RFM variables in developing prediction and classification models that can facilitate CRM. For instance, Cui et al. (2006) proposed a Bayesian Networks approach that uses RFM variables to predict a customer's response to direct marketing. In this study, customers' purchases were predicted with different algorithms such as Bayesian Networks, Artificial Neural Networks (ANN), Classification and Regression Trees (CART), and latent class regression. Data such as lifetime and consumer transaction variables along with RFM variables were used to train the models. The results show that Bayesian Networks have higher classification accuracy compared to the other models. While in Cheng and Chen (2009) research, customer value analysis was conducted by segmenting customers with K Means algorithm based on their consuming behavior which was represented with a RFM model. Classification rules were then generated using rough set Learning from Examples Module, version 2 (LEM2) algorithm; the accuracy of these rules were evaluated using Decision Tree, Artificial Neural Networks and Naive Bayes algorithms. The findings of this study showed positive results for the proposed procedure regardless of number of output classes as it has higher accuracy rate than the other listed models. Furthermore, Etzion et al. (2005) integrated RFM variables into their generalized Customer Lifetime Value (CLV) model approach, based on Markov Chain Models to predict customer behaviour in e-commerce. This study focuses on providing a comprehensive method to learn e-customers' behavior and calculate their respective lifetime value. In research from Zalaghi and Abbasnejad Varzi (2014), customer loyalty was determined using an extended RFM model and K-Means algorithm. Besides that, a multipurpose genetic algorithm was also implemented to minimize the number of attributes in this study. Once the attributes were selected, the RFM properties were assigned a Spearman's correlation coefficient proportionate to its importance. The suggested method was able to differentiate customer loyalty with good precision.

Besides that, application of machine learning algorithms on customer behaviour prediction is also widely adopted in various industries. Mozer et al. (2000) had implemented three machine learning algorithms, namely logit regression, C5.0 decision tree, and neural network to predict customer churn in the wireless telecommunications industry. The study found that all three predictors yield similar lift curves; however, the predictors produced significantly different estimated cost savings. Another example of comparing machine learning techniques in churn prediction in the telecommunication sector involved Artificial Neural Network, Support Vector Machines, Decision trees learning, Naive Bayes, and Logistic Regression (Vafeiadis et al., 2015). This study used a series of Monte Carlo simulation for each algorithm at different parameter settings and found that all algorithms performed significantly better in the boosted version and the best performing algorithm was SVM-POLY boosted with AdaBoost achieving almost 97% accuracy and over 84% F-measure. Next, Khodabandehlou & Zivari Rahman (2017) also compared various supervised machine learning techniques to predict customer churn in an Iran food store. The algorithms compared were Artificial Neural Network, Support Vector Machine and Decision Trees; the results found that Artificial Neural Network had the highest accuracy, while Decision Trees had the lowest accuracy. In another study carried out by Bahari and Elayidom (2015), customer behaviour towards direct bank marketing campaigns was predicted using Naive Bayes and Neural Networks. The results obtained indicates that the multi-layer perceptron neural network (MLPNN) model performed slightly better with an accuracy of 88.63% compared to the Naive Bayes model (87.97%).

Classification of imbalanced data is also a prominent issue in real world applications as it greatly impacts the learning performance of classification algorithms. Hence, a wide range of solutions had been proposed; in a study conducted by Ertekin et al. (2007), active learning with Support Vector Machine algorithm has been implemented to solve this issue and an early stopping criteria is used to enable the algorithm to converge faster. The performance of the proposed solution was compared to other resampling techniques and the results showed that the active learning has a higher predictive performance. While Sun et al. (2009) offered multiple research solutions from different perspectives, such as resampling the data space, boosting algorithms, using cost-sensitive learning, and adapting existing algorithms by introducing learning bias or applying one class learning. These solutions are also seen in other works such as using improved one-class support vector machine (SVM) in customer churn prediction by Zhao et al. (2005). The results showed that one-class SVM performed best with RBF kennel function and the algorithm also outperformed Artificial Neural Network, Decision Tree, and Naive Bayes with an accuracy of 87.15%. While a research done by Schölkopf et al. (2001) suggested adapting the Support Vector Machine algorithm to perform one-class classification. The features were first transformed via a kernel where they treated the origin as the only member of the second class. By introducing 'relaxation parameters', they separated the image of the one class from the origin.

3 Problem Formulation

As mentioned by Albers (2000), customers can be treated as individuals in the online environment because of their corresponding transactional data which allows organizations to provide various offerings tailored to the customer's preference. In contrast, too many distinct offerings as such in the offline environment would not be viable. As marketers are aware of the situation, this leads to the challenge of discovering the most suitable offering for each customer. These offerings include loyalty programmers, directing marketing campaigns, and promotions that require a great amount of time and money invested. This issue was also raised by Dogan et al. (2018) that marketers have difficulty identifying customer segments and effective marketing approaches for each segment which resulted in a waste of marketing resources.

Besides that, the problem is further amplified when there is an overwhelming volume of data and organizations lack the ability to transform these data into valuable information. There are only 15% of Fortune top 500 firms who are capable of analysing their customer data by combining data, analysis and technology according as stated in the research done by Panuš et al. (2016). Moreover, evaluation on the factors that affect customer purchasing behaviour is equally important because the results will affect the organization's choice of marketing strategies and tools which impacts the overall cost (Panuš et al., 2016). The underlying assumption is the more accurate the customer groups are segment, the more effective the directed marketing strategy can be plan. Therefore, a company needs to take the aforementioned problem into consideration in order to improve sales and develop insightful marketing strategies that will give a company the competitive edge in the industry.

4 Methodology

To develop the solution, Data Science methodology approach has been implemented. The following subsections represent each step in this process.

4.1 Data Acquisition

The dataset acquired is provided by a Brazilian e-commerce platform, Olist Store that is made publicly available on Kaggle. Olist connects small businesses and consumers from all over Brazil on one single platform. The merchants who are contracted with Olist will sell their products on the website and ship them directly to customers using Olist logistics partners. Once a product has been purchased, the merchants will get notified to fulfill the order. While customers will get a satisfaction survey via email where they can express their purchase experience after receiving the product, or when the estimated delivery date is due. The dataset consists of approximately 100k of customer transactional data from September of 2016 to October of 2018 that is stored in 8 entities, namely *olist_customer*, *olist_geolocation*, *olist_order_items*, *olist_order_payments*, *olist_order_reviews*, *olist_orders*, *olist_products*, and *olist_sellers*. The relationship between each entity is shown in Figure 1.

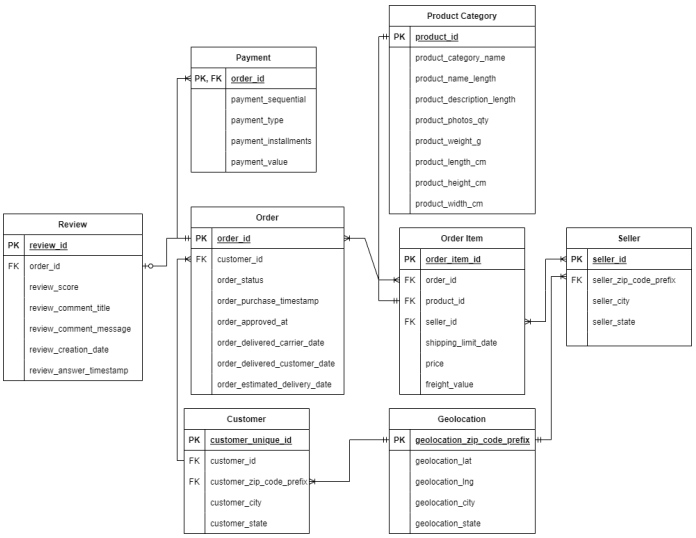

Fig. 1: Entity Relationship Diagram of the Dataset

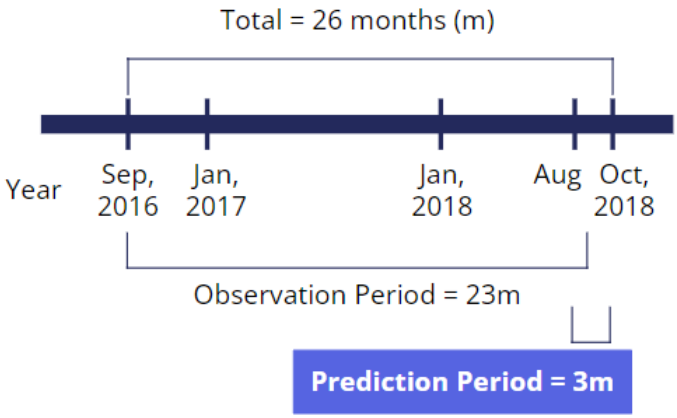

Fig. 2: Timeline of the Dataset

4.2 Data Preparation

4.2.1 Data Cleaning

Firstly, features such as *order_purchase_timestamp* and *order_delivered_customer_date* are converted from Python Object to Date objects with the datetime module. Therefore, it is easier to manipulate these objects with functions that handle, date, time, and time intervals. Secondly, all the missing values present in different features were imputed using median for continuous variables and mode for discrete variables. Median was selected because it is the measure of central tendency that is least affected by outliers. Thirdly, duplicated observations and unnecessary columns like *seller_id* and *order_id* were dropped from the overall dataset to reduce data redundancies and dimensionality.

4.2.2 Feature Engineering

To transform raw data into the desired form, new features such as *avg_review_score* and *avg_delivery_acc* were created from existing features. For example, the feature *avg_review_score* was calculated by taking the mean of the total review score of all the orders for each customer. While the binary target column was created based on the existing customers' order purchases in the prediction period (Figure 2). The three main features derived that represent a customer's purchasing behaviour are as follow:

- Recency: This feature represents the number of days since the last purchase made by a customer in the observation period. The higher the recency value of a customer, the higher the likelihood of the customer churning.
- Frequency: This feature represents the number of orders made by a customer during the observation period. The higher the ferquency value of a customer, the more engagement involved between the customer and the business.
- Monetary: This feature represents the total amount of money spent by a customer during the observation period. The higher the monetary value of a customer, the higher the amount of revenue gained by the business.

4.2.3 Label Encoding

The transformer, LabelEncoder from scikit-learn library was used to encode categorical features such as *customer_state* and *customer_city* with values from 0 to n_classes-1. This process is done because not all machine learning algorithms are able to handle categorical input data, therefore encoding takes place before fitting and evaluating a model.

4.2.4 Feature Scaling

Due to the negative influence of outliers on the sample mean and variance, RobustScaler from scikit-learn library was used to scale the features. This scaler uses the median and interquartile range to scale the features, thus it is robust to outliers. Furthermore, scaling takes place independently on each feature in the dataset.

4.2.5 Feature Selection

This is a process of selecting a subset of relevant and useful features to increase the model generalization capability. Therefore, a different combination of features as shown in Table 1 were used to train the model. The feature combination that yields the best model performance result would be selected.

4.2.6 Data Augmentation

The distribution of the target column is severely imbalanced as discovered through exploratory data analysis. Therefore, the minority target class in the training data is oversampled using SMOTE (Synthetic Minority Oversampling Technique). This technique synthesizes new observations from existing ones in the minority class by selecting data points that are close to each other in the feature space, drawing a line between the the points and generating a new observation at a point along the line (Brownlee, 2020b).

4.2.7 Train-Test Split

The data in the observation period was further divided into training data and testing data in the ratio of 8:2. This technique is implemented to test the model performance on general data. It is also used to prevent phenomenons such as overfitting or underfitting a model.

4.3 Modeling

Three algorithms were implemented for this stage and the primary reason these algorithms were selected is because the target column is severely imbalanced. Since the minority class is only 0.21% of the total sample, it is treated as outliers for Isolation Forest and One-Class Support Vector Machine. These algorithms are normally used for anomaly detection, however they are used for one-class classification in this project.

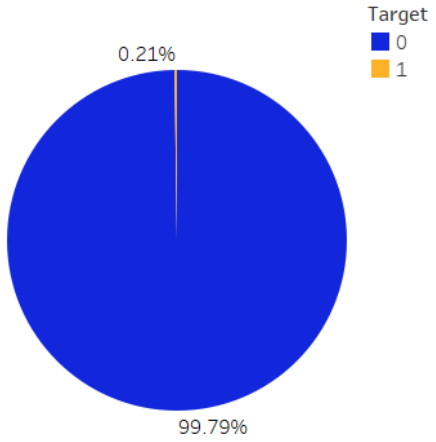

Fig. 3: Distribution of Target Classes

4.3.1 Weighted Logistic Regression

The default Logistic Regression algorithm is suitable for binary classification, but it is not effective in imbalanced classification. Therefore, the algorithm can be modified to take into account the skewed distribution by assigning weights to labels that are inversely proportional to the class frequencies. These weights will influence the coefficient value in the optimization algorithm that minimizes the negative log likelihood for the model. In other words, the weights influence the model to penalize less for errors on the majority class and vice versa during model training (Brownlee, 2020a). The loss function is as shown in equation 1:

$$\sum_{i=1}^{n} -(w_0 \times y_i log(\hat{y}_i) + w_1 \times (1 - y_i) log(1 - \hat{y}_i)) \quad (1)$$

where n is the number of observations, y_i is the true value for the *ith* observation, $\hat{y}_i$ is the predicted probability for the *ith* observation, w_0 and w_1 is the weights assigned to the minority and majority class respectively

4.3.2 Isolation Forest

According to Liu et al. (2008), Isolation Forest also known as iForest is an algorithm that explicitly isolates anomalies instead of profiling the normal data points. There are two processes involved; the first process is building isolation trees (iTrees) using subsamples of the training set to produces partial models. These iTrees are constructed by recursively partitioning the training set until data points are isolated or a specific tree height is reached. Data points with below average path lengths are considered to be anomalies and the two most important parameters is the sub-sampling size and number of tees. While the second process is passing the testing set through the isolation trees and obtaining the anomaly score that is derived from the expected path length, $E(h(x))$ for each data point. The $E(h(x))$ is obtained by passing the data point through every iTree in the iForest. An individual path length $h(x)$ is derived by counting the number of edges e from the root node to the terminating node as instance x tranverses through an iTree. This algorithm is highly efficient as it has a linear time complexity with low memomry requirements and low constant,thus making it an ideal algorithm for large databases.

4.3.3 One-Class Support Vector Machine

To predict customer churn, the improved version of one-class Support Vector Machine algorithm proposed by Li et al. (2003) was implemented. The algorithm classifies data points on the extremes of the density function as outliers by capturing the density of the majority class. The algorithm of One-Class Support Vector Machine is as follows: For any input x, the distance between the data point and center of the hypersphere is calculated if it follows the following condition:

$$||\phi(x) - x|| \leq R \quad (2)$$

The data point x will be considered as part of the +1 class if it belongs to the hypersphere, otherwise it belongs to the -1 class.

$$R^2 = 1 - \frac{2}{n}\sum_{k,i} a_i y_i K(x_k, x_i) + \sum_{i,j} a_i a_j y_i y_j K(x_i, x_j) \quad (3)$$

where x_k are the bounded vectors and n is the number of bounded vectors. Hence, the decision function can be written as:

$$f(x) = sgn(\sum_{i=1}^{l} a_i y_i K(x, x_i) + b) \quad (4)$$

where $b = -\frac{1}{n}\sum_{k,i} a_i y_i K(x_k, x_i)$

4.4 Evaluation

The following metrics were used to evaluate the model performance.

4.4.1 Recall

This metric is derived from the ratio of *tp / (tp + fn)* where *tp* is the number of true positives and *fn* is the number of false negatives. Recall represents the classifier's ability to find all the positive samples. The best value is 1 and the worst value is 0. Recall was selected to minimize false negatives as it will be more costly for the business. For example, if a model frequently predicts false positives, the business suffers small amount of marketing cost as it reaches out to churned customers. However, if the model frequently predicts false negatives, the business has a high possibility of losing revenue because it is not reaching out to potential loyal customers that will contribute much more to sales.

4.4.2 Balanced Accuracy

This metric is equivalent to accuracy metric, but it is specifically used to deal with imbalanced datasets. It is also defined as the average of recall from each class or the arithmetic mean of sensitivity and specificity. Since the default parameters were used where *adjusted = False*, the best value is 1 and the worst value is 0. Balanced accuracy score can be calculated according to equation 2:

$$\frac{1}{2}\left(\frac{TP}{TP+FN}+\frac{TN}{TN+FP}\right) \tag{5}$$

where *TP* is the number of true positives, *TN* is the number of true negatives, *FP* is the number of false positives, and *FN* is the number of false negatives.

4.4.3 Confusion Matrix

According to Pedregosa et al. (2011), this metric can be defined as $C_{i,j}$, where the number of observations known to be in group i and predicted to be in group j. This project is dealing with binary classification, thus $C_{0,0}$ is the count of true negatives, $C_{0,1}$ is the count of false positives, $C_{1,0}$ is the count of false negatives, and $C_{1,1}$ is the count of true positives. This metric is useful as it displays the count of a specific element in the confusion matrix.

4.5 Deployment

The data product developed is a dashboard that summarizes the sales performance of Olist Store, customers' preferences, and characteristics from the perspective of recency, frequency, and monetary. The name of the product is 'C.I.D.A.R', an acronym for '*Customer Is Definitely Always Right*'. It is built using Dash framework, which is written on top of Plotly.js and React.js and hosted on Heroku, a cloud platform.

5 Results

5.1 Variable Selection

The features in *Geo* represent the geographic location of the customers such as State and City; *RFM* represent the recency, frequency, and monetary value of the customers; *Avg* represent the average value of order features such as Review Score of customers; *Recall* represents the recall score and *Bal_Acc* represents the balanced accuracy score.

TABLE 1: Effects of Different Features based on Algorithm

Variable / Algorithm	Geo	RFM	Avg	Recall	Bal_Acc
Weighted Logistic Regression		✓		0.42	0.50
	✓			0.53	0.51
	✓		✓	0.55	0.57
	✓	✓	✓	**0.55**	**0.59**
Isolation Forest		✓		0.13	0.49
	✓			0.19	0.49
	✓		✓	0.23	0.53
	✓	✓	✓	**0.18**	**0.53**
One-Class Support Vector Machine		✓		**0.59**	**0.54**
	✓			0.52	0.51
	✓		✓	0.48	0.50
	✓	✓	✓	0.55	0.53

Based on the results in table I, Weighted Logistic Regression and Isolation Forest algorithm performed the best when all the features are used in model training. However, One-Class Support Vector Machine has the highest performance when only *RFM* features are incorporated. Thus, only *RFM* variables were used in the model training.

5.2 Model Performance

TABLE 2: Model Performance based on Algorithm

Algorithm / Metric	Weighted Logistic Regression	Isolation Forest	**One-Class Support Vector Machine**
Recall	0.42	0.13	**0.59**
Balanced Accuracy	0.50	0.49	**0.54**

According to the results displayed in Table II, One-Class Support Vector Machine has the highest performance in terms of recall (0.59) and balanced accuracy (0.54).

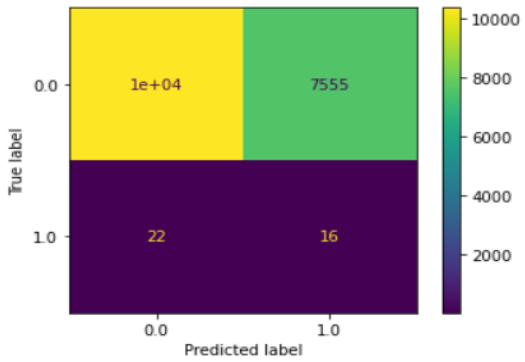

Fig. 4: Confusion Matrix of Weighted Logistic Regression

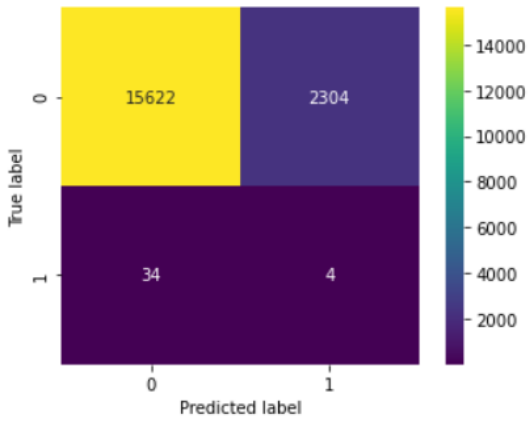

Fig. 5: Confusion Matrix of Isolation Forest

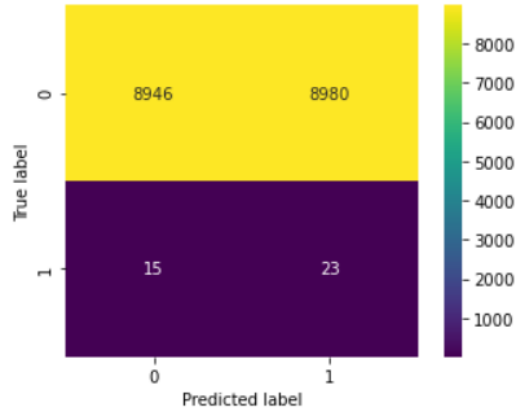

Fig. 6: Confusion Matrix of One-Class Support Vector Machine

Based on the confusion matrix shown, it can be observed that One-Class Support Vector Machine has the highest true positive count, while Isolation Forest has the highest true negative and false negative count.

5.3 Cut-Off Observation Period Analysis

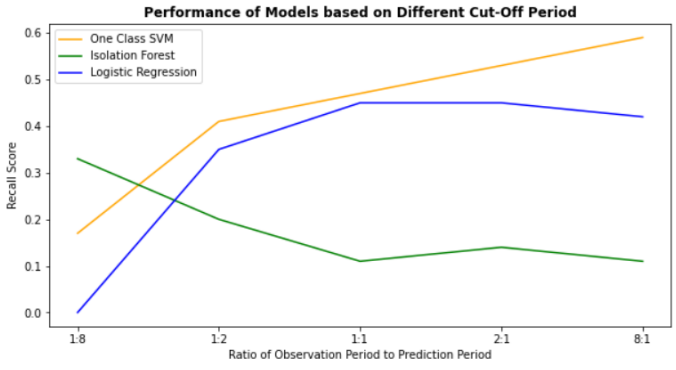

Fig. 7: Model Recall Score on Different Observation Period Lengths

The x-axis shows different ratios of observation period to prediction period. For example, the ratio of 1:8 represents 3 months to 23 months; 1:2 represents 8 months to 18 months; 1:1 represents 13 months to 13 months, and vice versa. The results show that different dataset size have an effect on the algorithms. Isolation Forest is able to perform the best when there is less data and a longer prediction period is needed. On the other hand, weighted Logistic Regression and one-class Support Vector Machine would be more suitable algorithms when the prediction period is shorer and there is a larger dataset size.

5.4 Characteristics of Customers Based on Recency, Frequency, Monetary

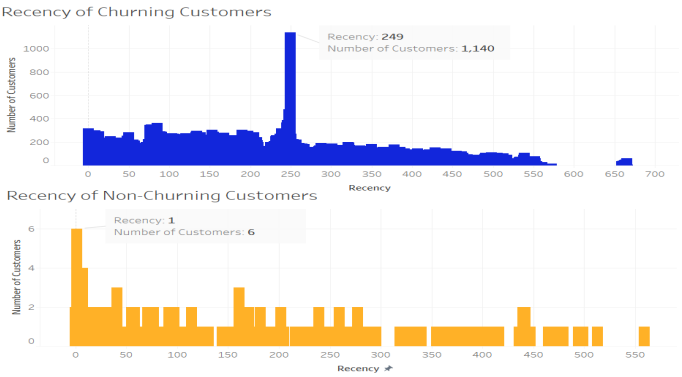

Fig. 8: Recency Value of Customers

Churning customers have higher recency than non-churning customers. There are 1,140 churning customers with a recency value of 249 and 6 non-churning customers with a recency value of 1. In other words, 1140 customers made their last purchase 249 days ago from the last day of the observation period. The high recency value indicates that the customer is inactive and has churned to other competitors.

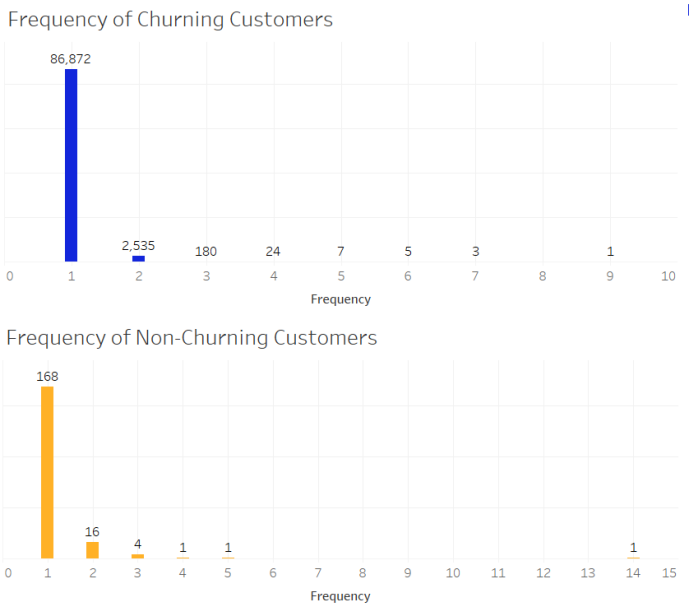

Fig. 9: Frequency Value of Customers

Both churning and non-churning customers have low frequency. 86,872 churning customers and 168 non-churning

customers have a frequency value of 1. Therefore, it can deduced that majority of the customers are one-time buyers.

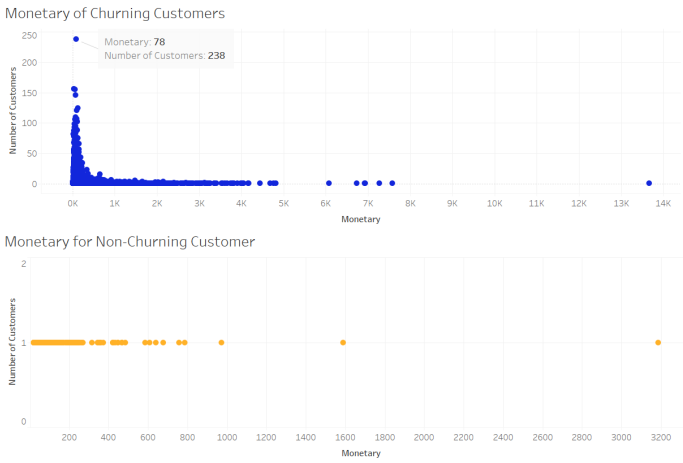

Fig. 10: Monetary Value of Customers

Churning customers have lower monetary value than non-churning customers. As shown in Figure 10, 238 churning customers have spent a total of 78 Brazilian real (BRL) throughout the observation period of 23 months, while the highest spender among non-churning customers contributed approximately 3200 BRL.

5.5 Data Product

The product developed is a multi-page dashboard that has the following features and it can be accessed via https://cidar.herokuapp.com/.

- Sign in Prompt: This feature provides security for the client as confidential customer data is analysed and presented in the dashboard. The default sign in credentials are *admin* as the username and *root* as the password (Figure 11).
- Overview: The sales performance of Olist Store is summarized in this page and yearly comparison can be done by choosing the year in the drop-down provided (Figure 12 - Figure 13). A brief comparison of the top earning sellers in terms of revenue and sales are also displayed (Figure 14).
- Customer Base: Customer preferences such as the preferred payment method and overall satisfaction rate are analysed as shown in Figure 15. Businesses can attract their customers better by devising promotional items based on their customers' preferences. A list of the customers based on specific attributes can also be obtained in a CSV file format by exporting the filtered table (Figure 16).
- Model Performance: A description of the features investigated is shown in Figure 17 to aid in future works for the improvement of this product development. The results obtained such as the recall score, balanced accuracy score, and confusion matrix of each algorithm and the graph that summarizes the effect of the dataset size on the model is also presented (Figure 18). Lastly, the characteristics of the customers are visualized from three perspectives that represent their purchasing behaviour (Figure 19).

6 Discussion

In the Variable Selection section, it was concluded that only *RFM* variables would be used to train the model even though Weighted Logistic Regression and Isolation Forest performed better with all the variables included. As shown in Table I, there is a significant increase of 0.13 in recall score for Weighted Logistic Regression algorithm in this scenario. Conversely, the additional variables act as noise for One-Class Support Vector Machine, thus resulting in a decrease of model performance. Future works can be done on exploring the effects of these variables in detail to improve One-Class Support Vector Machine by the same margin.

Furthermore, the usage of anomaly detection algorithms is proven effective in this project. The outcomes presented in Table II shows that One-Class Support Vector Machine has the best model performance. The anomaly detection algorithms are also useful in extremes scenarios such as small amount of data (3 months), but long prediction period (23 months), and vice versa. This is proven by the performance of Isolation Forest with a recall score of 0.33 and One-Class Support Vector with a recall score of 0.59 Machine in those scenarios respectively. However, if the required observation period and prediction period is equal, Weighted Logistic Regression can be considered over One-Class Support Vector Machine. This is because Weighted Logistic Regression offers a lower run time complexity of $O(nd)$ as mentioned by Banerjee (2020) as compared to non-linear Support Vector Machines that have a general run time complexity of $O(n^2)$ to $O(n^3)$ (Bottou and Lin, 2007; List and Simon, 2009).

7 Conclusion

In a nutshell, customer churn prediction is imperative for a business to better understand their customers and RFM analysis characterizes customers' purchasing behaviour well. The best performing algorithm is One-Class Support Vector Machine with a recall score of 0.59 and a balanced accuracy of 0.54. Furthermore, the effects of different dataset size on the model performance was explored and Isolation Forest performed significantly better when there is less data (3 months) and the objective is to predict customer churn in the next 23 months. Lastly, the purchasing behaviour of churning customers can be concluded as those with high recency, low frequency, and low monetary value. Regarding future works of this project, more transactional data can be collected to reduce data imbalance and different algorithms such as Artificial Neural Networks and Naive Bayes can be investigated.

Acknowledgment

I am immensely grateful to all of those whom have helped me in completing this project especially my supervisor, Dr. Hoo Wai Lam. He has provided me with extensive personal and professional guidance in this project and taught me

a great deal in general. Hence, I would sincerely like to thank him for all of his time and support for helping me accomplished my goal.

REFERENCES

Albers, S. (2000). Impact of types of functional relationships, decisions, and solutions on the applicability of marketing models. *International Journal of Research in Marketing*, 17(2-3):169–175.

Ali, Ö. G. and Arıtürk, U. (2014). Dynamic churn prediction framework with more effective use of rare event data: The case of private banking. *Expert Systems with Applications*, 41(17):7889–7903.

Bahari, T. F. and Elayidom, M. S. (2015). An efficient crm-data mining framework for the prediction of customer behaviour. *Procedia computer science*, 46:725–731.

Banerjee, W. (2020). Train/test complexity and space complexity of logistic regression.

BigCommerce (2021). Ecommerce 101 + history of online shopping (2021).

Bottou, L. and Lin, C.-J. (2007). Support vector machine solvers. *Large scale kernel machines*, 3(1):301–320.

Brownlee, J. (2020a). Cost-sensitive logistic regression for imbalanced classification.

Brownlee, J. (2020b). Smote for imbalanced classification with python.

Chen, D., Sain, S. L., and Guo, K. (2012). Data mining for the online retail industry: A case study of rfm model-based customer segmentation using data mining. *Journal of Database Marketing & Customer Strategy Management*, 19(3):197–208.

Chiang, D.-A., Wang, Y.-F., Lee, S.-L., and Lin, C.-J. (2003). Goal-oriented sequential pattern for network banking churn analysis. *Expert systems with applications*, 25(3):293–302.

Dogan, O., Ayçin, E., and Bulut, Z. (2018). Customer segmentation by using rfm model and clustering methods: a case study in retail industry. *International Journal of Contemporary Economics and Administrative Sciences*, 8.

Dursun, A. and Caber, M. (2016). Using data mining techniques for profiling profitable hotel customers: An application of rfm analysis. *Tourism management perspectives*, 18:153–160.

Ertekin, S., Huang, J., Bottou, L., and Giles, L. (2007). Learning on the border: active learning in imbalanced data classification. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 127–136.

He, X. and Li, C. (2016). The research and application of customer segmentation on e-commerce websites. In *2016 6th International Conference on Digital Home (ICDH)*, pages 203–208. IEEE.

Huang, B., Kechadi, M. T., and Buckley, B. (2012). Customer churn prediction in telecommunications. *Expert Systems with Applications*, 39(1):1414–1425.

Hung, S.-Y., Yen, D. C., and Wang, H.-Y. (2006). Applying data mining to telecom churn management. *Expert Systems with Applications*, 31(3):515–524.

Ivanovic, S., Mikinac, K., and Perman, L. (2011). Crm development in hospitality companies for the purpose of increasing the competitiveness in the tourist market. *UTMS Journal of Economics*, 2(1):59–68.

Jahromi, A. T., Stakhovych, S., and Ewing, M. (2014). Managing b2b customer churn, retention and profitability. *Industrial Marketing Management*, 43(7):1258–1268.

Li, K.-L., Huang, H., Tian, S. F., and Xu, W. (2003). Improving one-class svm for anomaly detection. *Proceedings of the 2003 International Conference on Machine Learning and Cybernetics (IEEE Cat. No.03EX693)*, 5:3077–3081 Vol.5.

Ling, R. and Yen, D. C. (2001). Customer relationship management: An analysis framework and implementation strategies. *Journal of computer information systems*, 41(3):82–97.

List, N. and Simon, H. U. (2009). Svm-optimization and steepest-descent line search. In *Proceedings of the 22nd Annual Conference on Computational Learning Theory*. Citeseer.

Liu, F. T., Ting, K. M., and Zhou, Z.-H. (2008). Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422.

Marcus, C. (1998). A practical yet meaningful approach to customer segmentation. *Journal of consumer marketing*.

Monalisa, S. and Kurnia, F. (2019). Analysis of dbscan and k-means algorithm for evaluating outlier on rfm model of customer behaviour. *Telkomnika*, 17(1):110–117.

Ngai, E. W., Xiu, L., and Chau, D. C. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert systems with applications*, 36(2):2592–2602.

Osisanwo, F., Akinsola, J., Awodele, O., Hinmikaiye, J., Olakanmi, O., and Akinjobi, J. (2017). Supervised machine learning algorithms: classification and comparison. *International Journal of Computer Trends and Technology (IJCTT)*, 48(3):128–138.

Panuš, J., Jonášová, H., Kantorová, K., Doležalová, M., and Horáčková, K. (2016). Customer segmentation utilization for differentiated approach. In *2016 International Conference on Information and Digital Technologies (IDT)*, pages 227–233. IEEE.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Reichheld, F. F. (1993). Loyalty-based management. *Harvard business review*, 71(2):64–73.

Reichheld, F. F. and Sasser, W. E. (1990). Zero defeofions: Quoliiy comes to services. *Harvard business review*, 68(5):105–111.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., and Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471.

Sun, Y., Wong, A. K., and Kamel, M. S. (2009). Classification of imbalanced data: A review. *International journal of pattern recognition and artificial intelligence*, 23(04):687–719.

Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., and Chatzisavvas, K. C. (2015). A comparison of machine learning techniques for customer churn prediction. *Simulation Modelling Practice and Theory*, 55:1–9.

Van den Poel, D. and Lariviere, B. (2004). Customer attrition

analysis for financial services using proportional hazard models. *European journal of operational research*, 157(1):196–217.

Wei, J.-T., Lin, S.-Y., and Wu, H.-H. (2010). A review of the application of rfm model. *African Journal of Business Management*, 4(19):4199–4206.

Zhao, Y., Li, B., Li, X., Liu, W., and Ren, S. (2005). Customer churn prediction using improved one-class support vector machine. In *International conference on advanced data mining and applications*, pages 300–306. Springer.

APPENDIX A
DATA PRODUCT

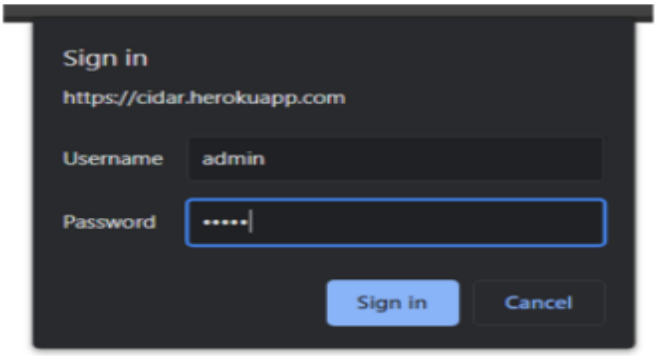

Fig. 11: Sign in Prompt

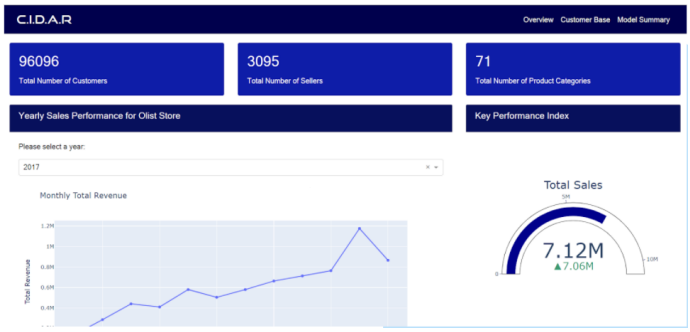

Fig. 12: Overview Page (Sales Performance)

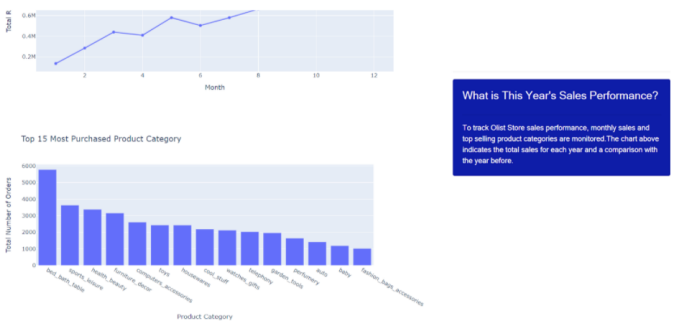

Fig. 13: Overview Page (Sales Performance)

Fig. 14: Overview Page (Seller Information)

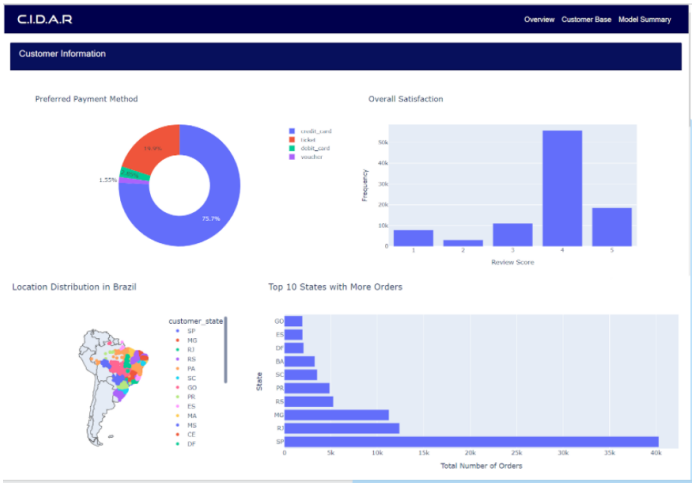

Fig. 15: Customer Base Page (Customer Preference)

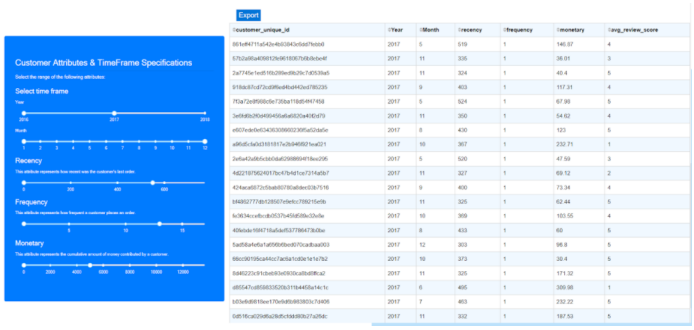

Fig. 16: Customer Base Page (Customer Details)

Fig. 17: Model Performance Page (Feature Descriptions)

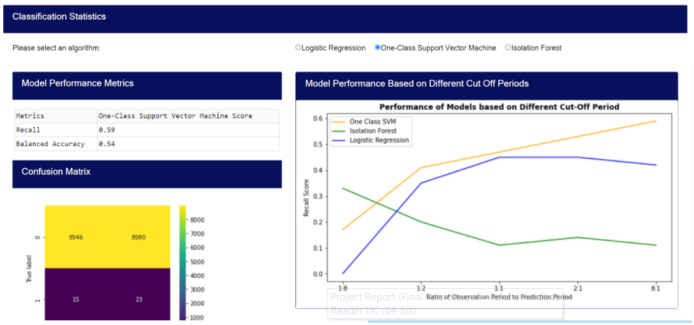

Fig. 18: Model Performance Page (Classification Statistics)

Fig. 19: Model Performance Page (Customers' Characteristics)