A recommendation system to avoid customer churn

Cho Zin Tun
National University of Singapore

Abstract

Long-term relationship with customers gains its importance and keeping customers has become major concern for modern enterprises. Nowadays, data mining technology is commonly used to analyse large quantities of customers behaviour data for better customer relationship management. This paper aims to propose the recommendation system for web-application using system logs to understand and avoid the customer churn. Various machine learning techniques are utilized to build the classification model for the recommendation system. Tree algorithm performs the best in terms of accuracy, log-loss, computation time and interpretation. The outcomes of the tree algorithm are found to be beneficial for suggesting marketing strategies as well as making strategic recommendation to avoid customer churn.

1. Introduction

The Visa Smart Debit/Credit (VSDC) service is Visa's chip-based solution that allows participants to combine the functionality of debit, credit and prepaid products with the flexibility of chip technology. Visa uses specifications developed by EMVCo, an association of Visa, MasterCard, JCB, American Express, Discover and UnionPay, as the foundation for VSDC. EMV Specifications provide a global foundation for chip-based payment services. VSDC provides a level of interoperability for cards and devices that support the same EMV options and offers enhanced security.

For card issuance, chip introduces many new features and complexities to the card personalization process. With the chip card having the ability to respond on behalf of the issuer at the point of transaction (POT), additional security and functional data elements need to be placed into the payment application of the card. In addition to the encoding and embossing data, there may be a significant number of chip data elements that need to be incorporated into the card personalization process.

Because VSDC offers such a large and complex set of options to choose from, Visa has developed the *Visa Personalization Assistant* (VPA) to assist issuers (banks) in choosing options applicable to their program. It benefits users by providing a user-friendly interface, including best practices and detailed help screens. Moreover, it allows better profile management for Visa and the issuer as well as helps to generate output that can be used during two key card personalization processes: data preparation and personalization validation.

This project emphasizes on how to classify different users into groups, based on recency, frequency and monetary (RFM) analysis on VPA application, by using machine-learning techniques. To realize whether the insights provided are substantiated with reasonable results, multiple models are valuated. In addition, possible marketing strategies are discussed to increase the revenue of the application by using the final model's outcome.

2. Problem Statement

The main problem to address in this report is to increase the revenue of VPA application, by maintaining the current loyal customers and winning back the lost customers. Currently, the application charges a flat fee for accessing the software annually. This is common pricing strategies for most of the software companies. Nevertheless, this does not offer much flexibility to the users. Regardless of number of usage, a client has to pay a flat fee. Thus, clients with low annual usage unsubscribe the application, resulting in the declination of the revenue.

Knowledge on the users' behaviors could provide the prediction on which types of users are likely to churn. With the help of the insight, sale team could perform some actions to stop the users from unsubscribing the application. Thus, in this project, data from the system log files are utilized to classify the users into different groups and various classification are explored, modelled, and evaluated.

3. State of the Art

3.1 Literature Review

Machine learning techniques have been gaining popularity and are widely used in different domains ranging from health care sectors to customer management (Subasi & Gursoy, 2010) (Chalmeta, 2006). In customer relationship management (CRM) sectors, many researches have been done to address the customers' needs, segment customers based on their behaviours as well as increase the profit of the company with the help of machine learning techniques. (Tsai & Chiu, 2004) developed market segmentation methodology based on product specific variables using purchase-based similarity measure, clustering algorithm and clustering quality function. (Kim, Jung, Suh, & Hwang, 2006) implemented the decision tree to analyse the customer life-time value and segment customers based on their values.

There are researches on segmentation of customers by integrating Recency, Frequency, and Monetary (RFM) analysis with machine learning techniques. (Christy, Umamakeswari, Priyatharsini, & Neyaa, 2018) performed RFM analysis on transactional data and clustered the customers based accordingly using K-means and Fuzzy C-means algorithm. (Chen, Chiu, & Chang, 2005) proposed a method for mining changes in customer behaviour and performing segmentation ,by integrating RFM scores extracted from transactional

database with customers' demographic variables, to assist managers in developing better marketing strategies.

In the digital age, web applications have become the source of data collection for companies and researchers are using those data to understand the customers, predict the customers' behaviours as well as cluster the customers into different segments. (Wang, Chiang, Hsu, Lin, & Lin, 2009) used decision tree to perform customer churn analysis to build the recommendation system for wireless network company. (Yuan & Chang, 2001) presented mixed-initiative synthesized learning approach which consists of various steps such as segmenting data sources into different clusters, and labelling the features of the clusters to gain better understanding on customers for web-based CRM. They used a combination of hierarchical automatic labelling, decision tree, and human tacit experience to implement the approach.

3.2 Literature Review on Classification Algorithm

3.2.1 Decision Tree

Decision Tree (DT) is non-parametric supervised learning method used for both regression as well as classification. With a set of if-else decision rules, decision tree learns from data to approximate a sine curve. It breaks down the data into smaller subsets by choosing one feature that can provide best split. Best split is judged either by entropy reduced or information gained. For each branch of the trees, the next feature is iteratively selected to best split the subsets until the stopping condition is met. The deeper the tree is, the more complex the decision rules are and the fitter the model is. However, in order to avoid overfitting, pruning is usually performed.

3.2.2 Random Forest

Random Forest (RF) is one of the ensemble-based methods in machine learning technique. This method was introduced by (Breiman, 2001) and it combines the basic principles of bagging with random feature selections to add additional robustness to the decision tree models. This alleviates the overfitting and pruning problem of tree. After the ensemble of trees (the forest) is generated, the model uses a vote to combine the trees' predictions. The key idea is to use only a small, random portion of the full feature set to build the trees on the train , instead of creating serval data sets, by bootstrapping. This lets random forest to handle extremely large number of attributes ("curse of dimensionality"), which could cause other models to fail. Figure 1 represents the architecture of random forest.

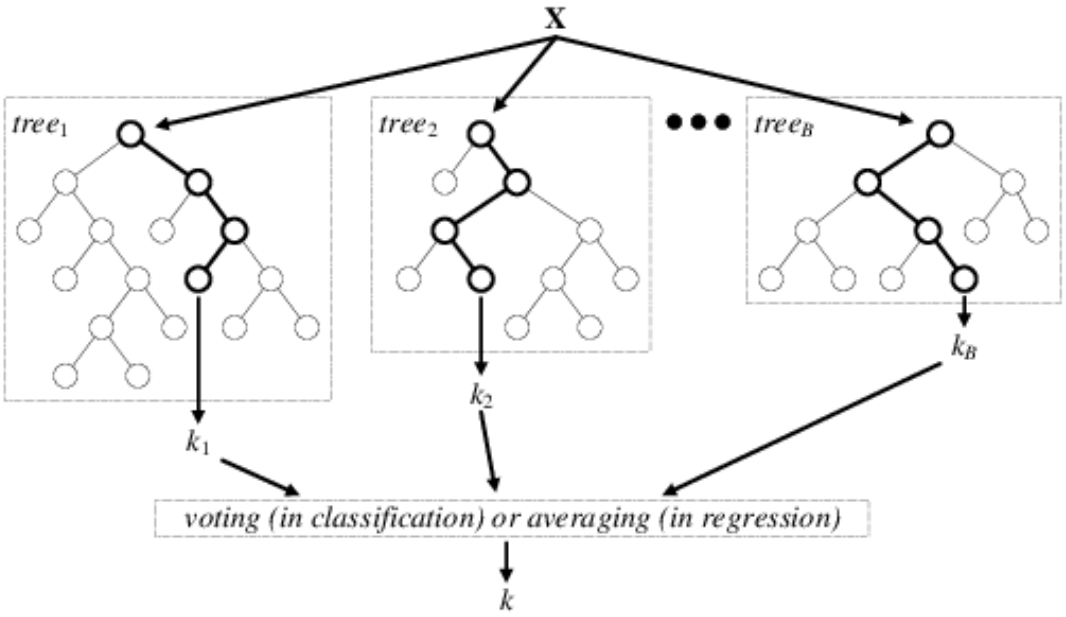

Figure 1- Typical Architecture of Random Forest (Antanas Verikas, 2016)

3.2.3 Support Vector Machine (SVM)

Support vector machine (SVM) proposed by (Vapnik, 1999) is based on the Structural Risk Minimization principle (Stitson, Weston, Gammerman, Vovk, & Vapnik, 1996). The goal of the Structural Risk Minimization is to find the hypothesis for which the lowest error is guaranteed on random test data. SVM uses geometric properties to achieve the optimal hyperplane keeping the training data as far as possible. SVM searches for the maximum marginal hyperplane (MMH) ,which provides the largest margin (distance) between classes, from support vectors (SVs). The SVs, which lie closest to the decision boundary, are critical training examples as they are representatives from each class and carry all the relevant information about the classification problem.

Key feature of SVMs is its ability to map the problem into a higher dimension space using kernel trick, in addition to slack variable which allows some examples to be misclassified. Figure 2 demonstrates how kernel trick transforms non-linear data into linearly separable dimensions.

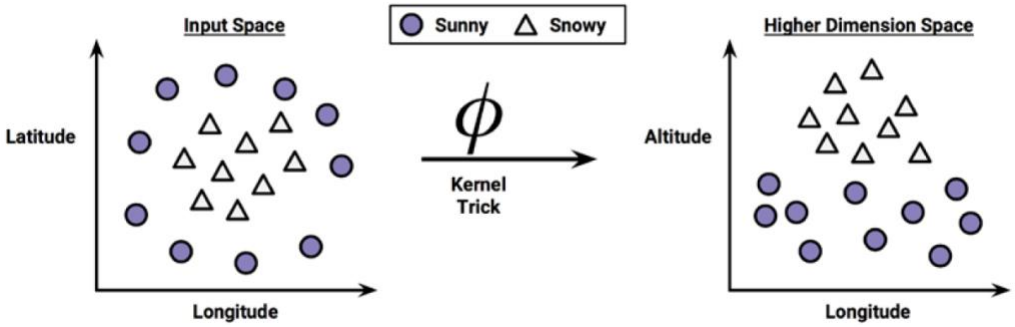

Figure 2- Support Vector Machine with Kernel Trick (Lantz, 2015)

3.2.4 Artificial Neural Network

Fundamental principle operation of the human neural system is the primary evolution of artificial neural network (ANN). Human neurons receive (input) signals through synapses located on the dendrites of the neuron and consequently, if they are high enough to surpass a certain threshold, the neuron will be activated and a (output) signal will be transmitted through the axon. This signal might be then sent to another synapse and might activate neuron. Figure 3 represents how am artificial neuron behaves like human biological neuron (Gershenson, 2003).

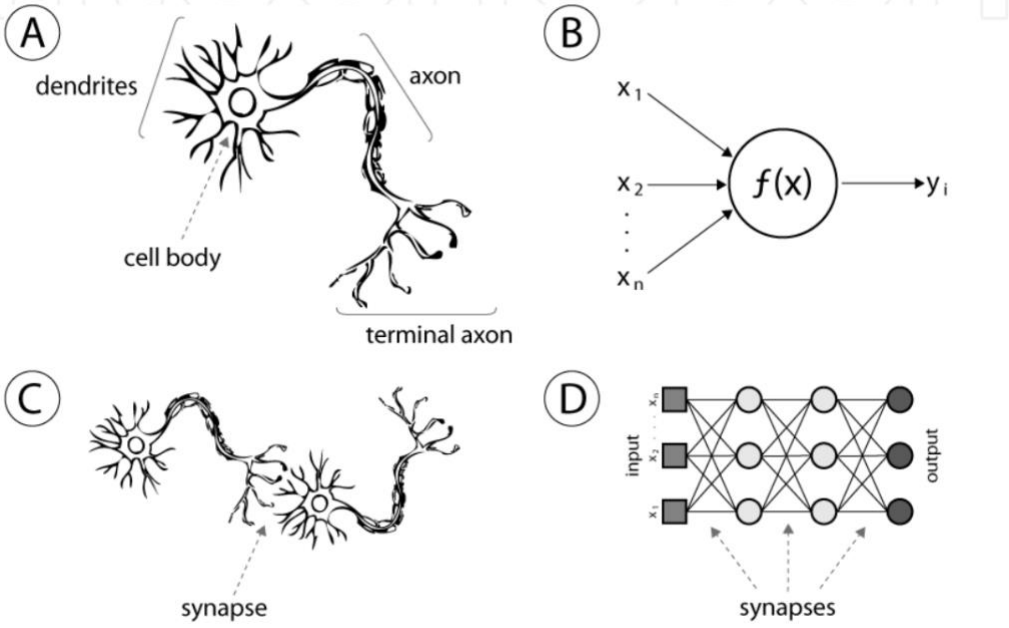

Figure 3- (A) Human Neuron; (B) artificial neuron or hidden unity; (C) biological synapse; (D) ANN synapses (Maltarollo, Honório, & Silva, 2013)

ANN depends on the stimulus-response activation functions that takes in inputs and obtains some outputs. Various ANN differs from each other based on the nature each neuron's activation function. Different sorts of

activation(threshold) functions are used in ANN in regard to the applications in which they are applied. Some examples of activation functions are Linear, sigmoid and gaussian. In addition, number of hidden layers can be chosen to get the desired system.

Figure 4 covers some types of ANN which vary based on connection pattern, the nature of activation functions, learning algorithm and number of hidden units.

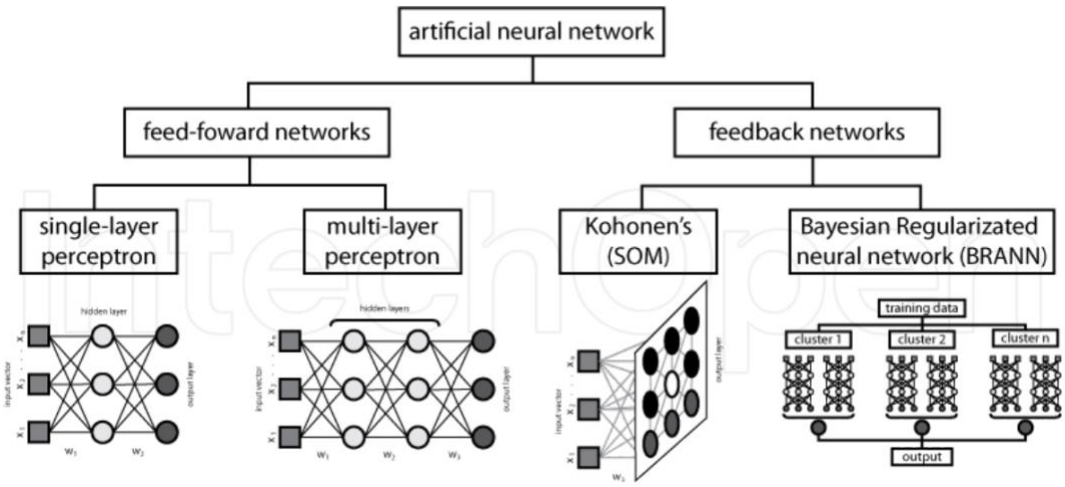

Figure 4- Some Types of Artificial Neural Network (Maltarollo, Honório, & Silva, 2013)

3.2.4 Multilayer Perceptron

Multilayer perceptron (MLP) is one of the feed-forward neural networks, where data information flows only in the forward direction. It is modified version of single-layer neural network and has several neurons arranged in different layers. It has at least three layers: an input layer, a hidden layer and an output layer.

4. Solution: Description of Approach

4.1 Database

Visa Online (VOL) is a centralized platform for Visa services and VPA application is one of them. Users have to enroll into VOL to access its services. VOL database stores all the information about different organizations, including which services they are subscribed to on which day. In this project, VOL database was used together with VPA database to extract features.

Features are extracted from multiple tables of these two databases by using different joins, grouping and aggregation methods with the help of Structured Query Language (SQL). The sample data in this study covers the application

usage from fiscal year 2018 to first quarter of 2019. 2018 data is used to train and validate models whereas the rest is used to evaluate the performance of those models. Table 1 shows the features and their respective description.

Features	Description
Orgid	Identification Number for Organization
isBilled	Indicator to check if organization has been billed
inVOL	Indicator to check if organization is in VOL
Recency	Last date that organization accessed VPA
isVE	Indicator to check if organization is from Visa Europe Region
Frequency	Number of time that organization accessed VPA
SinceYear	Number of years that organization subscribed to VPA
Asset	Number of confidential information that organization has in VPA

Table 1- Features with description

4.2 Data Processing

Data processing is done as per below steps:

1. Import data set
   Data retrieved from SQL is stored in the excel files and they are imported into python for further modelling and analysis.

2. Check out the missing data
   Missing data problem is common in practical applications and could have significant effect on the analysis. Thus, missing data is checked and handled accordingly, either by dropping or replacing with mean of the feature.

3. Check categorical values
   Since machine learning models are based on mathematical expressions, feeding categorial values to them could lead to problem. Therefore, categorical values are converted into numerical data.

4. Label data
   As the focus of this project is to forecast which clients are likely to churn or stay, labelled are assigned to train dataset before building the supervised learning model. The explanation for labelling data can be found in section 4.2.1.

5. Check data imbalance
   Imbalance data problem is as well common in practice and this could lead to build the biased models. Thus, checking on data imbalance is performed and section 4.2.2 explains how it is mitigated with the help of data re-sampling.

6. Split data into train and test data
   In order to make sure that the models built do not overfit and they

perform well in general for all sort of incoming data, data is split into train and test datasets with 70:30 ratio.

7. Feed into machine learning models
   Train dataset is used to train the models with 10-fold cross validation while test dataset is used to evaluate the performance.

4.2.1 Labelling data

Once features are ready, each tuple from training data is labelled into five corresponding different groups with the help of unsupervised clustering algorithm as well as business domain knowledge. Users from Europe Regions are grouped into "Group VE" (Class 2), while users who had paid for annual fee are assigned to "Group Billed" (Class 3) and users, who are no longer in VOL, are allocated to "Group NVOL" (Class 4).

K Means clustering algorithm is executed iteratively with different clusters to determine the number of groups for the rest of the training data. Features are normalized before executing the algorithm to make sure that values are adjusted to common scale. Based on Figure 5, two clusters are chosen and labelled as "Group Churn" (Class 1) and "Group Retain" (Class 0). Figure 6 displays the number of users belong to each group.

Each group has different values for marketing strategy as well as preventing customer churn. For example, currently, "Group VE" users are not charged but pricing strategy is planned to change in future. For "Group Churn" users, by discovering their usage patterns, it can be predicted if they are going to churn or not.

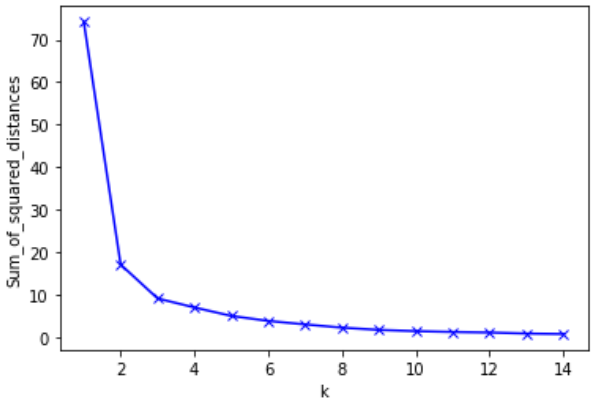

Figure 5 - Elbow Method For Optimal k

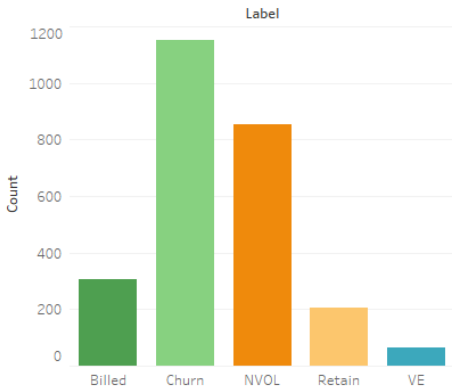

Figure 6 - Labels with corresponding counts

4.2.2 Re-sampling data

From Figure 6, it can be seen that there is imbalance in dataset. The class imbalance problem is common in practice and occurs when there are many more instances of classes than others. Generally, classification algorithms build the model by minimizing the overall error. Thus, under imbalance situation, the classifiers tend to be biased towards the large classes, and ignore the small classes while achieving maximum accuracy. In such cases, even though the model accuracy is high, it tells little information about the minority classes. (Ali, Shamsuddin, & Ralescu, 2015)

To mitigate the class imbalance problem, there are traditional ways of sampling data: over-sampling, duplication of minority class instances, and under-sampling, removing of majority class instances. The former method could lead to over-fitting problem whereas the latter could discard the useful information. In this paper, Synthetic Minority Over-sampling Technique (SMOTE) ,introduced by (Chawla, Bowyer, Hall, & Kegelmeyer, 2002), is used to resample the data. It is over-sampling method as well but instead of duplicating samples of minority classes, it creates synthetic samples by selecting similar records and altering that record on column at a time by a random amount within the difference to the neighbours. After resampling data, the data is split into train and test data with 70/30 ratio.
Figure 7 and Figure 8 illustrate the number of tuples belong to each class.

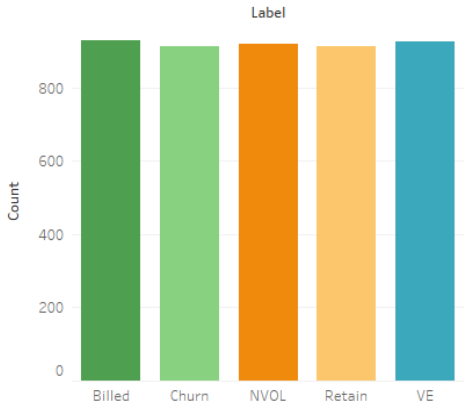

Figure 7 - Labels with corresponding counts after resample (Train data)

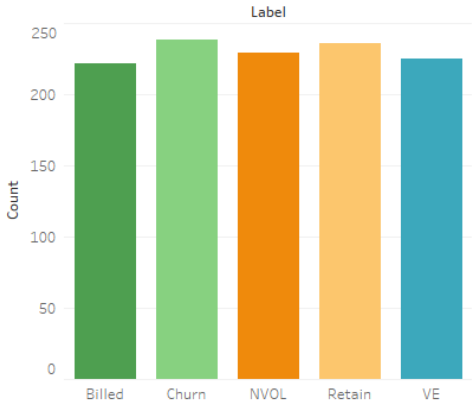

Figure 8 - Labels with corresponding counts after resample (Test data)

4.3 Model Building

Various machine learning models are built and evaluated to determine the group assignment for each user. Models include Random Forest (RF), Decision Tree (DT), K-Nearest Neighbours (KNN), Support Vector Machine (SVM), Artificial Neural Network (ANN) and Multilayer Perceptron Neural Network (MLP) .

4.3.1 Hyperparameter Tuning

Hyperparameters play an important role ,while building model, as they define the model architecture and have a considerable impact on the performance of

the trained model. Unlike model parameters ,which are learnt on their own during the process of model building, hyperparameters need to be initialized before building the model. Hyperparameter tuning step is essential while building the machine learning models since the right choice of hyperparameters value could produce a better model.

In this paper, grid-search is applied to find the best hyperparameter for different models ,out of common techniques like random-search and Bayesian optimization algorithm. Grid-search technique builds the model for each possible combinations of all the hyperparameters provided, evaluates each model by performing cross-validation, and selects the architecture which produces the best result. Table 2 below describes explanation of hyperparameters as well as their best values for each model, and how they are obtained.

Model	Best Parameters	How	Explanation
DT	max_features = sqrt(n_features) min_samples_split = 100 min_samples_leaf =2	Prune the tree until the criteria for min_samples_split and min_sample_leaf are met.	*Min_samples_split*: the minimum number of samples required to split the internal node *min_sample_leaf*: the minimum number of samples required to be at a leaf node; Both contribute to the complexity and size of the trees.
RF	max_features= sqrt(n_features) number of trees= 600	Loop through various number of trees with different options of max_features [“sqrt”,”log2”]	*max_features*: the number of features to consider when looking for best split (for classification, sqrt(n_features) is recommended)
SVM	Kernel =Radial Basis Function (RBF) C = 34 Gamma = 0.1	Iterate through C=[1..100] with gamma = [0.1, 0.01, 0.001, 0.0001, 0.00001] with 10-fold cross validation	*C*:penalty on soft margin; larger C, smaller training error,more chance to overfit *Gamma*: free parameter of the Gaussian RBF; large gamma leads to high bias and low variance models
ANN	One hidden layer with 25 neurons, activation function= sigmoid	Loop through neurons=[1..30]	Neurons: number of neuron present in layer. More neurons capture more complicated relationship
MLP	Batch_size = 32	Iterate through	Batch-size: number of samples

	Epochs = 100 Two hidden layers with 6 and 7 neurons, Output activation function = softmax	different batch_size and epochs with 10-fold cross validation	that will propagate through the network (higher batchsize requires more memory) Epoch: number of times updating weights (less epoch could lead to underfitting while higher epoch could lead to overfitting)
KNN	n_neighbours = 7	Loop through n_neighbours=[1…100]	n_neighbours: number of neighbours to determine the class

Table 2- Hyperparameter tuning

4.4 Performance Indices

Area Under ROC Curve - Receiver Operating Characteristic (AUC - ROC), accuracy, precision-recall and log-loss are commonly used as performance indices to evaluate the classification models. Significance of precision-recall is profound when the dataset is highly imbalanced (Saito T, 2015). Since the imbalance problem is being handled in this project, precision-recall index is omitted.

4.4.1 AUC – ROC curve

AUC – ROC curve measures the performance of the classification problem at various thresholds settings. ROC is a probability curve and it is plotted with true positive rate (TPR) against false positive rate (FPR). TPR is also known as sensitivity and is a measure of how many positive samples are being identified as positive (equation 1). (BRADLEY, 1996)

$$Sensitivity = \frac{TP}{TN} \qquad (1)$$

$$FPR = \frac{FP}{N} \qquad (2)$$

where TP = true positive, TN = true negative, FP = false positive, N =total.

AUC measures the two-dimensional area underneath the entire ROC curve. It provides an aggregate measure of performance across all possible classification thresholds. Higher the AUC, the better the model is and it can provide good measure of separability. However, when AUC is 0.5, it means the model has no capability to separate the classes.

As AUC – ROC curve is for binary classification and the data has more than two classes, sensitivity and FPR for each class are calculated against other classes.

4.4.2 Accuracy

Accuracy measures how often a classifier is correct and its calculation as shown in equation 3. Average accuracy across all the classes is used to compare the performance of different machine learning algorithms.

$$Accuracy = \frac{TP+TN}{N} \qquad (3)$$

4.4.3 Logarithmic loss

Logarithmic loss (logloss) measures the uncertainty of the probabilities of model by comparing them with the true labels. The larger the logloss, the mover diverted the predicted probability is from the actual label.

$$logloss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log(p_{ij}) \qquad (4)$$

Where N = number of rows in dataset, M = number of fault devliery classes, y_{ij} = 1 if observation belongs to class j; else 0 , p_{ij} = predicted probability that observation belongs to class j.

4.5 Result Interpretation

Various machine learning models are built and evaluated with test data. The result of models for train and test data in terms of accuracy and log-loss function can be found in Table 3. Figure 9 illustrates area under curve for different models.

Model	**Train**		**Test**	
	Accuracy	*Log-loss*	*Accuracy*	*Log-loss*
Decision Tree	0.9898	0.3109	0.9901	**0.2831**
Random Forest	**0.9903**	**0.3019**	**0.9907**	0.3013
SLP	0.9858	0.4891	0.9826	0.6007
SVM	0.9891	0.3776	0.9890	0.3804
MLP	0.9893	0.3690	0.9896	0.3604
KNN	0.9900	0.3432	0.9900	0.3203

Table 3- Result of different models for train and test data

For training data, random forest outperforms the other models by producing the highest accuracy and lowest log-loss. The difference in measurement metrics between models are trivial. K-nearest neighbours algorithm and decision tree provides slightly lesser accuracy and higher log-loss value than random forest. Ensemble models obtain marginally worse accuracy and log-loss value compare to tree algorithms. Out of all the models, single layer neural network with sigmoid activation function, gives the worst outcome. Even though both single layer perceptron and multilayer perceptron belong to neural network family, the discrepancy in performance could be due to the

choice of activation function, number of hidden layers and neurons.

The results obtained between training and test data are consistent. Tree algorithms provide the best performance by producing highest accuracy and lowest log-loss. Choosing model based on accuracy and log-loss value is subjective. If the absolute probabilistic difference from the actual label is main concern, decision tree which provides the lowest log-loss can be selected as the best performing model. Otherwise, random forest with the highest test accuracy is determined as the final model. This explains why most of the researches performing customer segmentation based on RFM analysis use the decision trees.

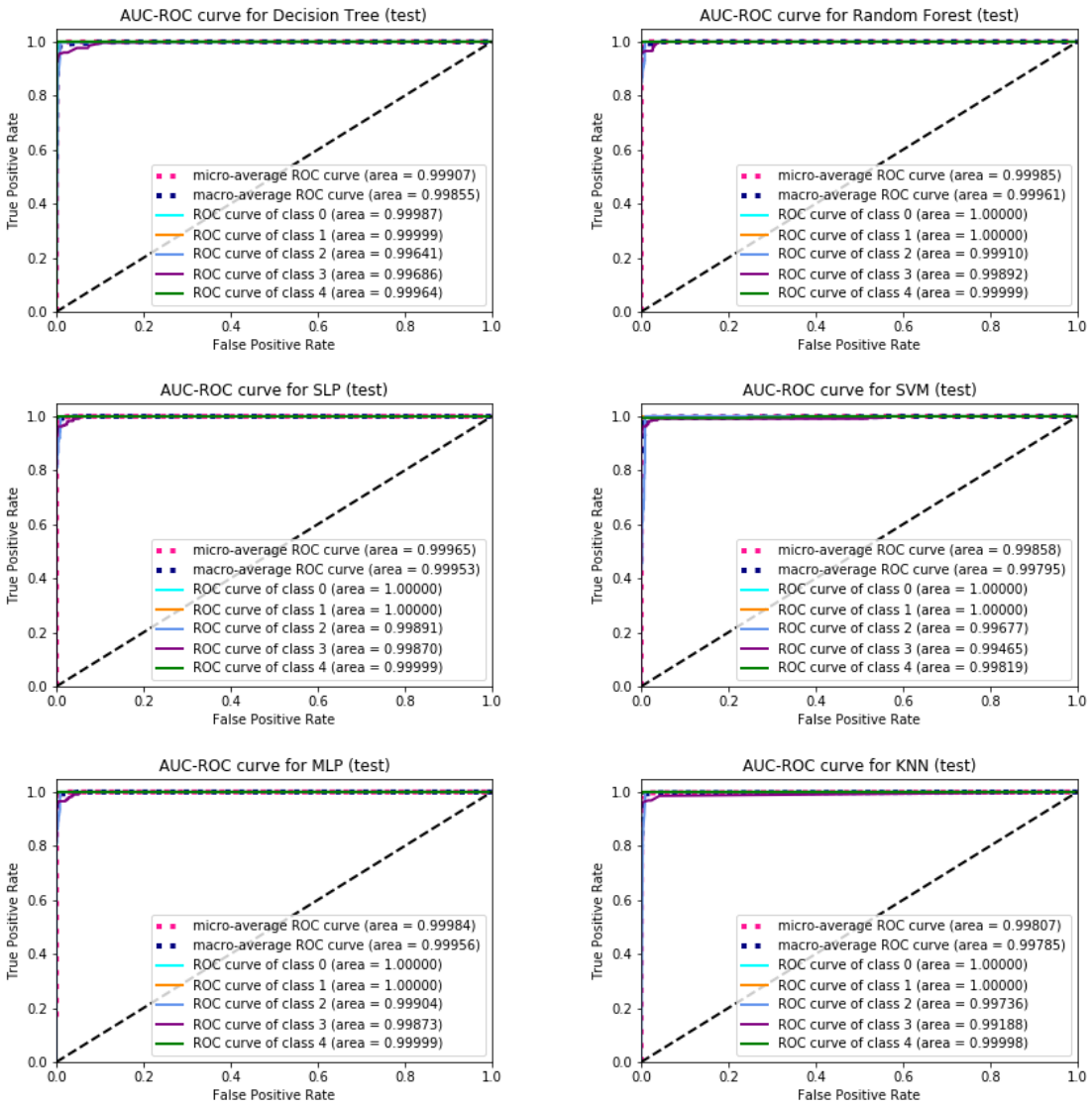

Figure 9 - ROC-AUC Curves for test data

5. Discussion

In customer segmentation based on RFM analysis problem, tree algorithms perform the best compared to neural networks, support vector machine and K nearest-neighbour models. Apart from the performance measurement metrics to evaluate the models, computation time also plays an important role. Tree models have acceptable training and hyperparameter tuning time even though they take slightly longer time than K-nearest neighbour. Support Vector machine and Neural Networks consume much more time as well as computation power. Another factor that should be taken into account is interpretation of outcome, especially for business recommendation. Ensemble models usually provide blackbox; no explanation about the result. On contrast, tree algorithms deliver the clear interpretation on how results are attained.

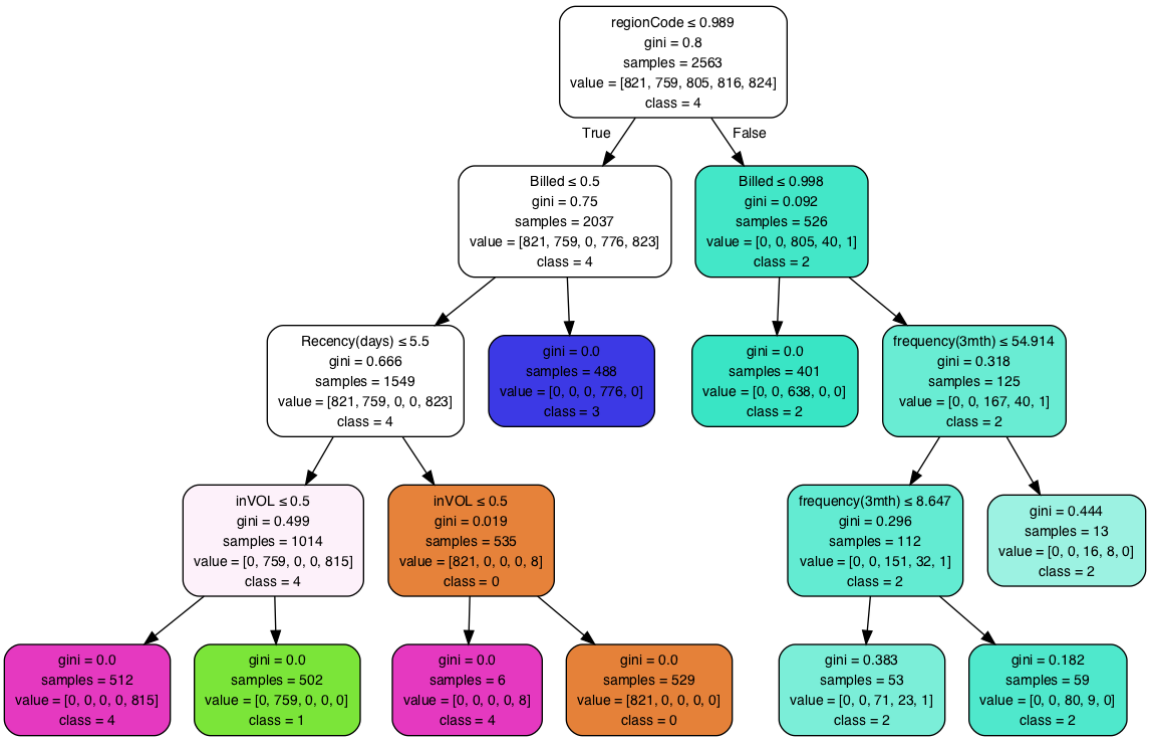

Figure 10 - Random Forest Decision Tree

Figure 10 displays how the customers are segmented based on features. For groupVE, it depends mainly on if the customer is from Europe region or not. Meanwhile, groupNVOL and groupBilled build upon customer's presence in VOL database and status of billing. If neither of the above conditions meet, RFM analysis scores are used to determine if a customer is likely to be churned. For every quarter or whenever needed, the sale or marketing team can feed the data of customers to the random forest model to perform customer segmentation. Subsequent actions can be taken to increase the revenue by the management.

5.1 Marketing Strategy

A company can plan marketing strategy, setup marketing strategy for each segment, and provide information in terms of grasping a whole picture of future marketing strategy by using the outcome of the tree algorithm. Furthermore, a company can setup targeted marketing plan for specific group of customers based on the importance.

5.1.1 Strategy to Win back Lost Customers

A customer with low frequency usage is the one who usually leaves the application. Current annual subscription plan is one of the reasons that forces customers churn. Customers felt that it is not worthy to pay for whole year while they are using application only for a few months. This means changing the subscription plan is a good way to win back lost customers. Instead of annual subscription, it is better for company to provide more variety plans such as quarterly or monthly subscription plans. It would be a win-win situation for both parties as customer with lower usage would pay less while company can earn more by regaining lost customers. Second degree price discrimination could be performed on each plan to earn more for company.

5.1.2 Strategy to Provide Better Service for Loyal Customers

As VPA customers (banks) can approach the processors (agents) to attain the same outcome, it is important to keep the current loyal customers within VPA. Losing customers could lead to revenue declination. In order to retain customers, it would be better to provide the additional unique service that the company is only capable of and could be of extra incentive for the customers. One of the strategies could be giving access to other related applications that would be useful for customers. For example, as VPA is an application for issuing cards, giving customers access to other VISA issuing related applications could be a big reason for loyal customers to stick with the company.

6. Conclusions and Further Work

This paper demonstrates the use of machine learning techniques to perform customer segmentation using application's system logs. The data is constructed from two VISA databases using various aggregation and transformation.

Various machine learning models are modelled and evaluated using accuracy as measurement metrics. The results obtained conclude that tree algorithms perform the best for segmenting customer using RFM analysis. They provide not only best accuracy but also the reasonable computation time as well as good interpretation. Ensemble models like support vector machine and neural networks seem to be a bit complicated for similar studies and unable to provide human understanding interpretation is one of the drawbacks for those

models.

Further work will be focused on performing price discrimination by imposing different pricing strategies on each group. This could be achieved by building a dynamic pricing problem, where relationship between demand and price is studied as well as revenue is optimized by tuning the price. Moreover, interactive dashboards could be built to give instant insights of different segmentations for stakeholders. Furthermore, automation of the project pipeline by integrating data extraction, and pre-processing with classification model as well as displaying outcome on dashboard could be done for future usage.

Bibliography

(Chalmeta, 2006) Ricardo Chalmeta, *Methodology for customer relationship management*, 2006

(Subasi & Gursoy, 2010) A.Subasi and M. Ismail Gursoy, *EEG signal classification using PCA, ICA, LDA and support vector machines*, 2010.

(Tsai & Chiu, 2004) Tsai, C.-Y., & Chiu, C.-C , *A purchase-based market segmentation methodology*, 2004

(Kim, Jung, Suh, & Hwang, 2006) Kim, S.-Y., Jung, T.-S., Suh, E.-H., & Hwang, H.-S,*Customer segmentation and strategy development based on customer lifetime value: A case study*, 2006

(Christy, Umamakeswari, Priyatharsini, & Neyaa, 2018) A. Joy Christy, A. Umamakeswari, L. Priyatharsini, A. Neyaa, *RFM ranking – An effective approach to customer segmentation,* 2018

(Chen, Chiu, & Chang, 2005) Chen, M.-C., Chiu, A.-L., & Chang, H.-H, *Mining changes in customer behavior in retail marketing. Expert Systems with Applications*, 28, 77–781, 2005.

(Yuan & Chang, 2001) Yuan, S.-T., & Chang, W.-L, Mixed-initiative synthesized learning approach for web-based CRM. Expert Systems with Applications, 20, 187–200, 2001.

(Wang, Chiang, Hsu, Lin, & Lin, 2009) Y.-F Wang, D.-A. Chiang, M.-H. Hsu, C.-J. Lin, I-L. Lin , *A recommender system to avoid customer churn*: A case study, 2009

(Breiman, 2001) L. Breiman, *RANDOM FORESTS*, 2001

(Antanas Verikas, 2016) Antanas Verikas, Evaldas Vaiciukynas, Adas Gelzinis, James Parker, and M. Charlotte Olsson, *Electromyographic Patterns during Golf Swing: Activation Sequence Profiling and Prediction of Shot Effectiveness,* 2016

(Vapnik, 1999) V.N.Vapnik*, An Overview of Statistical Learning Theory*, 1999

(Stitson, Weston, Gammerman, Vovk, & Vapnik, 1996) Stitson, M.O., Weston, J.A.E., Gammerman, A., Vovk, V., Vapnik, V, *Theory of support vector machines. Technical Report CSD-TR-96-17, Computational Intelligence Group, Royal Holloway, University of London,* 1996

(Lantz, 2015) B.Lantz, *Machine Learning with R*, 2015.

(Gershenson, 2003) C.Gershenson, *Artificial Neural Networks for Beginners*, 2003.

(Maltarollo, Honório, & Silva, 2013) V. G. Maltarollo, K. M. Honório and A. B. F. Silva ,*Applications of Artificial Neural Networks in Chemical Problems*, 2013.

(Ali, Shamsuddin, & Ralescu, 2015) A. Ali, S.M. Shamsuddin, and A.L. Ralescu ,*Classification with class imbalance problem: A Review* , 2015

(Chawla, Bowyer, Hall, & Kegelmeyer, 2002) Chawla, N.V., et al., *SMOTE: synthetic minority over-sampling technique. arXiv preprint arXiv:1106.1813*, 2002.

(Saito T, 2015) Saito T, Rehmsmeier M (2015) *The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets*. PLoS ONE 10(3): e0118432. https://doi.org/10.1371/journal.pone.0118432

(BRADLEY, 1996) Bradley, A.P , *The use of the area under the ROC curve in the evaluation of machine learning algorithms*, 1997