Project Report

Group 5 Charu Aggarwal

1. Mei-Chun Hung
2. Sukanya Aswini Dutta
3. Vaibhavi Gaekwad
4. Wasinee Sriapha
5. Yufei Wang

206-880-9298 (Tel of Student 1)
206-941-3762 (Tel of Student 2)
425-624-5609 (Tel of Student 3)
206-380-6328 (Tel of Student 4)
206-319-8422(Tel of Student 5)

Percentage of Effort Contributed by Student 1:_____20%_____
Percentage of Effort Contributed by Student 2:_____20%_____
Percentage of Effort Contributed by Student 3:_____20%_____
Percentage of Effort Contributed by Student 4:_____20%_____
Percentage of Effort Contributed by Student 5:_____20%_____

Signature of Student 1:____MH_____
Signature of Student 2:____SAD_____
Signature of Student 3:____VG_____
Signature of Student 4:____WOS_____
Signature of Student 5:____YW_____

Submission Date:_____04/24/20_____

Beyond Churn Analysis: Retail Customer Churn Prediction in a Leading American Wholesaler with the Retailer's Personalized Insights - an Observational Study

MC. Hung, S. Dutta, V. Gaekwad, W.Sriapha, Y.Wang
Northeastern University, IE 7275 - Data Mining in Engineering, April 2020

Problem Setting

Most American retailers spend a great amount of resources acquiring new subscriptions every year. Nevertheless, the current customers are those that earn the retailer highest revenue. As the number of purchases and store visits of the current customers increases, it is more likely for them to keep making future purchases. Another reason customer retention should be prioritized is that the customers who are loyal and invested in the retailer bring in more customers through the word-of-mouth marketing. The cost of customer churn includes but not limited to the loss in overall revenue as well as marketing budgets involved in replacing the lost customers with new sign-ups. Reducing the percentage of customer churn is undoubtedly a key business goal of every retailer as it is more challenging and more costly to recruit new customers than it is to maintain the currently registered customers.

This study sheds light on a leading American retailer known for selling bulk quantities of goods at discounted prices to its subscribers who pay an annual membership fee. The objective of the study is to statistically analyze accessible historical data in order to discover patterns of demographic characteristics and shopping behaviors that promote customer churn. The accuracy of the predictive model is critical for preventive actions. If the retailer is able to understand the reasons behind its former customers' discontinuation, the preventive plans can be implemented to avoid further loss.

Problem Definition

The study aims to predict the customers who are likely not to renew their memberships and to identify the most significant factors that might affect a customer's renewal decision. The findings will introduce the retailer suggestive hints towards handling customer renewal.

In order to achieve the above goals, the following questions were raised:

- What are the shopping behaviors that lead to customer churn?
- What are the common demographic characteristics of customers who left?
- Are repeat customers more likely to renew their memberships?
- Which region of the country have the highest numbers of churns?

Data Sources

Two data sources were accessed in this study:

Main dataset, MBRChurnModel_FirstYear_MSK, provided by a leading American wholesaler, and comprising historical membership information

Supportive dataset, free-zipcode-databse-Primary, used in the study for Zip codes and U.S. states information. Coven, D.S., (2012). Free Zipcode Database: Unique Zipcode [data file]. Retrieved from http://federalgovernmentzipcodes.us

Data Description

In the main dataset, MBRChurnModel_FirstYear_MSK, there were 22 columns and 120,450 rows in total. Each row had a customer account number as a primary key. The columns were assessed and categorized into the following groups.

Account Information:

- RENEW – the status of account renewal (renewed/churned)

- A2ACCIPK – the membership/account number
- A2ACCTYP – the account type (Individual / Business)
- M2EXCFLG – the membership type (Executive / Non-executive)
- B2BUSTYP – the business type code
- TENURE – duration of membership (0 as less than a year, 1 as 1+ years)
- MBRCOUNT – the number of cards under a single account

Other Personal Information:

- AGE – the age of the customer
- ZIPCODE – the zip code of customer's address
- DISTANCE – the distance between customer's address to the closest warehouse
- RECENTMOVING – whether the customer recently changed his/her address

Shopping Information:

- F2HOMRGN – region where customer shopped the last 3 times
- F2HOMFCY – warehouse id number (include eCommerce) where customer shopped
- EARLYFAREWELL – days since the customer last shopped
- HOMEFCTYCHANGE – whether customer shopped at home-city warehouse
- SHOP1YR/SHOP6M/SHOP3M – total amount of purchase in 1 year/6 months/3 months
- ECOMSHOP – the percentage of the total amount purchased in ecommerce
- GASSHOP – the percentage of the total amount purchased in gas
- MEDICALSHOP – the percentage of total amount purchased in medical products
- GRACERYSHOP – the percentage of total amount purchased in grocery

In the supportive dataset, free-zipcode-database-Primary, there were 20 columns and 42,523 rows. The following columns were used for data visualization in Tableau.

Location:

- Zipcode - zip code of each city in the United States
- State - the state of each city in the United States

Data Exploration

Statistical Visualization of the Data

The effort here was to remove missing values and redundant data as well as outliers to make sure they had no influence on the statistical analysis and that the data was ready for both visualization purposes and modeling.

A Correlation plot was created to investigate any dependencies among variables and to identify the most correlated numeric variables in a dataset. From the correlation plot, **Figure 1,** it was noted that:

- Account type (B2B) and Business type (A2A) were correlated - 0.86
- Shopping frequencies of 3 months, 6 months and 1 year were highly correlated.
    - Shop1yr and Shop6m - 0.96
    - Shop1yr and Shop3m - 0.93
    - shop6m and Shop3m - 0.97

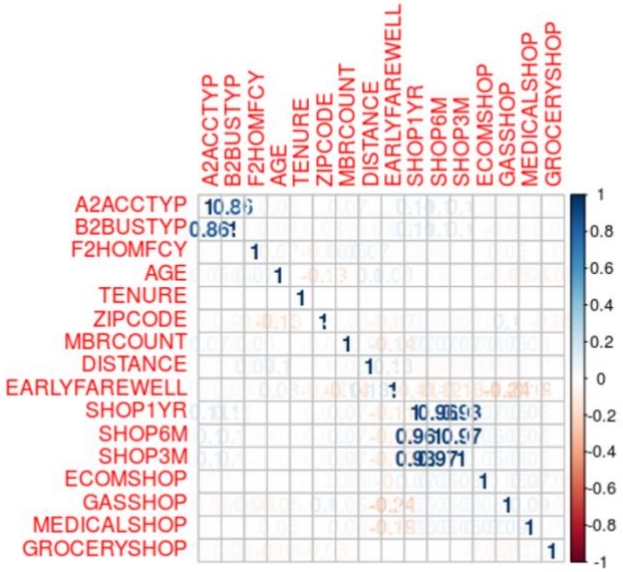

Figure 1: Correlation Plot for Numerical Variables

Next, the outliers were examined for various columns in the dataset. Theoretically, an outlier is defined as an observation more than 3 standard deviations from the mean. Having outliers can heavily skew any visualizations or models. Two kinds of outliers were found here i.e. **Figure 2:** univariate (SHOPIYR; SHOP6M; SHOP3M) and **Figure 3:** multivariate (SHOPIYR; SHOP6M; SHOP3M) vs. (Region).

Uncleaned dataset with outliers' plots can be seen below:

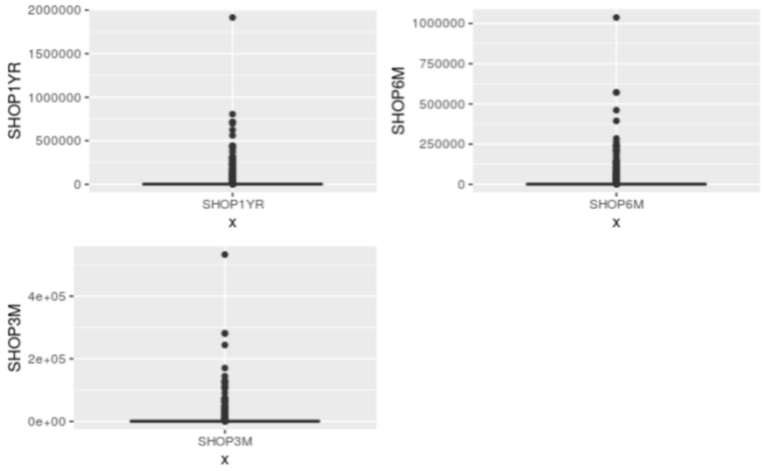

Figure 2: Plots Displaying Univariate Outliers from the Unclean Dataset

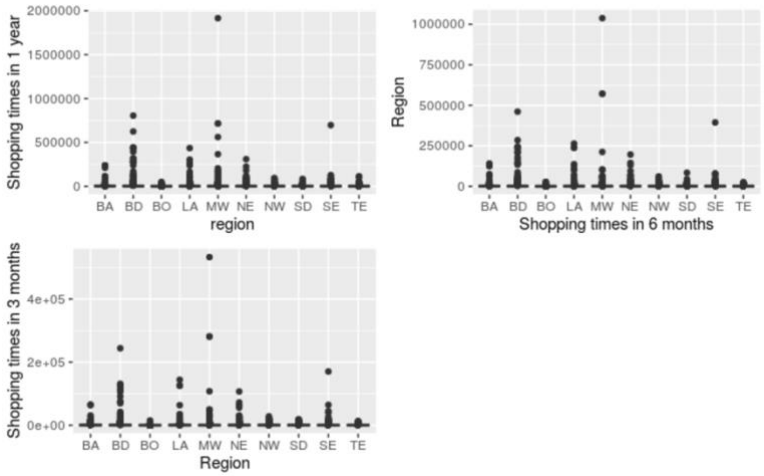

Figure 3: Plots Displaying Multivariate Outliers from the Unclean Dataset

One way to identify outliers is to determine the points with z-scores that are far from 0. The scores() function from the outliers package were used to complete this step. With the threshold =3, the condition was set to "TRUE" when outlier_scores was greater than 3 and "FALSE" if outlier_scores was less than negative 3. Here, rows with major outliers were removed from the dataset, and the graphs below display the cleaned dataset (**Figure 4 and Figure 5).**

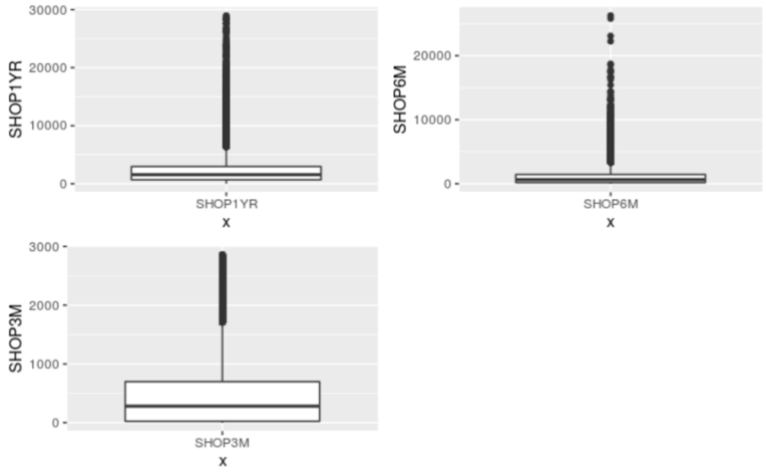

Figure 4: Plots Displaying Univariate Outliers from the Cleaned Dataset

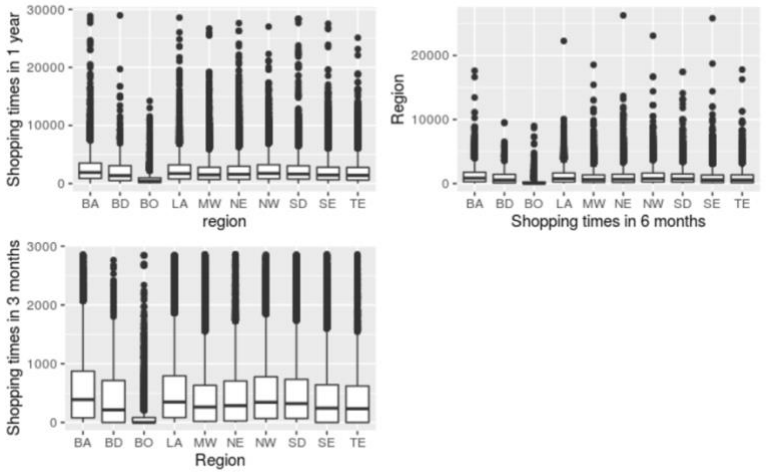

Figure 5: Plots Displaying Multivariate Outliers from the Cleaned Dataset

After removing the outliers, bar graphs of the categorical and numerical variables were plotted to observe the distribution and skewness of data. The categorical variables displayed in **Figure 6** are RENEW (Customer Churn), M2EXCFLG (Exclusive membership or not), F2HOMRGN

(Region), HOMEFCTYCHANGE (Customer changes home or not) and RECENTMOVING (Customer region). A few major observations that were noted from the graphs were:

- More customers did not renew their memberships than those that did
- Not many customers changed their residencies
- The retailer had many customers from the Midwest and the Southeast regions of the U.S.
- There were more customers with exclusive memberships than non-exclusive memberships

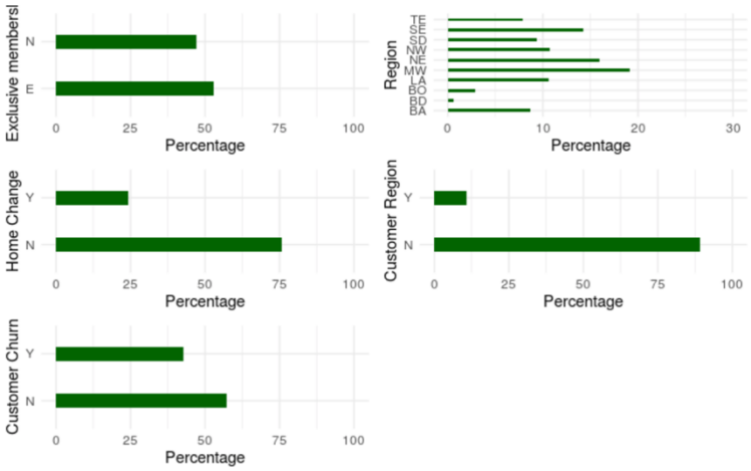

Figure 6: Data Visualization for Categorical Variables

The numerical variables displayed in **Figure 7a-b** are SHOPIYR (shopping in 1 year), SHOP6M (shopping in 6 months), SHOP3M (shopping in 3 months), MBRCOUNT (number of cards held by a single account), DISTANCE (miles to the warehouse), EARLYFAREWELL (number of days since the customer last shopped), and TENURE (length of membership). The dotted line in each plot represents the mean of the data. A few major observations can be made from the plots:

- The mean value for 1 year of shopping is 2153.639.
- The mean value for 6 months of shopping is 1006.032.
- The mean value for 3 months of shopping is 467.3689

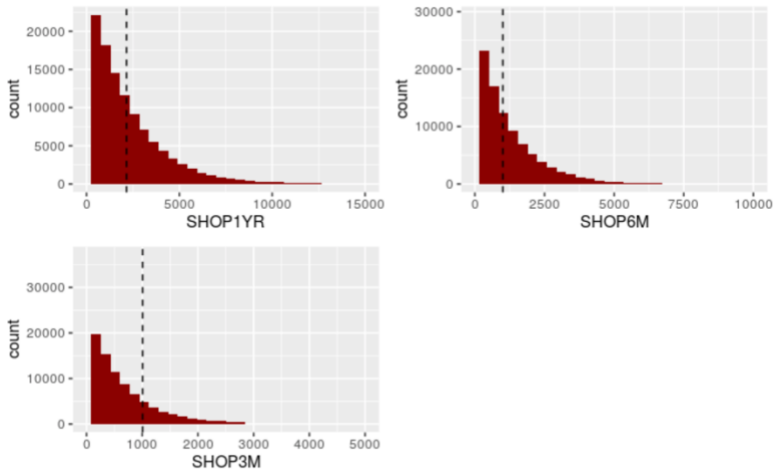

Figure 7a: Data Visualization for Numerical Variables

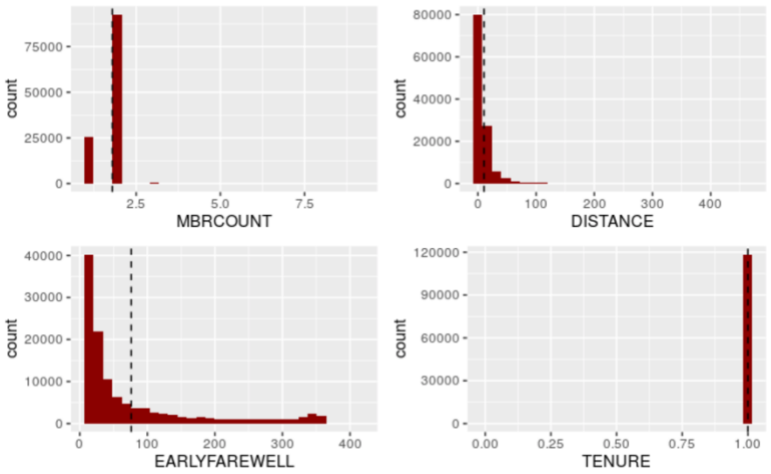

Figure 7b: Data Visualization for Numerical Variables

Tableau Visualization

In an effort to further explore the dataset, Tableau was utilized, and visualizations were presented side by side in the form of a dashboard with an interactive filtering to uncover fields that are not currently used in the view. A snapshot of the dashboard can be found in **Figure 8**, showing data from 4 main U.S. regions. Every region section, with the region's name at the top, is to be read from top to bottom. The interactive Age-range filter is set to apply the selected age range to all visualizations. Each type of visualizations will be discussed in the next paragraph.

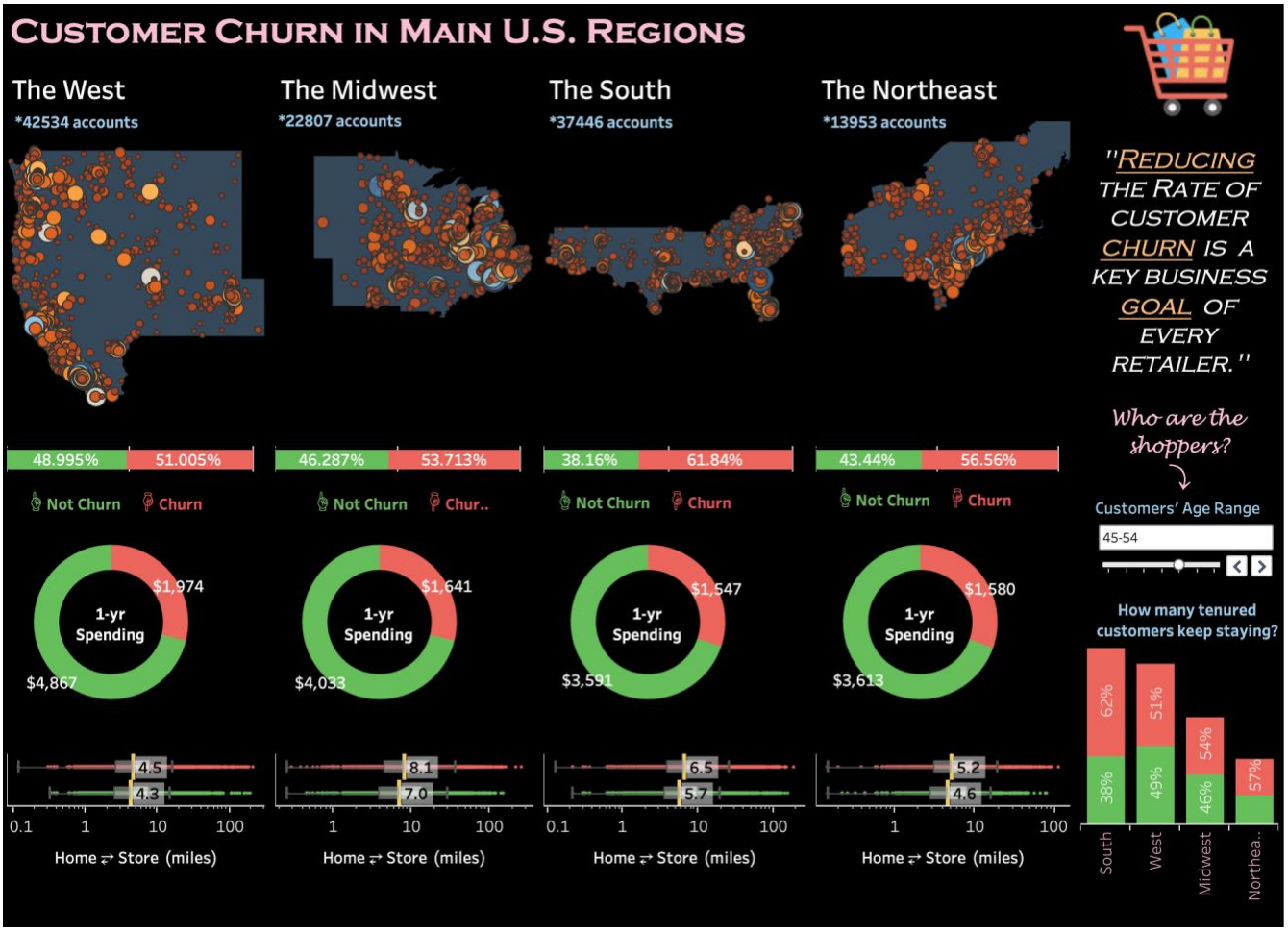

Figure 8: Data Exploration Dashboard

To visualize the distribution of the customers in each region, the account numbers were grouped by ZIP code and each area's total count was plotted onto a regional map. The bigger and darker shade of the circle indicates a higher number of accounts. The map can be zoomed in/out to better explore specific areas (see **Figure 9**).

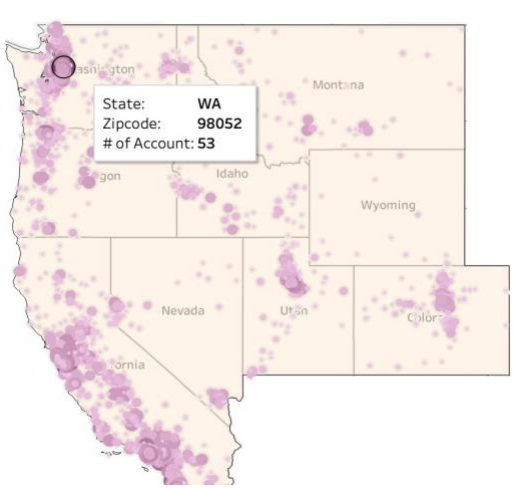

Figure 9: Number of Accounts in the Region by ZIP Code

To better interpret the difference between the proportion of regional customers who renewed their memberships and the proportion of regional customers who did not, the bar plot below was created (see **Figure 10**).

Figure 10: Proportions of Renewed/Churned Customers in the Region

To understand the relationship between the annual expenditure and membership renewal in each region, a donut chart was used to display an average-cumulative spending of the customers who renewed their memberships and that of the customers who churned (see **Figure 11**).

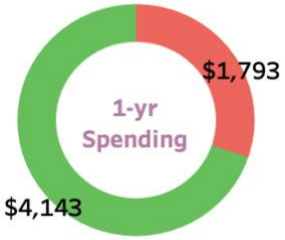

Figure 11: The Average 1-year Expenditure - Renewed vs. Churned

To assess the relationship between commuting distances and membership renewal in each region, a boxplot was created to display the distribution of miles traveled based on a median and 3 quartiles with an interquartile range (IQR). The median value of each group (renewed or churned) is shown and each distribution's outliers are observed after the third quartile (see **Figure 12**).

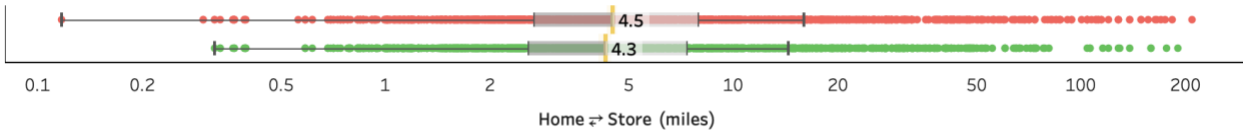

Figure 12: Distribution of Distances from Home to Warehouse - Renewed vs. Churned

To determine whether repeat customers have lower tendency to churn, the chart below was observed. Note that the dataset did not contain the numbers of years for the customers with "Tenure" status, thus making it impossible to completely identify if the length of membership after the first year had any effects on the decision to renew (see **Figure 13**).

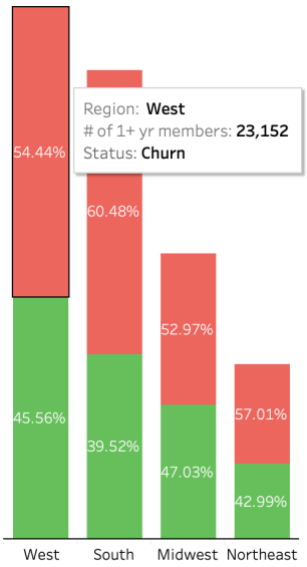

Figure 13: Numbers of Repeat Customers and Their Recent Renewal Statuses by Region

To understand the renewal behaviors of various customers' age ranges (10-year binning), the filter below was implemented in the dashboard to interact with all of the visualizations (see **Figure 14**).

Figure 14: Interactive Filter of Customers' Age Ranges

Data Mining Tasks

Data preprocessing is an essential process in building a learning model as the quality and the format of data can have a big impact on the learning ability and accuracy of the model. Further data preprocessing steps are discussed in this section.

Missing data imputation: There was no missing data in the dataset.

Data reduction: The irrelevant column to the model construction e.g., customer ID and the highly corelated columns were dropped from the dataset. As introduced in the section above, a few boxplots were created to examine the outliers. The outliers of the SHOP1YR column were removed. Several feature selection techniques such as the stepwise selection algorithm were used to study the relative importance of predictor variables to the response variable. Moreover, the numerical outcome of each predictor's importance from the Decision Tree was considered to reduce the data dimensionality and to prepare the dataset for the model construction.

Data transformation: The response variable was encoded into a factor variable of 1 and 0, as Renewed and Churned. Pivot tables were created to compute a mean value of the binary of outcome (RENEW) as a function of multiple categorical variables. The information from pivot tables was then used to group and reduce the number of categorical levels in each categorical predictor. Dummy variables were created in order to represent binary factors for categorical predictors.

Association Rules: Association rules were explored to uncover the behaviors of customers who did not renew their memberships as well as exploring some other interesting relationships in the dataset. Insightful associations with high confidence values are displayed below.

Table 1: Exploration of Association Rules

Antecedents		Consequents	Support	Confidence	Lift	Count
TENURE=1 and earlygroup=301 – 360 and shop1YrGROUP=0-1000	=>	RENEW=N	0.048717	0.92148	1.6321	5868
RENEW=Y and M2EXCFLG=N and F2HOMRGN=MW and MBRCOUNT=2 and age =30-34 and earlygroup=0 – 60 and DISTANCE=Less then 10	=>	shop1Yr = 1001-5000	0.002150	0.82747	1.5080	259

As the first row suggests, churned customers were repeat customers who had cumulative annual spending amounts of less than $1000 at the retailer and made their last store-visits over 300 days ago by the time of data extraction.

It was also noted in the second column of the table that customers who had cumulative annual spending amounts between $1001-5000 were 30-34 years old, non-executive customers who had 2 cards associated to their accounts, lived less than 10 miles away from the Midwest warehouses, made their last store-visits within the past 60 days, and most importantly they renewed their memberships.

Data Mining Models

Historical data is key for future churn prediction. With the knowledge of which customers left as a response variable, the information on their demographic characteristics and shopping behaviors can be examined and used as so-called predictors. The aims of the next few sections are to investigate the existing techniques in data mining and to propose the best predictive model for the customer churn prediction in addition to identifying churning factors. Five classifier models were developed and assessed to uncover the model that best classifies the customers into the churn and non-churn categories.

From the data preparation process, all of the relevant categorical variables were converted to dummy variables. The dataset was then partitioned rows of data randomly into training (70%) and validation (30%) sets. This same training set was used to fit all models, while the validation set was held out to assess the models' performance.

Logistic Regression

Since the response variable in the dataset constitutes binary outcome (having two possible classes), Logistic Regression was first to be evaluated as the theoretical model is known to efficiently extend linear regression concepts to accommodate the situation where the response variable is a binary categorical. By fitting available historical data to a statistical model that correlates the predictors to the response, potential churn candidates among existing customers can be predicted. In other words, the prediction is the output of a set of equations associating values affecting customer churn with an output value, the probability of churn.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n ,$$

Where:

y is the response variable for each customer account - a binary class label (renewed/churned)

β_0 is a constant or noise in the model

$\beta_{1\text{-}i}$ are coefficients or weights given to specific predictors

$X_{1\text{-}n}$ are predictor variables for customer account

To ensure the accuracy of the prediction, the following assumptions were addressed:

- The outcome was a binary – response class were renewed/churned
- The dataset was lacking extreme values or outliers in the continuous predictors – outliers in the dataset were removed during the data preparation process.
- There was no multicollinearity among the predictors – highly correlated columns were addressed during the data preparation process. The predictors were independent of each other by the time of model fitting.
- There was a linearity between each predictor and logit or log(odds) = $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$

The customer churn probabilities calculated by the Logistic Regression model yielded values between 0 and 1. In order to conclude with classifications into either 1 or 0 (renewed/churned), a threshold or cutoff value was applied. In the first modeling attempt, the default cutoff value of 0.5 was used. Before a classification of a customer in the validation dataset could be made, the information on his/her demographic characteristics and shopping behaviors was passed into the fitted equation. This yielded an estimated probability of renewing the membership, which was then compared to the cutoff value. When the probability of the customer renewing the membership was above the cutoff, that customer was classified as "renewed/non-churn". During an improvement process of the model, the attempt to compute an optimal cutoff was performed. Interestingly, the best cutoff was deemed to be 0.4993099, which was essentially 0.5.

In an attempt to reduce the complexity of the model without compromising the accuracy, stepwise regression was performed. Stepwise regression is an adaption between forward and backward selection techniques. It is an alteration of the forward selection in the way that after each step in which a predictor is added, all other predictors in the model are assessed to see whether their significance values have been reduced below a specified tolerance level. When a nonsignificant predictor is deemed, it is removed from the model. In this study, the simpler model that the stepwise regression returned was compared against the full/original Logistic Regression model by performance.

It is to note that the relationship between Y and the β parameters is nonlinear in Logistic Regression. Hence, β constraints of the churn predictors were not estimated using the method of least squares as in traditional linear regression. Instead, an evaluation method called Confusion matrix for maximum likelihood classifier was adopted. The performance result will be discussed in the following section.

Logistic regression is one of the most preferred classification algorithms as it generates more informative prediction than other classifiers by outlining the relationship between the response variable and predictors in the form of an equation with coefficients for the predictors. Nevertheless, the algorithm has its own limitations. First, it is required for each row of data to be independent of

all other rows. If observations are to some extent linked to one another, then the model can overweight the significance of those observations. In this dataset, this was a drawback as multiple observations came from the same customer account numbers. Recall that the retailer allows individual accounts to have 2 cards and unlimited numbers of cards for business accounts. Second, the logit algorithm tends to overconfidence and could show higher predictive power than it truly has as a result of sampling bias. In this dataset for example, a random sample of customers might lead the logit model to predict that all customers who live within 15 miles from the regional warehouses and have executive-type memberships will always renew their memberships. In reality, however, a small percentage of customers with these characteristics might churn. The Logistic Regression in this case would therefore be overconfident, meaning that the model exaggerates the accuracy of the prediction.

Decision Tree

Decision tree is the most commonly used tool for prediction and classification of future events as it can support both categorical and continuous data. It generates a tree-like structure that represents a set of decisions. The tree returns the probability scores of class membership. The development of the tree is done in two major steps: building and pruning. Both classification and regression trees are constructed by recursive splits of an instance into subgroups until a specified criterion has been met. During the first phase, the data set is partitioned recursively until most of the records in each partition contain identical value or until the decrease of Gini impurity falls below a user-defined threshold. The second and optional phase, pruning, then removes some branches which contain noisy data (branches with the largest estimated error rate). Each node in a Decision Tree is a test condition and branching is based on the value of the attribute being tested. The tree is representing a collection of multiple rule sets.

In the study, the function rpart.plot from the rpart.plot package was used to generate a Decision Tree on the customer churn training data. The function's control option minbucket is the minimum number of entities that are allowed in a terminal node. The default number which is computed as round(minsplit/3) was used. The minsplit parameter is the minimum number of entities in the parent node that can be further split. The ran default value was 20. The last parameter, maxdepth

prevents the growing tree from branching past a certain depth / height. The ran default value was 30. The complexity parameter (cp) in rpart is threshold complexity and used to optimize the splitting of the tree. In other words, it is pre-pruning the tree and acting as a stopping parameter to let the program know that if the split doesn't improve the fit, cross-validation should then be done. For the study, the default value of 0.001 was used.When evaluating the model, the classification was done by traversing through the tree until each leaf node was reached.

For pruning, insignificant rules such as repetitive rules are removed. This process adds quality and generates a more insightful tree. In addition, the tree may have overfit the training data and resulted in an overly complex model. Pruning cuts a certain unnecessary depth. The performance of the pre-pruning and post-pruning trees on the customer churn data will be discussed in the performance evaluation section.

Random Forest

Random forest is a classification model comprising of a huge number of various Decision Trees. Every Decision Tree with low correlation gives out a class prediction. The goal of random forest is to overcome the overfitting problem of the individual Decision Tree by using bagging (bootstrap aggregation) and feature randomness. The downside of the traditional Decision Tree is that they tend to be sensitive to the training data and any minor changes on the training dataset can generate significantly different tree structures. However, the bagging process in random forest allows every tree to randomly sample from the training dataset, resulting in different tree structures. The feature randomness helps each tree pick only from a random subset of features, resulting in training the model with more variety and diversity. After this, the final class prediction is the average class predictions from all trees.

Random forest is based on bagging and feature randomness with the idea of low bias and high variance and then averages them for the final class prediction. It performs well on a large dataset and can handle missing variable without variable deletion. Additionally, it is robust to outliers and can handle outliers and non-linear parameters efficiently. One main limitation of the random forest model is that a more accurate prediction model requires more trees, and it can make the algorithm

run very slow when testing the model. Meanwhile, unlike a single Decision Tree, random forest is more complex and is not easy to simply interpret the trees. It is also not suitable for the description of the relationships in the dataset.

For this study, random forest was applied to categorize the customers into renewed and churned groups by using a 70% partitioned dataset as a training dataset. The large size of the training dataset was suitable for model processing by forming different combinations of trees. In general, the importance of each predictor variable graph is plotted and a higher reduction in the mean value of Gini indicates a higher variable importance. The plot function can be applied to the random forest model, resulting in the relationship plot between the estimated error rate and trees included in the model. The accuracy in the class prediction can be improved by tuning the model and narrowing down the optimal numbers of trees. Additionally, this prediction model was also tested using a 30% partition dataset as the validation dataset, and model evaluation techniques such as confusion matrix were applied to find the effectiveness of the random forest classifier algorithm. The performance result will be discussed in the next section.

k - Nearest Neighbor (k-NN)

k-NN was used for classification problems in this study. Theoretically, the model relies on labeled input data to learn a function that produces an appropriate output when given new unlabeled data. The model is fitted with selected variables to generate an output of whether or not a customer will continue his/her membership.

To evaluate this type of model, three important aspects are considered: ease to interpret the output, which is the renewed/churned in this case, then the calculation time, and the predictive power. It is to note that the model represents a supervised classification algorithm that gives new data points according to the k number or the closest data points. The process is very easy to implement and allows an addition of new data seamlessly. However, this type of model does not work so well with large datasets that have missing values and noise. Moreover, it requires feature scaling before applying the algorithm to the dataset to prevent wrong predictions.

In this study, two k-NN models were built. The first one was built using the subset of the data with the top 10 most significant features found from Decision Tree algorithm. The second model was built using the sub-dataset with the most significant features based on stepwise Logistic Regression. For each model, the data was partitioned into training and validation in a 7:3 ratio. A function was written to compute for multiple k values on the validation data to find the best 'k'. Note that as the size of the k value decreases, the prediction becomes less stable. Ultimately, using the best value of k, the model was run on the validation dataset and the accuracy was calculated along with the confusion matrix.

Based on what features the Decision Tree model deemed most significant, the same set of identified features were used for the first k-NN model. The features were RENEW, F2HOMFCY, AGE, MBRCOUNT, DISTANCE, EARLYFAREWELL, SHOP1YR, ECOMSHOP, GASSHOP, MEDICALSHOP, GROCERYSHOP, M2EXCFLGE, HOMEFCTYCHANGEN, RECENTMOVINGN, F2HOMRGN_BOFALSE, F2HOMRGN_TEFALSE, F2HOMRGN_middleFALSE. The best value of k=53 generated a moderately high accuracy value of 0.7039.

Based on what features the stepwise regression for Logistic Regression deemed most significant, the same set of identified features were used in the second k-NN model The features were RENEW, F2HOMFCY, AGE, MBRCOUNT, DISTANCE, EARLYFAREWELL, SHOP1YR, ECOMSHOP, GASSHOP, MEDICALSHOP, GROCERYSHOP, M2EXCFLGE, HOMEFCTYCHANGEN, RECENTMOVINGN, F2HOMRGN_BOFALSE, F2HOMRGN_TEFALSE, and F2HOMRGN_middleFALSE. The best value of k=55 generated a moderately high accuracy value of 0.6932. This was almost identical with the accuracy from the first k-NN model.

Using the k-NN algorithm was advantageous, as it was easy to implement. However, the general downside of the model is that a big dataset will take the model a long time to run and having multiple independent variables will make the model process even slower. In this study, the variable selection process could not be done based on the domain knowledge. The features were pre-

selected using the Decision Tree and stepwise Logistic Regression, thus a considerable amount of model's fitting time was saved

Support Vector Machine (SVM)

Support Vector Machine (SVM), categorized as a supervised machine learning algorithm, is well-known in its ability to tackle both classification and regression tasks. In this study, SVM was adopted to solve the classification problem and the objective of the algorithm was to fit the training set and predict the target outcome, Renewed/churned during a validation process. Generally, SVM requires all variables to be numeric variables, whether they are the actual or the dummy. SVM finds a hyperplane with the maximal margin in higher dimensions to separate data into groups, renewed or churned in this case. Because the dataset was non-linear and the data could not simply be separated by a linear function, Radial basis function kernel (also known as RBF kernel or Gaussian kernel) was used in this study. The mathematic function is seen below.

$$K_{RBF}(x, y) = \exp\left[-\gamma \| x - y\|_2\right], \qquad \text{where}$$

γ is a hyper constraint or tuning constraint to set the spread of the kernel
x is the predictor variable
y is the response variable - Renewed/churned

There were a few benefits in using SVM for this study. First, SVM scales relatively well to high dimensional data and the dataset had 38 transformed variables. Secondly, SVM was perfect for this study's non-linearity in data. The kernel function, RBF kernel in this case, is designed to solve problems with more complexities. SVM, while being relatively memory efficient, is a suitable method to decrease the risk of overfitting due to generalization.

Nevertheless, there were still some challenges in using SVM for this study. First, SVM is typically not the best method for a large dataset and there were 84,315 rows in the study's training set. The fitting process was very time consuming due to the highly dimensional data. Secondly, SVM's performance power can be reduced when a dataset has more noise. It was observed during the

model's evaluation that the output target classes (renewed/churned) were somewhat overlapping. Lastly, the fact that SVM only places a data point above or below the classifying hyperplane resulted in the rows being classified as renewed/churned without any probabilistic explanation for the classification.

Performance Evaluation

In evaluating each model fitted in the study, Confusion Matrix and Receiver Operating Characteristic (ROC) Curve were implemented as validation techniques.

Confusion Matrix

One of the ways to gain a great perspective of what a classification model is getting correctly and how many incorrect predictions it is making, is to utilize the classification matrix or confusion matrix. The matrix recaps the correct and incorrect classifications that a model generated for a validation dataset.

The schematic confusion matrix can be seen below with n1,2 representing the number of records and classes Renewed/Churned customers.

Table 2: Mathematics Behind the Study's Confusion Matrix

		Actual Class	
		Renewed	Churned
Predicted Class	Renewed	n1,1 = number of Renewed records classified correctly	n2,1 = number of Churned records classified incorrectly as Renewed
	Churned	n1,2 = number of Renewed records classified incorrectly as Churned	n2,2 = number of Churned records classified correctly

The overall accuracy of a classifier is calculated by:

accuracy = (n1,1 + n2,2) / n,

where n is the total number of records in the validation dataset. The highest accuracy is 1.0 and the lowest is 0.

Additionally, two other accuracy measures that are commonly used are:

The first one is the sensitivity of a classifier which introduces the concept of measuring the ability to detect the important class customers correctly. This is estimated by n1,1/(n1,1 + n1,2), the percentage of renewed customers classified correctly.

The second one is the specificity of a classifier which introduces the concept of measuring the ability to rule out churned customers correctly. This is measured by n2,2/(n2,1 + n2,2), the percentage of churned customers classified correctly.

ROC (Receiver Operating Characteristic) Curve

ROC Curve is another well-known method for plotting the two measures, sensitivity of a classifier against specificity of a classifier. The curve plots the sensitivity and the specificity as the cutoff value descends from 1 to 0. The alternative arrangement of this plot has 1-specificity on the x-axis, which allows 0 to be located on the left end of the axis and 1 on the right. Greater performance is shown when the curves are plotted closer to the top-left corner. The comparative assessment curve is the diagonal line displaying the performance of the naive rule, using varying cutoff values.

ROC can also be interpreted with numeric AUC (Area Under the Curve) value. The value ranges from 1 given a model being the perfect algorithm for class segregation to 0.5 given a model being equal/less efficient compared to the naive rule.

Logistic Regression

During the performance evaluation process of Logistic Regression, the new model that the stepwise regression returned was compared against the full/original Logistic Regression model by measuring ROC's AUC and passing through the confusion matrix.

It was noted that both models resulted in approximately identical values for confusion matrix measures and area under the ROC (AUC). The full Logistic Regression gained 0.7761 for AUC and 70.86%, 76.90%, 63.00% for confusion matrix's accuracy, sensitivity, specificity respectively. The stepwise Logistic Regression's AUC value was 0.7762 and its confusion matrix's accuracy, sensitivity, specificity values were 70.85%, 76.92%, 62.97% respectively. The ROC plot of the new, feature-reduced model, **Figure A**, can be accessed in the Appendix section. Stepwise regression removed a total of 12 predictors and selected a set of 16 predictors for the new model. Here, it was proven that the stepwise regression process reduced the complexity of the original model without compromising its accuracy. Hence, the final model for this study's Logistic Regression effort will be the model returned by the stepwise regression.

Decision Tree

The most important attribute in deciding whether a customer would renew or not was found to be EARLYFAREWELL, which the model placed at the root node. The terminal nodes/leaves show fraction values of the records which were labeled by the Decision Tree. The unpruned classification tree, shown in **Figure 15**, was plotted using the training data with default parameter values as minbucket as 7, minsplit as 20 and the complexity parameter (cp) as 0.001 (default values). The list of rules generated by the rpart algorithm to sort the data can be seen in **Figure 16**. The developed model resulted in the confusion matrix's sensitivity of 76.54% and accuracy of 70.87%. The top most significant variables in the classification were found to be EARLYFAREWELL, SHOP1YR, GASSHOP, MEDICALSHOP, GROCERYSHOP.

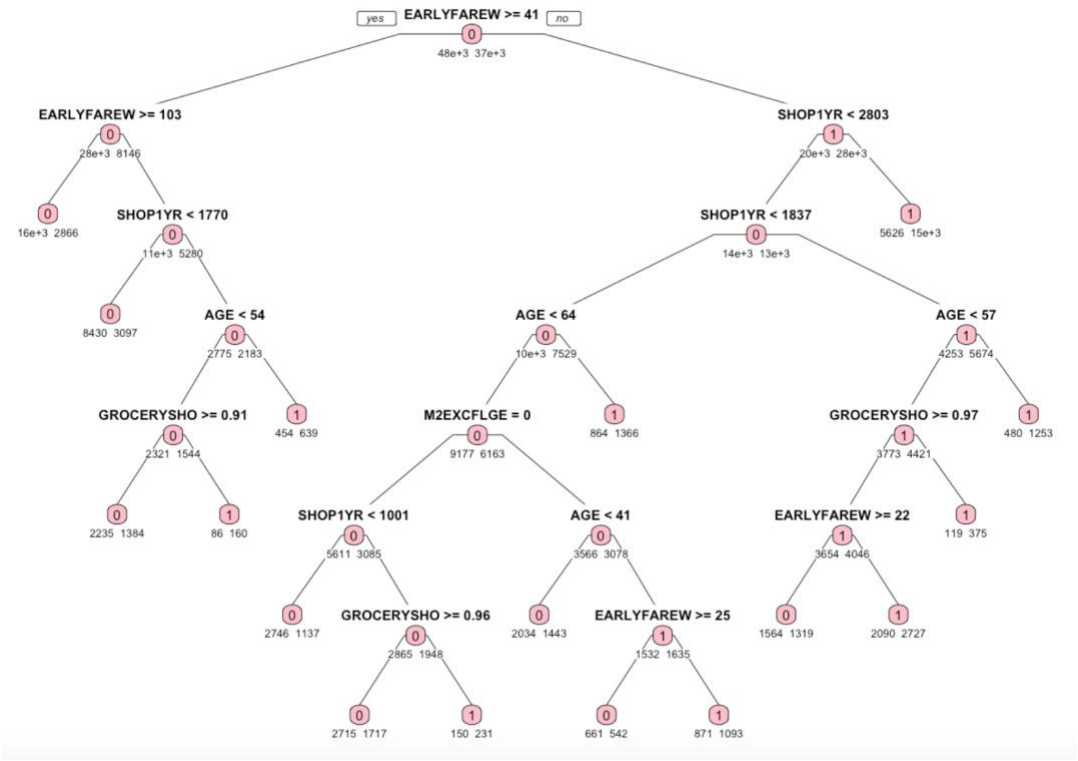

Figure 15: Unpruned Classification Tree

	RENEW																								
4	0.15	when	EARLYFAREW >=			103																			
10	0.27	when	EARLYFAREW is	41	to	103	&	SHOP1YR	<	1770															
96	0.29	when	EARLYFAREW <	41			&	SHOP1YR	<	1001			&	AGE	<	64							&	M2EXCFLGE	is	0
44	0.38	when	EARLYFAREW is	41	to	103	&	SHOP1YR	>=			1770	&	AGE	<	54			&	GROCERYSHOP	>=	0.91				
194	0.39	when	EARLYFAREW <	41			&	SHOP1YR	is	1001	to	1837	&	AGE	<	64			&	GROCERYSHOP	>=	0.96	&	M2EXCFLGE	is	0
98	0.42	when	EARLYFAREW <	41			&	SHOP1YR	<	1837			&	AGE	<	41							&	M2EXCFLGE	is	1
198	0.45	when	EARLYFAREW is	25	to	41	&	SHOP1YR	<	1837			&	AGE	is	41	to	64					&	M2EXCFLGE	is	1
104	0.46	when	EARLYFAREW is	22	to	41	&	SHOP1YR	is	1837	to	2803	&	AGE	<	57			&	GROCERYSHOP	>=	0.97				
199	0.56	when	EARLYFAREW <	25			&	SHOP1YR	<	1837			&	AGE	is	41	to	64					&	M2EXCFLGE	is	1
105	0.57	when	EARLYFAREW <	22			&	SHOP1YR	is	1837	to	2803	&	AGE	<	57			&	GROCERYSHOP	>=	0.97				
23	0.58	when	EARLYFAREW is	41	to	103	&	SHOP1YR	>=			1770	&	AGE	>=			54								
195	0.61	when	EARLYFAREW <	41			&	SHOP1YR	is	1001	to	1837	&	AGE	<	64			&	GROCERYSHOP	<	0.96	&	M2EXCFLGE	is	0
25	0.61	when	EARLYFAREW <	41			&	SHOP1YR	<	1837			&	AGE	>=			64								
45	0.65	when	EARLYFAREW is	41	to	103	&	SHOP1YR	>=			1770	&	AGE	<	54			&	GROCERYSHOP	<	0.91				
27	0.72	when	EARLYFAREW <	41			&	SHOP1YR	is	1837	to	2803	&	AGE	>=			57								
7	0.73	when	EARLYFAREW <	41			&	SHOP1YR	>=			2803														
53	0.76	when	EARLYFAREW <	41			&	SHOP1YR	is	1837	to	2803	&	AGE	<	57			&	GROCERYSHOP	<	0.97				

Figure 16: Rules from Unpruned Classification Tree

Even though the unpruned tree was moderately accurate, its structure was bulky, raising a concern of overfitting. Pruning is necessary to acquire trees that are less complex and ultimately optimizes the accuracy of the classification. **Figure 17** shows the pruned tree plotted on the training data with minbucket as 7 and minsplit as 20 uses the following rules deployed by the rpart function (seen in Fig xii):

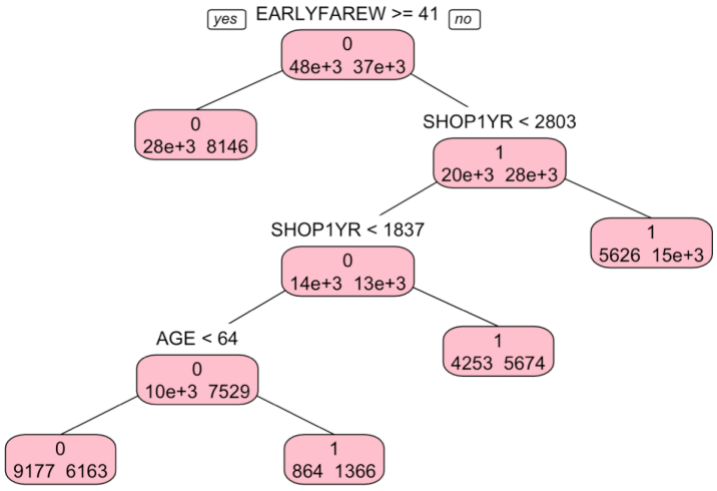

Figure 17: Pruned Classification Tree

	RENEW													
2	0.23	when	EARLYFAREW >=	41										
24	0.4	when	EARLYFAREW <	41	&	SHOP1YR	<	1837			&	AGE	<	64
13	0.57	when	EARLYFAREW <	41	&	SHOP1YR	is	1837	to	2803				
25	0.61	when	EARLYFAREW <	41	&	SHOP1YR	<	1837			&	AGE	>=	64
7	0.73	when	EARLYFAREW <	41	&	SHOP1YR	>=			2803				

Figure 18: Rules from Pruned Classification Tree

From the pruned model, it could be interpreted that there is a great possibility that the customers would churn when they did not shop at the store/warehouse for at least 41 days. Moreover, it is worth noting that a customer aged less than 64 who has annual shopping amount of less than $1837 might not renew his/her membership.

It can be hard to gauge levels of contribution variables have for a model. A statistical function variable.importance was used in the study to calculate the statistical significance values of predictors. From the unpruned tree, the five most significant were found to be EARLYFAREWELL, SHOP1YR, GASSHOP, MEDICALSHOP, and GROCERYSHOP. However, the pruned tree's variable significance result also contained DISTANCE. In fact, it was found that DISTANCE played a more important role in the class prediction than the GROCERYSHOP.

The results achieved from the pruned model indicated sensitivity to be 76.72% and accuracy as 70.26%. These numbers are mostly identical with those of unpruned tree. ROC curve for the pruned classification tree can be seen in Figure B in the Appendix section. Its AUC value was 0.7309.

Random Forest

Random Forest performs well on a large dataset and can automatically handle missing variables and outliers. Before the model was tuned, the result from the confusion matrix had already indicated that the sensitivity (80%) and accuracy (75.26%) were higher for the Random Forest classifier as compared to other types of models. The model evaluation was conducted by taking subsets of data that the trees within the model had never seen before to test how well those trees performed. This is the way to determine the out-of-bag (OOB) error rate, which is a useful metric to evaluate how well the model is performing. As observed in **Figure 19,** the optimal tree numbers are between 100 and 200, resulting in the minimal OOB error. The random forest model can be tuned by selecting the number of trees and the number of randomly chosen variables to be tested at each split within the trees (mtry). In the study, the tuned random forest function was applied on the validation dataset with tree numbers of 100,150,180, 185, and 200. Note that the goal of this function is to find an appropriate m(try) value that minimizes the OOB error. In **Figure 20,** the m(try) value should be 10 in order to minimize the OOB error, when the tree number is equal to 180. After the model was manually tested with different numbers of trees and mtry values, it was observed that random forest classifier performed better with tree numbers = 180 and mtry = 10. The result of this showed 75.25% accuracy and 80% sensitivity. **Table 3** displays distinctive results from random forest performance evaluation on various m(try) and number of trees**. Figure C** in the Appendix shows the relationship of the OOB error and number of trees for the tuned random forest classifier algorithm. The plot suggests that the tree numbers of 180 is reasonable for the tuning purpose. Furthermore, the ROC's AOC is observed in **Figure D** in the Appendix. The tuned random forest classifier performed very well compared to other algorithms with AUC value of 0.83.

Table 3: Random Forest Performance Evaluation on Various m(try) and Number of Trees

m(try)	Number of Trees	OOB Error	Accuracy
10	100	25.3%	75.11%
10	150	25.1%	75.21%
10	180	24.9%	75.25%
10	185	24.9%	75.23%
10	200	24.9%	75.19%

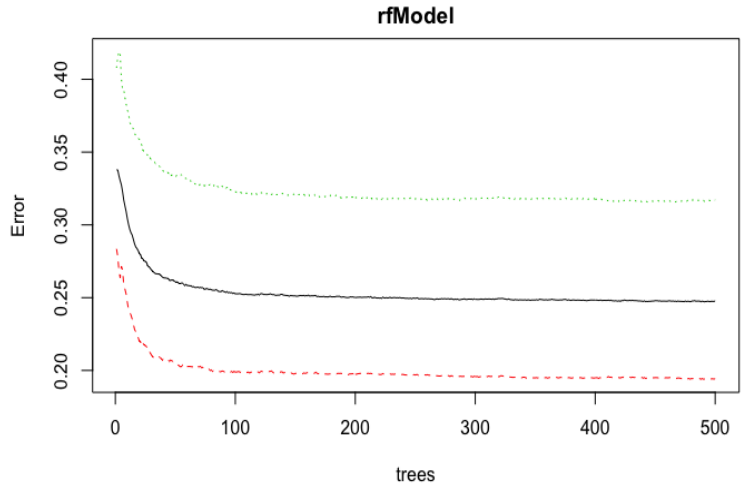

Figure 19: The Relationship between OOB Error and Number of Trees - Initial RF Model

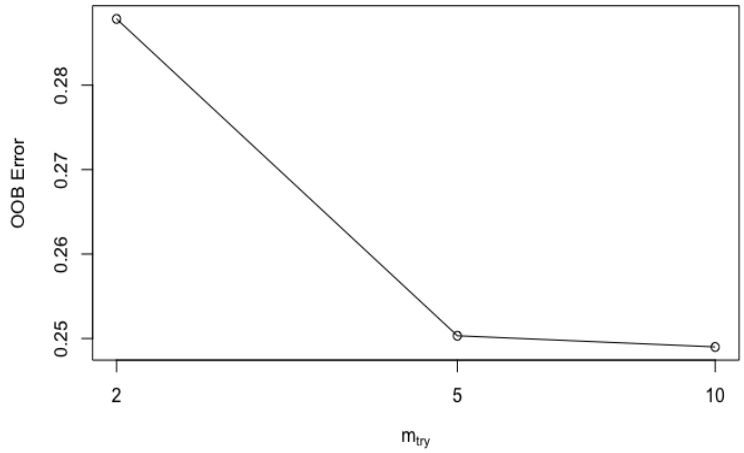

Figure 20: The Relationship between OOB Error and m(Try) Value, Tree = 180

k-Nearest Neighbor (k-NN)

The first k-NN model was built using the top 10 most significant variables found from the Decision Tree model. This k-NN model had an AUC value of 0.6960 which was the lowest when compared against AUC values of other model types. **Figure E** in the Appendix shows the model's ROC.

The second k-NN model was built using the significant variables found by stepwise Logistic Regression. It was noted that the AUC value of this model, 0.6826, was lower than the previous k-NN model with an AUC value of 0.6826. **Figure F** in the Appendix shows the model's ROC.

Looking at confusion matrixes for both models, the results were quite similar. The first model had 70.38%, 76.09%, 63.10% for the matrix's accuracy, sensitivity, specificity respectively. The second model's accuracy, sensitivity, specificity values were 69.32%, 77.01%, 59.52% respectively.

Support Vector Machine (SVM)

According to the result from a confusion matrix, the accuracy of the SVM model was 71.59% in predicting the renewal status in the validation set. The sensitivity was 76.90%, representing the rate of correctly identifying the renewed customers. The specificity was 64.68%, representing the rate of correctly identifying the churned customers. Both numbers were acceptable with the rate of sensitivity slightly higher. This indicates that the model is better at classifying churned customers. The ROC curve of SVM can be found in **Figure G** of the Appendix. The AUC was 0.7079, suggesting that the model performed fairly well in the study.

Project Results

Data Visualization

The Tableau visualization dashboard, **Figure 8,** presents the distribution of customers throughout the U.S. as well as suggesting the demographics and behaviors of both renewed and churned customers. It sections out into four vertical sections based on the major sales regions of the country. For each region, interactive charts show the proportions of customers who renewed or churned,

the average annual amounts spent by each type of customers, and the median distances traveled to the store for each type of customers. All these charts together can be filtered by the various age groups of the customers.

Interestingly, it can be seen that the most loyal customers were from the Midwest (approximately 47%) and the highest churn rate (approximately 60%) was in the South. The fact that the West shopped the most, with an average spending of $4143, could be linked to the smallest median distance traveled between home and store (4.2 miles). Overall, the Northeast and the South may not be the most profitable regions for the retailer based on the churn results and the average spending values. Some other interesting findings include:

- Least number of churns in the Midwest for age group of 65+
- The churn rate was the highest in the young adults of 15-24 age group. The rate was as high as 65-75%
- The West had the highest spending amounts
- As the values in the age group increases, the churn rate decreases

Prediction Models

In this project, various machine learning algorithms were proposed as churn prediction models; Logistic Regression, Decision Trees, Random Forest, k-NN, and SVM. The performance of each classifier was evaluated using the hold-out validation dataset. All of the models were assessed by standard evaluation metrics such as confusion matrix's accuracy, sensitivity, specificity; and ROC area (AUC). **Table 4** displays the summary of all classifier algorithms' performance in predicting customer churn.

Table 4: All Classifier Algorithms' Performance in Predicting Customer Churn.

Model	Accuracy	AUC
Stepwise Logistic Regression	70.85%	0.78
Pruned Decision Tree	70.26%	0.73

Tuned Random Forest	75.25%	0.83
K Nearest Neighbor	70.38%	0.70
Support Vector Machine	71.59%	0.71
Stepwise Random Forest	75.44%	0.83

The result table confirms that the best churn model was built using the Random Forest classifier algorithm with 75.25% accuracy in successfully classifying the customers into renewed and churned categories. The 0.83 AUC value proves that the Random Forest algorithm is robust for the binary outcome classification problem. Additionally, multiple feature selection techniques such as stepwise selection algorithm and variables importance plot from the tree algorithm were implemented. The top ten features with the highest variable importance were the following: EARLYFAREWELL,SHOP1YR,DISTANCE,GASSHOP,MEDICALSHOP,GROCERYSHOP, AGE, ECOMSHOP,M2EXCFLGE, M2EXCFLGN. It was noted that the Logistic Regression model gained a moderate performance accuracy and the second highest AUC value after the stepwise feature selection process was performed. This indicates that the selection method was effective, and the identified features were satisfactory to be used as starters in other models such as k-NN and Random Forest.

The robustness of the traditional Random Forest combined with the selection method resulted in the best overall model performance among all models. The Stepwise Random Forest model achieved 75.44% accuracy and 0.83 AUC, indicating an improvement compared to the traditional Random Forest model.

Factors behind the churning of customers were also examined by performing the Association Rules with 90% confidence. One key characteristic of the churned customers is that they had cumulative annual spending amounts of less than $1000 at the retailer and made their last store-visits over 300 days ago by the time of data extraction.

Impact of the Project Outcomes

Membership churn commonly links to customers' dissatisfaction and ultimately leads to a revenue loss. Some scholars have suggested that *"A long-term relationship with the customers is a very crucial factor in the logistics industry because of the innumerable aspects of service encounters which can easily be imitated by the competitors"* (Balasubramani, Rao, Puranik & Hegde, 2017, p.402). Member churn not only implies loss of business income, but also results in other negative impacts such as negative reputation. Although recruiting new members may sound like a solution in substituting the revenue loss, most of the times it costs more to advertise and recruit than to keep the current members from leaving. As the same group of scholars confirm *"A small improvement in churn will have a big impact on value over time - as gaining new customers is difficult and continuing to lose existing customers will affect the company's revenue and turnover"* (Balasubramani, et al, 2017, p.402). This is why it is necessary to understand what leads to member churn and further predict which customers will likely churn in order to apply proper business strategies/measures to prevent them from actually happening.

To predict customer churn and understand the demographic/behavioral characteristics of customers, the Random Forest classification with stepwise selection algorithm is highly recommended to generate a model with moderately high prediction accuracy.

Moreover, the association rules explored in this study gave hints to the specific wholesaler to shed light upon their current customers, repeat or new, who have not visited the store for over 300 days and spent less than $1000 during the calendar year.

A relatively accurate prediction model with approximately 76% accurate was successfully built for this study, however there were still some **challenges** faced and **limitations** to consider.

1. Understanding human behaviors is not straightforward, not only because each individual is unique but also various factors can influence and change individuals' behaviors whether for a short temporary period or a prolonged one. According to Campbell (2020), the author says "It's unlikely that your company will just be serving one type of customer. Different customers have

different needs, and these different needs translate to different behavioral patterns. Separating your customers into cohorts based on their user behavior relative to your product can be hugely helpful in analyzing, anticipating, and preventing churn."(para. 11)

2. Moreover, date/time is an important factor. The dataset used in this study did not include any time-related information. Learning models created using the same set of variables but with various time frames may likely yield slightly different results assuming an absence of abnormalities.

Lastly, these are the suggested **future works** that can be explored:

1. Build the proposed models with transaction profiles of customers as an additional data to the current dataset
2. Obtain the exact length/duration of each account to better understand behavioral patterns that customers might have at each specific range of membership duration. The dataset used in the study only had "less than 1 year" and "at least one year" as duration sub-categories.
3. Acquire date/time variable as it could provide insights to shopping patterns during various seasons of the year
4. Conduct interviews with professionals in the retail/wholesale industry to gain more domain knowledge that can enhance the feature selection process of the learning model construction.
5. Other advanced learning techniques such as Neuron Network can be explored.
6. In the marketing and business perspective, delve into measures that should be implemented to retain the customers.
7. If prolonged access to live database is feasible, explore the best frequency that the churn prediction should be made. How much time would be sufficient in implementing corrective actions to retain the predicted-to-churn customers?

Although churn prediction is complicated and labor-intensive, remember *"Using member churn analysis to look inward at what might at first be ugly and unpleasant is a critical step to growing your business. Churn rate improvement is very realistic and can have a tremendous impact on your revenue production."* a remark from S.Surdez (Surdez, 2019, para. 6).

Reference

[1] Balasubramani, Rao, Puranik & Hegde. (2017). Analysis of Customer Churn prediction in Logistic Industry using Machine Learning. *International Journal of Scientific and Research Publications*. 7(11), 401-403

[2] Campbell, P. (2020, Feb 07). "MOST IMPORTANT PROCESSES". Retrieved from https://www.profitwell.com/blog/churn-analysis

[3] Surdez, S. (2019, Nov 5). "How customer churn impacts your bottom line over the long term". Retrieved from https://www.illumine8.com/how-customer-churn-impacts-your-bottom-line-over-the-long-term

Appendix

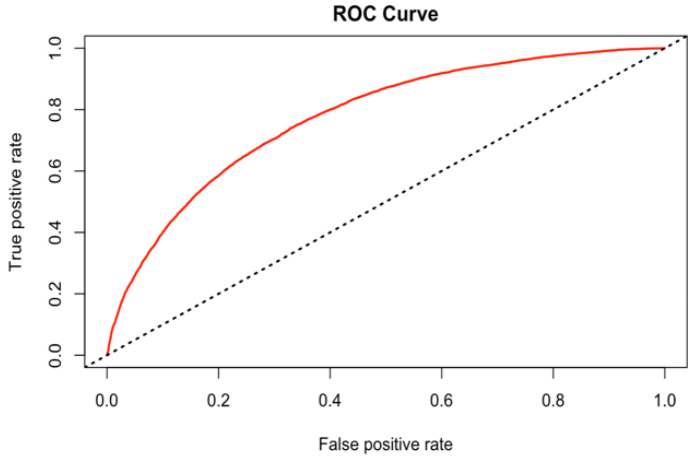

Figure A: ROC Plot for the Final Logistic Regression Model

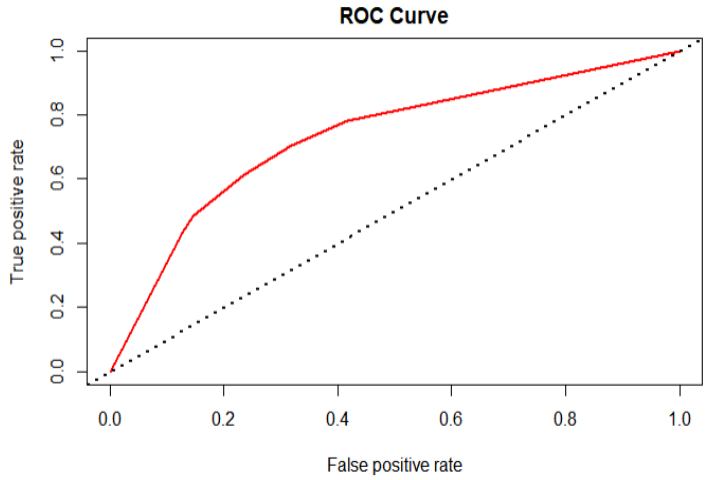

Figure B: ROC Curve for Pruned Classification Tree

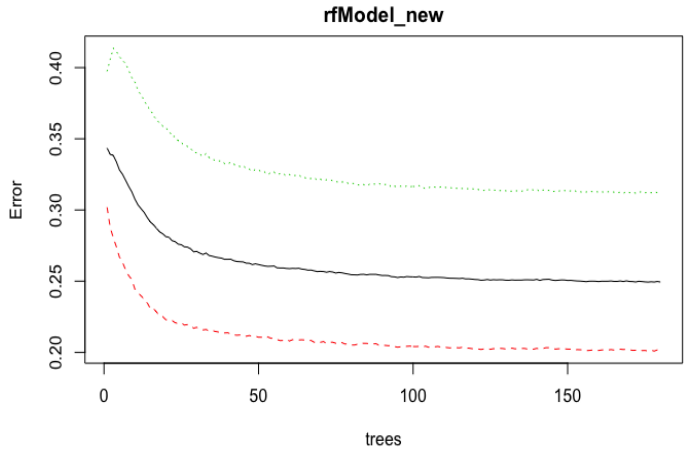

Figure C: The Relationship between OOB Error and Number of Trees - Tuned Random Forest

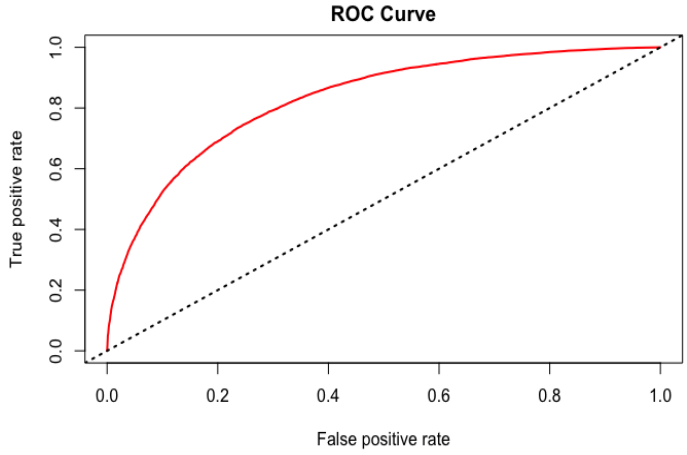

Figure D: ROC Plot for the Tuned Random Forest Classifier Algorithm

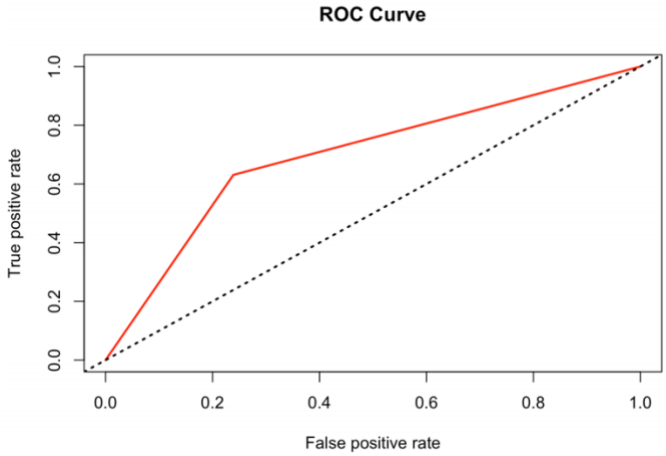

Figure E: ROC Plot for Tuned k-NN Algorithm Using Variables from Decision Tree

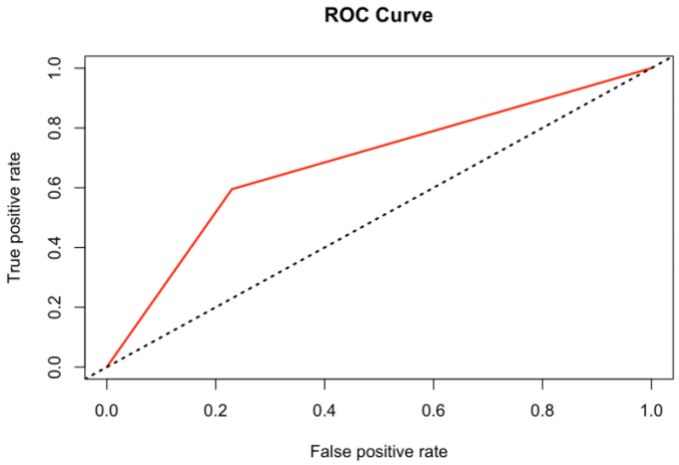

Figure F: ROC Plot for Tuned k-NN Algorithm Using Variables from Stepwise Logistic Regression

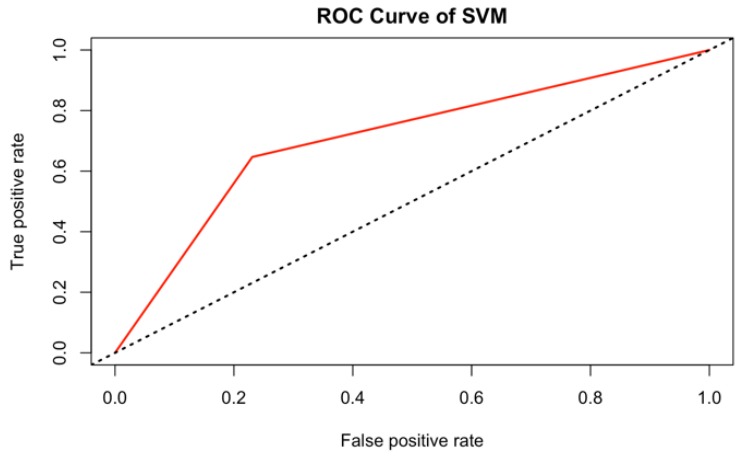

Figure G: ROC Plot for the Support Vector Machine (SVM)