## **Blue text on a black background Description automatically generated**

**MASTER OF DATA SCIENCE (SEMESTER 1 – 2023/2024)**

**FACULTY OF COMPUTER SCIENCE & INFORMATION TECHNOLOGY**

**WQD 7005 Data Mining**

**CASE STUDY (ALTERNATIVE ASSESSMENT 1)**

**NAME: TIMOTHY CHEN XIAN YII**

**MATRIC: 22056170**

**INSTRUCTOR: PROF DR TEH YING WAH**

**SUBMISSION DATE: 7th January 2024**

# **Introduction**

E-commerce is an ever-growing industry in modern society. Data analysis is critical in the fast-paced world of e-commerce for uncovering insights that fuel educated decision-making. The sheer volume of data created by online enterprises presents a goldmine of information begging to be studied.

In this case study, a hypothetical dataset containing the core attributes that are typically present in e-commerce dataset will be examined. The dataset will undergo pre-processing using Talend Data Integration and Talend Data Preparation. After that, the pre-processed dataset will undergo analysis using a decision tree to identify patterns. The dataset will also undergo modelling using ensemble and gradient boosting method. Both analysis and modelling will be carried out in SAS Enterprise Miner.

# **Dataset Description**

Two hypothetical datasets are created, “customer\_data.csv” and “customer\_marketing.csv” where customer data refers to the personal details of the customer which includes their basic information like age, gender, location. It also contains certain relevant information that pertains to their information as a consumer, such as total spendings, membership level etc. On the other hand, the customer marketing data is curated due to a recent marketing campaign where the customers are asked to subscribe to newsletters in the form or email and phone. It also asks the customers to rate their service as of the time of the marketing campaign.

With that in mind, a brief description of both datasets is provided below.

| **Customer\_data** | | |
| --- | --- | --- |
| **Parameters** | **Type** | **Description** |
| CustomerID | ID | The ID of each customer |
| Age | Integer | Age of the customer |
| Gender | Character | Gender of the customer  Available categories: Male, Female, Others |
| Location | Character | Location of the customer |
| MembershipLevel | Character | Membership level of the customer  Available categories: Bronze, Silver, Gold, Platinum |
| TotalPurchase | Integer | The number of items the customer has brought so far |
| TotalSpent | Float | The amount of money spent on purchases by customer so far |
| FavouriteCategory | Character | Most bought category of items from customer  Available categories: Books, Electronics, Sporting Goods, Home Goods, Clothing |
| LastPurchaseDate | Date | Most recent date that the customer brought items from shop |
| PreferredPaymentMethod | Character | Preferred payment method by customers  Available: Paypal, DebitCard, CreditCard, Cash |
| Churn | Binary | Whether the customer has churned or not |

| **Customer\_marketing** | | |
| --- | --- | --- |
| **Parameters** | **Type** | **Description** |
| CustomerID | ID | The ID of each customer |
| EmailSubscription | Binary | Whether the customer has subscribed to newsletter using email address |
| PhoneSubscription | Binary | Whether the customer has subscribed to newsletter using phone number |
| CustomerSatisfaction | Ordinal | Ranked from 1 to 5, with 5 being the best, the quality of service |

In total, there are 11 attributes in customer data and 4 attributes in customer marketing. Both dataset have 500 observations and they can be joined using CustomerID.

Our target in this analysis is to predict whether the customer will churn or not hence our target variable is in customer data, which is “Churn”. With the introduction of new data entries from customer marketing, the two datasets present here will give us insight into what could possibly cause a customer to churn.

# **Data Pre-Processing**

In this section, data preprocessing will be carried out. It is essential for the dataset to remove any potential errors that may be present in the original dataset, such as missing values, inconsistencies, and formatting errors. This process also involves combining two datasets (customer\_data & customer\_marketing) to generate a complete dataset for our analysis in the next few stages.

First, we combine the dataset. This is done using Talend Data Integration. Upon starting up the program, we can use the search function located at the top right of the window to look for our required functions. The functions involved are “tFileInputDelimited”, “tMap”, and “tFileOutputDelimited”.

Two “tFileInputDelimited” function is first dragged into the workspace. Both datasets are loaded into each node. The field separator is modified from “;” to “,” because the csv file is separated using comma instead of semicolon. After that, the schema is built for both datasets, which is just mapping each attribute into the schema, and the job is ready to run.

After loading the datasets into the workspace, we can combine them using the “tMap” function. Before running the job, we check the map editor and mapped the attributes that we want to combine. The desired output should be like the diagram shown below, where the individual rows of attributes from “row1” is dragged manually to “out1” located on the right side of the editor. To join both datasets, we can use “row1.CustomerID” as the key to join the attributes located in “row2”. Hence, we can add the attributes from “row2” to “out1” respectively.

A screenshot of a computer

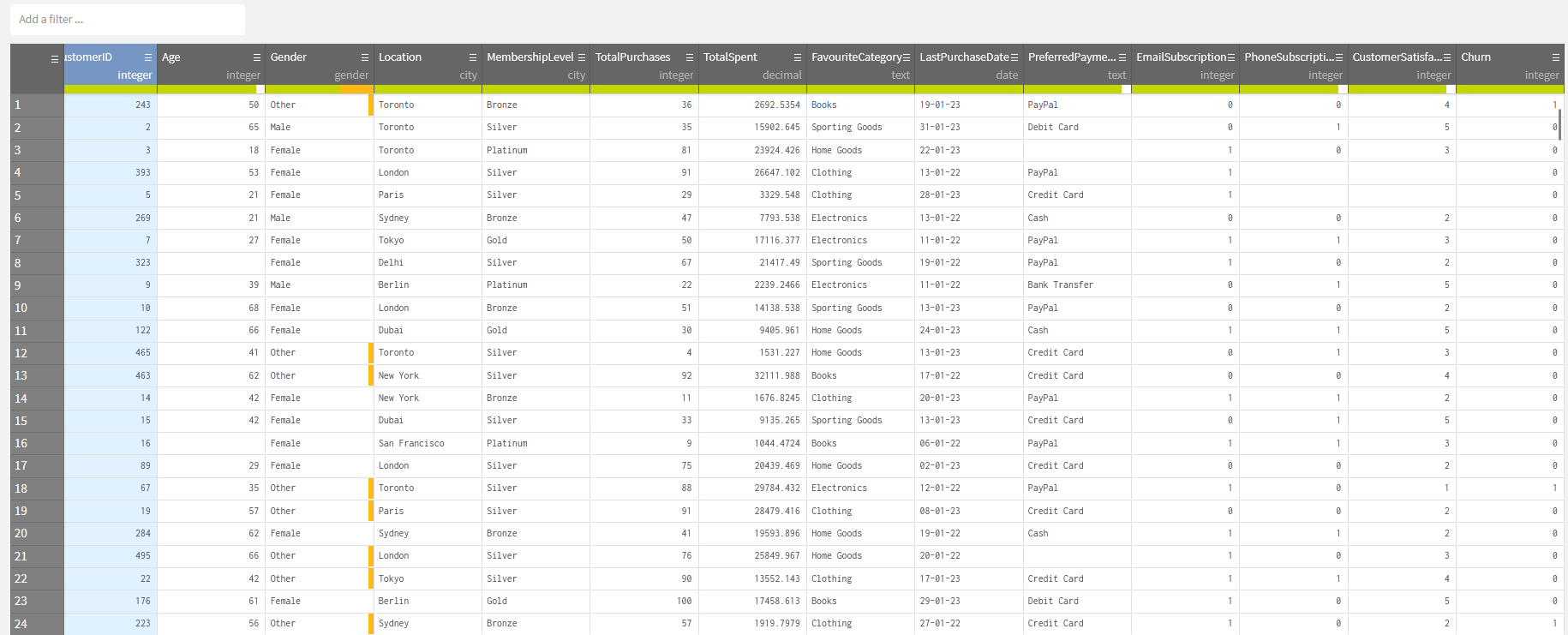
Description automatically generated

After mapping the attributes, we add the last function into the workspace, which is “tFileOutputDelimited”, which allows us to generate a new csv file with the updated mapped function from “tMap”. The final workspace diagram is shown below and the new csv file is generated in the local machine

A screenshot of a computer

Description automatically generated

The next step of preprocessing is to check for any errors in our newly unified dataset. This can be done using Talend Data Preparation. The dataset is loaded into the workspace and an overview is constructed.



It is observed that there are a few missing values located in some observations, particularly in Age, PreferredPaymentMethod, PhoneSubscription, and CustomerSatisfaction. For numerical variables like Age, we can impute the missing values with its median, that way we can safely inject new values into the column without skewing the existing distribution of the variable. Upon inspection, there are a total of 12 missing values in Age, hence all 12 of them is imputed with the median value of 44. The “Gender” errors are neglected because some countries do have “Others” as an option in the Gender information.

A screenshot of a number

Description automatically generated A screenshot of a computer

Description automatically generated

For PreferredPaymentMethod, there are a few missing values present in the variable, however it is categorical in nature and we can’t impute through mathematical means. We can use the mode however the distribution may be skewed as the dataset is not big enough for the mode imputation to be negligible. Hence, a new level is introduced. If there are missing values in PreferredPaymentMethod, we just impute it with “Any”, which means that the customer have no preference and are good with just any type of payment method.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated A screenshot of a phone

Description automatically generated

Next up is the missing values that was carried forward from the customer\_marketing dataset, which is PhoneSubscription and CustomerSatisfaction. Since we have the dataset now, the marketing campaign has most likely ended and we can’t follow up with the customer anymore. Hence, we can assume that the missing values in PhoneSubscription can be imputed as 0, meaning that they didn’t subscribe to newsletters with phone number.

As for the CustomerSatisfaction, the variable is ranked in an ordinal scale, with 5 being the best and 1 being the worst. The median will be used to impute missing values for the satisfactory score because it does not skew the distribution. The value of median is 3, which is the average customer experience.

A screenshot of a number

Description automatically generated

A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated

We also noticed that when PhoneSubscription is missing, the CustomerSatisfaction will also be missing. This is most likely because the marketing campaign is carried out remotely and the information for CustomerSatisfaction is generated using an application. Hence both missing values just means that the customer either does not have a smartphone or has not downloaded the app.

Finally, we reduce the number of decimal points located in TotalSpent, this is to ensure consistencies with the format of currency that typically contains just 2 decimal points. We can do this using the round value function and choose 2 precision points.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated

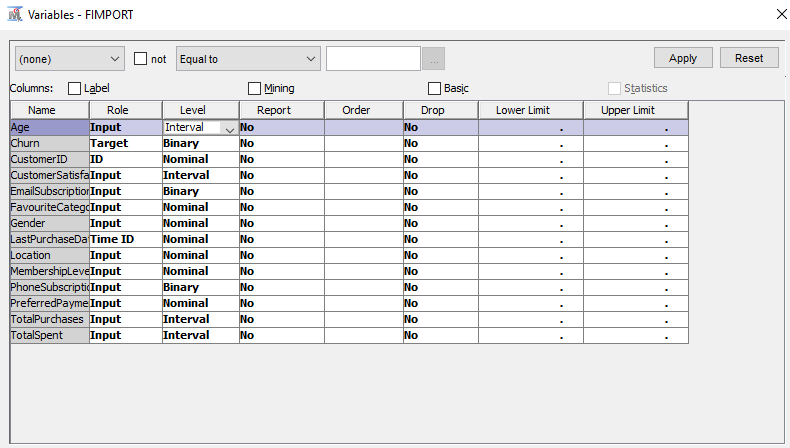
With that, we conclude our preprocessing phase of the case study. The final workflow in Talend Data Preparation is shown below. Basically, we have just made adjustments to impute missing values without the need of removing said observations, we also performed some adjustments in making sure the format is standardized such as the currency format. The final dataset is generated and will be ready for the next phase of the case study.

A screenshot of a computer

Description automatically generated

# **Importing Dataset into SAS**

The cleaned dataset is then imported into SAS Enterprise Miner using the File Import node. The variables are then examined using the “Edit Variables” option. It is found that we have to set “Churn” as our target variable and designate “CustomerID” as ID so that it won’t consider as an input in our modelling later on. The LastPurchaseDate was set to TimeID too because it contains dates. It is also observed that some variables have wrong levels such as PhoneSubscription, EmailSubscription, and Churn. Hence these 3 variables are set to “Binary” level because the attributes only have 0 and 1 as observations. The updated version will look like this.



We performed simple exploration in our dataset to visualize the distribution of each attribute in our dataset and found that most of them are fairly distributed across all the available categories and observations. The only probable issue was the imbalance found in the “Churn” variable.

A screenshot of a computer

Description automatically generated

After that, we add the Data Partition Node into the workspace. This is to split the dataset into train, validate and test dataset. In this case study, we only use the train and validate dataset, hence the splitting ratio will be set to 70:30 respectively, and it is ready for modelling.

A screenshot of a computer

Description automatically generated

# **Decision Tree Analysis**

First, a decision tree node was used to analyse churn behaviour of customers. After running the node, the tree map is visualized using an interactive decision tree. The tree is splitted slowly according to which variable that gives the most information gain when splitting. The optimally splitted map is shown below.

A computer screen shot of a diagram

Description automatically generated

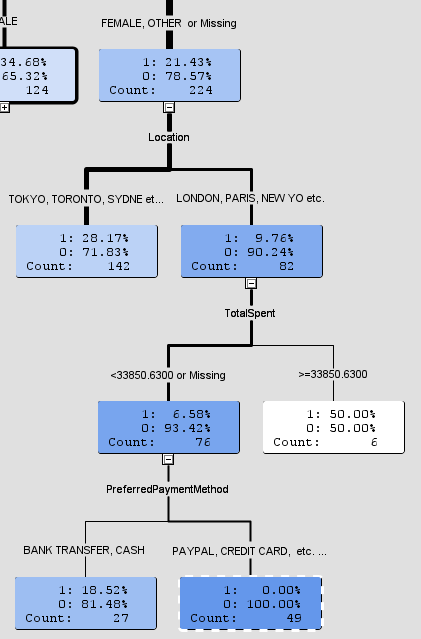
In general, the main demographic who don’t churn mainly consists of female shoppers who holds a silver or above membership level. In their most recent marketing campaign, these group of female shoppers most likely left a positive review and will be most likely to continue shopping at said establishments. As for the males, it depends whether they subscribed to newsletter with their phone numbers or not, if they don’t, generally they will give a more positive customer satisfactory score and retain as a customer.

If we pruned the branches after male branch and split them using the variable with 2nd highest information gain, which is “EmailSubscription” and “FavouriteCategory”, it is observed that male shoppers who don’t subscribe to newsletters with phone will not likely be subscribed using their emails too. It is also found that those who do subscribe to newsletter with their phones, tend to be interested in products like books, clothes, and home goods. However, there are also a handful of them interested in sporting goods and electronics too.

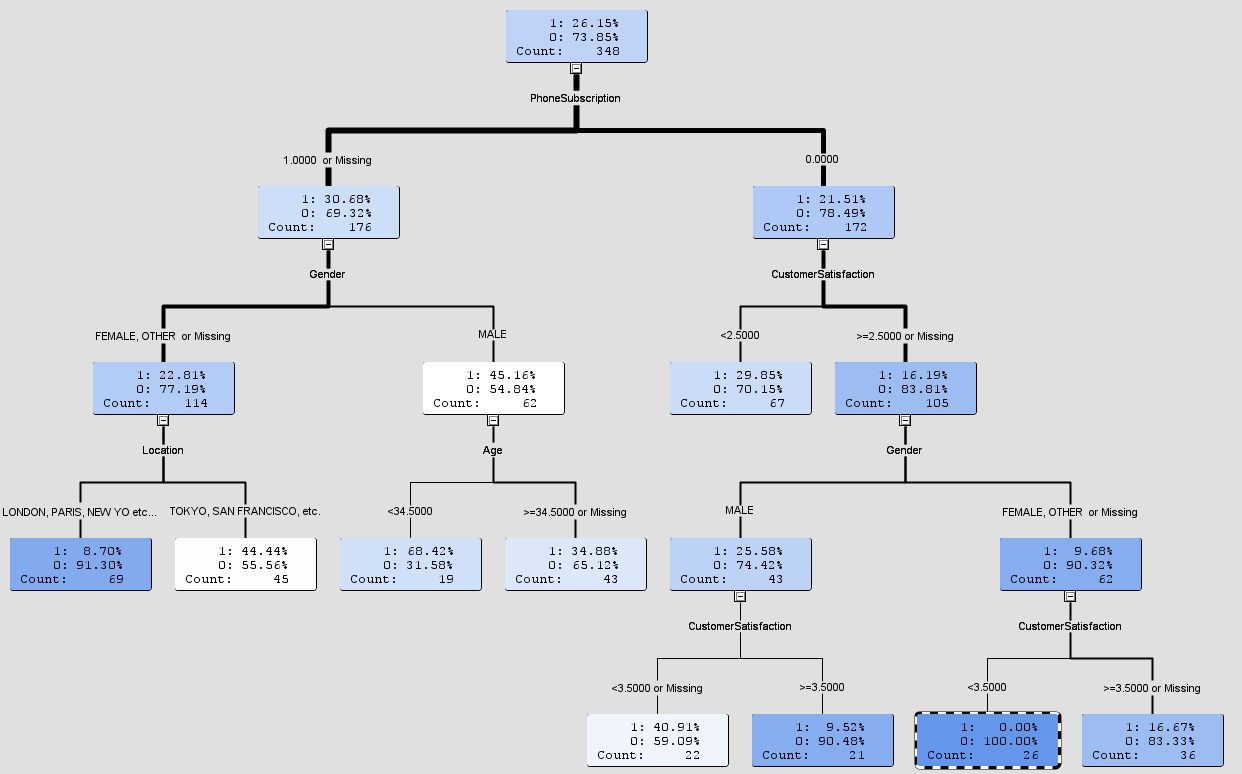
A diagram of a number

Description automatically generated

Similarly, if we pruned the branches for the females and replace it with the variable with 2nd highest information gain, it is evident that generally European female shoppers don’t churn as much as other female shoppers do in other regions. Most female shoppers in the European spend no more than 34 thousand dollars in purchases, which is the average consumption per customer due to the average income of the continent. Among those who have spent less than 34 thousand dollars so far and didn’t churn, their preferred method is usually cashless, which is PayPal and credit cards.



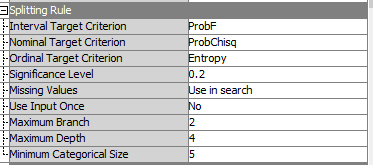
If we were to start back from the root and change the “Gender” splitting criteria to “PhoneSubscription”, which is the 2nd highest worth in information gain, the new treemap will look something like this.



It is inferred that those who subscribe using their phones are generally female shoppers located in European locations such as London, Paris etc. And for the males, it only gives insights about their age since most male subscribers are aged 34 above.

On the other hand, for shoppers who didn’t subscribe to newsletter with their phones has an above average customer satisfactory score, with female shoppers being the one who contributes most to that score, however those who don’t churn will give a score in between of 2.5 and 3.5, despite having high satisfactory score, some female shoppers end up churning. For the men, the pattern is obvious as the higher score they give, the more likely they will stay, or else they will churn, even when they gave a score which is deemed “above average”.

With the usage of the interactive decision tree map, not only we are able to mind insights from the tree map, but it also enables us to determine the number of branches and leaves that are suitable for the decision tree model for the most optimal results. Hence, in our decision tree model building, we set the maximum branch that the tree model could split into 2, and the depth is set to 4.



Another decision tree model is constructed, but the maximum branch is set to 3 to provide a more complex algorithm that can capture extra nuances within the data. The average square error for both tree models is shown below.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Both models’ performance is adequate to predict churn rate of customers. However, the tree models can smoothen out with the use of bagging and boosting method, which will be explored in the next section.

# **Gradient Boosting & Random Forest**

The Gradient Boosting node is then added to the workspace, and the model's depth parameter is increased to 4, resulting in a more complex network of weak learners than the default depth of 2. This change is intended to capture subtle patterns in the data and improve the model's forecasting skills for the specific characteristics of the problem at hand.

A screenshot of a computer

Description automatically generated

As a result, when paired with the ensemble technique, the two trees contribute to the model's decision-making process by utilizing their diverse views to create more accurate predictions. The random forest technique allows for a bagging technique for the decision tree models.

A screenshot of a computer

Description automatically generated

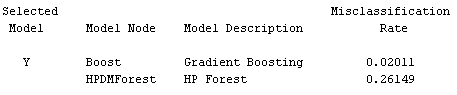
The final workflow diagram looks something like this, with the included “Model Comparison” node added later during model assessment.

A diagram of a flowchart

Description automatically generated

# **Model Comparison**

In this stage, the models constructed are compared with each other and the “Model Comparison” is used here. The results output that Gradient Boosting method performs the best with only having a misclassification rate of 0.02%, generally the lower the rate, the better. The score is followed by random forest which has a 0.26% misclassification rate, which is still acceptable.



Not only does the bagging method have a lower score than boosting method, but it also slightly increases the misclassification rate of the decision tree models.

Hence, the recommended model for predicting churn rate of customers is by Gradient Boosting for this dataset, however more models need to be assessed when the size of dataset grows and possibly more modifications to parameters needed to be done.

# **Conclusion**

As a conclusion, based on the decision tree analysis, several business strategies can be adopted to boost sales and prolong customer retention.

1. **Targeted Marketing Campaigns**:

* More marketing campaigns can be targeted for female shoppers who holds a silver membership, this not only encourage existing customer retention, but also attracts potential new customers who may sign up for the membership, enticing them to achieve silver level or above to enjoy the benefits during the campaign.

1. **Implement Region-Specific Promotions**

* Since many who don’t churn resides in the European region, the company can offer region specific promotions and offers that may increases sales in populated areas within Europe, such as having a sale on winter jacket during Winter season or promote local Europe cuisines in the establishments.

1. **Cashless Payments Cashbacks**

* With cashless payment being the preferred method of payment by most, the company can encourage more users to shop by offering cashbacks to those who shopped until a certain threshold in a single receipt. This not only encourages users to use cashless payments more, but also encourages sales.

With these methods in mind, the company is presented with several actionable options for preventing churn from customers and encouraging customer growth. The business can address certain customer groups proactively by utilizing customized marketing efforts and adapting regional considerations. The company not only retain existing consumers but also pave the road for sustainable customer growth and improved overall business performance.