**Name: Tneu Zhen Yang**
**ID: 22081192**
**<u>WQD7005 Alternative Assessment 1</u>**
**<u>Case Study - E-Commerce Customer Behavior Analysis</u>**

Github link: [https://github.com/zyang1611/WQD7005_AA](https://github.com/zyang1611/WQD7005_AA)
SAS link: [SAS]()

### Case Study Objective and Data Source

The objective of this case study is to apply machine learning techniques such as decision trees and ensemble learning techniques to predict if e-commerce customers will churn and investigate what factors are more likely to cause a customer to churn. Insights gained from such a case study can provide useful insights to e-commerces businesses and help them retain customers which is essential for the businesses continued growth and profitability.

The dataset used as a starting point for this case study in this was sourced from Kaggle.
*https://www.kaggle.com/datasets/shriyashjagtap/e-commerce-customer-for-behavior-analysis/data?select=ecommerce_customer_data_large.csv*
However, the dataset was slightly modified and adapted to suit the objectives of this case study. All manipulations and transformations that were conducted on the dataset are documented below.

### Part 1: Data Import and Preprocessing

The unedited dataset from Kaggle was first imported into Talend Data Preparation. The goal of this was to explore and gain an understanding of the dataset.

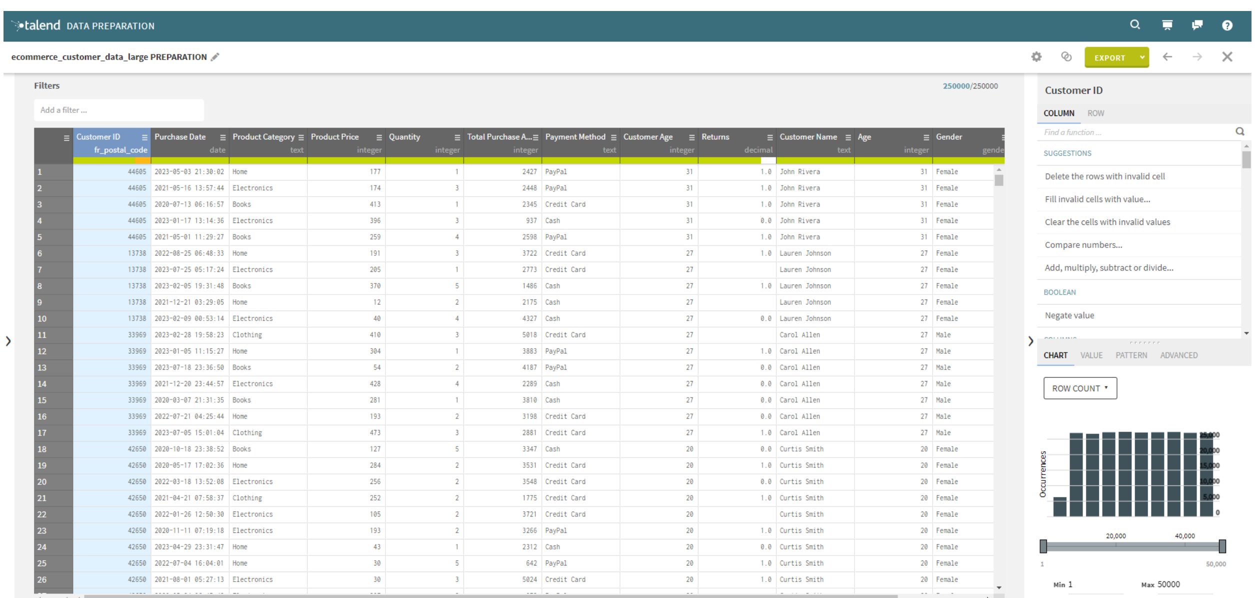



**Figure 1:** Screenshot of Talend Data Preparation on Kaggle Dataset.

It was found that the dataset consisted of a list of transactions with the attributes listed below.

**Table 1:** Attributes found in the Kaggle Dataset.

| | |
|---|---|
| Customer ID | A unique identifier for each customer |
| Customer Name | The name of the customer |
| Customer Age | The age of the customer |
| Gender | The gender of the customer |
| Age | The age of the customer |
| Purchase Date | The date of each purchase made by the customer. |
| Product Category | The category or type of the purchased product. |
| Product Price | The price of the purchased product. |
| Quantity | The quantity of the product purchased. |
| Total Purchase Amount | The total amount spent by the customer in each transaction |
| Payment Method | The method of payment used by the customer (e.g., credit card, PayPal) |
| Returns | Whether the customer returned any products from the order (binary: 0 for no return, 1 for return) |
| Churn | A binary column indicating whether the customer has churned (0 for retained, 1 for churned) |

The first thing we notice is that there are two columns that represent the age of the customer. Also we see that the dataset is a list of transactions but also contains details such as customer name, age and gender which in a real-world scenario would belong in another table that represents customer data. Furthermore, the dataset is aimed at predicting customer churn but does not contain typical features used for churn prediction such as Customer Lifetime Value (CLTV), Rate of Returns, Average Transaction Value or Tenure. The next steps will detail how the dataset is adjusted to reflect a real world e-commerce customer and transaction database as well as how a churn prediction dataset is created.

The Kaggle dataset was first split into two separate tables of *customers* and *transactions* to better reflect real-world operating conditions. *customers* is a table representing a list of customers and has a unique identifier for each customer while *transactions* is a table representing a list of transactions and has a unique identifier for each transaction.

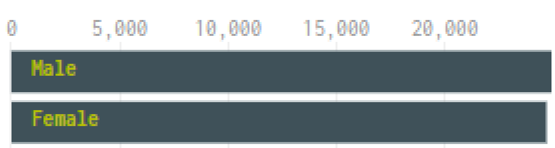
Both tables were then imported into Talend Data Preparation for data profiling, data cleaning and data validation.

## Data Profiling and Validation

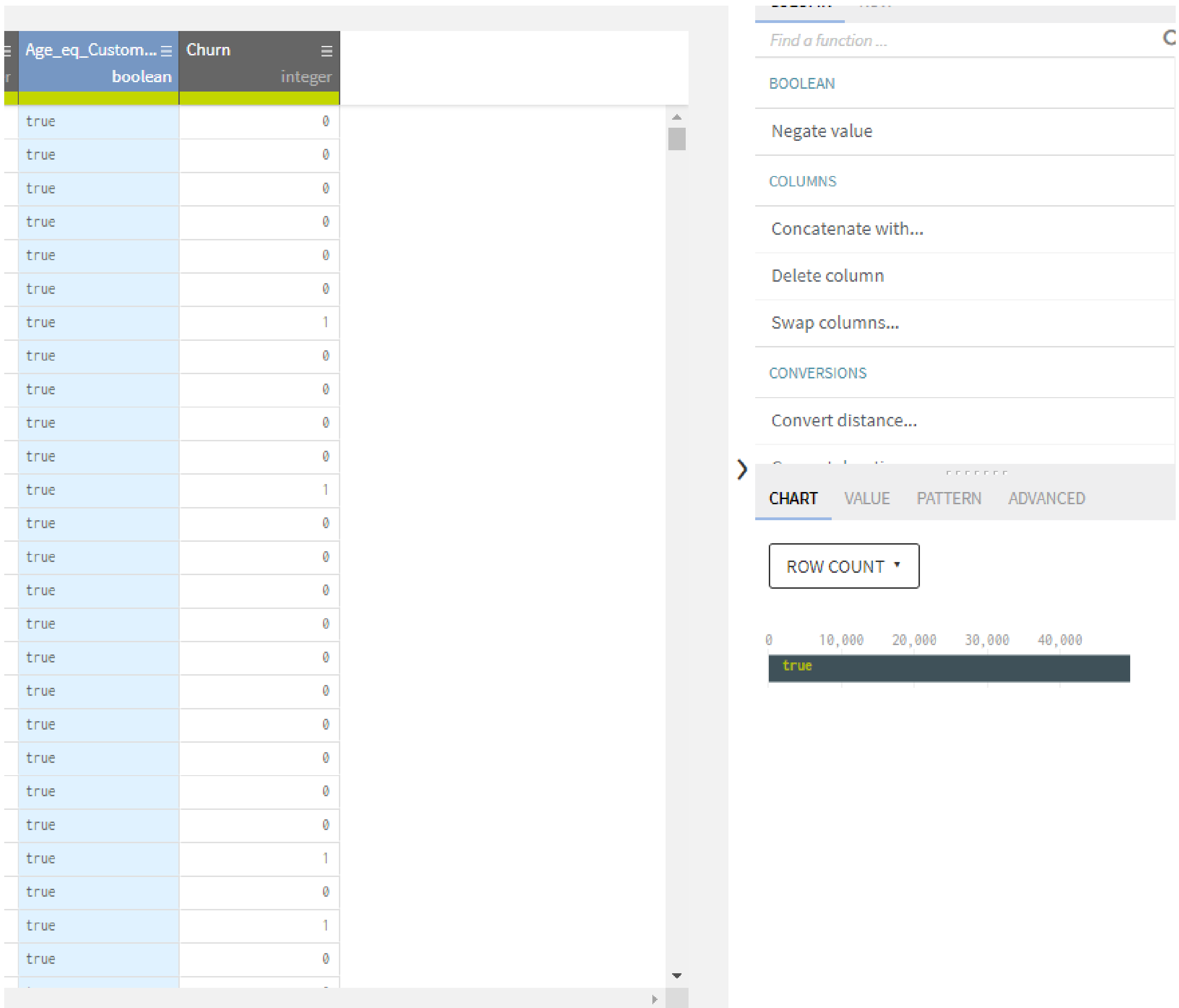
***customers***

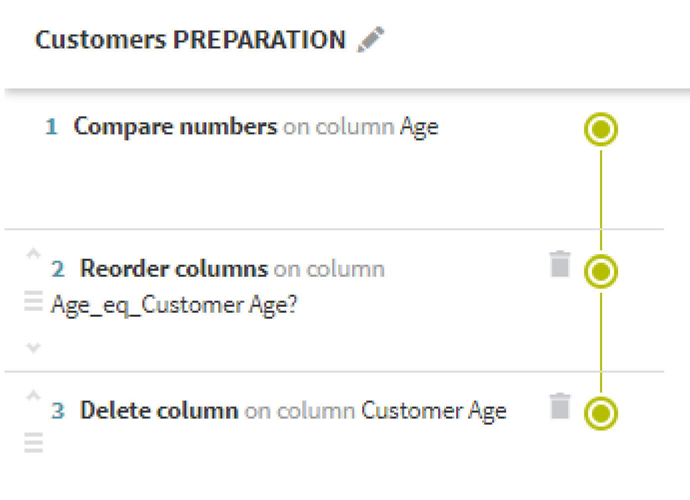
| | Customer ID fr_postal_code | Customer Name text | Customer Age integer | Age integer | Gender gender | Churn integer |
|---|---|---|---|---|---|---|
| 1 | 44605 | John Rivera | 31 | 31 | Female | 0 |
| 2 | 13738 | Lauren Johnson | 27 | 27 | Female | 0 |
| 3 | 33969 | Carol Allen | 27 | 27 | Male | 0 |
| 4 | 42650 | Curtis Smith | 20 | 20 | Female | 0 |
| 5 | 24053 | Jose Green | 53 | 53 | Male | 0 |
| 6 | 19676 | Linda Lee | 57 | 57 | Male | 0 |
| 7 | 19917 | Joshua Davis | 34 | 34 | Male | 1 |
| 8 | 16051 | Ellen Kent | 45 | 45 | Male | 0 |
| 9 | 23734 | Cheryl Espinoza | 18 | 18 | Male | 0 |
| 10 | 16921 | Cheyenne James | 54 | 54 | Male | 0 |
| 11 | 7796 | Lisa Dennis | 21 | 21 | Male | 0 |
| 12 | 21035 | Peter Watson | 50 | 50 | Female | 1 |
| 13 | 37125 | Terry Campos | 24 | 24 | Female | 0 |
| 14 | 41840 | Ronald Berger | 48 | 48 | Male | 0 |
| 15 | 2642 | Ian Kelley | 67 | 67 | Female | 0 |
| 16 | 1254 | Mr. David Morgan | 70 | 70 | Male | 0 |
| 17 | 13389 | Monica Ramos | 51 | 51 | Male | 0 |
| 18 | 24473 | Nicole Lewis | 38 | 38 | Female | 0 |
| 19 | 11230 | Rebecca Boyer | 23 | 23 | Male | 0 |
| 20 | 24741 | Steven Mcmillan | 50 | 50 | Female | 0 |
| 21 | 10191 | Walter Griffin | 66 | 66 | Male | 0 |
| 22 | 6829 | Chelsea Williams | 43 | 43 | Male | 0 |
| 23 | 16825 | Melissa Cabrera | 37 | 37 | Male | 1 |
| 24 | 18467 | Terri Carter | 23 | 23 | Male | 0 |
| 25 | 34857 | Alexander Morse | 57 | 57 | Male | 1 |
| 26 | 21801 | Anne Leblanc | 19 | 19 | Female | 0 |

Gender column was checked for spelling errors as well as distribution between males and females.

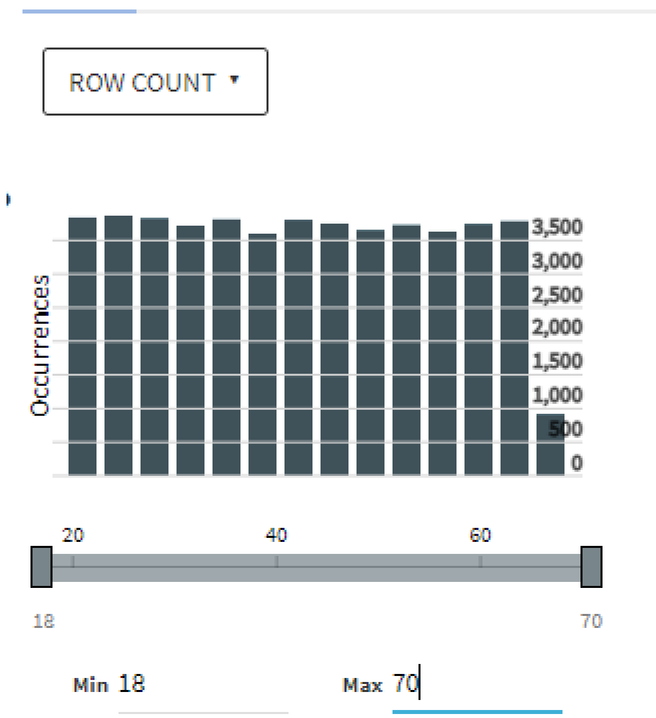



There are two columns that represent age, they were checked for equality then Customer Age was dropped.

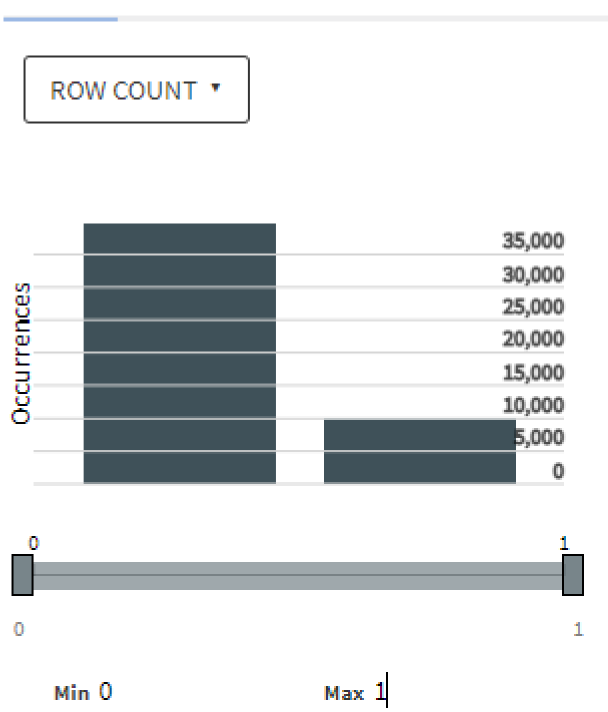


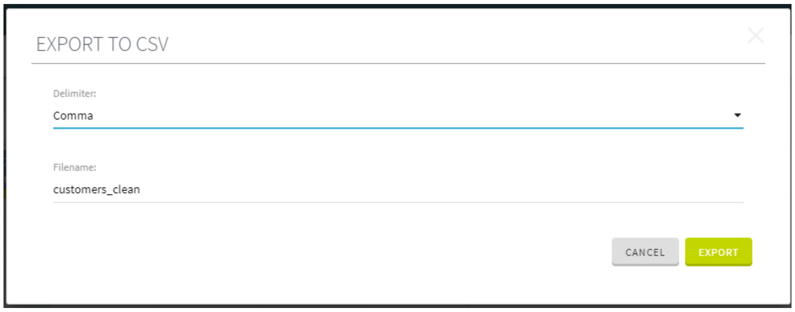

Next, the distribution of variables Age and Churn were investigated.




Age distribution.




Churn distribution is quite **imbalanced.**



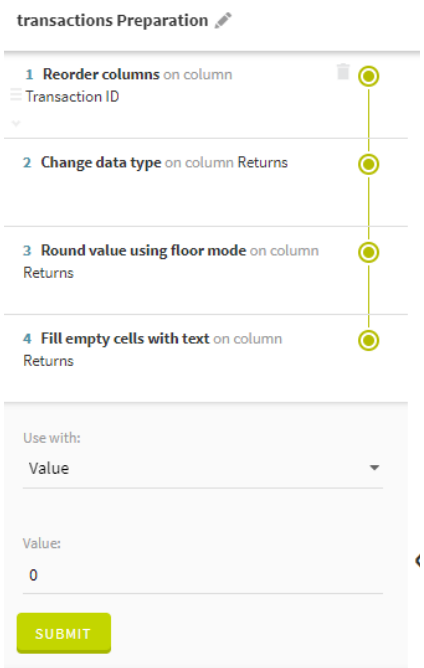

Cleaned data was exported to a csv file.

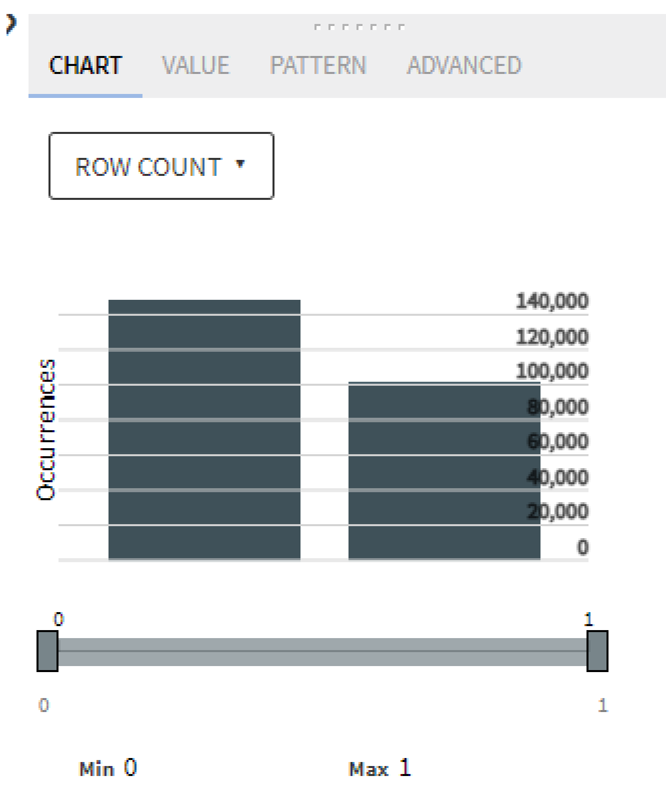
## transactions

| | Transaction ID integer | Customer ID fr_postal_code | Purchase Date date | Product Category text | Product Price integer | Quantity integer | Total Purchase A... integer | Payment Method text | Returns integer |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 100 | 44605 | 2023-05-03 21:30:02 | Home | 177 | 1 | 2427 | PayPal | 1 |
| 2 | 101 | 44605 | 2021-05-16 13:57:44 | Electronics | 174 | 3 | 2448 | PayPal | 1 |
| 3 | 102 | 44605 | 2020-07-13 06:16:57 | Books | 413 | 1 | 2345 | Credit Card | 1 |
| 4 | 103 | 44605 | 2023-01-17 13:14:36 | Electronics | 396 | 3 | 937 | Cash | 0 |
| 5 | 104 | 44605 | 2021-05-01 11:29:27 | Books | 259 | 4 | 2598 | PayPal | 1 |
| 6 | 105 | 13738 | 2022-08-25 06:48:33 | Home | 191 | 3 | 3722 | Credit Card | 1 |
| 7 | 106 | 13738 | 2023-07-25 05:17:24 | Electronics | 205 | 1 | 2773 | Credit Card | 2 |
| 8 | 107 | 13738 | 2023-02-05 19:31:48 | Books | 370 | 5 | 1486 | Cash | 1 |
| 9 | 108 | 13738 | 2021-12-21 03:29:05 | Home | 12 | 2 | 2175 | Cash | 2 |
| 10 | 109 | 13738 | 2023-02-09 00:53:14 | Electronics | 40 | 4 | 4327 | Cash | 0 |
| 11 | 110 | 33969 | 2023-02-28 19:58:23 | Clothing | 410 | 3 | 5018 | Credit Card | 2 |
| 12 | 111 | 33969 | 2023-01-05 11:15:27 | Home | 304 | 1 | 3883 | PayPal | 1 |
| 13 | 112 | 33969 | 2023-07-18 23:36:50 | Books | 54 | 2 | 4187 | PayPal | 0 |
| 14 | 113 | 33969 | 2021-12-20 23:44:57 | Electronics | 428 | 4 | 2289 | Cash | 0 |
| 15 | 114 | 33969 | 2020-03-07 21:31:35 | Books | 281 | 1 | 3810 | Cash | 0 |
| 16 | 115 | 33969 | 2022-07-21 04:25:44 | Home | 193 | 2 | 3198 | Credit Card | 0 |
| 17 | 116 | 33969 | 2023-07-05 15:01:04 | Clothing | 473 | 3 | 2881 | Credit Card | 1 |
| 18 | 117 | 42650 | 2020-10-18 23:38:52 | Books | 127 | 5 | 3347 | Cash | 0 |
| 19 | 118 | 42650 | 2020-05-17 17:02:36 | Home | 284 | 2 | 3531 | Credit Card | 1 |
| 20 | 119 | 42650 | 2022-03-18 13:52:08 | Electronics | 256 | 2 | 3548 | Credit Card | 0 |
| 21 | 120 | 42650 | 2021-04-21 07:58:37 | Clothing | 252 | 2 | 1775 | Credit Card | 1 |
| 22 | 121 | 42650 | 2022-01-26 12:50:30 | Electronics | 105 | 2 | 3721 | Credit Card | 2 |
| 23 | 122 | 42650 | 2020-11-11 07:19:18 | Electronics | 193 | 2 | 3266 | PayPal | 1 |
| 24 | 123 | 42650 | 2023-04-29 23:31:47 | Home | 43 | 1 | 2312 | Cash | 0 |
| 25 | 124 | 42650 | 2022-07-04 16:04:01 | Home | 30 | 5 | 642 | PayPal | 1 |
| 26 | 125 | 42650 | 2021-08-01 05:27:13 | Electronics | 30 | 3 | 5024 | Credit Card | 1 |

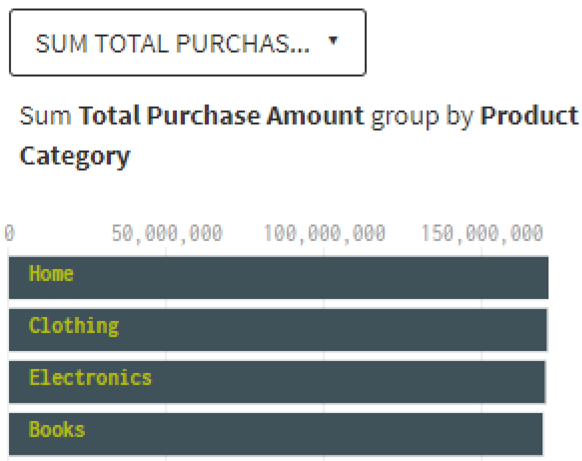
**Data Imputation**

It was found that the Returns column contained missing values, values were imputed with 0 using the procedure below. Decimal data type was also converted to integer type for the Return column.

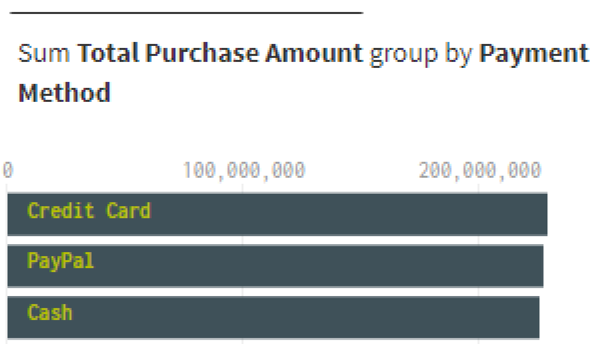


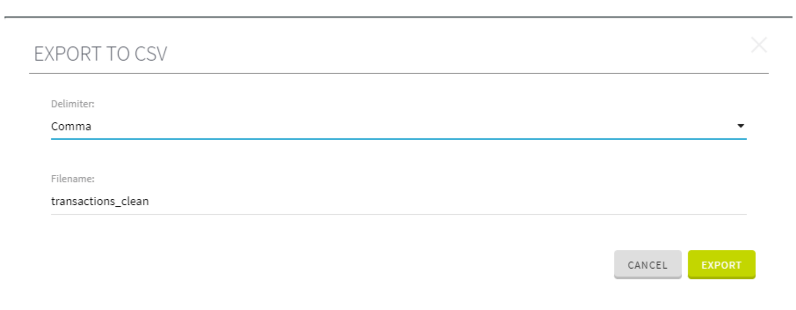
Returns distribution.




Explore Total Purchase Amount by Product Category.




Explore Total Purchase Amount by Payment Method.




Cleaned data was exported to a csv file.

### Talend Open Studio for Data Integration

The two separate tables customers and transactions were joined using Talend Open Studio to create a raw combined dataset. The output was saved to a csv file.

Import data into Talend Open Studio and define schema.

Schema of tFileInputDelimited_1

tFileInputDelimited_1

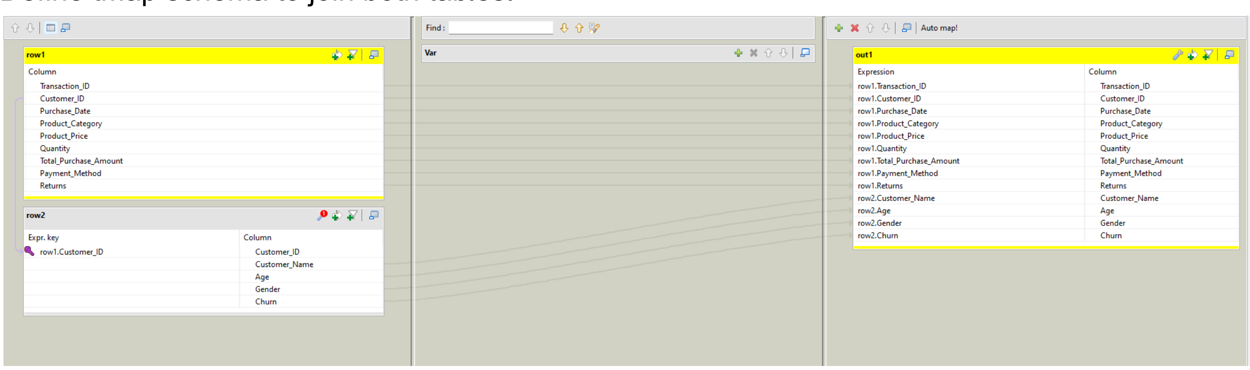
| Column | Key | Type | N.. | Date Pattern (Ct... | Length | Precision | Default | Comment |
|---|---|---|---|---|---|---|---|---|
| Customer_ID | ☐ | String | ☑ | | | | | |
| Customer_Name | ☐ | String | ☑ | | | | | |
| Age | ☐ | Integer | ☑ | | | | | |
| Gender | ☐ | String | ☑ | | | | | |
| Churn | ☐ | Integer | ☑ | | | | | |

Schema of tFileInputDelimited_2

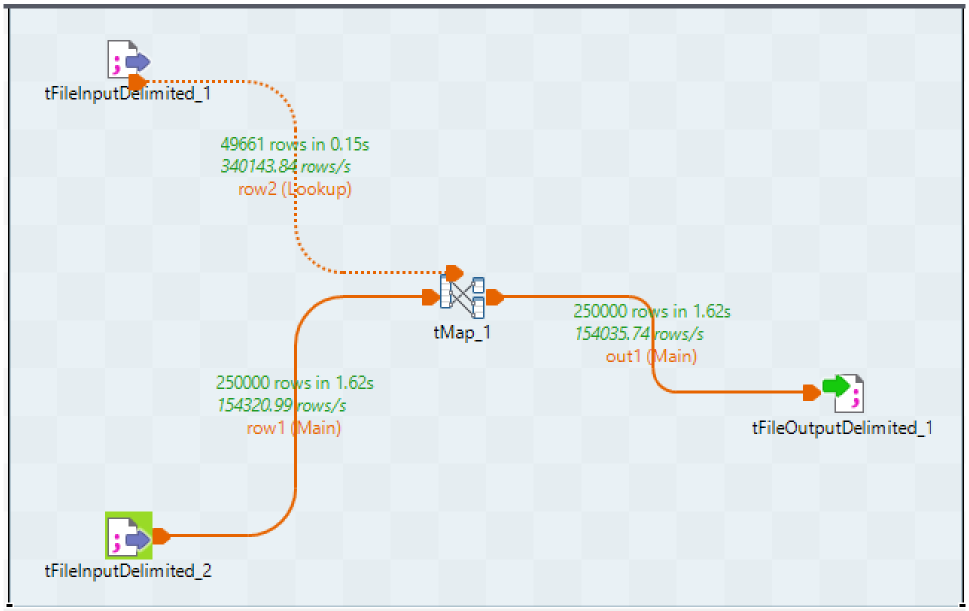
tFileInputDelimited_2

| Column | Key | Type | N.. | Date Pattern (C... | Length | Precision | Default | Comm... |
|---|---|---|---|---|---|---|---|---|
| Transaction_ID | ☐ | String | ☑ | | | | | |
| Customer_ID | ☐ | String | ☑ | | | | | |
| Purchase_Date | ☐ | Date | ☑ | "M/D/YYYY H... | | | | |
| Product_Category | ☐ | String | ☑ | | | | | |
| Product_Price | ☐ | Float | ☑ | | | | | |
| Quantity | ☐ | Integer | ☑ | | | | | |
| Total_Purchase_Amount | ☐ | Float | ☑ | | | | | |
| Payment_Method | ☐ | String | ☑ | | | | | |
| Returns | ☐ | Integer | ☑ | | | | | |

Define tMap schema to join both tables.



Job design diagram.




**Feature engineering for churn analysis**

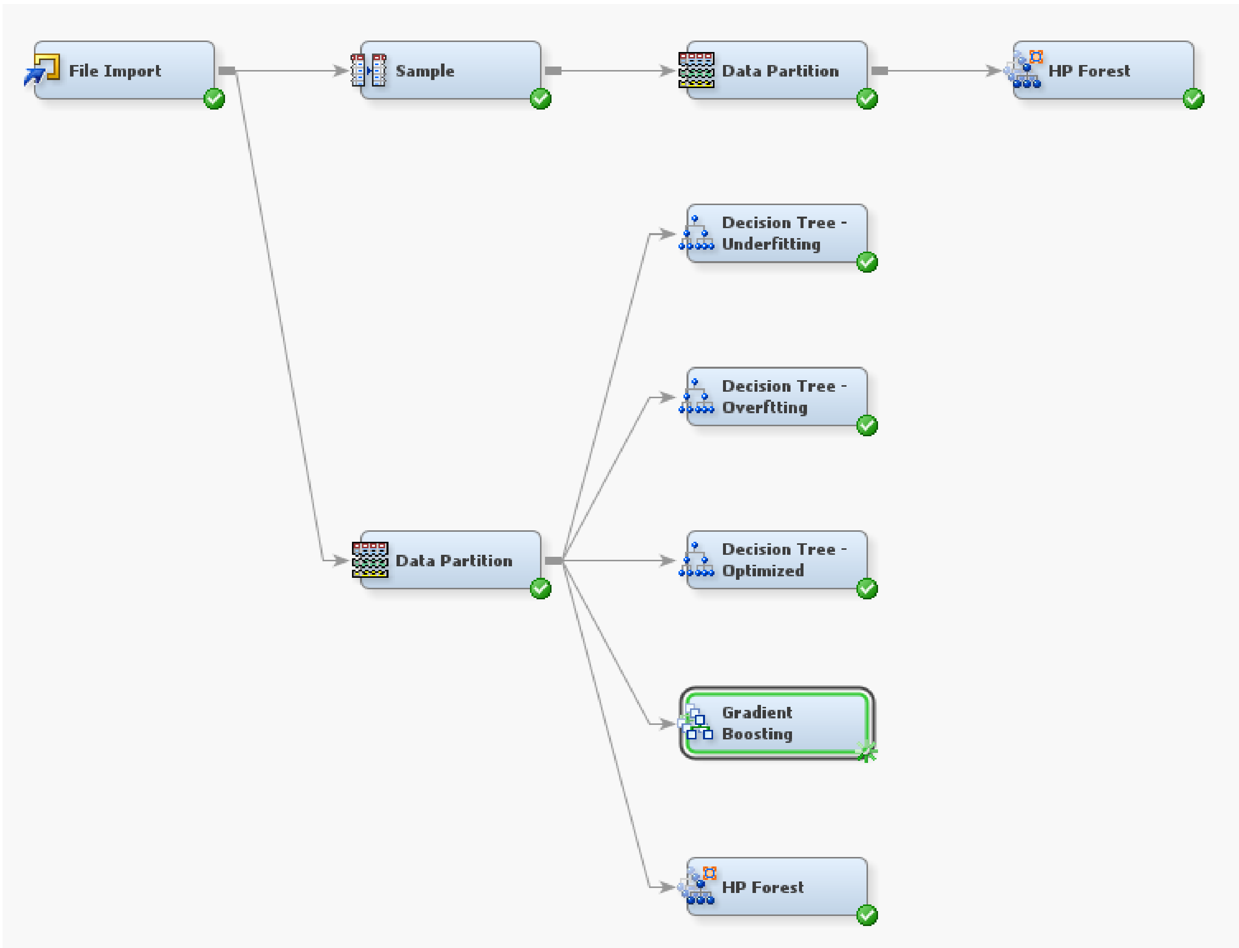
Python was used to conduct feature engineering and create the final churn_analysis dataset with attributes as below.

| | |
|---|---|
| Customer_ID | A unique identifier for each customer. |
| Age | The age of the customer |
| Gender | The gender of the customer |
| Days_Since_Last_Purchase | The number of days since the customer last made a purchase |
| CLTV | Customer Lifetime Value. Estimated as the total amount spent by the customer to date. |
| Avg_Txn_Value | The average spend per transaction for each customer. |
| Fav_Category | The customer’s most frequently purchased category. |
| Fav_Payment_Method | The customer’s most frequently used payment method. |
| Return Rate | The ratio of transactions with returns to total transactions made by the customer |

| | |
|---|---|
| Churn | A binary column indicating whether the customer has churned (0 for retained, 1 for churned) |

## <u>Part 2: SAS Enterprise Miner</u>

**Overall Diagram**




The diagram used in SAS EM is shown above. Each node including decision trees, underfitting and overfitting demonstrations, bagging and boosting will be shown and discussed below.

| Name | Role | Level | Report | Order | Drop | Lower Limit | Upper Limit |
|---|---|---|---|---|---|---|---|
| Age | Input | Interval | No | | No | . | . |
| Avg_Txn_Value | Input | Interval | No | | No | . | . |
| CLTV | Input | Interval | No | | No | . | . |
| Churn | Target | Interval | No | | No | . | . |
| Customer_ID | Rejected | Nominal | No | | No | . | . |
| Days_Since_Las | Input | Interval | No | | No | . | . |
| Fav_Category | Input | Nominal | No | | No | . | . |
| Fav_Payment_M | Input | Nominal | No | | No | . | . |
| Gender | Input | Nominal | No | | No | . | . |
| Return_Rate | Input | Interval | No | | No | . | . |

Define roles and levels for input and target variables.

## Part 3: Decision Tree Analysis

Decision trees are a supervising learning algorithm that makes decisions by splitting data into branches resembling a tree. Each branch is then further split recursively, until a stopping criteria is met. Decision trees have the advantage of being easily interpretable but are prone to overfitting and tend to low overall performance. For the purpose of this case study, both overfitting and underfitting will be demonstrated.

### Decision Tree Underfitting

Decision trees underfit data when the model is too simple to learn complex patterns in the data, in a similar way to overfitting, this results in poor model performance. Underfitting commonly occurs when decision trees are too shallow or pruning is conducted too aggressively. To demonstrate underfitting, the maximum depth of 3 was set in SAS EM. Since this dataset is imbalanced as demonstrated in the Data Profiling stage earlier, we investigate the performance of the decision tree at predicting both outcomes for Churn as shown in the confusion matrix and report below. The accuracy of the model at predicting churn is poor as seen by the low percentage values achieved.

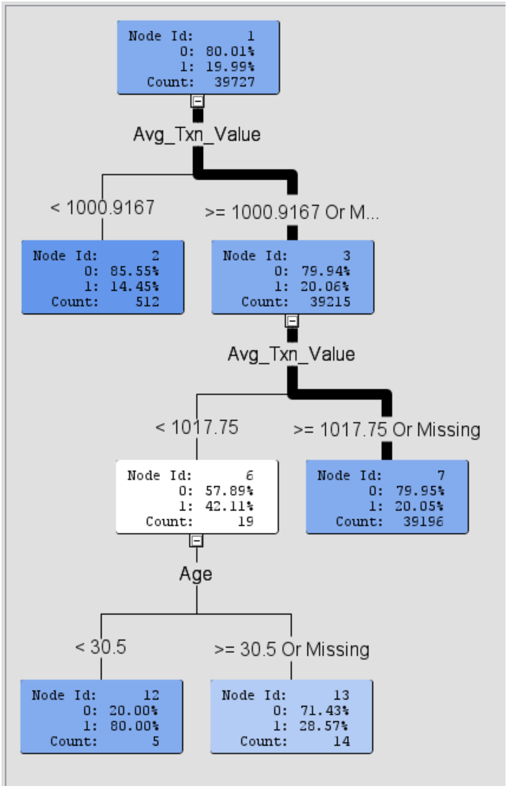
Event Classification Table

Data Role=TRAIN Target=Churn Target Label=' '

| False Negative | True Negative | False Positive | True Positive |
|---|---|---|---|
| 7938 | 31784 | 1 | 4 |

Data Role=TRAIN Target Variable=Churn Target Label=' '

| Target | Outcome | Target Percentage | Outcome Percentage | Frequency Count | Total Percentage |
|---|---|---|---|---|---|
| 0 | 0 | 80.0161 | 99.9969 | 31784 | 80.0060 |
| 1 | 0 | 19.9839 | 99.9496 | 7938 | 19.9814 |
| 0 | 1 | 20.0000 | 0.0031 | 1 | 0.0025 |
| 1 | 1 | 80.0000 | 0.0504 | 4 | 0.0101 |




**Decision Tree Overfitting**

Decision overfit the data when the tree becomes too complex, the model learns to “memorize” the training data including noise that is present in the data. This causes the model to perform worse on the test data than on the training data. In decision trees, overfitting commonly occurs when the tree is allowed to grow to excessive tree depths. To demonstrate this, a maximum tree depth of 20 was set in SAS EM. Target criterion was set to Gini for this Nominal classification task. The results show that the decision tree's performance on the test set is lower than on the training set.

| Splitting Rule | |
| --- | --- |
| Interval Target Criterion | ProbF |
| Nominal Target Criterion | Gini |
| Ordinal Target Criterion | Entropy |
| Significance Level | 0.5 |
| Missing Values | Use in search |
| Use Input Once | No |
| Maximum Branch | 2 |
| Maximum Depth | 20 |
| Minimum Categorical Size | 5 |

| Target | Target Label | Fit Statistics | Statistics Label | Train | Validation | Test |
|---|---|---|---|---|---|---|
| Churn | | _NOBS_ | Sum of Frequencies | 39727 | . | 9934 |
| Churn | | _MISC_ | Misclassification Rate | 0.175397 | . | 0.236863 |
| Churn | | _MAX_ | Maximum Absolute Err... | 0.983333 | . | 1 |
| Churn | | _SSE_ | Sum of Squared Errors | 10973.3 | . | 3634.348 |
| Churn | | _ASE_ | Average Squared Error | 0.138109 | . | 0.182925 |
| Churn | | _RASE_ | Root Average Squared... | 0.37163 | . | 0.427697 |
| Churn | | _DIV_ | Divisor for ASE | 79454 | . | 19868 |
| Churn | | _DFT_ | Total Degrees of Free... | 39727 | . | . |

Classification Table

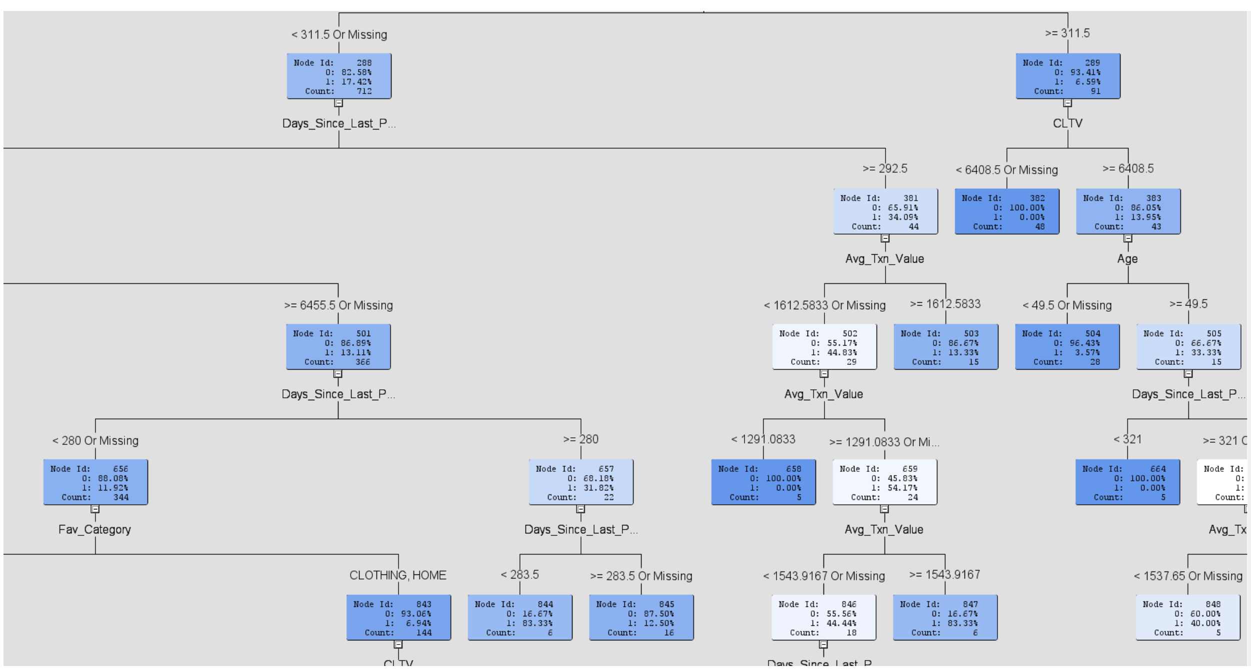
Data Role=TRAIN Target Variable=Churn Target Label=' '

| Target | Outcome | Target Percentage | Outcome Percentage | Frequency Count | Total Percentage |
|---|---|---|---|---|---|
| 0 | 0 | 85.6783 | 96.3851 | 30636 | 77.1163 |
| 1 | 0 | 14.3217 | 64.4800 | 5121 | 12.8905 |
| 0 | 1 | 28.9421 | 3.6149 | 1149 | 2.8922 |
| 1 | 1 | 71.0579 | 35.5200 | 2821 | 7.1010 |

Event Classification Table

Data Role=TRAIN Target=Churn Target Label=' '

| False Negative | True Negative | False Positive | True Positive |
|---|---|---|---|
| 5121 | 30636 | 1149 | 2821 |



Partial screenshot showing excessively deep trees.

Decision Tree Optimized

To balance overfitting and underfitting, a decision tree depth of 10 was used. Even so, performance of the model is poor

```
Classification Table

Data Role=TRAIN Target Variable=Churn Target Label=' '

                        Target      Outcome     Frequency      Total
Target     Outcome    Percentage   Percentage     Count      Percentage

  0           0         80.0539      99.9560      31771       79.9733
  1           0         19.9461      99.6726       7916       19.9260
  0           1         35.0000       0.0440         14        0.0352
  1           1         65.0000       0.3274         26        0.0654


Data Role=VALIDATE Target Variable=Churn Target Label=' '

                        Target      Outcome     Frequency      Total
Target     Outcome    Percentage   Percentage     Count      Percentage

  0           0         80.0343      99.9748       7945       79.9779
  1           0         19.9657      99.7484       1982       19.9517
  0           1         28.5714       0.0252          2        0.0201
  1           1         71.4286       0.2516          5        0.0503
```

The variable importances were also extracted to inform the business decisions of the e-commerce company. In this decision tree, it showed that average transaction value was the most important variable followed by the age of the user.

| Variable Name | Label | Number of Splitting Rules | Importance |
|---|---|---|---|
| Avg_Txn_Value | | 2 | 1.0000 |
| Age | | 1 | 0.6225 |
| Return_Rate | | 0 | 0.0000 |
| CLTV | | 0 | 0.0000 |
| Days_Since_Last_Purchase | | 0 | 0.0000 |
| Fav_Payment_Method | | 0 | 0.0000 |
| Fav_Category | | 0 | 0.0000 |
| Gender | | 0 | 0.0000 |

**Part 4: Ensemble Methods**

Bagging and Boosting are both examples of ensemble machine learning algorithms. To apply bagging in this case study, the HP Forest Node in SAS EM was used. For Boosting, the Gradient Boosting node was used.

**Bagging: HP Forest**

The Random Forests algorithm applies the bagging by creating multiple decision trees from different bootstrap samples of the dataset, where each sample is generated by randomly selecting data points with replacement. Each tree is built using a random subset of features at each split. The final prediction of the Random Forest model is obtained by aggregating the predictions of all individual trees built previously.

The HP Forest results are shown below. It can be seen that performance on the training data is good while performance on the test set is poor, indicating that the model is overfitting on the training data.

Classification Table

Data Role=TRAIN Target Variable=Churn Target Label=' '

| Target | Outcome | Target Percentage | Outcome Percentage | Frequency Count | Total Percentage |
|---|---|---|---|---|---|
| 0 | 0 | 99.975 | 100.000 | 27811 | 80.0063 |
| 1 | 0 | 0.025 | 0.101 | 7 | 0.0201 |
| 1 | 1 | 100.000 | 99.899 | 6943 | 19.9735 |

Data Role=VALIDATE Target Variable=Churn Target Label=' '

| Target | Outcome | Target Percentage | Outcome Percentage | Frequency Count | Total Percentage |
|---|---|---|---|---|---|
| 0 | 0 | 80.0121 | 99.6476 | 7918 | 79.7301 |
| 1 | 0 | 19.9879 | 99.6474 | 1978 | 19.9174 |
| 0 | 1 | 80.0000 | 0.3524 | 28 | 0.2819 |
| 1 | 1 | 20.0000 | 0.3526 | 7 | 0.0705 |

Fit Statistics

| Target | Target Label | Fit Statistics | Statistics Label | Train | Validation | Test |
|---|---|---|---|---|---|---|
| Churn | | _ASE_ | Average Squa... | 0.030754 | 0.165996 | 0.165769 |
| Churn | | _DIV_ | Divisor for ASE | 69522 | 19862 | 9938 |
| Churn | | _MAX_ | Maximum Abs... | 0.529111 | 0.99 | 0.956667 |
| Churn | | _NOBS_ | Sum of Frequ... | 34761 | 9931 | 4969 |
| Churn | | _RASE_ | Root Average ... | 0.175367 | 0.407426 | 0.407148 |
| Churn | | _SSE_ | Sum of Squar... | 2138.047 | 3297.013 | 1647.416 |
| Churn | | _DISF_ | Frequency of ... | 34761 | 9931 | 4969 |
| Churn | | _MISC_ | Misclassificati... | .0002014 | 0.201994 | 0.202053 |
| Churn | | _WRONG_ | Number of Wr... | 7 | 2006 | 1004 |

**HP Forest - Sampled**

Sampling was conducted to try and address the poor performance achieved in all models by balancing the distribution of the original dataset. Using the Sample node in SAS EM. Results show the Churn variable now has a distribution of 50:50 distribution. The HP Forest node was then applied to this sampled data. It can be seen that although the HP Forest algorithm is still overfitting on the training data, performance on the test data has improved when compared to the HP Forest node trained on imbalanced data.

```
Summary Statistics for Class Targets
(maximum 500 observations printed)

Data=DATA

            Numeric    Formatted    Frequency
Variable     Value       Value        Count       Percent     Label

 Churn         0           0          39732       80.0064
 Churn         1           1           9929       19.9936


Data=SAMPLE

            Numeric    Formatted    Frequency
Variable     Value       Value        Count       Percent     Label

 Churn         0           0           9929         50
 Churn         1           1           9929         50
```

Classification Table

Data Role=TRAIN Target Variable=Churn Target Label=' '

| Target | Outcome | Target Percentage | Outcome Percentage | Frequency Count | Total Percentage |
|---|---|---|---|---|---|
| 0 | 0 | 100 | 100 | 31784.80 | 80.0085 |
| 1 | 1 | 100 | 100 | 7942.00 | 19.9915 |

Data Role=VALIDATE Target Variable=Churn Target Label=' '

| Target | Outcome | Target Percentage | Outcome Percentage | Frequency Count | Total Percentage |
|---|---|---|---|---|---|
| 0 | 0 | 79.3395 | 44.3102 | 3521.42 | 35.4474 |
| 1 | 0 | 20.6605 | 46.1500 | 917.00 | 9.2307 |
| 0 | 1 | 80.5305 | 55.6898 | 4425.78 | 44.5510 |
| 1 | 1 | 19.4695 | 53.8500 | 1070.00 | 10.7709 |

**Boosting: Gradient Boosting**
SAS EM applies gradient boosting to decision trees by building trees sequentially, where each subsequent tree improves on the previous tree by fitting each new tree on the residual errors made by the previous trees. The final prediction is made by aggregating the predictions of all trees, which, when combined, provide a more accurate and generalized model than any single decision tree.

The gradient boosting results are shown below. Poor performance was seen in both the training and validation results. Indicating that the model is underfitting.

Event Classification Table

Data Role=TRAIN Target=Churn Target Label=' '

| False Negative | True Negative | False Positive | True Positive |
|---|---|---|---|
| 7942 | 31785 | 0 | 0 |

Data Role=VALIDATE Target=Churn Target Label='

| False Negative | True Negative | False Positive | True Positive |
|---|---|---|---|
| 1987 | 7947 | 0 | 0 |

Fit Statistics

| Target | Target Label | Fit Statistics | Statistics Label | Train | Validation | Test |
|---|---|---|---|---|---|---|
| Churn | | _NOBS_ | Sum of Frequencies | 39727 | 9934 | . |
| Churn | | _SUMW_ | Sum of Case Weights ... | 79454 | 19868 | . |
| Churn | | _MISC_ | Misclassification Rate | 0.199914 | 0.20002 | . |
| Churn | | _MAX_ | Maximum Absolute Err... | 0.81933 | 0.81933 | . |
| Churn | | _SSE_ | Sum of Squared Errors | 12696.18 | 3178.67 | . |
| Churn | | _ASE_ | Average Squared Error | 0.159793 | 0.159989 | . |
| Churn | | _RASE_ | Root Average Squared... | 0.399741 | 0.399987 | . |
| Churn | | _DIV_ | Divisor for ASE | 79454 | 19868 | . |
| Churn | | _DFT_ | Total Degrees of Free... | 39727 | . | . |

**Part 5: Conclusion**

In this case study, data was sourced from Kaggle to attempt to study and gain insights on e-commerce customer churn. In particular, the goal was to develop decision trees, random forests and gradient boosting models to predict if e-commerce customers will churn and then use insights gained from the model to inform business decisions for the e-commerce company. The data sourced from Kaggle was first cleaned, checked for errors, missing values were imputed in Talend Data Preparation. Talend Open Studio used to demonstrate how an e-commerce company would ingest data from two separate tables (transactions and customers) and join them into a single table before analysis. Feature engineering was then conducted using Python to create features necessary for a churn classification model. Finally SAS was used to create the prediction models.

Based on the results obtained from SAS, the HP Forest - Sampled model showed the best performance although the model still experienced a high degree of overfitting. Variable importance for the HP Forest model was extracted to give some insights on what factors affect customer churn. Compared to decision tree variable importance which come only from a single tree, variable importance outputs calculated from a random forest model are aggregated across all trees in the random forest, increasing consistency and reducing variability in the outcome. Based on this output, Days_Since_Last_Purchase is the most influential variable followed by CLTV and Avg_Txn_Size. The e-commerce company could reduce customer churn by sending discounts or promotions to customers that have made a purchase in a long time, especially customers who frequently make large purchases or those who have made many purchases over time. The least important variables are Fav_Payment_Method and Gender. This could inform the company that gender-specific marketing is not necessary as gender does not affect churn as much as other variables.

Variable Importance

| Variable Name | Number of Splitting Rules | Train: Gini Reduction | Train: Margin Reduction | OOB: Gini Reduction | OOB: Margin Reduction | Valid: Gini Reduction | Valid: Margin Reduction | Label |
|---|---|---|---|---|---|---|---|---|
| Days_Sinc... | 91226 | 0.080654 | 0.161309 | -0.17891 | -0.09691 | -0.17450 | -0.09277 | |
| CLTV | 73450 | 0.069281 | 0.138562 | -0.15380 | -0.08371 | -0.15383 | -0.08311 | |
| Avg_Txn_V... | 66962 | 0.063316 | 0.126633 | -0.14032 | -0.07691 | -0.14112 | -0.07768 | |
| Age | 53043 | 0.044937 | 0.089873 | -0.10232 | -0.05788 | -0.10307 | -0.05860 | |
| Return_Rate | 52971 | 0.038374 | 0.076748 | -0.09071 | -0.05021 | -0.09141 | -0.05103 | |
| Fav_Categ... | 7216 | 0.005007 | 0.010014 | -0.00817 | -0.00501 | -0.00836 | -0.00564 | |
| Gender | 5083 | 0.002964 | 0.005928 | -0.00577 | -0.00380 | -0.00554 | -0.00347 | |
| Fav_Payme... | 4039 | 0.002676 | 0.005351 | -0.00438 | -0.00245 | -0.00477 | -0.00282 | |

Other models such as decision trees and gradient boosting were unable to find good performance on this dataset with models either overfitting or underfitting. Furthermore, performance of all models suggest that a better dataset is required for the e-commerce company to address the issue of customer churn. To improve on this in the future, the company could embark on a data collection initiative by sending surveys or feedback forms to customers.