WQD7005 Data Mining

Semester 1, Session 2023/2024

Alternative Assessment 1

Matric Number	Name
S2197999	Chong Hun Yee

Lecturer: PROF. DR. TEH YING WAH

GitHub Link:- https://github.com/chonghunyee/WQD7005_AA1/tree/main

Project Overview

The project's goal is to predict customer churn in order to aid in proactive customer retention strategies. Customer information, purchase history, and behaviors are all included in the dataset. This study uses SEMMA methodology which stands for Sample, Explore, Modify, Model, and Assess using SAS e-Miner, Talend Data Integration as well as Talend Data Prep. Talend Data Integration is in charge of Extract, transform, and load (ETL) processes, whereas Talend Data Prep is in charge of data cleanliness to produce good quality dataset. SAS E-Miner is essential for sampling, detailed analysis, data exploration, imputation, feature engineering, data modification and modelling. The ultimate goal is to predict churn prediction effectively model that will allow businesses to identify and retain at-risk customers.

Objectives

The primary goal of this assignment is to leverage the churn dataset for predicting and understanding customer churn for the past few years. The objectives throughout this assessment is as follows:-

i) To identify the trend, patterns and the factors contributing to the customers' churn.

ii) To study the relationship between the customers' churn and other variables.

Dataset Description

The dataset in this project is obtained from Kaggle which is https://www.kaggle.com/datasets/shriyashjagtap/e-commerce-customer-for-behavior-analysis?select=ecommerce_customer_data_large.csv. There are 2 dataset files this Kaggle website, and ecommerce_customer_data_custom_ratios CSV dataset is chosen. The dataset is then modified by separating the dataset into a sales data and a general dataset. Besides, some changes are also done in this dataset by adding and removing the dataset attributes, as well as renaming the attribute names in order to be different from other course mates who might take the same dataset. The meaning of the dataset columns are explained as follows:-

CustomerID: A unique identifier for each customer.
Age: The age of the customer.
Gender: The gender of the customer.
Location: The place where customer stays.
MembershipLevel: There are 4 categories which are Silver, Bronze, Platinum and Gold.
TotalPurchase: The quantity of the product purchased.

TotalSpent: The total amount spent by the customer in each transaction.
FavouriteCategory: The category or type of the purchased product.
LastPurchaseDate: The date of each last purchase made by the customer.
ProductPrice: The price of the purchased product.
PaymentMethod: The method of payment used by the customer (e.g., credit card, PayPal).
ProductReturns: 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).

Tools' Roles and Justification

Talend Data Integration

In this project, Talend Data Integration is chosen because it allows task like merging 2 different datasets together easily. In doing this, I do not need to combine the dataset one by one manually with will take a long time. Besides, after merging, the dataset is huge whereby it consists of 250,000 rows and 13 columns.

Talend Data Preparation

Talend Data Prep tool is selected due to its ability in preparing data, including data cleaning and transformation steps. In this project, this tool is used to handle data inconsistencies as well as to transform data which is to transform the date format to YYYY-MM-DD.

SAS Enterprise-Miner

SAS Enterprise-Miner is used because of its ability to do stratified sampling, data imputation, data modification, data modelling and model assessing. To further elaborate, for sampling phase, stratified random sampling is applied where only 10% of the dataset is used for this project. This can be seen where it helps to reduce the number of rows from 250,000 to 25,000 rows. Besides, it also allows data modification to drop variable such as CustomerID which is not significant in this project and will not affect the customer churn prediction. In addition, data imputation is also done by imputing the missing values of "ProductReturns" columns with "count" or mode. Lastly, this tool assists in building data mining models, including decision trees, HP forest, and Gradient Boosting. It also provides models' performances evaluation which is under the "Assess" phase in SEMMA methodology.

Steps In Using Each Tool

Talend Data Integration

1) Open Talend Data Integration Studio.
2) Launch the Talend Data Integration tool and create a new project.
3) Right-click in the Repository panel on the left under Job Designs.
4) Choose Create job. Provide a name for the job which is “combine two files” and a description. Click Finish.
5) Add 2 tFileInputDelimited Components in order to read 2 delimited CSV files which are the general file and sales file.
6) From the Palette panel on the right, type "tFileInputDelimited" into the search bar.
7) For input Data, drag and drop tFileInputDelimited components from the palette onto the workspace for both CSV files. Configure each component to point to the respective CSV files. Ensure you define the schema correctly for each CSV.
8) tMap Configuration. Drag and drop a tMap component onto the workspace. Link both tFileInputDelimited components to the tMap component. Double click the tMap component to open its editor.
9) For data joining, On the left side of the tMap editor, you will see the two input datasets. Drag the CustomerID column from task1_customers to the other dataset where it will combine using the CustomerID key. On the right side of the tMap editor, define your output structure.
10) Configure tFileOutputDelimited. Double-click on the component. Set the file path where you want to save the filtered data. Ensure the output schema matches the input schema. Define the CSV format, e.g., delimiter as ",".
11) Run the job.
12) Save the job.
13) Click on the "Run" tab at the bottom. Click on the "Run" button. The job will process and merge 2 files into a CSV file.

Talend Data Preparation

1) Open Talend Data Prep tool and import the merged dataset into this tool.
2) For the Gender column, replace from Female to “F” and Male to “M”.
3) For Location column, replace USA to US
4) For FavouriteCategory column, change “cloth” to “Clothing”.
5) For LastPurchaseDate column, transform date column to “yyyy-MM-dd”.

6) Export the dataset into CSV file for further analysis in SAS Enterprise Miner, as well as export to local Tableau file for reporting purposes.

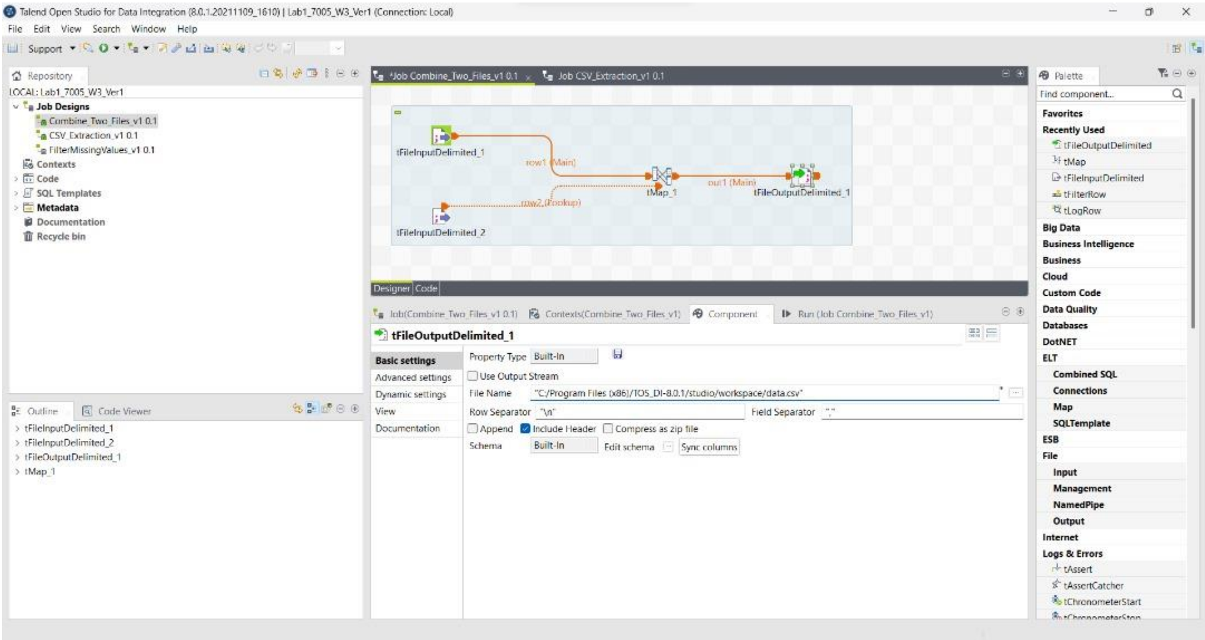
SAS Enterprise-Miner

1) Open and launch SAS Enterprise-Miner which is the local SAS version, and SAS On Demand for academics which the cloud version.
2) Create a new diagram named “Final Exam”, upload the dataset into the cloud version as well as create a new library.
3) In the local SAS, create a new data source by editing the data type of each variable clearly and state the target variables.
4) By using the “Sample” node, do stratified random sampling using 10% of the original dataset. Right click and run.
5) Drag the “Drop” node, whereby “CustomerID” column is dropped. Right click and run.
6) Then, drag the “Impute” node, to handle missing data for “ProductReturns” column by using “count”. Right click and run.
7) Then, split the dataset into 70% of train, 30% of validation and 0% for test on the left panel. Right click and run.
8) For modelling, Decision Tree, HP Forest and Gradient Boosting nodes are dragged from the model section. Right click and run.
9) Under the Assess section, drop the “Model Comparison” and connect all the models to it. Right click and run.

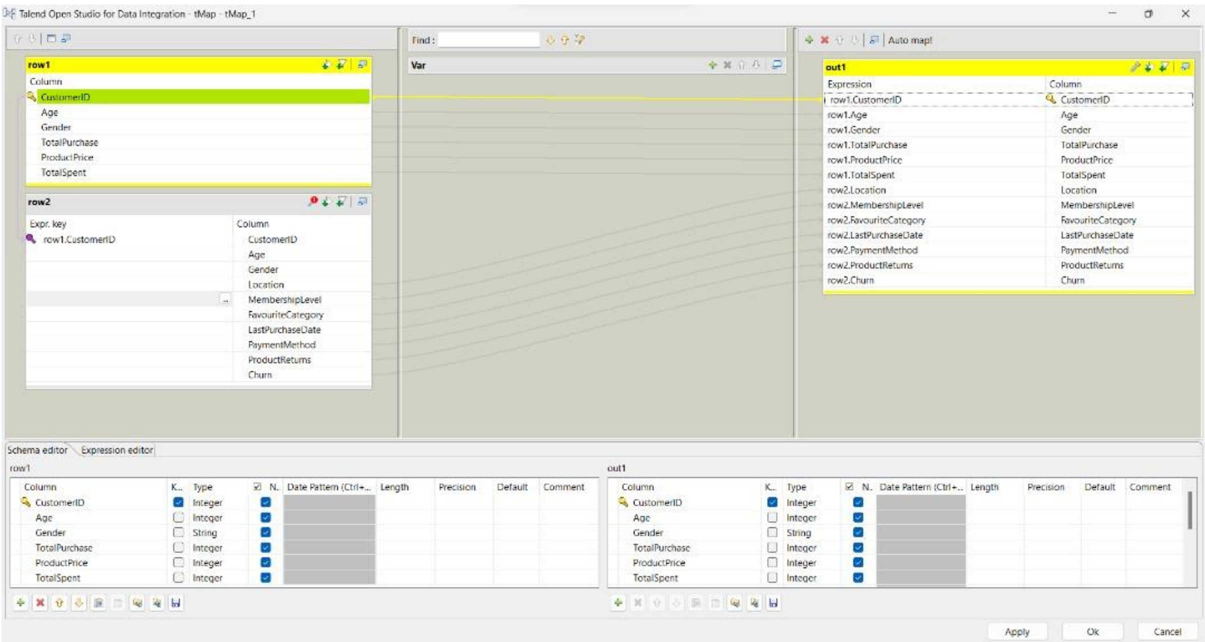
Implementation, Screenshots and Explanation

Talend Data Integration

Firstly, dataset is merged using Talend Data Integration. Merge both sales data with general data csv files together using Talend Integration tool. The general pipelines are as shown below:-

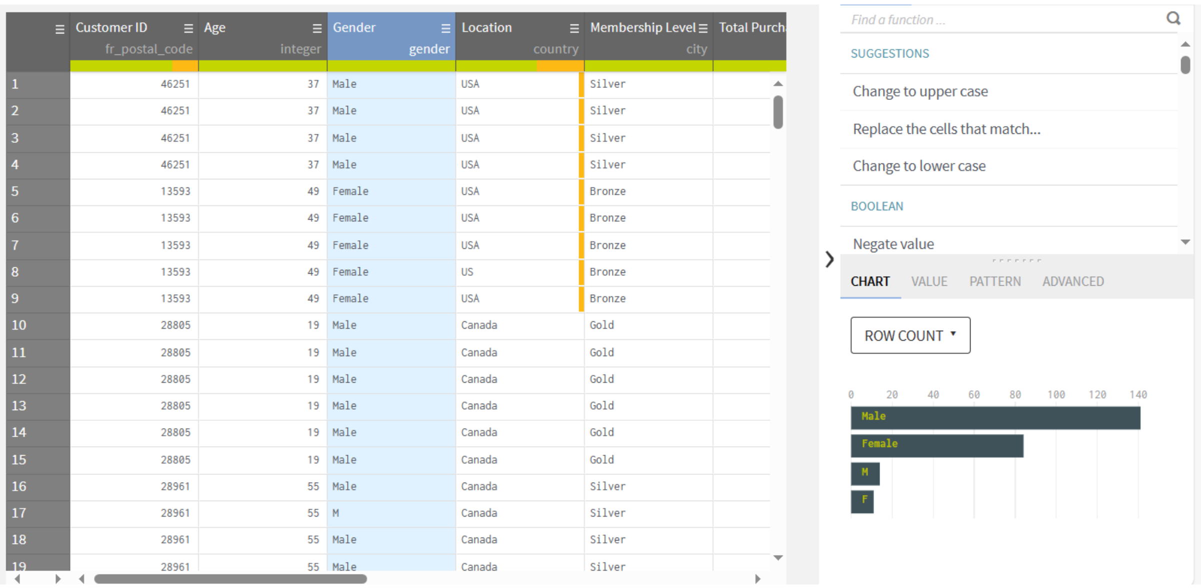


In the tMap node, configure and connect both dataset using the key column which is “CustomerID”.

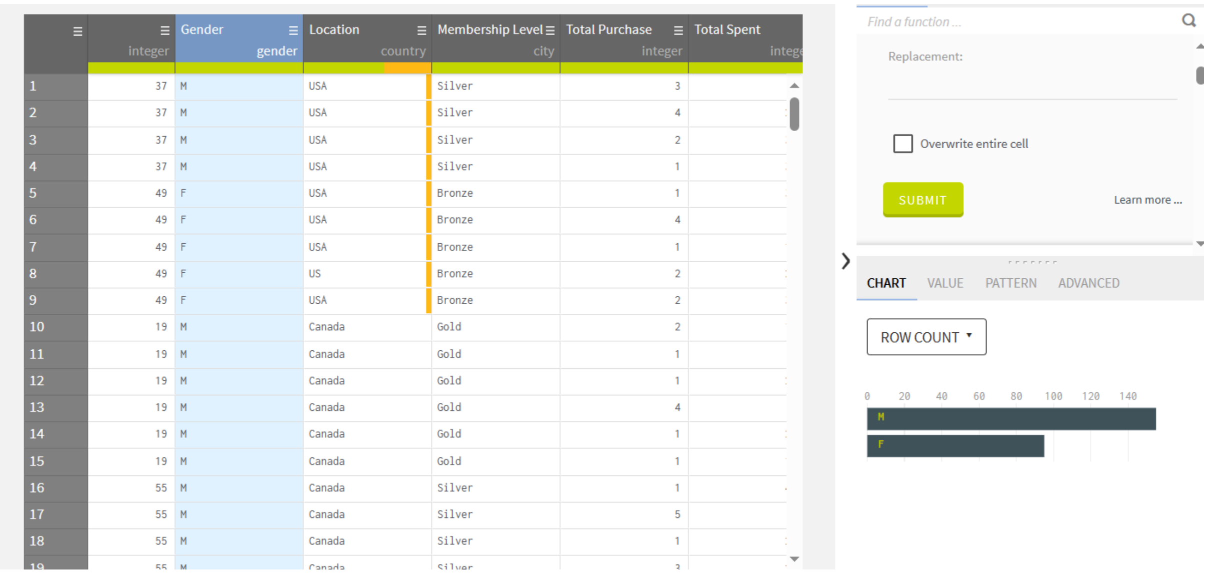


Talend Data Prep

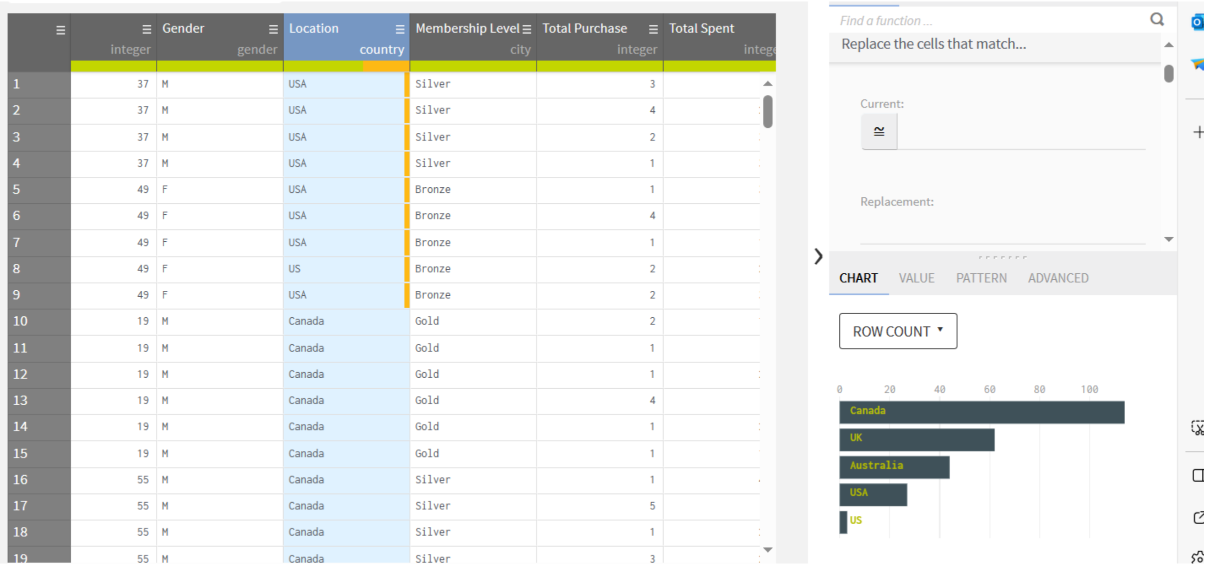
1) Change from Female to “F” and Male to “M”.



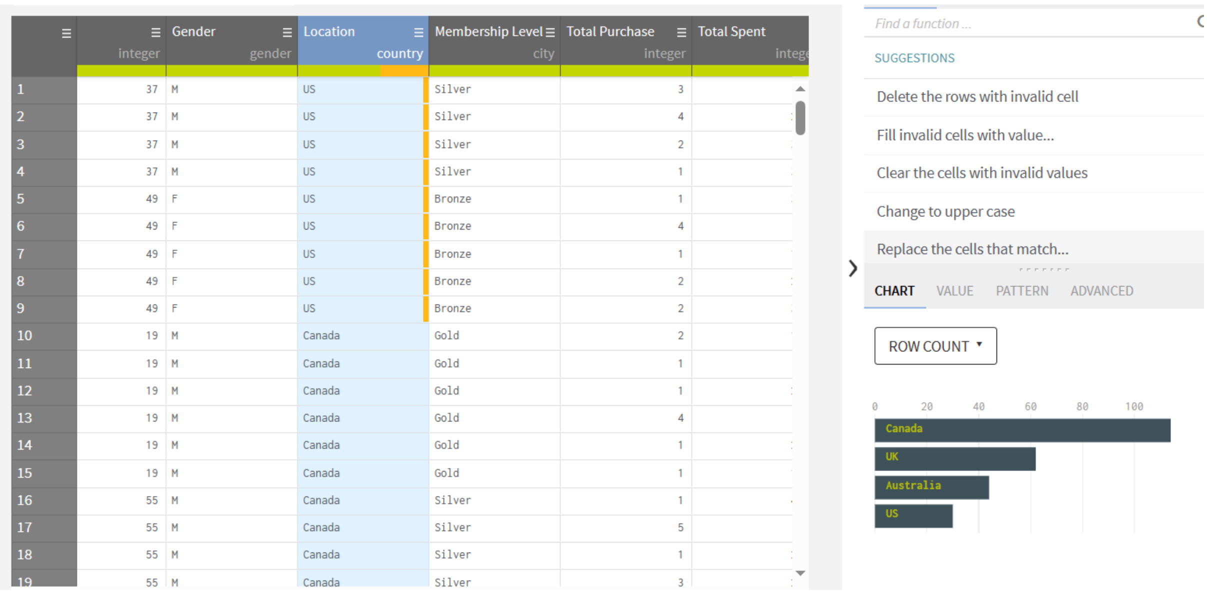
Result:



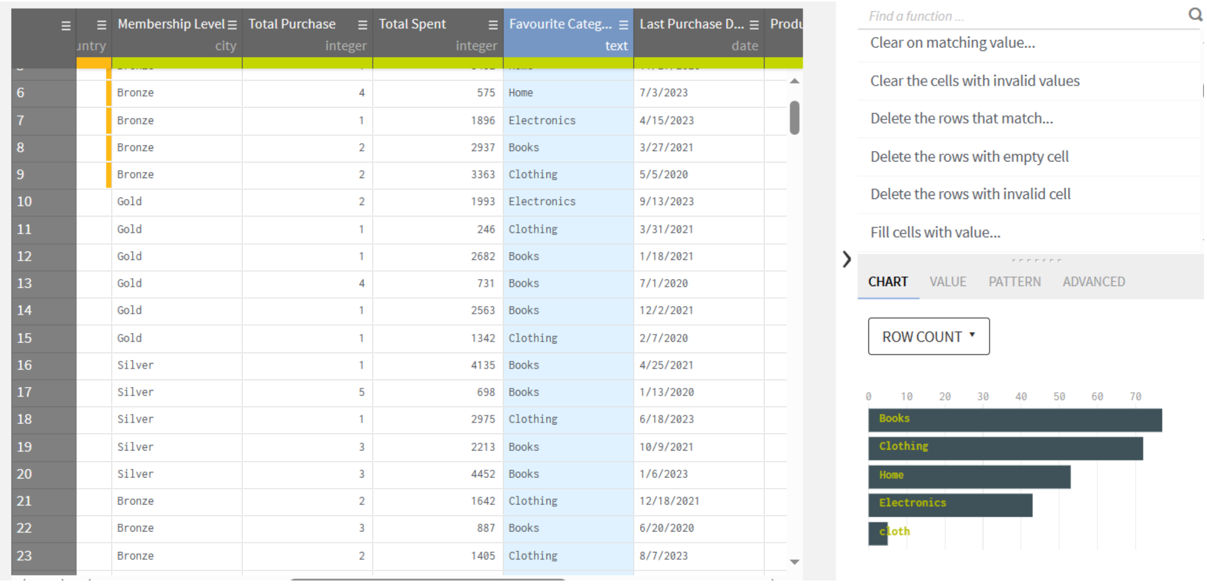
2) Change USA to US



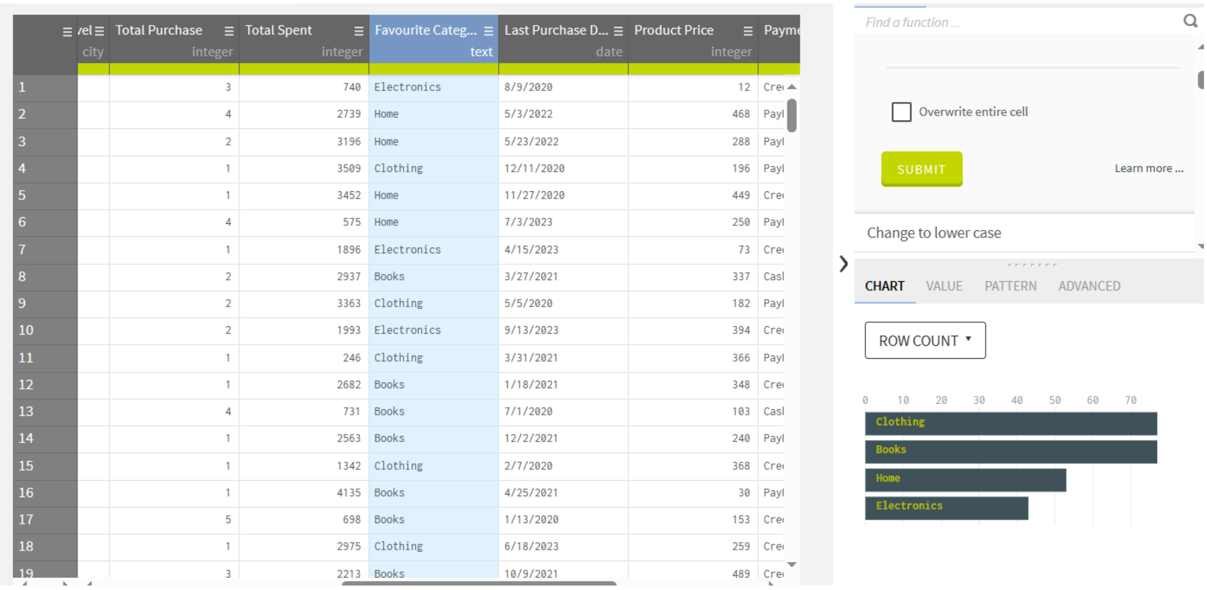
Result:-



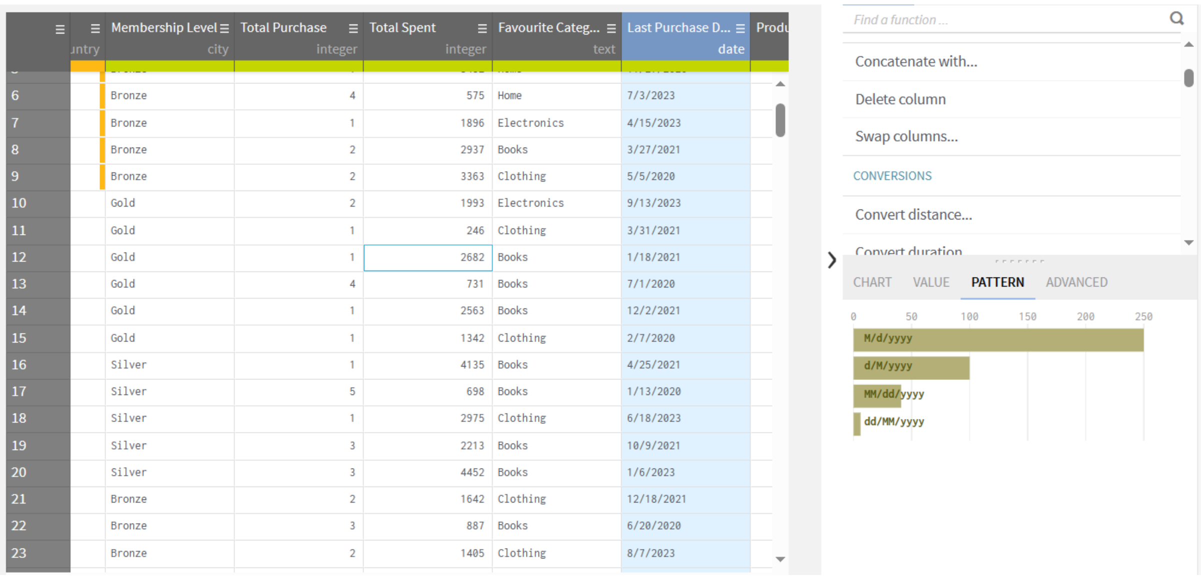
3) Change cloth to Clothing



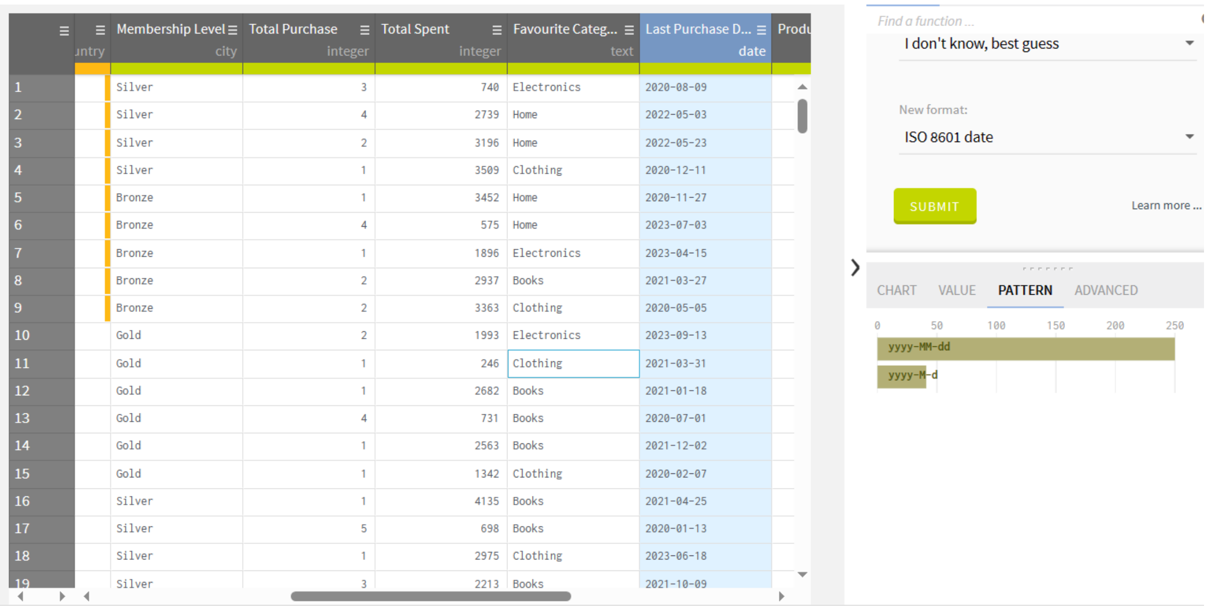
Results:-



4) Change date format to “yyyy-MM-dd”.

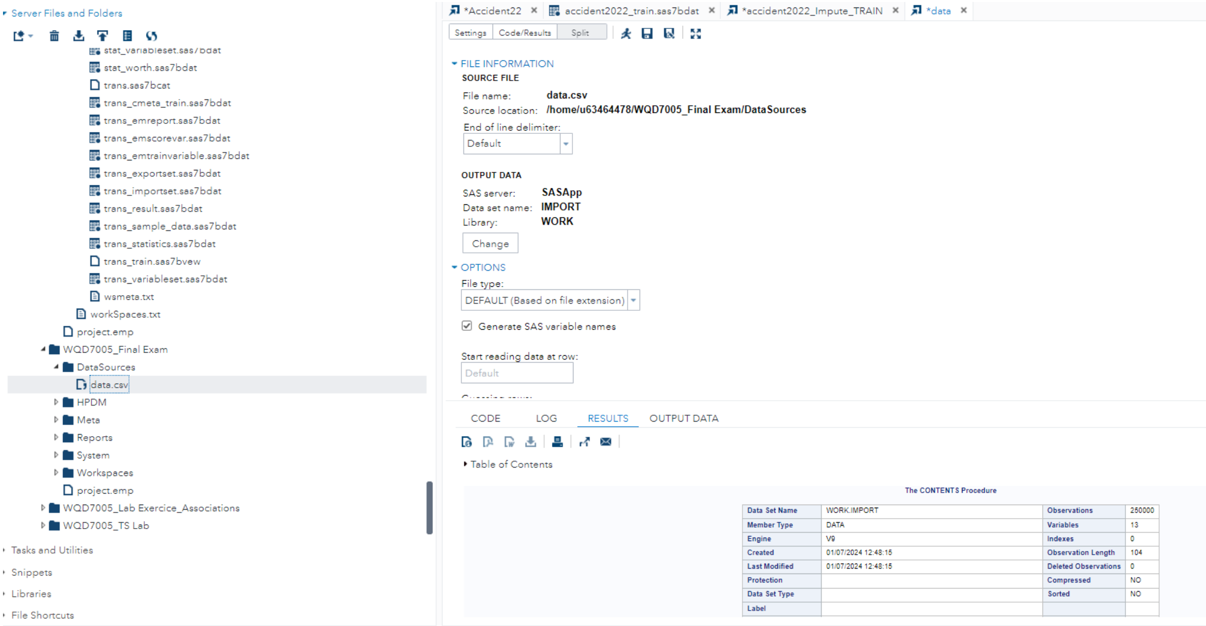


Result:-

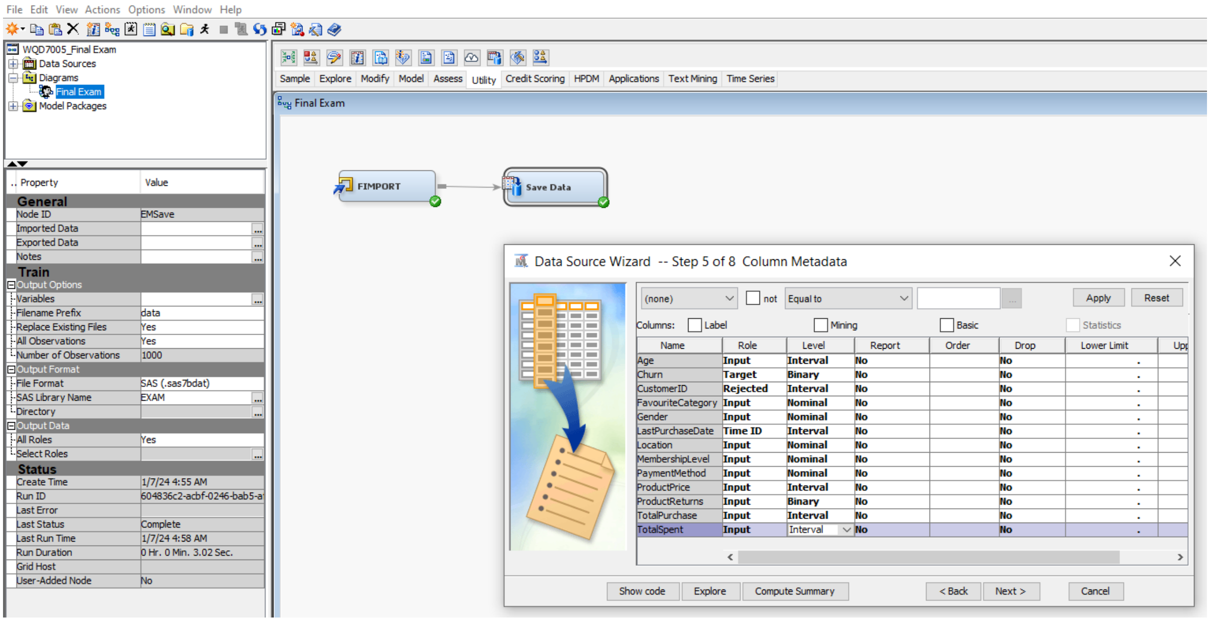


SAS Enterprise Miner and SAS Studio On Demand

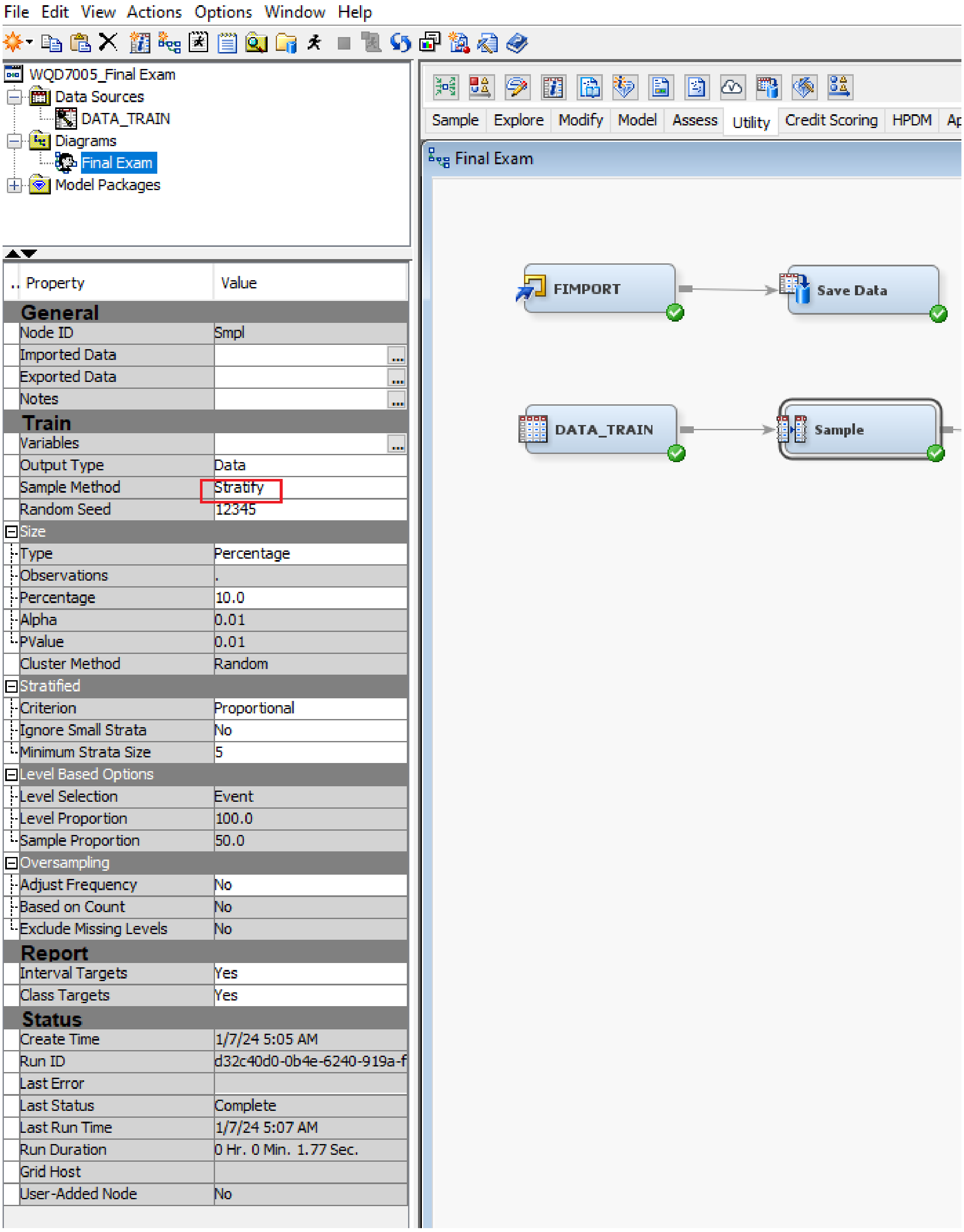
Create project and diagram in SAS Miner Enterprise. Then, import the dataset into SAS.



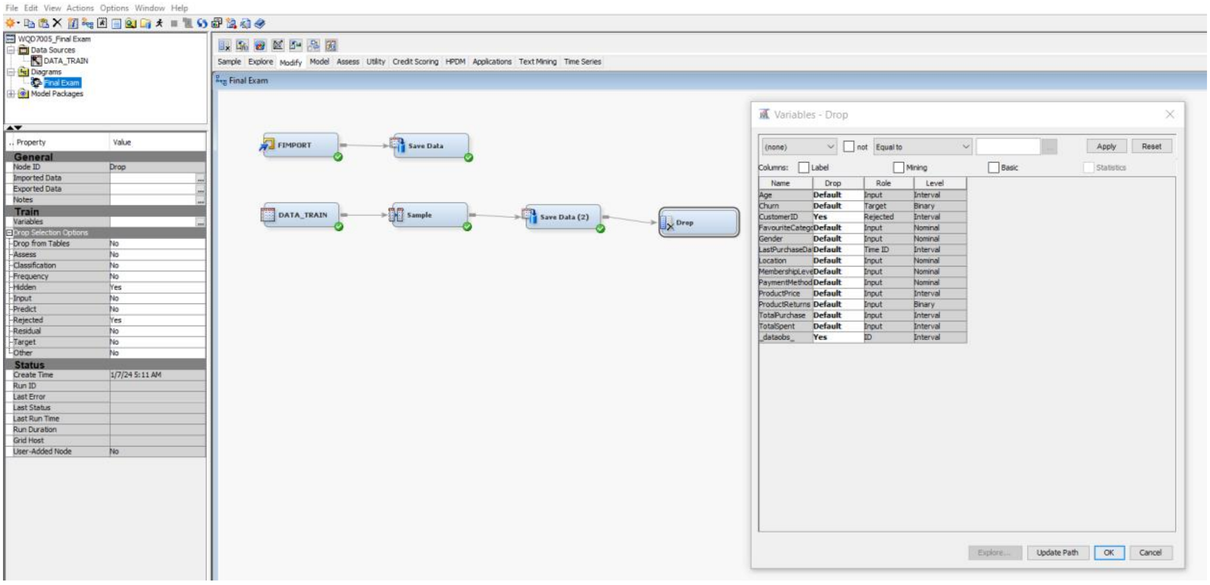
"File Import" node is dragged to the diagram to import the dataset in SAS Enterprise-Miner. Edit the data type of each variable clearly, state the "Churn" as target variable.



Since it is know that SEMMA methodology will be applied, sampling will need to do as a first step. Sampling using stratify sampling with 10% using the “Sample” node.

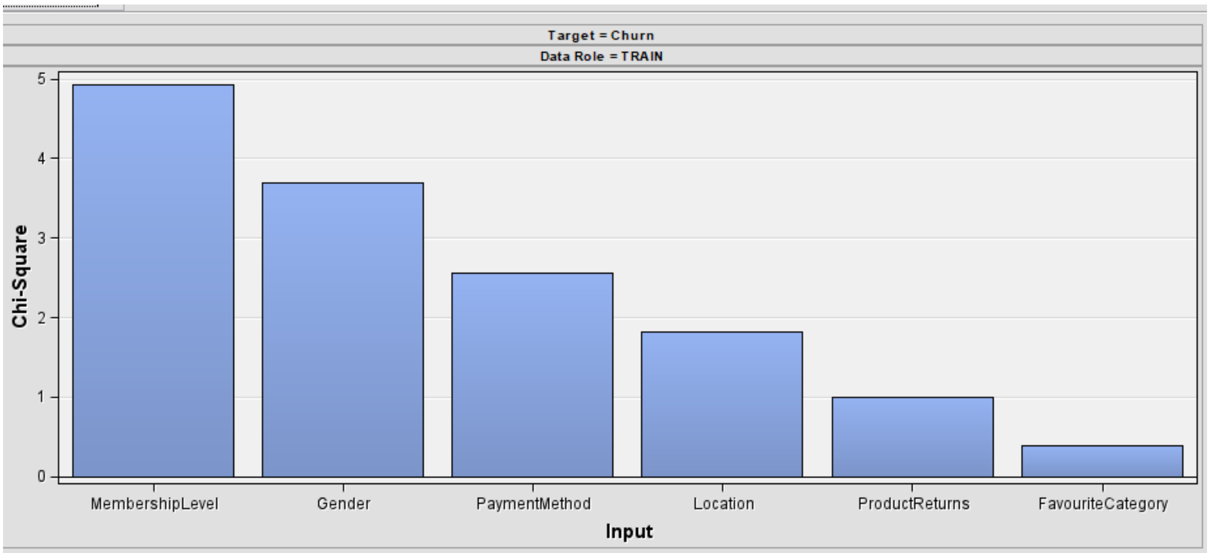


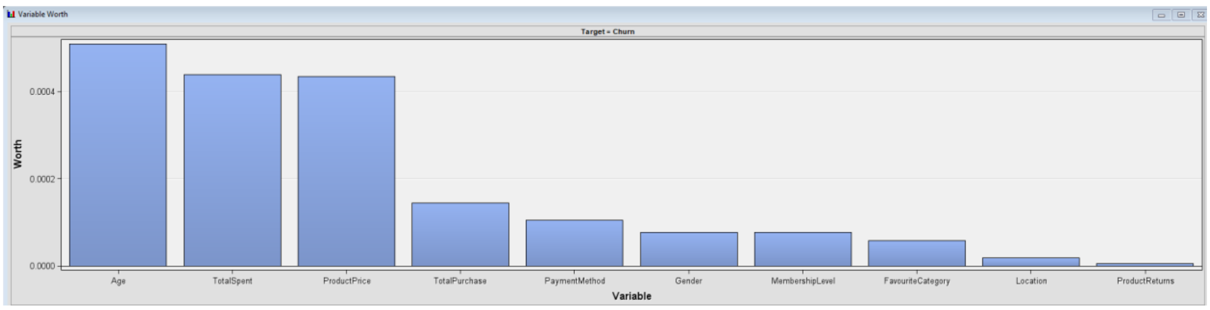
Drop unnecessary variables such as “CustomerID”. However, there is a new attribute generated by SAS named “_dataobs_” which also need to be dropped. This can be done by using the “Drop” node.



Exploration

By using the “StatExplore”, the summarized of the sampled data report will be generated.





The summary statistics table is shown as follows:-

Variable Summary

Role	Measurement Level	Frequency Count
INPUT	BINARY	1
INPUT	INTERVAL	4
INPUT	NOMINAL	5
TARGET	BINARY	1

Variable Levels Summary
(maximum 500 observations printed)

Variable	Role	Frequency Count
Churn	TARGET	2

Class Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	FavouriteCategory	INPUT	4	0	Books	29.85	Clothing	29.48
TRAIN	Gender	INPUT	2	0	F	50.30	M	49.70
TRAIN	Location	INPUT	4	0	UK	25.49	US	25.15
TRAIN	MembershipLevel	INPUT	4	0	Gold	25.44	Bronze	25.00
TRAIN	PaymentMethod	INPUT	4	0	Credit Card	39.59	PayPal	30.45
TRAIN	ProductReturns	INPUT	3	4806	0	40.51	1	40.26
TRAIN	Churn	TARGET	2	0	0	80.05	1	19.95

Distribution of Class Target and Segment Variables
(maximum 500 observations printed)

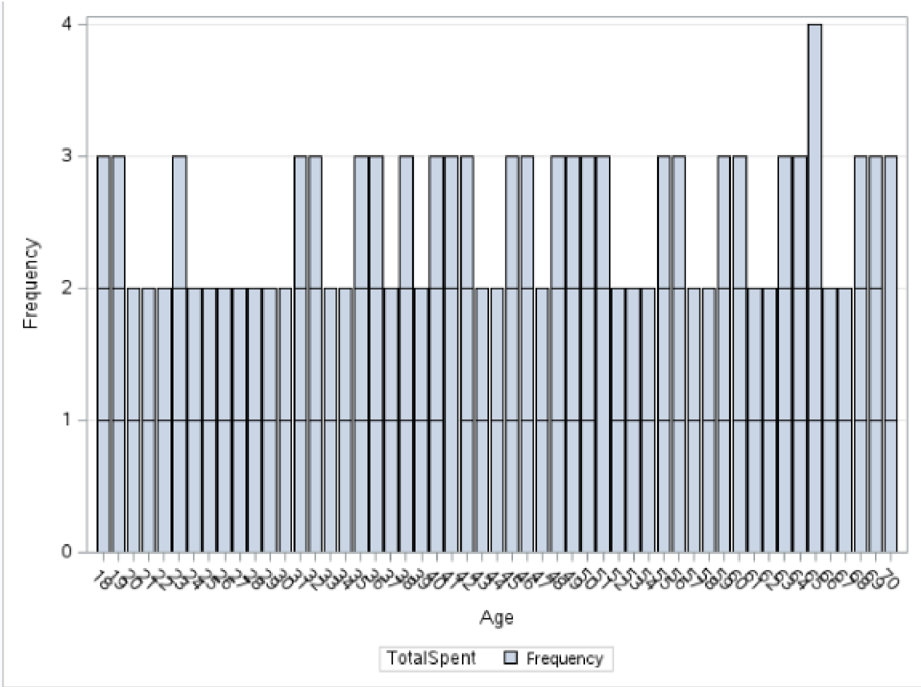
Data Role=TRAIN

Data Role	Variable Name	Role	Level	Frequency Count	Percent
TRAIN	Churn	TARGET	0	20013	80.052
TRAIN	Churn	TARGET	1	4987	19.948

From the above, there are 2 attributes which are binary data type, 4 interval data type attributes as well as 5 nominal data type. Besides, it can be seen that there are missing values in “ProductReturns” column with 4806 missing data, whereas others have no missing values.

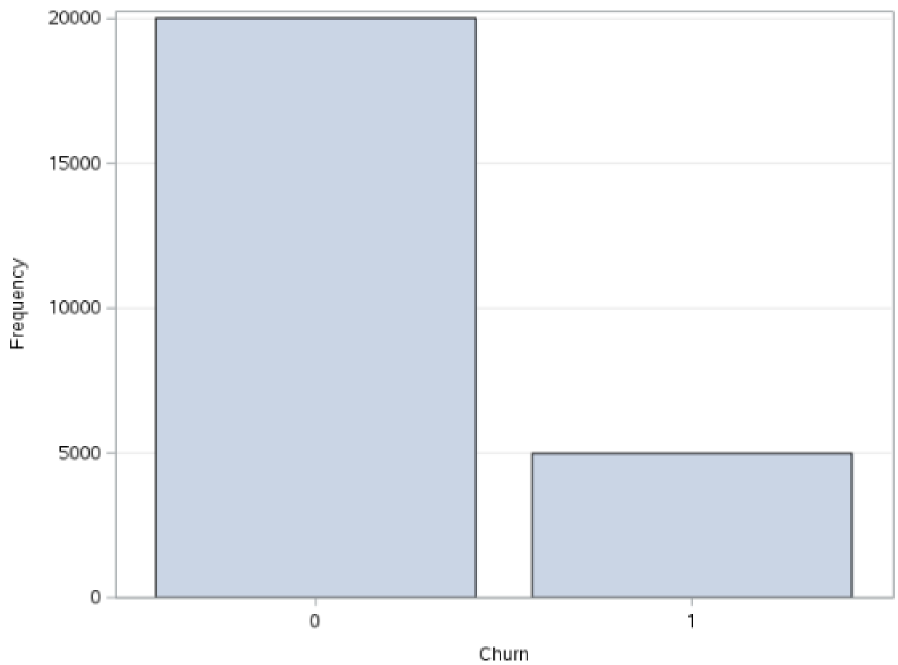
For data exploration, this project will start by analyzing some important attributes to have a better understanding of the dataset which will be shown below.

Bar graph of TotalSpent against Age



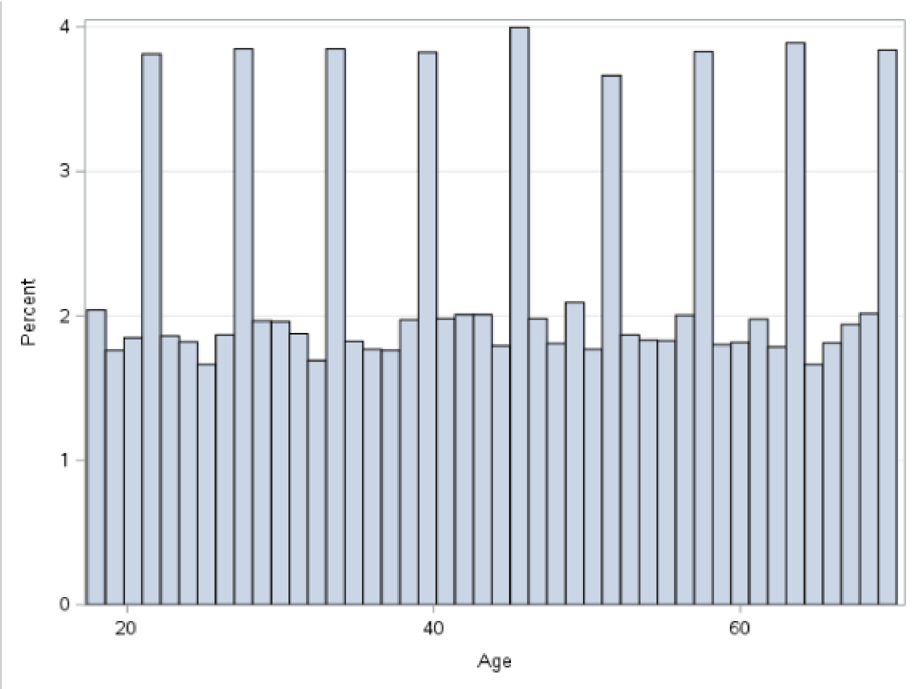
From the above, it can be seen that the "TotalSpent" against "Age" distribution is not normal.

Box plot to check the outliers of the target variable "Churn"



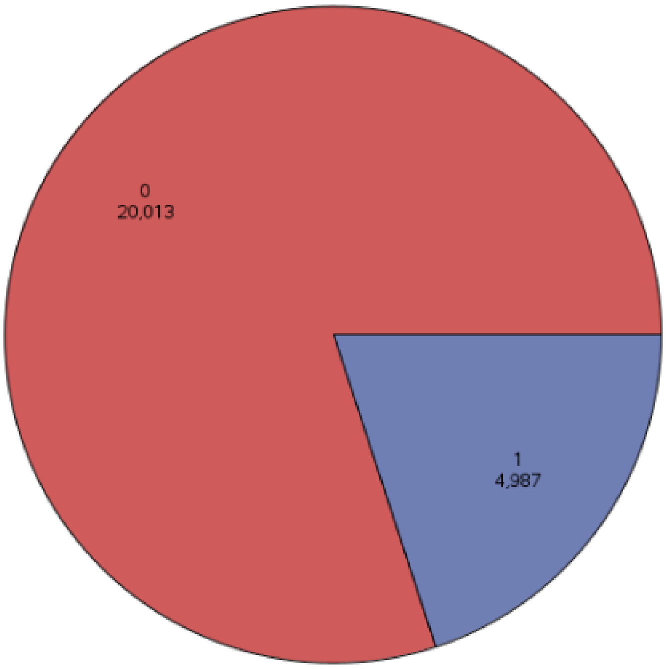
From the above, it can be seen that there are around 20,000 with churn of value "0". On the other hand, there are about 5000 with churn of value "1".

Distribution of age using histogram



From the above histogram, the age distribution is not normal.

Check the Target variable using Pie Chart

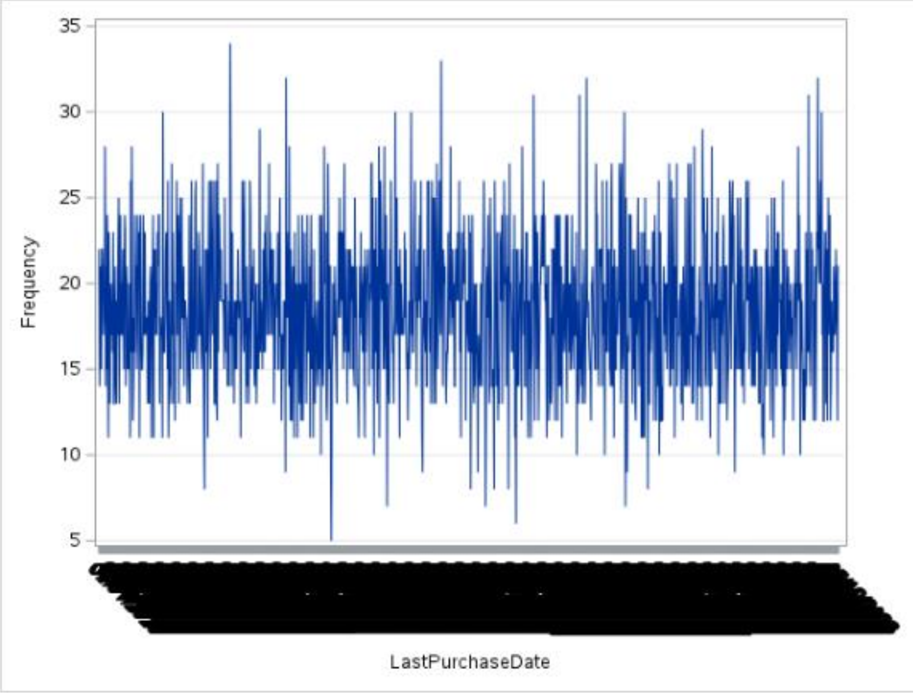


From the above pie chart, it can seen that there are the proportion of 20.013 for churn value of "0", whereas 4.987 for the churn value of "1". This illustrates that the churn value of "0" is more than the churn value of "1".

Boxplot of the target variable

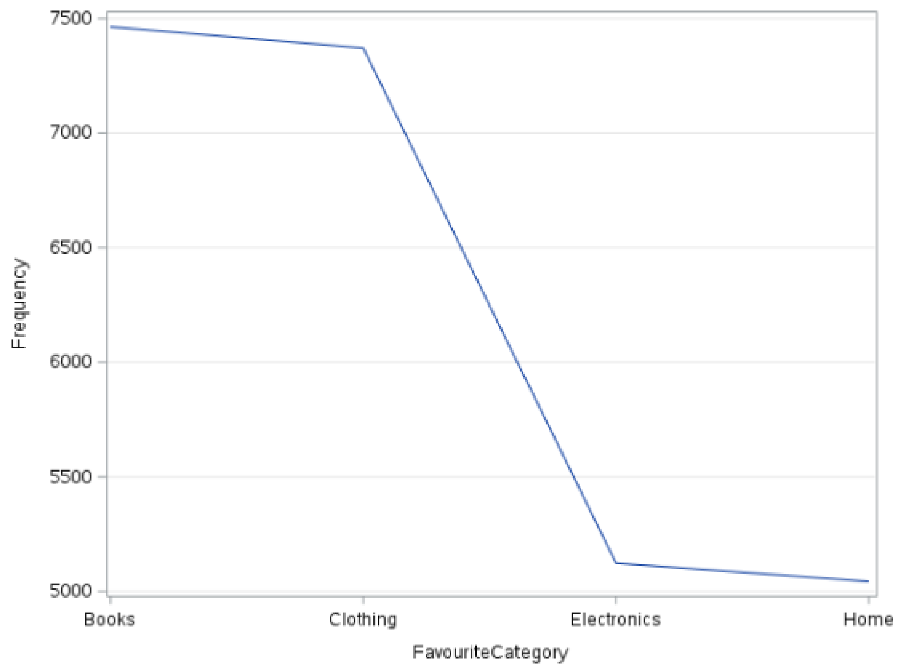
From the above, it can be concluded that there are no outliers for the target variable “Churn”.

LastPurchaseDate using Line Graph



From the line graph above, it can be seen that the distribution of LastPurchaseDate is not normal.

Line graph of the FavouriteCategory column



From the above line graph, it can be summarized that the Home is the least "FavouriteCategory" at 5000. On the contrary, books are the highest favourite category fat 7500 for the customers.

Results and Analysis - Exploration

Overall, from all the different types of graphs plotted above, the visualizations provide useful information about various aspects of the dataset. The line graph reveals significant customer preferences, with "Home" being the least preferred category and "Books" being the most preferred, emphasizing distinct spending patterns. Examining the distribution of "LastPurchaseDate" reveals an out-of-the-ordinary pattern, implying potential irregularities in purchasing behaviour over time. The pie chart shows a significant imbalance in the "Churn" variable, with approximately 20.013% indicating no churn "0" and only 4.987% indicating churn "1". This implies that instances with no churn predominate. This trend is reinforced by the histogram depicting age distribution, which shows a higher number of instances with "Churn" value "0" compared to "1." Finally, the non-normal distribution of "TotalSpent" versus "Age" highlights potential complexities in the relationship between age and total spending. Overall, these visualizations help to provide a more complete understanding of customer behavior, churn patterns, and potential areas for additional analysis or model refinement.

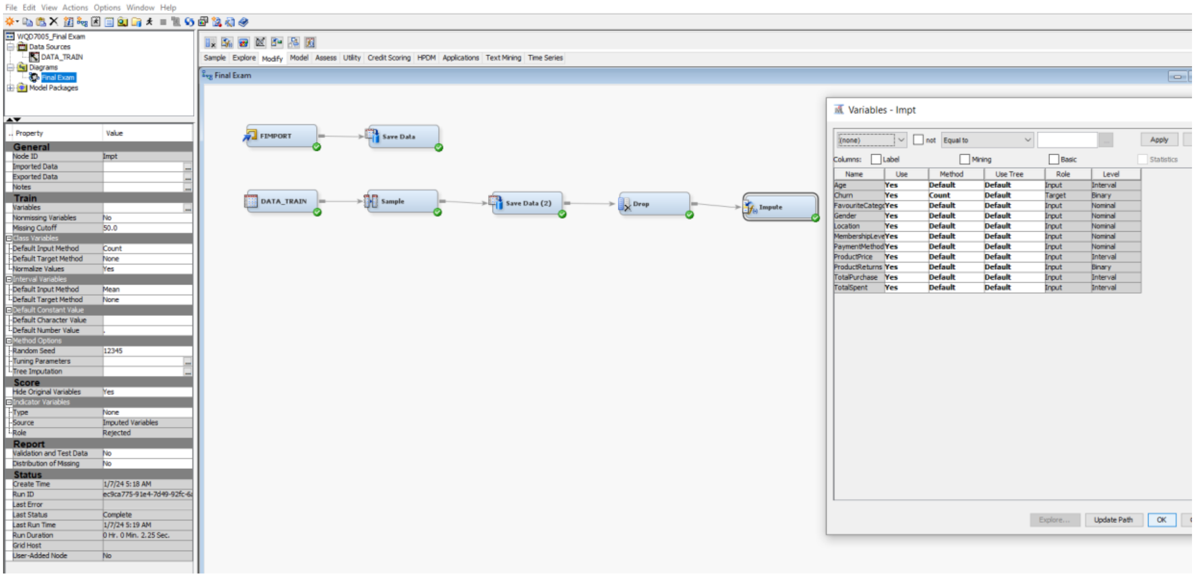
Modify Phase

Imputation

Since there are missing values in “ProductReturns” column, imputation will be done.

```
Output

Variable Summary

            Measurement    Frequency
 Role          Level         Count

INPUT       BINARY           1
INPUT       INTERVAL         4
INPUT       NOMINAL          5
TARGET      BINARY           1
TIMEID      INTERVAL         1


*------------------------------------------------------------*
* Score Output
*------------------------------------------------------------*


*------------------------------------------------------------*
* Report Output
*------------------------------------------------------------*




Interval Variable Summary Statistics

                                                                                  Standard
Variable         Label      Missing         N    Minimum    Maximum       Mean   Deviation        Skewness    Kurtosis

Age                               0     25000         18         70      44.05       15.28    -.002626630    -1.18833
ProductPrice                      0     25000         10        500     254.32      141.81    0.006763591    -1.19704
TotalPurchase                     0     25000          1          5       2.99        1.42    0.008806459    -1.30572
TotalSpent                        0     25000        101       5338    2730.78     1446.11    0.004605113    -1.19664




Class Variable Summary Statistics

                                      Number
                                        of
Variable              Label    Type   Levels    Missing

Churn                           N        2          0
FavouriteCategory               C        4          0
Gender                          C        2          0
Location                        C        4          0
MembershipLevel                 C        4          0
PaymentMethod                   C        4          0
ProductReturns                  N        2       4806
```

Impute the missing data in Churn column with count or mode. This can be shown in the diagram below.



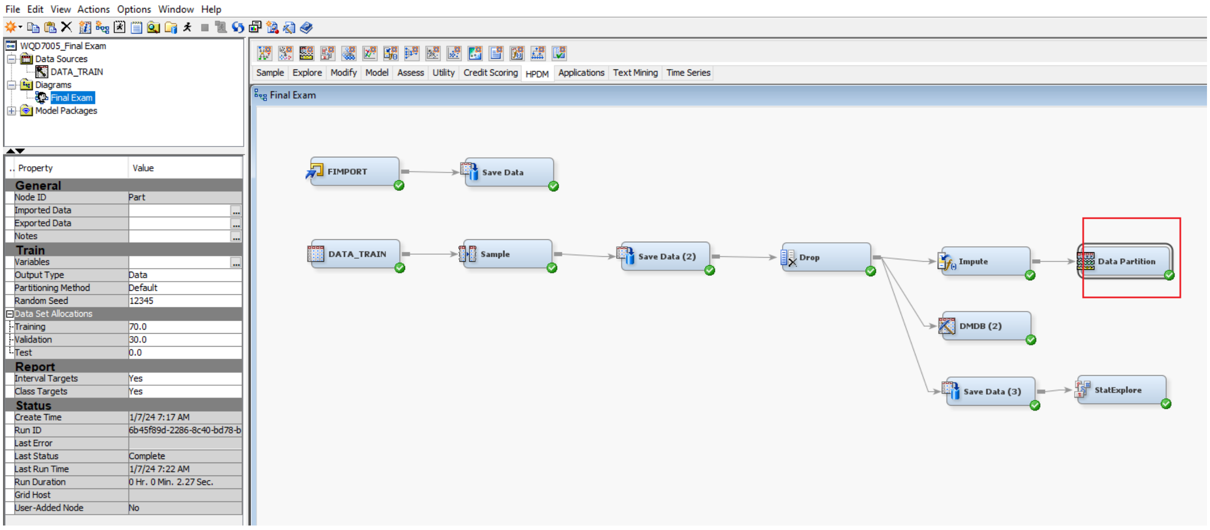
Output

```
Variable Summary

         Measurement    Frequency
 Role       Level         Count

INPUT     BINARY            1
INPUT     INTERVAL          4
INPUT     NOMINAL           5
TARGET    BINARY            1
TIMEID    INTERVAL          1


*------------------------------------------------------------*
* Score Output
*------------------------------------------------------------*


*------------------------------------------------------------*
* Report Output
*------------------------------------------------------------*




Interval Variable Summary Statistics

                                                                          Standard
Variable        Label      Missing          N    Minimum    Maximum      Mean   Deviation      Skewness    Kurtosis

Age                              0      25000         18         70     44.05       15.28   -.002626630    -1.18833
ProductPrice                     0      25000         10        500    254.32      141.81   0.006763591    -1.19704
TotalPurchase                    0      25000          1          5      2.99        1.42   0.008806459    -1.30572
TotalSpent                       0      25000        101       5338   2730.78     1446.11   0.004605113    -1.19664




Class Variable Summary Statistics

                                                   Number
                                                     of
Variable                     Label            Type   Levels   Missing

Churn                                          N        2        0
FavouriteCategory                              C        4        0
Gender                                         C        2        0
IMP_ProductReturns    Imputed ProductReturns   N        2        0
Location                                       C        4        0
MembershipLevel                                C        4        0
PaymentMethod                                  C        4        0
```

From the above, after imputation, there are no more missing values for the "ProductReturns" column. Since it is a synthetics dataset, there is no outliers as shown in the exploration section.

Hence, data partition will be done to split the dataset into 70% of training and 30% testing dataset.

<u>Data Partition</u>



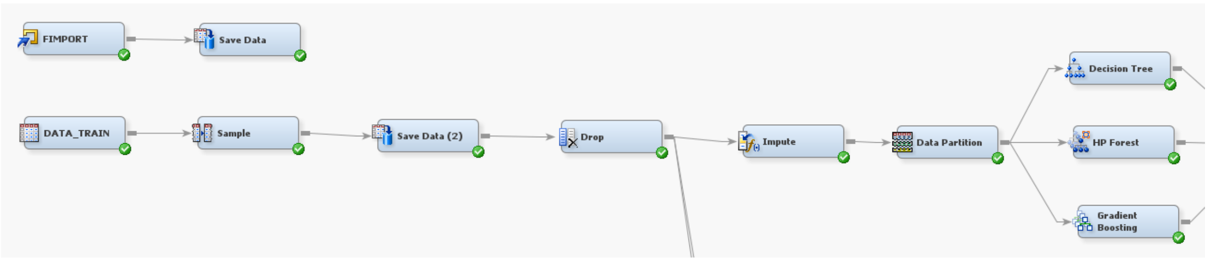
Below are the summary of the data partition node. Additionally, the dataset is split 17,498 training data, whilst 7,502 for validation data.

```
Variable Summary

          Measurement    Frequency
 Role        Level         Count

INPUT      BINARY            1
INPUT      INTERVAL          4
INPUT      NOMINAL           5
TARGET     BINARY            1
TIMEID     INTERVAL          1



Partition Summary

                                    Number of
Type            Data Set           Observations

DATA        EMWS1.Impt_TRAIN           25000
TRAIN       EMWS1.Part_TRAIN           17498
VALIDATE    EMWS1.Part_VALIDATE         7502
```

```
Summary Statistics for Class Targets

Data=DATA

              Numeric     Formatted     Frequency
Variable       Value        Value         Count        Percent     Label

 Churn           0            0           20013         80.052
 Churn           1            1            4987         19.948


Data=TRAIN

              Numeric     Formatted     Frequency
Variable       Value        Value         Count        Percent     Label

 Churn           0            0           14008        80.0549
 Churn           1            1            3490        19.9451


Data=VALIDATE

              Numeric     Formatted     Frequency
Variable       Value        Value         Count        Percent     Label

 Churn           0            0            6005        80.0453
 Churn           1            1            1497        19.9547
```

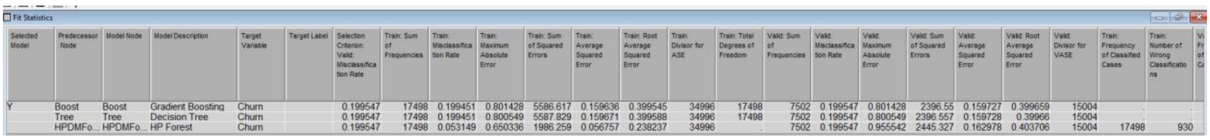
<u>Modelling</u>

This project will be using models such as decision trees, HP Forest as bagging model, and Gradient Boosting as boosting model. The nodes are dragged as shown in the diagram below.



<u>Assess</u>

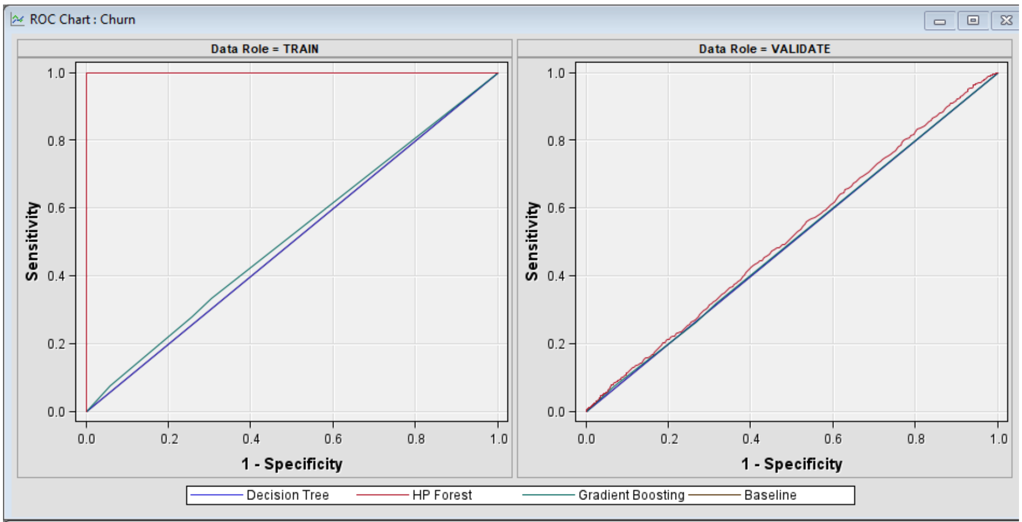
Model comparison. It can be concluded that Gradient Boosting is the best model.

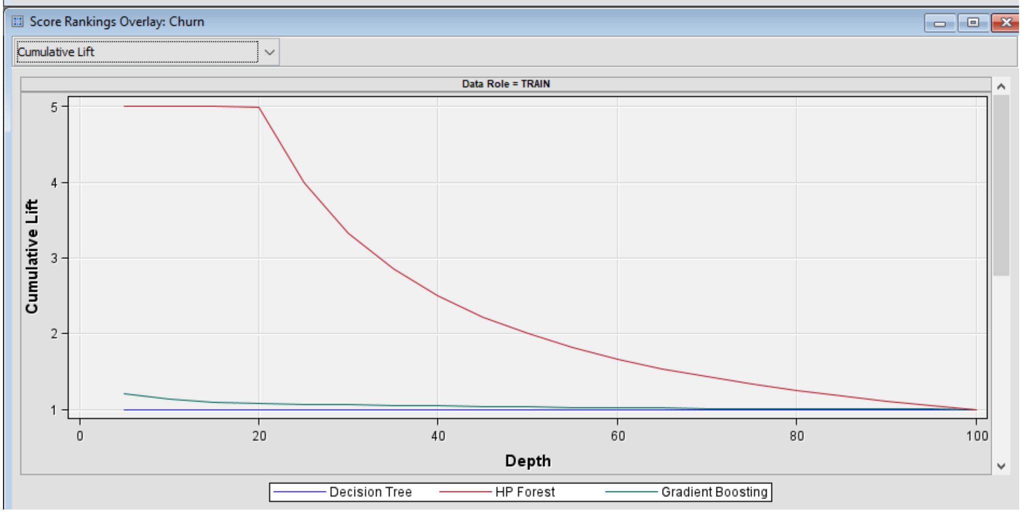

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Train: Sum of Frequencies	Train: Misclassification Rate	Train: Maximum Absolute Error	Train: Sum of Squared Errors	Train: Average Squared Error	Train: Root Average Squared Error	Train: Divisor for ASE	Train: Total Degrees of Freedom	Valid: Sum of Frequencies	Valid: Misclassification Rate	Valid: Maximum Absolute Error	Valid: Sum of Squared Errors	Valid: Average Squared Error	Valid: Root Average Squared Error	Valid: Divisor for VASE	Train: Frequency of Classified Cases	Train: Number of Wrong Classifications
Y	Boost	Boost	Gradient Boosting	Churn		0.199547	17498	0.199451	0.801428	5586.617	0.159636	0.399545	34996	17498	7502	0.199547	0.801428	2396.55	0.159727	0.399659	15004	.	.
	Tree	Tree	Decision Tree	Churn		0.199547	17498	0.199451	0.800549	5587.829	0.159671	0.399588	34996	17498	7502	0.199547	0.800549	2396.557	0.159728	0.39966	15004	.	.
	HPDMFo...	HPDMFo...	HP Forest	Churn		0.199547	17498	0.053149	0.650336	1986.259	0.056757	0.238237	34996	.	7502	0.199547	0.955542	2445.327	0.162978	0.403706	15004	17498	930

From the above result for model comparison, it can be seen that Gradient Boosting has the best performance compared to Decision Tree and followed by HP Forest.

Below are ROC chart for all the 3 models.





Fit Statistics
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Selected Model	Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Y	Boost	Gradient Boosting	0.19955	0.15964	0.19945	0.15973
	Tree	Decision Tree	0.19955	0.15967	0.19945	0.15973
	HPDMForest	HP Forest	0.19955	0.05676	0.05315	0.16298

Fit Statistics Table
Target: Churn

Data Role=Train

Statistics	Boost	Tree	HPDMForest
Train: Bin-Based Two-Way Kolmogorov-Smirnov Probability Cutoff	0.20	0.00	0.40
Train: Kolmogorov-Smirnov Statistic	0.03	0.00	1.00
Train: Average Squared Error	0.16	0.16	0.06
Train: Roc Index	0.51	0.50	1.00
Train: Cumulative Percent Captured Response	11.32	10.00	50.14
Train: Percent Captured Response	5.31	5.00	25.07
Selection Criterion: Valid: Misclassification Rate	0.20	0.20	0.20
Train: Total Degrees of Freedom	17498.00	17498.00	.
Train: Frequency of Classified Cases	.	.	17498.00
Train: Divisor for ASE	34996.00	34996.00	34996.00
Train: Gain	13.21	0.00	401.38
Train: Gini Coefficient	0.03	0.00	1.00
Train: Bin-Based Two-Way Kolmogorov-Smirnov Statistic	0.02	0.00	1.00
Train: Kolmogorov-Smirnov Probability Cutoff	0.20	.	0.34
Train: Cumulative Lift	1.13	1.00	5.01
Train: Lift	1.06	1.00	5.01
Train: Maximum Absolute Error	0.80	0.80	0.65
Train: Misclassification Rate	0.20	0.20	0.05
Train: Sum of Frequencies	17498.00	17498.00	17498.00
Train: Root Average Squared Error	0.40	0.40	0.24
Train: Cumulative Percent Response	22.58	19.95	100.00
Train: Percent Response	21.16	19.95	100.00
Train: Sum of Squared Errors	5586.62	5587.83	1986.26
Train: Sum of Case Weights Times Freq	34996.00	.	.
Train: Number of Wrong Classifications	.	.	930.00

From the above result, although all these 3 models have the same misclassification rate, Gradient Boosting is still performed as the best model. This is because, in terms of the average squared error, HP Forest has the highest which is at 0.1630. This followed by Decision Tree average squared error at 0.159728, which this model has the second-best performance. Lastly, Gradient Boosting performs the best with the lowest average squared error at 0.159727.

Confusion Matrix

Event Classification Table
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Model Node	Model Description	Data Role	Target	Target Label	False Negative	True Negative	False Positive	True Positive
Tree	Decision Tree	TRAIN	Churn		3490	14008	0	0
Tree	Decision Tree	VALIDATE	Churn		1497	6005	0	0
HPDMForest	HP Forest	TRAIN	Churn		930	14008	.	2560
HPDMForest	HP Forest	VALIDATE	Churn		1496	6004	1	1
Boost	Gradient Boosting	TRAIN	Churn		3490	14008	0	0
Boost	Gradient Boosting	VALIDATE	Churn		1497	6005	0	0

For the confusion matrix above, it can be concluded that HP Forest predicted 1 data wrong as positive.

Future Strategies/ Improvement

For future business strategies, it is critical to ensure that the dataset used for analysis is representative of real-world scenarios. This improves the findings relevance and applicability to practical situations in real business world. A closer alignment with real-world data allows for a more accurate understanding of customer behaviour. As a result, more effective solutions to real-world problems. Furthermore, given more time and resources, a strategic next step could involve SAS feature selection. The goal of this process is to identify and prioritize the most important variables in predicting customer churn. We can improve the model's accuracy and interpretability by utilizing advanced feature selection techniques within SAS, resulting in more robust and reliable predictions. This step is consistent with our ongoing commitment to improving our model's predictive capabilities and practical utility in real-world applications. In doing all these business strategies, the future researchers or businessman can manage and analyze their data well in this business field.

Reflections or Learning Outcomes

Throughout this project, the most important outcome to me is to implement the 3 tools which I have learnt in class into this alternative assessment or project. These 3 tools which I use in this project are Talend Data Integration, Talend Data Prep, and SAS Enterprise Miner. I gained valuable insights and learning outcomes throughout the course of this project. The SEMMA (Sample, Explore, Modify, Model, Assess) methodology was used to provide a structured approach to predicting customer churn.

On the other hand, I successfully used data mining models in SAS Enterprise Miner by utilizing decision trees, HP Forest as bagging model, and Gradient boosting as boosting model. With this tool, I am able to compare all these 3 models' performances easily without doing codes.

Besides, Talend Data Integration is very helpful in merging 2 separate datasets together with ease. By using this tool, the dataset becomes complete and ready for analysis.

Moreover, Talend Data Prep was helpful in data cleaning and transformation. To further elaborate on this, it helps to tackle issues such as data inconsistency and date format discrepancies in an easy-to-understand manner. Additionally, it has a very friendly interface where I can use it and clean my data easily.

In my opinion, this project allows me to show my project management as well as problem solving skills efficiently and effectively. Furthermore, the project demonstrated the application of acquired knowledge to a real-world Kaggle dataset, emphasizing the ability to apply theoretical concepts into practical solutions for predicting customer churn.

In a nutshell, reporting skill is also crucial to present the project results by using a word document as well as GitHub. GitHub is a well-know website used by all researchers, data scientists and data analysts all around the globe. By publishing my work into GitHub, it allows other people to follow my step-by-step implementation using SEMMA methodology with 3 tools. They are able to follow as well as correct my mistakes. This project also provides valuable insights with visualization for readers to understand well. Overall, the project provided a comprehensive understanding of the end-to-end data analysis process as well as its practical application in solving real-world business problems.