Alternative Assessment 1

WQD 7005

Name: TSU HIAO PING

Student ID: 22106817

Git Hub Link for SAS Folder: https://github.com/Thping99/WQD7005_AA1

Git Hub Link for Excel/CSV File: https://github.com/Thping99/WQD7005_AA1_Excel

Report

By creating our own dataset, we can tailor it to match the specific objectives and requirements of this assignment. This ensures that the data is directly aligned with the skills and concepts we are expected to demonstrate. Therefore, I have created my own dataset (up to 205 observations and 12 attributes) using a data generator. The objective of this project is to provide insights into customer behavior and suggestions for business strategy

The 12 attributes I have created:

- **CustomerID**
- **Age**
- **Gender**
- **Location (Malaysia)**
- **Membership (Bronze, Silver, Gold, Platinum)**
- **TotalPurchases**
- **TotalSpent**
- **FavoriteCategory**
- **LastPurchaseDate**
- **Occupation**
- **WebsiteVisitsPerMonth**
- **Churn: Indicates whether the customer has stopped purchasing (1 for churned, 0 for active).**

Firstly, we import our CSV file, which is generated by our own self, into Talend Data Prep. Talend Data Prep automates the data profiling process, enabling users to quickly identify and understand the distribution of spelling errors within their datasets. This purposeful automation ensures a more efficient and accurate cleaning process. Users can explore their data visually to identify issues such as missing values, spelling errors and inconsistent date formats.

For example, we have found misspellings under two attributes (*MembershipLevel* and *Favorite Category*), as shown in Figures 1 and 2 below. Through the function "Find and Group Similar Text" under Transformation Suggestions, we can **fix spelling errors**. We replace misspellings with "Silver" and "Platinum" accordingly. Besides, we also replace the spelling mistakes for "*Favorite Category"* with "Clothing" and "Electronics" accordingly.

Figure 3 displays that there are **date formatting errors** when glancing through our dataset. To solve this problem, we will use the function "Change Date Format" and redefine the date format into "dd-MM-yyyy". This is done using Talend Data Prep, as shown in Figure 4.

By lowering the possibility of mistakes and inconsistencies, standardizing date formats aids in maintaining data integrity. The accuracy of analyses and interpretations can be jeopardized by inconsistent date formats, which can cause errors in computations and other areas. Hence, this is one of the important steps for data cleaning process.

Moreover, we can find out the number of missing values under the "Value" panel. In our dataset, there are 4 missing values under "Location", 2 missing values under the "LastPurchaseDate", and 3 missing values under "TotalSpent".

Under Talend data prep, the final step is to export the dataset and name it under "Dateset_WQD7005_AA1_DataPrep_v1".

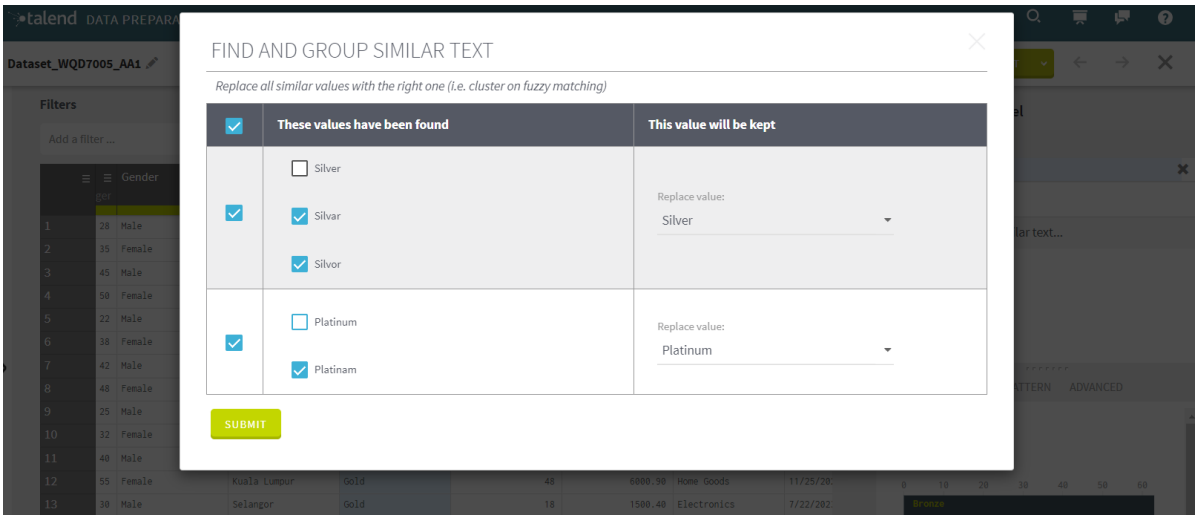

Figure 1: Spelling errors under the attribute of "MembershipLevel".

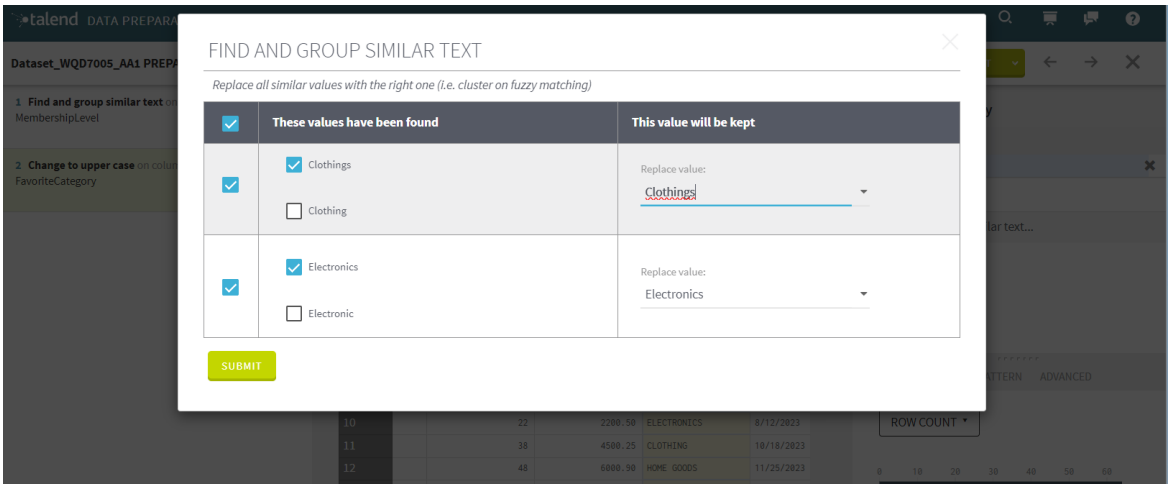

Figure 2: Spelling errors under the attribute of "Favorite Category".

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	CustomerID	Age	Gender	Location	Membersh	TotalPurch	TotalSpen	FavoriteCa	LastPurchaseDate	Occupatio	WebsiteVi	Churn	
2	101	28	Male	Kuala Lum	Bronze	15	1200.5	Electronic	6/15/2023	Software E	10	0	
3	102	35	Female	Penang	Silver	25	2000.75	Clothing	9/28/2023	Marketing	8	0	
4	103	45	Male	Johor	Silvor	30	3500.2	Clothing		Architect	12	0	
5	104	50	Female	Selangor	Platinum	40	5000.3	Clothing	10/20/2023	Doctor	15	0	
6	105	22	Male	Kuala Lum	Bronze	10	800.6	Home Goo	5/18/2023	Student	5	1	
7	106	38	Female	Selangor	Silver	20	1800.4	Clothing	7/10/2023	Engineer	9	0	
8	107	42	Male	Johor	Silver	35	4000.15	Clothing	8/15/2023	Business (	11	0	
9	108	48	Female	Selangor	Platinum	45	5500.8	Clothing	9/22/2023	Lawyer	14	0	
10	109	25	Male	Kuala Lum	Bronze	12	1000.3	Clothing	6/25/2023	Graphic D	6	1	
11	110	32	Female	Kuala Lum	Silvar	22	2200.5	Electronic	Saturday, August 12, 2023	Teacher	7	0	
12	111	40	Male	Kuala Lum	Gold	38	4500.25	Clothing	Wednesday, October 18, 2023	Chef	10	0	
13	112	55	Female	Kuala Lum	Gold	48	6000.9	Home Goo	Saturday, November 25, 2023	Artist	16	0	
14	113	30	Male	Selangor	Gold	18	1500.4	Electronic	Saturday, July 22, 2023	Nurse	8	0	
15	114	36	Female	Johor	Gold	28	2500.6	Clothing	9/22/2023	Scientist	10	1	
16	115	48	Male	Kuala Lum	Platinum	42	4800.3	Home Goo	Tuesday, September 12, 2023	Financial A	12	0	
17	116	52	Female	Kuala Lum	Gold	50	7000.5	Electronic	Tuesday, November 28, 2023	Psycholog	18	0	
18	117	28	Male	Kuala Lum	Platinum	14	1100.6	Clothing	Tuesday, June 20, 2023	Researche	7	1	

Figure 3: Date Formatting under the attribute of "LastPurchaseDate".

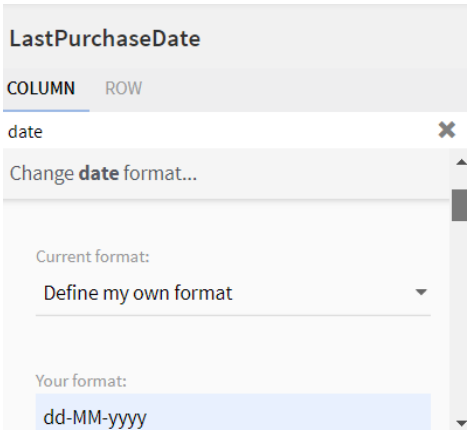

Figure 4: Redefining the Date Format.

To demonstrate the way to handle missing values, we can use Talend Data Integration for this report.

- **Talend Data Integration: Filter out null values**

We will use Talend Data Integration to demonstrate how to filter out missing values. As mentioned above, the missing values found are listed again below:

Attributes	**Number of Missing Values**
Location	4
LastPurchaseDate	2
TotalSpent	3

Firstly, we drag and drop a "tFileInputDelimited" component from the palette to the job design area to read the CSV file. Under the 'Component' panel, we add the file stream to our CSV file "Dateset_WQD7005_AA1_DataPrep_v1". Then, we adjust the Row Separator as "\n", Field Separator as "," and Header as "1", as shown in Figure 5. The next step is to define the schema by clicking on "Edit Schema" and mapping the attributes, as displayed in Figure 6.

Since the number of our missing values is very limited and these missing values do not significantly impact the overall analysis, we decided to perform filtering out rows with missing values. However, we might use SAS Enterprise Miner if the number of observations (output) after filtering out is significant. This ensures that the remaining data set contains only complete and reliable information, preserving the integrity of the available data.

To filter out these missing values, we drag and drop tFilterRow to the workspace. In the Condition field of **tFilterRow**, we specify the condition to filter out rows with missing values. For example, we filter out rows where the columns have missing values, as demonstrated in Figure 7 below. After that, we drag two tFileOutputDelimited to the workspace. This component is used to save the filtered results. Then, we link the tFilterInputDelimited to tFilterRow, and then link to tFileOutputDelimited_1. Meanwhile, the tFileOutputDelimited_3 is the rejected case for null values. We run the jobs after linking all these components. This process is shown in Figure 8.

Figure 9 displays that the rejected case is stored under the name as output2.csv, while for the cleaned data without any null values is stored under output1.csv. On the other hand, Figure 10 shows that there are 5 observations that have been filtered out. Since the deleted observations are 5 out of 205. We will proceed the process for "specify variable roles" under SAS Enterprise Miner.

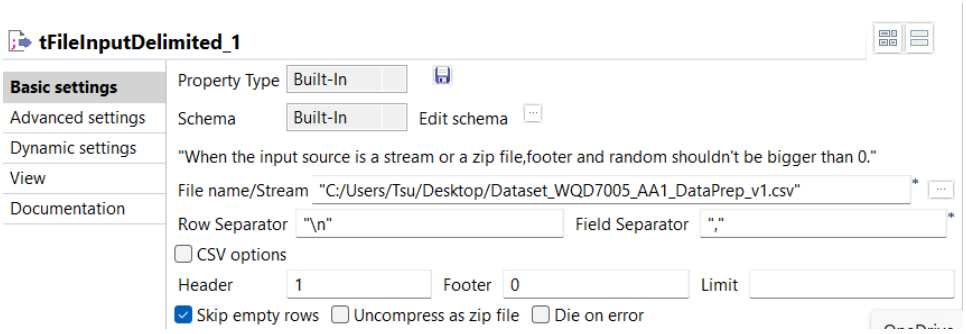

Figure 5: The Component Panel under tFileInputDelimited_1.

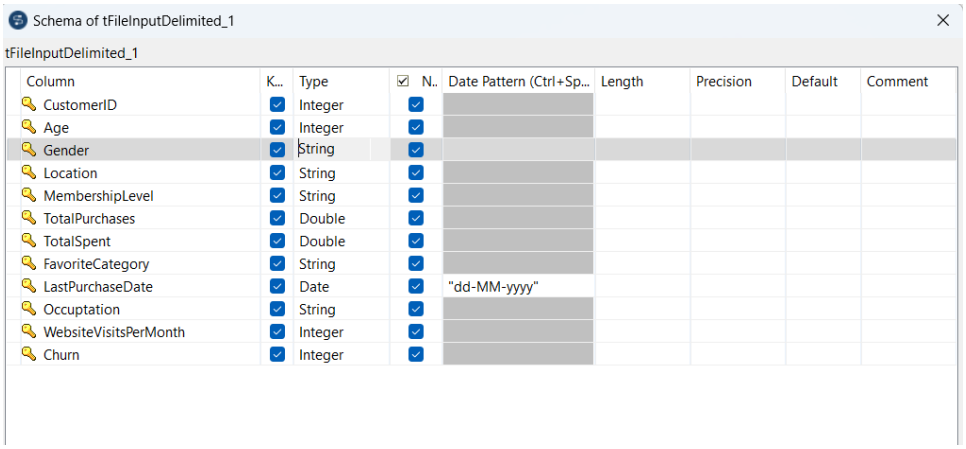

Figure 6: Scheme of tFileInputDelimited_1.

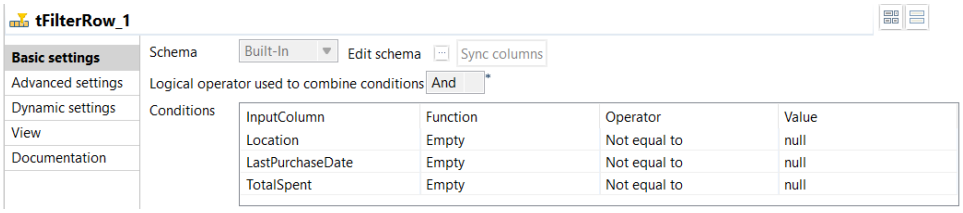

Figure 7: tFilterRow.

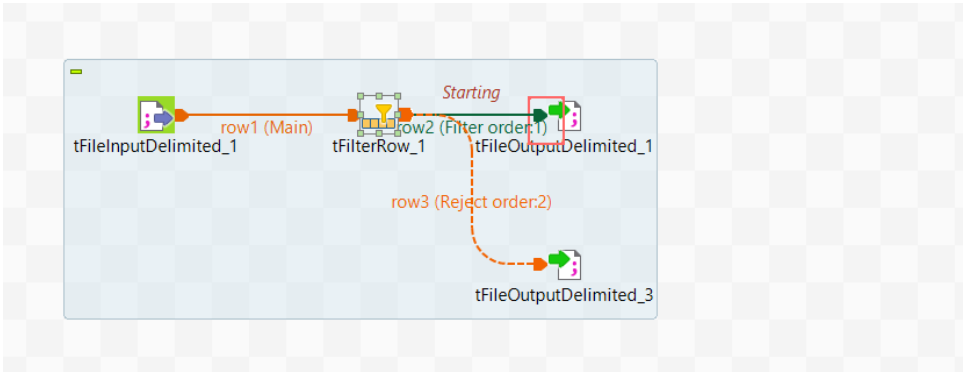

Figure 8: The Mapping.

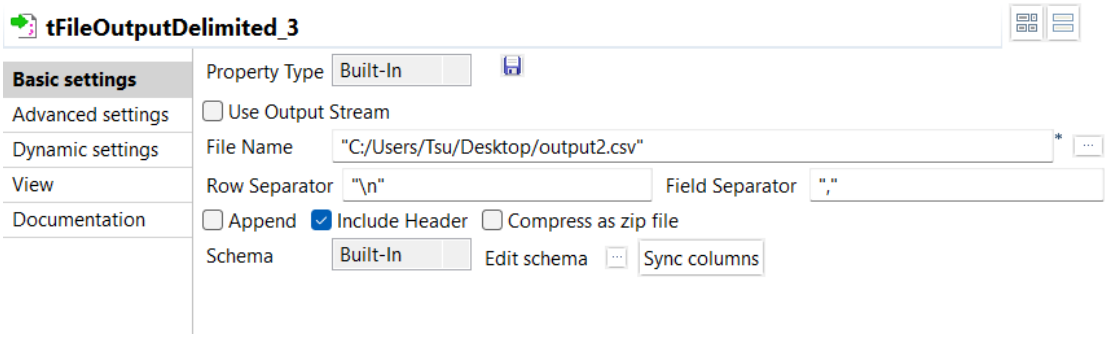

Figure 9: tFileOutputDelimited_3 and the File name is output2.csv.

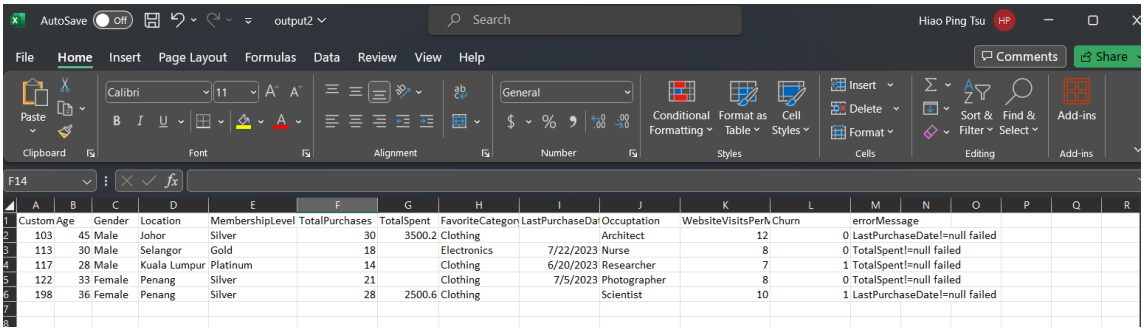

Figure 10: output2.csv (5 observations are filtered out).

SAS Enterprise Miner

As mentioned above, the cleaned data is named as "output1.csv". We created a new project and named it "AA1_Tsu Hiao Ping", as shown in Figure 11. The challenge faced during executing SAS Enterprise Minder is mapping CSV data to SAS variables is a critical step in the integration process. We will use CSV file, instead of in the format under "SAS dataset". To import csv file, we first create a new diagram named "csv" and then introduce "File Import" component to carry our CSV file (output1.csv). This step is demonstrated in Figure 12.

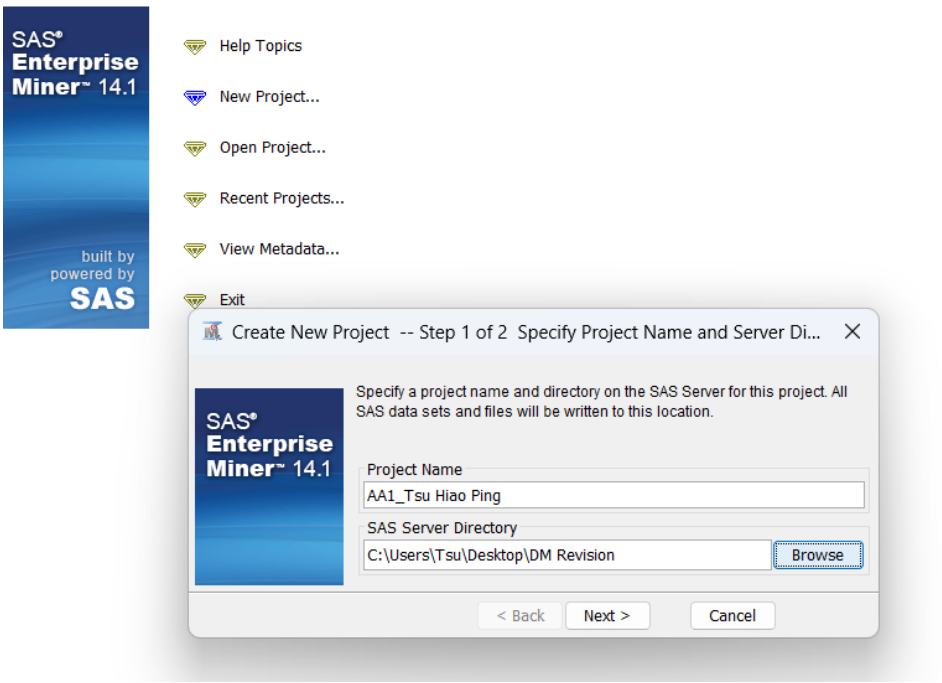

Figure 11: Create New Project for running SAS Enterprise Miner.

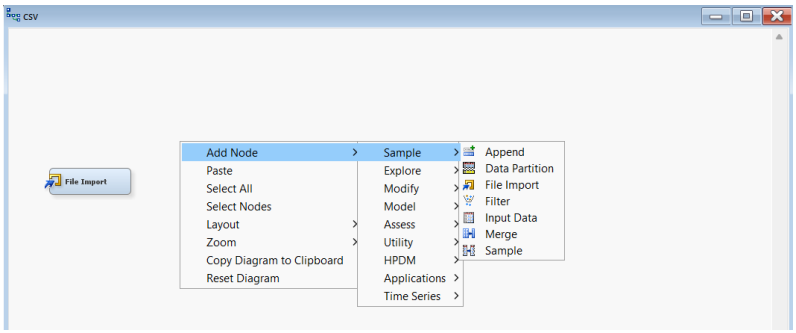

Figure 12: Importing CSV file.

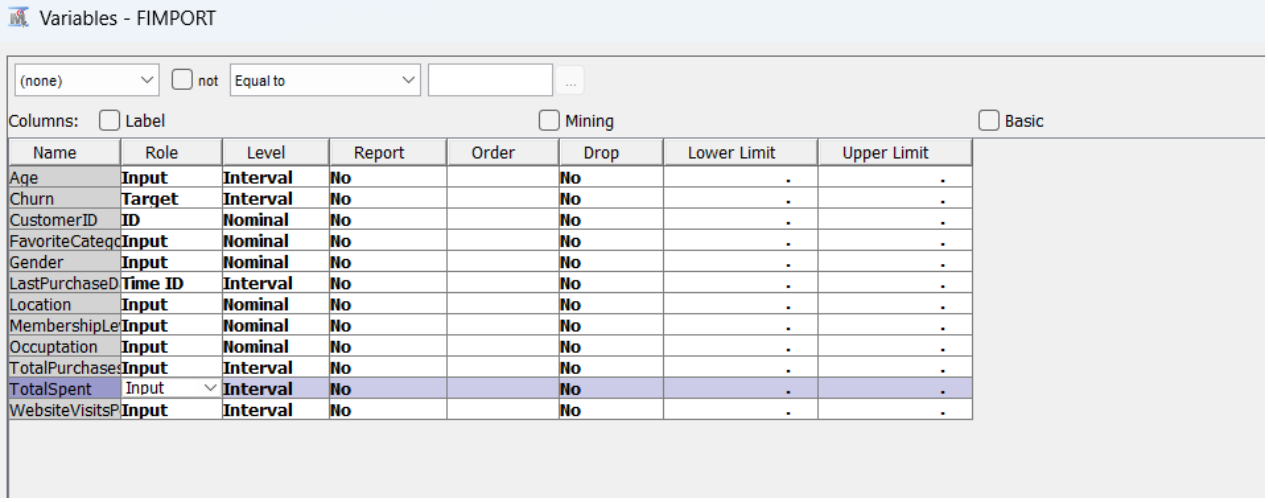

Figure 13: Specifying variable roles.

The next step is to specify the variable roles. I have adjusted the roles as shown in Figure 13 above.

In my dataset, the term "Churn" refers to the phenomenon where customers stop buying or interacting with the e-commerce platform. The goal of churn analysis is often to identify patterns and factors that are associated with customer attrition. By understanding the characteristics and behavior of customers who churn, businesses can implement targeted strategies to retain customers and reduce overall churn rates. The attributes related to churn analysis in my dataset might include:

- **Churn (Target Variable):**

This is likely a binary variable where 1 indicates that a customer has stopped making purchases or interacting with the platform (churned), and 0 indicates that the customer is still active.

- **LastPurchaseDate:**

The date of the last purchase made by the customer. This attribute can be useful in determining the recency of customer activity.

- **TotalPurchases:**

The total number of purchases made by the customer. A decrease in this value over time may be an indication of potential churn.

- **TotalSpent:**

The total amount spent by the customer. Similar to TotalPurchases, a decrease in spending may signal churn.

- **WebsiteVisitsPerMonth:**

The frequency of website visits per month. A decline in visits may suggest decreasing engagement, potentially leading to churn.

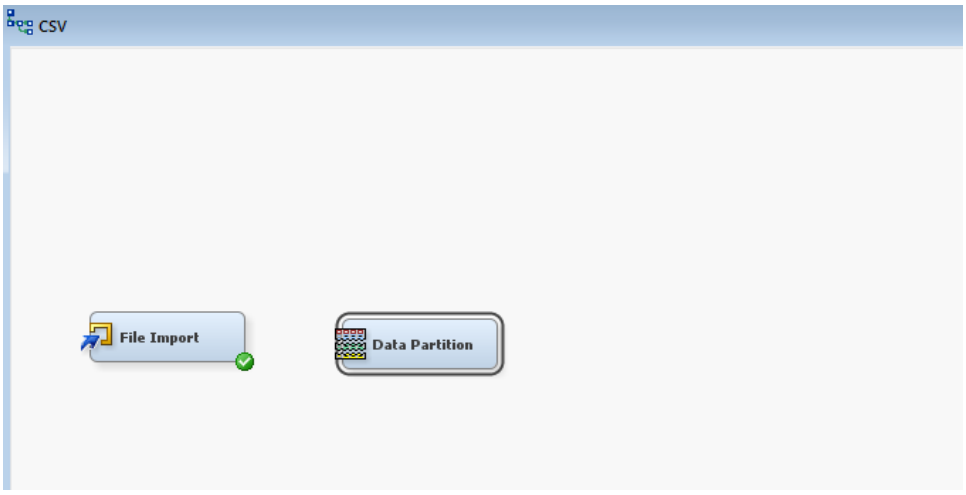

Figure 14: Data Partition N ode.

As shown in Figure 14, we use the "Data Partition" node to split the dataset into training, validation, and test sets. This is essential for model building and evaluation. The node allows to customize the proportions of the dataset allocated to training and validation. In this report, we allocate 70% of the data for training and 30% for validation.

When working with the "Data Partition" node in SAS Enterprise Miner, we configure the roles of variables by double-clicking on the node to access its settings. In the "Variables" tab, a comprehensive list of all variables in my dataset is presented. For each variable, we designate the role as "Input" for predictor variables, signifying the features or attributes used to predict the target variable. However, we did not reject attributes for this report. Additionally, the target variable, often representing binary outcomes such as '1' for churned and '0' for retained customers, is assigned the role of "Target".

We then drag “Decision Tree” under the “Model” panel and link it to Data Partition node. After that, we run the Decision Tree node, as shown in Figure 15 below. The variables under the Decision tree are displayed in Figure 16.

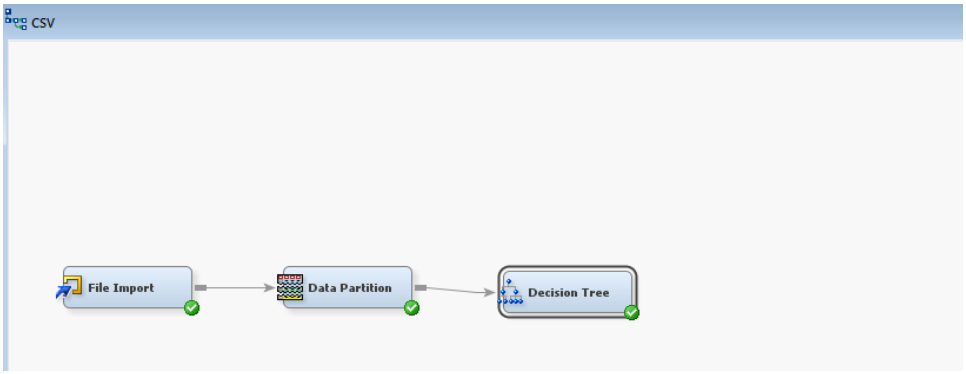

Figure 15: Decision Trees.

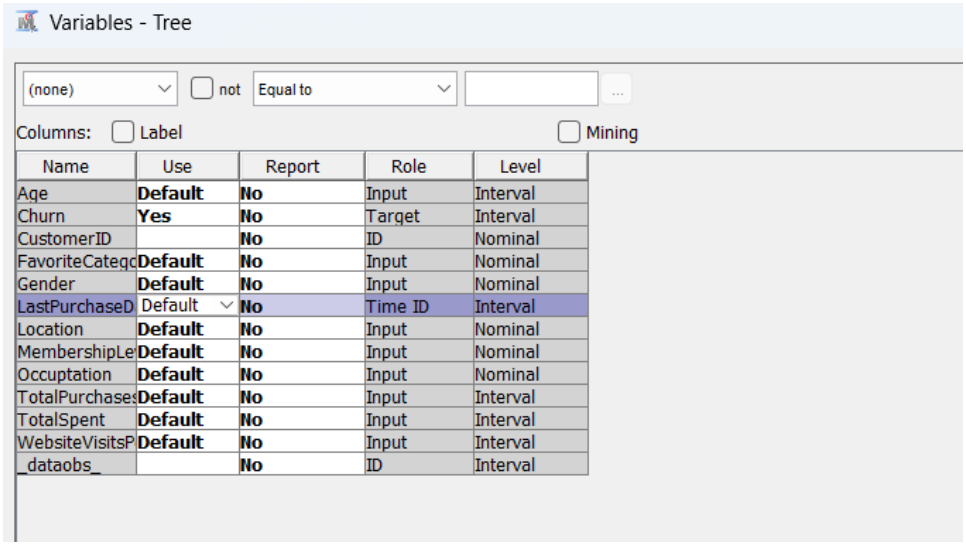

Figure 15: The Variables under Decision tree.

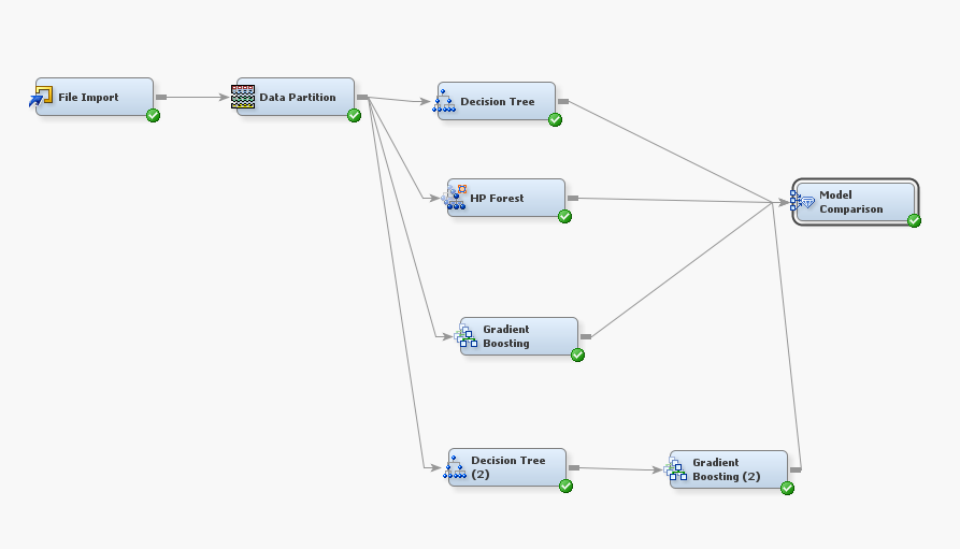

Figure 16: Adding HP Forest Node and Gradient Boosting.

The subsequent step is to add HP Forest node and Gradient Boosting for model comparison. Gradient boosting builds an ensemble of decision trees sequentially, correcting errors from previous trees. We have then added the fourth model, which is adding Gradient Boosting node sequentially after the decision tree model to help in underfitting issues. Meanwhile, random Forest is an ensemble method that builds multiple decision trees. It can increase the number of trees in the forest to enhance model complexity. It can also experiment with other hyperparameters to find the optimal balance between bias and variance. Since this Alternative Assessment (report) is asked to use Random Forest as a Bagging example, we proceed with building an ensemble of decision trees. Random Forest aggregates multiple decision trees, reducing the risk of underfitting and improving predictive performance.

Fit Statistics
Model Selection based on Valid: Average Squared Error (_VASE_)

Selected Model	Model Node	Model Description	Valid: Average Squared Error	Train: Average Squared Error
Y	Tree	Decision Tree	0.002937	0.027407
	Boost	Gradient Boosting	0.017837	0.023896
	HPDMForest	HP Forest	0.020683	0.036711

Figure 17: The average squared error.

Fit Statistics
Model Selection based on Valid: Average Squared Error (_VASE_)

Selected Model	Model Node	Model Description	Valid: Average Squared Error	Train: Average Squared Error
Y	Tree	Decision Tree	0.002937	0.027407
	Boost2	Gradient Boosting (2)	0.017552	0.026142
	Boost	Gradient Boosting	0.017837	0.023896
	HPDMForest	HP Forest	0.020683	0.036711

Figure 18: The average squared error when adding the fouth model.

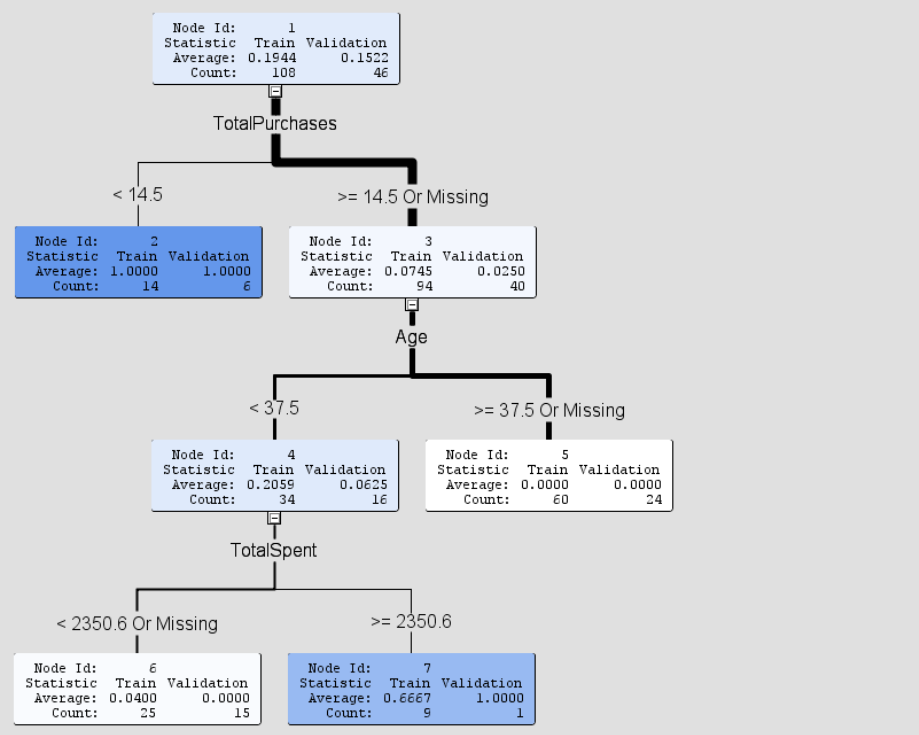

Figure 19: The Decision Tree Model (Underfitting).

However, under the Model Comparison node, the lowest average squared error for these three models is the decision tree model, as shown in Figure 17 and 18 above. Since the decision tree model (the first model) has the lowest average squared error among the Decision Tree, Random Forest, and Gradient Boosting models, it suggests that, in this specific case, the decision tree is performing better in terms of prediction accuracy when evaluated using mean squared error. The result of decision tree model is displayed in Figure 20 below. More details will be upload on GitHub Link.

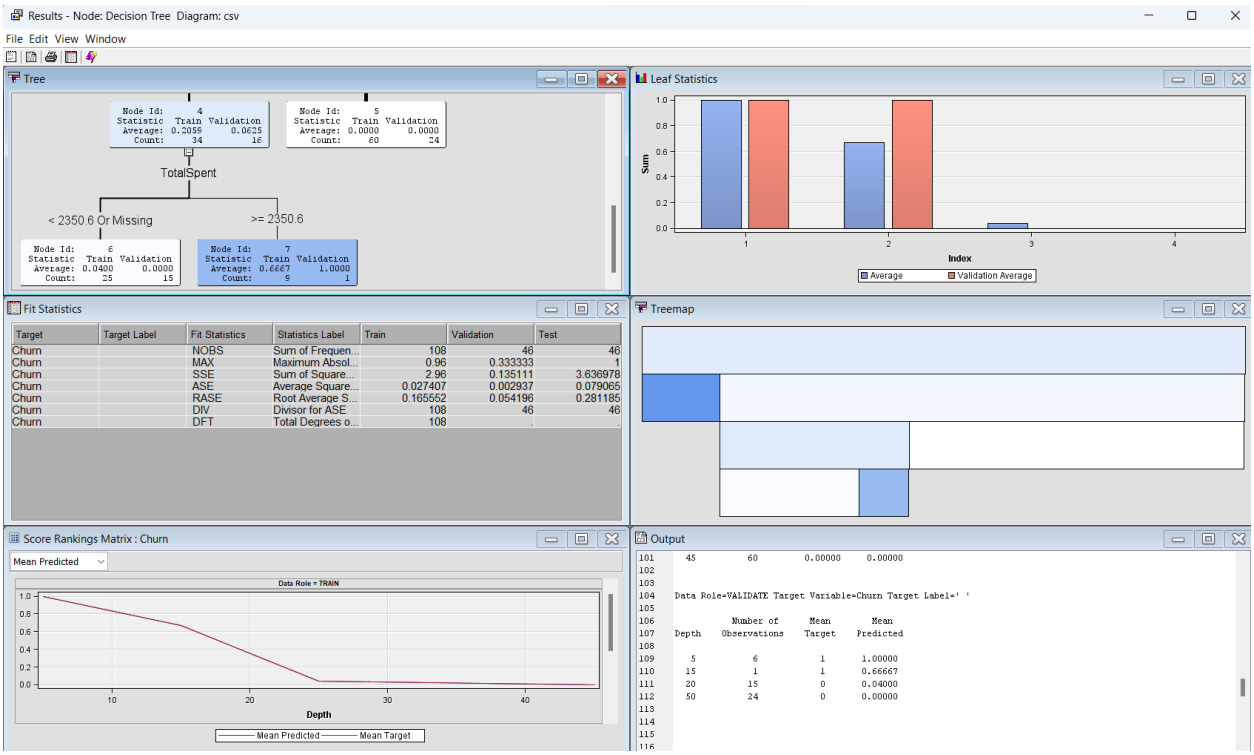

Figure 20: The Result of the Decision Model.

Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
TotalPurchases		1	1.0000	1.0000	1.0000
TotalSpent		1	0.4990	0.4784	0.9587
Age		1	0.2969	0.0000	0.0000

Figure 21: Variable Importance Plot.

Figure 19 and Figure 21 show that the variables that are significant for this analysis are the “TotalPurchases”, “Total Spent” and “Age”. The significance of the variables "TotalPurchases," "Total Spent," and "Age" in the context of churn analysis within an e-commerce platform dataset can be explained based on common patterns and behaviors observed in online retail environments. Combined with the findings in Talend Data Prep (*Figure 22 below*), the observation that LastPurchaseDate occurs most frequently in December 2023 may also contribute to the understanding of customer behavior during that period. The fact that LastPurchaseDate occurs most frequently in December aligns with common trends related to the holiday shopping season. December is commonly linked to heightened online shopping endeavors, driven by festive

occasions such as Christmas and New Year. During this period, customers tend to actively participate in the platform, making purchases for gifts, capitalizing on promotional offers, and participating in holiday sales events.

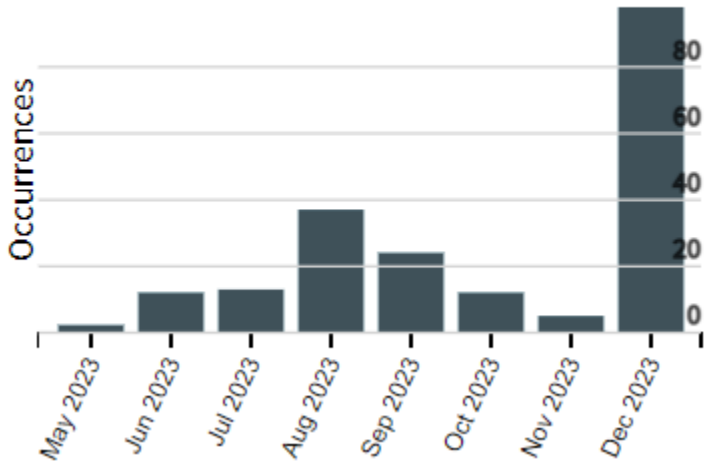

Figure 22: LastPurchaseDate occurs most frequently in December 2023.

To explain this, customers who make a higher number of total purchases are likely to be more engaged with the e-commerce platform. A lower total number of purchases might indicate a lack of interest or reduced engagement, which could be associated with potential churn.
Additionally, the total amount spent by a customer is a key indicator of their financial commitment to the platform. Customers who exhibit higher spending levels are typically more valuable to the business. Conversely, a reduction in total spending could indicate a diminishing customer value and may serve as an early signal of potential churn.

The third variable to be discussed is the Age. The age of the customer can be a relevant factor in churn analysis. Older customers who have been with the platform for an extended period may exhibit different behavior compared to newer customers. It is essential to comprehend the preferences and actions of customers in various phases of their association with the platform as it plays a critical role in anticipating and averting churn.

Regarding recommendations on business strategy, the platform may leverage the insights from significant variables to design targeted marketing campaigns. For example, offer personalized promotions to customers with declining Total Purchases or Total Spent to re-engage them. On top of that, considering customer segmentation based on their age groups to tailor marketing strategies

that resonate with different demographics. Younger customers may respond well to different incentives compared to older, more established customers. In addition, capitalizing on the knowledge that LastPurchaseDate peaks in December by planning special holiday promotions, discounts, or loyalty programs to encourage customer retention during this crucial shopping season.