

Yeo Yea Ern

S2165943

WQD7005 Alternative Assessment 1

Github link: https://github.com/YeaErn/wqd7005_aa1

Introduction

Customer churn analysis is essential in an e-commerce organisation to identify the customers who are more likely to churn. This is helpful for the company to identify the potential issues that result in customer churn and plan appropriate actions to be taken to secure customers' loyalty.

Objectives

1. To analyse the customer churn dataset using exploratory data analysis.
2. To clean the customer churn dataset thoroughly based on the findings in exploratory analysis.
3. To build a Decision Tree model and an ensemble model that predicts the churn of a customer given sets of attributes.
4. To assess the models built and select the best model for customer churn prediction.

Dataset

There are two datasets being used for customer churn analysis and prediction. Both are synthetic datasets. The first dataset consists of 19 independent variables, 1 target variable and 5630 observations in total. The variables are described in Table 1.

Table 1: First dataset variables description

Variables	Description
CustomerID	A unique identifier for each customer
Churn	An indicator of whether a customer has churned or not.
Tenure	Tenure of customer in the e-commerce
PreferredLoginDevice	The device preferred by customer to log into the platform
CityTier	City tier of where customer lives
WarehouseToHome	Distance between warehouse and customer's shipping address
PreferredPaymentMode	Payment method preferred by customer
Gender	Customer's gender
HourSpendOnApp	The number of hours spend on using the platform
NumberOfDeviceRegistered	Total number of devices registered to log into the platform
PreferredOrderCat	Customer's preferred order category in last year

SatisfactionScore	Satisfactory score of customers on the full service
MaritalStatus	Marital status of customer
NumberOfAddress	Total number of addresses recorded by customer
Complain	An indicator of whether a complaint has been raised in last year
OrderAmountHikeFromlastYear	Percentage increases in order from last year
CouponUsed	Total number of coupons used in year
OrderCount	Total number of orders placed in last year
DaySinceLastOrder	Number of days passes since customer's last order
CashbackAmount	Average cashback used in last year

The second dataset is generated using Python's Faker() library. It consists of 5 independent variables and 5630 observations in total. The variables are described in Table 2.

Table 2: Second dataset variables description

Variables	Description
CustomerID	A unique identifier for each customer
Name	Customer's name
Email	Customer's email address
Occupation	Customer's job
Phone	Customer's contact number

Data Preparation with Talend Data Preparation

The initial checking of data is performed on the first dataset by using the Talend Data Preparation tool. It is observed that 7 variables have missing values and 3 variables have different values referring to the same item. An example for the latter case is seen for "PreferredLoginDevice", in which there are "Mobile Phone", "Computer" and "Phone" present in the column. As "Mobile Phone" and "Phone" are technically referring to the same device, one of the value is replaced by the other to standardise the values. The "replace match string with..." technique in Talend Data Preparation tool is used to resolve such inconsistencies. The adjusted data is exported as .csv.

talend DATA PREPARATION

Customer_churn_dataset PREPARATION

1 Replace the cells that match on column PreferredPaymentMode

2 Replace the cells that match on column PreferredPaymentMode

3 Replace the cells that match on column PreferredOrderCat

4 Replace the cells that match on column PreferredLoginDevice

Current: Mobile Phone

Replacement: Phone

☒ Overwrite entire cell

SUBMIT

Filters

Add a filter ...

	Integer	PreferredLoginD... text	CityTler Integer	WarehouseToHo... Integer	PreferredPayme... text	Gender gender	HourSpenc
4	0	Phone	3	15	Debit Card	Male	
5	0	Phone	1	12	Credit Card	Male	
6	0	Computer	1	22	Debit Card	Female	
7		Phone	3	11	Cash on Delivery	Male	
8		Phone	1	6	Credit Card	Male	
9	13	Phone	3	9	E wallet	Male	
10		Phone	1	31	Debit Card	Male	
11	4	Phone	1	18	Cash on Delivery	Female	
12	11	Phone	1	6	Debit Card	Male	
13	0	Phone	1	11	Cash on Delivery	Male	
14	0	Phone	1	15	Credit Card	Male	
15	9	Phone	3	15	Credit Card	Male	
16		Phone	2	12	UPI	Male	
17	0	Computer	1	12	Debit Card	Female	
18	0	Phone	3	11	E wallet	Male	
19	0	Computer	1	13	Debit Card	Male	
20	19	Phone	1	20	Debit Card	Female	
21	0	Phone	3	12	Debit Card	Male	

PreferredLoginDevice

Find a function ...

Replacement:

☐ Overwrite entire cell

CHART VALUE PATTERN ADVANCED

ROW COUNT

0 1,000 2,000 3,000

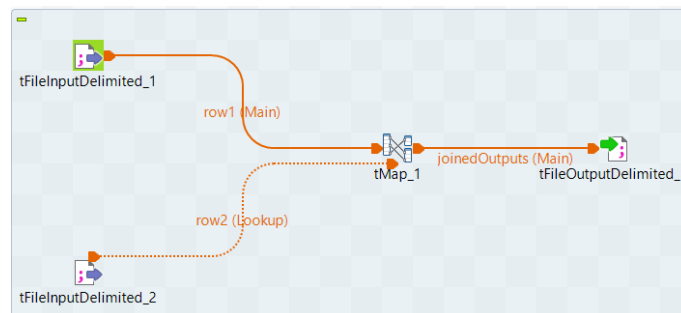
Phone

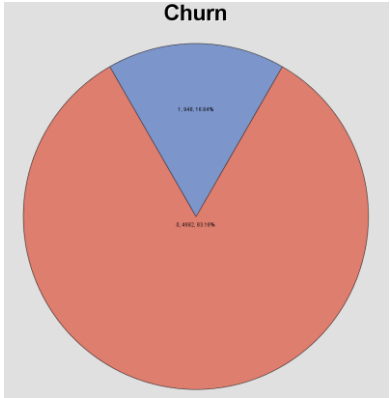
Computer

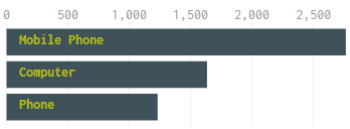

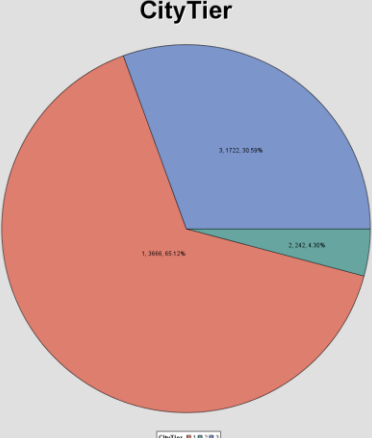
The findings, resolutions and outputs are summarised in Table 3.

Data Integration with Talend Data Integration

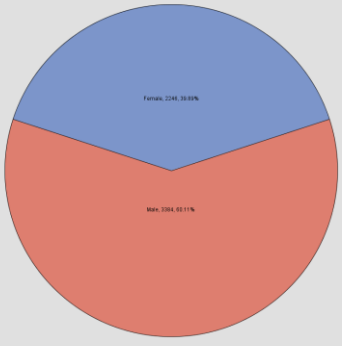
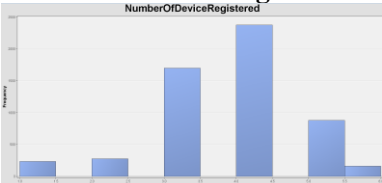
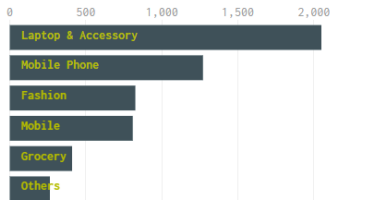
After preparing the first dataset, the Talend Data Integration tool is used to integrate first dataset and second dataset. Both files are first imported into the created job using “tFileInputDelimited” component. Each of the component has its schema being defined explicitly according to the dataset. Then, “tMap” is used to map each item in the dataset by using the “ConsumerID” key. The mapped output is being exported as .csv using the “tFileOutputDelimited” component.


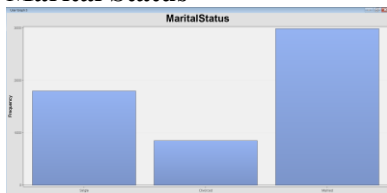
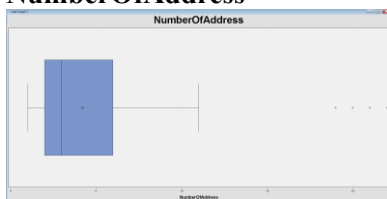


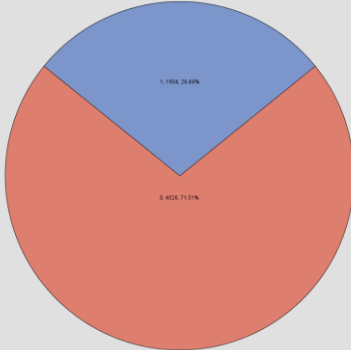
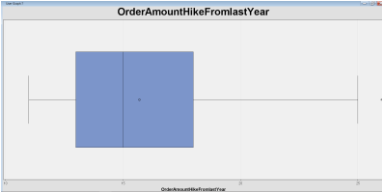
Variables	Findings	Resolutions	Outputs
CustomerID Count: 5630 Min: 50001 Distinct: 5630 Max: 55630 Duplicate: 0 Mean: 52815.5 Valid: 5630 Variance: 2641877.5 Empty: 0 Median: 52815.5 Invalid: 0 Lower quantile: 51407.75 Upper quantile: 54223.25	<ul style="list-style-type: none"> All customers are unique. No duplicate observations found. 	<ul style="list-style-type: none"> Duplication handling is not required. 	<ul style="list-style-type: none"> N/A
Churn 	<ul style="list-style-type: none"> 948 (16.84%) of churn = 1. 4682 (83.16%) of churn = 0. Target class is highly imbalance with majority class on non-churn customers. 	<ul style="list-style-type: none"> Reclassify the variable to binary. 	<ul style="list-style-type: none"> Reclassified to binary. Changed its role to Target.
Tenure Count: 5630 Min: 0 Distinct: 37 Max: 61 Duplicate: 5593 Mean: 10.19 Valid: 5366 Variance: 73.23 Empty: 264 Median: 9 Invalid: 0 Lower quantile: 2 Upper quantile: 16	<ul style="list-style-type: none"> 264 missing values. 	<ul style="list-style-type: none"> Model-based imputation. 	<ul style="list-style-type: none"> 264 missing values are imputed.


<p>PreferredLoginDevices</p> 	<ul style="list-style-type: none"> There are 3 types of devices being recorded. “Mobile Phone” and “Phone” should be categorised as one. 	<ul style="list-style-type: none"> Replace “Mobile Phone” with “Phone” <div data-bbox="1057 263 1518 595"> <p>Current:</p> <div>≡ Mobile Phone</div> <p>Replacement:</p> <div>Phone</div> <p><input checked="" type="checkbox"/> Overwrite entire cell</p> </div>	
<p>CityTier</p> 	<ul style="list-style-type: none"> There are 3 different city tier levels. 	<ul style="list-style-type: none"> Reclassify the variable to ordinal. 	<ul style="list-style-type: none"> Reclassified to ordinal.
<p>WarehouseToHome</p>	<ul style="list-style-type: none"> 251 missing values. 	<ul style="list-style-type: none"> Model-based imputation. 	<ul style="list-style-type: none"> 251 missing values are imputed.

<div>Count: 5630</div> <div>Min: 5</div> <div>Distinct: 35</div> <div>Max: 127</div> <div>Duplicate: 5595</div> <div>Mean: 15.64</div> <div>Valid: 5379</div> <div>Variance: 72.79</div> <div>Empty: 251</div> <div>Median: 14</div> <div>Invalid: 0</div> <div>Lower quantile: 9</div> <div>Upper quantile: 20</div>			
<div><div>PreferredPaymentMethod</div><div><div><div>0</div><div>500</div><div>1,000</div><div>1,500</div><div>2,000</div></div><div><div>Debit Card</div><div>Credit Card</div><div>E wallet</div><div>UPI</div><div>COD</div><div>CC</div><div>Cash on Delivery</div></div></div></div>	<ul style="list-style-type: none">7 types of payment methods are recorded, but 2 of which are repeating types.“CC” and “Credit Card” “should be categorised as one.“COD” and “Cash on Delivery” should be categorised as one.	<ul style="list-style-type: none">Replace “CC” with “Credit Card”. <div><div>replace</div><div><div>Current:</div><div><div>=</div><div>CC</div></div></div><div><div>Replacement:</div><div>Credit Card</div></div><div><div><input checked="" type="checkbox"/></div>Overwrite entire cell</div></div> <ul style="list-style-type: none">Replace “COD” with “Cas on Delivery”. <div><div>replace</div><div><div>Current:</div><div><div>=</div><div>COD</div></div></div><div><div>Replacement:</div><div>Cash on Delivery</div></div><div><div><input checked="" type="checkbox"/></div>Overwrite entire cell</div></div>	<ul style="list-style-type: none">5 types of payment methods. <div><div><div>0</div><div>500</div><div>1,000</div><div>1,500</div><div>2,000</div></div><div><div>Debit Card</div><div>Credit Card</div><div>E wallet</div><div>Cash on Delivery</div><div>UPI</div></div></div>
<div>Gender</div>	<ul style="list-style-type: none">Male customers are 20% more than female customers.	<ul style="list-style-type: none">N/A	<ul style="list-style-type: none">N/A

<p>Gender</p>  <p>Gender: 0 Female 1 Male</p>			
<p>HourSpendOnApp</p> <p>Count: 5630 Min: 0</p> <p>Distinct: 7 Max: 5</p> <p>Duplicate: 5623 Mean: 2.93</p> <p>Valid: 5375 Variance: 0.52</p> <p>Empty: 255 Median: 3</p> <p>Invalid: 0 Lower quantile: 2</p> <p>Upper quantile: 3</p>	<ul style="list-style-type: none"> 255 missing values. 	<ul style="list-style-type: none"> Model-based imputation. 	<ul style="list-style-type: none"> 255 missing values are imputed.
<p>NumberOfDeviceRegistered</p> 	<ul style="list-style-type: none"> Most people registered 3 to 4 devices. 	<ul style="list-style-type: none"> N/A 	<ul style="list-style-type: none"> N/A
<p>PreferredOrderCat</p> 	<ul style="list-style-type: none"> 6 types of categories are recorded. “Mobile Phone” and “Mobile” should be categorised as one. 	<ul style="list-style-type: none"> Replace “Mobile Phone” with “Mobile” 	<ul style="list-style-type: none"> 5 types of order categories.

		<div>Current: Mobile Phone</div> <div>Replacement: Mobile</div> <div><input checked="" type="checkbox"/> Overwrite entire cell</div>	<div><div>05001,0001,5002,000</div><div>Mobile</div><div>Laptop & Accessory</div><div>Fashion</div><div>Grocery</div><div>Others</div></div> <div><ul style="list-style-type: none">“Mobile” and “Laptop & Accessory” being the top two leading preferred order categories.</div>										
<div>SatisfactionScore</div> <div></div>	<ul style="list-style-type: none">5 different satisfaction score ranking from 1 through 5.Score of 3 being the most frequently rated.	<ul style="list-style-type: none">Reclassify the variable to ordinal.	<ul style="list-style-type: none">Reclassified to ordinal.										
<div>Marital Status</div> <div></div>	<ul style="list-style-type: none">3 different marital statuses.Customers who are married are twice as much as single, which are twice as much as divorced customers.	<ul style="list-style-type: none">N/A	<ul style="list-style-type: none">N/A										
<div>NumberOfAddress</div> <div></div>	<ul style="list-style-type: none">4 potential outliers identified.	<ul style="list-style-type: none">Replace the outliers with “Missing” using “Replacement” node.Then impute the “Missing” values with model-based imputation.	<div><ul style="list-style-type: none">Replacement<div>Replacement Counts</div><table><tr><th>Obs</th><th>Variable</th><th>Label</th><th>Role</th><th>Train</th></tr><tr><td>1</td><td>NumberOfAddress</td><td>NumberOfAddress</td><td>INPUT</td><td>4</td></tr></table><ul style="list-style-type: none">Imputation – 4 missing values are imputed.</div>	Obs	Variable	Label	Role	Train	1	NumberOfAddress	NumberOfAddress	INPUT	4
Obs	Variable	Label	Role	Train									
1	NumberOfAddress	NumberOfAddress	INPUT	4									
<div>Complain</div>	<ul style="list-style-type: none">A binary class variable.1604 (28.49%) customers had	<ul style="list-style-type: none">Reclassify the variable’s level to binary.	<ul style="list-style-type: none">Reclassified to binary.										

<p>Complain</p> 	<p>complained in the last year</p> <ul style="list-style-type: none"> 4026 (71.51%) customers had not complained in the last year 		
<p>OrderAmountHikeFromlastYear</p> <p>Count: 5630 Min: 11 Distinct: 17 Max: 26 Duplicate: 5613 Mean: 15.71 Valid: 5365 Variance: 13.51 Empty: 265 Median: 15 Invalid: 0 Lower quantile: 13 Upper quantile: 18</p> 	<ul style="list-style-type: none"> 265 missing values. The one value out of the box-plot is not considered as outlier since its value (26) is close to the maximum (25). 	<ul style="list-style-type: none"> Model-based imputation. 	<ul style="list-style-type: none"> 265 missing values are imputed.
<p>CouponUsed</p>	<ul style="list-style-type: none"> 256 missing values. 	<ul style="list-style-type: none"> Model-based imputation. 	<ul style="list-style-type: none"> 256 missing values are imputed.

<p>Count: 5630 Min: 0</p> <p>Distinct: 18 Max: 16</p> <p>Duplicate: 5612 Mean: 1.75</p> <p>Valid: 5374 Variance: 3.59</p> <p>Empty: 256 Median: 1</p> <p>Invalid: 0 Lower quantile: 1</p> <p>Upper quantile: 2</p>			
<p>OrderCount</p> <p>Count: 5630 Min: 1</p> <p>Distinct: 17 Max: 16</p> <p>Duplicate: 5613 Mean: 3.01</p> <p>Valid: 5372 Variance: 8.64</p> <p>Empty: 258 Median: 2</p> <p>Invalid: 0 Lower quantile: 1</p> <p>Upper quantile: 3</p>	<ul style="list-style-type: none"> • 258 missing values. 	<ul style="list-style-type: none"> • Model-based imputation. 	<ul style="list-style-type: none"> • 258 missing values are imputed.
<p>DaySinceLastOrder</p> <p>Count: 5630 Min: 0</p> <p>Distinct: 23 Max: 46</p> <p>Duplicate: 5607 Mean: 4.54</p> <p>Valid: 5323 Variance: 13.35</p> <p>Empty: 307 Median: 3</p> <p>Invalid: 0 Lower quantile: 2</p> <p>Upper quantile: 7</p> 	<ul style="list-style-type: none"> • 307 missing values. • There are 2 values with days > 30, but it is still within acceptable range and should not be consider as outlier. 	<ul style="list-style-type: none"> • Model-based imputation. 	<ul style="list-style-type: none"> • 307 missing values are imputed.

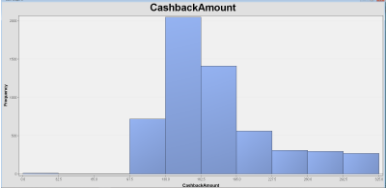
CashbackAmount 	<ul style="list-style-type: none"> • Most customers used cashback amount over 97.5 last year. • Only 9 customers used cashback amount less than 32.5. 	<ul style="list-style-type: none"> • N/A 	<ul style="list-style-type: none"> • N/A
VAR25	<ul style="list-style-type: none"> • A new column that is introduced after importing data into the library. • It is irrelevant to the customer churn analysis at any point 	<ul style="list-style-type: none"> • Reclassify the variable to rejected. 	<ul style="list-style-type: none"> • Rejected.
Name Email Phone	<ul style="list-style-type: none"> • Unique value in each observation 	<ul style="list-style-type: none"> • Reclassify the variable to rejected. 	<ul style="list-style-type: none"> • Rejected.

Table 3: Findings of each important variable

Reclassification of Role and Level of Variables

After the exploratory data analysis, it is identified that certain variables need to have their role or level being reclassified. The role and level reclassification of these variables are shown below.

Variables	Role	Level		Variables	Role	Level
Churn	Input	Interval	➔	Churn	Target	Binary
CityTier	Input	Interval		CityTier	Input	Ordinal
Complain	Input	Interval		Complain	Input	Binary
CustomerID	Input	Interval		CustomerID	Rejected	Nominal
SatisfactionScore	Input	Interval		SatisfactionScore	Input	Ordinal
VAR25	Input	Interval		VAR25	Rejected	Interval
Name	Input	Nominal		Name	Rejected	Nominal
Email	Input	Nominal		Email	Rejected	Nominal
Phone	Input	Nominal		Phone	Rejected	Nominal

Using tree-imputation for missing input variables

A model-based imputation technique is chosen because this method imputes missing values based on a model fitted to generate the value for replacement. It is more flexible, adaptive and accurate than other imputation methods such as mode and mean imputation. Mode imputation is usually used for categorical variables and therefore it is not applicable to the missing value cases here. On the other hand, mean imputation is heavily dependent on the data distribution and therefore an incorrect value may be imputed if the data is skewed.

- Tree-based imputation settings and configurations

Property	Value
General	
Node ID	Impt
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Nonmissing Variables	No
Missing Cutoff	50.0
Class Variables	
Default Input Method	Count
Default Target Method	None
Normalize Values	Yes
Interval Variables	
Default Input Method	Tree
Default Target Method	None
Default Constant Value	
Default Character Value	
Default Number Value	.
Method Options	

Tree Imputation

Property	Value
Leaf Size	5
Maximum Branch	2
Maximum Depth	6
Minimum Categorical Size	2
Number of Rules	5
Number of Surrogate Rules	2
Split Size	.

Minimum Categorical Size
Minimum number of observations for a categorical value. A category must occur in at least the specified number of observations for it to be used in the split search.

OK
Cancel

- Tree-based imputation result

Imputation Summary
Number Of Observations

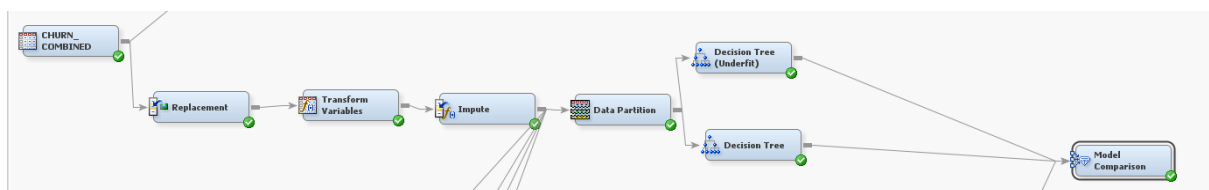
Variable Name	Impute Method	Imputed Variable	Impute Value	Role	Measurement Level	Label	Number of Missing for TRAIN
CouponUsed	TREE	IMP_CouponUsed	.	INPUT	INTERVAL		256
DaySinceLastOrder	TREE	IMP_DaySinceLastOrder	.	INPUT	INTERVAL		307
HourSpendOnApp	TREE	IMP_HourSpendOnApp	.	INPUT	INTERVAL		255
OrderAmountHikeFromlastYear	TREE	IMP_OrderAmountHikeFromlastYear	.	INPUT	INTERVAL		265
OrderCount	TREE	IMP_OrderCount	.	INPUT	INTERVAL		258
REP_NumberOfAddress	TREE	IMP_REP_NumberOfAddress	.	INPUT	INTERVAL	Replacement: NumberOfAddress	4
Tenure	TREE	IMP_Tenure	.	INPUT	INTERVAL		264
WarehouseToHome	TREE	IMP_WarehouseToHome	.	INPUT	INTERVAL		251

Analysing customer behaviour using Decision Tree model

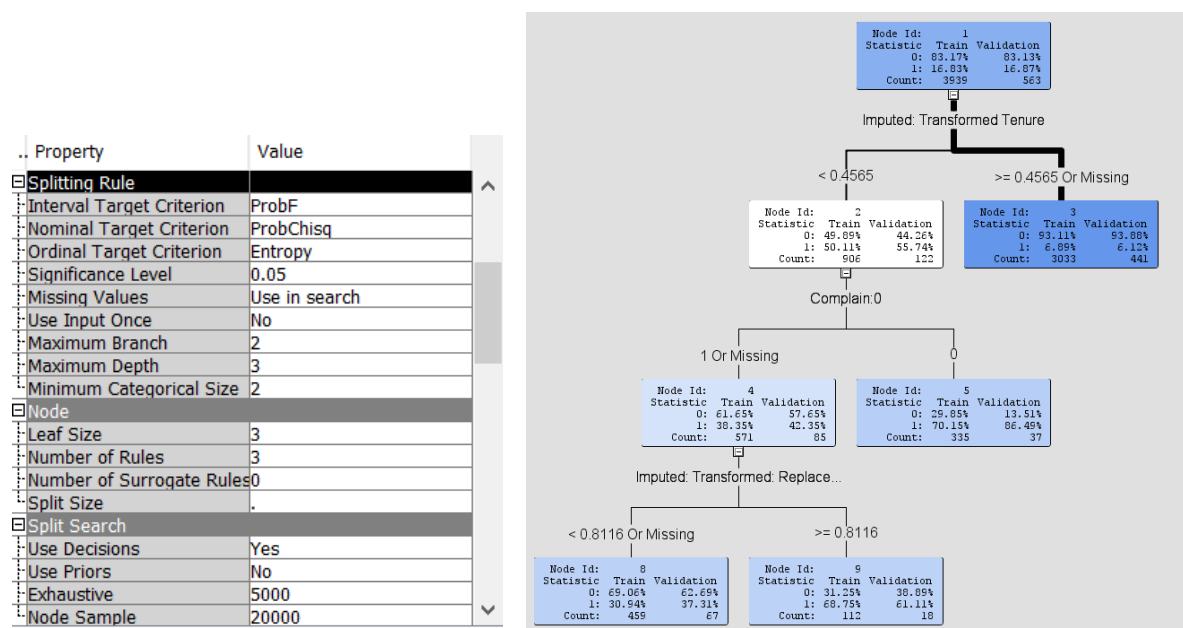
After a series of data preprocessing steps, two decision tree models are built upon the entire data. One decision is used to demonstrate underfitting scenario, while the other decision tree is to simulate the model at its optimum parameters.

In this case, the full dataset is used because each observation in the dataset is unique. Moreover, the dataset size is just right to be handled by the machine used to operate all the analysis, preprocessing and modelling tasks. Before the modeling phase, the data is split into training, validation and testing set in a ratio of 7:1:2. A validation set is used to fine-tune the model and to ensure the model is not biased. The model is fine-tuned a few times with different values of maximum branch, maximum depth, minimum categorical size, leaf size and number of rules to reach to identify the optimum parameters.

The figure below shows the steps towards building decision tree models and finally the assessment.

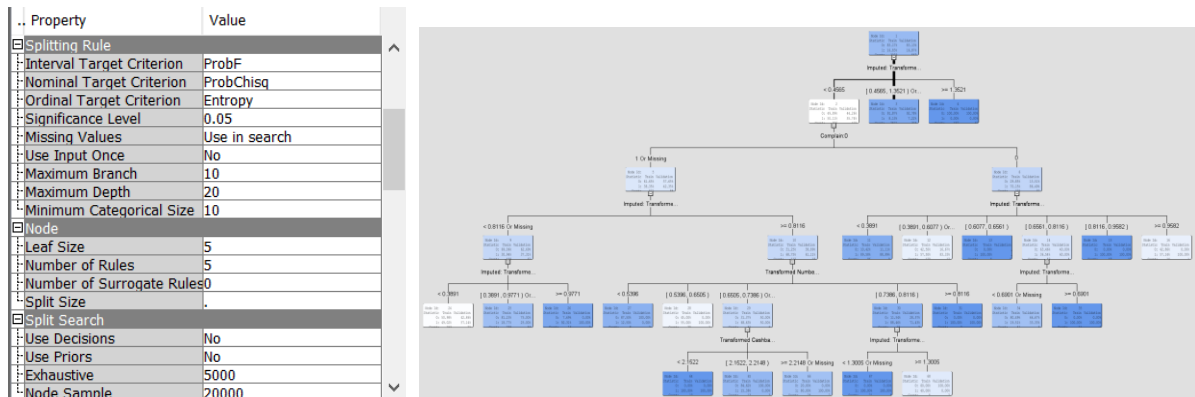


- Underfitting decision tree:



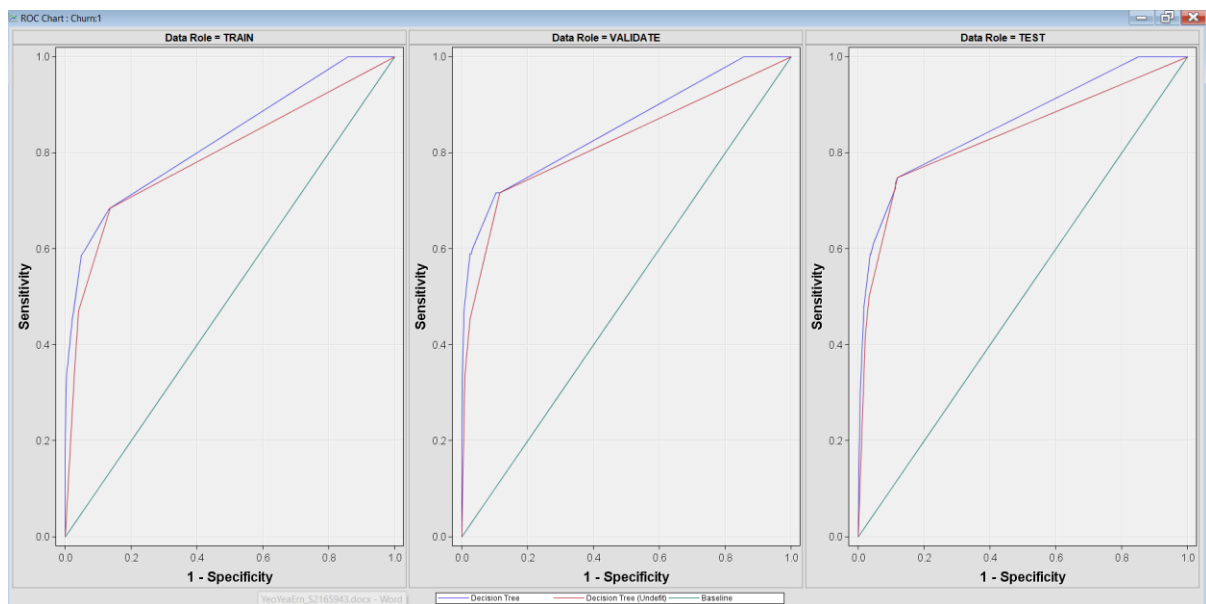
The figures show that the decision tree is very shallow.

- A more comprehensive decision tree:



The figures show that the decision tree is more in depth with more rules.

- Comparison of optimum decision tree and underfitted decision tree:



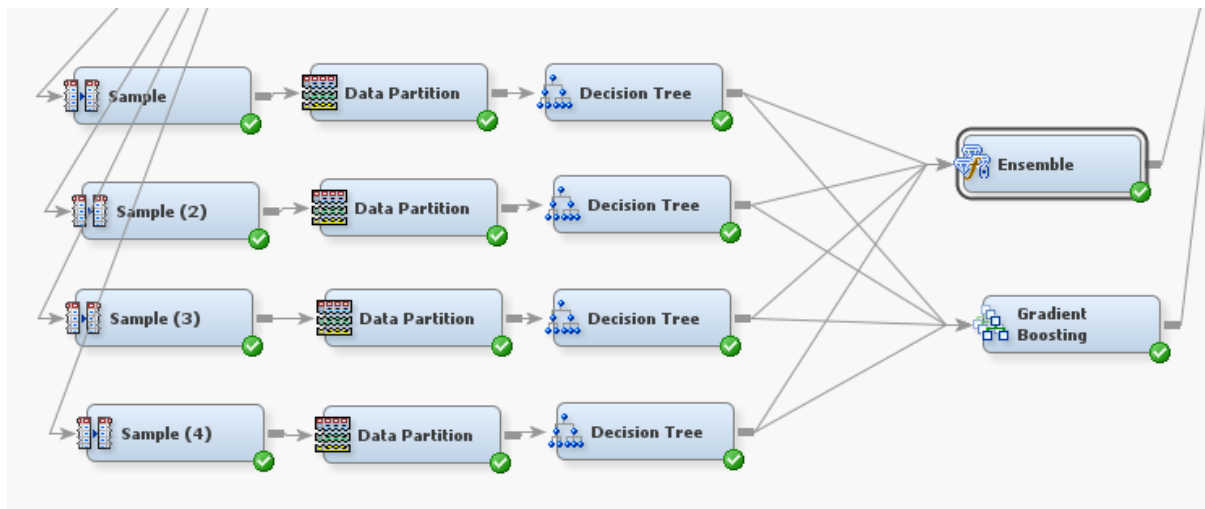
This graph proves that a decision tree with optimum depths and branches (blue line) has a better performance than a decision tree that is underfit (red line).

- Analysis of the optimum decision tree:

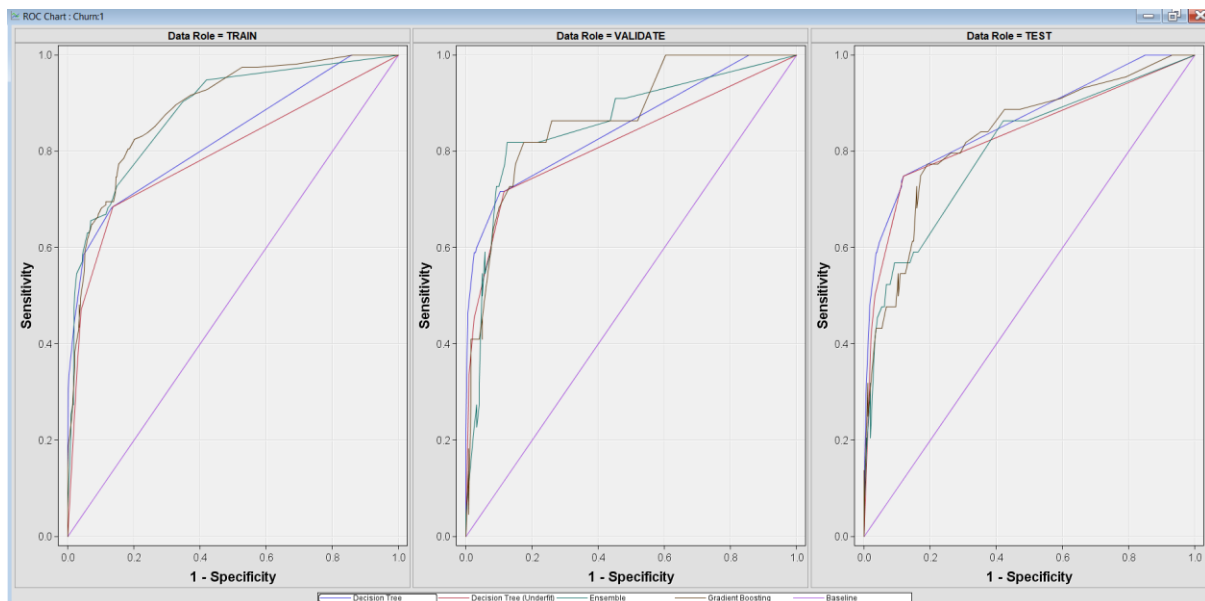
The variables that play significant roles in determining a customer's churn are Tenure, Complain, NumberOfAddresses, DaySinceLastOrder, NumberOfDeviceRegistered, CashbackAmount and OrderAmountHikeFromLastYear

Analysing customer behaviour using Bagging/Boosting method

An ensemble model is built based on 4 decision trees as shown below. Random sampling method is used to sample 25% of the whole dataset on each sample node. The same Data Partition and Decision Tree node are copied over from the original nodes used to train the models earlier. The Decision Tree node selected is the one that has a better performance in the test earlier.



- Comparison of optimum decision tree, underfitted decision tree, bagging and boosting method:



The ensemble method, bagging, is usually used to reduce overfitting in a model. As shown in the graph above, the bagging method (green line) and boosting method (brown line) achieve a better result in the training and validation dataset as compared to the two decision tree models earlier. However, when evaluating against the test dataset, the bagging method shows a significant drop in the performance whereas the boosting method shows a slight drop. This indicates that the bagging and boosting method are not ideal to be used in this dataset for customer churn analysis especially when the dataset is imbalanced.

Business Insights

In short, the optimum decision tree is the best model that can be used to identify customer churn. The variables that give significant impact to customer churn include Tenure, Complain, NumberOfAddresses, DaySinceLastOrder, NumberOfDeviceRegistered, CashbackAmount and OrderAmountHikeFromLastYear.

While Tenure being the most impactful attribute, it can be inferred that having membership feature and membership promotions or discounts will be advantageous to maintain customers' loyalty.

Reviewing and understanding customer complaints towards the platform or the products is beneficial for the organization to enhance their services and maintain their products quality to keep customers from churning.

Analysing the cashback amount helps to assess the appropriate cashback amount to keep the company profitable while still be able to please the customers and prevent them from churning.

Reflections and Learning Outcomes

This project's main purpose is to apply different data mining tools such as SAS-EM, Talend Data Integration and Talend Data Preparation, to analyse the customer churn of an e-commerce dataset. The use of tools instead of coding from scratch has indeed simplified the data mining problems. In this project, I have learned how to inspect and prepare data with Talend Data Preparation tool. I have also gotten the skill to integrate different data sources into one output file using the Talend Data Preparation tool. Undoubtedly, knowing how to apply the SEMMA methodology in SAS-EM is the biggest achievement that I have achieved.

From the analysis, I discovered that bagging or boosting methods do not always outperform a decision tree. While many people claim that Random Forest (ensemble of trees with bagging/boosting) is better than decision tree, I believe it is a situation dependent on the datasets and how the tree parameters are being set up.

The most challenging part of the project is the limited time given to conduct the entire SEMMA methodology in one day, alongside producing a well-structured report. Due to the time-constraint, it is difficult to explore the tools to their fullest.

Appendix

Talend Data Preparation

Preparing to replace values of specific columns

1

Replace the cells that match on column PreferredPaymentMode

2

Replace the cells that match on column PreferredPaymentMode

3

Replace the cells that match on column PreferredOrderCat

4

Replace the cells that match on column PreferredLoginDevice

Export Preparation File to Local CSV

EXPORT TO CSV

Delimiter:

Comma

Filename:

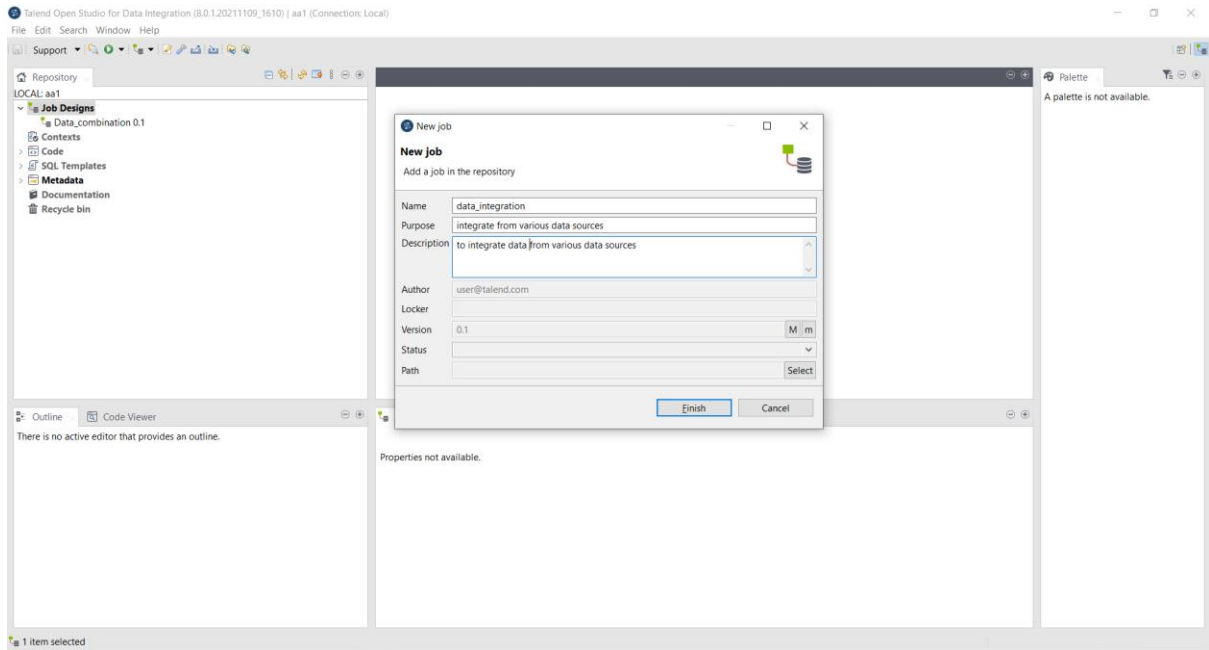
Customer_churn_dataset_prepared

CANCEL

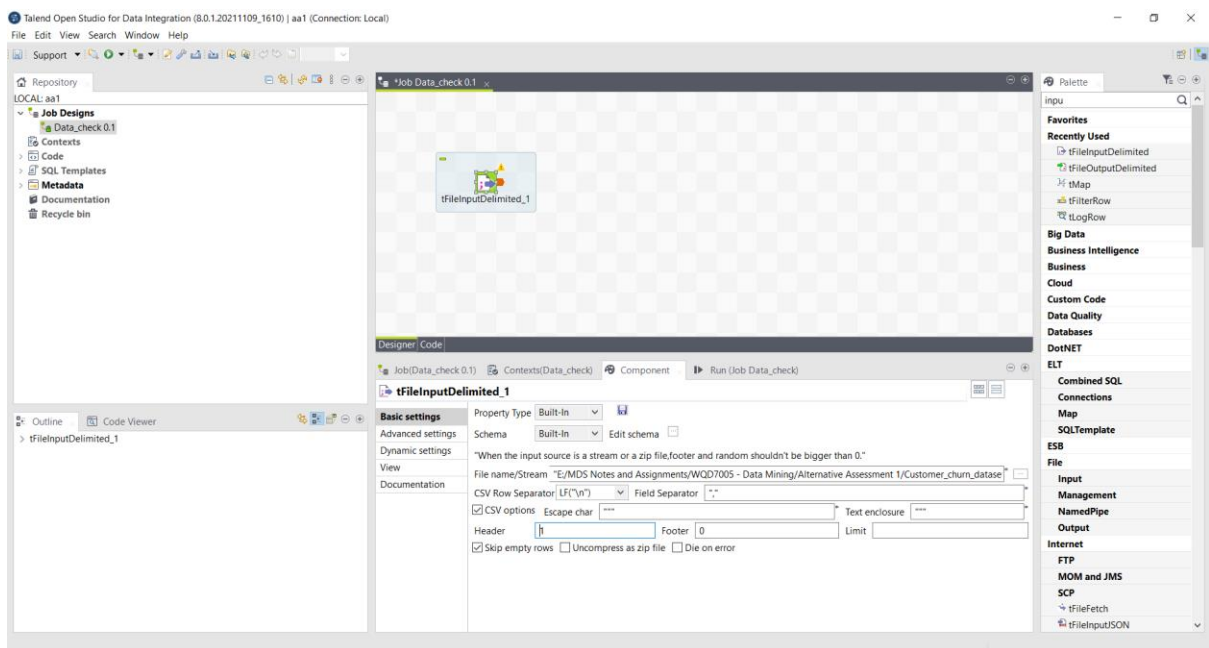
EXPORT

Talend Data Integration

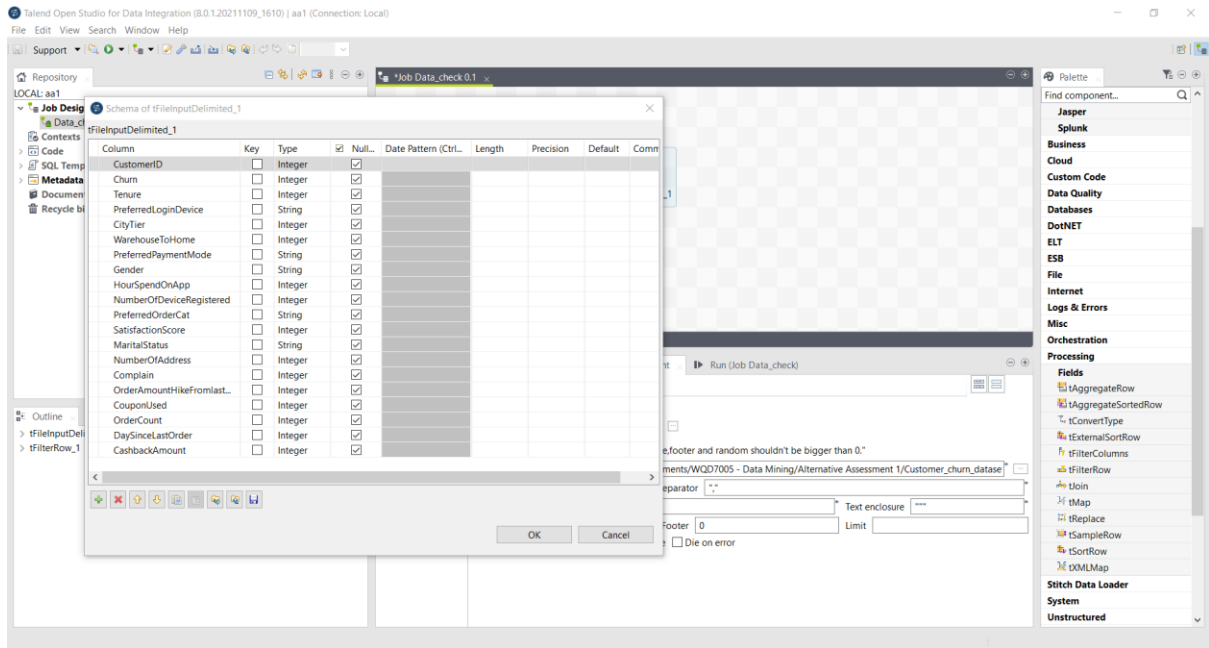
Create new job.



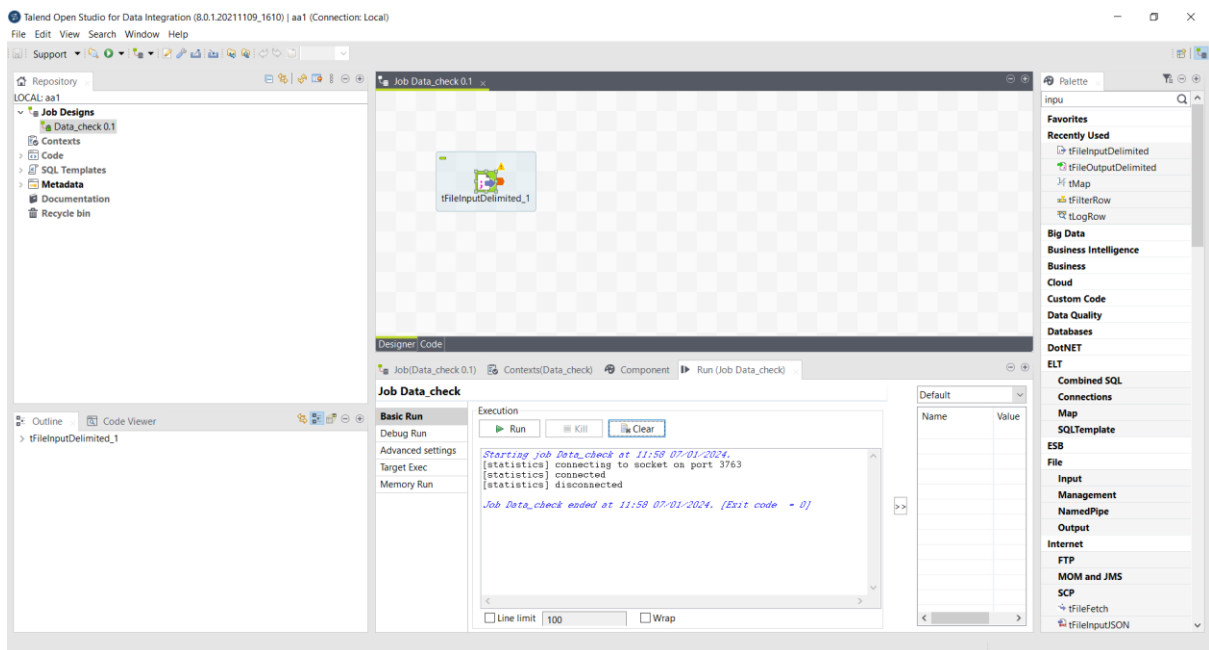
Import file to job.



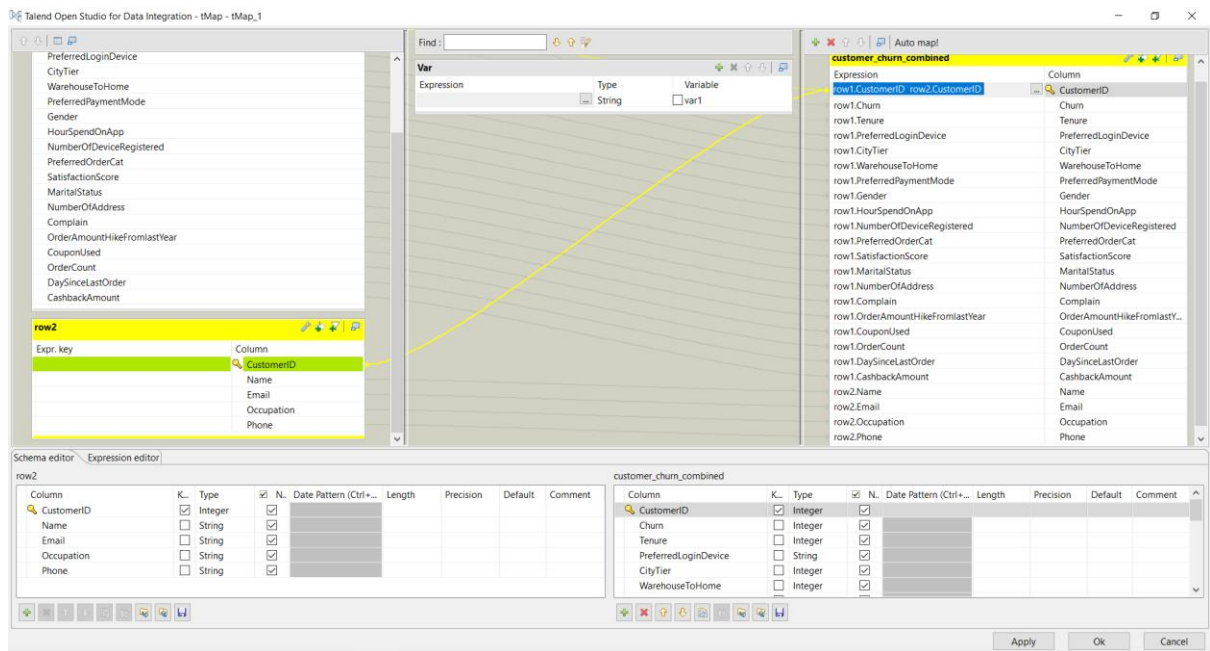
Define dataset schema.



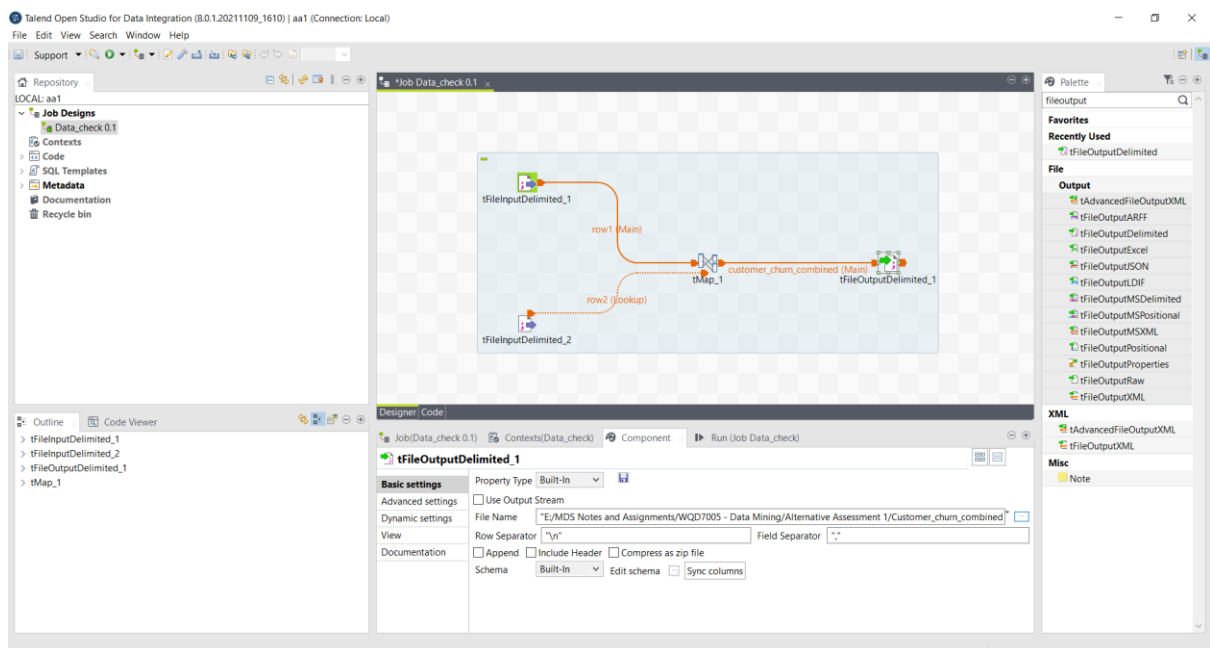
Run job.



Integrate two data sources.

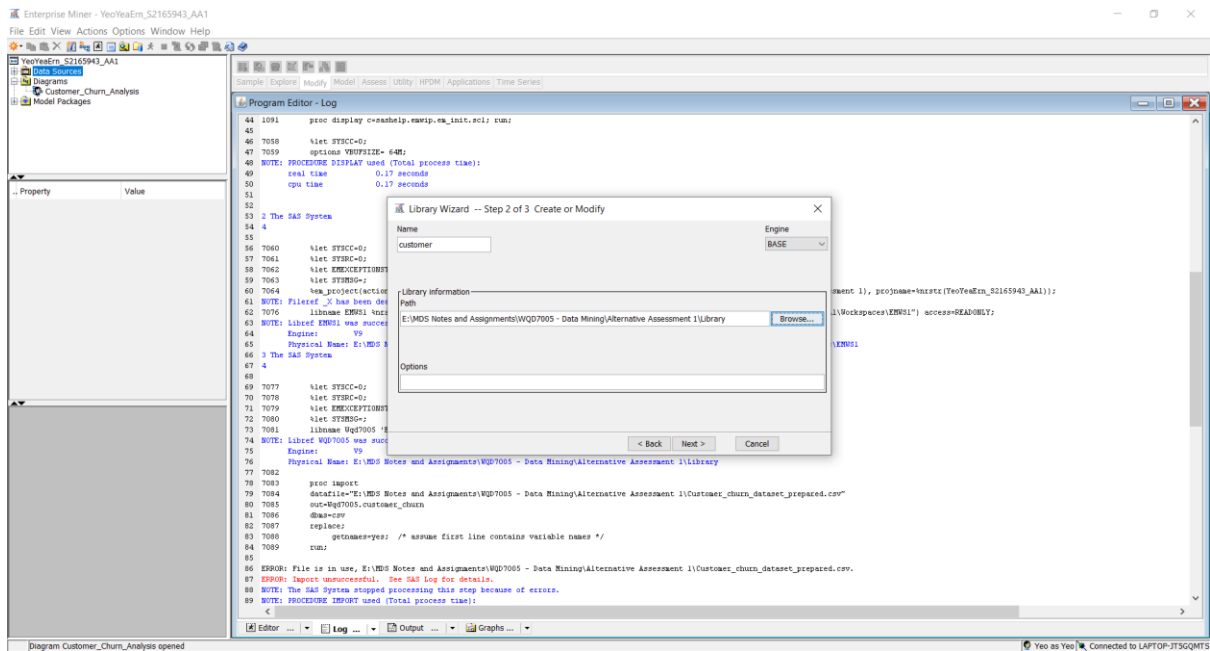
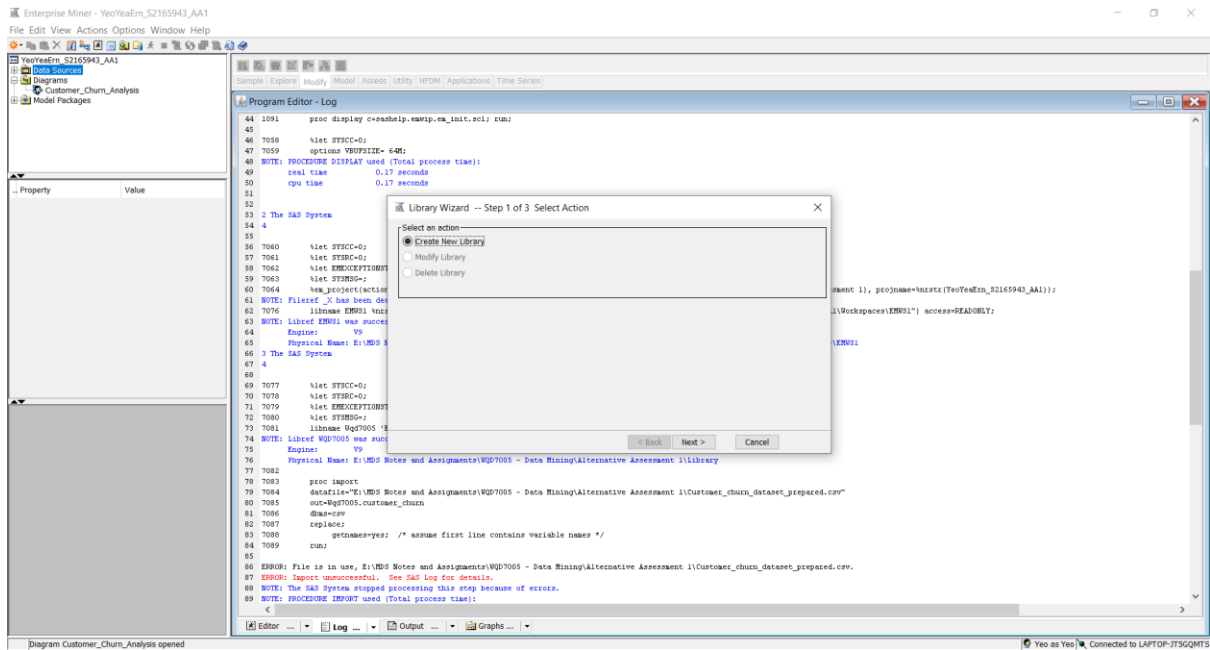


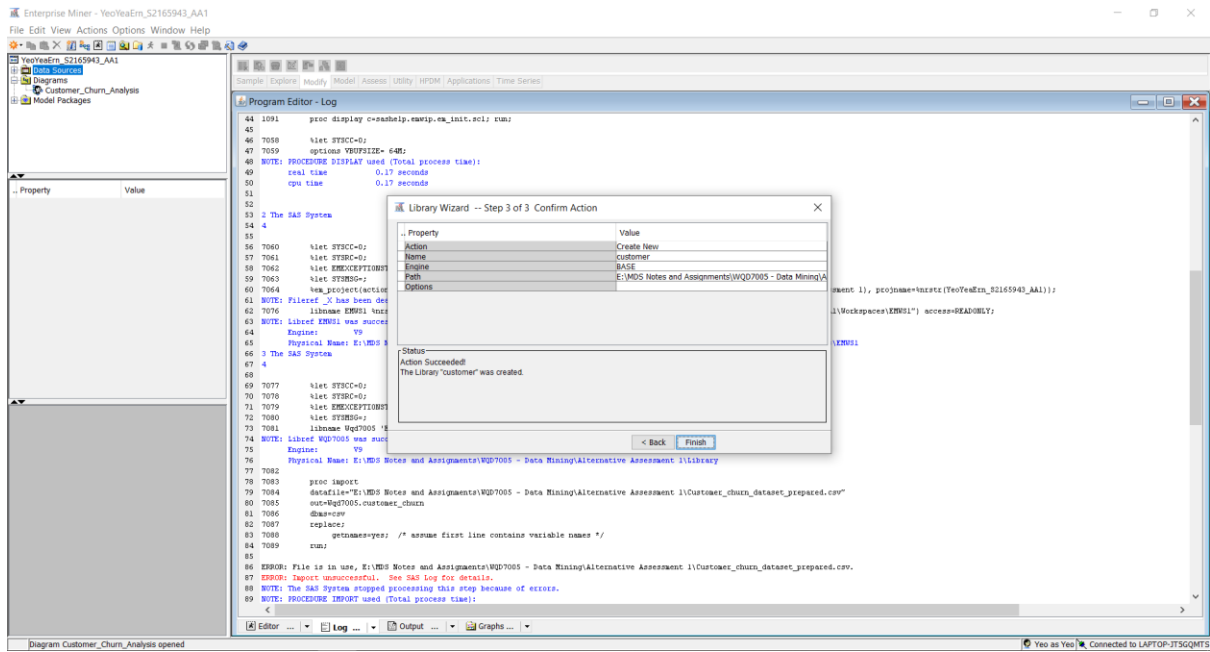
Output combined data into csv.



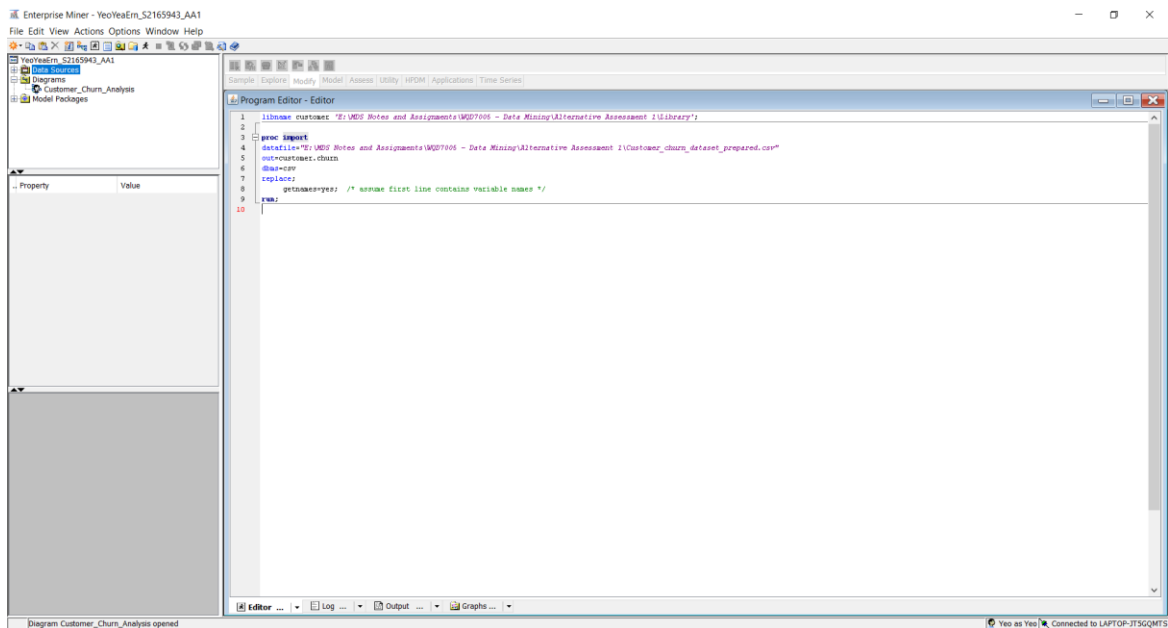
SAS Enterprise Miner

Create a new library.





Import Prepared Dataset to Library



Import data source.

