WQD7005 – AA1

Nur Natisya Binti Abdul Yazid (S2195163)

https://github.com/s2195163/AA1.git

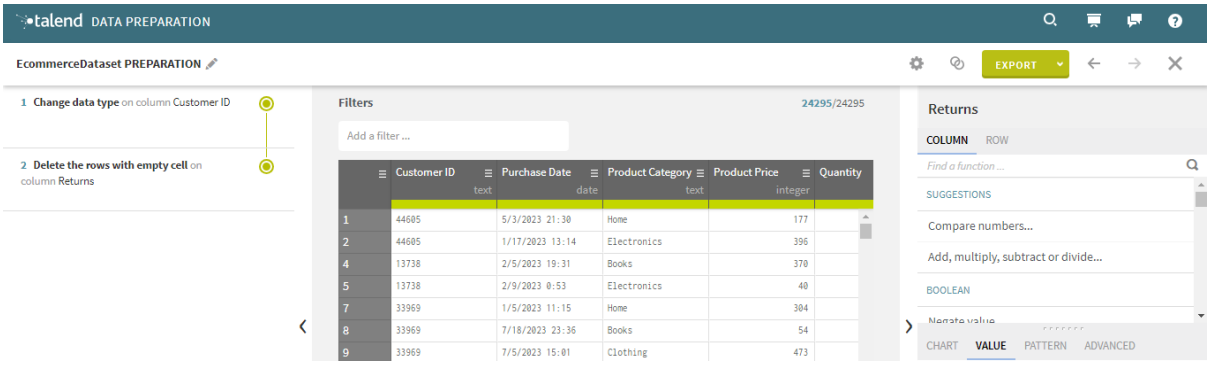
I provide a thorough analysis of consumer behaviour in this paper using data that was downloaded from the Kaggle website. Using ensemble methods and decision tree analysis, aim to extract insights. This report will cover every step of the procedure, providing a justification for the decisions made and addressing any difficulties encountered during the study. The objective of this case study is to indicate whether the customer I provide a thorough analysis of consumer behaviour in this paper using data that was downloaded from the Kaggle website. Using ensemble methods and decision tree analysis, aim to extract insights. This report will cover every step of the procedure, providing a justification for the decisions made and addressing any difficulties encountered during the study. The objective of this case study is to indicate whether the customer had stopped purchasing or not (1 Churn, 0 active) producing predictive modelling like decision trees and ensemble methods such as gradient boosting and random forest.

had stopped purchasing or not (1 Churn, 0 active) producing predictive modelling like decision trees and ensemble methods such as gradient boosting and random forest.

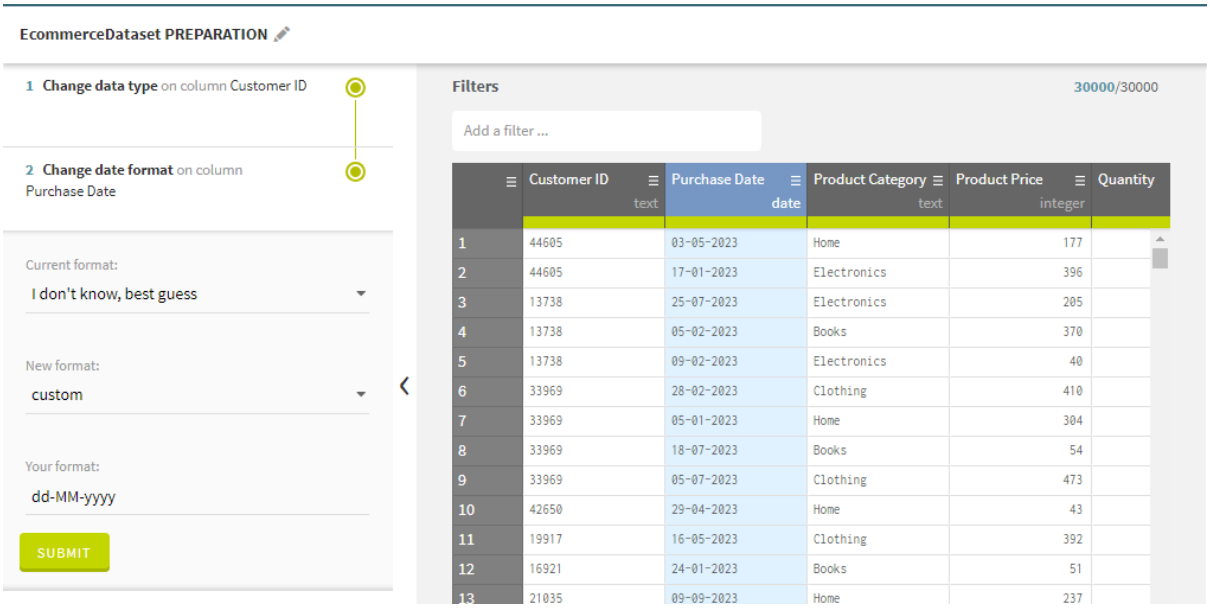
Talend Data Preparation

1. **Import Data:**

After importing data into Talend Data Preparation, it identified the customer ID as a postal code, and I changed it to an integer instead. Apparently, it only imported 30000 observations as that is the limit for Talend Data Preparation.



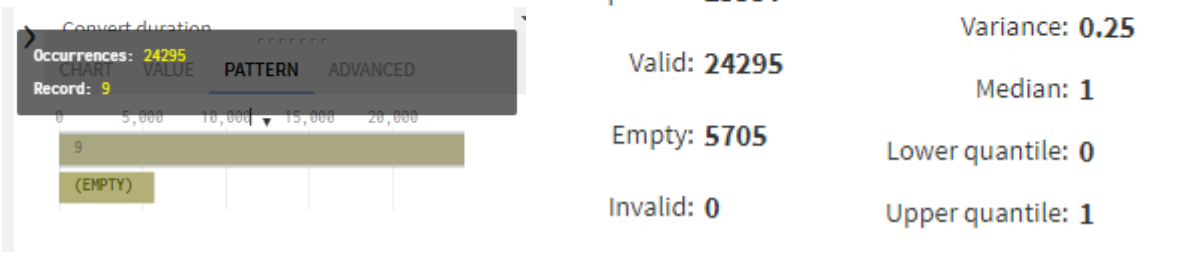
2. **Change Date Format**



Changing the date format for better understanding.

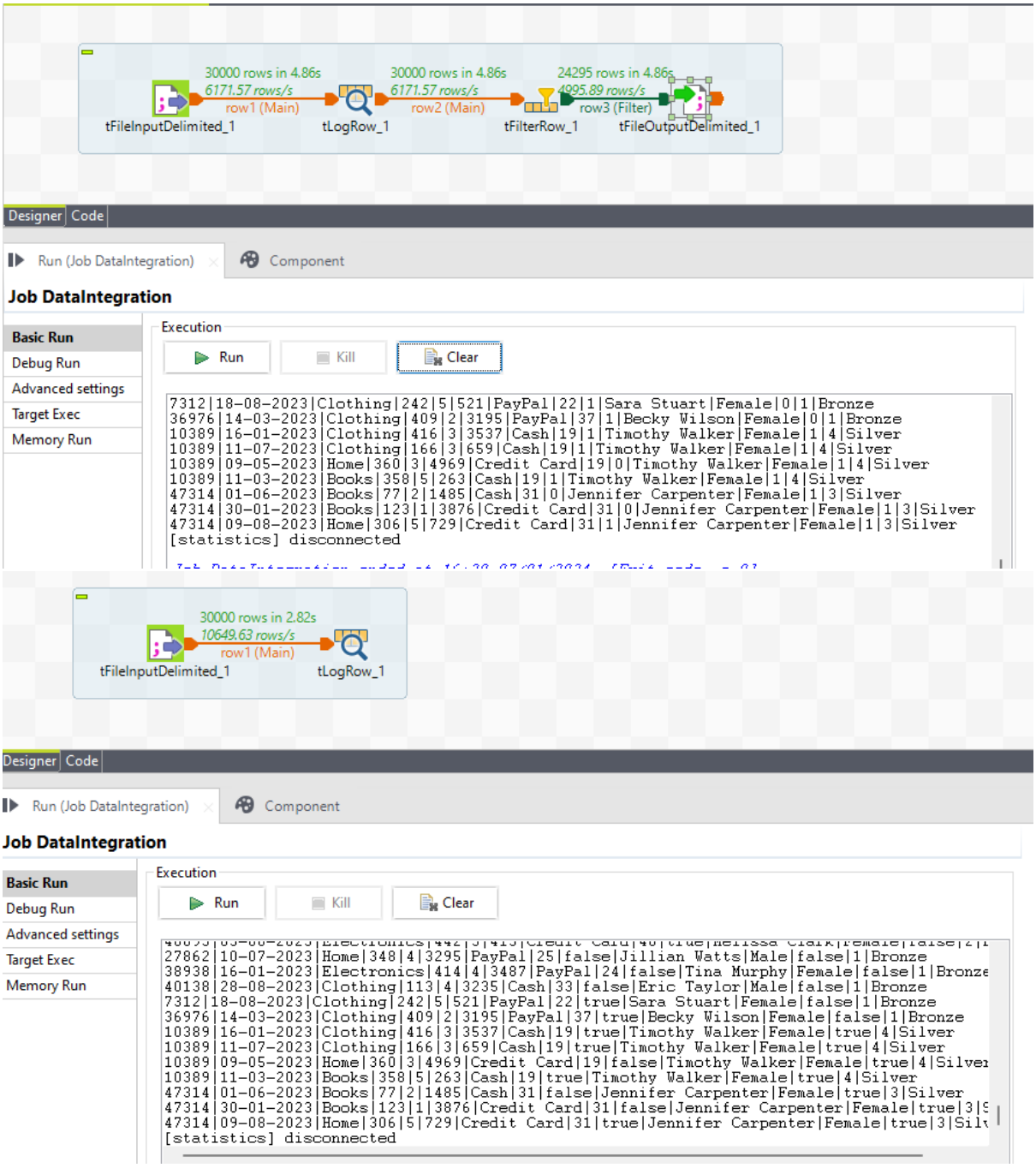
3. Missing Values:

It occurs that there are 5705 missing values for the “Return” variable. The “Return” variable refers to as if the customer has returned the purchased item or not. It is either 1 or 0. Therefore, having it deleted would be less appropriate because it could indicate whether the customer did return the item or not as the transaction of items has occurred.

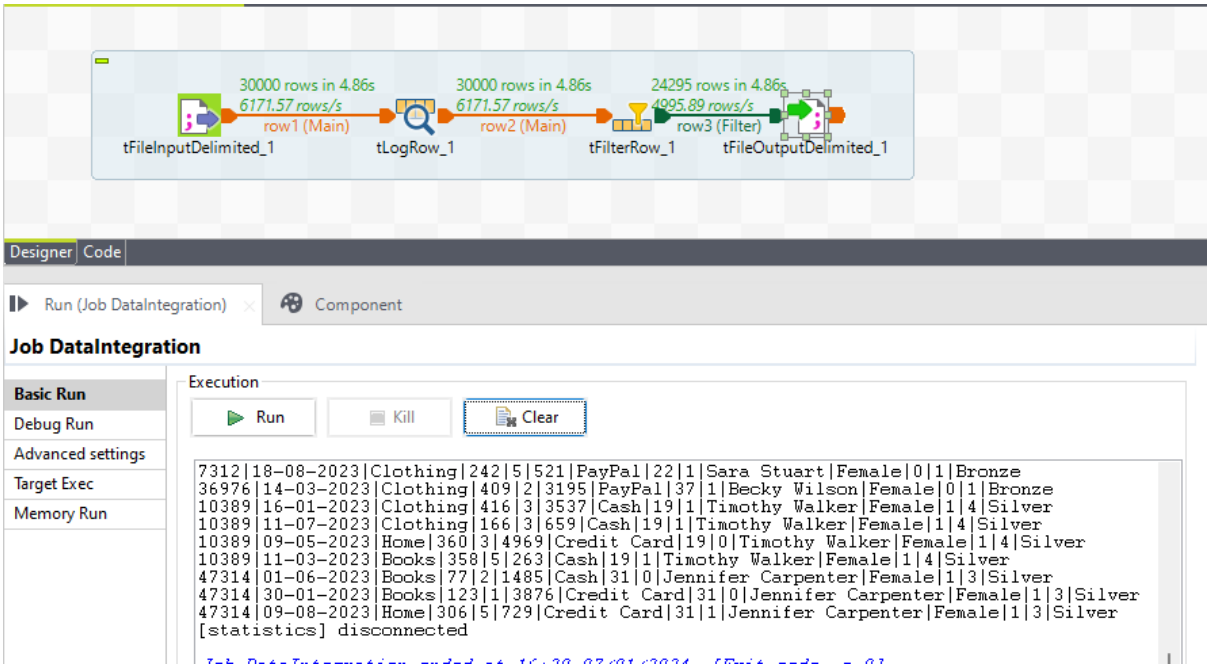


Talend Data Integration

1. **Read CSV file**
   Reading CSV file that has been process in the Talend Data Preparation to check on the missing values and removing them.



2. **Filter Out Missing Values**

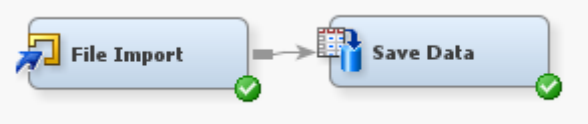


As removing missing values would not be appropriate as 0 and 1 for “Returns” variable hold a meaning whether customer have returned the item or not. In this situation, imputation would be implemented.

SAS Enterprise Miner

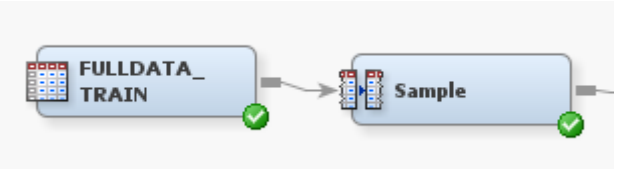
1. Importing Data

The dataset, which includes a variety of variables and purchase history over the past year, was imported and saved as SAS data using SAS Enterprise Miner.



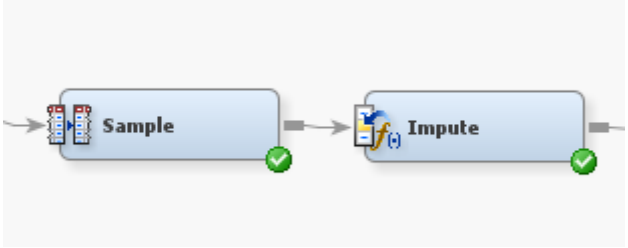
2. Sampling

Sampling is an alternative towards big data as it could represent the whole. The dataset consists of 47237 observations. The sampling size was determined using 10% of the population (4724).



3. Addressing Missing Values

Missing values in the dataset can be imputed using SAS Miner's Impute node. Able to select from a range of imputation techniques, such as custom values, mean, median, and mode. It is also possible to impute values depending on the variable's distribution using the impute node. In the data, there are 871 missing values under the “Return” variable. Since the “Return” variable is a binary (categorical), using the mode method would be appropriate.



Variable Name	Impute Method	Imputed Variable	Impute Value	Role	Measurement Level	Label	Number of Missing for TRAIN
Returns	COUNT	IMP_Returns	0	INPUT	BINARY	Returns	871

4. Specify Variable Roles

Target Variable: Churn

Name	Role	Level
Churn	**Target**	**Binary**
Customer_Age	**Input**	**Interval**
Customer_ID	**ID**	**Nominal**
Gender	**Input**	**Nominal**
MembershipLeve	**Input**	**Nominal**
Payment_Metho	**Input**	**Nominal**
Product_Catego	**Input**	**Nominal**
Product_Price	**Input**	**Interval**
Purchase_Date	**Input**	**Interval**
Quantity	**Input**	**Interval**
Returns	**Input**	**Binary**
Total_Purchase_	**Input**	**Interval**
WebsiteVisit	**Input**	**Interval**

Modeling:

1. **Decision Tree:**

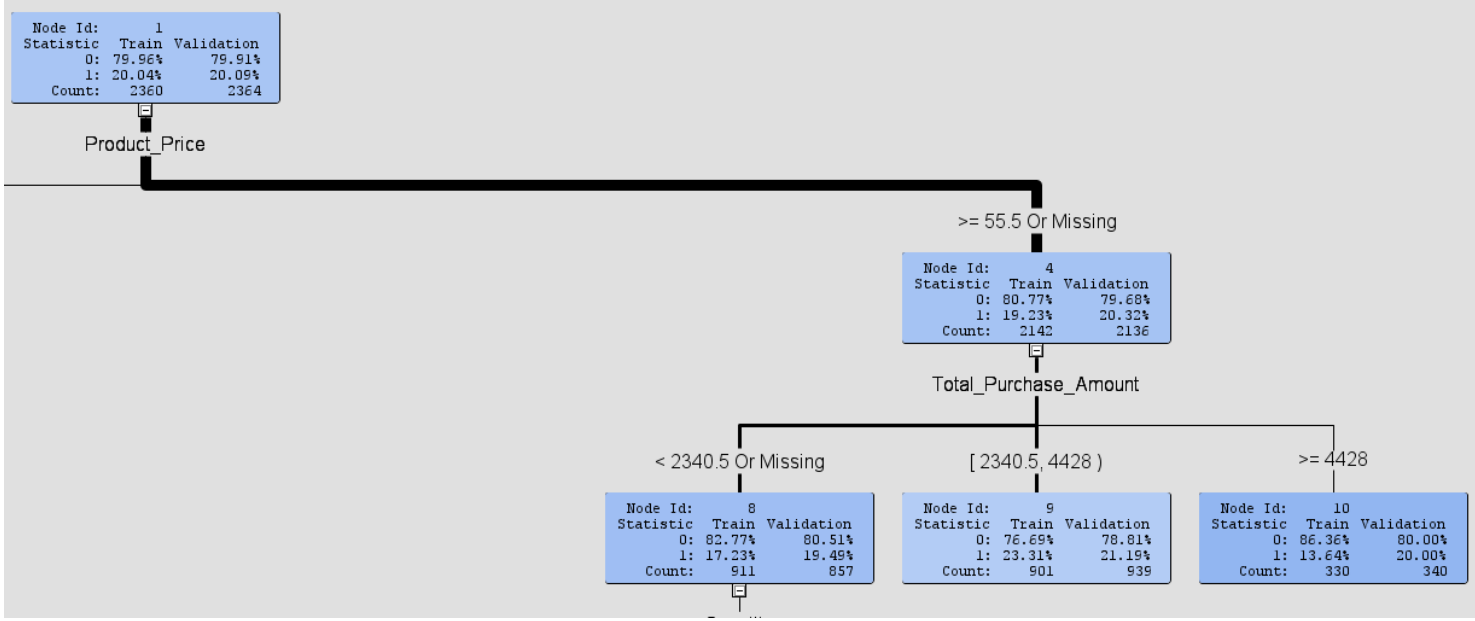
   Data Partition

   Building and assessing decision trees require the critical step of data split. This ensures that the model performs properly when applied to fresh, untested data. The "Partition" node in SAS Miner is commonly used to separate data before creating a decision tree. The dataset is split into training and testing sets by this node. Split data into train and valid (50:50).

Partition Summary

Type	Data Set	Number of Observations
DATA	EMWS1.Impt4_TRAIN	4724
TRAIN	EMWS1.Part4_TRAIN	2360
VALIDATE	EMWS1.Part4_VALIDATE	2364

Decision Tree Model



Created a decision tree model using SAS Enterprise Miner, which gave important insights into consumer behaviour. To optimise the interpretability of the resulting tree, great thought was given to the selection of variables and parameters. The diagram shows that a product price above 55.5$ has an 80.77% confidence that it predicts active customers and unseen data shows 79.68%. Following with a total purchase amount of less than 2340.5$ shows an 82.77% confidence that it predicts active customers and 80.51% for unseen data.

Target Variable: Churn

Fit Statistics

Target=Churn Target Label=Churn

Fit Statistics	Statistics Label	Train	Validation
NOBS	Sum of Frequencies	2360.00	2364.00
MISC	Misclassification Rate	0.20	0.20
MAX	Maximum Absolute Error	0.89	1.00
SSE	Sum of Squared Errors	731.13	789.38
ASE	Average Squared Error	0.15	0.17
RASE	Root Average Squared Error	0.39	0.41
DIV	Divisor for ASE	4720.00	4728.00
DFT	Total Degrees of Freedom	2360.00	.

Output

In terms of the highest difference between projected and actual churn probabilities, the model appears to be producing quite accurate predictions, as evidenced by the relatively low maximum absolute error (0.89 for training and 1.00 for validation). In general, lower values of SSE, ASE, and RASE are better since they show that the model's predictions are more in line with the real results. One especially helpful tool for evaluating the overall forecast accuracy is the average squared error or ASE. There may be a problem with overfitting since the model performs better on the training set than on the validation set.

2. Ensemble Method

Multiple model predictions are combined during ensembling, which can lessen overfitting and increase generalisation.

Boosting using Gradient Boosting Model

Variable Importance

Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
Total_Purchase_A...		24	1	0	0
Product_Price		23	0.979748	0	0
Purchase_Date		26	0.960233	0	0
Customer_Age		9	0.602852	0.92774	1.538917
Quantity	Quantity	8	0.530214	0.548663	1.034795
Product_Category		5	0.477966	0	0
IMP_Returns	Imputed: Returns	3	0.397545	0	0
Payment_Method		4	0.371292	0	0
Gender	Gender	2	0.292516	0	0
WebsiteVisit	WebsiteVisit	2	0.248388	1	4.025954
MembershipLevel	MembershipLevel	1	0.20722	0	0

“Total Purchased Amount” is the most influential variable in making predictions.

Fit Statistics

Target=Churn Target Label=Churn

Fit Statistics	Statistics Label	Train	Validation
NOBS	Sum of Frequencies	2360.00	2364.00
SUMW	Sum of Case Weights Times Freq	4720.00	4728.00
MISC	Misclassification Rate	0.20	0.20
MAX	Maximum Absolute Error	0.87	0.89
SSE	Sum of Squared Errors	714.68	769.97
ASE	Average Squared Error	0.15	0.16
RASE	Root Average Squared Error	0.39	0.40
DIV	Divisor for ASE	4720.00	4728.00
DFT	Total Degrees of Freedom	2360.00	.

Output

The 20% misclassification rate indicates that there is potential for improvement in the distinction between situations that are churn and those that are not. The biggest potential inaccuracy in predictions is indicated by the maximum absolute error values of 0.89 (validation) and 0.87 (training). These values are small, indicating that the model's predictions are generally within a reasonable range of the observed values. The training and validation average squared error values of 0.15 and 0.16, respectively, show that, on average, there are not many squared differences between the predicted and actual values. This shows that the model's ability to anticipate churn is generally correct. On the training set, the gradient boosting model shows good predictive performance with low errors and misclassification rates. Higher errors and misclassification rates on the validation set show that the model has trouble generalising to new data. The training and validation sets' performance differs noticeably, which could be an indication of overfitting.

Random Forest

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
Total_Purc...	7567	0.067657	0.135314	-0.07004	-0.00407	-0.06906	-0.00158	
Product_Pri...	5900	0.055237	0.110475	-0.05319	0.00105	-0.05789	-0.00330	
Purchase_...	4238	0.043591	0.087182	-0.04449	-0.00130	-0.04407	-0.00098	
Customer_...	4174	0.035359	0.070717	-0.03803	-0.00169	-0.03834	-0.00262	
WebsiteVisit	3564	0.016770	0.033540	-0.02111	-0.00271	-0.01982	-0.00071	WebsiteVisit
Quantity	2969	0.019627	0.039254	-0.02037	0.00055	-0.02323	-0.00265	Quantity
Product_Ca...	1105	0.005449	0.010897	-0.00592	0.00044	-0.00556	0.00093	
Gender	1025	0.004413	0.008826	-0.00364	0.00084	-0.00456	-0.00074	Gender
Payment_M...	944	0.004693	0.009385	-0.00356	0.00078	-0.00433	0.00032	
IMP_Returns	748	0.003608	0.007217	-0.00250	0.00107	-0.00336	-0.00018	Imputed: R...
Membershi...	480	0.001967	0.003935	-0.00229	-0.00006	-0.00189	0.00014	Membershi...

“Total Purchased Amount” is the most influential variable in making predictions.

Output

Fit Statistics

Target=Churn Target Label=Churn

Fit Statistics	Statistics Label	Train	Validation
ASE	Average Squared Error	0.04	0.17
DIV	Divisor for ASE	4720.00	4728.00
MAX	Maximum Absolute Error	0.61	0.98
NOBS	Sum of Frequencies	2360.00	2364.00
RASE	Root Average Squared Error	0.21	0.41
SSE	Sum of Squared Errors	211.08	794.42
DISF	Frequency of Classified Cases	2360.00	2364.00
MISC	Misclassification Rate	0.01	0.20
WRONG	Number of Wrong Classifications	31.00	483.00

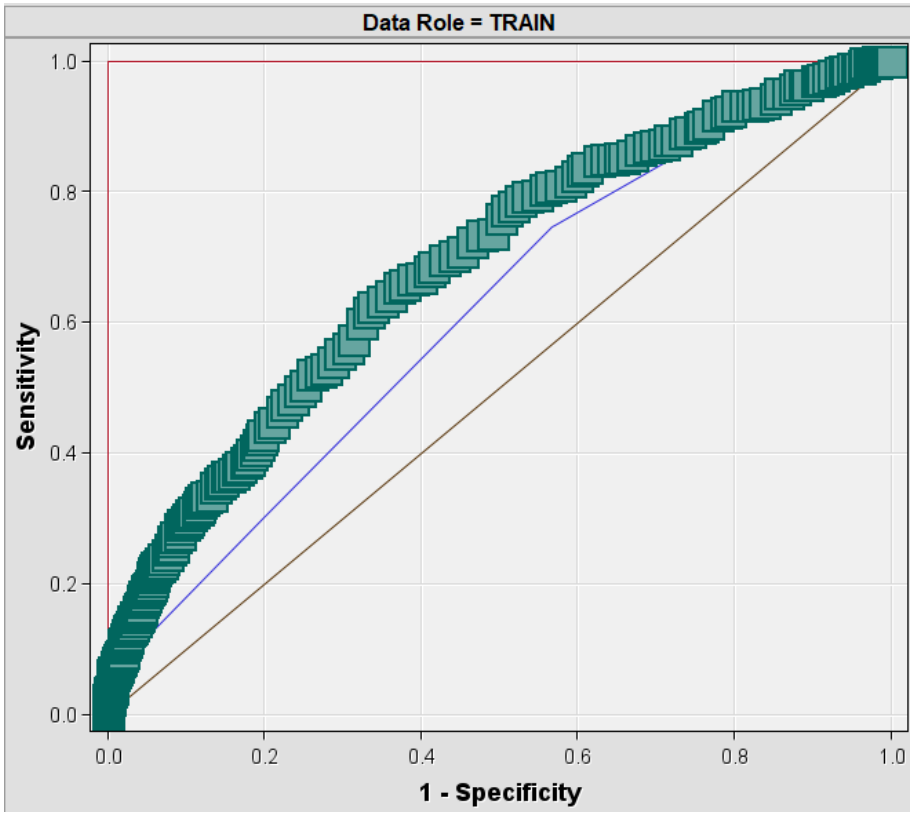
With a low maximum absolute error, low average squared error, and low misclassification rate, the model performs admirably on the training set. The model performs worse on the validation set, as evidenced by increased maximum absolute error, average squared error, and misclassification rate. This may indicate that the training data were overfitted. The training and validation sets perform noticeably differently, suggesting that more optimisation of the model may be necessary to enhance its generalisation.

Model Comparison

Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Train: Sum of Frequencies	Train: Misclassification Rate
Boost2	Boost2	Gradient Bo...	Churn	Churn	0.200508	2360	0.198729
Tree4	Tree4	Decision Tr...	Churn	Churn	0.200931	2360	0.200424
HPDMFore...	HPDMFore...	HP Forest	Churn	Churn	0.204315	2360	0.013136

It shows that Gradient Boosting produced the best in predicting churn customer. On the training set, the model achieves low maximum absolute error, low average squared error, and a very low misclassification rate. This suggests that the model has successfully picked up on the relationships and patterns seen in the training set. The capacity of gradient boosting to identify intricate connections and non-linear patterns in data is well established. The boosting technique, which combines several weak learners to produce a powerful predictive model, probably helps the model.

ROC Curve



Based on the ROC curve, it has been proven that Gradient Bosting is the best.

Challenges:

It is difficult for the model to extrapolate its prediction capabilities to the validation set. The validation set exhibits a significant rise in the maximum absolute error, average squared error, and misclassification rate when compared to the training set. Overfitting may be present based on the variations in performance metrics observed between the training and validation sets. The model might not generalise well to fresh, untested data since it is overly adapted to the unique features of the training set. larger errors, such as a larger root average squared error and misclassification rate, are seen in the validation set. This suggests that the accuracy of the model's predictions on fresh data is lower than that of its training set.

To deal with overfitting, think about using regularisation strategies and adjusting hyperparameters. The generalisation performance of the model can be affected by varying parameters such as regularisation strength, tree depth, and learning rate. Investigate further feature engineering to improve the model's capacity to identify pertinent patterns in the data. In order to more accurately depict underlying relationships, this may entail adding new features or altering current ones. Throughout the training process, keep an eye on the model's performance on the validation set. This can assist in determining the starting point of overfitting and direct modifications to stop it. To further enhance generalisation, think about utilising ensemble methods or other model-ensembling strategies. To get a final model that is more reliable and broadly applicable, this may entail merging the predictions of various models.

In summary, even though the gradient boosting model performs well on the training set, care must be taken to make sure it can adapt to new data. To effectively refine the model for churn prediction across a variety of datasets, regularisation, feature engineering, and close observation of validation set performance are essential components.