Master of Data Science (2023/2024 – Semester 1)

Faculty of Computer Science & Information Technology

WDQ 7005 Data Mining

ALTERNATIVE ASSESSMENT 1

Matric No:	Name:
22069349	Jiajia Jiang

Table of Contents

1. Introduction

This case study revolves around the analysis of customer behavior within the context of an e-commerce platform. The dataset at hand comprises customer transaction records collected over the past year, encompassing various customer attributes and purchase history. This analysis aims to gain valuable insights into customer behavior, with the ultimate goal of formulating effective business strategies to enhance customer satisfaction, reduce churn, and boost sales.

In today's highly competitive e-commerce landscape, understanding customer behavior is crucial for the success of any online business. Customer behavior analysis allows us to comprehend the preferences, trends, and tendencies of our customer base. Armed with this knowledge, businesses can make informed decisions regarding marketing strategies, product recommendations, and personalized experiences.

2. Dataset structure

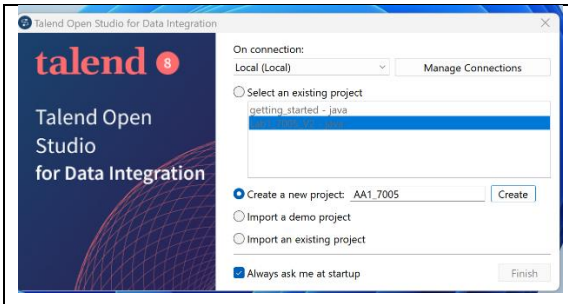
No	Attributes	Description
1	CustomerID:	A unique identifier for each customer.
2	Age:	The age of the customer.
3	Gender:	The gender of the customer.
4	Location:	The geographic location of the customer.
5	MembershipLevel:	Indicates the membership level of the customer (e.g., Bronze, Silver, Gold, Platinum).
6	TotalPurchases:	The total number of purchases made by the customer.
7	TotalSpent:	The total amount spent by the customer.
8	FavoriteCategory:	The category in which the customer most frequently shops (e.g., Electronics, Clothing, Home Goods).
9	LastPurchaseDate	: The date of the customer's last purchase.
10	Occupation:	The occupation of the customer.
11	Frequency of Website Visits:	Indicates how often the customer visits the e-commerce website.
12	Churn:	A binary indicator (1 for churned, 0 for active) that signals whether the customer has stopped purchasing.

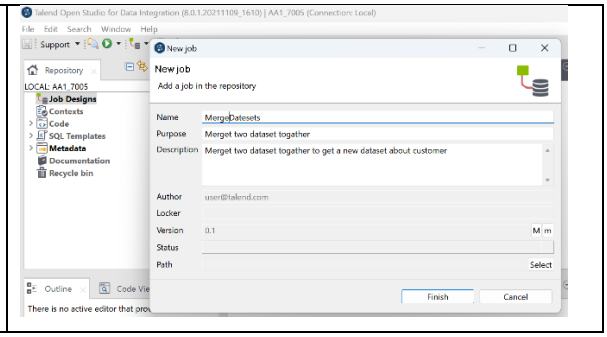
3. Data Import and Preprocessing

3.1 Talend Open Studio for Integration

The Origin dataset are two separate datasets with same attributes, so we need to merge the two datasets into one dataset. Talend Open Studio for Integration was chosen for data merging.

Firstly, Create Project and new job in the Talend.

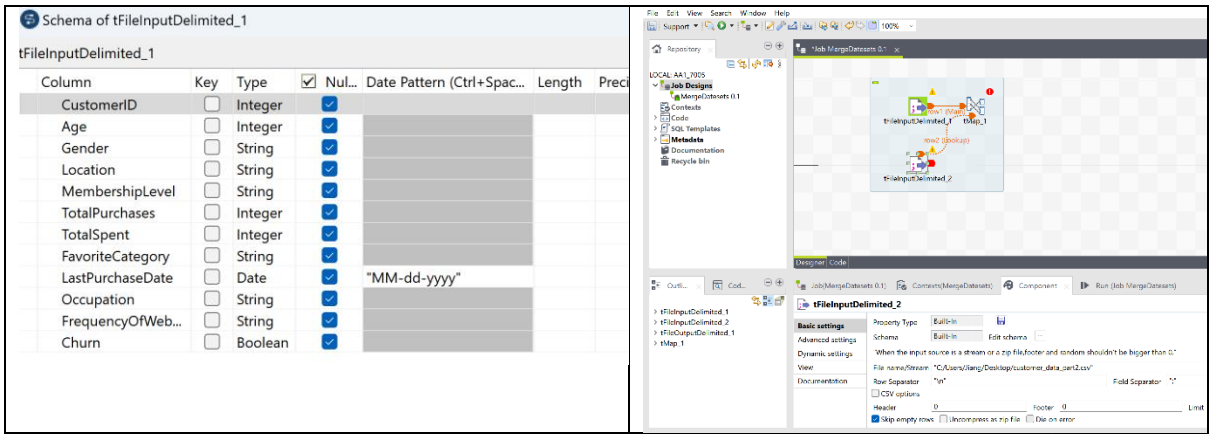




After that add a file input component. Drag and drop two "tFileInputDelimited" components from the component panel to the job design space. These components will be used to load the CSV files.

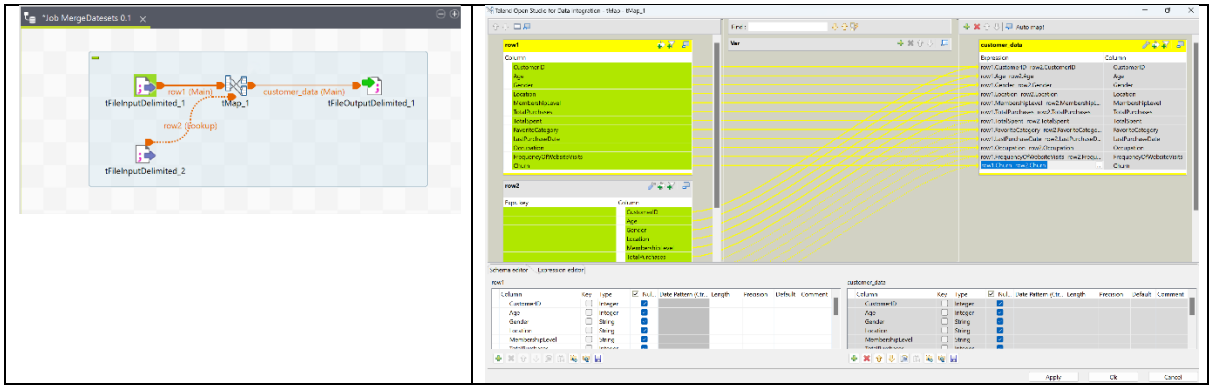
Then configure the file input components: Double-click each "tFileInputDelimited" component, specify the path to the CSV file, set the field delimiter, and configure any other necessary settings.

In the next step add a merge component: Drag and drop a "tMap" component from the component panel to the job design space and connect both file input components to the "tMap" component.



After that add a file output component: Drag and drop a "tFileOutputDelimited" component to the design space and connect the output from the "tMap" component to it.

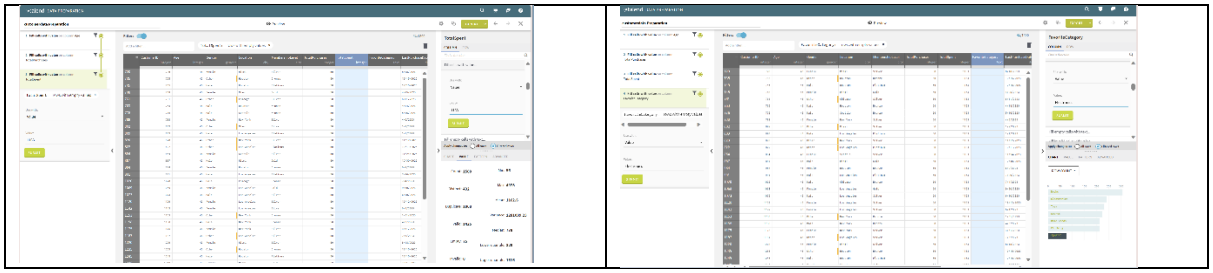
Configure the file output component: Double-click the "tFileOutputDelimited" component, specify the path, filename, and other settings such as the field delimiter for the output file.



In the end, click "Run" (F6) to execute the job. This will merge the two files and output them to the specified location.

3.2 Talend Data Preparation

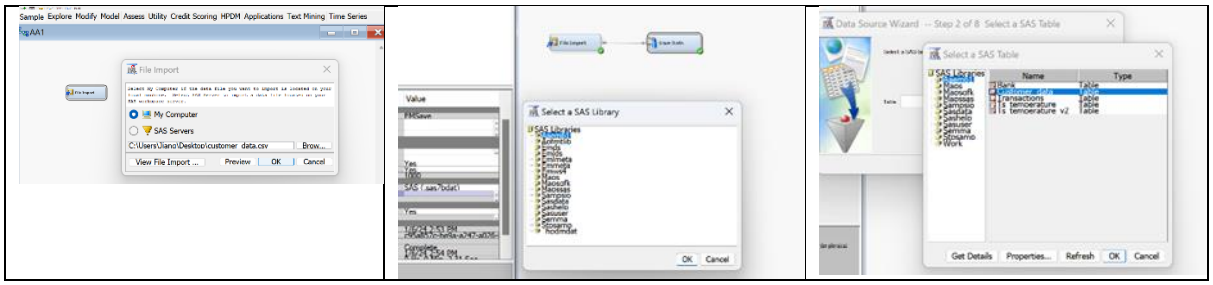
After merging the two dataset, the data preparation step is needed to do. In the Talend data preparation website page, we check the miss value first. It find there have two kind of missing class and interval. They are FavoriteCategory (75 missing values), Age(75 missing values), TotalPurchases (75 missing values), TotalSpent(75 missing values), then we impute the value, for FavoriteCategory the class the missing value it use the most frequentable one, in this case we choose books. For the other interval missing in this case we choose mean value to impute. After impute all the missing value export the file.



3.3 SAS Enterprise Miner

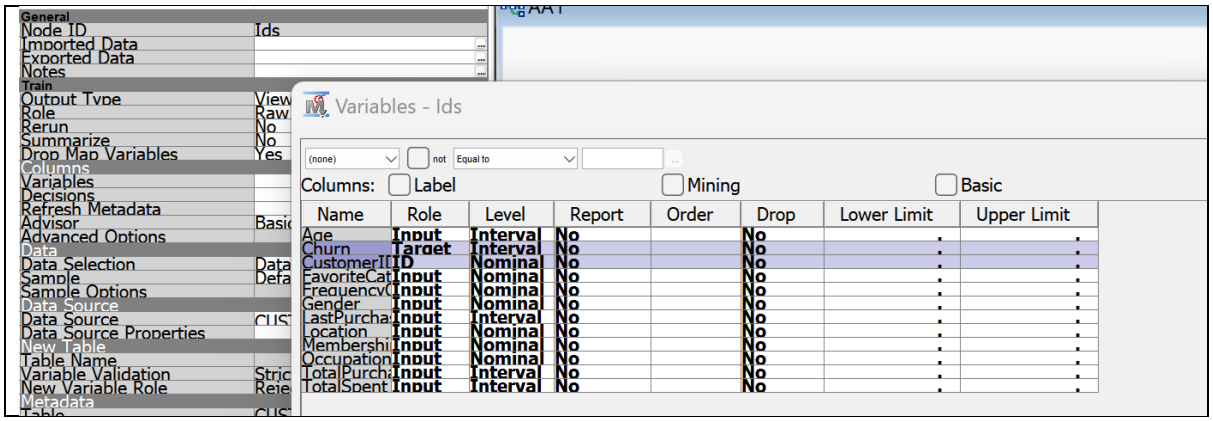
3.3.1Import Processing:

Firstly in the SAS Enterprise Miner we create a diagram then drag the file import node upload the customer_data.csv after that drag the save file node to save the file into AAEM61 library. Then create new data source from the AAEM61 library.



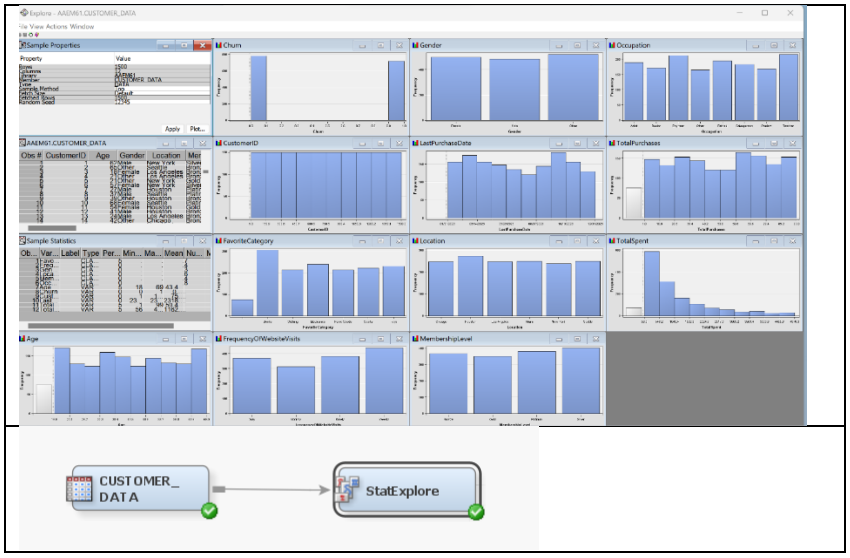
3.3.2 Variable Role Specification:

After creating the new data source, drag the data to the diagram and edit the variables set Churn as Target and Customer as ID.

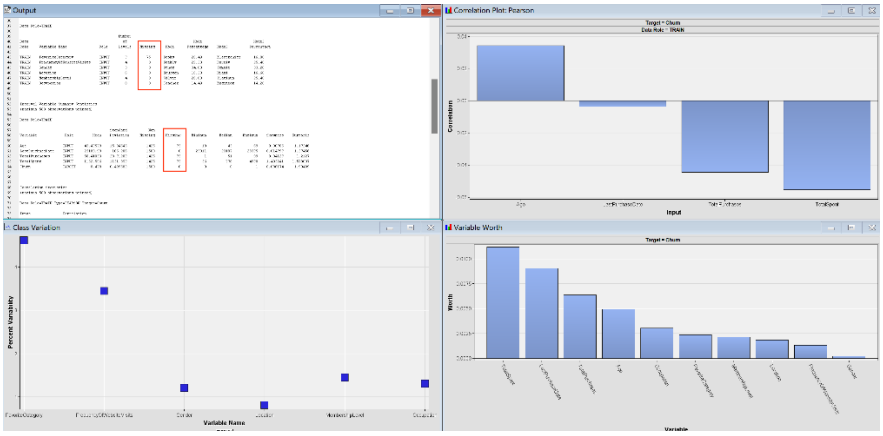


3.3.3 Dataset Explore and Check missing values

Firstly, explore all the variables, check each attributes data distribution and whether there is a missing value.

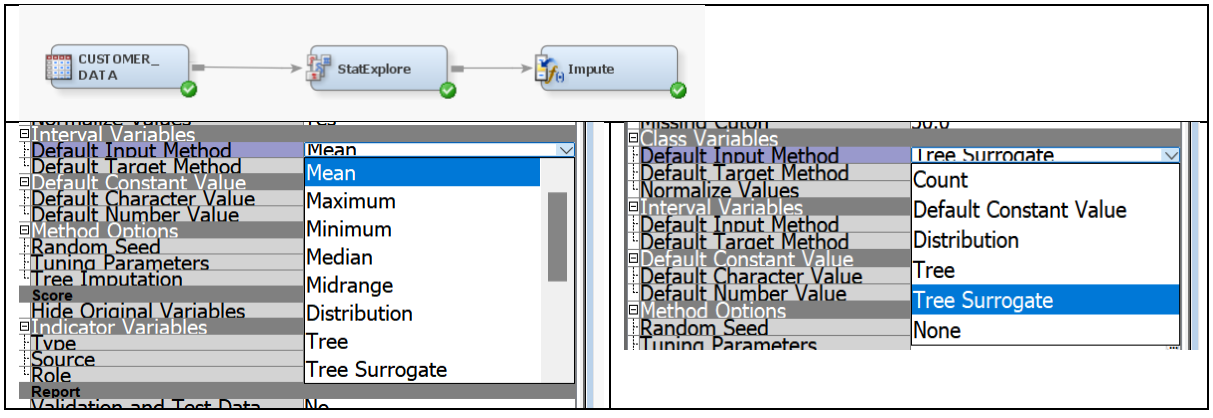


Then connect to StatExplore node check the output report then we find there have two kinds of missing values, one is class variable another is interval variable. They are FavoriteCategory (75 missing values), Age(75 missing values), TotalPurchases (75 missing values), TotalSpent(75 missing values).

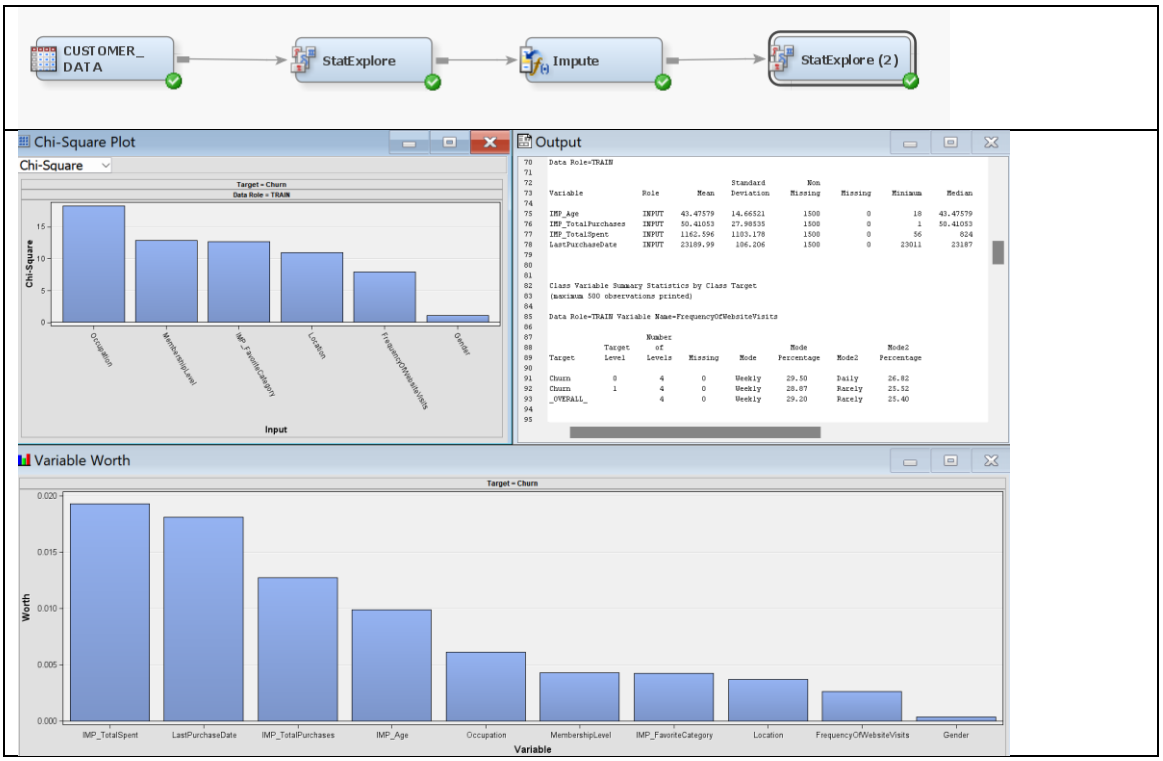


3.3.4 Handle missing value

For the missing value we drag the impute node to the diagram, for Interval missing value chose mean and Tree Surrogate for Class Variables.



After imputing the missing value, check the dataset again. We can see there has no missing value.



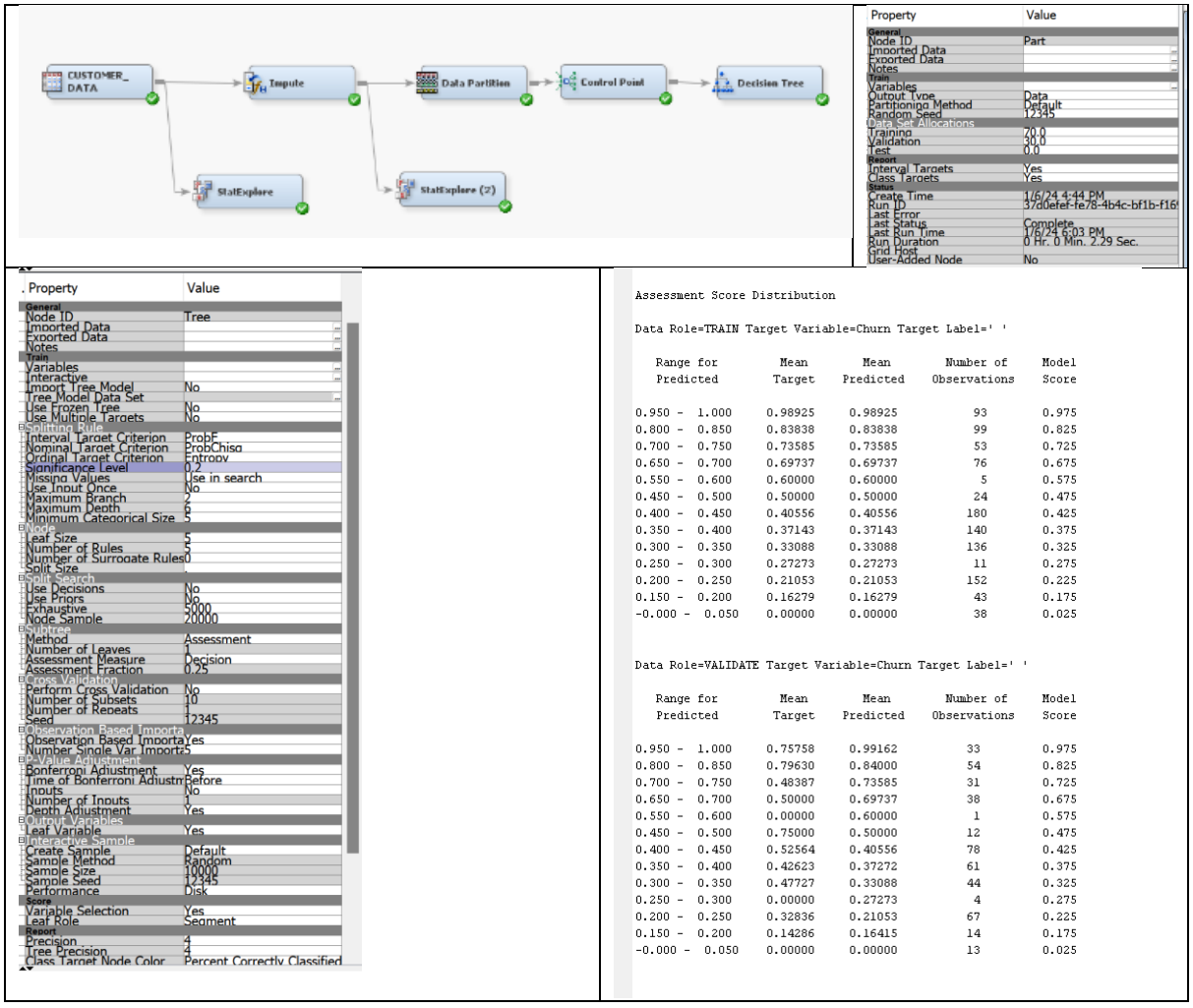
4. Decision Tree Analysis

4.1 Create a decision tree model in SAS Enterprise Miner

Decision tree is a versatile data mining algorithm that enables both classification and regression tasks. It models decisions and their possible consequences as a tree-like structure, where each internal node represents a "test" on an attribute, each branch represents the outcome of the test, and each leaf node represents a decision or prediction. This structure emulates the human decision-making process, making decision trees one of the most intuitive and widely used algorithms in analytics.

Before creating the decision tree model, drag the data partition node and control point. The data partition we set 70% for train and 30% for validation.

For the decision tree model, we set the max branch as 2, maximum tree depth is 6 and maximum categorical size is 5. For the node properties the leaf size is 5, number of rules is 5 and number of surrogate rules is 0.



The results from the decision tree model show its performance in classifying customers' churn probabilities on both training and validation data.

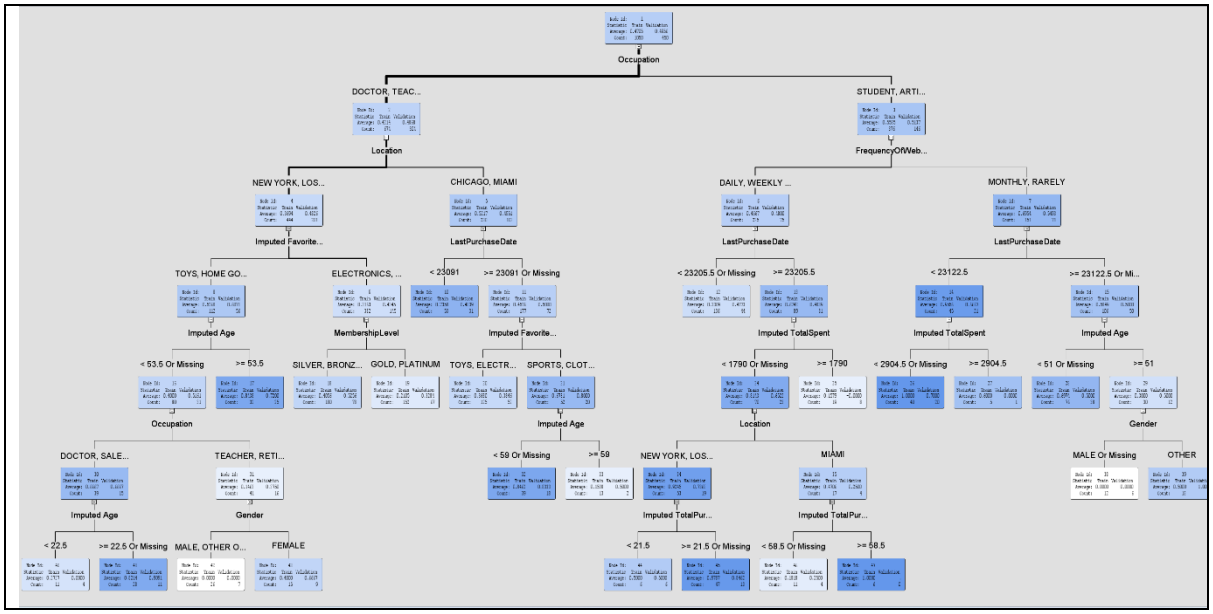
In the training set, the model has strong alignment between the predicted probabilities and actual churn rates, indicating good calibration. For example, in the highest probability bin (0.950 - 1.000), the model perfectly matches the Mean Predicted with the Mean Target at 0.98925, suggesting high accuracy for the most certain predictions of churn.

However, in the validation set, there are discrepancies. In the highest predicted churn range (0.950 - 1.000), the Mean Target drops to 0.75758, while the Mean Predicted stays high at 0.99162, indicating an overestimation of churn risk. This suggests the model may not generalize as well to unseen data, a common challenge known as overfitting.

The model's performance, indicated by the model score, remains high in extreme ranges (close to 0 or 1) but is lower in the middle ranges, reflecting less certainty in predictions where the churn probability is around 50%.

Overall, the decision tree demonstrates strong training performance but may require adjustments to improve its predictive accuracy on new, unseen data.

4.2 Analyse customer behaviour



To mitigate customer churn, it is essential to understand the underlying factors that contribute to a customer's likelihood of disengaging. Leveraging a decision tree analysis, we can unearth patterns and predictors within customer data that signal churn risk. This predictive model slices through layers of demographic, behavioral, and transactional data to reveal key attributes ranging from occupation and geographic location to spending habits and product preferences that are instrumental in forecasting churn. With this knowledge, we can devise targeted interventions designed to bolster retention and foster enduring customer relationships.

4.2.1 Occupational Impact on Churn:

The decision tree places occupation as a significant indicator of churn risk. The differentiation between professions such as Doctors, Teachers, and Salespeople against Students, Artists, and other occupations suggests a correlation between occupation type and customer retention. For example, busy professionals may have less time for extensive shopping and might prefer a streamlined, reliable service, potentially leading to lower churn if their expectations are met. Conversely, Students and Artists might be more price-sensitive and could churn if they find better deals elsewhere. Tailoring loyalty programs and customer service to the specific needs and behaviors of each occupational group could help mitigate churn risks.

4.2.2 Geographical and Engagement Insights:

Geographical location plays a pivotal role in customer churn, with separate branches for metropolitan areas versus other cities, indicating different churn dynamics. Customers in larger cities might have access to more competitive alternatives, which could influence their loyalty. Engagement metrics, such as the recency of the last purchase and the frequency of website visits, are directly tied to churn, where infrequent visits and a longer time since the last purchase signal a higher likelihood of churn. Identifying at-risk customers through these

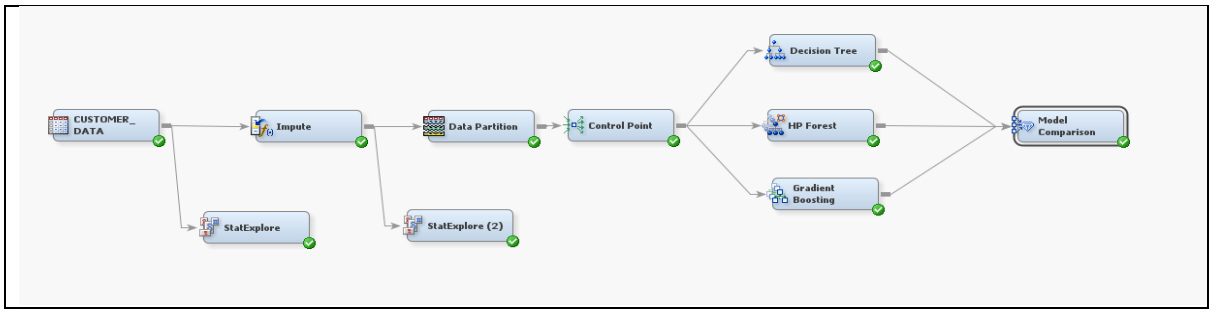
metrics allows for timely intervention strategies, such as personalized promotions or reminders, to re-engage them.

4.2.3 Demographic and Behavioral Predictors of Churn:

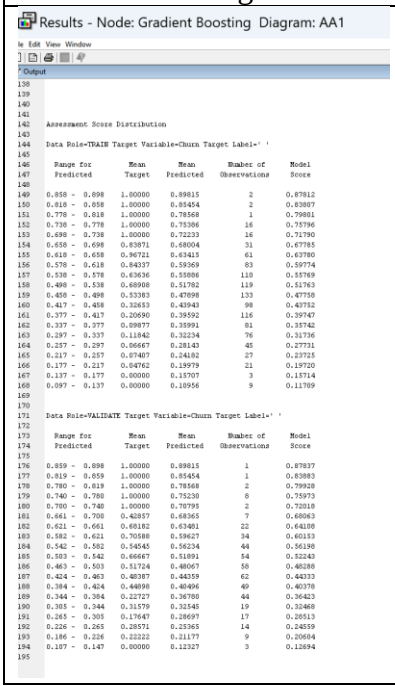
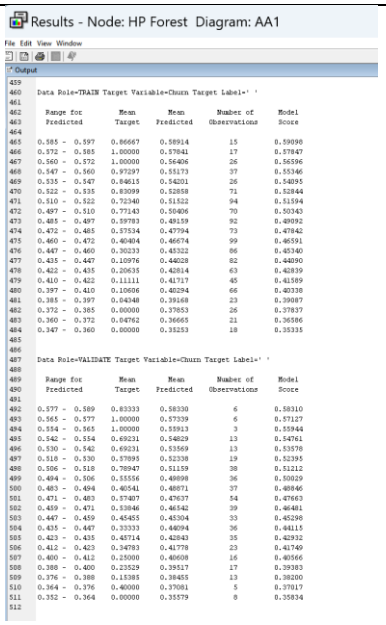
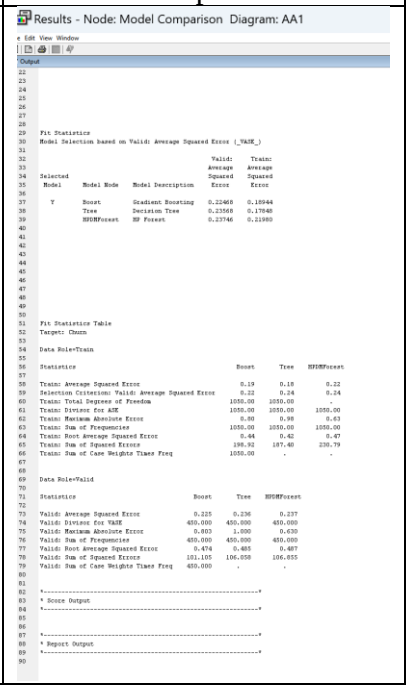
The decision tree also highlights demographic factors like age and gender, which can be instrumental in predicting churn. Different age groups may have varying levels of engagement and brand loyalty, influencing their churn behavior. Additionally, spending behavior and product preferences are strong indicators of churn; customers who spend above certain thresholds or those who purchase specific product categories may exhibit different churn rates. Understanding these spending patterns can be crucial for developing targeted retention campaigns, such as offering special deals on favorite categories or appreciation rewards that encourage continued patronage.

In summary, the decision tree analysis sheds light on various factors that contribute to customer churn. By addressing these areas with focused customer retention strategies, companies can proactively reduce churn rates. This might include personalized engagement based on occupational needs, regional marketing strategies, and targeted offers that align with customer demographics and spending behaviors. Ultimately, leveraging this decision tree analysis can lead to more effective churn prevention and an overall improvement in customer loyalty.

5. Ensemble Methods



In addition to decision trees, the use of ensemble methods like Boosting and Bagging with the Random Forest algorithm can significantly enhance the predictive power of your customer churn models by reducing variance, bias, or improving predictions through aggregation. In this section, we will analyze the results achieved by Boosting and Bagging with the Random Forest algorithm, assessing their respective strengths and weaknesses. Subsequently, we will perform a comparative analysis among these three models: the decision tree, Boosting, and Bagging with the Random Forest algorithm, to gain insights into their performance and suitability for addressing the customer churn prediction problem.

Gradient Boosting Result	HP Forest Result	3 Model Comparison Result
		

5.1 Boosting: (Gradient Boost)

In the top probability range (0.858 - 0.898), the Gradient Boosting model predicts with high confidence, as indicated by the "Mean Predicted" value of 0.8915, which is very close to the "Mean Target" of 1.0000. This suggests that, for the 2 observations in this range, the model is almost certain they will churn.
For the mid-probability ranges, such as (0.618 - 0.658), the "Mean Predicted" value is 0.63415, with a "Mean Target" of 0.96721 for 83 observations. Here, the model is underestimating the churn risk compared to the actual outcome.
Lower probability ranges, such as (0.177 - 0.217), show a "Mean Predicted" of 0.19797 and a "Mean Target" of 0.04762 for 21 observations, indicating the model's conservative prediction for these cases.

5.2 Bagging (Random Forest model in HP Environment)

The Random Forest model, in a similar top probability range (0.560 - 0.597), has a "Mean Predicted" of 0.58914 and a "Mean Target" of 0.86667 for 15 observations, which shows that the model is moderately confident about the churn prediction.
In a lower range (0.435 - 0.472), the "Mean Predicted" value is 0.45322 with a "Mean Target" of 0.30976 for 86 observations, here the model is slightly overestimating the churn likelihood.
For the lower probability range (0.347 - 0.360), the "Mean Predicted" is 0.35253, with a "Mean Target" of 0.00000 for 18 observations, suggesting that in this range, the model predicts no churn, which aligns with the actual outcome.

5.3 Model Comparison Analysis

In the comparative analysis of the three models based on their Average Squared Error (ASE), we can deduce their performance as follows:

The Gradient Boosting model showcases commendable predictive performance with a training ASE of 0.18944, implying a high degree of accuracy in fitting the training data. When applied to the validation data, the model maintains its robustness, yielding an ASE of 0.22468, the lowest among the three models. This superior performance on the validation set suggests that Gradient Boosting is effectively capturing the underlying patterns in the data, making it a reliable choice for predicting customer churn in diverse scenarios.

The Decision Tree model achieves an impressive fit on the training data, indicated by a training ASE of 0.17848, the best among the models. However, this model's performance falters on the validation set, where the ASE rises to 0.23568. The notable increase in error from training to validation signals a potential overfitting issue, where the model has learned specific details of the training data that do not generalize well to unseen data, limiting its practical applicability for churn prediction.

The Random Forest model, marked by an ASE of 0.21980 on the training data, does not fit the training set as closely as the Decision Tree model. Its validation ASE further increases to 0.23746, the highest observed in this model comparison. This indicates a more modest performance in generalizing beyond the training data. Despite this, the Random Forest's intrinsic mechanism of bagging multiple decision trees can still provide a more stable and generalized prediction in certain cases, although it may require further tuning to enhance its predictive capabilities in the churn context. Considering the ASE values, the Gradient Boosting model appears to strike the best balance between fitting the training data and maintaining performance on the validation set, pointing to its superior ability to generalize. The Decision Tree, while excellent at learning from the training set, seems to overfit and thus performs less well on validation data. The Random Forest model demonstrates a consistent performance between training and validation but does not reach the predictive accuracy of the Gradient Boosting model in this particular instance.

In conclusion, the Gradient Boosting model exhibits the strongest performance for predicting customer churn on the validation dataset, as indicated by the lowest Average Squared Error (ASE). This suggests that it is the most accurate at generalizing from the training data to unseen data. However, given the signs of potential overfitting, it would be beneficial to monitor this model closely in a live environment and perhaps retrain regularly with new data.

The Decision Tree model shows a good fit to the training data but does not perform as well on the validation set, indicating that it may not generalize as effectively. This could be a result of the model's tendency to overfit to the training data, capturing noise rather than the underlying patterns.

The HP Random Forest model, while not outperforming the Gradient Boosting in this instance, still offers valuable characteristics such as robustness against overfitting and the ability to handle diverse data types and structures. Its performance might be enhanced with further tuning of hyperparameters or by using a larger dataset for training.

Each model has its strengths and can be chosen based on the specific requirements of the task at hand. If the highest predictive accuracy on new data is the priority, Gradient Boosting would be the preferred choice. If interpretability or computational efficiency is more important, or if the data changes frequently, a Decision Tree might be more appropriate. For scenarios where the primary concern is avoiding overfitting and ensuring model stability, Random Forest could be the best option.

6. Conclusion

Throughout this project, we've embarked on a comprehensive examination of various predictive modeling techniques to address the critical business challenge of customer churn. By employing a trio of machine learning models—Gradient Boosting, Decision Tree, and Random Forest—we sought to identify the most accurate and reliable method for predicting customer turnover.

Our analysis revealed that the Gradient Boosting modeL outshined its counterparts in terms of validation performance, showcasing its prowess in handling complex patterns and its robustness against overfitting when compared to the simpler Decision Tree model. This suggests that for intricate datasets where customer behavior is nuanced, Gradient Boosting is a formidable tool capable of providing deep insights and accurate predictions.

Conversely, the Decision Tree model offered a stark contrast with its exceptional fit to the training data but a notable decline in accuracy on the validation set. This highlighted the model's tendency to overfit and its limitations when extrapolating to broader, unseen datasets.

The HP Random Forest model demonstrated a middle ground, with a performance that did not reach the heights of the Gradient Boosting model but still offered a level of stability and generalization, thanks to its ensemble nature. Though not the top performer in this instance, its method of aggregating multiple decision trees could offer valuable insights, especially with further tuning and in conjunction with larger datasets.

The endeavor underscores the importance of model selection in predictive analytics and the trade-offs between interpretability, accuracy, and the ability to generalize. It also emphasizes the necessity of model validation and the potential benefits of ensemble methods in achieving both high predictive performance and robustness against overfitting.

As we conclude this project, the overarching takeaway is that an iterative approach to model building and selection, paired with diligent validation, is crucial in developing an effective churn prediction strategy. The insights gleaned from the Gradient Boosting model, with its superior validation performance, have laid a foundation for a strategic approach to customer retention, enabling targeted interventions aimed at maintaining a loyal customer base.

Reference:

Link to GitHub repository: https://github.com/JIANGJJ12/7005-AA1