**Using Talend DataPrep for data exploration**

A screenshot of a computer

Description automatically generated

The dataset's columns with missing data include 'Age,' 'Gender,' 'Location,' 'Membership Level,' 'TotalPurchases,' 'TotalSpent,' 'FavoriteCategory,' and 'Occupation.' Missing data in these fields can significantly impact customer profiling, demographic analysis, and behavioral insights. For instance, unknown ages or locations hinder demographic segmentation, while missing purchase details affect spending pattern analysis. Additionally, gaps in gender, membership level, and favorite category data can challenge targeted marketing efforts and personalized customer experiences. Overall, these missing values present obstacles in understanding and effectively engaging with the customer base.

The dataset also shows inconsistencies in three columns: 'Age,' 'TotalPurchases,' and 'TotalSpent.' Age values containing decimals are unusual since age is typically a whole number. Negative values in 'TotalPurchases' and 'TotalSpent' are also illogical as they should represent non-negative quantities like number of purchases or amounts spent. 'Gender' includes 'Unknown' values, which might indicate missing or unrecorded gender data. These inconsistencies suggest data quality issues, necessitating data cleaning and validation to ensure accurate analysis.

**Data transformation**:

A screenshot of a computer

Description automatically generated

Utilizing Talend Data Preparation for data transformation for the dataset is the best choice due to its enhanced visualization features and ease of use. In Talend, we can quickly adjust the 'Age' column to remove decimal points, ensuring age data is represented as whole numbers. For 'TotalPurchase', Talend's functionality allows for the removal of decimals and rounding of values, creating consistency in purchase data. Additionally, Talend can effectively drop all negative values from the 'TotalSpent' column, which is logical for expenditure data. This decision over SAS Miner is influenced by Talend's more straightforward steps for data manipulation and its superior capabilities in identifying data inconsistencies and missing values, making it a more user-friendly and efficient tool for these specific data transformations.

**Load the DataSet to SAS Miner and specify variable roles**

A screenshot of a computer

Description automatically generated

In the context of SAS Miner, setting 'FavoriteCategory' as the target variable aligns to predict customer preferences regarding the product. By defining 'FavoriteCategory' as the target, the model focuses on identifying patterns and factors that influence a customer's likelihood to purchase from or become a member of a specific category. This approach enables more targeted marketing strategies and a deeper understanding of customer behavior, helping in personalized customer engagement and improved business decision-making.

Using the “StatExplore” node to view data class and interval variables in the dataset

A screenshot of a computer

Description automatically generated

StatExpore node shows the input and target variables, class variable statistics indicating frequency and mode, and interval variable statistics like mean, median, and standard deviation. Chi-square statistics are provided for each input variable, indicating their significance in predicting the target. This comprehensive analysis, including summaries by class target and overall statistics, offers insights into customer preferences and behaviors, crucial for strategic business decisions.

**Impute missing values with SAS Miner**

**A screenshot of a computer

Description automatically generated**

**A computer screen shot of a computer

Description automatically generated**

The above report highlights an imputation summary where missing values in the dataset are addressed in Sas Enterprise Miner. For instance, 'Age' and 'TotalPurchases' use median values for imputation, while other variables like 'Gender' and 'Location' use the most frequent value (mode). Each imputed variable had 200 missing observations in the training data. The model focuses on predicting the 'FavoriteCategory', with all variables fully accounted for in the training data.

**Decision Tree Analysis without imputing missing data**

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

The decision tree model summary shows a focus on the "FavoriteCategory" target variable, using a training set of 1800 observations. The model shows a high misclassification rate of approximately 62.72%, indicating a significant portion of predictions are incorrect. The maximum absolute error and other error metrics like SSE (Sum of Squared Errors), ASE (Average Squared Error), and RASE (Root Average Squared Error) suggest obvious discrepancies between predicted and actual values. The lack of data for validation and test phases implies limited assessment scope, and the high error rates point towards the model's suboptimal performance in accurately predicting the target variable.

**Decision tree Analysis after imputing missing data**

A screenshot of a computer

Description automatically generated

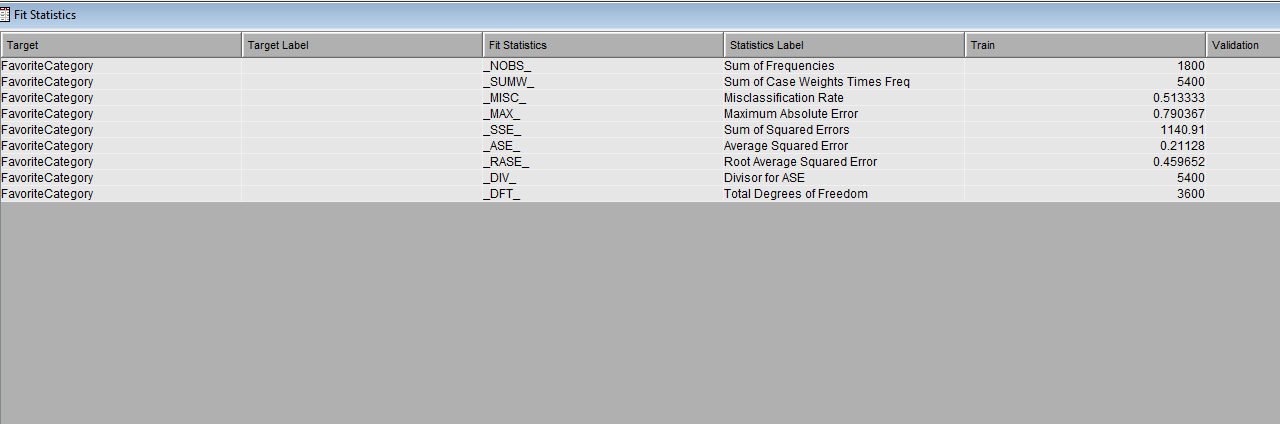
A screenshot of a computer

Description automatically generated

The analysis of the imputed "FavoriteCategory" shows a misclassification rate of 44.4%, which is an improvement compared to the previous model. However, the maximum absolute error remains high. The sum of squared errors and the root average squared error indicate that while the model's performance has improved with imputation, there is still a notable level of prediction error. These results suggest that while imputation techniques can enhance model accuracy, understanding customer preferences in shopping categories remains complex and may require further refinement of the model or additional data features.

From the two sets of observations (before and after imputing missing), it's evident that the decision tree model for predicting "FavoriteCategory" has a high misclassification rate, around 51-63%, indicating a significant portion of the predictions are incorrect. This can be seen by high values in maximum absolute error and sum of squared errors. The lack of validation and test data suggests the model's evaluation is based only on training data, which limits the understanding of its generalization capabilities. Overall, these metrics show that the decision tree may not be very effective or accurate in predicting the target variable which is the FavoriteCategory.

**Gradient Boosting before imputing missing data:**



The summary of the decision tree model for predicting "FavoriteCategory" shows it was trained on 1800 data points, with a misclassification rate of about 51.33%. The model's maximum absolute error and the sum of squared errors are high, indicating significant prediction inaccuracies. The average squared error and its root (RASE) further underscore these prediction challenges. No data is provided for validation or testing phases, emphasizing the model's limited performance evaluation based solely on training data.

**Gradient Boosting after imputing missing data:**

**A screenshot of a computer

Description automatically generated**

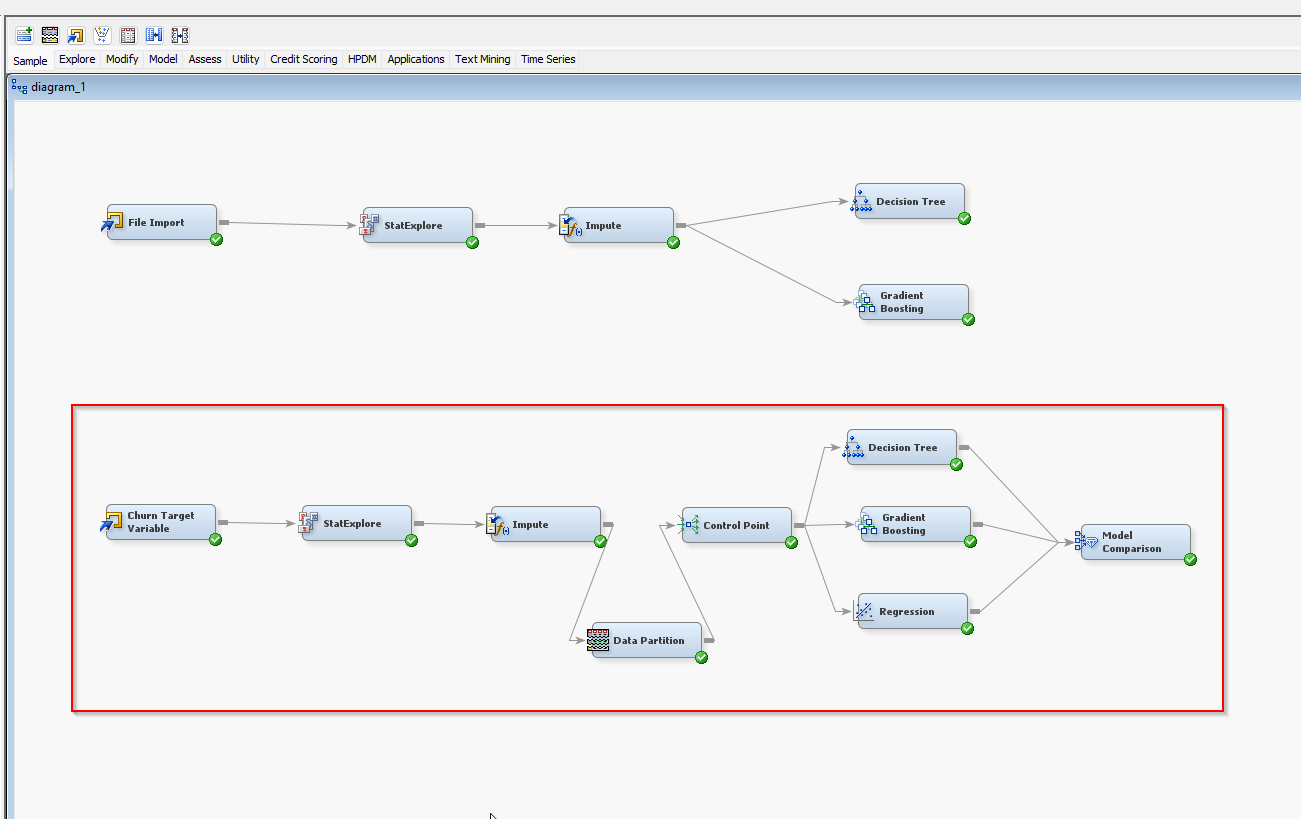
After imputing the missing data, the maximum absolute error and the sum of squared errors are still high, indicating significant discrepancies in predictions. The average squared error and root average squared error metrics further reflect these prediction challenges. These statistics, particularly the high misclassification rate, suggest Gradient Boosting might be struggling to accurately predict the favorite category, even after addressing missing data through imputation.

**Conclusion on model performance**

The high misclassification rate in the decision tree analysis and gradient boosting suggests that predicting a customer's favorite shopping category is complex and influenced by multiple factors. This could imply that customer preferences are diverse and not easily categorized, highlighting the need for a more specific approach to understanding customer behavior. The data indicates potential gaps in the current model or a lack of sufficient or relevant features for accurate prediction. For businesses, this suggests investing in more comprehensive data collection to better understand and cater to varied customer preferences.

**Changing the Target Variable**

Shifting the business objective to predict customer churn and selecting 'Churn' as the target variable is a strategic change, given that 'FavoriteCategory' was not yielding optimal modeling results. By creating a new flow in SAS Miner with 'Churn' as the target, the aim is to assess if models perform better with this variable. This approach includes adding a Data Partition node to divide the data into training and validation sets. Additionally, incorporating a logistic regression model alongside decision tree and gradient boosting models allows for a comprehensive comparison. A Model Comparison node will then evaluate the performance of each model, determining the most effective approach for predicting customer churn.



In the data partition node, a 70/30 split is used, allocating 70% of the data for training and 30% for validation. This ratio is a standard practice in machine learning as it provides a substantial amount of data for training the model, ensuring it learns effectively. Simultaneously, the 30% validation portion is adequate to evaluate the model's performance and its ability to generalize to unseen data. This balance helps in achieving a comprehensive understanding of the model's capabilities and limitations in real-world scenarios.

A screenshot of a computer

Description automatically generated

Result:

A white screen with black text

Description automatically generated

After analyzing three models in SAS Miner from the above Fit Statistic result, the Decision Tree (Tree2) emerged as the most effective, demonstrating the lowest misclassification rate on validation data. This better performance in validation suggests better generalization capability. The Gradient Boosting (Boost3) model, while slightly less effective than the Decision Tree, also showed a commendable performance. In contrast, the Regression (Reg) model exhibited the highest misclassification rate, indicating it was less suited for this dataset compared to the other models.

The observed performance of the models indicates better results compared to the initial analysis where the dataset wasn't split. However, a deeper look reveals that each model's validation misclassification rate is higher than its training rate. This pattern is indicative of overfitting, where models perform well on training data but less so on unseen validation data. It suggests the models are too tailored to the specifics of the training set, forcing further tuning to enhance their generalization to new, unseen data.

A white background with black text

Description automatically generated

Besides the misclassification rate, we can also look at the Event Classification Table report that shows the performance metrics for Regression, Gradient Boosting, and Decision Tree models on training and validation data for 'Churn' prediction. In training, Regression correctly identified fewer true positives and negatives compared to Gradient Boosting and Decision Tree models. In validation, the Decision Tree model has a higher rate of true positives and negatives than Regression, indicating better performance. Gradient Boosting shows significant predictive accuracy in both training and validation, particularly in identifying true positives and negatives.

**Challenges**

Using dummy datasets poses several challenges. Firstly, they often lack real-world complexity, leading to oversimplified models that don't accurately reflect true data variability. Secondly, there's a high risk of overfitting, as models might perform well on these simplified datasets but fail to generalize to real-world data. Thirdly, dummy datasets may not offer the diversity and volume required for advanced models to learn effectively, limiting their scope for learning. Additionally, such datasets can introduce bias if they don't adequately represent the actual data's diversity. Lastly, dummy datasets might not encompass all possible scenarios or edge cases, limiting the effectiveness of testing.