1. Import the CSV file using “File Import” node. Save it as a SAS file.

图示

描述已自动生成

2.Creat Library and datasource

图片包含 图形用户界面

描述已自动生成

Drag the data source into a new diagram and perform operations.

文本

描述已自动生成

Right click on case\_study and select Edit Variables. In order to pay attention to customer churn, I set "Churn" as the target variable.

表格

描述已自动生成

**2.Decision Tree Modelling using SAS Enterprise Miner**

Create a Data Partition node and divide the data. 70% is used for train data and 30% is used for validation data.

表格

描述已自动生成

手机屏幕截图

描述已自动生成

图示

描述已自动生成

图形用户界面

描述已自动生成

图形用户界面

低可信度描述已自动生成

Word

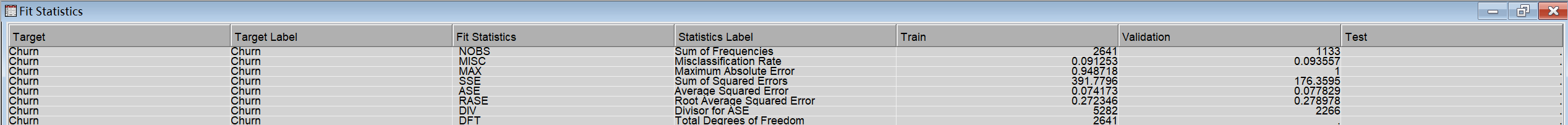
中度可信度描述已自动生成

图形用户界面, 应用程序

描述已自动生成

图表, 折线图

描述已自动生成



Based on Fit Statistics, misclassification rate is 0.0912 for training dataset and 0.0935 for validation dataset.

表格

描述已自动生成

The Variable Importance Plot displays the importance of each predictor variable in the model. Only 8 out of 18 input variables are important to the pruned decision tree model.

**3.Ensemble Methods: Apply Bagging and Boosting, using the Random Forest algorithm as a Bagging example.**

**3.1 Using the Random Forest algorithm as a Bagging**

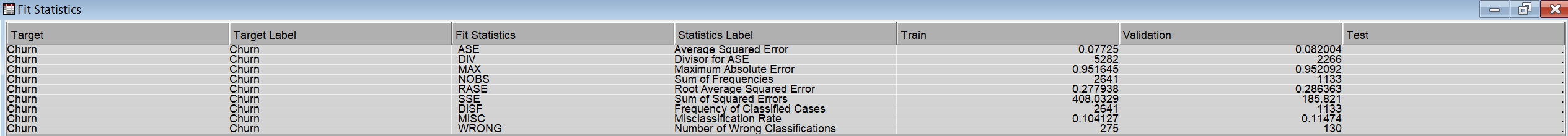
图形用户界面, 应用程序

描述已自动生成

图形用户界面, 图表

描述已自动生成

Based on Iteration Plot, misclassification rate plateaued out when number of trees reaches 25.



Based on Fit Statistics, misclassification rate is 0.1041 for training dataset and 0.1147 for validation dataset.

图表, 折线图

描述已自动生成

**3.2 Gradient Boosting**

图形用户界面

低可信度描述已自动生成

图表, 折线图

描述已自动生成

图片包含 表格

描述已自动生成

It can also be seen that the most important variable of the Boosting model is Tenure. A total of 13 variables are used by the Boosting model.

表格

描述已自动生成

表格

描述已自动生成

表格

描述已自动生成

表格

描述已自动生成

As can be seen from the above figure, the misclassification rate is only 0.1, so the model has better effect.

**4.Compare models**

Use the model comparison node Model Compare to compare the results of the three models. The result is as follows:

图形用户界面

中度可信度描述已自动生成

The above is the setting process of the model comparison node. The misclassification rate is selected as the criterion for selecting the best model. The results are as follows:

图片包含 图表

描述已自动生成 表格

描述已自动生成

As can be seen from the figure, the misclassification rate of Decision tree is 0.0912, boost is 0.0954, and bagging is 0.1041. The Decision Tree has the lowest misclassification rate (0.0912), making it the best-performing model among the three based on the given metric. Lower misclassification rates usually suggest better predictive performance.

Examining decision tree and ensemble models, specifically in the context of customer behavior, offers valuable insights for shaping business strategies. The Decision Tree model, boasting low Root Average Squared Error (RASE) and Sum of Squared Errors (SSE), lays a robust foundation for comprehending factors influencing customer loyalty and churn. Crucial factors like "Tenure," "Preferred Login Device," and "Satisfaction Score" emerge as pivotal in shaping customer decisions. Meanwhile, the Boosting model, despite slightly higher RASE and SSE, delves into nuanced patterns, adding depth to the analysis. On the contrary, the HPDM, likely a hyperparameter-tuned Decision Tree, presents higher complexity with an elevated RASE, prompting careful consideration.

Strategic recommendations involve prioritizing insights from the Decision Tree, harnessing the nuanced findings of the Boosting model, and thoughtfully evaluating the benefits of hyperparameter tuning. Businesses can refine customer retention strategies, tailor promotions based on factors like "Coupon Usage" and "Order Frequency," and implement targeted engagement approaches, taking into account "Days Since Last Order."

In essence, a thorough examination of decision tree and ensemble models equips businesses with actionable insights to elevate customer retention strategies, optimize promotional initiatives, and enhance overall customer engagement.