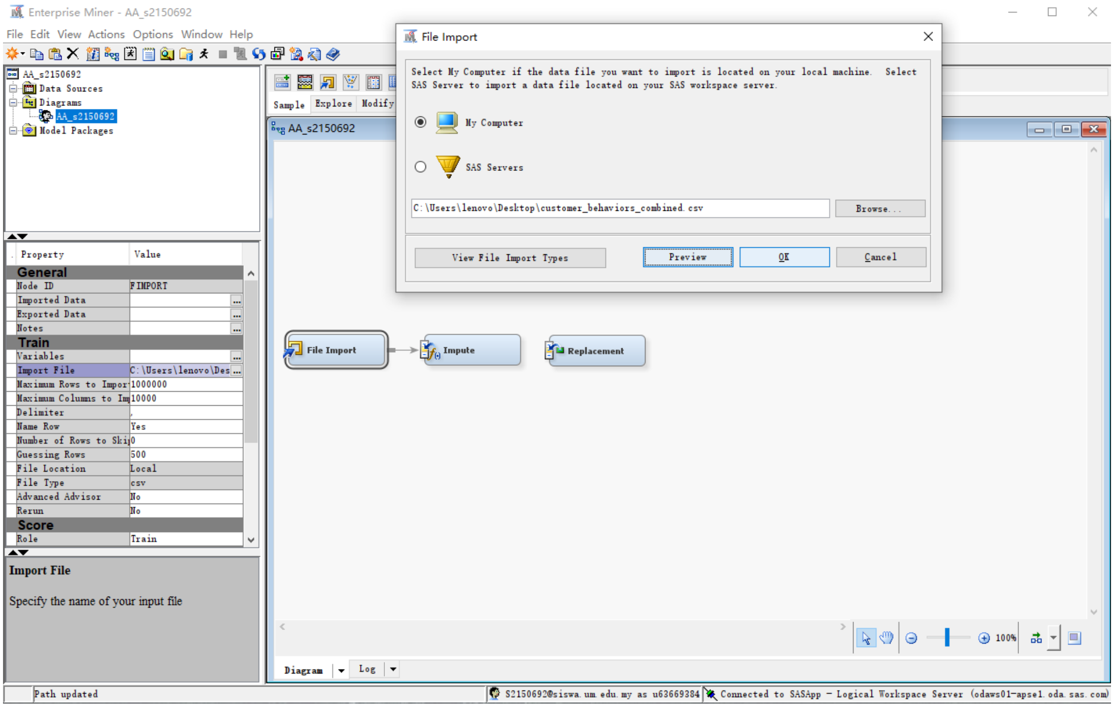
WQD7005 AA1

Name: YinQiXiang ID: S2150692

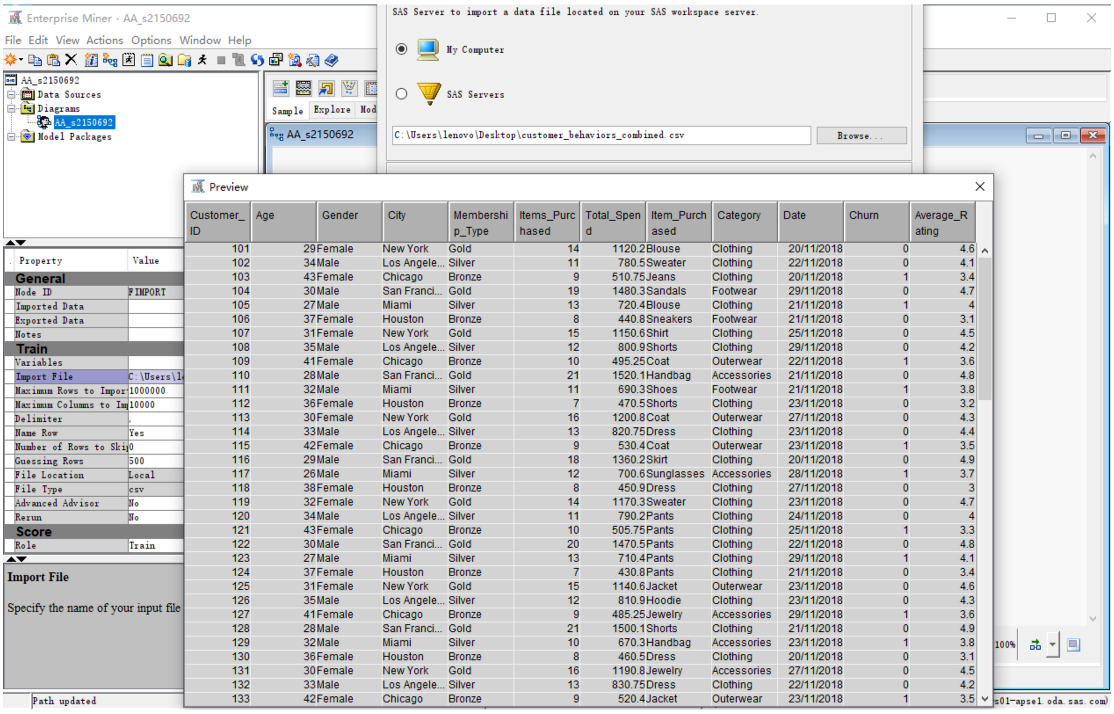
Github: https://github.com/SumerYin/YinQiXiang_S2150692

Data Import and Preprocessing: Import your dataset into SAS Enterprise Miner, handle missing values, and specify variable roles.

Import Dataset



Import Dataset from the local computer .csv file into the SAS EM.

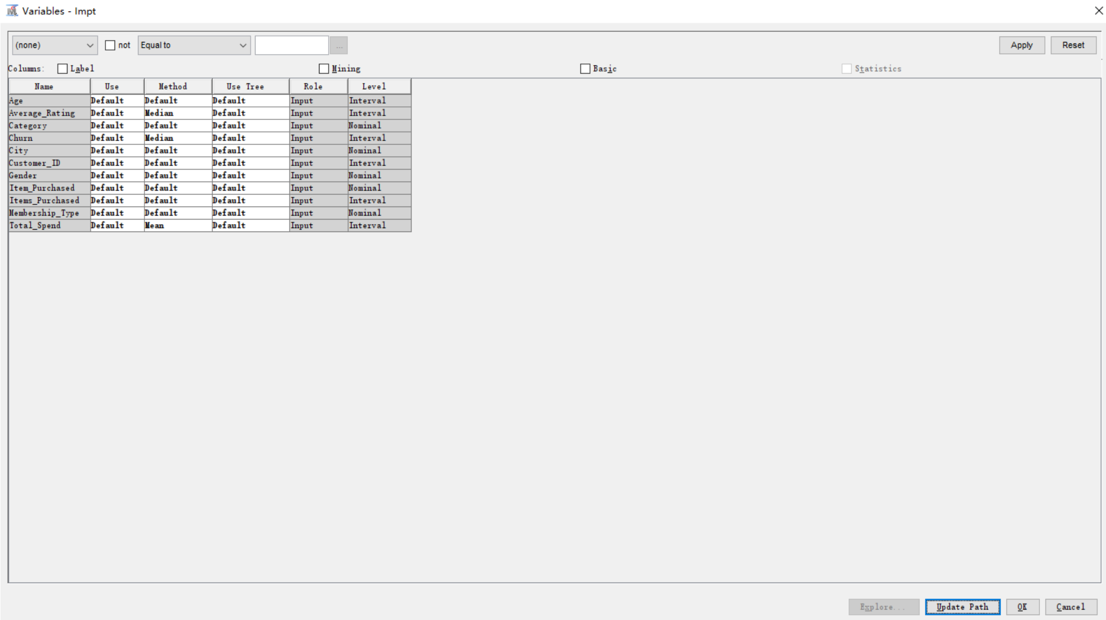


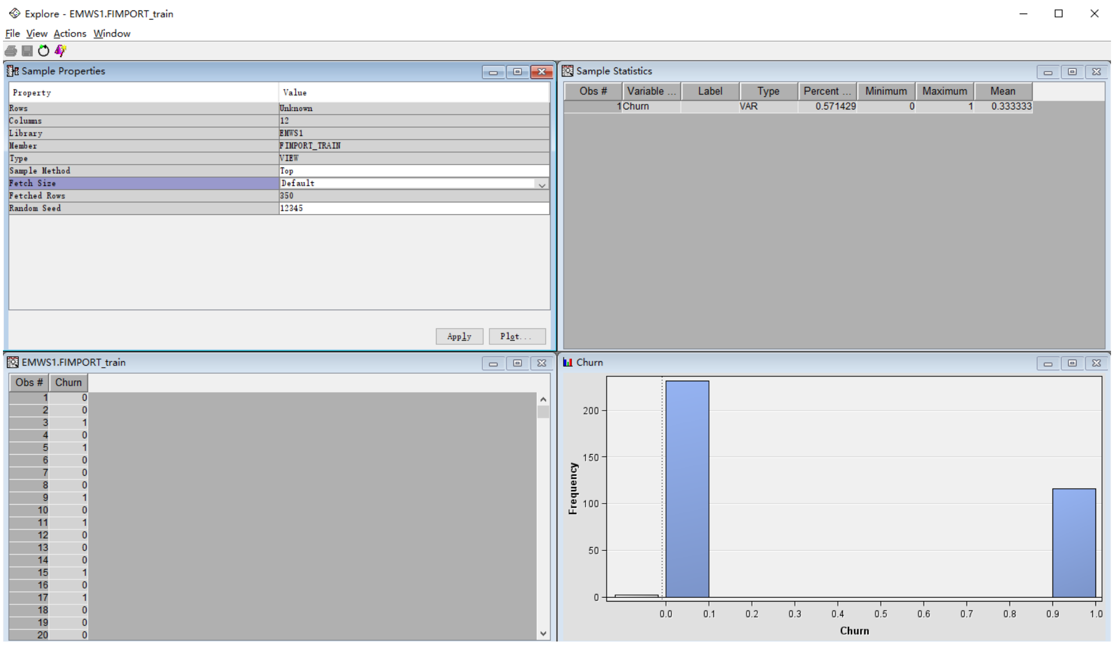
Customer_ID	Age	Gender	City	Membership_Type	Items_Purchased	Total_Spend	Item_Purchased	Category	Date	Churn	Average_Rating
101	29	Female	New York	Gold	14	1120.2	Blouse	Clothing	20/11/2018	0	4.6
102	34	Male	Los Angele...	Silver	11	780.5	Sweater	Clothing	22/11/2018	0	4.1
103	43	Female	Chicago	Bronze	9	510.75	Jeans	Clothing	20/11/2018	1	3.4
104	30	Male	San Franci...	Gold	19	1480.3	Sandals	Footwear	29/11/2018	0	4.7
105	27	Male	Miami	Silver	13	720.4	Blouse	Clothing	21/11/2018	1	4
106	37	Female	Houston	Bronze	8	440.8	Sneakers	Footwear	21/11/2018	0	3.1
107	31	Female	New York	Gold	15	1150.6	Shirt	Clothing	25/11/2018	0	4.5
108	35	Male	Los Angele...	Silver	12	800.9	Shorts	Clothing	29/11/2018	0	4.2
109	41	Female	Chicago	Bronze	10	495.25	Coat	Outerwear	22/11/2018	1	3.6
110	28	Male	San Franci...	Gold	21	1520.1	Handbag	Accessories	21/11/2018	0	4.8
111	32	Male	Miami	Silver	11	690.3	Shoes	Footwear	21/11/2018	1	3.8
112	36	Female	Houston	Bronze	7	470.5	Shorts	Clothing	23/11/2018	0	3.2
113	30	Female	New York	Gold	16	1200.8	Coat	Outerwear	27/11/2018	0	4.3
114	33	Male	Los Angele...	Silver	13	820.75	Dress	Clothing	23/11/2018	0	4.4
115	42	Female	Chicago	Bronze	9	530.4	Coat	Outerwear	23/11/2018	1	3.5
116	29	Male	San Franci...	Gold	18	1360.2	Skirt	Clothing	20/11/2018	0	4.9
117	26	Male	Miami	Silver	12	700.6	Sunglasses	Accessories	28/11/2018	1	3.7
118	38	Female	Houston	Bronze	8	450.9	Dress	Clothing	27/11/2018	0	3
119	32	Female	New York	Gold	14	1170.3	Sweater	Clothing	23/11/2018	0	4.7
120	34	Male	Los Angele...	Silver	11	790.2	Pants	Clothing	24/11/2018	0	4
121	43	Female	Chicago	Bronze	10	505.75	Pants	Clothing	25/11/2018	1	3.3
122	30	Male	San Franci...	Gold	20	1470.5	Pants	Clothing	22/11/2018	0	4.8
123	27	Male	Miami	Silver	13	710.4	Pants	Clothing	29/11/2018	1	4.1
124	37	Female	Houston	Bronze	7	430.8	Pants	Clothing	21/11/2018	0	3.4
125	31	Female	New York	Gold	15	1140.6	Jacket	Outerwear	23/11/2018	0	4.6
126	35	Male	Los Angele...	Silver	12	810.9	Hoodie	Clothing	23/11/2018	0	4.3
127	41	Female	Chicago	Bronze	9	485.25	Jewelry	Accessories	29/11/2018	1	3.6
128	28	Male	San Franci...	Gold	21	1500.1	Shorts	Clothing	21/11/2018	0	4.9
129	32	Male	Miami	Silver	10	670.3	Handbag	Accessories	23/11/2018	1	3.8
130	36	Female	Houston	Bronze	8	460.5	Dress	Clothing	20/11/2018	0	3.1
131	30	Female	New York	Gold	16	1190.8	Jewelry	Accessories	27/11/2018	0	4.5
132	33	Male	Los Angele...	Silver	13	830.75	Dress	Clothing	22/11/2018	0	4.2
133	42	Female	Chicago	Bronze	9	520.4	Jacket	Outerwear	23/11/2018	1	3.5

Preview the Dataset first to check each Column and Variables in the Dataset.

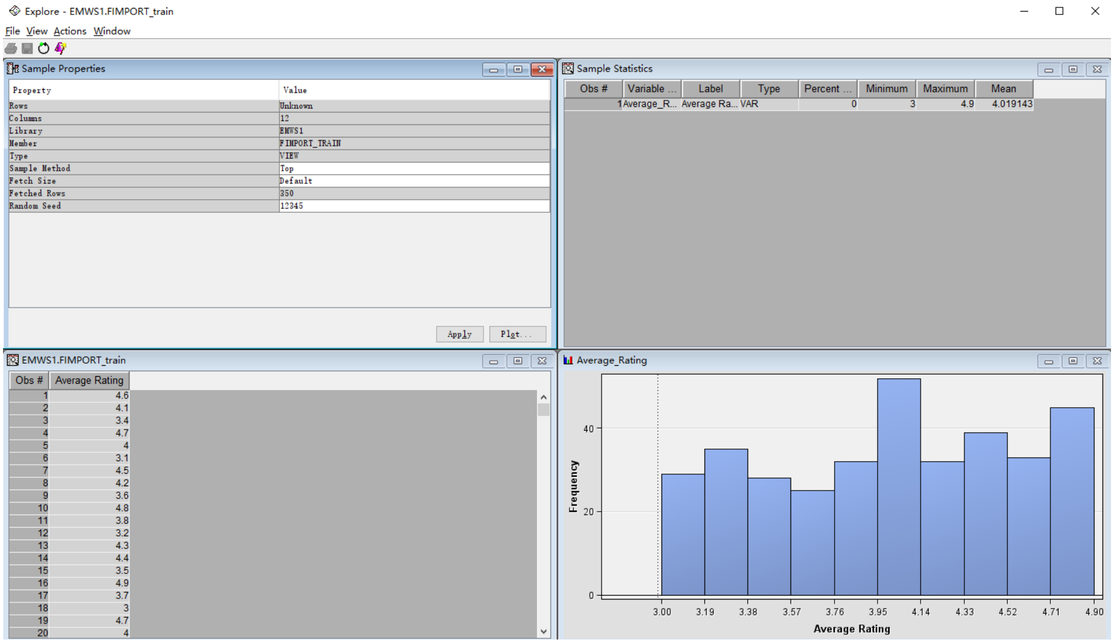
Handle missing values, and specify variable roles.

Impute

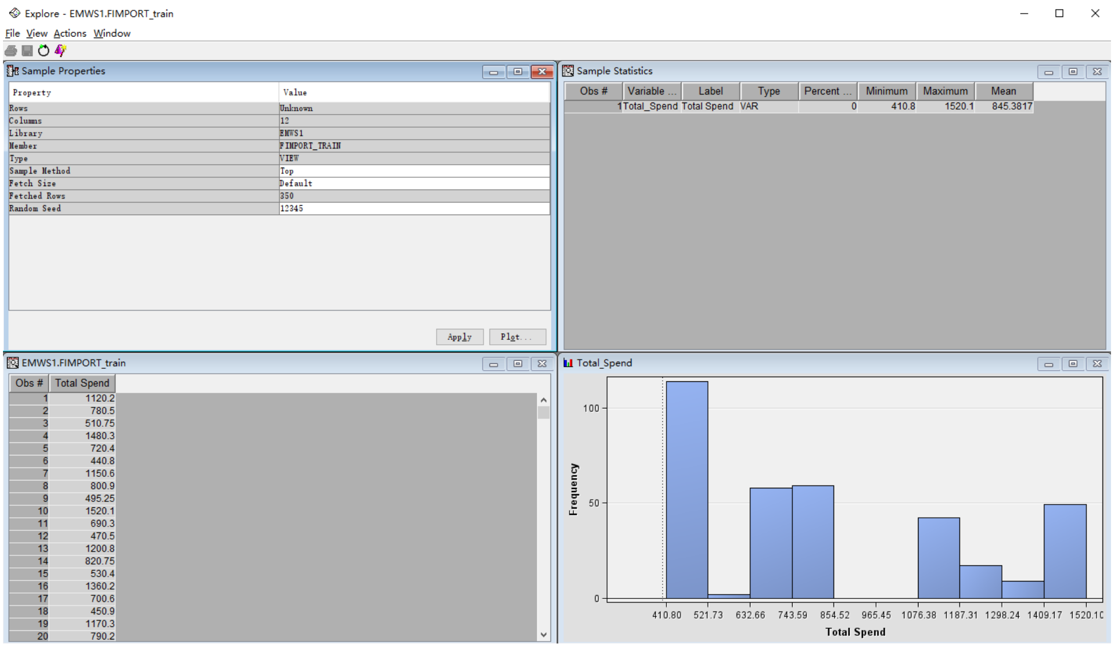




Impute the missing value for churn which is our target in this project using median Method for churn. For missing interval variables, it can use the mean or median of the entire variable to fill in the missing values. This method can maintain the overall distribution trend of the data and may be reasonable for most cases. In here it's applied median method.

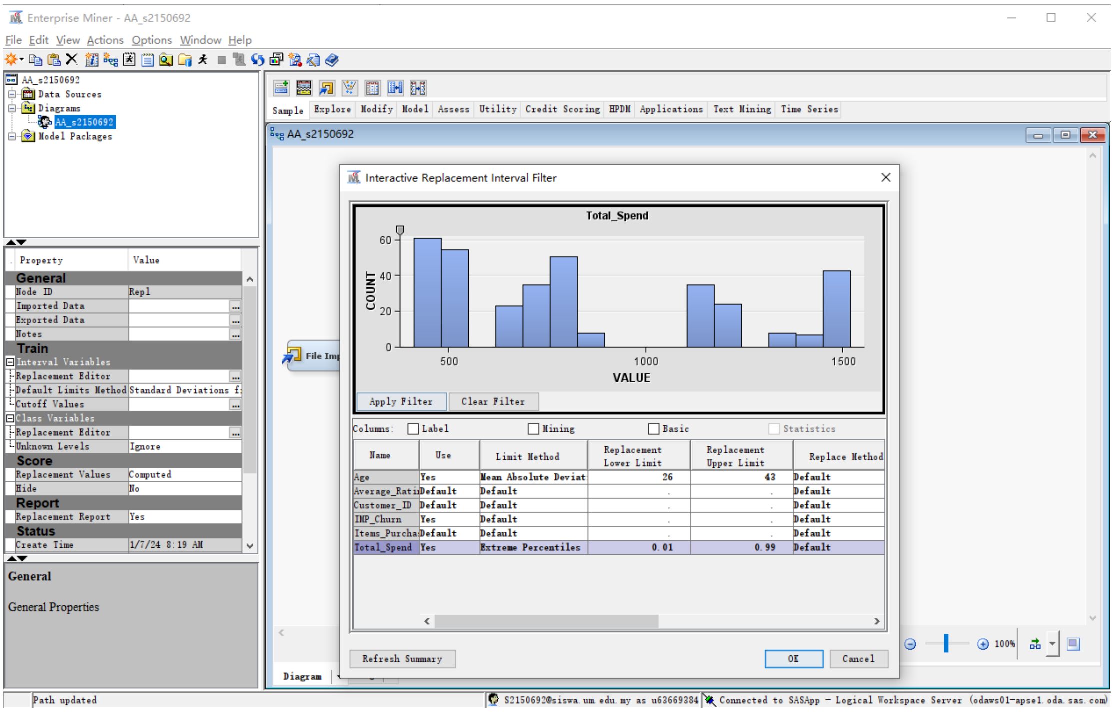


Also including the average_Rating for interval Variable using median method to handle

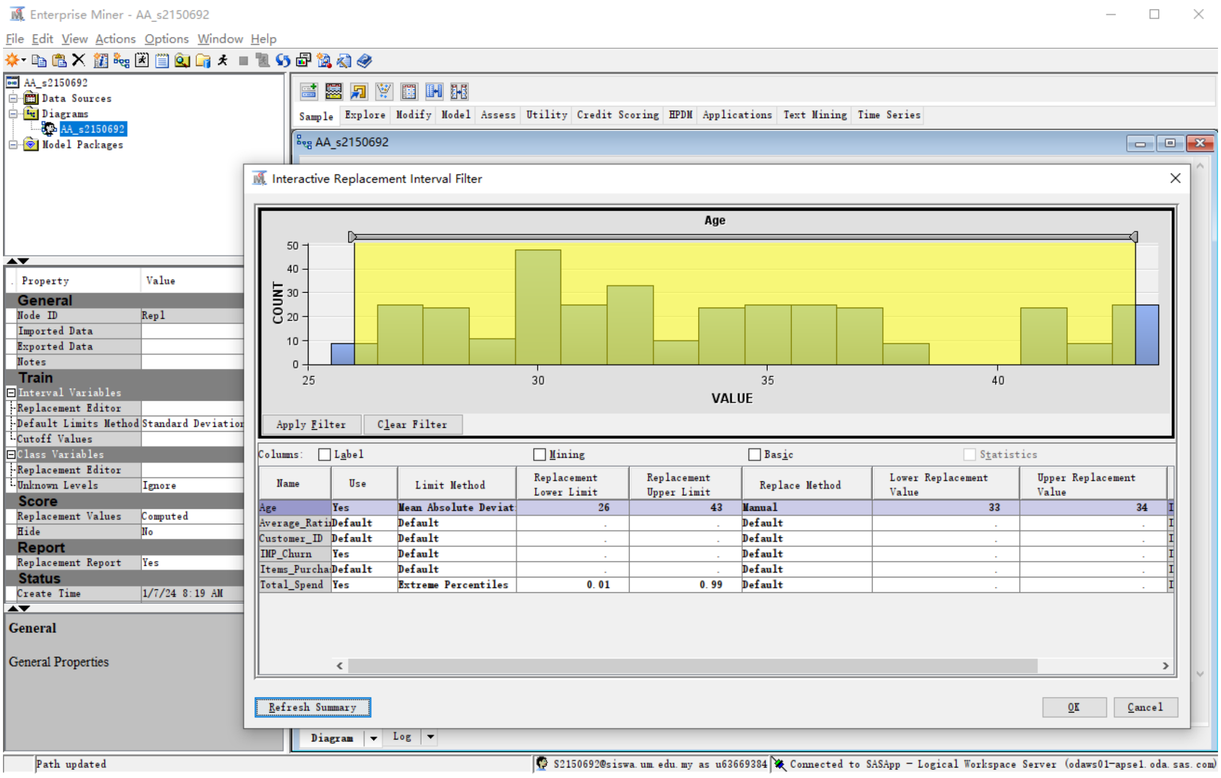


Also Explore the total_spend using median method is more suitable to deal with it. Check the distribution of total_spend. The distribution of the data is relatively asymmetric and has too many extreme values. Median filling, etc., fills missing values more accurately

Replacement



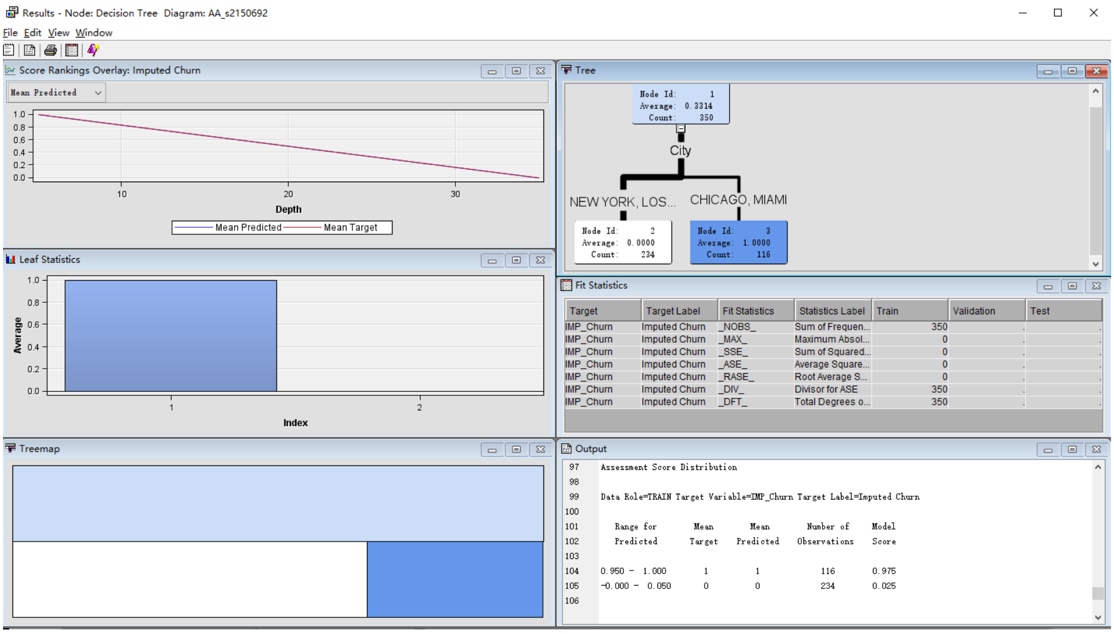
For Total_spend using the Extreme percentiles to deal with the Extreme data in this Column. The "Extreme percentiles" method is often used to replace extreme values or outliers when working with interval variables. For interval variables like Total_spend, using extreme percentiles to replace possible outliers is a reasonable approach. By choosing appropriate percentiles to replace extreme values, the impact of extreme values on modeling can be reduced and the robustness of the model can be improved. Here we chose 1% and 99



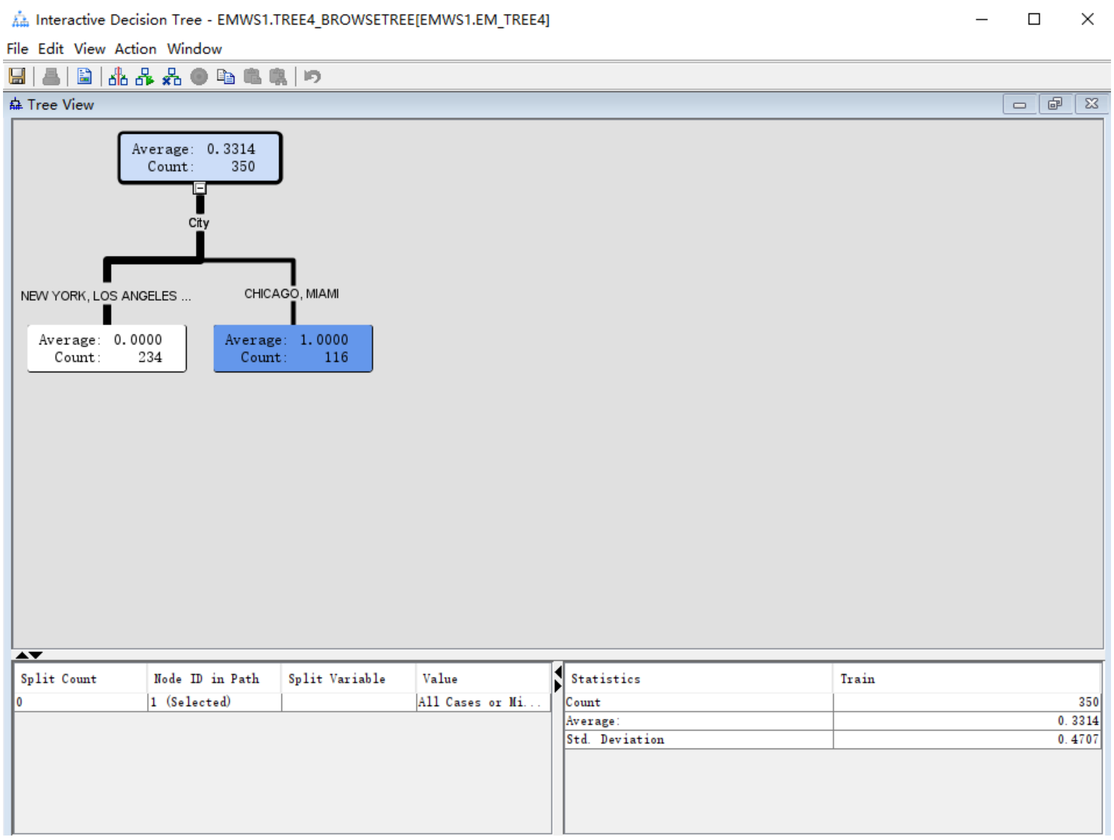
For Age Column it use Mean Absolute Deviate method to handled. "Mean Absolute Deviation" is used to measure the degree of dispersion of data. For variables such as Age, you can consider using the mean absolute deviation to identify and deal with possible outliers. Replace the Age values identified as outliers with the mean, mean is 33.5. Replace the minimum value with 33 and the maximum value with 34. Using the mean absolute deviation to deal with outliers of the Age variable can help improve the robustness of the model and reduce the interference of outliers on modeling.

Decision Tree Analysis: Create a decision tree model in SAS Enterprise Miner to analyse customer behaviour.

Decision Tree



Decision tree 1 result



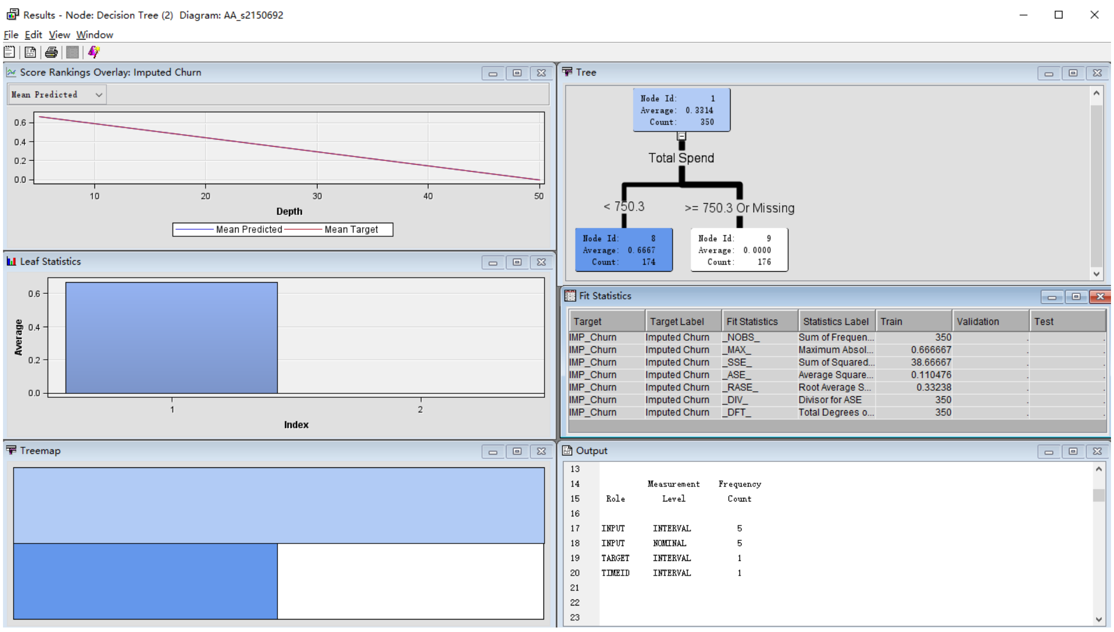
Decision tree 1: The tree structure reveals that geographic location, specifically the city, is a major determinant in predicting customer churn. Customers from New York and Los Angeles show a tendency not to churn (average churn of 0), whereas customers from Chicago and Miami show a definite churn (average churn of 1).

Model Performance and Predictions:

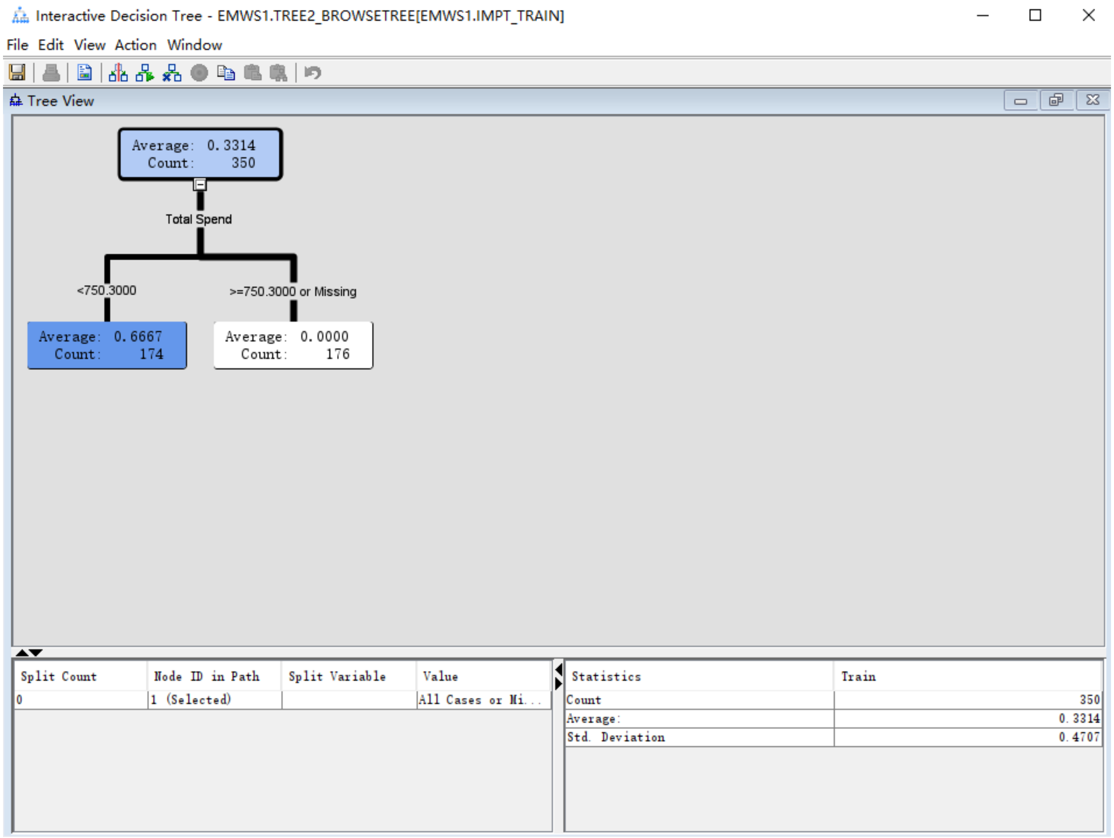
The Score Rankings Overlay graph shows that as we move deeper into the tree (increasing tree depth), the mean predicted churn probability converges closely with the mean target (actual churn), indicating that the model is performing well in separating churners from non-churners.

The Leaf Statistics bar graph reflects that one leaf node predicts no churn (average = 0) for a large group of customers, while another predicts churn (average = 1) for a smaller group, suggesting that the model has identified patterns that strongly differentiate between churners and non-churners.

From these observations, the analysis suggests that customer churn is heavily influenced by their city of residence. The model's accuracy in the training phase is reasonably good, as indicated by the closeness of the predicted and actual churn rates. The identified patterns in customer behavior based on geographic segmentation can be used to tailor specific retention strategies. For instance, companies could deploy targeted marketing campaigns or customer service enhancements in Chicago and Miami to address the higher churn rates in those cities.



Decision tree 2 result



The decision tree indicates that 'Total Spend' is a key variable in predicting customer churn. The tree splits on this variable, suggesting that spending behavior is predictive of churn.

Customers with a 'Total Spend' less than 750.3 have a higher likelihood of churning (average = 0.6667) compared to those with a 'Total Spend' greater than or equal to 750.3 or missing data, who have a lower propensity to churn (average = 0.0000).

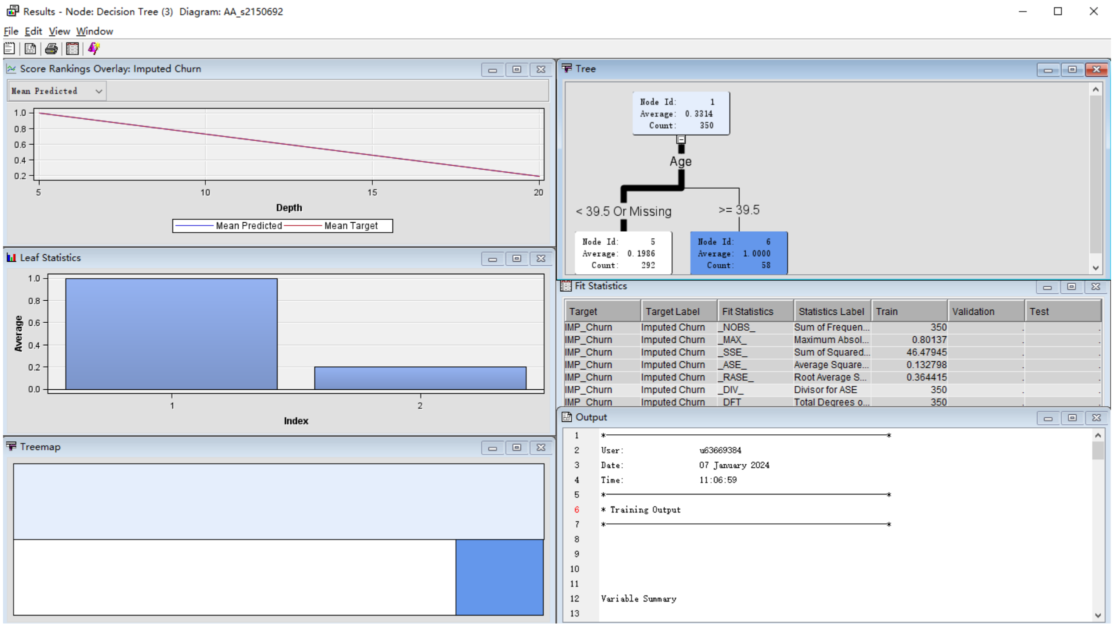
Model Performance Evaluation:

The Score Rankings Overlay graph shows that the mean predicted churn probability is relatively stable across different depths of the tree. The mean target (actual churn) line and the mean predicted line are close, which indicates that the model has a consistent prediction capability. The Leaf Statistics graph shows that the model can distinctly classify customers into high-risk (churn) and low-risk groups based on their total spending.

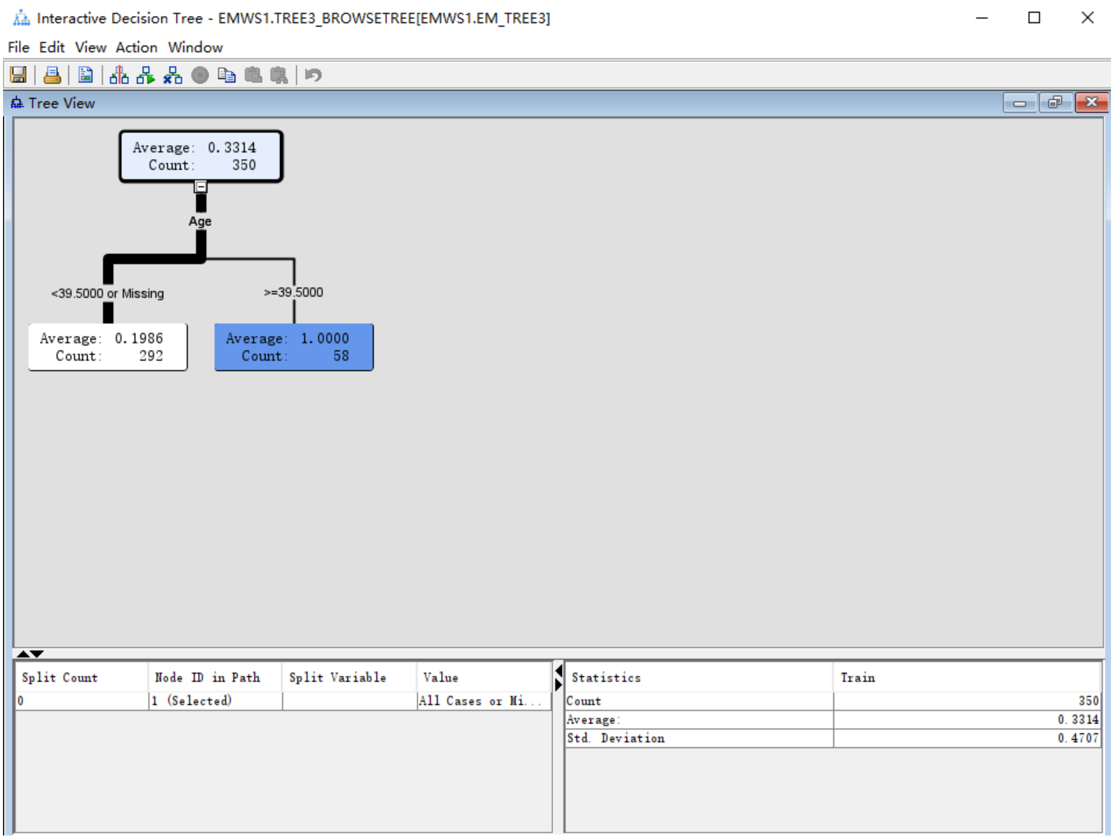
Interpretation and Strategic Implications:

The clear separation of churn risk based on 'Total Spend' suggests that spending levels are strongly associated with churn likelihood. This analysis could imply that customers with lower spending are at a higher risk of churning and may require targeted engagement and retention strategies. Conversely, customers with higher spending or those with missing spend data (which could suggest new customers or data capture issues) are not churning and may represent a stable or satisfied customer base.

In conclusion, the Decision Tree model provides actionable insights, highlighting 'Total Spend' as a critical factor in customer retention efforts. This finding can guide the development of differentiated strategies, such as personalized promotions for lower-spending customers to increase their engagement and reduce churn risk.



Decision tree 3 result



The decision tree uses 'Age' as a predictor for customer churn. The tree splits customers into two groups: those younger than 39.5 years and those 39.5 or older, including missing data on age. Customers younger than 39.5 have a lower churn rate (average = 0.1986), whereas all customers 39.5 and older have a churn rate of 1, indicating they all churned.

Model Performance Evaluation:

The Score Rankings Overlay graph shows a stable mean predicted probability of churn across tree depths, with the prediction line closely following the actual churn line, indicating good model prediction capability.

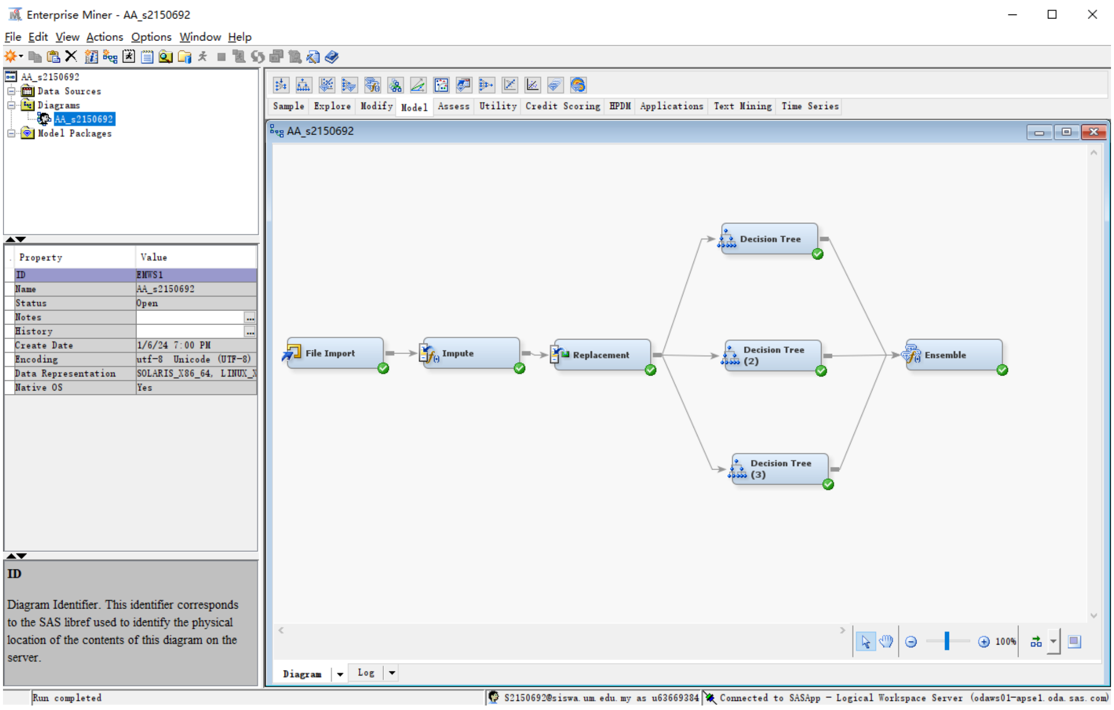
The Leaf Statistics graph depicts two groups, with the younger customer group having a significantly lower average churn rate compared to the older or missing age data group.

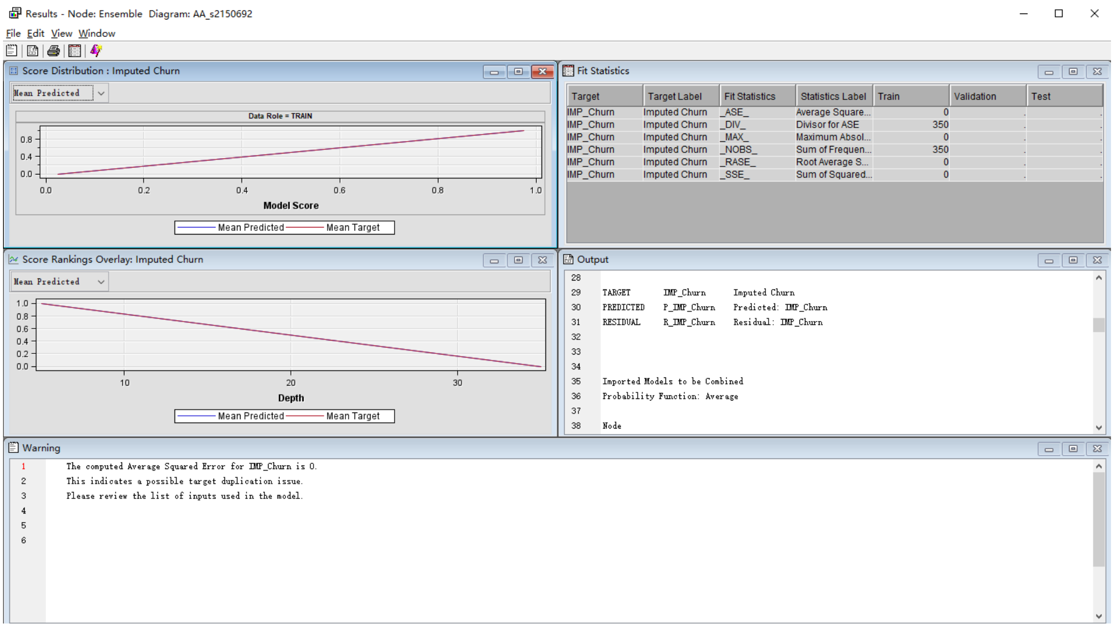
Interpretation and Strategic Implications:

The analysis reveals age as a critical factor in predicting churn, with older customers being at a much higher risk of churning. This could be reflective of different needs or service expectations that are not being met for the older demographic. This insight could guide the development of age-specific customer engagement and retention strategies, such as tailoring services or communication to meet the preferences and expectations of older customers.

In conclusion, the Decision Tree model identifies age as a significant determinant of churn, suggesting the need for targeted strategies to improve customer retention among older customers.

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





The ensemble model's analysis in SAS Enterprise Miner indicates a high predictive accuracy for customer churn, as evidenced by the close alignment between predicted and actual churn probabilities.

Model Performance:

The Score Distribution graph indicates that the mean predicted churn probability is consistent across all model scores, suggesting a stable model. The close alignment of the mean predicted line with the mean target line across the entire score range demonstrates that the model has a good fit.

Score Rankings Overlay:

The Score Rankings Overlay graph shows the mean predicted churn probability against the depth of the ensemble model. The overlay indicates that the model's predictions are closely aligned with the actual churn, reinforcing the model's predictive accuracy.

Fit Statistics:

The Fit Statistics section indicates zero average squared error (ASE), which suggests that the model predictions are very close to the actual outcomes. However, it also notes a possible target duplication issue, prompting a review of the input variables to ensure they are correctly specified and there are no duplicates that could be influencing the model's predictions.

Strategic Insights:

Despite the warning, the ensemble model appears to be highly predictive of churn behavior. This can provide confidence in identifying customers at risk of churn.

The insights from this model should be used to inform customer retention strategies, such as personalized interventions for customers predicted to have a high probability of churning.

In conclusion, while the ensemble model shows strong predictive performance, the warning message suggests a need for further investigation into the input variables and model configuration to ensure the reliability of the predictions before taking strategic actions based on this analysis.