**Predictive Analytics for Loan Default Risk: A SAS Enterprise Miner Approach**

**Gelin Gangoue Mawa**

**MIS530 Predictive Analytics**

**Colorado State University (CSU) Global Campus**

**May 11th, 2025**

**Business Problem**

This project addresses the problem of predicting whether a borrower will experience a serious loan default within the next two years. By leveraging predictive analytics, financial institutions can proactively assess credit risk and make more informed lending decisions. The dataset used in this analysis originates from the publicly available “Give Me Some Credit” dataset on Kaggle, which offers real-world borrower credit behavior and serves as the foundation for building predictive models (Zhu et al., 2023).

**Description of Dataset Fields**

Table 1 summarizes the primary variables and their roles in the modeling process.

**Table 1**

*Variables description and roles*

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Description** | **Role** |
| SeriousDlqin2yrs | Binary target indicating default in next 2 years  (1 = Yes, 0 = No) | Target |
| RevolvingUtilizationOfUnsecuredLines | Credit used vs. available | Input |
| age | Borrower's age | Input |
| NumberOfTime30-59DaysPastDueNotWorse | 30-59 days late payment  occurrences | Input |
| DebtRatio | Debt payments as a % of income | Input |
| MonthlyIncome | Monthly income | Input |
| NumberOfOpenCreditLinesAndLoans | Active credit lines and loans | Input |
| NumberOfTimes90DaysLate | 90+ day late payment occurrences | Input |
| NumberRealEstateLoansOrLines | Real estate-related loans or lines | Input |
| NumberOfTime60-89DaysPastDueNotWorse | 60-89 days late payment  occurrences | Input |
| NumberOfDependents | Number of dependents | Input |

**Hypotheses**

The following hypotheses were formulated to guide the analysis of the data and evaluate the significance of borrower characteristics in predicting loan default.

Null Hypothesis (H₀): There is no significant relationship between borrower characteristics (e.g., income, age, credit utilization) and the likelihood of serious loan delinquency.   
Alternative Hypothesis (H₁): There is a significant relationship between borrower characteristics and the likelihood of serious loan delinquency.   
 These hypotheses are evaluated using statistical modeling techniques implemented in SAS Enterprise Miner and are supported by recent findings that emphasize the importance of borrower behavior in predicting defaults (Addy et al., 2024; Madaan et al., 2021).

**SAS Enterprise Miner Steps – Importing the cs-training.csv file and partitioning the dataset**

The data analysis process in SAS Enterprise Miner began by importing the Excel version of the cs-training.csv file using the Import Node (see Appendix, Fig. 1). This allowed the raw dataset to be loaded into the workspace. Next, the dataset was saved in SAS-compatible format (.bdat) using the Save Data Node (see Appendix, Fig. 6) enabling it to be used throughout the project without compatibility issues. Following this, a Data Source was created (see Appendix, Fig. 10), during which the roles and levels of each variable were assigned appropriately – a crucial step for defining the metadata (see appendix, Fig. 11) that informs how SAS handles each variable in subsequent analysis. After configuring the data source, the Data Source Node (see Appendix, Fig 16) was run to generate initial descriptive statistics and validate the setup. Lastly, the data source was partitioned into two part to build the model.

**Descriptive Statistics**

Descriptive statistics will serve as a foundation for understanding the underlying patterns in the dataset. They will help identify the range, distribution, and variability in key borrower characteristics. For instance, preliminary findings showed that financial behavior variables such as DebtRatio, MonthlyIncome, and delinquency history are expected to be among the most predictive features. Reviewing the distribution of the target variable, ‘SeriousDlqin2yrs’, will also help determine whether class imbalance exists in the data. Any issues such as missing values, extreme outliers, or data skewness will be noted for resolution during data preparation and cleaning.

Descriptive statistics were generated in SAS Enterprise Miner following variable role assignments. These included the mean, median, standard deviation, and range for all input variables. This step is essential for understanding the distribution and central tendencies of the data. The analysis provided insight into variable distributions and helped identify issues such as skewness, extreme values, and missing data (Kwak & Kim, 2017).

**Partition Summary and Class Distribution**

**Table 2**

*Sample data partition results*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset | Default (1) | No Default (0) | Total | % Default |
| All Data | 10,026 | 139,974 | 150,000 | 6.68% |
| Training Set | 6,015 | 83,983 | 89,998 | 6.68% |
| Validation Set | 4,011 | 55,991 | 60,002 | 6.68% |

The dataset was successfully partitioned (see Appendix, Fig. 20) using SAS Enterprise Miner into 60% training and 40% validation sets. The class distribution for defaults was preserved across both subsets, ensuring a representative evaluation for both training and validation stages (Addy, Ajayi-Nifise, & Bello, 2024). That is the partitioned datasets maintain a consistent class distribution for the target variable SeriousDlqin2yrs, which represents whether a borrower defaulted on a loan. The target variable SeriousDlqin2yrs revealed a significant class imbalance, with only about 6.7% of borrowers marked as defaulters, which is typical in credit risk data (Burez & Van den Poel, 2009).

**Data Preparation in SAS Enterprise Miner**

**Imputation**

Impute node was applied to handle missing values using the median for interval variables such as MonthlyIncome and NumberOfDependents. These variables were later rejected due to system constraints. Median imputation was selected to reduce the influence of outliers, a common and effective practice in credit scoring models (Templ & Ulmer, 2024; Elsharif Karrar, 2022).

**Filtering and outliers' treatment**

The Filter node was used to apply user-specified limits to remove outliers. For example, NumberOfOpenCreditLinesAndLoans was capped at 18, reducing skewness from 1.18 to 0.40. Other variables such as NumberOfTime30-59DaysPastDueNotWorse were similarly filtered, improving data quality and reducing kurtosis. Detecting outliers in real-world datasets is crucial for enhancing data quality and minimizing their influence on analysis (Arimie et al., 2020, p. 3).

**Transformation**    
 Variables were discretized using optimal binning into four categories. This step simplified variable relationships and supported the modeling process by reducing noise and increasing interpretability – an established best practice in predictive analytics (McCarthy et al., 2022).

**Model 1: Logistic Regression**

The logistic regression model was built using stepwise selection based on the validation misclassification rate. Deviation coding was used for inputs, and interaction or higher-order terms were excluded. This configuration supports interpretability while focusing on key predictors. Logistic regression remains a widely adopted method for credit risk analysis due to its transparency and robustness (McCarthy et al., 2022).

**Fit Statistics**   
 The logistic regression model demonstrated strong performance based on its fit statistics (see Appendix, Fig 32). The validation Average Squared Error (ASE) was 0.0503, the misclassification rate was 0.0607, and the root ASE stood at 0.2243. These values suggest the model is well-calibrated and generalizes effectively to unseen data.

**Cumulative lift**   
 The cumulative lift chart (see Appendix, Fig. 33) indicated strong predictive ability, with a lift value exceeding 6 in the top decile. This demonstrates that the model effectively identifies borrowers most likely to default, supporting effective targeting strategies (Bekkar, Djemaa, & Alitouche, 2013).

**Iteration plot**   
 The iteration plot (see Appendix, Fig. 34) showed a steady reduction in ASE throughout model selection. The final model achieved the lowest validation ASE. This pattern confirms the utility of stepwise selection for refining predictor sets and improving generalization (Addy et al., 2024). This improves model stability and accuracy.

**Model 2: Decision Tree**

**Configuration**

The decision tree model was constructed using the Gini criterion for binary splits. The maximum tree depth was capped at 5, with a minimum leaf size of 5. Bonferroni corrections were applied to p-values to reduce the risk of overfitting (Madaan et al., 2021).

**Fit statistics**   
 The decision tree model demonstrated slightly improved performance over the logistic regression model based on its fit statistics (see Appendix, Fig. 38). It achieved a validation ASE of 0.0483, a validation misclassification rate of 0.0614, and a root ASE of 0.2197. While the differences are modest, these results indicate that the tree model offers slightly better predictive accuracy and calibration on the validation data. 

**Cumulative lift**   
 The cumulative lift chart (see Appendix, Fig. 41) mirrored the pattern observed in the logistic regression model. The lift value exceeded 6 in the top decile, indicating effective identification of high-risk borrowers.

**Leaf statistics**   
 Leaf statistics (see Appendix, Fig. 40) showed a consistent distribution of defaulters, indicating strong generalizability. Decision trees are particularly useful for understanding segmentation logic and borrower subgroups (Addy et al., 2024). This generalizability is important to ensure reliable performance when applied to new applicants.

**Variable importance**   
 Variable importance results (see Appendix, Fig. 39) identified NumberOfTimes90DaysLate, RevolvingUtilizationOfUnsecuredLines, and NumberOfTime60-89DaysPastDueNotWorse as the top predictors. These variables are closely aligned with established financial risk indicators, reinforcing the practical relevance and interpretability of the model. These results align with prior research on the predictive value of delinquency patterns (Zhu et al., 2023).

**Subtree assessment plot**   
 Lastly, the subtree assessment plot (see Appendix, Fig. 42) suggested that the optimal number of leaves was around 37. At this point, the validation ASE reached its minimum before beginning to increase again, indicating that the model achieved an effective balance between complexity and overfitting. This suggests the final tree was pruned appropriately to support robust generalization and model stability.

**Model Comparison and Discussion**

Both models demonstrated strong predictive performance with validation ASEs under 0.06. Logistic regression offered greater interpretability through coefficient estimates, while the decision tree provided superior segmentation logic and interaction detection (Addy et al., 2024; Burez & Van den Poel, 2009). According to McCarthy et al. (2022, Section 8.2), decision trees are ideal for uncovering data structure, while logistic regression remains superior for coefficient-based interpretation.   
 Model choice should be based on organizational needs. Logistic regression is ideal where transparency and regulatory justification are required. Decision trees are better suited for automated credit scoring and operational segmentation (Hlongwane, Ramaboa, & Mongwe, 2024).

**Conclusion**

This project successfully developed and evaluated two predictive models for classifying loan default risk using SAS Enterprise Miner. Data preprocessing techniques – including imputation, outlier handling, and transformation – were effectively applied to ensure model integrity. The decision tree model showed slightly better predictive performance, while logistic regression maintained strong interpretability. Both models are viable for implementation depending on institutional priorities. Decision trees offer clarity in segmentation and automation, while logistic regression supports explainability and statistical inference. These models support data-driven risk mitigation strategies in lending, aligning with current trends in credit analytics and regulatory compliance (Addy et al., 2024).

**References**

Arimie, C.O., Biu, E.O. and Ijomah, M.A. (2020) Outlier Detection and Effects on Modeling. *Open Access Library Journal*. <https://www.scirp.org/pdf/oalibj_2020091516194323.pdf>

McCarthy, R. V., McCarthy, M. M., & Ceccucci, W. (2022). *Applying Predictive Analytics: Finding Value in Data*. (2nd Edition). Prospect Press.

Addy, W. A., Ugochukwu, C. E., Oyewole, A. T., Ofodile, O. C., Adeoye, O. B., & Okoye, C. C. (2024). Predictive analytics in credit risk management for banks: A comprehensive review. *GSC Advanced Research and Reviews*, 14(3), 54–68. [Predictive analytics in credit risk management for banks: A comprehensive review](https://gsconlinepress.com/journals/gscarr/sites/default/files/GSCARR-2024-0077.pdf)

Addy, W., Ajayi-Nifise, A., & Bello, B. (2024). AI in credit scoring: A comprehensive review of models and predictive analytics. *Global Journal of Engineering and Technology Advances*, 18(2), 118–129. <https://doi.org/10.30574/gjeta.2024.18.2.0029>

Bekkar, M., Djemaa, H. K., & Alitouche, T. A. (2013). Evaluation measures for models assessment over imbalanced data sets. *Journal of Information Engineering and Applications*.  <https://core.ac.uk/download/pdf/234677037.pdf>

Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. *Expert Systems with Applications*, 36(3), 4626–4636. <https://www.sciencedirect.com/science/article/abs/pii/S0957417408002121>

Elsharif Karrar, A. (2022). The effect of using data pre-processing by imputations in handling missing values. *Indonesian Journal of Electrical Engineering and Informatics*, 10(2), 375–384. <https://www.researchgate.net/publication/361094086_The_Effect_of_Using_Data_Pre-Processing_by_Imputations_in_Handling_Missing_Values>

Hlongwane, R., Ramaboa, K. K. M., & Mongwe, W. (2024). Enhancing credit scoring accuracy with a comprehensive evaluation of alternative data. *PLOS ONE*, 19(5). <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0303566>

Kwak, S. K., & Kim, J. H. (2017). Statistical data preparation: Management of missing values and outliers. *Korean Journal of Anesthesiology*, 70(4), 407–411. <https://pmc.ncbi.nlm.nih.gov/articles/PMC5548942/>

Madaan, M., Kumar, A., Keshri, C., Jain, R., & Nagrath, P. (2021). Loan default prediction using decision trees and random forest: A comparative study. *IOP Conference Series: Materials Science and Engineering*, 1022(1), 012042. <https://www.researchgate.net/publication/348604301_Loan_default_prediction_using_decision_trees_and_random_forest_A_comparative_study>

Templ, M., & Ulmer, M. (2024). The impact of misclassifications and outliers on imputation methods. *Journal of Applied Statistics*, 51(15), 2894–2928. <https://www.tandfonline.com/doi/full/10.1080/02664763.2024.2325969#d1e149>

Zhu, X., Chu, Q., Song, X., Hu, P., & Peng, L. (2023). Explainable prediction of loan default based on machine learning models. *Data Science and Management.* <https://www.sciencedirect.com/science/article/pii/S2666764923000218>

**Appendix**

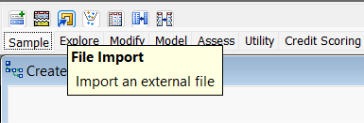
The following figures present screenshots from SAS Enterprise Miner, including node configurations, flow diagrams, and model outputs.

**SAS Enterprise Miner Steps**

**Step 1. Import an Excel version of cs-training.csv file**

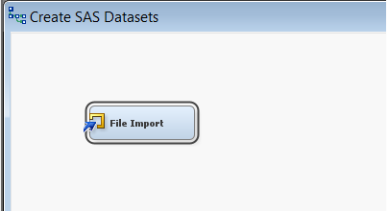
**Figure 1**

*Import node*



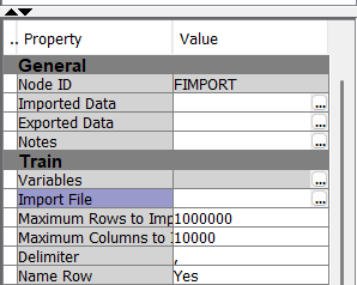
**Figure 2**

*Process flow diagram with import node*



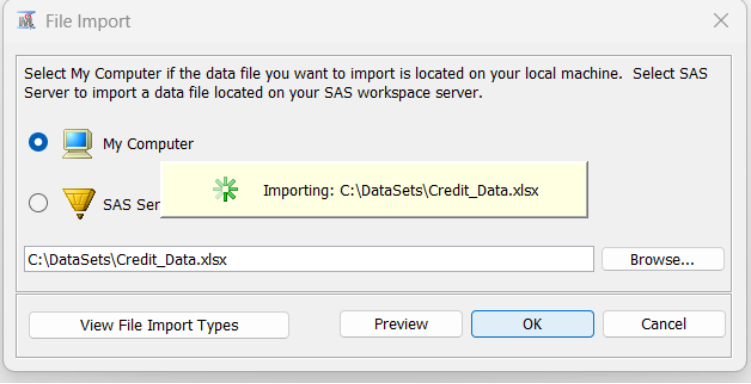
**Figure 3**

*Import node properties*



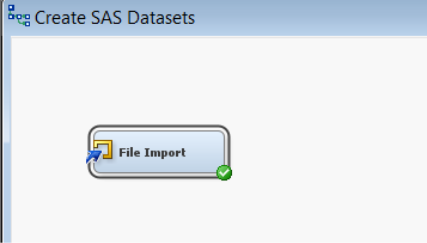
**Figure 4**

*File import window*



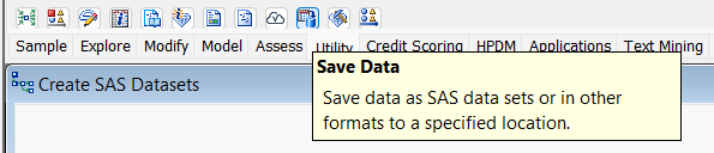
**Figure 5**

*File import*



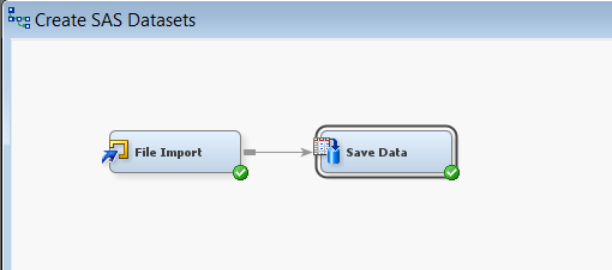
**Figure 6**

*Save data node*



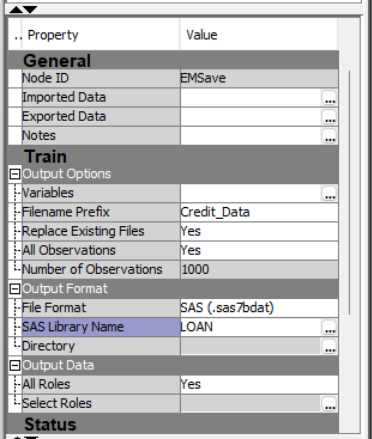
**Figure 7**

*Process flow diagram with save data node*



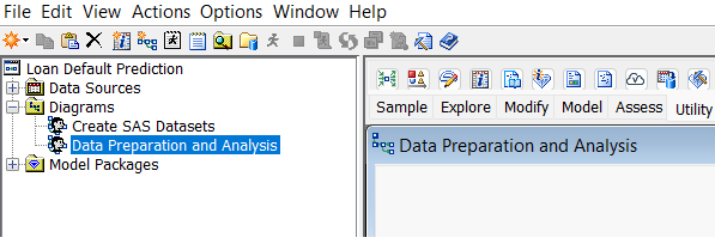
**Figure 8**

*Save data node properties*



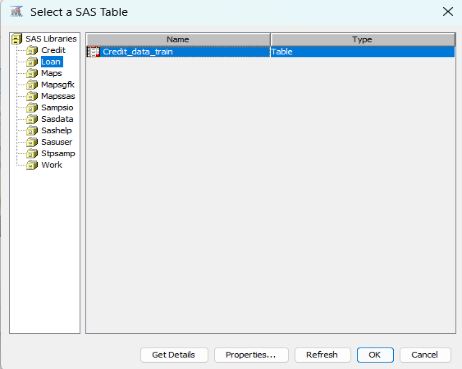
**Figure 9**

*Data preparation and analysis diagram*



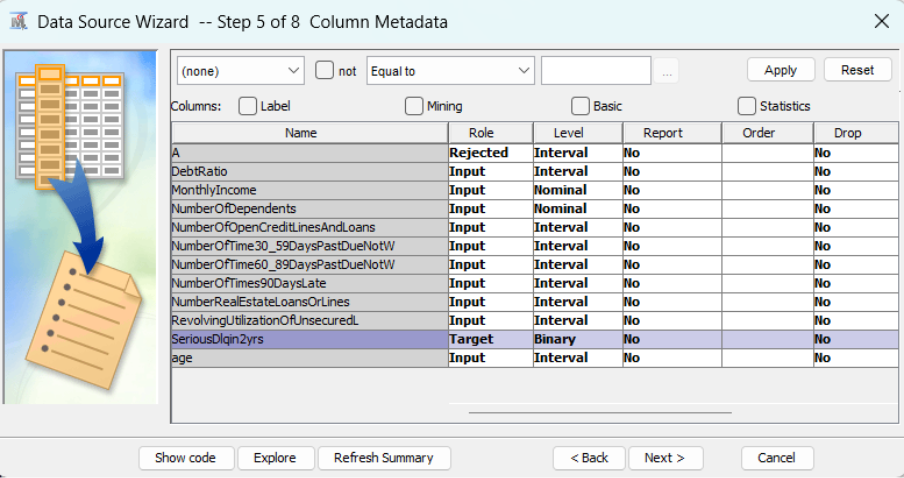
**Figure 10**

*Data source wizard – SAS table*



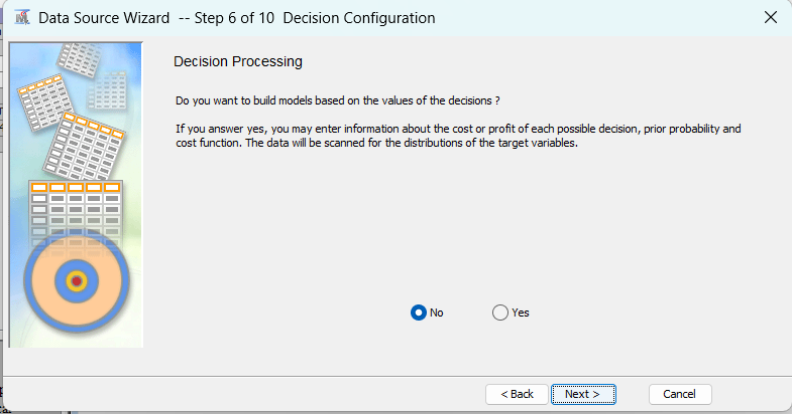
**Figure 11**

*Data source wizard – step 5 setting the metadata roles*



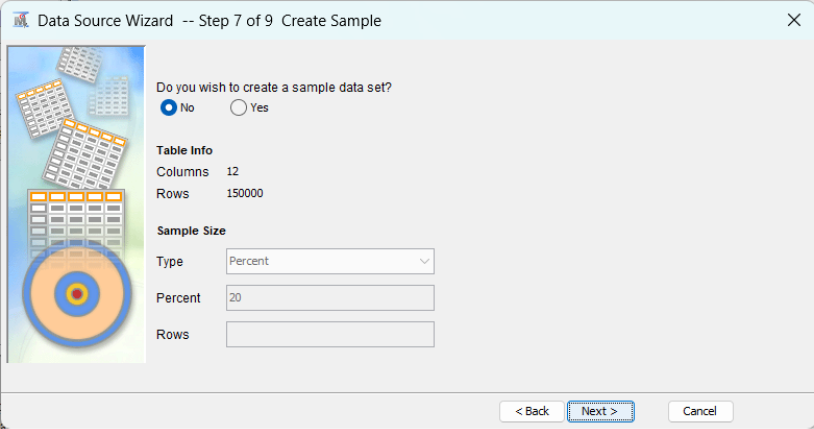
**Figure 12**

*Data source wizard – Step 6 Decision processing*



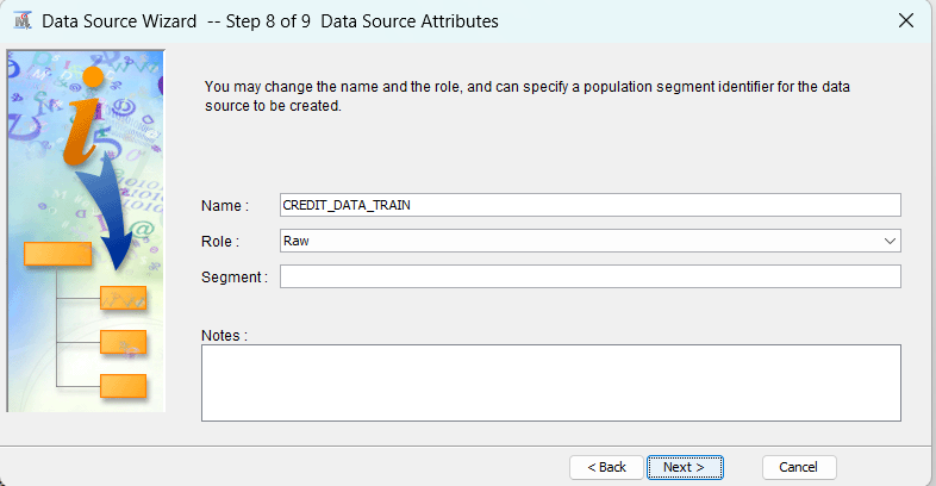
**Figure 13**

*Data source wizard – Step 7 Sample data set*



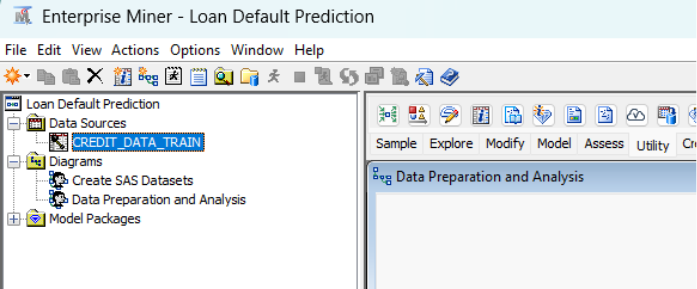
**Figure 14**

*Data source wizard – Step 8 Data source attributes*



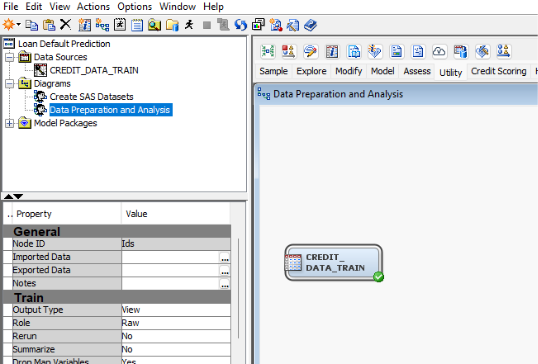
**Figure 15**

*Project panel with new data source – Credit\_Data\_Train*



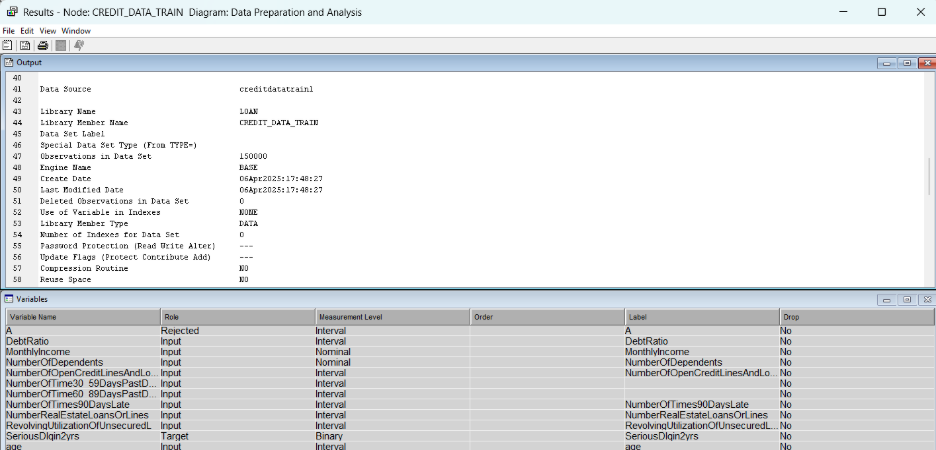
**Figure 16**

*Data source node – Credit\_Data\_Train*



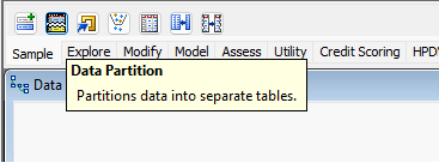
**Figure 17**

*Data source node results*



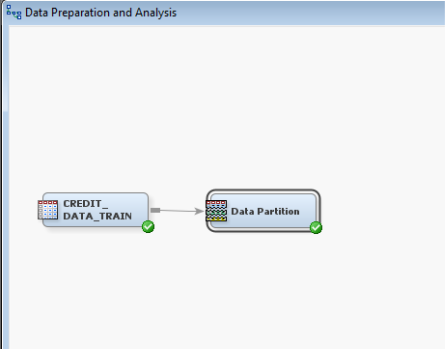
**Figure 18**

*Data partition node*



**Figure 19**

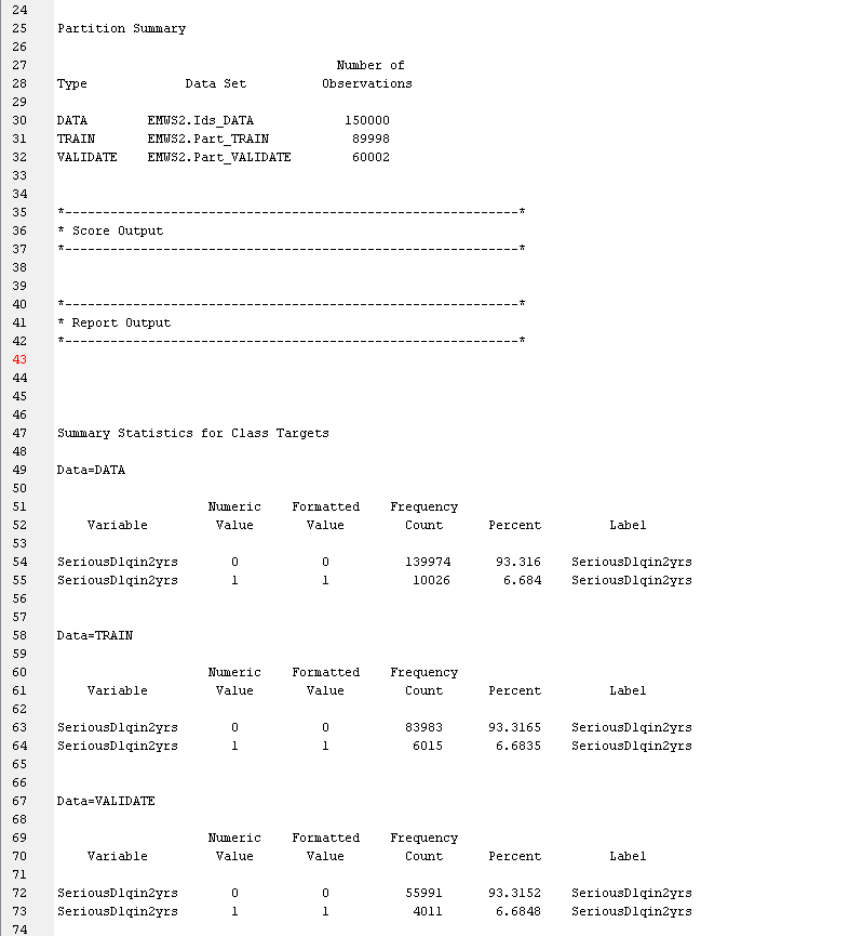
*Process flow diagram with data partition*



**Partition Summary and Class Distribution**

**Figure 20**

*Data partition results*



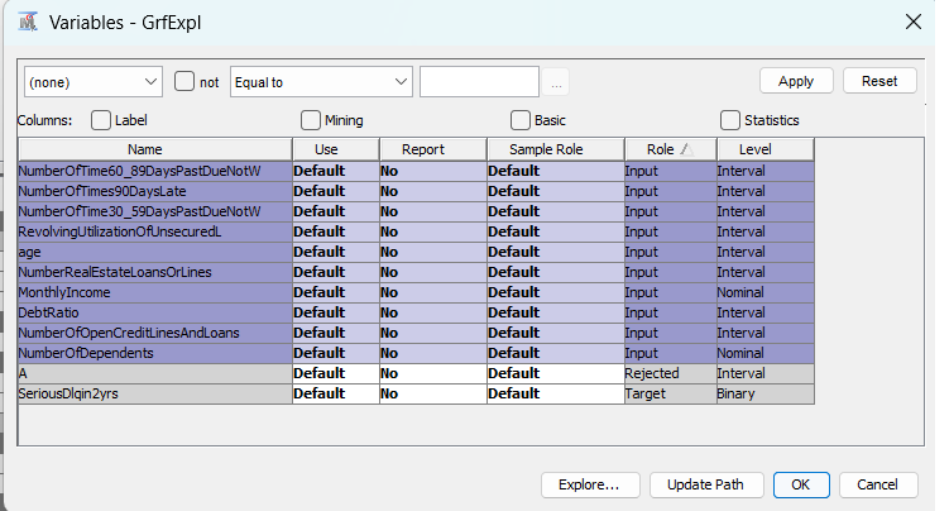
**Table 2**

*Sample data partition results*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset | Default (1) | No Default (0) | Total | % Default |
| All Data | 10,026 | 139,974 | 150,000 | 6.68% |
| Training Set | 6,015 | 83,983 | 89,998 | 6.68% |
| Validation Set | 4,011 | 55,991 | 60,002 | 6.68% |

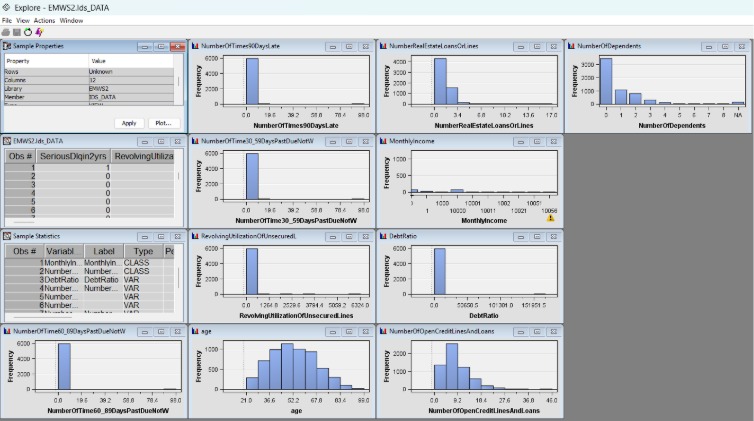
**Figure 21**

*Graph Explore – Variable selection*



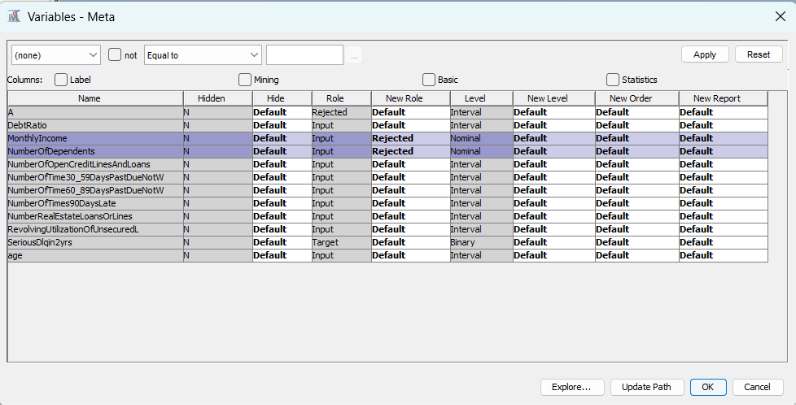
**Figure 22**

*Histograms of input variables*



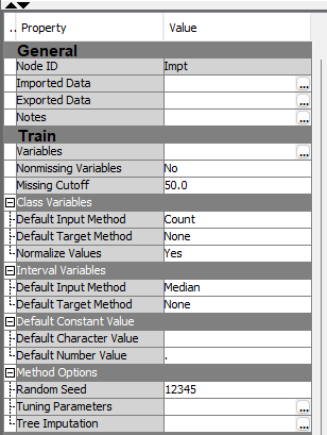
**Figure 23**

*Metadata – Rejecting two variables*



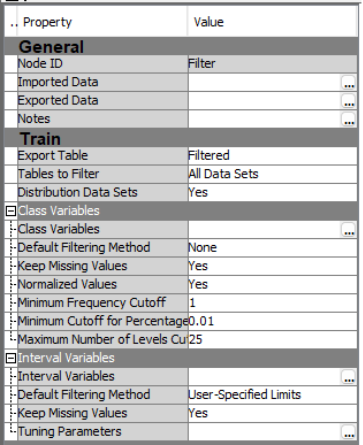
**Figure 24**

*Impute node settings*



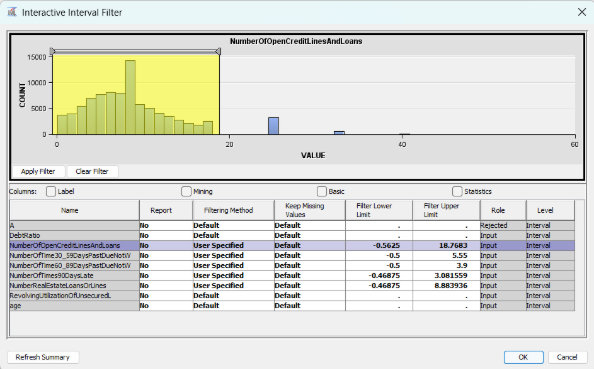
**Figure 25**

*Filter node with User-Specified Limits*



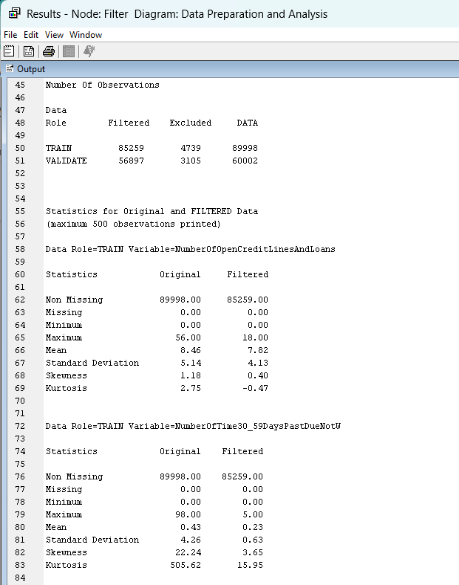
**Figure 26**

*User-Specified limits*



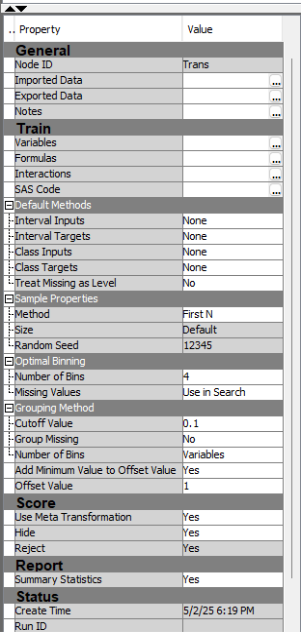
**Figure 27**

*User-specified filter results*



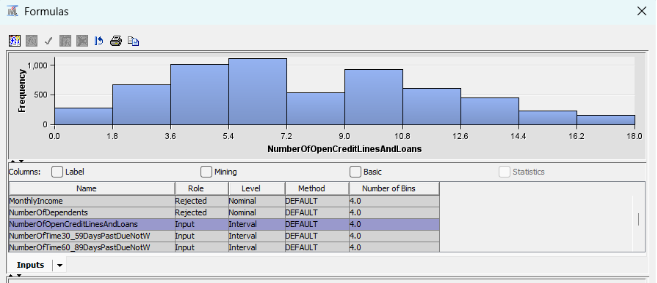
**Figure 28**

*Transform variables properties*



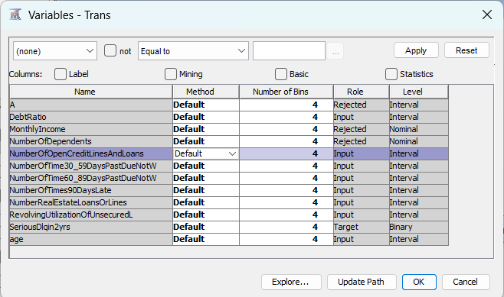
**Figure 29**

*Formulas window – Number of open credit lines and loans*



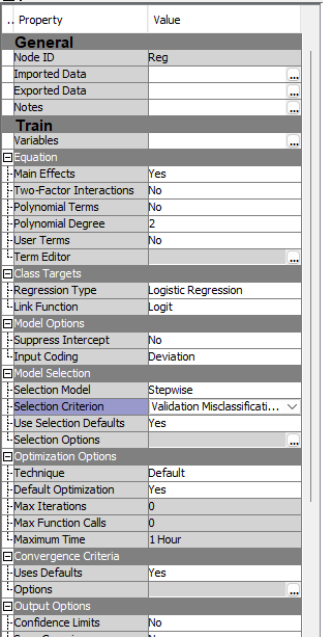
**Figure 30**

*Transform variables – Formulas window*



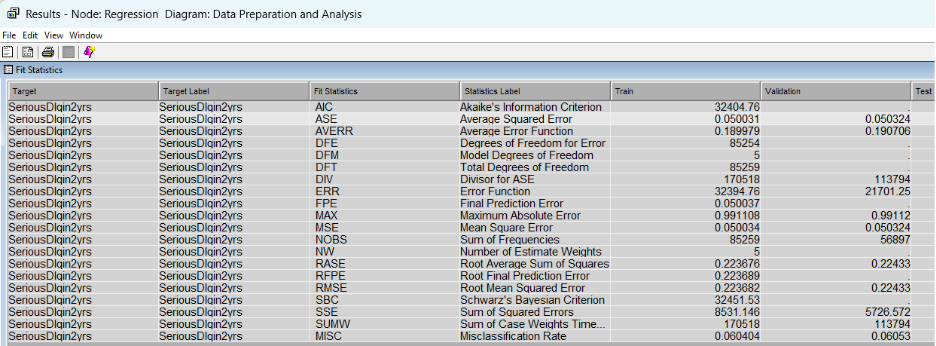
**Figure 31**

*Regression node properties*



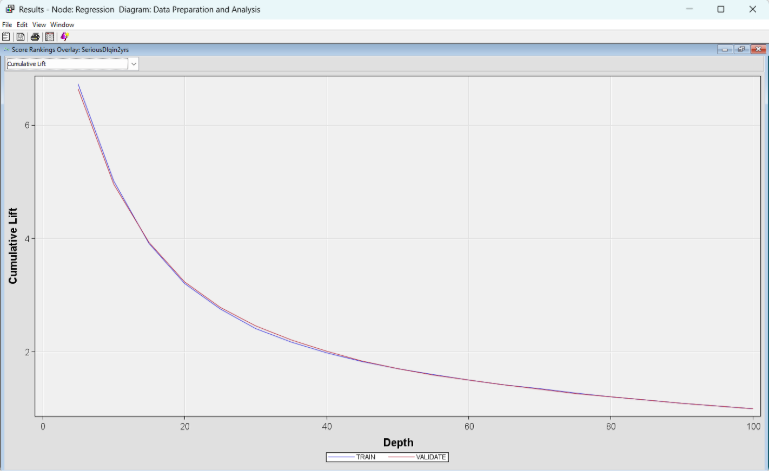
**Figure 32**

*Logistic regression – Fit statistics*



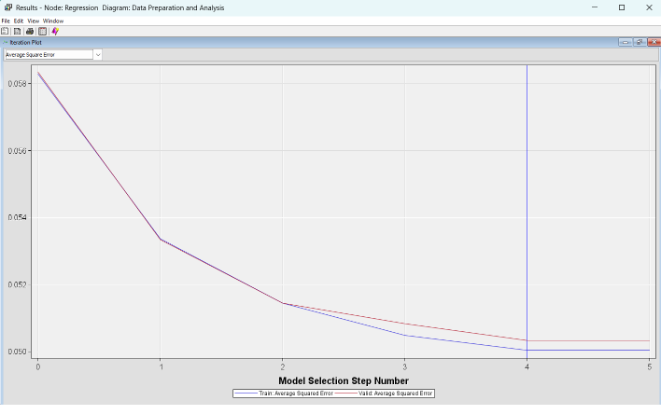
**Figure 33**

*Logistics regression – Cumulative lift*



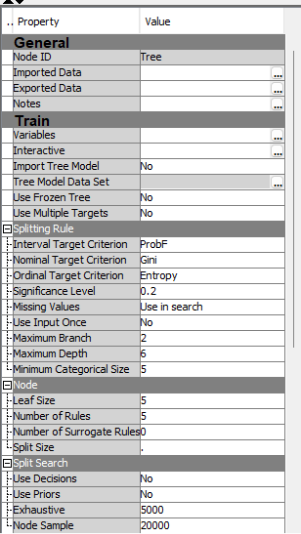
**Figure 34**

*Logistic regression – Iteration plot*



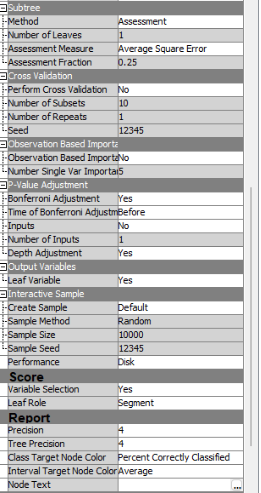
**Figure 35**

*Decision tree properties*



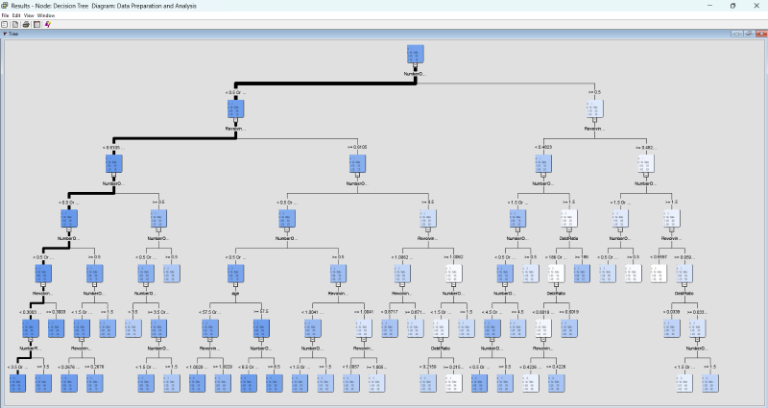
**Figure 36**

*Additional decision tree properties*



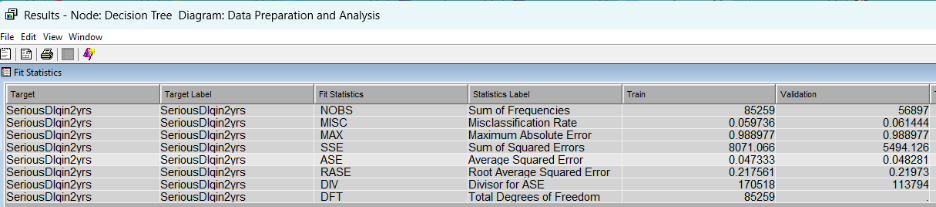
**Figure 37**

*Decision tree diagram results*



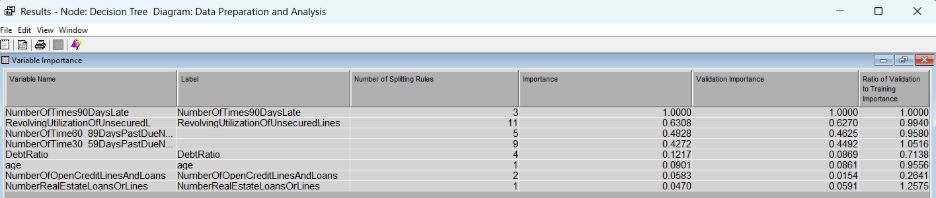
**Figure 38**

*Decision tree fit statistics*



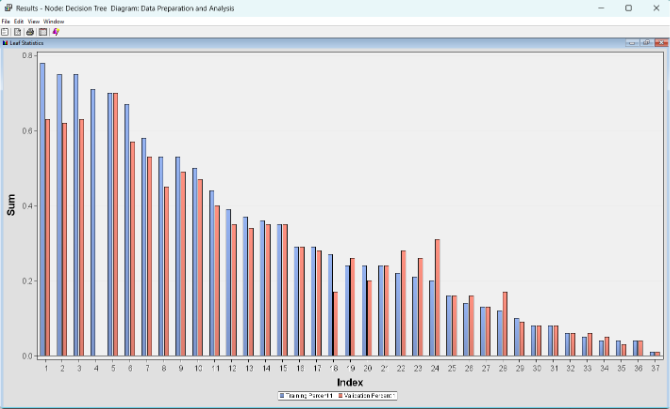
**Figure 39**

*Variable importance table*



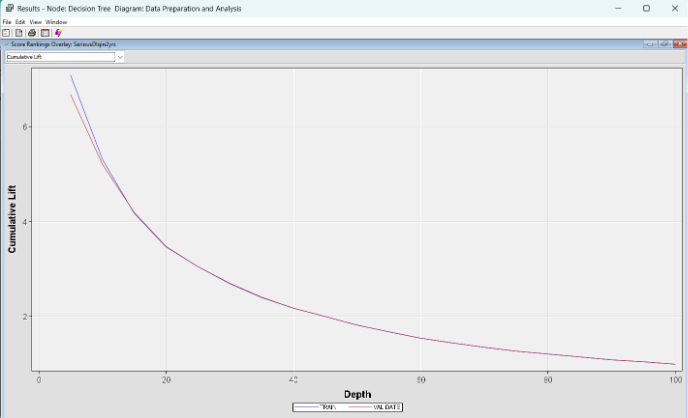
**Figure 40**

*Leaf statistics*



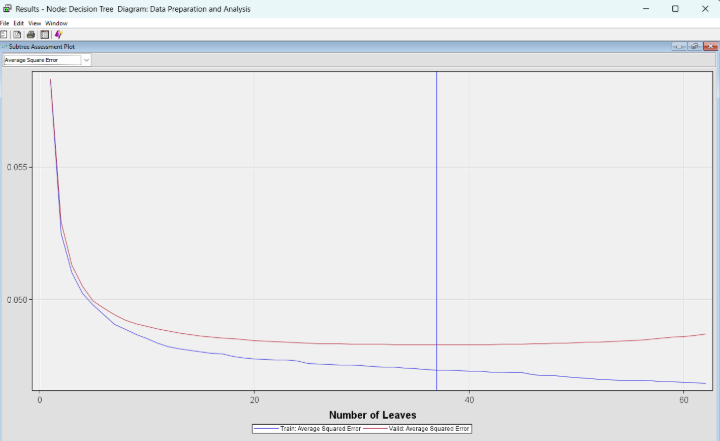
**Figure 41**

*Cumulative lift chart*



**Figure 42**

*Subtree assessment plot*



**Figure 43**

*Process flow diagram*

