**Creating Data Source**

1. **Upload CSV and Convert to SAS Table in SAS Studio**:

* In **SAS Studio**, upload your CSV file from the **Server Files and Folders** pane.
* Use the **Import Data** task to load the CSV, converting it into a SAS table. This table is now available in your default library in SAS Studio.

1. **Create a Library in SAS Enterprise Miner**:

* In SAS Enterprise Miner, go to the New option in File Tab, then select **New Library**.
* Name the library (e.g., plib) and specify the folder path where the SAS table from SAS Studio is stored.
* This library links to the data location, allowing Enterprise Miner to access the SAS table.

1. **Define the Data Source in Enterprise Miner**:

* Right-click **Data Sources** in Enterprise Miner and choose **Create Data Source**.
* In the wizard, select your new library (e.g., plib) and choose the imported SAS table.
* Complete the wizard steps to configure the dataset, making it ready for analysis in your project.

**Assessing Data**

1. **What is the distribution of credit limit across different age groups?**

**Steps:**

1. Right-click the data source and click on the Explore option.

2. In the explore window click on the Actions tab and choose the plot option.

3. Select a suitable graph, we chose boxplot.

4. Select age as x variable and Credit limit as y variable and finish the process.

5. Determine relationship and distribution of credit limit across different ages from the boxplot.

**Boxplot:**

A screen shot of a graph

Description automatically generated

**Interpretation:**

Based on the boxplot analysis, credit limits generally increase with age, peaking around the 30-40 age range, where limits are highest and most consistent. Younger clients (20s) tend to have lower credit limits, likely due to limited credit history, while older clients (50+) show more variability in credit limits, possibly due to varied financial situations. There are also notable outliers across age groups, indicating individuals with exceptionally high credit limits, especially in the younger and older demographics. Overall, credit limits appear to stabilize in middle age before becoming more varied in later years.

1. **How does the repayment status vary across different educational levels?**

**Solution**

**Steps:**

1. Right-click the data source and select the Explore option.
2. In the Explore window, go to the Actions tab and choose the Plot option.
3. Select a suitable graph; we chose a bar chart for this analysis.
4. Set “Education” as the Group variable and “Repayment Status (PAY\_0)” as the Category variable. This setup allows us to compare repayment behaviors across educational levels.
5. Run the plot and examine the chart to observe how repayment status differs by education level.

**Bar Chart:**

A screen shot of a computer

Description automatically generated

**Interpretation**

The bar chart reveals notable differences in repayment status across educational levels:

1. **On-Time Payments (PAY\_0 = 0)**:
   * Clients who paid on time (category “0”) form the largest group across all educational levels, especially for those with **University** and **Graduate School** education. This suggests that higher education levels may be associated with more consistent on-time repayment behavior.
2. **Paid in Full (PAY\_0 = -1)**:
   * A significant portion of clients with **University** and **Graduate School** education fully paid their balances, as seen in the high bars in the “-1” category. This trend indicates that higher educational levels might also be linked with financial responsibility or capacity to clear balances completely.
3. **Delayed Payments (PAY\_0 > 0)**:
   * As the number of months delayed increases (PAY\_0 = 1, 2, etc.), the bars get shorter, meaning fewer clients have extended delays.
   * In the delayed payment categories, there’s still a noticeable representation of **University** and **Graduate School** clients, but **High School** and **Other** education groups also appear more frequently, suggesting they may be at higher risk of delayed payments.
4. **Key Insight**:
   * Clients with **higher education levels** (University and Graduate School) tend to have a higher percentage of on-time and fully-paid balances, whereas clients with lower education levels (such as High School) show a higher tendency for delayed payments.

**Conclusion**

The analysis indicates that **education level has an impact on repayment behavior**, with higher education levels generally associated with more favorable repayment patterns (on-time or paid-in-full). This trend could reflect greater financial stability or awareness among clients with higher education.

1. **Give summary statistics for interval variables. What is the mean age of the people who have defaulted?**

**Steps:**

1. Connect a StatExplore Node in the diagram

2. Run the node and and click on results.

3. In the results window click on view tab select summary statistics and click on interval variables.

4. Observe the results.

**Summary Stats for interval variables:**

**A screenshot of a computer

Description automatically generated**

The mean age of people who have defaulted is 35.7.

1. **Give summary statistics for class variables. What is the count of married people who have defaulted?**

**Steps:**

1. Connect a StatExplore Node in the diagram

2. Run the node and and click on results.

3. In the results window click on view tab select summary statistics and click on class variables.

4. Observe the results.

**Summary Stats for Class Variables:**

**A screenshot of a computer

Description automatically generated**

3206 married people have defaulted.

1. **How does the distribution of defaulters look like based on sex?**

**Steps:**

1. Connect a Multiplot Node to the datasource

2. Run the node and and click on results.

3. In the results window find Sex by default.

4. Observe and analyze

A graph of frequency and frequency

Description automatically generated

The chart shows that there is a larger population of female customers (F) compared to male customers (M) in the dataset, as indicated by the taller bars for females in both default ("Y") and non-default ("N") categories. Among females, approximately 18,112 did not default, with a smaller portion defaulting. For males, about 11,868 did not default, but the proportion of defaulters is slightly higher compared to females, suggesting that males may have a marginally higher likelihood of default despite being a smaller segment overall.

This could imply that gender has a potential influence on credit risk, with males showing a relatively higher tendency to default. However, further statistical analysis would be necessary to confirm this relationship. This observation could guide further investigation into how demographic or financial characteristics interact with gender to impact default rates.

1. **How does the credit limit (LIMIT\_BAL) relate to the likelihood of default in the dataset?**

**Steps:**

1. Connect a Multiplot Node to the datasource

2. Run the node and and click on results.

3. In the results window find LIMIT\_BAL by default.

4. Observe and analyze

A graph of a number and a number

Description automatically generated with medium confidence

This chart reveals the distribution of customers across different credit limit ranges (LIMIT\_BAL) and their corresponding default status. The majority of customers have lower credit limits, as shown by the higher frequency bars on the left side of the chart, with a noticeable peak in the first few credit limit categories. Within these lower credit limits, there is a significant proportion of customers who defaulted, as indicated by the red segments at the top of each bar.

As the credit limit increases, the frequency of customers decreases, and the proportion of defaulters generally decreases as well. In the higher credit limit ranges (far right of the chart), defaulting becomes rare, with most customers not defaulting. This suggests a trend where customers with lower credit limits have a higher likelihood of defaulting compared to those with higher limits.

This pattern could indicate that lower credit limits are associated with higher financial risk, possibly because these customers are more financially constrained or have less financial stability. The chart suggests that credit limit could be a useful variable in predicting default risk.

1. **How does the chi-square plot look for this dataset?**

**Steps:**

1. Make sure the chi-square plot is enabled in the properties panel(set to yes).
2. In the results window after running the StatExplore node you should see the chi-square plot.
3. Observe and analyze.

**Chi-square plot:**

**A screenshot of a computer

Description automatically generated**

**Interpretation:**

The Chi-Square plot shows that:

* **Education** has the strongest association with default risk, suggesting higher educational levels might relate to better financial stability.
* **Gender** has a moderate impact, indicating slight differences in default patterns between males and females.
* **Marital Status** has the weakest association, implying it’s less influential in predicting defaults compared to education and gender.

In summary, education level is the most significant demographic factor in predicting defaults, followed by gender and marital status.

**Introduction to Predictive Modeling: Decision Trees**

1. **Make a decision tree based on just repayment behavior (Pay\_0 to Pay\_6)**

Steps:

1. Add a Data Partition node to the diagram and connect it to your data source.
2. Set the partition to allocate 70% for training and 30% for validation.
3. Add the Decision Tree node and connect it to the Data Partition node.
4. Set default as the target variable and PAY\_0 to PAY\_6 as input variables.
5. Run the Decision Tree node.

A diagram of a graph

Description automatically generated with medium confidence

This model effectively uses PAY\_0, PAY\_6, and PAY\_3 to classify defaulters, with a stable performance across training and validation datasets. The model achieves a reasonable lift in the top deciles but shows diminishing returns beyond a depth of 5. With a misclassification rate of 18% and consistent error metrics, the decision tree model provides an interpretable, stable baseline for identifying potential defaulters based on repayment behavior.

1. **What is the optimal depth for the decision tree to avoid overfitting?**

* **Steps**:
  + In the **Decision Tree** node properties, adjust the **Maximum Depth** parameter to values such as 3, 5, and 7.
  + Run the model each time, noting the accuracy and misclassification rate on the validation data for each depth level.

A screenshot of a computer

Description automatically generated

**Tree with max depth=7:**

**A screenshot of a computer

Description automatically generated**

As we can see even with max depth set to 7 the tree only goes to depth 3 so the optimal depth for this tree is 3.

1. **Create a decision tree using all the variables available while minimizing misclassification rate.**

* **Steps**:
  + Add a new decision tree node with default settings.
  + Set assessment measure as misclassification rate in the subtree panel of the properties panel
  + Run the node and see results.

**Decision Tree:**

A diagram of a computer

Description automatically generated

The model has a misclassification rate of 0.18.

1. **Analyze variable importance in the decision tree and how it differs from the previous decision tree with only repayment behavior as the variables.**

* **Steps**:
  + Go to the results window of the Decision Tree.
  + Maximize the output section and find variable importance section.

**A number of spitting rules

Description automatically generated**

The payment status variables (PAY\_0, PAY\_2, PAY\_5, PAY\_6) are the strongest predictors of default, with PAY\_0 being the most critical. This suggests that recent payment behavior is highly indicative of default risk. Demographic factors such as AGE and financial measures like BILL\_AMT1 have relatively low importance, indicating that payment history is a stronger predictor of default in this model than these other variables.

In conclusion, focusing on recent payment behaviors could provide the most accurate predictions for credit default in this model, as they are the primary drivers of the decision splits.

1. **How does the misclassification rate subtree assessment plot look like?**

* **Steps**:
  + Open the results window of the Decision Tree.
  + Click on the view tab.
  + Select model -> Subtree assessment plot.
  + Select misclassification rate for the graph in the upper left corner.

**Misclassification rate subtree assessment plot:**

**A screenshot of a computer

Description automatically generated**

**Interpretation:**

The ideal model complexity for this decision tree is achieved with around **3 to 5 leaves**. Beyond this range, the model’s performance on validation data does not improve and may even worsen, suggesting that a simpler model generalizes better on unseen data.

1. **How does the cumulative lift change with depth, and what does this tell us about the model's effectiveness at different ranking levels?**

* **Steps**:
  + Open the results window of the Decision Tree.
  + You will se the cumulative lift plot.

**A graph showing a line

Description automatically generated with medium confidence**

The cumulative lift represents the model's ability to identify default cases compared to random selection at each increasing depth. At depth 5, the cumulative lift is 3.50826, indicating that cases ranked in the top 5% are over 3.5 times more likely to default than cases selected randomly. As we increase the depth, the cumulative lift decreases, showing diminishing returns; by depth 100, the cumulative lift falls to 1.00000, equivalent to a random selection. This decrease in cumulative lift as depth increases suggests that while the model is highly effective at ranking high-risk cases at the top, its predictive power becomes less concentrated as more cases are included. This insight is valuable for deciding an optimal cut-off depth for targeting interventions on high-risk cases.

**Predictive Modeling: Regression**

1. **Can we predict credit limit based on demographic features using linear regression?**

* **Steps**:
  + Create a new flow for this.
  + Take the data source node again.
  + Edit variables so that credit limit is target and the rest of the variables are input.
  + Connect Data partition node.
  + Connect the regression node.
  + Edit variables so that only demographic variables like age, sex, marriage and education are the inputs.
  + Set regression type as linear regression in the properties panel.
  + Run the node and observe the results.

**Fit Statistics:**

**A screenshot of a computer

Description automatically generated**

**Conclusion:**

**No**, we cannot effectively predict credit limit (LIMIT\_BAL) based solely on demographic features using linear regression. With an **R-Square of 0.1208**, the model explains only about **12.08% of the variance** in credit limit, which is quite low. This indicates that demographic features alone are not sufficient for accurate predictions, and using this model for prediction is not preferred. Additional financial or behavioral variables would likely improve predictive accuracy.

1. **Can we use logistic regression to predict default?**

* **Steps**:
  + Connect new regression node to the data partition node of the previous flow where default was target.
  + Set regression type as logistic regression.
  + Run the node and observe results.

**Fit statistics:**

A screenshot of a computer

Description automatically generated

**Conclusion:**

The average squared error is 0.144 and misclassification rate is 0.18. So yes we can effectively predict defaulters using logistic regression.

**For the following 5 questions (13 to 18) analyze the output section in the results window of the regression node:**

1. **Which variables are the strongest predictors of default, and what does this imply about customer risk factors?**

A white paper with black text

Description automatically generated A table of numbers and letters

Description automatically generated with medium confidence

The strongest predictors of default are identified by examining the Wald Chi-Square values and odds ratios. Variables with the highest Chi-Square values, like PAY\_0 (recent payment status), are particularly impactful. For instance, PAY\_0 has a very high Chi-Square value, indicating that customers with recent payment delays are at a much higher risk of default. This implies that recent payment behavior is a crucial factor in assessing credit risk, as customers who fall behind on payments are more likely to default.

1. **How does a customer’s education level impact their likelihood of default?**

Education level has a significant impact on default risk. By examining the Chi-Square value for EDUCATION (33.0021) and the odds ratios for different education levels, it’s evident that higher education levels, such as university and graduate school, are associated with increased odds of default. For example, university graduates have an odds ratio of 3.516 compared to customers with unknown education levels. This suggests that individuals with higher education may have higher credit utilization, affecting their likelihood of default. Therefore, education level plays a role in financial behavior and credit risk.

1. **What is the relationship between marital status and default risk?**

Marital status is associated with varying levels of default risk. By examining the Chi-Square value for MARRIAGE (28.9385) and the odds ratios, we find that married individuals are 55.9% more likely to default compared to single individuals (odds ratio of 1.559). This could indicate that married customers may have additional financial commitments, impacting their ability to meet credit obligations. Thus, marital status can be an important factor in assessing credit risk.

1. **How does the credit limit (LIMIT\_BAL) affect default probability, and is this impact practically significant?**

The credit limit (LIMIT\_BAL) has a statistically significant relationship with default risk, as indicated by its Chi-Square value (16.1111). However, its odds ratio is very close to 1.000, which implies that while LIMIT\_BAL is statistically significant, its practical impact on default risk is minimal. This means that changes in credit limit alone do not substantially alter default risk, suggesting that credit limit by itself may not be a strong predictor of default when other factors are considered.

1. **What does the likelihood ratio test indicate about the overall significance of the predictors in the model?**

A close-up of a white card

Description automatically generated

The likelihood ratio test provides insights into the collective significance of the model’s predictors. With a Likelihood Ratio Chi-Square value of 2668.6973 and a p-value < 0.0001, the test strongly suggests that the included predictors are statistically significant in predicting default. This high Chi-Square value confirms that the model’s variables collectively contribute valuable insights, making the model effective for identifying default risk factors.

1. **How does the cumulative lift at each depth help in assessing the model's ability to rank high-risk cases effectively, and what can be inferred from the lift values at different depths?**

A graph with a line

Description automatically generated

The cumulative lift metric at each depth indicates how effectively the model identifies high-risk cases compared to a random selection. At depth 5, the cumulative lift is 3.27288, meaning that cases within the top 5% are over three times more likely to be identified as default cases than if selected randomly. As depth increases, the cumulative lift gradually decreases, reaching 1.00000 at depth 100. This decrease reflects that as more observations are included, the model's ability to distinctly separate high-risk cases diminishes, approaching random selection by depth 100. This information is valuable for setting a threshold depth (e.g., depth 5 or 10), where the model is most effective at distinguishing high-risk cases, allowing for more targeted intervention or decision-making.

This is similar to the cumulative lift plot of the decision tree.

1. **Which variables have the largest absolute effects on the likelihood of default, and what does this indicate about the primary factors influencing credit risk?**

**Effects plot:**

**A screenshot of a graph

Description automatically generated**

The Effects Plot shows the absolute coefficients of variables, with higher bars indicating stronger influence on the likelihood of default. The variables with the largest absolute coefficients (i.e., the tallest bars on the left) are the most impactful in the model. Typically, variables like PAY\_0 (most recent payment status) and other past payment behaviors appear to have the highest absolute effects, suggesting that recent and consistent delays in payments significantly increase the likelihood of default. The color coding indicates whether these effects are positive or negative:

* **Blue bars** (negative coefficients): Variables that reduce the risk of default, such as higher recent payments.
* **Red bars** (positive coefficients): Variables that increase the risk, such as higher delays in payments.

These insights highlight that recent payment history and payment behaviors are the primary factors influencing credit risk in this model.

**Predictive Modeling: Neural Networks**

**For the next section of questions steps are:**

* 1. **Connect neural network node after data partition.**
  2. **Analyze the results window and the output section.**

1. **What are the fit statistics for a neural network for this dataset.**

**Fit statistics:**

**A screenshot of a computer

Description automatically generated**

The neural network model has a consistent misclassification rate of around **18%** in both training and validation sets, with an average error (ASE) of **0.1357** on validation. The RMSE is **0.3684**, and **1614** cases are misclassified in validation. While stable, accuracy could still be improved.

1. **What does the Maximum Absolute Error indicate for both training and validation sets, and how might this affect the model’s usability in critical applications?**

A graph on a computer screen

Description automatically generated

The Maximum Absolute Error for both training and validation sets is **0.95**, which is relatively high. This suggests that the model's worst-case prediction can deviate significantly from the true value. In critical applications where precision is crucial, such as high-stakes financial predictions, this level of error could lead to substantial misjudgments.

1. **Examine the Root Final Prediction Error (RFPE) and Final Prediction Error (FPE) values. What do they indicate about the model’s stability?**

A graph on a white background

Description automatically generated

A graph on a white background

Description automatically generated

The RFPE for the training set is **0.37**, matching the Root Mean Squared Error (RMSE). The FPE for the training set is **0.14**. These values indicate that the model's predictive performance is relatively stable, as the RFPE is consistent with the RMSE, showing the model does not vary drastically when presented with new data.

1. **Analyze the gain and lift values at the 15th depth for both training and validation sets. What does this tell you about the model’s effectiveness in ranking high-risk cases?**

A table of numbers with numbers

Description automatically generated

At the 15th depth:

* **Training set**: Gain is **187.38** with a lift of **2.25**.
* **Validation set**: Gain is **185.30** with a lift of **2.21**.

These values indicate that the model is approximately twice as effective as random chance in identifying high-risk cases (defaults) up to this depth. This demonstrates that the model has a reasonable ability to rank and prioritize high-risk clients but could benefit from further tuning for more precision.

1. **Given that the Average Error Function (AVERR) for the validation set is 0.43, how does this error compare to the model's average squared error, and what does it imply about error distribution?**

A screen shot of a graph

Description automatically generated

The Average Error Function (AVERR) of **0.43** is much higher than the ASE of **0.14** for the validation set. This discrepancy suggests that the model has a few cases with larger errors, skewing the average error function upwards. It implies that the error distribution is not uniform, with outliers or extreme cases that the model struggles to predict accurately.

1. **What is the significance of the posterior probabilities in the score distribution at the topmost range (0.90-0.95) for training and validation, and how does this reflect on the model's confidence?**

A table of numbers and text

Description automatically generated with medium confidence

The posterior probability range of **0.90-0.95** has:

* **Training set**: 19 events with a mean posterior probability of **0.91134**.
* **Validation set**: 8 events with a mean posterior probability of **0.91088**.

These high probabilities indicate that the model is highly confident in classifying these cases as defaults. However, the small number of cases with such high confidence suggests that the model can confidently identify only a limited number of high-risk cases, which may limit its practical utility in a larger, more varied dataset.

1. **Evaluate the Total Degrees of Freedom (DFT) and Degrees of Freedom for Error (DFE) in this neural network. What can these degrees tell us about model complexity and fit?**

A screenshot of a statistics report

Description automatically generated

The **Total Degrees of Freedom (DFT)** is **20999**, and the **Degrees of Freedom for Error (DFE)** is **20905**, with a **Model Degrees of Freedom (DFM)** of **94**. This shows that the model uses 94 parameters to fit the data, which is a relatively small portion of the total degrees of freedom. This suggests that the model isn’t overly complex, which may help prevent overfitting. However, this also indicates limited capacity for capturing complex relationships, given the number of predictors involved.

1. **How does the cumulative lift at the 25th depth compare between the training and validation sets, and what does it reveal about model generalizability?**

**A graph on a screen

Description automatically generated**

At the 25th depth:

* **Training set cumulative lift**: **2.30135**
* **Validation set cumulative lift**: **2.30245**

The near-identical cumulative lift values between training and validation sets suggest strong generalizability. The model’s performance in identifying defaults remains consistent as we move deeper into the model, showing that it does not lose predictive accuracy significantly when applied to new data.

1. **How do the ASE and RMSE metrics for the validation set reflect on the model's ability to handle new data, particularly in relation to error size and stability?**

A screen shot of a graph

Description automatically generated

A graph showing a line

Description automatically generated with medium confidence

The **Average Squared Error (ASE)** for the validation set is **0.14**, and the **Root Mean Squared Error (RMSE)** is **0.37**. These values suggest that while individual predictions may deviate slightly, the error remains stable across multiple cases. The model appears to maintain consistent accuracy when applied to new data, without large fluctuations in prediction quality, demonstrating robustness in handling data not seen during training.

1. **What does the weights plot indicate about the influence of input variables on the hidden layers (H11, H12, H13) in predicting credit default?**

**Steps:**

**1. Click on the view tab in the results window.**

**2. Select model->weights\_final.**

**A screenshot of a computer

Description automatically generated**

The weights plot shows how each input variable contributes to the hidden layers (H11, H12, H13) and ultimately to the prediction of defaultY. The color coding (red for positive and blue for negative weights) and intensity indicate the strength and direction of influence for each variable on each hidden node.

1. **Strong Positive Influences**: Variables such as LIMIT\_BAL and EDUCATIONUniversity have strong red (positive) weights, especially in nodes like H12 and H13. This suggests that higher credit limits and a university education level contribute positively to activating these nodes, possibly indicating an increased risk of default associated with these factors in certain network paths.
2. **Strong Negative Influences**: Some variables, like AGE, show a strong blue (negative) weight, especially in H11. This suggests that as age increases, there’s a negative effect in certain paths within the network, potentially correlating older age with lower risk of default in these hidden layers.
3. **Balanced Contributions**: Variables like BILL\_AMT series (BILL\_AMT1 to BILL\_AMT6) have a more balanced influence across different nodes, with both red and blue weights in various nodes, suggesting that bill amounts contribute variably to different paths in the model. This variability may indicate that bill amounts interact with other factors in a complex way, affecting the model differently depending on other inputs.

Overall, this weights plot reveals that variables impact the model differently across hidden layers, with credit limit and education level showing particularly strong influences in predicting default. This indicates the neural network captures nuanced relationships in the data that may not be linear or straightforward.

**Model Assessment and Comparison**

**For the following questions (29 to 41) steps are:**

**1. Connect a model comparison node to the decision tree, regression and neural network node.**

**2. Run the node and observe the results.**

**A diagram of a company

Description automatically generated with medium confidence**

**Fit statistics in the result of the model comparison are as follows:**

**A document with numbers and text

Description automatically generated**

1. **How does the Gini Coefficient vary between models (Decision Tree, Neural Network, and Regression), and what does this tell us about model discrimination?**

The Gini Coefficient is a measure of a model's ability to distinguish between defaulters and non-defaulters. In the output, the Gini for the Neural Network is 0.56, the Decision Tree is 0.39, and the Regression model is 0.45. This suggests that the Neural Network, with the highest Gini, has the strongest discrimination ability, followed by Regression, with the Decision Tree showing the weakest discrimination. **[Ref: “Train: Gini Coefficient” under Fit Statistics Table]**

1. **What can we infer about model complexity from the degrees of freedom for each model?**

Degrees of freedom (DF) indicate model complexity; higher DF typically means a more complex model that can capture more patterns. In the output, the Neural Network has 94 degrees of freedom, Regression has 30, while Decision Tree doesn’t specify. This suggests that the Neural Network is the most complex and capable of capturing more relationships, while Regression, being simpler, may be less prone to overfitting. **[Ref: “Train: Model Degrees of Freedom” in Fit Statistics Table]**

1. **What does the Lift value tell us about each model’s ability to identify defaulters in the training set?**

A screen shot of a graph

Description automatically generated

Lift measures the model's ability to rank high-risk cases. In the output, Lift is 3.01 for the Neural Network, 2.97 for the Decision Tree, and 2.92 for Regression. The higher lift of the Neural Network suggests it is slightly better at ranking defaulters, making it potentially the best choice if capturing high-risk cases accurately is a priority. **[Ref: “Train: Lift” under Fit Statistics Table]**

1. **Why might the Neural Network be preferred over the Decision Tree based on the Misclassification Rate for validation?**

Both Neural Network and Decision Tree models show a misclassification rate of 0.18 on validation, indicating similar accuracy in distinguishing defaulters. However, the Neural Network’s higher Gini Coefficient and better discrimination power make it a more robust choice for complex scenarios where detailed patterns in data are necessary. **[Ref: “Misclassification Rate” and “Gini Coefficient” for Valid in Fit Statistics Table]**

1. **What insight can we gain from the Kolmogorov-Smirnov (KS) statistic about the models’ performance on the validation set?**

The KS statistic, which assesses the model's ability to differentiate between positive and negative classes, is highest for the Neural Network (0.42) and lower for both the Decision Tree and Regression (0.38). This implies that the Neural Network separates defaulters from non-defaulters more effectively, potentially making it the preferred model when prioritizing risk classification. **[Ref: “Valid: Kolmogorov-Smirnov Statistic” under Fit Statistics Table]**

1. **How does the Percent Captured Response compare between models, and what does this indicate about each model’s sensitivity to defaulters?**

A screen shot of a graph

Description automatically generated

Percent Captured Response reveals each model's sensitivity to detecting defaulters. For the Decision Tree, this is 14.88%, while it’s 15.11% for the Neural Network and 13.81% for Regression in validation. The higher capture rate of the Neural Network implies slightly better sensitivity in identifying defaulters, making it advantageous in risk assessment. **[Ref: “Percent Captured Response” for Valid in Fit Statistics Table]**

1. **How consistent is each model in identifying high-risk cases, as shown by model gains for both training and validation sets?**

A screen shot of a graph

Description automatically generated

Gains between training and validation are consistent across all models, with the Decision Tree showing 224.00 and 217.36, while the Neural Network has 218.46 and 217.45, respectively. This consistency, especially in the Neural Network, implies stability in identifying high-risk cases, which is beneficial for reliable performance on new data. **[Ref: “Gain” values in Fit Statistics Table for Train and Valid]**

1. **How significant is the difference in Root Mean Squared Error (RMSE) between the training and validation sets for each model?**

The RMSE is 0.37 in both training and validation for the Decision Tree and Neural Network, while it’s 0.38 for Regression. The minimal difference across all models suggests consistent error across training and validation sets, which is an indicator of strong generalization to new data. **[Ref: “Root Mean Squared Error” for Train and Valid in Fit Statistics Table]**

1. **How does the cumulative lift in the top decile compare across models in the validation set, and why is this important?**

A screen shot of a graph

Description automatically generated

In the top 10% of ranked cases, the Neural Network’s cumulative lift slightly outperforms the other models in validation, indicating that it better ranks high-risk defaulters. This is crucial when the focus is on prioritizing top cases for follow-up or intervention. **[Ref: “Cumulative Lift” for Valid in Fit Statistics Table]**

1. **What do the Kolmogorov-Smirnov probability cutoffs indicate for each model, and how might they affect cutoff selection for identifying defaulters?**

The KS probability cutoff is highest for the Neural Network at 0.30, compared to 0.26 and 0.28 for the Decision Tree and Regression. This indicates that with the Neural Network, setting a higher cutoff may better balance precision and recall, making it a more effective threshold for identifying high-risk cases. **[Ref: “Bin-Based Two-Way Kolmogorov-Smirnov Probability Cutoff” in Fit Statistics Table]**

1. **What might the Neural Network’s high Gini Coefficient indicate about its suitability for this classification task?**

The Neural Network’s Gini Coefficient of 0.56 suggests high discriminatory power, making it well-suited for classification tasks where accurately distinguishing between defaulters and non-defaulters is essential. This is especially valuable in high-stakes financial applications. **[Ref: “Gini Coefficient” for Train in Fit Statistics Table]**

1. **How does the Sum of Squared Errors (SSE) in the validation set reflect the models’ ability to generalize?**

The SSE is lowest for the Neural Network at 2443.44 in the validation set, compared to 2524.12 for the Decision Tree and 2594.11 for Regression. This lower SSE indicates the Neural Network’s ability to generalize better on new data, making it less prone to high error rates outside the training set. **[Ref: “Sum of Squared Errors” for Valid in Fit Statistics Table]**

1. **What can be inferred about model interpretability versus accuracy based on the Decision Tree’s lower Gini Coefficient and simpler structure?**

While the Decision Tree has a lower Gini Coefficient (0.39), it is often more interpretable, with clear paths for each classification. This suggests that, despite its lower discrimination power, the Decision Tree might be preferable in scenarios where interpretability is valued over raw accuracy, particularly for users who require transparency. **[Ref: “Gini Coefficient” in Fit Statistics Table]**

1. **Based on the ROC curve for each model (Decision Tree, Regression, and Neural Network) in both training and validation sets, which model shows the highest sensitivity at lower false-positive rates, and what does this imply about the model's performance?**

A graph of a function

Description automatically generated with medium confidence

The ROC curve shows that the Neural Network (green line) consistently has the highest sensitivity across both the training and validation sets compared to the Decision Tree (blue line) and Regression (red line). This is particularly noticeable at lower false-positive rates (the left side of the ROC plot), where the Neural Network achieves higher sensitivity more quickly. This indicates that the Neural Network is better at distinguishing between defaulters and non-defaulters, especially when minimizing false positives is crucial.

In terms of model performance, the higher area under the curve (AUC) for the Neural Network implies stronger discrimination power, making it the preferred model for identifying high-risk defaulters accurately. This advantage in sensitivity and overall performance suggests the Neural Network is the best choice when prioritizing accurate classification in both training and unseen (validation) data.

**Ensemble Model**

1. **Create an ensemble model from decision tree, regression and neural network and give analyze fit statistics.**

**Steps:**

1. Connect an ensemble node to decision tree, regression and neural network nodes.

2. Run the node and click on results.

3. Observe the results and analyze the fit statistics from it.

A screenshot of a computer

Description automatically generated

The ensemble model shows strong performance with consistent metrics across datasets. Key statistics include:

* **Average Squared Error (ASE):** 0.1361 (Train), 0.1366 (Validation), 0.1366 (Test)
* **Maximum Absolute Error:** 0.9329 (Train), 0.9307 (Validation), 0.9371 (Test)
* **Root Average Squared Error (RASE):** 0.3689 (Train), 0.3696 (Validation), 0.3696 (Test)
* **Misclassification Rate:** 17.8% (Train), 17.9% (Validation), 17.9% (Test)
* **Wrong Classifications:** 3738 (Train), 1612 (Validation)

These figures indicate high accuracy and consistent performance across all data splits.

1. **How does the ensemble model perform in comparison to the Decision Tree, Neural Network, and Regression models in terms of error metrics and classification accuracy?**

Based on the validation data, the ensemble model demonstrates competitive performance across several metrics:

1. **Misclassification Rate:** The ensemble model achieves a misclassification rate of **0.179**, which is equivalent to the Decision Tree (0.178) and Neural Network (0.179) models but slightly better than the Regression model (0.189). This suggests that the ensemble model is as effective at minimizing classification errors as the best-performing individual models.
2. **Average Squared Error (ASE):** The ensemble model has an ASE of **0.1366**, similar to the Neural Network’s 0.1357 and slightly better than the Decision Tree’s 0.1402 and the Regression’s 0.1441. This indicates that the ensemble model has a slight edge in terms of prediction accuracy over the Decision Tree and Regression models, although it’s comparable to the Neural Network.
3. **Kolmogorov-Smirnov (KS) Statistic:** For the validation dataset, the ensemble model’s KS statistic is **0.42**, closely matching the Neural Network’s 0.42 and higher than both the Decision Tree’s 0.38 and Regression’s 0.38. The KS statistic is a measure of the model’s ability to separate positive and negative cases, and a higher value generally indicates better discriminative power.
4. **ROC Index (AUC):** The ROC Index for the ensemble model is **0.76** on the validation dataset, which is marginally lower than the Neural Network’s 0.77 but higher than both the Decision Tree (0.70) and Regression (0.73). A higher ROC Index means that the model is better at ranking observations correctly in terms of probability of default.
5. **Classification Accuracy (Event Classification Table):** In terms of correctly classifying events, the ensemble model performs comparably with other models. It correctly classifies **6717 true negatives** and **672 true positives** in the validation dataset, which is slightly less than the Neural Network’s 6668 true negatives and 719 true positives but better than the Regression model’s 6818 true negatives and 489 true positives.

In conclusion, the ensemble model achieves a balanced performance, combining the strengths of various models. While it performs slightly below the Neural Network in some metrics, it outperforms or matches other models, particularly in metrics such as ASE, KS statistic, and ROC Index, making it a robust choice for this classification task.

1. **How does the ensemble model's performance compare to the individual models (Decision Tree, Regression, and Neural Network) in terms of ROC curve on the training and validation datasets?**

A screenshot of a computer

Description automatically generated

The ROC curves show that the Neural Network model has a slightly higher curve than the ensemble model in both the training and validation datasets, indicating it may have the best performance in terms of sensitivity and specificity. However, the ensemble model still outperforms the Decision Tree and Regression models, with its ROC curve consistently above these two models. This implies that the ensemble model provides a balanced performance by combining the strengths of different models, though the Neural Network model, by itself, appears to be the most effective at distinguishing between defaulters and non-defaulters.

1. **Which is the best-performing model among the Decision Tree, Neural Network, Regression, and Ensemble models, and which should be used?**

A screenshot of a computer

Description automatically generated

The **Neural Network model** is the best-performing model based on several key metrics, although the **Ensemble model** is also highly competitive and may be preferred in certain scenarios. Here’s a breakdown of the performance metrics to support this conclusion:

1. **ROC Index (AUC):** The Neural Network has the highest ROC Index in both the training (0.78) and validation (0.77) datasets. This metric indicates that the Neural Network is the best at ranking observations by their probability of default, which is crucial in risk prediction.
2. **Kolmogorov-Smirnov (KS) Statistic:** In the training data, the Neural Network and Ensemble models both achieve a high KS statistic (0.43 for Neural and 0.42 for Ensemble), while in the validation data, both models score similarly with the Ensemble and Neural Network models at around 0.42. This shows that both models have strong discriminative power, but the Neural Network has a slight edge in separating default vs. non-default cases.
3. **Average Squared Error (ASE):** The Neural Network has the lowest ASE (0.1357 on validation), followed closely by the Ensemble model (0.1366), indicating slightly higher prediction accuracy for the Neural Network.
4. **Misclassification Rate:** Both the Neural Network and Ensemble models have similar misclassification rates (0.179 for Neural Network and 0.178 for Ensemble) on the validation dataset, showing that they are equally effective at reducing classification errors.
5. **Stability and Robustness:** The Ensemble model, by combining multiple approaches, offers a balanced performance and may provide more robustness across different datasets. If the slight advantage in ASE and ROC Index for the Neural Network is not critical, the Ensemble model may be preferred for its stability.

**Conclusion:** The **Neural Network model** technically outperforms the others in accuracy and ranking ability, making it the best choice if pure predictive accuracy is prioritized. However, the **Ensemble model** is also a strong candidate, offering robust, balanced performance across metrics. It may be preferred in cases where stability and generalizability are essential, particularly when deploying in production settings where data variability is a concern.

**Pattern Discovery: Unsupervised Learning & Clustering**

1. **Can we perform clustering on this dataset?**

Yes, clustering can be performed on this dataset. Given the variety of numerical variables related to billing amounts (BILL\_AMT1 to BILL\_AMT6), payment amounts (PAY\_AMT1 to PAY\_AMT6), and other demographic features, clustering can help identify groups of customers with similar financial behaviors, payment patterns, or billing histories.

**Steps:**

1. Connect a clustering node to the data source

2. Run the node and click on results.

A cartoon of two people

Description automatically generated

1. **What insights do the segment plots for billing and payment amount variables provide about customer clusters?**

A screenshot of a computer

Description automatically generated  
The segment plots for billing (BILL\_AMT1 to BILL\_AMT6) and payment amount variables (PAY\_AMT1 to PAY\_AMT6) reveal distinct distribution patterns across various segments, providing insights into customer clustering based on their financial behavior.

1. **Bill Amount Variables (BILL\_AMT1 to BILL\_AMT6):**
   * The billing amount variables show considerable variation across segments, suggesting that customer billing behavior varies widely across clusters.
   * Segments such as Segment 1 and Segment 2 show a concentration of billing amounts within a specific range (indicated by a dominant color), suggesting that customers in these segments have relatively consistent billing patterns.
   * Segments with more diverse color distributions, such as Segments 8-15, indicate higher variability in billing amounts within those clusters, suggesting that customers in these segments have more fluctuation in their billing.
2. **Payment Amount Variables (PAY\_AMT1 to PAY\_AMT6):**
   * The payment amount variables also show distinctive patterns across segments. Segments like Segment 1 and Segment 2 have more consistent colors, reflecting consistent payment amounts, which may indicate regular or predictable payment behavior.
   * Conversely, segments with a wider variety of colors, such as Segment 8, show that customers in these clusters have a range of payment behaviors, reflecting variability in payment amounts across customers.

These plots reveal that some segments (e.g., Segment 1 and Segment 2) consist of customers with stable billing and payment patterns, whereas other segments exhibit more variability in both billing and payment amounts, suggesting diverse financial behaviors.

1. **How can the insights from segment distribution patterns in billing and payment behavior guide targeted strategies for customer management?**

A screenshot of a computer

Description automatically generated

The segment plots, which show distinct distribution patterns across segments, can guide targeted customer management strategies by identifying clusters with specific financial behaviors:

1. **Segment Distribution Patterns:**
   * Segments like Segment 1 and Segment 2 display consistent distributions across most variables, likely representing customers with stable billing and payment behavior. These could represent low-risk customers who may benefit from loyalty programs or personalized offers.
   * Other segments, such as Segment 10, Segment 12, and Segment 14, show more variability in their billing and payment behaviors. These segments likely contain customers with fluctuating financial patterns and may require closer monitoring or flexible payment options.
2. **Interpretation and Applications:**
   * Homogeneous segments (with concentrated colors) may represent financially stable or low-risk customers, suitable for regular credit products.
   * Diverse segments (with varied colors) may indicate customers with inconsistent financial behavior, potentially higher-risk profiles. These customers could be offered more tailored solutions, such as payment reminders or credit limit adjustments, to manage risk.
   * These insights could also support credit risk assessment and help design customized credit policies to cater to different customer financial behaviors effectively.

This segmentation approach allows for more informed, targeted strategies—such as adjusting credit limits, offering flexible repayment plans, or implementing specific monitoring protocols for high-variability segments, which can enhance customer experience and reduce financial risk.

1. **What is the segment size like and what does it tell us?**

**A pie chart with different colors

Description automatically generated**

This pie chart provides a visual representation of the relative sizes of segments in your dataset. Here’s an analysis based on the chart:

1. **Dominant Segments**:
   * Segments 1 and 2 occupy the largest portions of the chart, with Segment 1 being the largest, followed by Segment 2. This suggests that a significant portion of the data points fall into these two clusters.
   * These large segments likely represent a substantial group of customers with similar characteristics, possibly indicating the most common profiles within the dataset.
2. **Smaller Segments**:
   * Other segments, like 3, 4, 10, 12, and 17, occupy relatively small portions of the chart, indicating fewer data points in these clusters.
   * These smaller segments might represent more unique or less common customer profiles, possibly outliers or groups with distinct behaviors or characteristics.
3. **Implications for Analysis**:
   * The large size of Segments 1 and 2 suggests these segments could have an outsized impact on overall trends or behaviors observed in the dataset. This could be important if you are looking to target typical behaviors or dominant trends.
   * Smaller segments could represent niche groups that may require different strategies, especially if you are segmenting customers for targeted interventions, marketing, or risk assessment.
4. **Further Investigation**:
   * It may be beneficial to explore the characteristics of the dominant segments to understand what defines these large clusters.
   * Additionally, analyzing the unique attributes of the smaller segments could provide insights into specific customer behaviors or patterns that differ from the majority.

In summary, this pie chart indicates that the dataset has a few dominant customer profiles (Segments 1 and 2), with smaller, potentially unique profiles in other segments. This distribution can guide focused analyses or strategies based on the prevalence of these profiles.

1. **What do the eigenvalues of the covariance matrix indicate about the dimensionality of the dataset?**

A table of numbers with numbers

Description automatically generated

The eigenvalues of the covariance matrix provide insights into the variance explained by each principal component, which can guide dimensionality reduction efforts. Here, the first few eigenvalues account for most of the variance:

* The first eigenvalue (2.16188E10) explains approximately 82.43% of the total variance.
* The second eigenvalue adds an additional 10.08% to the cumulative variance, making the first two eigenvalues together account for around 92.51%.
* The first five eigenvalues cumulatively account for over 97% of the total variance.

This indicates that the majority of the dataset's variability can be captured by a small number of principal components, suggesting that the data lies within a lower-dimensional subspace. This allows for dimensionality reduction techniques like Principal Component Analysis (PCA) to effectively reduce the dataset's complexity by focusing on the top few components without significant loss of information.

1. **What can we deduce about the dataset's structure from the R-Square and Pseudo F values in the cluster history?**

A table of numbers with a number on it

Description automatically generated with medium confidence

The R-Square and Pseudo F values in the cluster history provide insight into the compactness and separation of the clusters formed at each step.

* **R-Square Values**: At the initial stages, the R-Square values are close to 1, indicating that clusters are well-separated with minimal overlap. As clustering progresses and the number of clusters decreases, the R-Square values gradually drop, suggesting that clusters are merging and variance within clusters is increasing. This decrease is natural in hierarchical clustering as similar clusters are combined to form larger ones.
* **Pseudo F Values**: Higher Pseudo F values at the initial stages signify well-defined clusters. As the number of clusters decreases, the Pseudo F value decreases, indicating less clear separation and cohesion. For instance, at 49 clusters, the Pseudo F value is extremely high (14E7), implying strong separation among initial clusters. As we move to fewer clusters (e.g., around 10 clusters), the Pseudo F value drops, indicating that the clusters are becoming less distinct.

This information suggests that the dataset has some distinct segments initially, but as clustering continues, it becomes harder to maintain clear boundaries. Therefore, the optimal cluster solution might be at an intermediate point, where both the R-Square and Pseudo F values indicate balance between compactness within clusters and separation between clusters.

1. **Which variables are most influential in predicting outcomes based on the variable importance table, and what does this imply about the dataset?**

A table of numbers with text

Description automatically generated with medium confidence

The most influential variables, as indicated by the variable importance scores, are primarily the billing amount variables such as BILL\_AMT3, BILL\_AMT1, and BILL\_AMT4, with importance scores close to 1.0. Specifically, BILL\_AMT3 has the highest importance score of 1.0, followed by BILL\_AMT1 (0.98172) and BILL\_AMT4 (0.95295). This suggests that the billing amounts at various time points are highly predictive in the model and likely have a strong relationship with the target outcome, possibly indicating financial risk or behavior.

In contrast, the payment amount variables (PAY\_AMT1 to PAY\_AMT6) have significantly lower importance scores, with values ranging from 0.09035 to 0.26995. This implies that while payment amounts contribute to the model, they are less critical than the billing amounts in predicting the target outcome.

This pattern suggests that in this dataset, historical billing information may be a more reliable indicator of the target outcome, possibly related to credit default risk, compared to the amount or frequency of payments made.

1. **What does the Cluster Proximities plot reveal about the relative positions of clusters, and how could this information be useful?**

A graph with numbers and points

Description automatically generated

The Cluster Proximities plot shows the spatial distribution of clusters based on their proximity in a reduced-dimensional space (Dimensions 1 and 2). Clusters 7 and 18 are notably distant from the main grouping of clusters, suggesting they are outliers or distinct groups with characteristics quite different from the other clusters. Most of the other clusters are grouped more tightly together near the center, indicating that they may share more similar characteristics or behaviors.

This information is useful for identifying unique segments within the dataset. For instance, clusters that are far apart (such as 7 and 18) could represent customer segments with atypical financial behaviors or demographic attributes. Understanding these outlier clusters can provide insights into specific customer needs or risks that differ from the main customer base, which could be valuable for tailored interventions or targeted marketing strategies.

1. **What does the CCC (Cubic Clustering Criterion) plot suggest about the optimal number of clusters for this dataset?**

A graph with blue squares

Description automatically generated

The CCC plot shows that the Cubic Clustering Criterion increases significantly as the number of clusters grows, peaking around 50 clusters. However, the CCC values begin to show noticeable jumps around the 35–40 cluster range, indicating that the model is identifying increasingly distinct clusters at that level. This suggests that a clustering solution with 35 to 40 clusters might balance interpretability and granularity well. Selecting a higher number of clusters, while potentially more precise, could also lead to overfitting or overly granular segments that may not be practically useful. Thus, the CCC plot suggests examining clustering solutions in the 35–40 cluster range to find a balance between complexity and interpretability.

1. **What does the input means plot reveal about the distribution and influence of different variables across the segments?**

A screen shot of a computer

Description automatically generated

The input means plot reveals how various variables are distributed across different segments of the dataset. Each colored square in the plot corresponds to a specific segment, and its position along the horizontal axis shows the normalized mean for each variable in that segment.

In this plot:

1. **Payment History Variables (PAY\_0 to PAY\_6):** These variables show diverse normalized mean values across segments, indicating that payment behavior varies significantly among different groups. For example, some segments show higher normalized means for certain payment variables, potentially representing higher delayed payment tendencies in those groups.
2. **Demographic Variables (SEX, AGE, MARRIAGE, and EDUCATION):** Demographic variables show distinct patterns. For instance, the normalized mean for SEX=F is higher in certain segments than for SEX=M, suggesting that these segments may have a greater proportion of female participants. Similarly, segments with higher means for MARRIAGE=Single or EDUCATION=University might indicate groups with higher proportions of single or university-educated individuals, respectively.
3. **Bill Amount Variables (BILL\_AMT1 to BILL\_AMT6):** The bill amount variables also show variation across segments, with some segments having higher normalized means. This may indicate that these groups have higher average billing amounts, which could correlate with higher credit limits or spending behaviors.
4. **Credit Limit (LIMIT\_BAL):** Some segments show higher normalized means for LIMIT\_BAL, suggesting these segments have participants with higher credit limits. This can be associated with different financial behaviors and possibly lower default risks in higher credit limit groups.

In summary, this plot provides a snapshot of how each segment is characterized by different averages for key variables. These differences help identify the unique characteristics of each segment, which can be useful for targeted analysis, intervention, or model building.

### Conclusion for the Project

This project thoroughly examines credit risk assessment by analyzing customer demographics, billing, and payment behaviors. By leveraging various statistical and machine learning methods, including clustering, decision trees, and neural networks, the analysis has identified critical patterns and relationships within the dataset. Segmenting customers based on financial behaviors has provided actionable insights that could inform credit policies and risk management strategies.

The project effectively uses both supervised and unsupervised learning approaches, demonstrating how different analytical techniques can complement each other to build a comprehensive understanding of default risk. The questions explored in this project cover a range of analyses, from demographic influences to detailed financial behaviors, offering a holistic view of the factors impacting credit default. Each section contributes cohesively, allowing for a step-by-step evaluation of default predictors and clustering tendencies.