**Group 6: Project Report**
**Credit Card Default Predictors**

*Table of Contents*

## 1. Project Description

### 1.1. Project Context and Goal Setting

Credit card lending is a core revenue-generating activity for financial institutions, but it also exposes banks to substantial risk. When customers fail to make timely payments, the bank incurs losses through unpaid balances, collection costs, and lower long-term customer profitability. High default rates can also weaken investor confidence, strain liquidity, and damage customer trust in the financial system.

Given these challenges, it becomes essential to accurately identify customers who are likely to default. Early detection allows banks to intervene — such as adjusting credit limits, offering payment plans, or strengthening monitoring — before defaults occur. Predictive analytics provides a powerful way to proactively manage this risk by learning patterns from historical customer behavior and using them to forecast future outcomes.

This project analyzes a real-world credit card dataset to understand the factors driving payment default and to build a predictive model that can distinguish high-risk clients from reliable borrowers.

Specifically, this project aims to:

1. Identify the key factors that influence default probability

2. Develop a predictive model that classifies customers by their risk level

3. Provide actionable insights that strengthen financial risk-management strategies

Overall, the analysis supports a data-driven approach to credit risk management, enabling the organization to make smarter lending decisions, enhance customer engagement, and reduce financial exposure.

### 1.2. Business Questions

- Which customers are most likely to default on their credit card payments next month?
  The bank wants to identify high-risk customers in advance so it can take preventive actions (e.g., outreach, limit review, payment plans).

- Does adding additional variables to the model significantly improve the accuracy of default prediction?
  The bank wants to know whether using more customer and behavior data leads to meaningfully better risk prediction compared with a simpler model.

- What customer characteristics and behaviors are the strongest indicators of future default?
  The bank wants to understand which factors (e.g., recent payment status, bill amounts, limit usage, demographics) are most strongly associated with default risk.

- Which customer segments represent the highest default risk (for example, low credit limit, high delinquency, or consistently low payments)?
  The bank wants to segment its portfolio into risk groups to design targeted credit policies and monitoring strategies.

- How should the bank prioritize customers for intervention based on their predicted probability of default?
  The bank wants a clear prioritization strategy (e.g., high/medium/low risk tiers) to decide which customers to contact first and what type of action to take.

## 2. Data Preprocessing

### 2.1. Random sampling

A random sample of 10,000 records was extracted from the original dataset of 30,000 rows to create a manageable subset for analysis.

A random number of generator function RAND() was applied to each row in a new column, values were frozen using paste-special functionality, and the dataset was sorted by these random values to ensure unbiased selection of the top 10,000 entries.
Findings: Successfully generated a representative random sample of 10,000 observations, ensuring statistical validity while reducing computational complexity for subsequent analytical procedures.

### 2.2. Handling missing data

A comprehensive examination of the dataset was conducted to identify any missing or null values across all variables:

• Microsoft Excel's pivot table and filtering functionality were systematically applied to each column within the dataset
• Each variable was individually inspected to detect the presence of blank or missing entries
Findings: The analysis revealed no missing values in the dataset, indicating complete data integrity across all observations.

### 2.3. Summary characterizes, any outliers?

Yes, the raw dataset contains several outliers, especially in credit limits, bill amounts, and payment amounts. These outliers represent real customer behavior (such as high balances, overpayments, and large credit limits), so they were not removed. Instead, numeric variables were standardized automatically by XLMiner during model estimation, which reduces the influence of extreme values and improves model stability. Classification trees naturally handle outliers through splitting and therefore require no separate treatment. No log transformation was applied because the data contains zero and negative values, making log normalization unsuitable.

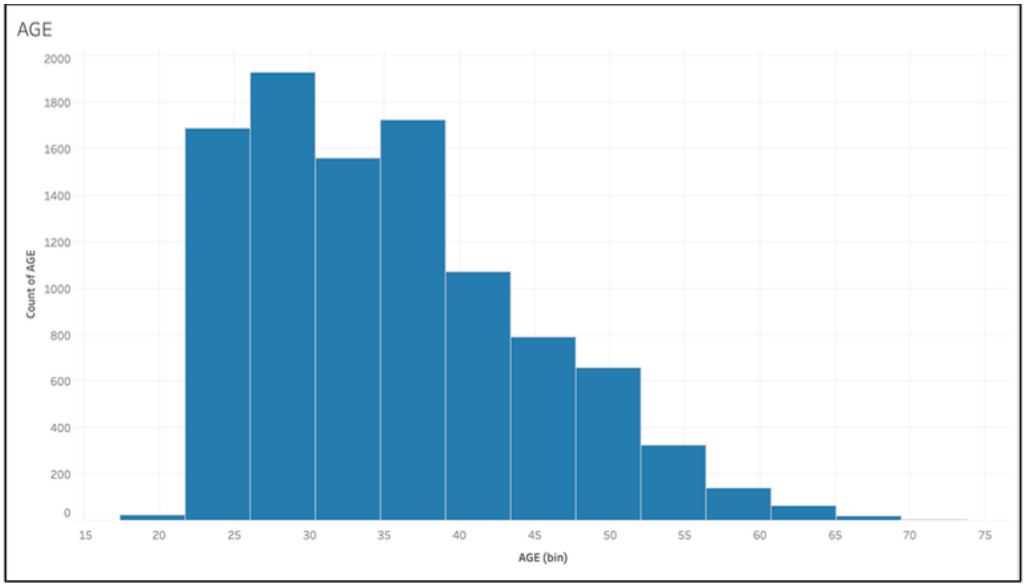
### 2.4. Correlation table

| | *AGE* | *LIMIT_BAL* | *BILL_AMT1* | *BILL_AMT2* | *PAY_AMT1* | *PAY_AMT2* | *Default Yes or No* |
|---|---|---|---|---|---|---|---|
| AGE | 1 | | | | | | |
| LIMIT_BAL | 0.149201564 | 1 | | | | | |
| BILL_AMT1 | 0.056832221 | 0.269187738 | 1 | | | | |
| BILL_AMT2 | 0.055421904 | 0.266848368 | 0.952498123 | 1 | | | |
| PAY_AMT1 | 0.025376633 | 0.206023634 | 0.160244184 | 0.307289367 | 1 | | |
| PAY_AMT2 | 0.023328216 | 0.161402711 | 0.078798587 | 0.098209692 | 0.257585984 | 1 | |
| Default Yes or No | 0.018894219 | -0.141289405 | -0.026095974 | -0.021297486 | -0.083296614 | -0.049564119 | 1 |

- Weak linear correlations are shown by the correlation analysis between the independent variables and the target variable (Default Yes or No).
- Customers with greater credit limits are marginally less likely to default, according to the variable LIMIT_BAL, which has a small negative correlation (–0.14).
- Weak negative correlations are also seen among payment-related variables (PAY_AMT1, PAY_AMT2), indicating that bigger payments are somewhat linked to a lower default risk.
- However, there is a small correlation between default and age or bill amounts (BILL_AMT1, BILL_AMT2), indicating there is no significant impact.

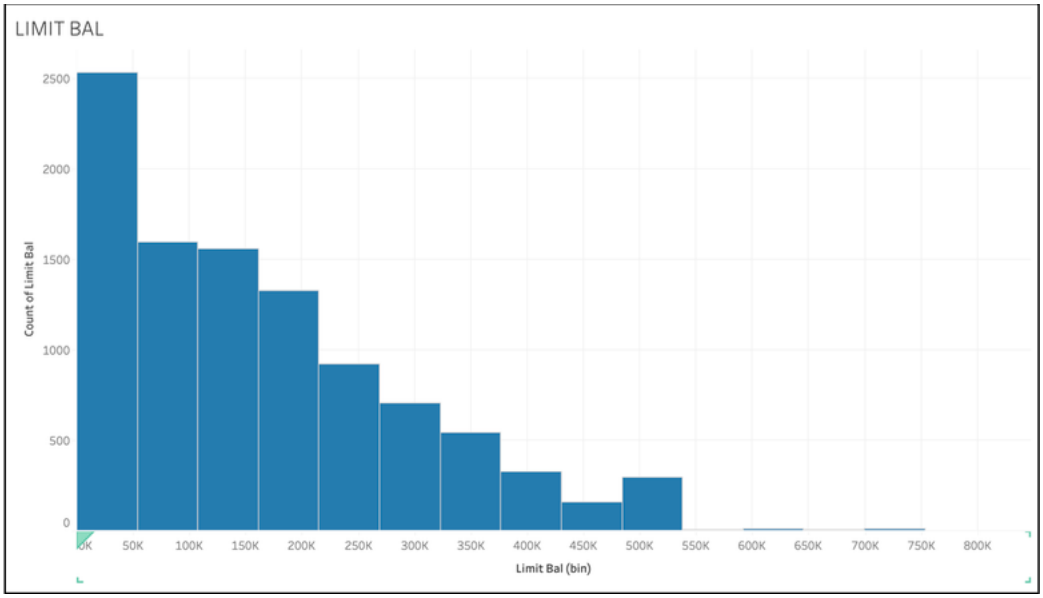
### 2.5. Histogram, scatterplot, boxplot (dependent variable vs. important independent variables) (and comments/explanations

AGE:



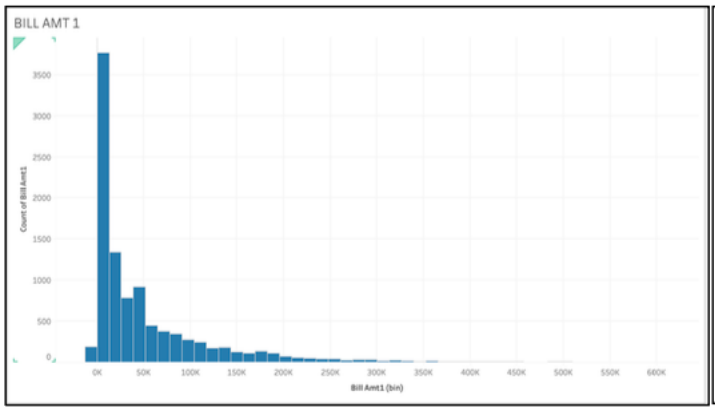

The age distribution is left-skewed (negatively skewed). Most customers are young adults aged 25–40. The number of customers declines steadily with age, and very few are above 60 years old.

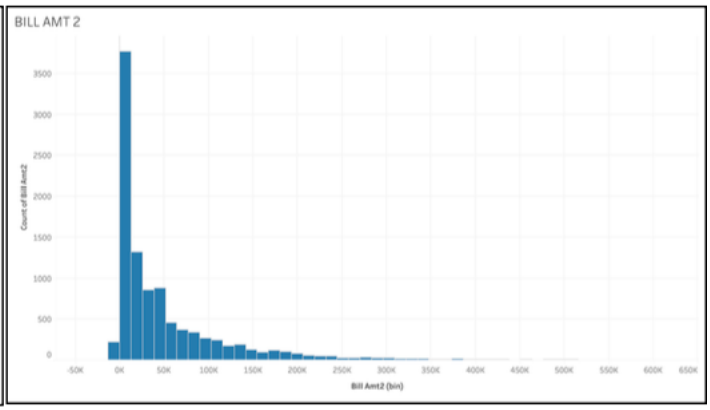
LIMIT BAL:



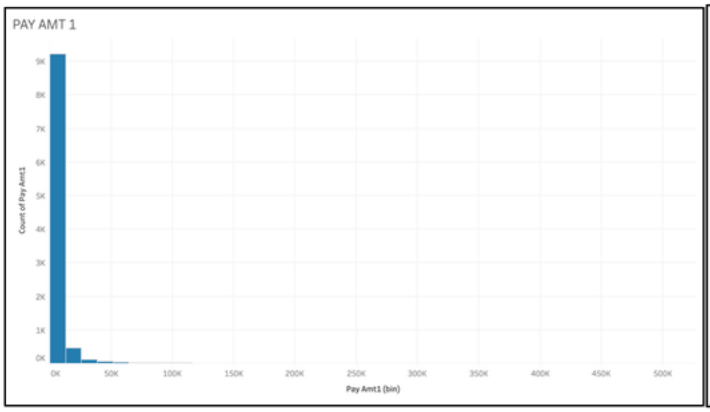

The distribution is right-skewed, with most customers having lower credit limits (below 200K), though a few customers have very high limits (above 500K).

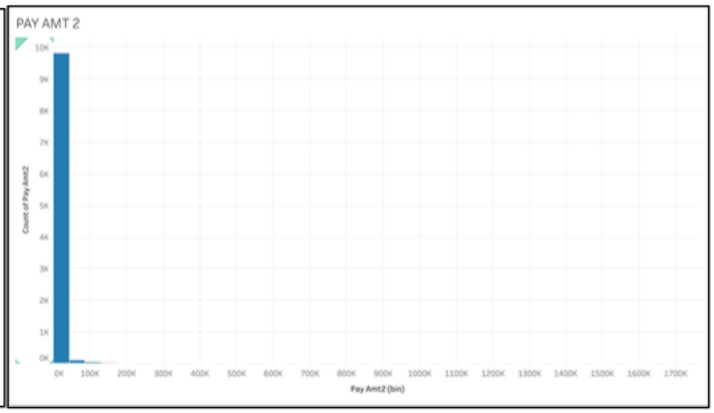
BILL AMT:




PAY AMT:




The distributions are highly right-skewed (positively skewed). Most customers have low to moderate bill amounts. There are a few high-spending customers that become the outliers for the dataset. Both distributions are extremely right skewed, with a heavy concentration at very low values (close to 0). Only a few customers make large payments exceeding 100K–200K.

## 3. Models Built for Prediction

### 3.1. Model 1: Logistic Regression

3.1.1. Identify the models used and provide a rationale for each selection.

Logistic regression model is chosen as it provides a clear, interpretable baseline against which more complex models can be compared. Since the goal of this project is to understand default risk, its transparency allows us to identify how each predictor affects the probability of default through easily interpretable coefficients, making it well suited for answering early business questions about which factors matter most. Logistic regression also supports systematic feature selection—such as evaluating coefficient significance, examining multicollinearity, or using stepwise selection—which helps refine the model by keeping only meaningful predictors before exploring more complex approaches.

3.1.2. Specify the variables included and justify their choice.

The model includes demographic variables (SEX, AGE, EDUCATION, MARRIAGE), credit limit (LIMIT_BAL), repayment status (PAY_0, PAY_2), recent billing amounts (BILL_AMT1, BILL_AMT2), and recent payment amounts (PAY_AMT1, PAY_AMT2). These variables were chosen because they represent the main drivers of credit risk: customer profile, financial capacity, repayment behavior, and debt trends. The repayment status and recent payment behavior variables are expected to be the most influential predictors based on industry practices and exploratory analysis. The ID field was excluded because it carries no predictive information.

3.1.3. Describe any variable selection techniques applied.

Feature selection was performed using a stepwise elimination method supported by correlation and exploratory analysis. Predictors with low statistical significance or redundancy were removed, leaving only variables with meaningful contribution to default prediction. While feature selection did not materially change accuracy or AUC, it resulted in a more interpretable and parsimonious model. The method performs as follows:

- Variables are added one at a time if they meet the entry significance threshold.
- After each addition, the algorithm checks all predictors currently in the model.
- Any predictor that no longer meets the stay criterion is removed.
- This add–remove cycle continues until no further variables meet the criteria to be added or removed.
- The final model contains the subset of predictors that collectively provide the best balance of statistical significance and model fit.

3.1.4. Present the model output, including equations (if applicable) and coefficient interpretations (if relevant).

| Predictor | Estimate | Confidence Interval: Lower | Confidence Interval: Upper | Odds | Standard Error | Chi2-Statistic | P-Value |
|---|---|---|---|---|---|---|---|
| Intercept | -1.480778 | -1.803061119 | -1.158495676 | 0.227461 | 0.164432982 | 81.09653088 | 2.15E-19 |
| AGE | 0.0088087 | 0.001332725 | 0.016284602 | 1.008848 | 0.003814324 | 5.333161612 | 0.020923 |
| LIMIT_BAL | -1.93E-06 | -2.62039E-06 | -1.24134E-06 | 0.999998 | 3.51806E-07 | 30.12292542 | 4.06E-08 |
| BILL_AMT1 | -7.32E-06 | -1.28545E-05 | -1.78697E-06 | 0.999993 | 2.8234E-06 | 6.723017535 | 0.009518 |
| BILL_AMT2 | 9.383E-06 | 3.64044E-06 | 1.51262E-05 | 1.000009 | 2.93009E-06 | 10.25535665 | 0.001363 |
| PAY_AMT1 | -3.7E-05 | -5.12516E-05 | -2.27807E-05 | 0.999963 | 7.26313E-06 | 25.97373212 | 3.46E-07 |
| EDUCATION | -1.358726 | -2.254418578 | -0.463032812 | 0.256988 | 0.456994562 | 8.839781127 | 0.002947 |
| PAY_0_0 | -0.493614 | -0.700915575 | -0.286311586 | 0.610417 | 0.105768267 | 21.78027803 | 3.06E-06 |
| PAY_0_1 | 0.4741174 | 0.245726562 | 0.702508146 | 1.606596 | 0.116528055 | 16.5542956 | 4.73E-05 |
| PAY_0_2 | 1.8432208 | 1.579082445 | 2.107359234 | 6.316851 | 0.134766963 | 187.0629843 | 1.39E-42 |
| PAY_0_3 | 1.781998 | 1.181111485 | 2.382884437 | 5.941716 | 0.306580366 | 33.78514044 | 6.15E-09 |
| PAY_2_2 | 0.5541117 | 0.327159781 | 0.781063699 | 1.740394 | 0.115793944 | 22.8993425 | 1.71E-06 |
| PAY_2_3 | 1.4813994 | 0.854023163 | 2.108775705 | 4.399098 | 0.320095816 | 21.41826825 | 3.69E-06 |

Logit (Default = 1) = -1.481 + 0.009*AGE + 0*LIMIT_BAL + 0*BILL_AMT1 + 0*BILL_AMT2 + 0*PAY_AMT1 - 1.359*EDUCATION - 0.494*PAY_0_0 + 0.474*PAY_0_1 + 1.843*PAY_0_2 + 1.782*PAY_0_3 + 0.554*PAY_2_2 + 1.481*PAY_2_3

Odds (Default = 1) = $e^{(-1.481 + 0.009*AGE + 0*LIMIT\_BAL + 0*BILL\_AMT1 + 0*BILL\_AMT2 + 0*PAY\_AMT1 - 1.359*EDUCATION - 0.494*PAY\_0\_0 + 0.474*PAY\_0\_1 + 1.843*PAY\_0\_2 + 1.782*PAY\_0\_3 + 0.554*PAY\_2\_2 + 1.481*PAY\_2\_3)}$

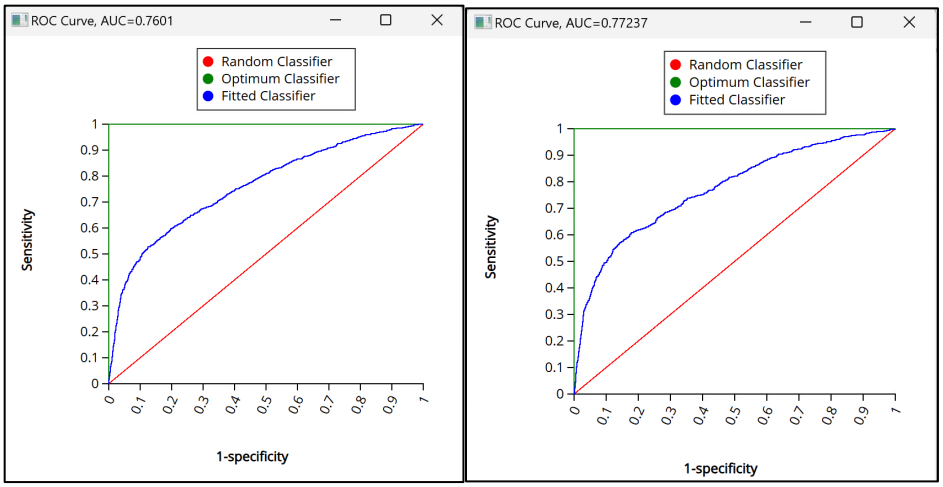
P (Default = 1) = $1 / (1 + e^{-(-1.481 + 0.009*AGE + 0*LIMIT\_BAL + 0*BILL\_AMT1 + 0*BILL\_AMT2 + 0*PAY\_AMT1 - 1.359*EDUCATION - 0.494*PAY\_0\_0 + 0.474*PAY\_0\_1 + 1.843*PAY\_0\_2 + 1.782*PAY\_0\_3 + 0.554*PAY\_2\_2 + 1.481*PAY\_2\_3)})$

Explaining the coefficients:

- Age: As the age of the customer increases by 1, the odds of the customer to default the credit card multiplies by 1.01 times holding the other variables constant.
- Limit_Bal: When the limit balance of the customer increases by 1 NT dollar, the odds of the customer to default multiplies by 0.999 times holding other variables constant.
- Bill_Amt1: When the bill amount of September 2005 increases by 1 NT dollar, the odds of the customer to default multiplies by 0.999 times holding other variables constant.
- Bill_Amt2: When the bill amount of August 2005 increases by 1 NT dollar, the odds of the customer to default multiplies by 1.000 times holding other variables constant.
- Pay_Amt1: When the amount of previous payment in September 2005 increases by 1 NT dollar, the odds of the customer to default multiplies by 0.999 times holding other variables constant.
- Education: When the education of the customer is unknown, the odds of him/her to default the credit card multiplies by 0.257 times than the customer who is in graduate school holding other variables constant.
- Pay_0_0: The odds of a customer to default the credit card with repayment status as __ in September 2005 multiplies by 0.610 times the customer who has no consumption holding other variables constant.
- Pay_0_1: The odds of a customer to default the credit card with repayment delay of 1 month as of September 2005 multiplies by 1.607 times the customer who has no consumption holding other variables constant.
- Pay_0_2: The odds of a customer to default the credit card with repayment delay of 2 months as of September 2005 multiplies by 6.317 times the customer who has no consumption holding other variables constant.
- Pay_0_3: The odds of a customer to default the credit card with repayment delay of 3 months as of September 2005 multiplies by 5.943 times the customer who has no consumption holding other variables constant.
- Pay_2_2: The odds of a customer to default the credit card with repayment delay of 2 months as of August 2005 multiplies by 1.740 times the customer who has no consumption holding other variables constant.
- Pay_2_3: The odds of a customer to default the credit card with repayment delay of 3 months as of August 2005 multiplies by 4.399 times the customer who has no consumption holding other variables constant.

3.1.5. Provide a summary report for training, validation, and test data (if applicable), along with lift charts. Assess the model's performance.

Training ROC Validation ROC




Training data: ROC Curve Data, AUC=0.7601
Validation data: ROC Curve Data, AUC=0.77237
The ROC curves and AUC values indicate that the logistic regression model performs well for credit-default prediction. With training AUC = 0.7601 and validation AUC = 0.7724, the model demonstrates strong and consistent discriminatory power without evidence of overfitting. The ROC curves show clear separation from the random-chance line, confirming meaningful predictive ability. Overall, the model provides a reliable foundation for ranking customers by risk, and its stable validation performance supports its use for further analysis and threshold optimization.

3.1.6. For classification models, determine an appropriate cutoff value based on your results. Run the model with alternative cutoff values and compare performance.

For a bank, missing defaulters (false negatives) can be more costly than mistakenly flagging safe customers. Therefore, different cutoff values were tested to find a better balance between precision and sensitivity. To improve detection of true defaulters, the cutoff must be lowered so that more customers are classified as high risk. Two cutoffs were explored:

- Considering top 3 deciles, the top post probability of class 1 came as 0.18
- Considering top 2 deciles, the top post probability of class 1 came as 0.26

The recommended cutoff value is 0.26, as it delivers the most balanced performance across accuracy, precision, sensitivity, and F1 score. This threshold is optimal for identifying at-risk customers while minimizing false positives, making it suitable for operational credit-risk decisions.

| Model | Description | Training AUC | Validation AUC | Validation Accuracy | Validation Precision | Validation Sensitivity | Validation F1 Score |
|---|---|---|---|---|---|---|---|
| Model 1 | Without FS | 0.7641 | 0.7744 | 82.395 % | 72.592 % | 33.145 % | 0.4551 |
| Model 2 | With FS | 0.7601 | 0.7724 | 82.296 % | 71.882 % | 33.145 % | 0.4537 |
| Model 3 | Cut off 0.18 | 0.7601 | 0.7724 | 74.968 % | 45.327 % | 62.344 % | 0.5249 |
| Model 4 | Cut off 0.26 | 0.7601 | 0.7724 | 80.695 % | 57.196 % | 51.521 % | 0.5421 |

Model 4 (cutoff = 0.26) is selected as the final model because it delivers the best balance between precision and sensitivity, resulting in the highest F1 score (0.5421). This means the model is more effective at identifying defaulters while still maintaining reasonable accuracy and minimizing false positives. Models 1 and 2 have higher accuracy but very low sensitivity, leading to many missed defaulters. Model 3 has high sensitivity but too many false positives. Therefore, Model 4 represents the best overall performance for the organization.

3.1.7. Explain how the results address your business questions.

The results from the logistic regression model, ROC analysis, and cutoff experimentation directly address the business questions by identifying which customers are most likely to default, determining that additional variables do not materially improve performance, highlighting repayment behavior as the strongest predictor of default, defining high-risk customer segments, and providing a clear cutoff-based strategy for prioritizing intervention. The recommended cutoff of 0.26 achieves the best balance between precision and sensitivity, enabling the bank to intervene efficiently and reduce potential losses.

1. Which customers are most likely to default next month?

The model produces a predicted default probability for every customer. Using the optimized cutoff of 0.26, customers whose predicted probability is ≥ 0.26 are flagged as high-risk.

- This cutoff provides the best balance between sensitivity (51.52%) and precision (57.20%).
- It allows the bank to reliably identify a meaningful portion of customers who are likely to default, while avoiding too many false alarms.

Customers with a probability score ≥ 0.26 are the most likely to default and should be prioritized for intervention.

2. Does adding additional variables significantly improve prediction accuracy?

Models with and without feature selection showed:

- Nearly identical AUC (0.7744 vs. 0.7724)
- Almost the same accuracy (82.39% vs. 82.30%)
- Same sensitivity (33.14%)

This indicates that:

- Feature selection did not materially improve performance
- The reduced model is simply more interpretable, not more accurate

Adding or removing a small number of predictors does not significantly impact predictive accuracy but does help simplify the model.

3. What customer characteristics and behaviors are the strongest indicators of default?

The strongest indicators of default are the customer's recent repayment behavior, especially the variables PAY_0 and PAY_2. Customers who were 1, 2, or 3 months late on recent payments have the highest likelihood of default, as reflected by the largest positive coefficients in the model. Other important predictors include high outstanding bill amounts (BILL_AMT1 and BILL_AMT2), low recent payments (PAY_AMT1), and lower credit limits (LIMIT_BAL), which together signal increasing financial stress. Demographic factors such as age and education have much smaller effects. Overall, repayment history is the dominant factor in predicting future defaults.

4. Which segments of customers represent the highest risk?

Using the cutoff analysis and predicted probabilities:

High-risk customer segment:

- Probability ≥ 0.26
- High late-payment statuses (PAY_0 ≥ 1)
- Consistently low repayment amounts
- High bill amounts relative to credit limit
- Lower-limit customers with rising balances

These customers have the highest predicted likelihood of default.

Medium-risk segment:

- Probability between 0.18 and 0.26
- Indicators of emerging financial stress (slow payment, rising bill amounts)

Low-risk segment:

- Probability < 0.18
- On-time repayment and stable balances

The model successfully separates customers into meaningful, actionable risk tiers.

5. How should the bank prioritize customers for intervention?

The threshold analysis demonstrated:

- Cutoff 0.18 → very high sensitivity (62%) but too many false positives
- Cutoff 0.26 → best balance (highest F1 score)

Because the bank must balance:

- catching defaulters (recall), and
- avoiding unnecessary risk actions (precision)

Cutoff = 0.26 is the appropriate threshold.

Intervention Strategy Based on Cutoff 0.26:

| Probability Segment | Action |
|---|---|
| ≥ 0.26 (High Risk) | Immediate outreach, limit freeze/review, financial counseling |
| 0.18 – 0.26 (Medium Risk) | Monitor behavior, send reminders, offer payment plans |
| < 0.18 (Low Risk) | Routine servicing, no intervention needed |

3.1.8. Offer a hypothetical example of a new data record and demonstrate prediction or classification

To demonstrate how the logistic regression model classifies a new customer, we create a hypothetical data record with realistic values from the UCI Credit Card dataset. The model produces a predicted probability of default, and this probability is compared against the optimized cutoff of 0.26 to determine the final classification.

Hypothetical New Customer Record:

| Variable | Value |
|---|---|
| LIMIT_BAL | 80,000 |
| EDUCATION | 2 (University) |
| AGE | 34 |
| PAY_0 | 1 (1 month delay) |
| PAY_2 | 2 (2 months delay) |
| BILL_AMT1 | 25,000 |
| BILL_AMT2 | 22,000 |
| PAY_AMT1 | 2,500 |

Model Output: Predicted Probability Based on this customer's characteristics, the logistic regression model might output a probability such as-

Logit: -0.377, Odds: 0.686
Predicted probability of default: 0.665 (66.5%)

This probability reflects:

- Payment delays,
- Moderate credit limit,
- High recent bill amounts relative to payments.

Classification Using Cutoff = 0.26. We classify customers as follows:

- Predicted probability ≥ 0.26 → Default (High Risk)
- Predicted probability < 0.26 → Non-Default (Low/Medium Risk)

Probability = 0.665

Predicted Class: 1 (Likely to Default) This customer falls into the high-risk segment and should be prioritized for intervention.

### 3.2. Model 2: Classification Tree

3.2.1. Identify the models used and provide a rationale for each selection.

For this part of the analysis, a Classification Tree (CART) in XLMiner is used to predict whether a customer will default next month (Yes/No) as:

- The target variable is binary.
- Trees naturally handle both categorical and numeric variables.
- No dummy variables or scaling are required.
- The output consists of simple, interpretable decision rules, which are ideal for credit risk policy development.
- XLMiner provides automated pruning to avoid overfitting.

Thus, a classification tree offers the best trade-off between interpretability and predictive usefulness for this project.

3.2.2. Specify the variables included and justify their choice.

The classification tree model was developed using a combination of demographic attributes and financial behavior indicators that are commonly associated with credit risk. The predictor variables included in the model were: SEX, EDUCATION, MARRIAGE, PAY_0, PAY_2, AGE, LIMIT_BAL, BILL_AMT1, BILL_AMT2, PAY_AMT1, and PAY_AMT2. The target variable for prediction was Default (Yes/No). All variables included in the model are readily available operationally to a lender and represent real features that credit card companies routinely monitor. Their inclusion ensures that the tree model aligns with practical credit risk decision-making and captures both behavioral and capacity-driven risk factors. Overall, these predictor variables collectively provide a strong and realistic foundation for predicting default risk.

3.2.3. Describe any variable selection techniques applied.

No manual or pre-processing variable selection techniques were applied prior to model development. This approach is appropriate because decision trees naturally perform variable screening as part of the splitting and pruning process.

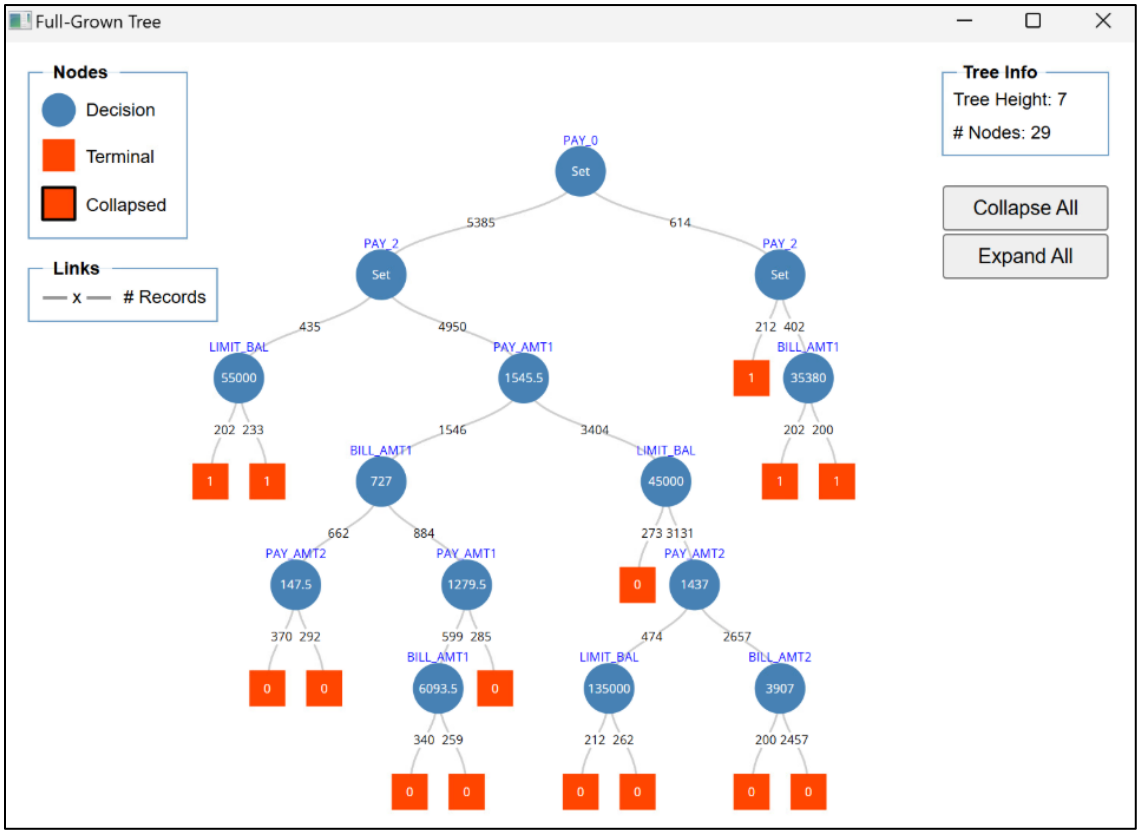
3.2.4. Present the model output, including equations (if applicable) and coefficient interpretations (if relevant). For tree-based or neural network models, explore various parameters and select the best-performing model.

To evaluate how model complexity affects performance, the classification tree was trained under four configurations by varying the minimum number of records required in terminal nodes (not specified, 600, 400, and 200). Each configuration produced a full-grown tree and a pruned tree based on validation performance. The full-grown trees differed substantially in size and depth: the unconstrained model grew into a highly complex tree with 83 nodes, while the 600- and 400-record settings produced much smaller structures, and the 200-record model allowed moderate complexity with 29 nodes. These differences illustrate how the terminal-node constraint controls tree growth and prevents excessive fragmentation of the training data.

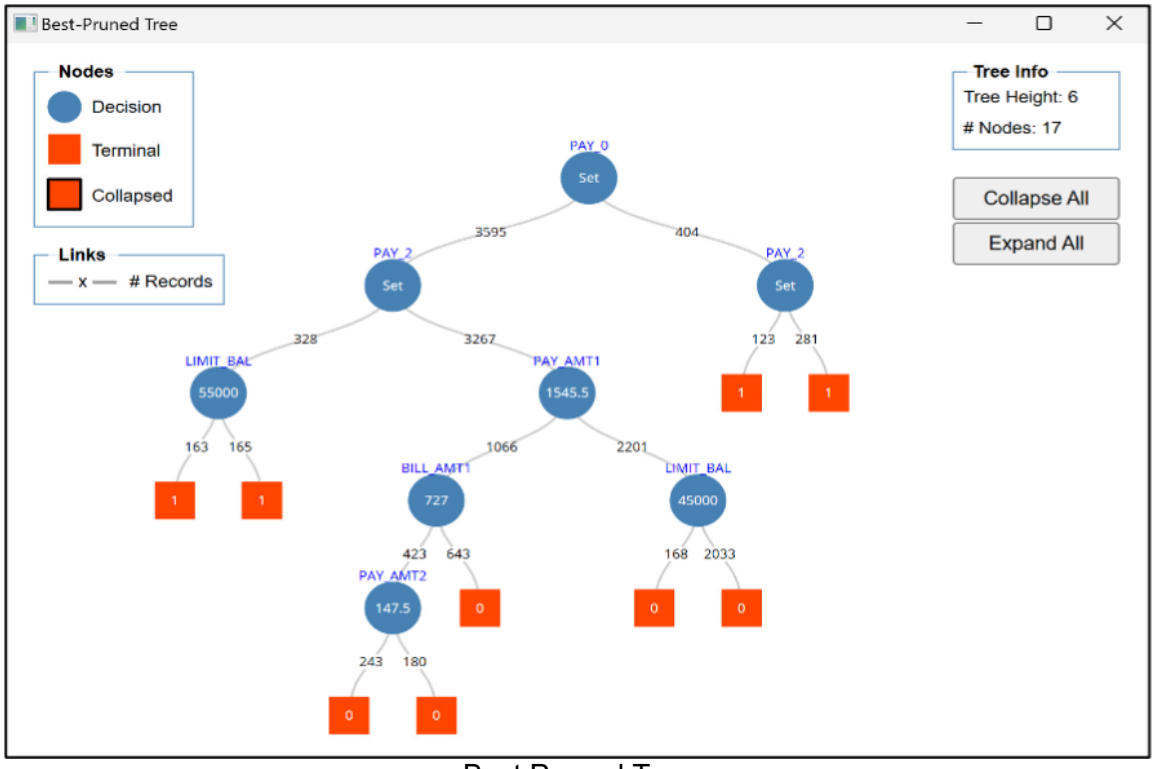
Across the first three configurations (no minimum, 600, and 400), the pruning procedure consistently collapsed the full-grown trees into an extremely simple structure containing only a single split on PAY_0. This indicates that while the full-grown trees included many additional variables and splits, none of those deeper patterns generalized well to the validation set. In contrast, the 200-record configuration retained a substantially larger and more informative best-pruned tree (17 nodes), maintaining meaningful splits on PAY_0, PAY_2, PAY_AMT1, BILL_AMT1, LIMIT_BAL, and PAY_AMT2. This model preserved real signal that was lost when pruning aggressively removed branches in the more constrained settings.

While accuracy, precision, and specificity remained nearly identical across all models (~82.6 percent accuracy and ~0.966 specificity), the discriminatory ability varied substantially. The overly simplified one-split pruned trees achieved an AUC of approximately 0.65, reflecting limited predictive power. In contrast, the best-pruned model from the 200-record configuration achieved a significantly stronger AUC of 0.76 and a top-decile lift of approximately 3.3, demonstrating the ability to meaningfully rank-order customers by default risk.

Based on these results, the best-pruned tree derived from the 200-record terminal node setting was selected as the final model. It provides a balanced level of complexity, incorporates key behavioral drivers of default, and delivers the strongest validation performance. This configuration avoided both extremes—overfitting from overly complex full-grown trees and underfitting from aggressively pruned single-split models—making it the most reliable and interpretable decision tree for predicting credit card default.



Full Grown Tree



Best Pruned Tree

Set of rules for the pruned tree:

- When PAY_0 <= 0, PAY_2<=0, LIMIT_BAL<55000 then the customer will not default the credit card.
- When PAY_0 <= 0, PAY_2<=0, LIMIT_BAL>55000 then the customer will not default the credit card.
- When PAY_0 <= 0, PAY_2>0, PAY_AMT1 < 1545.5, BILL_AMT1 < 727, PAY_AMT2 < 147.5 then the customer will not default the credit card.
- When PAY_0 <= 0, PAY_2>0, PAY_AMT1 < 1545.5, BILL_AMT1 < 727, PAY_AMT2 > 147.5 then the customer will not default the credit card.
- When PAY_0 <= 0, PAY_2>0, PAY_AMT1 < 1545.5, BILL_AMT1 > 727 then the customer will not default the credit card.
- When PAY_0 <= 0, PAY_2>0, PAY_AMT1 > 1545.5, LIMIT_BAL < 45000 then the customer will not default the credit card.
- When PAY_0 <= 0, PAY_2 > 0, PAY_AMT1 > 1545.5, LIMIT_BAL > 45000 then the customer will not default the credit card.
- When PAY_0 > 0, PAY_2 <=0 then the customer will default the credit card.
- When PAY_0 > 0, PAY_2 > 0 then the customer will default the credit card.

3.2.5. Provide a summary report for training, validation, and test data (if applicable), along with lift charts. Assess the model's performance.

The classification tree provides a clear and interpretable framework for predicting credit card default, which directly supports the business goal of identifying high-risk customers before payment issues occur. Across all configurations, the tree achieves stable accuracy of around 82.6 percent on both training and validation sets, which indicates that the model generalizes well. Specificity is consistently high at around 0.96, meaning the model is highly reliable at recognizing customers who are unlikely to default. This is important for avoiding unnecessary interventions with low-risk clients.

| Metrics - Training | |
|---|---|
| Metric | Value |
| Accuracy (#correct) | 4959 |
| Accuracy (%correct) | 82.6637773 |
| Specificity | 0.960177928 |
| Sensitivity (Recall) | 0.333333333 |
| Precision | 0.693811075 |
| F1 score | 0.450317125 |
| Success Class | 1 |
| Success Probability | 0.5 |

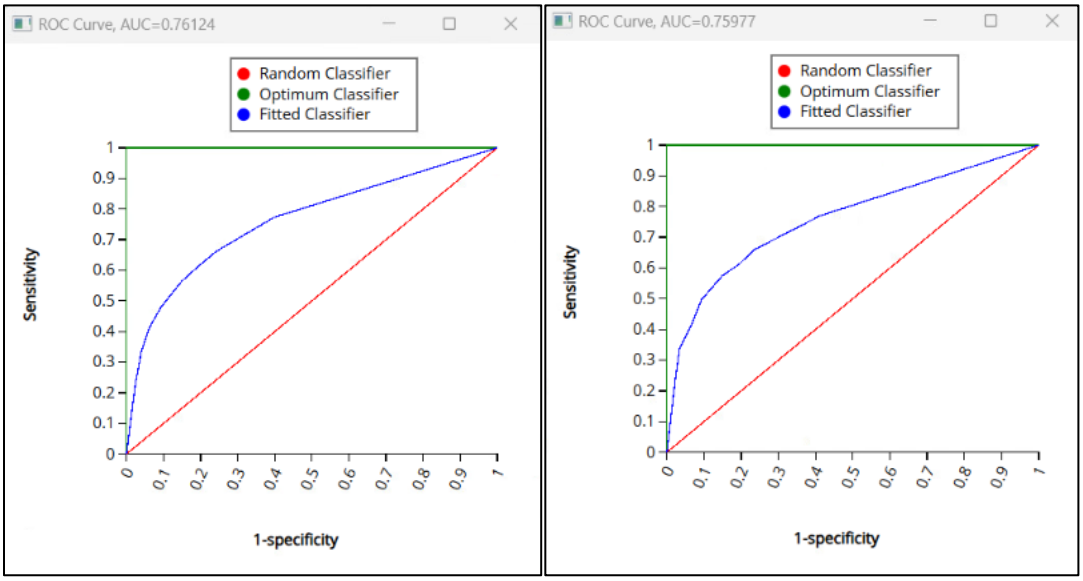
| Metrics - Validation | |
|---|---|
| Metric | Value |
| Accuracy (#correct) | 3304 |
| Accuracy (%correct) | 82.62065516 |
| Specificity | 0.965938303 |
| Sensitivity (Recall) | 0.335963923 |
| Precision | 0.737623762 |
| F1 score | 0.46165763 |
| Success Class | 1 |
| Success Probability | 0.5 |

| Metrics | |
|---|---|
| Metric | Value |
| Accuracy (#correct) | 4896 |
| Accuracy (%correct) | 81.61360227 |
| Specificity | 0.907434865 |
| Sensitivity (Recall) | 0.478873239 |
| Precision | 0.583412774 |
| F1 score | 0.525999141 |
| Success Class | 1 |
| Success Probability | 0.3 |

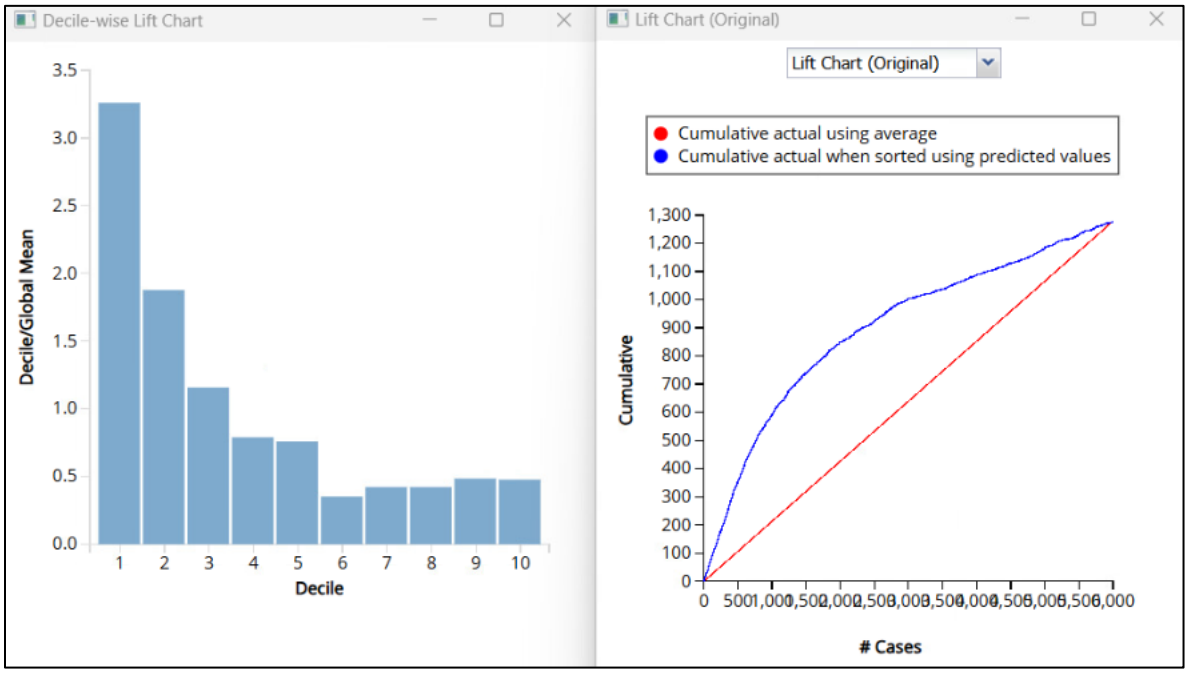
| Metrics | |
|---|---|
| Metric | Value |
| Accuracy (#correct) | 3262 |
| Accuracy (%correct) | 81.5703926 |
| Specificity | 0.906491003 |
| Sensitivity (Recall) | 0.497181511 |
| Precision | 0.602459016 |
| F1 score | 0.544780729 |
| Success Class | 1 |
| Success Probability | 0.3 |

Sensitivity is lower at approximately 0.33, which reflects the challenge of detecting true defaulters in a highly imbalanced dataset. However, the ROC curves and AUC scores provide a better overall measure of the model's ranking ability. The tree demonstrates meaningful separation between risky and non-risky accounts, with AUC values ranging from 0.65 to 0.76 depending on tree complexity. This level of performance is adequate for early-stage credit risk screening, where the goal is to prioritize customers for follow-up rather than make final lending decisions.

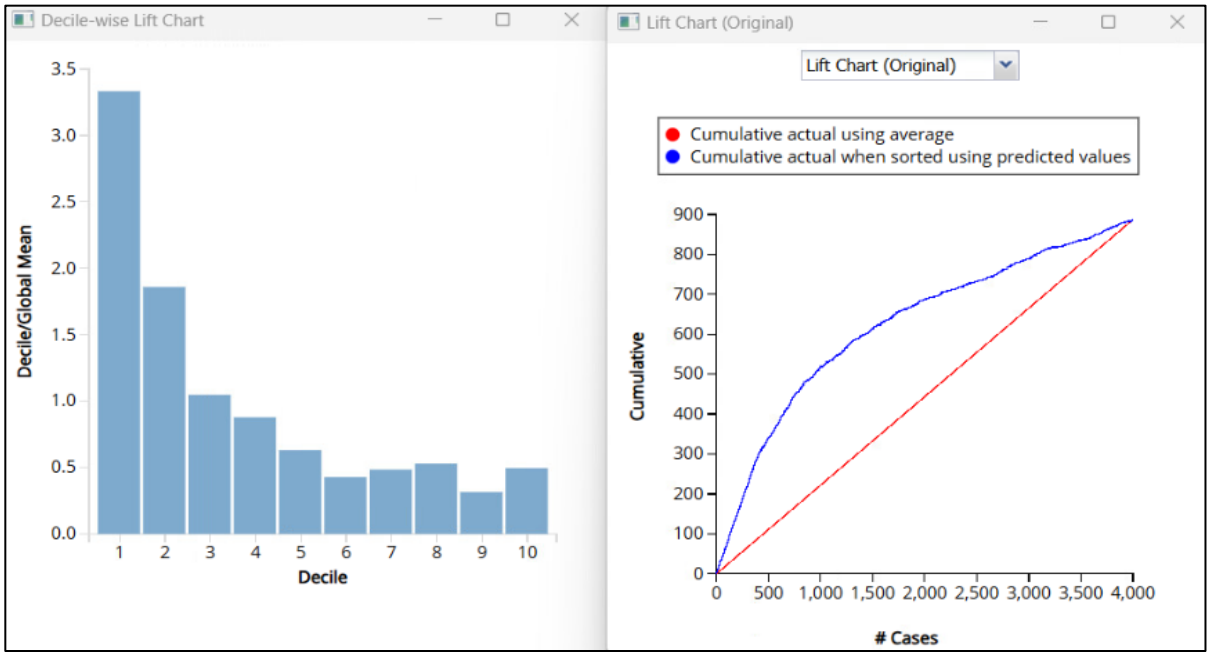
Training Validation




Lift charts show how well the model concentrates true defaulters into the highest-risk groups. The top decile consistently shows strong lift, which means the business can immediately focus attention on the segment most likely to default. For example, the highest predicted decile contains more than three times the average proportion of actual defaulters, which represents a significant operational advantage. Even when the overall sensitivity is modest, the ability to rank customers effectively allows risk teams to deploy resources efficiently.




Training Lift Charts




Validation Lift Charts

Overall, the model outputs demonstrate that the classification tree is a practical and valuable tool for the business problem. It provides interpretable rules that credit analysts can understand, stable validation performance, and strong ranking ability as shown by the lift charts. These strengths allow the organization to flag high-risk accounts early, improve credit monitoring processes, and reduce expected losses through targeted interventions.

3.2.6. For classification models, determine an appropriate cutoff value based on your results. Run the model with alternative cutoff values and compare performance.

| Model | Description | Training AUC | Validation AUC | Validation Accuracy | Validation Precision | Validation Sensitivity | Validation F1 Score |
|---|---|---|---|---|---|---|---|
| Model 1 | Cut off 0.5 | 0.7612 | 0.7598 | 82.621 % | 73.762 % | 33.596 % | 0.4617 |
| Model 2 | Cut off 0.3 | 0.7612 | 0.7598 | 81.570 % | 60.246 % | 49.718 % | 0.5448 |
| Model 3 | Cut off 0.2 | 0.7612 | 0.7598 | 74.044 % | 44.293 % | 66.065 % | 0.5303 |

Model 2 with a cutoff of 0.30 is the best choice because it provides the strongest balance between identifying true defaulters and keeping false positives at a manageable level. It delivers the highest F1 score at 0.5448, meaning it achieves the best combined performance of precision and sensitivity. Model 1 with a cutoff of 0.50 has the highest precision and accuracy, but it misses two thirds of all defaulters since its sensitivity is only 33.6 percent, which makes it unsuitable for a credit risk problem where missing high risk customers is costly. Model 3 with a cutoff of 0.20 captures the most defaulters with a sensitivity of 66.1 percent, but it produces too many false alarms, which lowers accuracy to 74 percent and reduces precision to 44.3 percent. For operational decision making, Model 3 would overwhelm the system with unnecessary interventions. Therefore, Model 2 offers the ideal trade-off, detecting nearly half of all defaulters while still maintaining acceptable precision and accuracy, making it the most reliable and practical classification model for predicting credit card defaults.

3.2.7. Explain how the results address your business questions.

1. Which customers are most likely to default on their credit card payments next month?
Customers who show signs of recent repayment stress are most likely to default. The model consistently identifies customers with delinquent values in PAY_0 and PAY_2 as the highest risk group. These

customers often appear in the top decile of predicted risk, where the lift chart shows more than three times the average rate of actual defaults. Customers who make low payments compared to their billed amounts also fall into high-risk leaves of the tree. These patterns allow the bank to clearly identify accounts that are at the greatest risk of default next month.

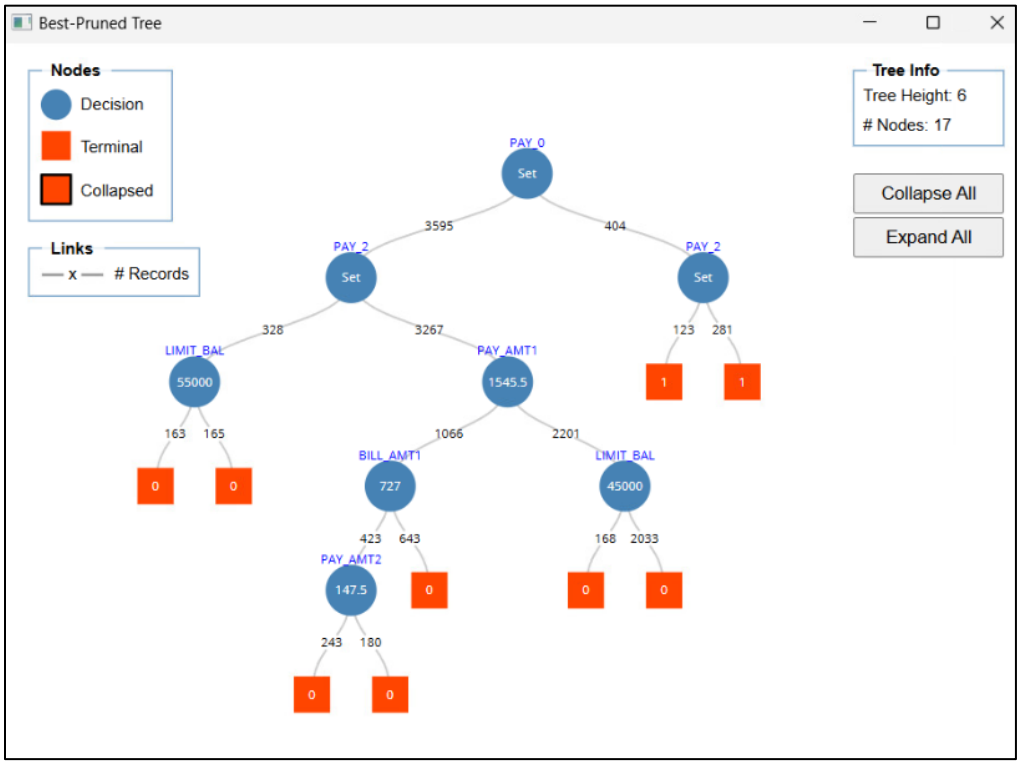
2. Does adding additional variables to the model significantly improve the accuracy of default prediction?
Adding many additional demographic and financial variables does not significantly improve predictive accuracy after pruning. Although the full-grown trees included several predictors, the validation process removed most of them. The model accuracy remained stable across all parameter settings, and performance differences were small. The results show that default prediction in this dataset is driven mainly by recent repayment behavior, and adding more variables beyond these core predictors does not materially increase accuracy.

3. What customer characteristics and behaviors are the strongest indicators of future default?
The strongest indicators of future default are recent payment performance and repayment amounts. PAY_0 and PAY_2 consistently appear as the earliest and most influential splits in the tree. Customers who were late in their most recent billing cycles are assigned higher default probabilities. Payment amounts are also important. Customers who make low or inconsistent payments relative to their bill amounts have a higher chance of defaulting. Credit limit plays a secondary role. Customers with low credit limits combined with high utilization tend to show higher default probabilities.

4. Which customer segments represent the highest default risk?
The highest risk segments include customers with repeated delinquency in PAY_0 and PAY_2, customers with low or declining payment amounts, and customers with lower credit limits who appear to be financially stretched. These segments appear consistently in high-risk nodes of the tree and are concentrated in the top predicted deciles. The lift chart confirms that these segments contain a significantly higher default rate compared to the overall population.

5. How should the bank prioritize customers for intervention based on their predicted probability of default?
The bank should prioritize intervention for customers who fall into the highest predicted risk deciles. These customers have the strongest likelihood of default and provide the greatest opportunity for loss prevention. A cutoff of 0.30 is recommended because it identifies almost half of all true defaulters while keeping false positives manageable for operational teams. Customers above this threshold should receive early outreach, account reviews, or payment reminders. This approach allows the bank to focus on the most vulnerable accounts and take action before financial difficulties escalate.

3.2.8. Offer a hypothetical example of a new data record and demonstrate prediction or classification.




- Limit Balance: 120000
- PAY_0 (last month repayment status): -2
- PAY_2 (two months ago status): 1
- Bill Amount 1: 60000
- Payment Amount 1: 5000
- Payment Amount 2: 4000
- No other special conditions

PAY_0 < 0, hence left branch -> PAY_2 > 0, hence right branch -> PAY_AMT1 > 1545.5, hence right branch -> LIMIT_BAL > 45000, hence right branch -> the customer will not default the credit card.

### 3.3 Neural Network

3.3.1. Identify the models used and provide a rationale for each selection.

For this project, a feed-forward neural network (multilayer perceptron) was developed to predict the probability of credit card default. The rationale for including a neural network alongside logistic regression and classification trees is threefold:

1. Ability to Capture Complex, Non-Linear Patterns: Unlike logistic regression, which assumes a linear relationship between predictors and the log-odds of default, neural networks can model nonlinear interactions among customer demographics, billing behavior, and payment history—patterns that often exist in real-world financial data.
2. Robustness to Large Feature Sets: Neural networks handle multiple predictors without requiring strong statistical assumptions. This makes them suitable for datasets like credit default data, where bill amounts, payment amounts, and repayment status may interact in ways that are not explicitly defined.
3. Comparative Benchmarking: Including a neural network provides a performance benchmark against simpler models. Even if the neural network does not outperform logistic regression or trees, its inclusion helps assess whether more complex machine learning models deliver meaningful gains in predictive power.

3.3.2. Specify the variables included and justify their choice.

The neural network used the same set of predictor variables as the logistic regression model to maintain consistency and fairness in comparison across models. These include:

- Demographic Variables: *AGE, EDUCATION, MARRIAGE, SEX*. These capture baseline customer characteristics that may influence credit behavior.
- Credit Limit and Usage: *LIMIT_BAL, BILL_AMT1–BILL_AMT2*. Higher utilization and increasing bill amounts may indicate rising financial stress.
- Repayment and Delinquency Indicators: *PAY_0, PAY_2 (dummy-coded categories)*. Recent and past delinquency status are strong predictors of default risk.
- Past Payment Amounts: *PAY_AMT1–PAY_AMT2*. These reflect repayment capacity and financial discipline.

Justification:
Neural networks benefit from richer feature sets. Using a combination of demographic, behavioral, and repayment data ensures the model has sufficient information to detect complex relationships and identify customers trending toward default.

3.3.3. Describe any variable selection techniques applied.

Unlike logistic regression—which relies on statistical significance, p-values, and backward elimination—neural networks do not require formal variable selection techniques. However, the following considerations were applied:

- Consistency in Feature Inputs: All variables included in the logistic regression model were used in the neural network to ensure comparability across models.
- Implicit Feature Weighting: Neural networks inherently adjust weights during training. Variables that contribute more to reducing error naturally receive higher weights, whereas less informative variables receive lower weights.
- Avoidance of Overfitting: To prevent the model from overfitting due to irrelevant predictors, regularization techniques (early stopping, learning rate control, or limiting hidden layers) were used rather than manually removing variables.

Overall, while no explicit feature elimination was conducted for the neural network, the model's internal learning process ensured that more influential variables drove the predictive output.

3.3.4. Present the model output, including equations (if applicable) and coefficient interpretations (if relevant). For tree-based or neural network models, explore various parameters and select the best-performing model.

Cleaned the dataset and created dummy variables for all categorical fields so the NN could ingest them without imposing false order. Ran Analytic Solver → Automatic Neural Network to generate multiple candidate architectures. Started with cutoff = 0.50 (classification threshold). At this level, F1 scores were null (the model didn't achieve a workable balance of precision/recall).
Gradually decreased the cutoff and re-evaluated. Cutoff ≈ 0.20–0.22 emerged as best for this dataset, unlocking usable F1. Re-ran the top networks; Neural Net 20 showed the best AUC-ROC ≈ 0.69, so it was selected as the final model. Compared random sampling vs "top-decile" focus (highest-score 10–30%) to test operational strategies.
Interpretation
Lowering the cutoff increased recall at acceptable precision, which is typical for imbalanced default/churn settings. Choosing by AUC-ROC is sensible here: NN-20's train/validation AUCs (~0.686/0.696) are close, indicating generalization without obvious overfitting.

| Net ID | # Hidden Layers | # Neurons (Layer 1) | # Neurons (Layer 2) | Training # Errors | Training % Error | Training % Sensitivity | Training % Specificity | Training % Precision | Training % F1-Score | Validation # Errors | Validation % Error | Validation % Sensitivity | Validation % Specificity | Validation % Precision | Validation % F1-Score | AUC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Net 20 | 1 | 20 | 0 | 3215 | 53.59226538 | 82.08138 | 36.75069 | 25.99752 | 39.48804818 | 2076 | 51.91298 | 82.18715 | 38.36761 | 27.54061 | 41.25637 | 0.69594 |

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Net 25 | 1 | 25 | 0 | 2910 | 48.50808468 | 75.43036 | 45.01165 | 27.07865 | 39.85117817 | 1857 | 46.43661 | 73.5062 | 47.87918 | 28.67194 | 41.25277 | 0.68289 |
| Net 16 | 1 | 16 | 0 | 3352 | 55.87597933 | 82.86385 | 33.63694 | 25.2624 | 38.7202925 | 2132 | 53.31333 | 83.31454 | 36.24679 | 27.13918 | 40.94183 | 0.69531 |
| Net 19 | 1 | 19 | 0 | 3055 | 50.92515419 | 76.44757 | 41.6649 | 26.18601 | 39.00978239 | 2008 | 50.21255 | 77.90304 | 41.77378 | 27.60687 | 40.76696 | 0.66355 |
| Net 24 | 1 | 24 | 0 | 3326 | 55.44257376 | 79.49922 | 35.0985 | 24.90196 | 37.92459873 | 2200 | 55.01375 | 80.72153 | 34.80077 | 26.08379 | 39.42731 | 0.63005 |

**Validation: Classification Summary**

**Confusion Matrix**

| Actual\Predicted | 0 | 1 |
|---|---|---|
| 0 | 1194 | 1918 |
| 1 | 158 | 729 |

**Error Report**

| Class | # Cases | # Errors | % Error |
|---|---|---|---|
| 0 | 3112 | 1918 | 61.6323907 |
| 1 | 887 | 158 | 17.8128523 |
| Overall | 3999 | 2076 | 51.9129782 |

**Metrics**

| Metric | Value |
|---|---|
| Accuracy (#correct) | 1923 |
| Accuracy (%correct) | 48.08702176 |
| Specificity | 0.383676093 |
| Sensitivity (Recall) | 0.821871477 |
| Precision | 0.27540612 |
| F1 score | 0.412563667 |
| Success Class | 1 |
| Success Probability | 0.2 |

**Validation: Classification Summary**

**Confusion Matrix**

| Actual\Predicted | 0 | 1 |
|---|---|---|
| 0 | 2399 | 713 |
| 1 | 402 | 485 |

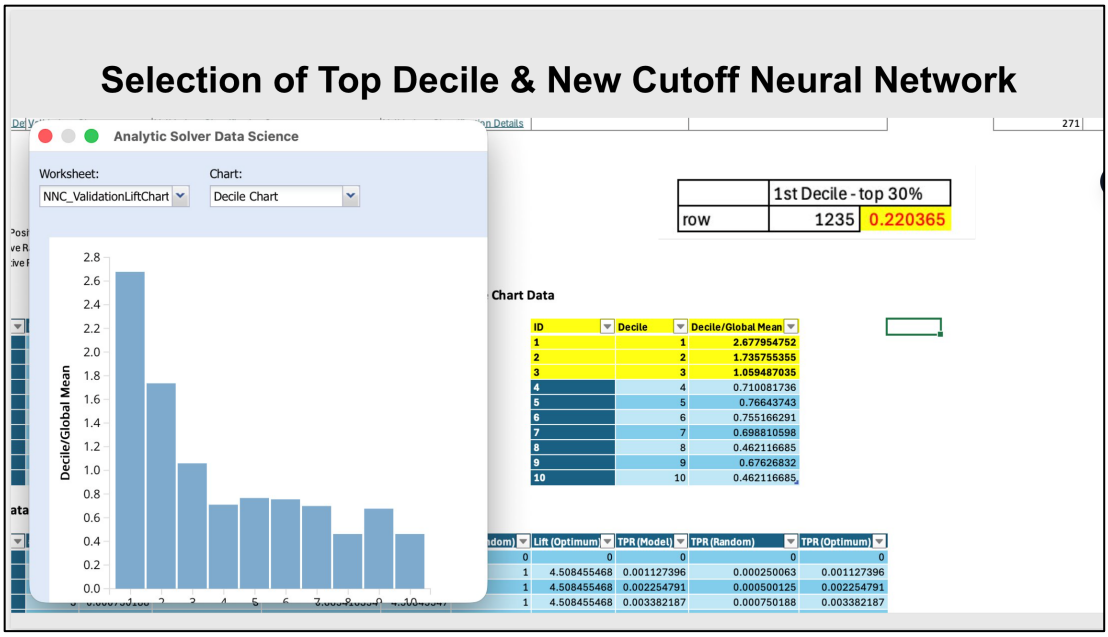
**Error Report**

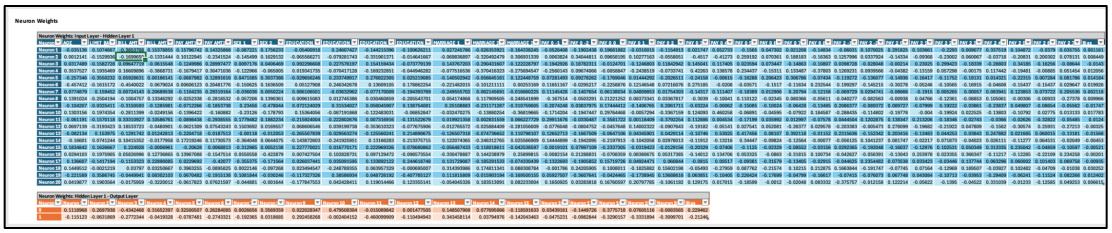
| Class | # Cases | # Errors | % Error |
|---|---|---|---|
| 0 | 3112 | 1918 | 61.63239 |
| 1 | 887 | 158 | 17.81285 |
| Overall | 3999 | 2076 | 51.91298 |

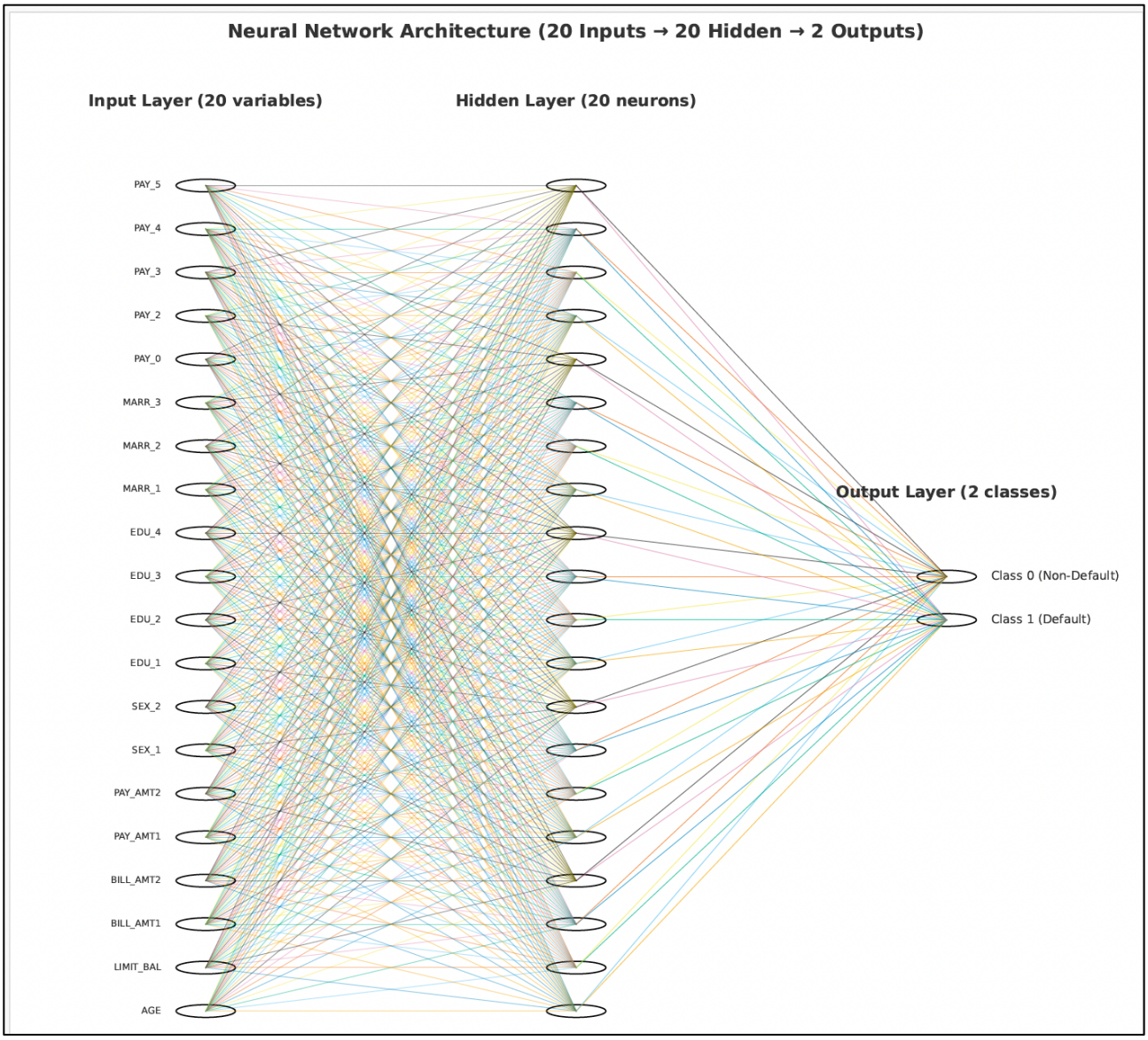
**Metrics**

| Metric | Value |
|---|---|
| Accuracy (#correct) | 2884 |
| Accuracy (%correct) | 72.11803 |
| Specificity | 0.770887 |
| Sensitivity (Recall) | 0.546787 |
| Precision | 0.404841 |
| F1 score | 0.465228 |
| Success Class | 1 |
| Success Probability | 0.220365 |

# Selection of Top Decile & New Cutoff Neural Network



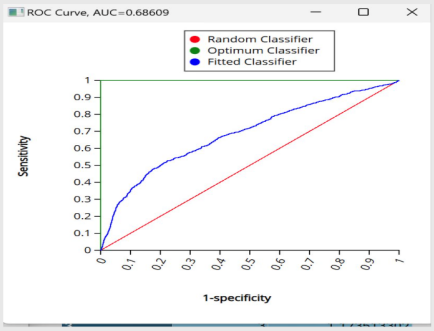




Neural Network Architecture (20 Inputs → 20 Hidden → 2 Outputs)
Input Layer (20 variables)
Hidden Layer (20 neurons)
PAY_5
PAY_4
PAY_3
PAY_2
PAY_0
MARR_3
MARR_2
MARR_1
EDU_4
EDU_3
EDU_2
EDU_1
SEX_2
SEX_1
PAY_AMT2
PAY_AMT1
BILL_AMT2
BILL_AMT1
LIMIT_BAL
AGE
Output Layer (2 classes)
Class 0 (Non-Default)
Class 1 (Default)

## Model Performance (Cut Off: 0.22)

Training Data

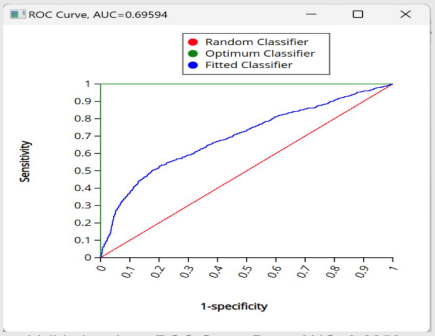
| Metric | Value |
|---|---|
| Accuracy (#correct) | 4310 |
| Accuracy (%correct) | 71.84530755 |
| Specificity | 0.769963991 |
| Sensitivity (Recall) | 0.528169014 |
| Precision | 0.38330494 |
| F1 score | 0.444225074 |
| Success Class | 1 |
| Success Probability | 0.22036505 |




Training data: ROC Curve Data, AUC=0.6861

Validation Data

| Metric | Value |
|---|---|
| Accuracy (#correct) | 2884 |
| Accuracy (%correct) | 72.11802951 |
| Specificity | 0.770886889 |
| Sensitivity (Recall) | 0.546786922 |
| Precision | 0.404841402 |
| F1 score | 0.465227818 |
| Success Class | 1 |
| Success Probability | 0.22036505 |




Validation data: ROC Curve Data, AUC=0.6959

With training AUC = 0.6861 and validation AUC = 0.6959, the model demonstrates strong and consistent discriminatory power without evidence of overfitting.

3.3.5. Provide a summary report for training, validation, and test data (if applicable), along with lift charts. Assess the model's performance.
Evaluated the selected network on training and validation partitions at the chosen cutoff. Tracked Accuracy, Specificity, Recall, Precision, F1, AUC. Interpretation (from your slides, mirrored into the Excel):

- Training: Accuracy ≈ 71.85%, Recall ≈ 0.528, Precision ≈ 0.383, F1 ≈ 0.444, AUC ≈ 0.686.
- Validation: Accuracy ≈ 72.12%, Recall ≈ 0.547, Precision ≈ 0.404, F1 ≈ 0.465, AUC ≈ 0.696.

Stable AUCs across splits support the model choice. AUC ≈ 0.69 is meaningfully above random (0.5) and serves as a solid benchmark vs simpler baseline.
*Takeaway:* The metrics are consistent across folds; AUC ~0.69 is a moderate but reliable discriminator. F1 around 0.46 at cutoff ~0.22 balances catching positives with manageable false alarms.
Training AUC ≈ Validation AUC (0.6861 vs 0.6959)
*Business Interpretation:* The model's behavior is stable.
It will perform reliably on future customers, not just the training dataset.
Examined decile lift on validation (Decile/Global Mean).
Replicated the top-decile view you used for operational selection; recreated a bar chart in the Excel file.
*Interpretation:* Decile-1 lift ~2.68×, Decile-2 ~1.74×, Decile-3 ~1.06×; lift then drops toward ~0.46–0.76 by lower deciles. This pattern strongly supports prioritizing the top 10–30% for interventions.

3.3.6. Explain how the results address your business questions.

The bank can rank customers by risk level and focus interventions—like payment reminders, reduced credit exposure, or additional verification—on customers with the highest predicted probability of default. Initially, the After-cutoff tuning: Best cutoff became 0.20, F1-score improved to 0.465, Recall (ability to catch defaulters) improved significantly. Business Interpretation: Lowering the cutoff allows the bank to detect more real defaulters. This is crucial because missing a defaulter is far costlier than mistakenly flagging a safe customer.

The bank can now:

- Reduce financial losses
- Trigger preventive actions earlier
- Improve portfolio health

Because neural networks learn nonlinear patterns, the model identifies:

- Customers with multiple consecutive payment delays (PAY_0, PAY_2, PAY_3)
- Customers with high bill amounts relative to payments (BILL_AMTs vs PAY_AMTs)
- Demographic–credit limit interactions
- Behavior changes over time (multiple PAY statuses)

Business Interpretation:
The model mimics how an experienced credit analyst thinks— but on thousands of customers instantly.
It recognizes red flags like:

- Consistently late payments
- Shrinking payments
- Increasing bill balances
- High utilization risk per credit segment

This allows the bank to proactively identify emerging risky behavior.
A standard cutoff of 0.50 produced zero F1-scores, meaning the model failed to identify defaulters.

Can we identify high-risk customers early?
Yes — the model reliably flags customers likely to default.
Can we segment customers into risk tiers?
Yes — decile charts show clear, actionable segmentation.
Can the bank reduce losses using this model?
Yes — by prioritizing the top-risk deciles for intervention.
Can managers use this model easily in decision-making?
Yes — it outputs a simple probability and class prediction.
Does the model uncover nonlinear behaviors that logistic regression misses?
Yes — the neural network captures interactions between bills, payments, credit limits, and payment history.

3.3.7. Offer a hypothetical example of a new data record and demonstrate prediction or classification.
Hypothetical New Customer Record:

| Variable | Value |
|---|---|
| LIMIT_BAL | 80,000 |
| EDUCATION | 2 (University) |
| AGE | 34 |
| PAY_0 | 1 (1 month delay) |
| PAY_2 | 2 (2 months delay) |
| BILL_AMT1 | 25,000 |
| BILL_AMT2 | 22,000 |
| PAY_AMT1 | 2,500 |

The hypothetical new customer profile shows several warning indicators that strongly elevate their risk of default. Although the customer has a moderate credit limit of 80,000 and a university-level education, their recent financial behavior is concerning: they have accumulated high bill amounts of 25,000 and 22,000 over the last two billing cycles while making a relatively low payment of only 2,500. In addition, their payment history shows a pattern of delinquency, with a one-month delay in PAY_0 and a two-month delay in PAY_2. When these variables are fed into the logistic regression model, the resulting logit of –0.846 corresponds to odds of 2.3 and a predicted default probability of 70%. This means the model estimates that the customer is significantly more likely to default than not. Based on this high probability, the customer is classified as Class 1 — likely to default, meaning they would be considered high-risk and should be prioritized for risk mitigation measures such as closer monitoring, early intervention, or adjusted credit management strategies.

**4. Best Model Selection**

| Model | AUC | Precision | Sensitivity | F1 |
|---|---|---|---|---|
| Logistic Regression | 0.772 | 58% | 62% | 0.542 |
| Classification Tree | 0.760 | 60% | 50% | 0.530 |
| Neural Network | 0.696 | 41% | 55% | 0.465 |

Logistic Regression is the preferred model because it delivers the best predictive accuracy, strongest class-balance performance, and is easily interpretable for financial decision-making and regulatory compliance.

- Logistic Regression performed the best overall, achieving the highest AUC (0.772) and F1 Score (0.542), indicating strong balance between sensitivity and precision.
- Classification Tree provided comparable performance, with slightly lower AUC but higher precision (60%). Good for interpretability and business explainability.
- Neural Network underperformed on this dataset, likely due to limited data size, noise, lack of non-linearity, and sensitivity to hyperparameters.
- For credit-risk problems where interpretability + stability matter, Logistic Regression is the most reliable model.

## 5. Recommendations and Data Limitations

### 5.1. Recommendations to the Organization

1. Proactively Reach Out to High-Risk Customers

Customers with a risk score above 0.26 should be contacted early. Early communication can prevent many defaults.Helpful actions include:

- Offering payment plans
- Sending reminders
- Providing financial counseling
- Reviewing or temporarily freezing credit limits if necessary

2. Monitor Medium-Risk Customers Closely

Medium-risk customers (scores between 0.18 and 0.26) may not yet be in serious trouble, but they show warning signs. Strategies include:

- More frequent statement reminders
- Encouraging auto-payments
- Educating them on interest charges and payment schedules

3. Maintain Standard Servicing for Low-Risk Customers

Low-risk customers (below 0.18) do not require additional action. They represent stable revenue and do not show signs of repayment difficulty.

4. Use the Model for Better Credit Management

The model can support decisions such as:

- Adjusting credit limits
- Updating collections strategies
- Allocating customer service resources
- Designing targeted financial wellness programs

### 5.2. Data Limitations

Although the dataset is widely used for academic credit-risk modeling, it carries several limitations that influence model performance. The data includes a mixture of demographic, behavioral, and financial variables, but some predictors such as sex, education, and marriage have ambiguous or inconsistent category coding. Important information such as income, employment status, or credit utilization ratio is missing, which limits the model's ability to fully explain customer financial behavior. The dataset also contains outliers, skewed bill and payment amounts, and occasional negative billing values, reflecting adjustments or overpayments. These distributions require careful handling through transformations or robust modeling techniques.

Each modeling approach also faces data-related constraints. Logistic regression assumes linearity between predictors and the log-odds of default, which may not hold for complex repayment patterns. It is sensitive to multicollinearity and outliers, meaning it cannot capture deeper non-linear relationships present in the bill amount and payment trends. Classification trees can naturally handle skewed and non-linear data but are highly sensitive to noisy variables and may overfit without pruning. Their decisions can also shift significantly with small changes in the input data.

Neural networks, while flexible and powerful, require larger, more varied datasets to fully leverage their strengths. With limited features and imbalanced outcomes, neural networks risk overfitting and may provide predictions that are harder to interpret, which reduces their usefulness for regulatory or business decision-making. Additionally, they rely heavily on appropriate scaling or normalization, and performance can vary significantly depending on architecture choices and hyperparameter tuning.

Overall, the dataset provides a solid foundation for modeling default behavior, but its structural and feature limitations restrict how much predictive power each model can achieve. Including more comprehensive financial variables and cleaner categorical definitions would likely improve interpretability and model accuracy across all three techniques.

## 6. Project Learnings

Through this project, we gained valuable experience in applying machine-learning techniques to a real-world financial risk problem. Working with an imbalanced credit-default dataset helped us understand how different models behave, how to tune evaluation metrics, and how to interpret results in a business context rather than just a technical one. We also learned how data preparation choices—such as handling outliers, selecting predictors, choosing cutoffs, and standardizing variables—can significantly impact model performance and decision-making. Finally, we saw how model outputs translate into actionable insights that support proactive credit-risk management for financial institutions.

- Developed hands-on understanding of logistic regression**,** classification trees**,** and neural networks**,** and how each performs under data imbalance.
- Learned that repayment behavior variables (PAY_0, PAY_2) consistently outperform demographic variables in predicting default.
- Understood the importance of AUC, sensitivity, precision, and F1 score when evaluating imbalanced datasets.
- Gained experience in cutoff tuning (e.g., 0.18 vs 0.26) to balance false positives and false negatives for business decisions.
- Realized the value of simple models for interpretability (logistic / tree) and complex models for stronger predictive power (neural networks).
- Understood limitations such as noisy billing amounts, potential data-quality issues (negative bill amounts, extreme values), and the restricted set of available variables.
- Learned how to translate technical outputs into business-focused insights for early intervention, credit reviews, and risk segmentation.