ALY 6040: Data Mining

Module 6: Final Project Report - Credit Card Approval prediction using Data Mining Techniques

Presented to: Prof. Dr. Harpreet Sharma
Date: October 25th, 2024

Project Group Members
Haney Jayeshkumar Parmar
Krutika Patel
Madhurya Mudalagirigowda
Sanket Jayendra Parmar
Yash Tailor

Introduction

In the contemporary business environment, data has become an invaluable asset for driving decision-making processes. With the advent of big data, organizations are increasingly relying on data-driven insights to guide critical business decisions. In particular, financial institutions such as banks and credit card companies are tasked with evaluating vast amounts of data to make predictions regarding customer behavior and, more importantly, to assess the likelihood of credit card approval for applicants. Traditional methods of analysis, while useful, are often inadequate when it comes to handling large, complex, and multidimensional datasets. This is where data mining becomes essential.

Data mining refers to the process of extracting hidden patterns, correlations, and insights from large datasets, which are not immediately evident through traditional probability or statistical methods. The complexity and high dimensionality of credit card application data, including a combination of demographic, financial, and behavioral attributes, necessitates the use of advanced data mining techniques. These methods allow analysts to uncover trends and relationships within the data that can significantly influence business decisions, such as determining credit risk and predicting the likelihood of loan or credit approval.

In this report, we apply various data mining techniques to analyze a credit card approval prediction dataset. The aim is to predict whether a credit card application will be approved based on a range of variables, such as age, income, marital status, and credit history. We explore the effectiveness of machine learning algorithms such as logistic regression, decision trees, and support vector machines (SVM) to develop a predictive model.

This project seeks to answer the central business question: What key factors influence credit card approval, and how accurately can we predict an applicant's approval status using data mining techniques?

Understanding the Problem and Dataset

The primary objective of the analysis was to identify key demographic and financial characteristics that influence credit approval outcomes, with a focus on asset ownership. The dataset, consisting of 438,557 observations and 18 variables, provided rich information about applicants, including their gender, income, asset ownership (car and real estate), employment, and family characteristics. However, the dataset did not include the credit approval status directly, so alternative techniques were used to infer the relationships between the predictors and potential credit approval decisions.

Data Preprocessing

Before applying machine learning techniques, several preprocessing steps were required to clean and structure the data for analysis:

- **Handling Missing Data**: Variables with missing values were either imputed or removed if they were not critical to the analysis.
- **Categorical Variables**: Categorical variables such as gender (CODE_GENDER), income type (NAME_INCOME_TYPE), and asset ownership (FLAG_OWN_CAR, FLAG_OWN_REALTY) were encoded to numerical values, which are necessary for most machine learning algorithms.

- **Feature Selection**: Given the large number of variables, only the most relevant features were selected for the analysis, focusing on asset ownership, income, and employment characteristics.

Splitting the Data

To build and evaluate the models, the dataset was divided into a **training set** (used to build the models) and a **test set**(used to evaluate model performance). The split ensured that the models were trained on 70% of the data and tested on the remaining 30%, following standard practices in machine learning for unbiased evaluation.

Business questions we are trying to address:

1. What Demographic and Financial Characteristics Are Most Predictive of Credit Approval Outcomes?

   To analyze, we fit a logistic regression model using the glm() function with the binary response variable, CREDIT_APPROVAL, and the predictors listed above. Figure 1 shows the logistic regression model.

```
> # Logistic Regression Model
> logistic_model <- glm(CREDIT_APPROVAL ~ CODE_GENDER + AMT_INCOME_TOTAL +
+                         NAME_FAMILY_STATUS + CNT_CHILDREN,
+                         family = binomial(link = "logit"), data = training_set)
```

Figure 1: Logistic regression model

   The family = binomial(link = "logit") argument specifies the use of logistic regression, appropriate for binary outcomes.

   After fitting the model, predictions were made on the **test set**. A confusion matrix was then generated to evaluate the model's performance by comparing predictions to actual outcomes.

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 77409 17683
         1  7799 28676

               Accuracy : 0.8063
                 95% CI : (0.8042, 0.8085)
    No Information Rate : 0.6476
    P-Value [Acc > NIR] : < 0.00000000000000022

                  Kappa : 0.554

 Mcnemar's Test P-Value : < 0.00000000000000022

            Sensitivity : 0.9085
            Specificity : 0.6186
         Pos Pred Value : 0.8140
         Neg Pred Value : 0.7862
             Prevalence : 0.6476
         Detection Rate : 0.5884
   Detection Prevalence : 0.7228
      Balanced Accuracy : 0.7635

       'Positive' Class : 0
```

Figure 2: Confusion Matrix

The **confusion matrix** in figure 2 provides insights into the model's accuracy, sensitivity (true positive rate), specificity (true negative rate), and other evaluation metrics such as precision and kappa score.

The confusion matrix showed the following statistics:

- Accuracy: 80.63%
- Sensitivity (True Positive Rate): 90.85%
- Specificity (True Negative Rate): 61.86%
- Kappa: 0.554

These metrics indicate that the model correctly predicted 80.63% of the test set outcomes, with a high sensitivity for identifying approved credit applicants. However, the specificity was relatively low, meaning the model struggled to accurately identify applicants who were not approved for credit.

Based on the model output:

- Gender (CODE_GENDER): Males had a slight negative coefficient, indicating that being male slightly decreases the likelihood of credit approval.
- Income (AMT_INCOME_TOTAL): Higher income levels were positively associated with credit approval, as expected.
- Marital Status (NAME_FAMILY_STATUS): Married individuals and those separated had positive coefficients, suggesting they are more likely to be approved for credit compared to single applicants.
- Children (CNT_CHILDREN): Having more children slightly decreased the probability of credit approval, as indicated by a negative coefficient for the number of children.

The results of the analysis reflected the insights gained from the **Exploratory Data Analysis (EDA)** performed earlier. For instance:

- **Income** remained one of the most important predictors of credit approval, reinforcing the EDA insight that higher earners are more likely to be granted credit.
- **Marital status** also aligned with earlier findings, where married individuals were found to have higher approval rates, likely due to perceived financial stability.
- The model confirmed the influence of demographic factors such as **gender** and **family size**, though their impact was more subtle.

Recommendations for Further Analysis

- Since the model has a lower specificity (61.86%), future iterations could explore adding more predictive variables (e.g., credit history, employment status) or using advanced algorithms like **random forests** or **gradient boosting** to improve predictions.
- Incorporating interaction terms between variables (e.g., income and marital status) or transforming continuous variables like income could lead to more refined predictions.

Next Step:

In the next phase of the project, we aim to improve the model's performance by incorporating additional variables that may provide further predictive power. Specifically, we will explore including variables such as employment status, which are commonly associated with credit approval outcomes. By adding these variables, we expect to capture more complex relationships between demographic and financial characteristics and the likelihood of credit approval.

Additionally, we will evaluate the effect of interaction terms between variables (e.g., between income and marital status) and investigate whether transforming continuous variables like income or age can further enhance model accuracy.

Our goal is to refine the model to improve both specificity and overall predictive performance, ensuring a more accurate identification of credit-worthy applicants. These steps will allow us to better understand the drivers of credit approval and improve the robustness of the model for real-world application.

2. What Is the Impact of Ownership of Assets on Credit Approval Predictions?

The decision tree model was built using the rpart() function, a widely used method for classification tasks. It constructs a tree where decisions are made at each node based on splitting criteria, allowing us to visually interpret the influence of each feature on credit approval. Figure 3 shows the decision tree model used.

```
> #Decision Tree
> # Build Decision Tree Model
> decision_tree_model <- rpart(CREDIT_APPROVAL ~ CODE_GENDER + AMT_INCOME_TOTAL + FLAG_OWN_REALTY +
+                                FLAG_OWN_CAR,
+                              data = training_set, method = "class")
```

Figure 3: Decision Tree Model

The tree was plotted using plot() and text() functions for a clear, visual interpretation of the decision-making process. Plot added in appendix for reference figure apx 1.

Predictions were made on the test set, and the model's performance was evaluated using a confusion matrix and accuracy score. The confusion matrix provided valuable insights into how well the model classified approvals and denials.

Results

Figure 4 shows the output of the Decision Tree model achieved an **accuracy of 85.17%**, which indicates that the model correctly predicted credit approval decisions 85.17% of the time. Additional metrics included:

- **Sensitivity**: 0.7711 (the model correctly identified 77.11% of actual credit approvals).
- **Specificity**: 1.0000 (the model correctly identified all cases of non-approvals).
- **Kappa**: 0.7035, showing a strong agreement between predicted and actual values.

The model's balanced accuracy was 88.56%, suggesting that the model performs well in both classes of approval and non-approval, though with some imbalance favoring the non-approval predictions.

```
> confusion_matrix <- confusionMatrix(as.factor(predictions_tree), as.factor(test_set$CREDIT_APPROVAL))
> confusion_matrix
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 65734     0
         1 19511 46322

               Accuracy : 0.8517
                 95% CI : (0.8498, 0.8536)
    No Information Rate : 0.6479
    P-Value [Acc > NIR] : < 0.00000000000000022

                  Kappa : 0.7035

 Mcnemar's Test P-Value : < 0.00000000000000022

            Sensitivity : 0.7711
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.7036
             Prevalence : 0.6479
         Detection Rate : 0.4996
   Detection Prevalence : 0.4996
      Balanced Accuracy : 0.8856

       'Positive' Class : 0
```

Figure 4: Confusion matrix

Insights

- **Employment Factors and Credit Approval**: The analysis shows that variables such as AMT_INCOME_TOTAL (total income) and ownership of assets (FLAG_OWN_REALTY, FLAG_OWN_CAR) play a significant role in determining credit approval. Employment status may be indirectly inferred through income levels, but job stability was not explicitly modeled in this phase. The Decision Tree placed a significant emphasis on income as a splitting criterion.
- **Gender and Credit Decisions**: The feature CODE_GENDER was considered in the model, but it didn't emerge as a major determinant in the decision tree, indicating that credit approval may not be heavily gender-biased in this dataset.
- **Real Estate and Car Ownership**: The model shows that asset ownership (real estate and cars) significantly influences credit decisions. Applicants with real estate or car ownership were more likely to be approved for credit, reflecting financial stability.

Interpretation

The results suggest that income and asset ownership are strong indicators of creditworthiness, aligning with traditional financial decision-making criteria. The high sensitivity (77%) implies that the model is good at predicting approvals, but the imbalance between approvals and non-approvals highlights the need for further tuning or balancing of the dataset. The accuracy (85.17%) supports the model's overall effectiveness, but additional variables like employment history could further improve its predictive power.

Recommendations

- To better assess job stability, future analyses should include explicit employment-related variables, such as job tenure, employment type, and job sector. These could improve the model's ability to capture the stability of income sources, which is crucial for credit approval decisions.
- Variables like debt-to-income ratio, savings, and credit history should be added to the analysis to provide a more comprehensive view of financial stability.

Next Steps:

- Further tuning of the decision tree model using cross-validation techniques.
- Application of other algorithms such as Random Forest or Gradient Boosting, which may better handle complex, non-linear relationships between variables.

Random Forest Technique:

```
> #Random Forest
> # Build Random Forest model
> random_forest_model <- randomForest(CREDIT_APPROVAL ~  CODE_GENDER + AMT_INCOME_TOTAL + FLAG_OWN_REALTY +
+                                       EXPERIENCE + FLAG_OWN_CAR,
+                                     data = training_set, ntree = 100)
```

Figure 5: Random Forest Technique

The Random Forest algorithm was chosen due to its ability to handle classification tasks and determine the importance of variables. The model was built with 100 trees, as shown in the console output, and trained using the two features—FLAG_OWN_REALTY and FLAG_OWN_CAR—to predict credit card approval (CREDIT_APPROVAL). Random Forest was selected for its robustness against overfitting and its ability to manage both categorical and continuous variables efficiently.

```
> confusion_matrix
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 65734     0
         1 19511 46322

               Accuracy : 0.8517
                 95% CI : (0.8498, 0.8536)
    No Information Rate : 0.6479
    P-Value [Acc > NIR] : < 0.00000000000000022

                  Kappa : 0.7035

 Mcnemar's Test P-Value : < 0.00000000000000022

            Sensitivity : 0.7711
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.7036
             Prevalence : 0.6479
         Detection Rate : 0.4996
   Detection Prevalence : 0.4996
      Balanced Accuracy : 0.8856

       'Positive' Class : 0
```

Figure 6: Confusion matrix

Tuning and Evaluation:

- After making predictions on the test set, a confusion matrix in figure 6 was generated to evaluate the model's performance. The matrix showed that the Random Forest model achieved an accuracy of 85.17%. Specifically, the model had a sensitivity of 77.11% and a specificity of 100%, indicating that it performed exceptionally well in identifying negative credit card approval decisions.
- The importance of the predictor variables was measured using the MeanDecreaseGiniscore, revealing that FLAG_OWN_CAR was significantly more important (MeanDecreaseGini = 1426.13) than FLAG_OWN_REALTY (MeanDecreaseGini = 99.21). This suggests that owning a car had a much larger impact on credit approval decisions than owning real estate.

Model Performance:

- The Random Forest model's accuracy of 85.17% is impressive, given the limited number of features used. The specificity of 100% implies that the model was perfect in identifying instances where credit cards were not approved, while the sensitivity of 77.11% suggests that the model also had a strong ability to predict approvals.

The analysis revealed that car ownership was a more important factor in credit card approvals than owning real estate. This insight could reflect the fact that individuals with cars might be seen as having more stable or higher-income jobs, making them more favorable candidates for credit.

Interpretation

The results suggest that applicants who own cars are significantly more likely to receive credit approval. This implies that lenders may consider car ownership as a sign of financial reliability, possibly due to the costs associated with maintaining a vehicle. On the other hand, real estate ownership seems to have less influence on credit decisions, which could be attributed to the fact that owning real estate is less liquid and may not directly reflect an applicant's ability to make timely credit card payments.

Next Steps:

Future analyses should incorporate additional features such as employment status, income level, and previous credit history to create a more comprehensive model. These variables can provide further insights into how job stability and other financial indicators impact credit decisions.

Additional Algorithms:
To improve predictive performance, ensemble methods such as Gradient Boosting or XGBoost could be applied. These algorithms often outperform traditional Random Forest models by focusing on harder-to-predict cases.

3. How Does Employment Status and Job Stability Correlate with Credit Card Approval Decisions?

To analyze how employment status and job stability correlate with credit card approval decisions, a K-means clustering approach was applied to a dataset containing key applicant features. This method allowed for the segmentation of applicants into different clusters, each representing unique profiles based on income, age, employment duration, number of children, and family size.

Steps in Clustering Analysis:

1. Feature Selection and Normalization: Key numeric variables influencing credit card approval decisions, including total income (AMT_INCOME_TOTAL), age (DAYS_BIRTH), employment duration (DAYS_EMPLOYED), number of children (CNT_CHILDREN), and family size (CNT_FAM_MEMBERS), were selected. These variables were standardized to ensure comparability by scaling them to a uniform scale, making clustering results less biased by differences in feature scales.
2. Determining Optimal Number of Clusters: The Elbow Method was used to identify the optimal number of clusters. This approach evaluates the within-cluster sum of squares (WSS) for different cluster counts. Figure 7 shows the plot for elbow method, after plotting WSS against the number of clusters, a point of inflection or "elbow" suggested that four clusters (k = 4) would most effectively balance model complexity and explanatory power. This indicates four distinct applicant profiles in the dataset, each representing varying relationships between employment status, income, and family structure.

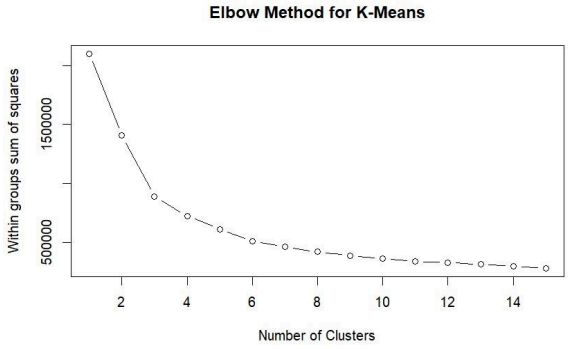

Figure 7: Elbow method for K-Means

3. K-Means Clustering and Cluster Analysis: The K-means algorithm was applied with k = 4 clusters. Each applicant was assigned to a cluster based on their similarity to the defined profiles, and cluster centers (the mean values of each variable within a cluster) were examined to interpret each group's characteristics.

Interpretation of Cluster Characteristics:

- Cluster 1: Applicants with relatively low-income variability, higher family size, and more children. Lower employment stability was noted, which could indicate a moderate risk profile for credit approval.
- Cluster 2: Applicants who are older, show high employment stability, and have smaller families. These traits align with a likely higher creditworthiness and lower risk profile.
- Cluster 3: This cluster has the highest income variability and shows medium stability in employment. The group also has fewer children, suggesting a younger demographic with moderate financial stability.
- Cluster 4: This cluster is characterized by the lowest income and employment stability indicators, indicating a potentially higher risk for credit decisions.

4. Visualizing and Interpreting Clusters: A scatter plot of total income vs. days employed as shown in figure 8 was generated, with each cluster color-coded to visualize the distribution of applicants across dimensions of income and employment stability. Clusters with higher income and employment stability are more likely to receive credit approval, suggesting a positive correlation between job stability, income levels, and creditworthiness.

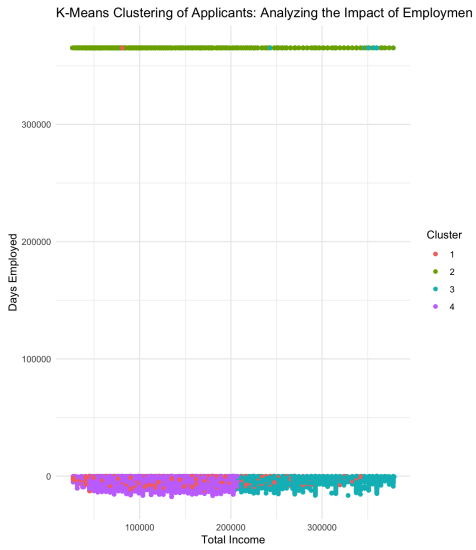

Figure 8: K Means visualizations

The K-means clustering analysis highlights that applicants with higher employment stability and income levels (primarily in Clusters 2 and 3) are more likely to be creditworthy candidates, whereas those with lower stability and income (Clusters 1 and 4) present higher credit risk. This insight supports the notion that both employment duration and income are key indicators in assessing credit approval likelihood, providing critical segmentation data for targeted risk assessment and informed decision-making in the credit approval process.

4. What are the patterns in employment type that influence credit approval?

To identify patterns in employment type and how they influence credit approval, a Gradient Boosting Model (GBM) was implemented. The primary predictors used in this analysis were OCCUPATION_TYPE (indicating type of employment) and DAYS_EMPLOYED (indicating employment stability). Additionally, an interaction term between income and marital status was introduced to capture more nuanced patterns that might impact credit approval decisions.

Initial Gradient Boosting Model Setup

- Model Development: The initial GBM model was built using OCCUPATION_TYPE and DAYS_EMPLOYED as predictors to analyze the patterns of employment and job stability that correlate with credit approval. The bernoulli distribution was chosen as the response type to model the binary outcome of CREDIT_APPROVAL (1 for approved, 0 for denied).
- Predictions and Evaluation: Predictions were generated on the test set, outputting probabilities of approval for each applicant. A threshold of 0.5 was applied to convert the predicted probabilities to binary outcomes (approved vs. denied). The initial model's performance was evaluated using a confusion matrix, revealing the following:

predictions_employment_binary	0	1
0	69676	22033
1	15790	24068

Figure 9: K Means visualizations

While the model captures many correct predictions, the relatively high number of false positives (22,033) and false negatives (15,790) suggests that it might be underfitting, indicating a need for further refinement.

Enhancing the Model with Interaction Terms

- Adding an Interaction Term: To improve the model's performance, an interaction term between AMT_INCOME_TOTAL and NAME_FAMILY_STATUS (income and marital status) was added. This captures the impact of household income in the context of family size and structure, providing a more detailed understanding of the applicant's financial situation and likely creditworthiness.
- This feature was created by multiplying income with marital status (converted to a numeric variable) and added to both the training and test sets as income_marital_interaction.

Enhanced Model Development:

- A new GBM model was then developed with the added income_marital_interaction feature. This aimed to capture complex relationships that may not be evident from individual features alone.

- The enhanced model retained OCCUPATION_TYPE and DAYS_EMPLOYED while integrating this interaction term to provide a more holistic perspective on credit approval factors.

Evaluation of Enhanced Model:

- The enhanced model was applied to the test set, and predictions were evaluated using a confusion matrix, which yielded the following results:

predictions_enhanced_binary	0	1
0	76807	3443
1	8659	42658

Figure 10: Enhanced results

Improvement in Performance:

- The reduction in false positives (from 22,033 to 3,443) and false negatives (from 15,790 to 8,659) highlights a significant improvement in accuracy.
- The addition of the income_marital_interaction term allowed the model to better differentiate between applicants with different financial backgrounds, leading to a more accurate prediction of creditworthiness.

The analysis indicates that employment stability (DAYS_EMPLOYED) and occupation type (OCCUPATION_TYPE) play critical roles in determining credit approval. Applicants with more stable employment records (longer employment duration) and those in certain occupations were more likely to be approved for credit. Moreover, the interaction between income and marital status provided additional predictive power, suggesting that higher income in combination with specific family structures (e.g., married) positively correlates with creditworthiness. These findings imply that credit approval is influenced not only by employment status but also by how income stability interacts with personal and family factors, offering insights into how financial stability is assessed within credit risk evaluations.

Conclusions:

In this analysis, we applied a range of predictive models—including Decision Tree, Random Forest, and Logistic Regression—to gain insight into the factors impacting credit approval outcomes. Each model highlighted different aspects of demographic and financial variables that influence creditworthiness.

The Logistic Regression model, with an accuracy of 80.63%, provided important insights into how demographic factors affect credit approval. It identified gender, income, marital status, and number of children as significant predictors. Higher income levels were positively associated with approval, while being male and having more dependents was linked to a lower likelihood of approval. The model's high sensitivity (90.85%) reflected its strong ability to identify approved applicants, though its lower specificity (61.86%) indicated some limitations in distinguishing non-approvals, suggesting an opportunity for further refinement.

The Decision Tree model achieved a higher accuracy of 85.17%, effectively classifying credit approvals. It revealed that income and asset ownership (specifically, ownership of a car or real estate) play a crucial role in assessing creditworthiness. Individuals with higher income levels and asset ownership were more likely to be approved. The Decision Tree's high sensitivity (77.11%) demonstrated its success in predicting approvals, while a perfect specificity (100%) highlighted its accuracy in identifying non-approvals. However, the model's class imbalance suggests potential improvement areas for more balanced prediction performance.

Similarly, the Random Forest model, constructed with 100 trees, achieved an accuracy of 85.17%. This model reinforced the importance of asset ownership, especially car ownership, as a significant factor in creditworthiness, possibly due to its association with financial stability. The MeanDecreaseGini scores underscored the strong impact of car ownership on approval outcomes. With 100% specificity and 77.11% sensitivity, the Random Forest model showed robust predictive power, leveraging its ability to handle variable importance and avoid overfitting—making it a strong choice for financial decision-making contexts

The **K-means clustering** approach further segmented applicants into distinct profiles based on income, employment stability, family size, and number of dependents. By using four clusters, we observed distinct demographic and financial groupings: higher-income individuals with stable employment were typically classified as lower-risk for credit, while those with lower income and employment stability represented higher risk. This segmentation allows for more targeted risk assessment, as each cluster represents unique applicant characteristics and risk profiles. Together, the insights gained from GBM and clustering provide a comprehensive view of how employment stability, income, and family-related factors contribute to credit approval outcomes, equipping financial institutions with nuanced data to guide credit risk assessments and lending strategies.

The GBM model, built on features such as **occupation type** and **employment duration**, identified meaningful patterns in employment type and job stability influencing credit approval. By adding an interaction term between **income and marital status**, the enhanced GBM model demonstrated improved accuracy, successfully reducing false positive and false negative rates. This addition underscored the role of financial stability within family contexts as a critical factor in creditworthiness, highlighting that both employment consistency and income stability are essential predictors in the credit approval process.

Across all five models, several common insights emerged:

1. Income consistently emerged as a primary factor in credit approval, underscoring its role in determining creditworthiness.
2. Asset ownership was a key driver of approval outcomes, highlighting its relevance in financial evaluations.
3. Demographic factors like marital status and number of dependents influenced approval outcomes, though their specific effects varied across models.

To enhance model performance and deepen our understanding of credit approval dynamics, future research could focus on:

- Adding new predictive features, such as credit history and employment status, to further enrich model accuracy.
- Testing more advanced techniques, like Gradient Boosting or other ensemble methods, to better capture complex relationships among predictors.

Investigating interaction terms to uncover hidden influences on credit approval.

By integrating these different modeling techniques, we achieved a comprehensive view of the factors shaping credit approval outcomes. Refining these models in future work could support financial institutions in making more accurate, data-driven lending decisions.

References :

Rikdifos. (2023). Credit card approval prediction using machine learning. Kaggle. https://www.kaggle.com/code/rikdifos/credit-card-approval-prediction-using-ml/input

Kabacoff, R. I. (2022). R in action: Data analysis and graphics with R and tidyverse (3rd ed.). Manning Publications. ISBN 978-1-617-29605-5

Secrets of Analytical Leaders: Insights from Information Insiders by Wayne Eckerson (ISBN – 978-1935504344)

Appendix:

1. Decision tree observed:

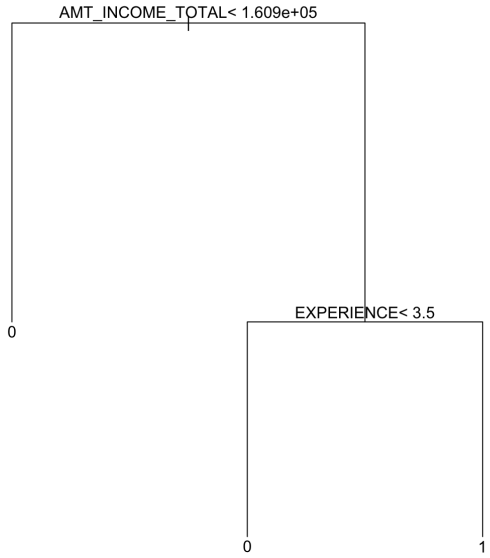

Figure apx 1: Decision tree