Project Title: Loan Approval Prediction: A Data-Driven Approach to Credit Risk

Business Domain: Financial Services

Group Members:

Hessa Khalfan

Nourah Abdulla Alghfeli

Maryam Ali Abdulla

Date of Submission: [24/11/2025]

Table of Contents

1. Introduction and Business Case

1.1 Background and Problem Description

In the highly competitive financial services sector, lending institutions from large banks to specialized finance companies face a fundamental and persistent challenge: managing credit risk. The core of their business model relies on profitably lending money, but every loan issued carries the inherent risk of default, where the borrower fails to repay the debt, leading to significant financial losses. At the same time, being overly cautious is also problematic. A lending strategy that is too strict will lead to rejecting applicants who are actually capable of repaying their loans. This leads to **lost revenue opportunities** and can cause the institution to lose its competitive edge in the market.

The primary decision-maker in this context is the **Credit Risk Manager**. Their daily responsibility is to oversee the approval or rejection of numerous loan applications. The critical decision they must make for each application is:

"Based on the applicant's profile, how likely that they will repay this loan? Should we **approve** this application to generate profit, or **reject** it to avoid a potential loss?"

This decision is crucial because an effective credit risk assessment system directly impacts the institution's profitability, market competitiveness, and operational efficiency. An inaccurate process leads to financial losses and missed opportunities, making it one of the most important challenges to solve in the financial domain.

1.2 Business Objectives

From a business perspective, the main goals of this project are to:

1. **Predict and Minimize Loan Defaults:** The primary objective is to accurately identify applicants with a high probability of defaulting on their loans, allowing the institution to reject these applications and reduce financial losses.

2. **Increase Approval Accuracy for Good Applicants:** The system must also reliably identify applicants who are low risk. This will minimize the rejection of qualified applicants, ensuring the institution capitalizes profitable lending opportunities.

3. **Enhance Decision-Making Efficiency:** To provide the Credit Risk Manager with a data-driven tool that automates the initial screening process, enabling faster and more consistent decisions.

1.3 Data Mining Objectives

To achieve the stated business objectives, the business problem is translated into a specific data mining task:

1. **Develop a Classification Model:** The core technical goal is to build a robust classification model. This model will predict the loan_status for each application, classifying it as either "Approved" or "Rejected" based on the applicant's financial and personal attributes.

2. **Compare Different Algorithms:** To build and evaluate at least two different classification models specifically, a **Decision Tree** and a **Logistic Regression** model to determine which algorithm provides the best performance in terms of accuracy, precision, and recall for this specific business case.

3. **Identify Key Influential Factors:** To use data mining techniques to analyze and identify the most significant factors (e.g., CIBIL score, income, asset value) that influence the loan approval decision, providing actionable insights to the business.

2. Data Understanding and Dataset Description

2.1 Data Source and Origin

- **Dataset Source (Kaggle):** Loan Approval Prediction Dataset
    - **Link:** https://www.kaggle.com/datasets/architsharma01/loan-approval-prediction-dataset
- **Original Owner / Author:** Archit Sharma
- **Number of Records (rows):** 4,269
- **Number of Variables (columns):** 13

2.2 Description of Key Variables (Data Dictionary)

The table below describes the key variables used in this project for predicting loan approval.

Variable Name	Type (Numeric/Categorical)	Description	Role (Input/Target/ID)
loan_id	Categorical (ID)	A unique identifier for each loan application.	ID
loan_status	Categorical	The final outcome of the loan application. This is the variable we aim to predict.	Target
cibil_score	Numeric	The applicant's credit score, indicating their creditworthiness. A higher score is better.	Input
income_annum	Numeric	The applicant's total annual income.	Input

Variable Name	Type (Numeric/Categorical)	Description	Role (Input/Target/ID)
loan_amount	Numeric	The total amount of money requested in the loan.	Input
loan_term	Numeric	The duration of the loan in years.	Input
no_of_dependents	Numeric	The number of dependents the applicant supports financially.	Input
education	Categorical	The applicant's highest education level (Graduate or Not Graduate).	Input
self_employed	Categorical	Indicates if the applicant is self-employed (Yes or No).	Input
residential_assets_value	Numeric	The total value of the applicant's residential assets.	Input
commercial_assets_value	Numeric	The total value of the applicant's commercial assets.	Input
luxury_assets_value	Numeric	The total value of the applicant's luxury assets.	Input
bank_asset_value	Numeric	The total value of the applicant's savings and other bank assets.	Input

2.3 Initial Data Exploration

An initial exploratory data analysis (EDA) was conducted using Tableau to understand the dataset's characteristics and identify preliminary patterns.

- **Target Variable Distribution:** The analysis began by examining the distribution of the *loan_status* variable. The dataset is reasonably balanced, with **62.2%** of applications marked as "**Approved**" and **37.8%** as "**Rejected**." This balance is advantageous as it mitigates the need for specialized techniques to handle class imbalance.

- **Distributions of Key Variables:**

    - **cibil_score:** The histogram of CIBIL scores showed a wide and relatively uniform distribution, indicating that the applicant pool is diverse, with individuals from various credit backgrounds. However, when segmented by loan status, a clear pattern emerged: approved loans had a significantly higher average CIBIL score (approx. 704) compared to rejected loans (approx. 430).

    - **income_annum:** The box plot for annual income showed that the income distribution for both approved and rejected applicants was nearly identical. This surprising finding suggests that income alone is not a primary factor for approval or rejection, but as the scatter plot revealed, it is likely used to determine the maximum allowable loan amount.

- **Data Quality Issues:** The dataset was found to be of high quality with no missing values reported in the key fields, which simplified the data preparation phase. The primary data cleaning step involved the exclusion of the loan_id column, as it serves only as an identifier and has no predictive value.

2.4 Assumptions and Limitations

- **Assumptions:**
    - It is assumed that the data provided is an accurate and representative sample of the financial institution's real world loan applications.
    - We assume that the definitions of the variables (e.g., how "luxury assets" are valued) are consistent across all records.
- **Limitations:**
    - **Lack of Temporal Information:** The dataset lacks any date or time information, which prevents time-series analysis. Consequently, it is not possible to analyze how lending standards or default rates may have evolved over time.
    - **Potential for Hidden Bias:** The model is trained on historical data that reflects past human decisions. There is a risk that this data contains hidden biases, which the model could inadvertently learn and perpetuate in its future predictions.
    - **External Economic Factors:** The model's predictions are based solely on applicant data and do not account for external macroeconomic factors, such as changes in interest rates or economic downturns. These external events can significantly impact a borrower's ability to repay a loan.

3. Data Preparation

Explain in detail how you cleaned, transformed, and reduced your data before modeling.

3.1 Data Cleaning

The initial assessment of the dataset revealed it to be of high quality, requiring minimal cleaning.

- **Missing Values:** A thorough check was conducted on all key variables, and no missing values were found. This eliminated the need for imputation techniques (such as filling with mean, median, or mode).
- **Duplicate Records:** The dataset was checked for duplicate rows, and none were identified.
- **Invalid Values:** All numerical fields (e.g., cibil_score, income_annum) were reviewed for logical consistency. No invalid entries, such as negative values or obvious data entry errors, were detected. The data appeared clean and valid.

3.2 Data Transformation

Data transformation was a necessary step, particularly for the Logistic Regression model, which has specific input requirements.

- **Categorical Variable Encoding:**
    - The dataset contained two key categorical features: education (Graduate/Not Graduate) and self_employed (Yes/No).
    - For the **Logistic Regression** model, which requires all inputs to be numerical, the **Nominal to Numerical** operator in AI Studio was used. This operator performs one-hot encoding, creating new binary columns (dummy variables) for each category (e.g., education = Graduate, self_employed = Yes).

    - The **Decision Tree** model, however, can handle categorical variables natively, so this transformation was not required for that specific workflow.

- **Feature Engineering (Creation of New Features):**
    - To capture the applicant's overall wealth, a new feature named **Total Asset Value** was engineered in Tableau for visualization purposes. This was created by summing the values of the four asset-related columns:

Total Asset Value = bank_asset_value + luxury_asset_value + commercial_asset_value + residential_asset_value

    - This new feature provides a more holistic view of an applicant's financial standing than any single asset column alone.

3.3 Data Reduction / Feature Selection

The primary goal of data reduction is to remove irrelevant or redundant features to improve model performance and reduce complexity.

- **Removal of Irrelevant ID Variable:**
    - The loan_id column was identified as a unique identifier with no predictive value. Including such an attribute can mislead the model and cause overfitting.
    - Therefore, the **Select Attributes** operator in AI Studio was used to explicitly **exclude** the loan_id column from the dataset before it was passed to the modeling algorithms. This decision was based on domain knowledge.

- **Final Feature Set:**
    - After the reduction step, the final set of features used for building the predictive models included all original columns except for loan_id. This

ensured that the models were trained only on meaningful, predictive information.

3.4 Train Validation Test Split

To properly evaluate the performance of the models and ensure they can generalize to new, unseen data, the dataset was partitioned using the **Split Data** operator in AI Studio.

- **Splitting Ratio:** A standard **70/30** split was used.
    - **Training Set (70%):** This larger portion of the data was used to train the models, allowing them to learn the underlying patterns between the input features and the target variable (loan_status).
    - **Testing Set (30%):** This smaller, completely separate portion of the data was held back and used exclusively to evaluate the final performance of the trained models. This provides an unbiased estimate of how the models would perform in a real-world scenario.
- **Sampling Method:** Stratified sampling was used to ensure that the proportion of "*Approved*" and "*Rejected*" applications was the same in both the training and testing sets as it was in the original dataset. This is crucial for preventing bias in the evaluation, especially with the target variable.

4. Data Mining Methodology

4.1 Chosen Methodology (e.g., CRISP-DM / SEMMA)

For this project, we followed the **Cross-Industry Standard Process for Data Mining (CRISP-DM)** methodology. This is a widely adopted and proven framework that breaks down a data mining project into a structured sequence of six iterative phases. In our own words, these phases are:

1. **Business Understanding:** The starting point of any project. It involves understanding the client's objectives and requirements from a business perspective and then converting this knowledge into a data mining problem definition.
2. **Data Understanding:** This phase begins with initial data collection and proceeds with activities to get familiar with the data, identify data quality problems, discover first insights, or detect interesting subsets to form hypotheses.
3. **Data Preparation:** This phase covers all activities to construct the final dataset from the initial raw data. Tasks include cleaning, transforming, and selecting data.
4. **Modeling:** In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values.
5. **Evaluation:** Before proceeding to final deployment, it is crucial to thoroughly evaluate the model and review the steps executed to be certain it properly achieves the business objectives.
6. **Deployment:** The knowledge gained through the model is organized and presented in a way that the business can use. This can range from generating a report to implementing a repeatable data mining process.

4.2 Justification for Methodology

CRISP-DM was the most suitable methodology for our loan approval project for several key reasons:

- **Business-Centric Approach:** CRISP-DM starts with **Business Understanding**, which perfectly aligned with our project's need to first define the problem from the perspective of a Credit Risk Manager. It forced us to translate a business goal (reducing default risk) into a specific data mining task (classification).
- **Iterative Nature:** The framework is not strictly linear; it allows for moving back and forth between phases. For example, during the Modeling phase, we might discover that we need to go back to Data Preparation to create a new feature. This iterative flexibility was essential for refining our approach.
- **Comprehensive and Structured:** It provides a clear, step-by-step roadmap from start to finish, ensuring that no critical step is missed. This structure was invaluable for a project with multiple deliverables, including analysis, modeling, and visualization.

4.3 Mapping Phases to Project Steps

The table below maps the phases of the CRISP-DM methodology to the specific activities we performed in this project.

Phase (CRISP-DM)	What We Did in This Project
Business Understanding	We defined the problem of credit risk for a financial institution. We identified the **Credit Risk Manager** as the key decision-maker and framed their need to accurately approve or reject loan applications. Our business objective was set to **predict loan defaults** to minimize financial loss and maximize revenue.
Data Understanding	We sourced the "Loan Approval Prediction Dataset" from Kaggle. We performed an initial Exploratory Data Analysis (EDA) in Tableau to understand the variables,

Phase (CRISP-DM)	What We Did in This Project
	creating histograms and box plots. We analyzed the distribution of the target variable (loan_status) and identified key relationships, such as the strong correlation between cibil_score and the loan outcome.
Data Preparation	We performed several key preparation steps in AI Studio and Tableau. This included: • **Cleaning:** Verifying that there were no missing or invalid values. • **Transformation:** Creating a Total Asset Value feature and converting categorical variables to numerical for the Logistic Regression model. • **Reduction:** Removing the non-predictive loan_id attribute.
Modeling	We selected **Classification** as our data mining task. We built and trained two different models in AI Studio: a **Decision Tree** and a **Logistic Regression** model. We used a 70/30 train-test split to train the models on the majority of the data.
Evaluation	We evaluated both models on the unseen 30% testing set. We compared their performance using key metrics like **Accuracy, Precision, and Recall**. We analyzed the confusion matrix for each model to understand the business impact of their errors (False Positives vs. False Negatives) and ultimately selected the Decision Tree as the superior model.
Deployment	For the context of this academic project, the "deployment" phase involved organizing and presenting our findings. This included: • Developing a comprehensive **written report** detailing all project steps and results. • Creating an interactive **Tableau Dashboard** and **Storyboard** to visually communicate our key findings to a business audience. • Preparing a **PowerPoint presentation** to deliver the final recommendations.

5. Model Development (Algorithms)

This section details the development of the two predictive models built in AI Studio to classify loan applications.

5.1 Overview of Algorithms Used

Two different algorithms were selected to tackle this classification task, allowing for a robust comparison of their performance and suitability for the business problem.

- **Decision Tree** – Classification
- **Logistic Regression** – Classification

5.2 Algorithm 1: Decision Tree

5.2.1 Purpose and Intuition

The Decision Tree is a powerful and highly interpretable classification algorithm. It works by creating a tree-like model of decisions. In our own words, it learns a series of "if-then" rules from the data by splitting it into smaller and smaller subsets based on the most significant input features. For our problem, it will learn rules like "IF cibil_score is less than 550, THEN Reject" or "IF cibil_score is greater than 750 AND loan_amount is less than 5M, THEN Approve". Its main advantage is its "white-box" nature, meaning its decision-making process is easy to visualize and explain to business stakeholders.

5.2.2 Input and Target Variables

- **Input**

  Features: cibil_score, income_annum, loan_amount, loan_term, education, self_employed, no_of_dependents, and all asset-related columns.

- **Target Variable:** loan_status (Categorical: "Approved" / "Rejected").

5.2.3 Model Settings / Hyperparameters

The default settings for the **Decision Tree** operator in AI Studio were used to establish a baseline model. The key settings were:

- **Criterion:** Gini Index (This is a measure of impurity used to select the best feature for splitting the data at each node).
- **Maximal Depth:** 10 (This parameter limits the number of consecutive rules the tree can create, which helps prevent overfitting).
- **Apply Pruning:** Enabled (This technique automatically removes branches from the tree that provide little predictive power, improving its ability to generalize to new data).

5.2.4 AI Studio Process Design

The workflow for the Decision Tree model in AI Studio was designed as follows:

1. **Data Input:** The loan_approval_dataset was loaded using the Retrieve operator.
2. **Preprocessing:** The Select Attributes operator was used to exclude the loan_id. The Set Role operator defined loan_status as the target (label).
3. **Data Splitting:** The Split Data operator partitioned the data into a 70% training set and a 30% testing set using stratified sampling.
4. **Model Training:** The Decision Tree operator was connected to the training data output.
5. **Model Application & Evaluation:** The trained model was applied to the testing data using the Apply Model operator. The Performance (Classification) operator was then used to evaluate the predictions against the actual values from the test set.

5.2.5 Training Details

The model was trained on the 70% training partition of the dataset. The evaluation was performed on the held-out 30% testing partition to provide an unbiased assessment of its performance on unseen data. The training process was very fast, completing in a matter of seconds on the available hardware.

5.3 Algorithm 2: Logistic Regression

5.3.1 Purpose and Intuition

Logistic Regression is a statistical algorithm used for binary classification. Unlike a Decision Tree which creates rules, Logistic Regression calculates the probability of an outcome occurring. In our own words, it learns a mathematical equation that relates the input features to the probability of a loan being approved. It assigns a "weight" or "coefficient" to each feature, indicating how much that feature influences the final decision and in which direction (positive or negative). It is a powerful and efficient algorithm, often used as a benchmark for classification problems.

5.3.2 Input and Target Variables

- **Input Features:** Same as the Decision Tree, but with categorical variables (education, self_employed) converted into numerical dummy variables.
- **Target Variable:** loan_status (Categorical: "Approved" / "Rejected").

5.3.3 Model Settings / Hyperparameters

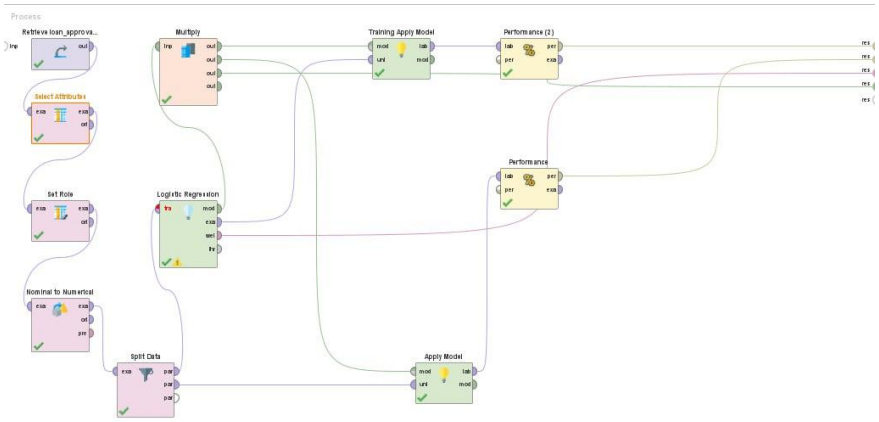
The default settings for the **Logistic Regression** operator in AI Studio were used:

- **Solver:** auto (Allows AI Studio to choose the best optimization algorithm).
- **Regularization:** L2 (A technique used to prevent overfitting by penalizing overly large coefficients).

5.3.4 AI Studio Process Design

The workflow for the Logistic Regression model was similar to the Decision Tree, with one critical addition:

1. **Data Input, Attribute Selection, Role Setting, and Splitting:** These steps were identical to the Decision Tree workflow.
2. **Data Transformation:** A **Nominal to Numerical** operator was added after the Split Data operator (on the training partition) to convert categorical features into a numerical format that the algorithm can process. The same transformation was applied to the testing data.
3. **Model Training:** The Logistic Regression operator was connected to the transformed training data.
4. **Model Application & Evaluation:** The trained model was applied to the transformed testing data, and the Performance (Classification) operator was used for evaluation.



5.3.5 Training Details

Similar to the Decision Tree, the model was trained on the 70% training set and evaluated on the 30% test set. The training process was also extremely fast, completing within a few seconds. The key difference was the necessity of the data transformation step before the model could be trained.

6. Model Evaluation and Comparison

6.1 Evaluation Metrics

To assess the performance of the classification models, the following standard metrics were used, with "Approved" designated as the positive class:

- **Confusion Matrix:** A table that provides a detailed breakdown of predictions into True Positives (correct approvals), True Negatives (correct rejections), False Positives (incorrect approvals, representing financial risk), and False Negatives (incorrect rejections, representing missed opportunities).
- **Accuracy:** The overall percentage of correct predictions. It provides a general measure of model correctness.
- **Precision:** Measures the reliability of positive predictions (Precision = TP / (TP + FP)). High precision minimizes the risk of approving bad loans.
- **Recall (Sensitivity):** Measures the model's ability to identify all actual positive cases (Recall = TP / (TP + FN)). High recall maximizes the capture of creditworthy applicants.

6.2 Results for Each Model

The performance of both models was measured on the unseen 30% testing dataset. The results are summarized in the table below.

Model	Accuracy	Precision (for "Approved")	Recall (for "Approved")
Decision Tree	96.72%	97.37%	97.37%
Logistic Regression	91.02%	94.29%	91.09%

6.3 Model Comparison and Selection

Based on the evaluation results, a direct comparison reveals a clear winner.

- **Performance:** The **Decision Tree model is unequivocally superior** across all key metrics. It achieved a significantly higher Accuracy (96.72% vs. 91.02%) and demonstrated better performance in both Precision and Recall. This indicates it is not only more correct overall but also better at both minimizing risk (high precision) and capturing opportunities (high recall).
- **Interpretability:** As a "white-box" model, the Decision Tree's rule-based logic is transparent and can be easily explained to business stakeholders, which is a major advantage over the less interpretable "black-box" nature of Logistic Regression.
- **Robustness & Overfitting:** The Decision Tree's performance on the testing data (96.72%) was very close to its performance on the training data (99.43%). This small gap suggests that the model generalizes well and is not significantly overfitted, thanks to the pruning mechanisms used.

Selection Justification: The **Decision Tree model is selected as the final, recommended model**. It is chosen not only for its superior predictive performance but also for its high degree of interpretability, which is a critical requirement in a regulated industry like finance where decisions must be justifiable.

6.4 Error Analysis (Optional but Recommended)

An analysis of the confusion matrices for both models provides deeper insights into their business implications.

- **Decision Tree Errors:**
    - **False Positives (21 cases):** The model incorrectly approved 21 risky applications. This represents a direct but manageable **financial risk**.
    - **False Negatives (21 cases):** The model incorrectly rejected 21 creditworthy applicants. This represents a small number of **missed**

business opportunities. The low and balanced number of errors makes this model highly reliable.

- **Logistic Regression Errors:**
    - **False Positives (44 cases):** This model exposed the bank to double the financial risk compared to the Decision Tree.
    - **False Negatives (71 cases):** This is the model's most significant weakness. It failed to identify 71 good applicants, which translates to a substantial loss of potential revenue and could harm customer growth.

This error analysis reinforces the choice of the Decision Tree. Its ability to minimize both types of errors, especially the costly False Negatives, makes it a much more valuable tool for the business.

7. Business Insights and Recommendations

7.1 Key Insights from the Models

1. **CIBIL Score is the Most Dominant Factor:** The applicant's CIBIL score is the single most critical factor in determining the loan outcome. The significant gap in average scores between approved (704) and rejected (430) applicants confirms that credit history is the primary driver of lending decisions.

2. **Income's Role is to Set Limits, Not to Approve:** An applicant's income was not a strong independent predictor of the approval decision. Instead, it appears to function as a ceiling for determining the maximum affordable loan amount, rather than a direct criterion for approval.

3. **Applicant Demographics are Not Decisive:** Factors such as an applicant's education level, employment status (salaried vs. self-employed), and number of dependents had almost no impact on the final loan decision. The approval and rejection rates were nearly identical across these different groups. This suggests the current lending model is almost purely financial and does not significantly weigh personal circumstances.

4. **Total Wealth (Assets) is a Hidden Influential Factor:** While not as dominant as the CIBIL score, our analysis of asset values indicates that applicants with higher total assets have a better chance of approval. This suggests the bank considers an applicant's overall wealth as a secondary indicator of financial stability and a potential safety net.

7.2 Recommendations to Decision Makers

Based on these insights, we provide the following actionable recommendations to the Credit Risk Manager:

1. **Deploy the Decision Tree Model for Automated Pre-Screening:** We recommend integrating the developed Decision Tree model into the initial loan application workflow. It can serve as a highly accurate, automated pre-screening tool.

   Action: Applications with a very high probability of approval can be "fast-tracked" for minimal human review. Applications with a very high probability of rejection can be automatically declined, freeing up valuable analyst time.

2. **Focus Manual Reviews on "Gray Area" Applicants:** Human expertise is most valuable for applications that are not clear-cut.

   Action: Use the model to flag "borderline" cases for example, applicants with a moderate CIBIL score but high income or significant assets. These are the applications where a human analyst should invest their time to make a nuanced judgment.

3. **Refine Marketing and Customer Acquisition Strategies:** The insights show who the bank is currently approving. This information can be used to target the right customers.

   Action: Launch marketing campaigns targeted at individuals with high CIBIL scores, as they are the most likely to be approved. This increases the efficiency of the marketing budget and improves the quality of incoming applications.

4. **Review Business Rules for Demographic Factors:** Since demographic factors currently have no impact, the bank should formalize this.

   Action: Officially state in the lending policy that factors like education and number of dependents are not used as primary decision criteria, which promotes fair and transparent lending practices.

7.3 Limitations and Risks

While the model is powerful, it is crucial to understand its limitations and the risks associated with its use.

- **Risk of Perpetuating Historical Bias:** The model learns from historical data. If past lending decisions contained any hidden human biases (e.g., unconsciously favouring certain profiles), the model will learn and automate these biases. It is a reflection of the past, not necessarily a perfect predictor of the future.
- **Inability to Adapt to Changing Economic Conditions:** The model is static and does not account for external macroeconomic factors. A sudden economic downturn could increase default rates across all CIBIL score ranges, and the model would not be able to predict this without being retrained on new data.
- **The "False Negative" Risk:** The model, while highly accurate, will still incorrectly reject a small number of creditworthy applicants. Over-reliance on the model without any human oversight could lead to a permanent loss of these potential customers.
- **Ethical and Privacy Concerns:** The use of personal data for automated decision-making requires strict adherence to data privacy regulations. The model's decisions must be explainable (which is why the Decision Tree was chosen), and customers should have a right to understand why their application was rejected. The data must be stored securely and used only for its intended purpose.

8. Tableau Dashboard and KPIs

8.1 Dataset for Visualization

The same **Loan Approval Prediction Dataset** used for the data mining part was also used for the visualization in Tableau. This dataset is highly suitable for dashboarding because:

- **Rich in Dimensions and Measures:** It contains a healthy mix of categorical dimensions (like education, self_employed, loan_status) and numerical measures (like cibil_score, income_annum, loan_amount). This allows for the creation of a wide variety of charts and analyses.
- **Clear Business Relevance:** Every variable in the dataset is directly related to the business problem of loan approval, making it easy to define meaningful Key Performance Indicators (KPIs) that a Credit Risk Manager would find valuable.
- **Sufficient Data Volume:** With over 4,000 records, the dataset is large enough to reveal significant patterns and trends when visualized, while still being small enough to allow for fast, interactive performance in Tableau.

8.2 Defined KPIs

To provide a comprehensive overview of the lending process, the following Ten Key Performance Indicators (KPIs) were defined and visualized on the dashboard:

1. **Overall Loan Approval Rate:** Calculates the percentage of approved loans out of the total applications. This is the highest-level indicator of the bank's lending posture (aggressive vs. conservative).
2. **Distribution of CIBIL Scores:** Analyzes the creditworthiness of the entire applicant pool by showing the frequency of applicants across different CIBIL score ranges.

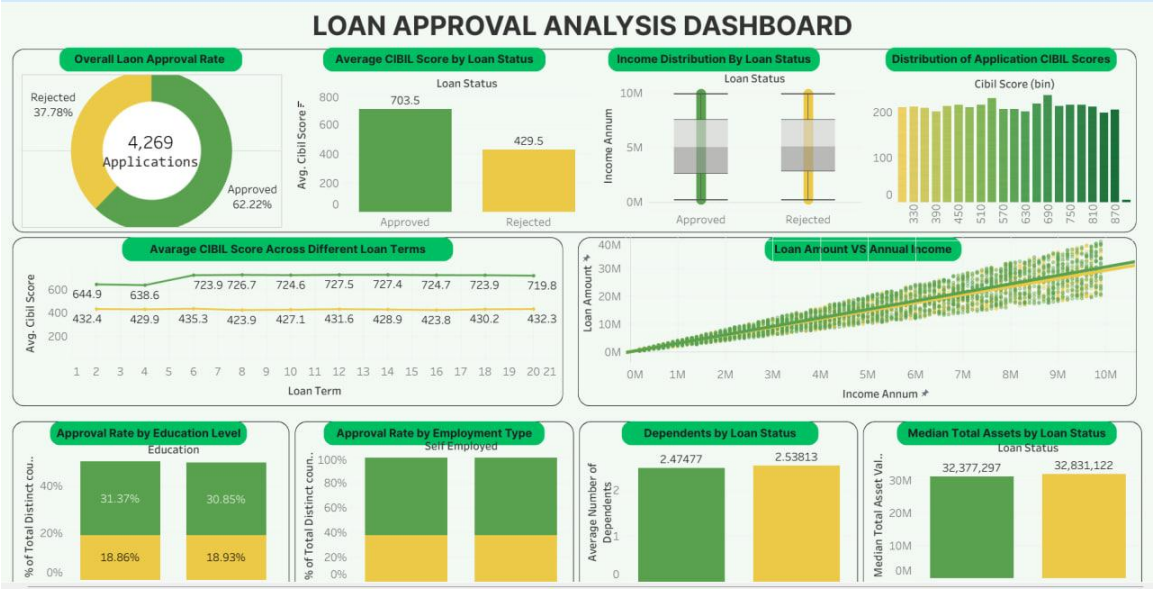
3. **Average CIBIL Score by Loan Status:** Directly compares the average CIBIL score for approved loans versus rejected loans to quantify the impact of credit history.
4. **Impact of Education Level on Approvals:** Examines the approval and rejection counts for graduates versus non-graduates to see if education is a factor.
5. **Impact of Employment Type on Approval Rate:** Compares the approval rates for salaried versus self-employed applicants to determine if one group is favored.
6. **Relationship Between Income and Loan Amount:** Explores the correlation between an applicant's annual income and their requested loan amount to understand affordability rules.
7. **Distribution of Annual Income by Loan Status:** Compares the income ranges of approved versus rejected applicants to see if income is a primary decision factor.
8. **Impact of Loan Term on CIBIL Score:** Investigates if the required CIBIL score for approval changes for loans with different repayment terms.
9. **Average Number of Dependents by Loan Status:** Analyzes if the number of dependents (a proxy for financial burden) influences the loan decision.
10. **Impact of Total Asset Value on Loan Approval:** Calculates the average total asset value for approved versus rejected applicants to determine if overall wealth acts as a secondary factor in the bank's lending assessment.

8.3 Dashboard Design

The dashboard was designed to tell a clear story, with the most important information placed prominently at the top.

- **Layout:** A grid layout was used. The top row contains the most critical, high-level KPIs: the overall approval rate (Donut Chart) and the distribution of applicant credit scores (Histogram). Income Distribution by Loan Status (Box Plot). Distribution of Applicant CIBIL Scores (Histogram) This top-down arrangement allows a decision-maker to grasp the main findings instantly before exploring the secondary factors in the rows below. The most powerful, conclusive evidence he direct comparison of average CIBIL scores is placed second from the left in this top row for immediate impact.
- **Chart Selection:** A variety of charts were used, each chosen for its specific purpose:
- **Donut Chart:** Used for the Overall Loan Approval Rate to provide a quick, high-level percentage view of the total 4,269 applications.
- **Bar Charts:** Used for direct, powerful comparisons, such as Average CIBIL Score by Loan Status and Median Total Assets by Loan Status.
- **Scatter Plot:** Used for Loan Amount VS Annual Income to clearly show the relationship and implicit lending ceiling between these two continuous variables.
- **Box Plot:** Used for Income Distribution by Loan Status to compare the distribution (median, quartiles, range) of income for approved vs. rejected applicants, revealing their similarity.
- **Line Chart:** Used for Average CIBIL Score Across Different Loan Terms to track trends over the duration of the loan.
- **Stacked Bar Charts:** Used for Approval Rate by Education Level and Approval Rate by Employment Type to show the proportional breakdown within each category.

Interactivity: The dashboard is designed to be interactive. A user (e.g., the Credit Risk Manager) can click on a segment in one chart (like the "Approved" bar) and see all other charts on the dashboard filter to show data for that specific segment only, enabling dynamic and deeper analysis.



8.4 Storyboard and Business Interpretation

To guide the decision-maker through the findings in a logical narrative, a Tableau Storyboard was created. The story unfolds across several key steps, each on a separate tab, visually confirming the insights from the data mining model.

1. **Overall Loan Rate:** The story begins with the Donut Chart, answering the first question: "What is our overall loan approval rate?" The accompanying text explains: *"Out of 4,269 applications, 62% were approved and 38% were rejected. This shows the bank approves most loans but still rejects a significant portion to manage risk."* This sets the high-level context.

2. **Average CIBIL Score by Loan:** The next and most critical step focuses on the Bar Chart showing the huge gap in average CIBIL scores. The narrative is direct: *"Approved applicants have a CIBIL score of 703, while rejected ones have 429. This 274-point gap shows that credit score is the main reason for approval or rejection."* This visually confirms the model's most important finding.

3. **Loan Amount VS Annual Income:** The story then moves to the Scatter Plot to clarify the role of income. The interpretation is concise: *"People who earn more get approved for larger loans. The bank checks if the loan amount matches your*

income." This explains that income's role is to set a limit, not to decide the outcome.

4. **Debunking Myths (Demographics):** The subsequent steps group the charts for education, employment, and dependents to test common assumptions.
    - **Education Level:** The stacked bar chart shows nearly identical approval/rejection splits. The text notes: *"More educated people have slightly better approval chances, but it's not the main factor."*
    - **Employment Type:** This chart shows a slight advantage for salaried employees. The text states: *"Salaried employees get approved more often than self-employed people. Stable jobs = easier approval."*
    - **Dependents:** The bar chart shows almost no difference. The narrative confirms: *"Both approved and rejected applicants have about 2.5 dependents on average. Number of dependents doesn't significantly affect approval."*
5. **Secondary Financial Factors:** The final steps explore other financial metrics.
    - **Median Total Assets:** The bar chart shows similar asset values for both groups. The text concludes: *"Approved and rejected applicants have similar asset values. Having assets doesn't guarantee approval — the bank cares more about your income and credit score."*
    - **CIBIL Score Across Loan Term:** The line chart shows two flat lines. The narrative is clear: *"The required CIBIL score stays about the same whether you borrow for 1 year or 21 years. Loan duration doesn't change credit requirements."*

This storyboard complements the data mining model by visually confirming its findings. The model identified cibil_score as the most important feature, and the Tableau story makes it visually undeniable why that is the case, while also clarifying the secondary roles of other factors.

8.5 Tableau Public Link

https://public.tableau.com/views/FinalGraph_17638895054630/Story1?:language=en-US&publish=yes&:sid=&:redirect=auth&:display_count=n&:origin=viz_share_link

9. Conclusion and Future Work

9.1 Summary of Work

This project successfully addressed the critical business problem of credit risk assessment for a financial institution. Starting with the goal of helping a Credit Risk Manager make more accurate and efficient lending decisions, we followed the CRISP-DM methodology to execute a comprehensive data mining project.

We utilized a **Loan Approval Prediction dataset** from Kaggle, which we prepared by cleaning, transforming, and selecting the most relevant features. We then developed and compared two classification models in AI Studio: a **Decision Tree** and a **Logistic Regression** model. The evaluation revealed that the **Decision Tree model was superior**, achieving an outstanding **accuracy of 96.72%** on the unseen test data.

The key business insight derived from both the model and an interactive **Tableau dashboard** was that the **CIBIL score is the single most dominant factor** driving loan approvals. Other factors like income were found to set lending limits, while demographic attributes had minimal impact. The project concluded by providing actionable recommendations, including the deployment of the Decision Tree model as an automated pre-screening tool.

9.2 What Worked Well

- **High-Performing and Interpretable Model:** The Decision Tree model not only delivered excellent predictive accuracy but was also highly interpretable. Its "white-box" nature makes it easy to explain and justify to business stakeholders, which is a major strength.
- **Synergy Between Modeling and Visualization:** The findings from the data mining model were strongly supported and clearly illustrated by the Tableau

dashboard. The model identified the importance of the CIBIL score mathematically, and the visualizations made this insight visually undeniable, creating a powerful and cohesive narrative.

- **Actionable Business Insights:** The project went beyond technical results to produce clear, actionable recommendations. The insights about the roles of CIBIL score, income, and demographic factors can be used immediately to refine business strategy, from marketing to operational workflows.

9.3 Possible Improvements / Future Work

While this project was successful, there are several avenues for future work that could further enhance the credit risk assessment process:

1. **Incorporate More Sophisticated Data:**

Transactional Data: Integrating an applicant's transaction history with the bank could provide powerful behavioural insights (e.g., saving habits, payment history) that are not available in the current dataset.

External Economic Data: Adding macroeconomic indicators (e.g., unemployment rate, inflation) could help make the model more robust and adaptable to changing economic climates.

2. **Advanced Feature Engineering:**

Future work could focus on creating more complex features, such as a **debt-to-income ratio** (loan_amount / income_annum) or an **assets-to-loan ratio** (Total Asset Value / loan_amount). These engineered features often have more predictive power than raw variables alone.

3. **Explore More Advanced Models:**

While the Decision Tree performed well, more advanced ensemble models like **Random Forest** or **Gradient Boosting (e.g., XGBoost)** could be explored. These models often

provide a slight edge in accuracy, although it may come at the cost of reduced interpretability.

4. **Real-World Deployment and Monitoring:**

The next logical step would be to move towards a real-world deployment. This would involve A/B testing the model's recommendations against human decisions, and implementing a **monitoring system** to track the model's performance over time and trigger alerts if its accuracy begins to degrade (a concept known as "model drift").

10. Team Collaboration and Individual Contribution

The successful completion of this project was the direct result of a structured, collaborative, and communicative team effort. Our group established a clear strategy from the outset to ensure all project requirements were met efficiently and to the highest possible standard.

10.1. Collaboration Strategy

Our teamwork was founded on three core principles:

1. **Strategic Task Allocation:** We began the project by collectively reviewing the project brief and breaking down the deliverables into distinct modules. Tasks were then assigned to team members based on their individual skills and strengths, ensuring that each component was handled by the person best equipped for it. This allowed for parallel work streams and efficient progress.

2. **Consistent Communication and Synchronization:** We maintained constant communication using a dedicated WhatsApp group for quick updates and file sharing via word for collaborative writing. Weekly sync-up meetings were held to review progress, provide peer feedback, and collaboratively solve any roadblocks. This agile approach ensured the team remained aligned and that any challenges were addressed promptly.

3. **Integrated Peer Review Process:** A "no-silo" policy was adopted, meaning no task was considered complete until it was reviewed by the other team members. Models built in AI Studio were cross-checked, Tableau visualizations were evaluated for clarity, and report sections were proofread by the entire group. This iterative feedback loop was crucial for ensuring quality, consistency, and accuracy across all project deliverables.

10.2. Individual Contributions

While the project was a unified effort, each member made significant contributions in their assigned roles:

- **Hessa Khalfan:** Hessa Khalfan played a crucial dual role, bridging the technical modeling with the final visual output. She was responsible for the entire data mining lifecycle in AI Studio, where she designed the process workflows and built both the Decision Tree and Logistic Regression models. In Tableau, she created the high-level summary charts, including the **Donut Chart** Overall Loan Approval Rate and the **Histogram** Distribution of Applicant CIBIL Scores, **Bar Chart** Average Number of Dependents by Loan Status, and took the lead on assembling and organizing the final interactive dashboard. She also reviewed and proofread the written content in Word, contributed to the publication tasks, helped in the PowerPoint preparation.

- **Nourah Abdulla:** handled the creation of several core analytical charts that formed the backbone of our visual story. She developed the **Line Graph** to analyze CIBIL scores over time, the **Scatter Plot** to reveal the relationship between income and loan amount, and the **Bar Chart** that highlighted the impact of CIBIL scores and the **Bar Chart** for the total asset values on loan approval. She also organized and formatted the Word document in a clear and professional manner and created the storyboard that guided the structure and flow of the final visual narrative. In addition, she served as the documentation lead and assisted in preparing PowerPoint.

- **Maryam Ali**: was responsible for creating several advanced visualizations in Tableau to compare variables. Her key contributions included developing the **initial Stacked Bar Chart** Approval Rate by Education Level, the more advanced **100% Stacked Bar Chart** Approval Rate by Employment Type, for fair rate comparisons, and the insightful **Box Plot** Income Distribution by Loan Status, to analyze income distribution. She also created the storyboard and assisted in building the dashboard. In addition, she designed PowerPoint, helped format the Word document and tables, and contributed to organizing the team's meetings.

11. References

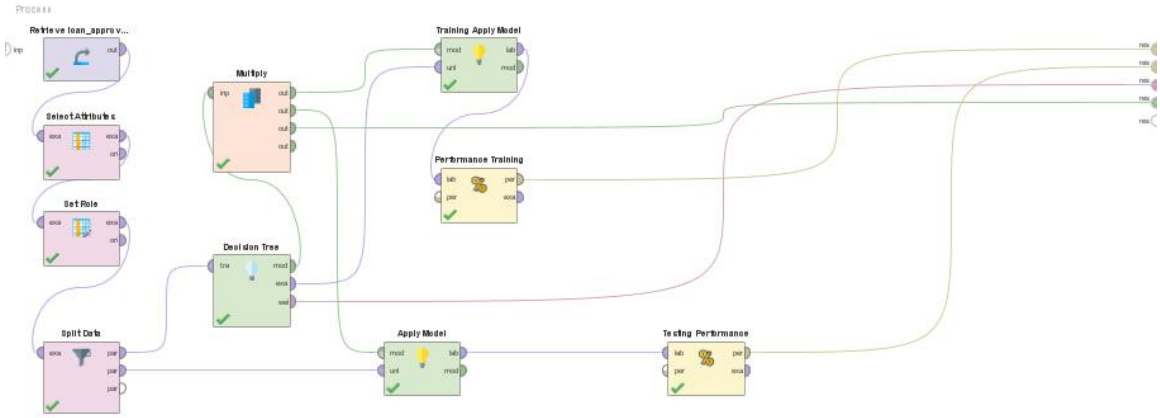
[List all external sources using a consistent style (e.g., APA/IEEE): Kaggle dataset, academic papers, books, websites, documentation for AI Studio and Tableau, etc.]

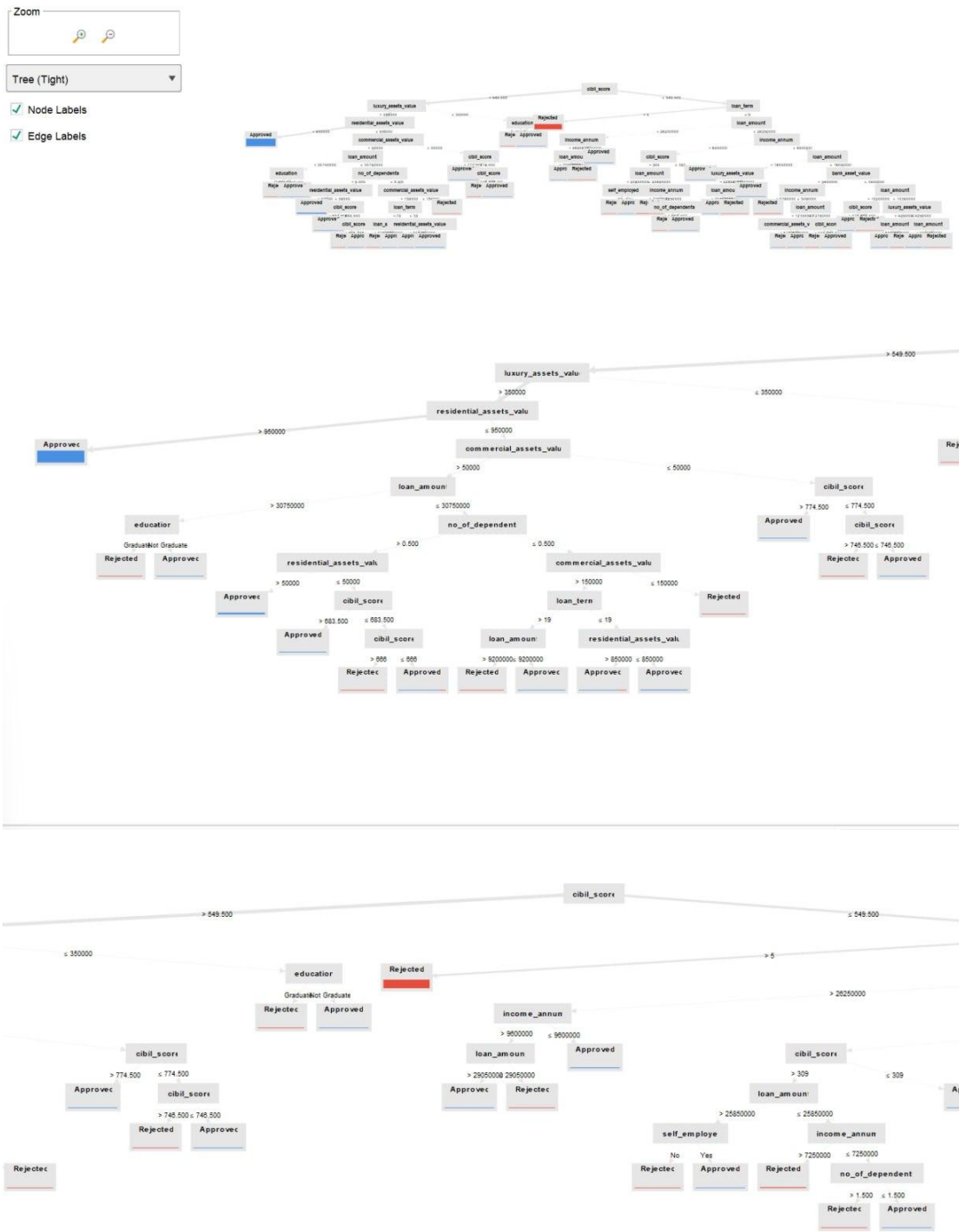
1- **architsharma01. (2025).** ***Loan Approval Prediction Dataset*** **[Data set]. Kaggle. https://www.kaggle.com/datasets/architsharma01/loan-approval-prediction-dataset**

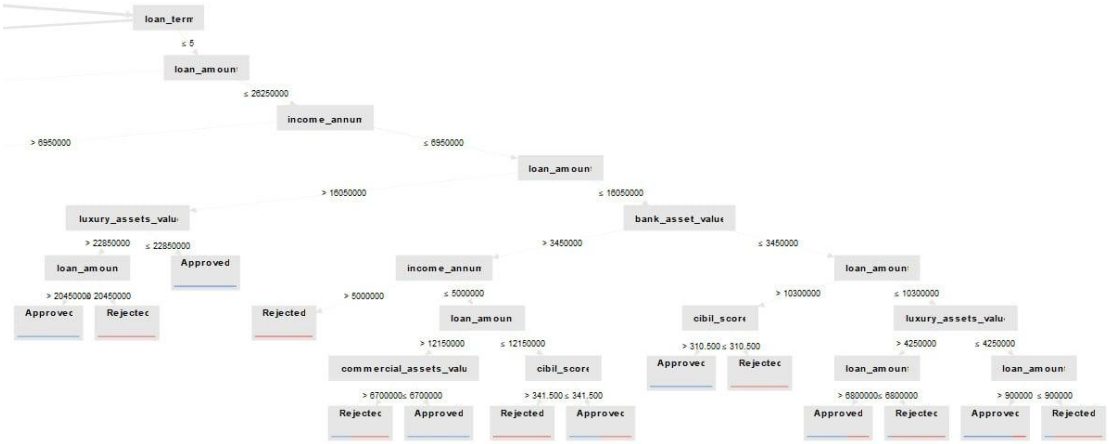
12. Appendices

• Appendix A: AI Studio process diagrams and screenshots

A.1 Decision tree



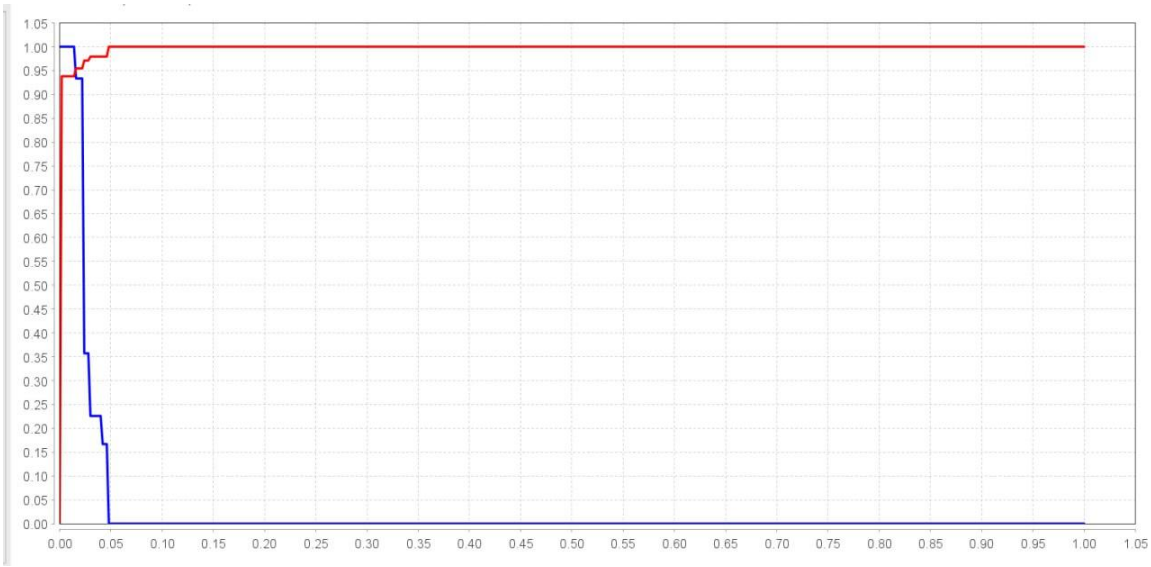

Zoom
Tree (Tight)
Node Labels
Edge Labels
luxury_assets_value
residential_assets_value
commercial_assets_value
loan_amount
education
no_of_dependents
cibil_score
loan_term
income_annum
self_employed
bank_asset_value
Approved
Rejected
Not Graduate
Graduate
Yes
No



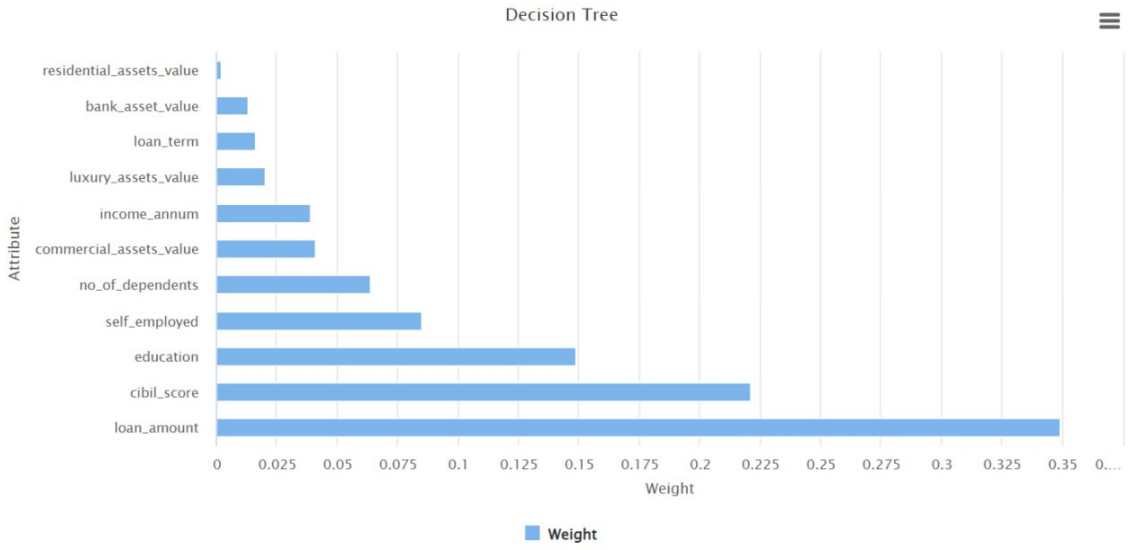
Additional evaluation tables, confusion matrices, ROC curves

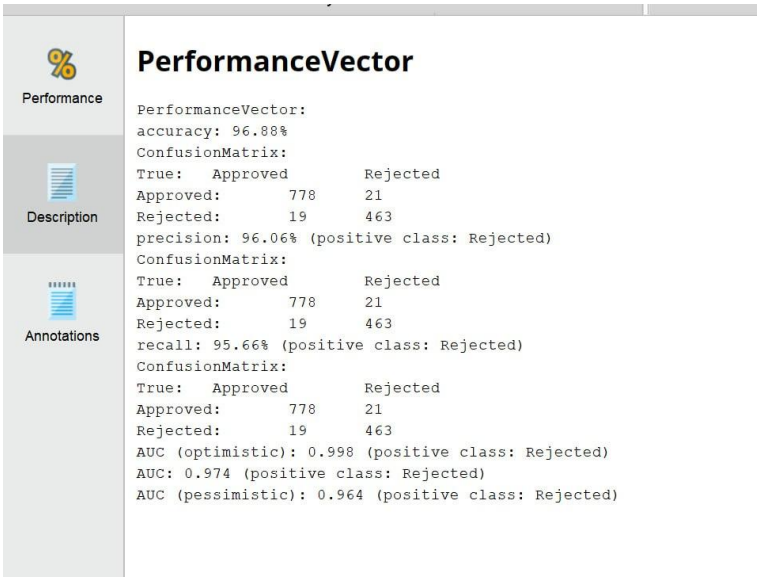
accuracy: 96.88%

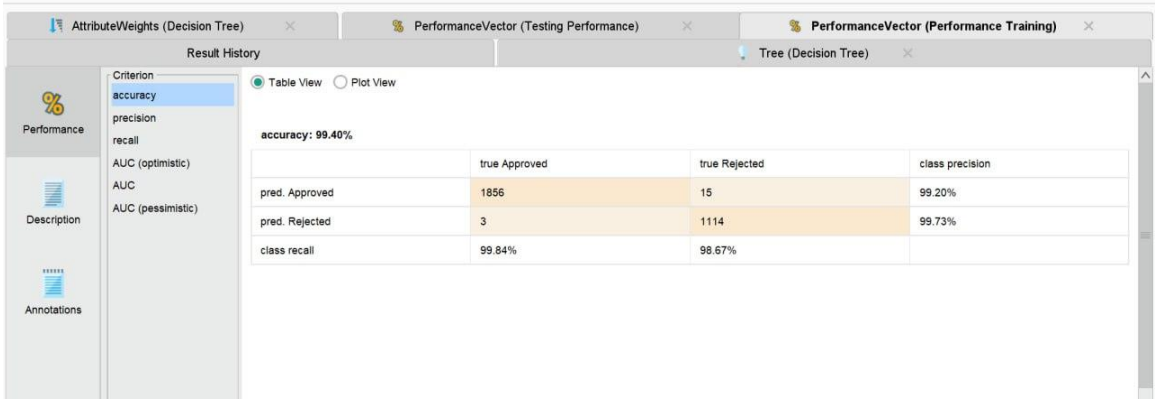
	true Approved	true Rejected	class precision
pred. Approved	778	21	97.37%
pred. Rejected	19	463	96.06%
class recall	97.62%	95.66%	



attribute	weight
education	0.149
income_...	0.039
bank_as...	0.013
commerc...	0.041
loan_term	0.016
self_emp...	0.085
residenti...	0.002
loan_am...	0.349
cibil_score	0.221
no_of_d...	0.064
luxury_a...	0.020




Performance
Description
Annotations
PerformanceVector
PerformanceVector:
accuracy: 96.88%
ConfusionMatrix:
True: Approved Rejected
Approved: 778 21
Rejected: 19 463
precision: 96.06% (positive class: Rejected)
ConfusionMatrix:
True: Approved Rejected
Approved: 778 21
Rejected: 19 463
recall: 95.66% (positive class: Rejected)
ConfusionMatrix:
True: Approved Rejected
Approved: 778 21
Rejected: 19 463
AUC (optimistic): 0.998 (positive class: Rejected)
AUC: 0.974 (positive class: Rejected)
AUC (pessimistic): 0.964 (positive class: Rejected)


AttributeWeights (Decision Tree)
PerformanceVector (Testing Performance)
PerformanceVector (Performance Training)
Result History
Tree (Decision Tree)
Performance
Description
Annotations
Criterion
accuracy
precision
recall
AUC (optimistic)
AUC
AUC (pessimistic)
Table View
Plot View
accuracy: 99.40%
true Approved
true Rejected
class precision
pred. Approved
1856
15
99.20%
pred. Rejected
3
1114
99.73%
class recall
99.84%
98.67%

Tree

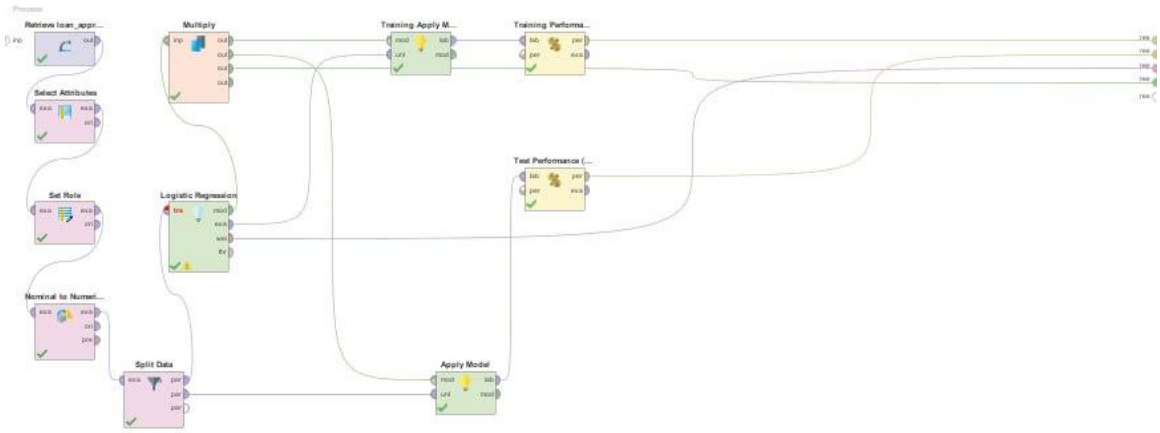
```
cibil_score > 549.500
|   luxury_assets_value > 350000
|   |   residential_assets_value > 950000: Approved {Approved=1513, Rejected=0}
|   |   residential_assets_value ≤ 950000
|   |   |   commercial_assets_value > 50000
|   |   |   |   loan_amount > 30750000
|   |   |   |   |   education = Graduate: Rejected {Approved=0, Rejected=1}
|   |   |   |   |   education = Not Graduate: Approved {Approved=3, Rejected=0}
|   |   |   |   loan_amount ≤ 30750000
|   |   |   |   |   no_of_dependents > 0.500
|   |   |   |   |   |   residential_assets_value > 50000: Approved {Approved=151, Rejected=0}
|   |   |   |   |   |   residential_assets_value ≤ 50000
|   |   |   |   |   |   |   cibil_score > 683.500: Approved {Approved=13, Rejected=0}
|   |   |   |   |   |   |   cibil_score ≤ 683.500
|   |   |   |   |   |   |   |   cibil_score > 666: Rejected {Approved=0, Rejected=1}
|   |   |   |   |   |   |   |   cibil_score ≤ 666: Approved {Approved=6, Rejected=1}
|   |   |   |   |   no_of_dependents ≤ 0.500
|   |   |   |   |   |   commercial_assets_value > 150000
|   |   |   |   |   |   |   loan_term > 19
|   |   |   |   |   |   |   |   loan_amount > 9200000: Rejected {Approved=0, Rejected=1}
|   |   |   |   |   |   |   |   loan_amount ≤ 9200000: Approved {Approved=3, Rejected=0}
|   |   |   |   |   |   |   loan_term ≤ 19
|   |   |   |   |   |   |   |   residential_assets_value > 850000: Approved {Approved=5, Rejected=1}
```

```
|   |   |   |   |   no_of_dependents ≤ 0.500
|   |   |   |   |   |   commercial_assets_value > 150000
|   |   |   |   |   |   |   loan_term > 19
|   |   |   |   |   |   |   |   loan_amount > 9200000: Rejected {Approved=0, Rejected=1}
|   |   |   |   |   |   |   |   loan_amount ≤ 9200000: Approved {Approved=3, Rejected=0}
|   |   |   |   |   |   |   loan_term ≤ 19
|   |   |   |   |   |   |   |   residential_assets_value > 850000: Approved {Approved=5, Rejected=1}
|   |   |   |   |   |   |   |   residential_assets_value ≤ 850000: Approved {Approved=30, Rejected=0}
|   |   |   |   |   |   commercial_assets_value ≤ 150000: Rejected {Approved=0, Rejected=1}
|   |   |   commercial_assets_value ≤ 50000
|   |   |   |   cibil_score > 774.500: Approved {Approved=6, Rejected=0}
|   |   |   |   cibil_score ≤ 774.500
|   |   |   |   |   cibil_score > 746.500: Rejected {Approved=0, Rejected=2}
|   |   |   |   |   cibil_score ≤ 746.500: Approved {Approved=4, Rejected=0}
|   luxury_assets_value ≤ 350000
|   |   education = Graduate: Rejected {Approved=0, Rejected=1}
|   |   education = Not Graduate: Approved {Approved=1, Rejected=0}
cibil_score ≤ 549.500
|   loan_term > 5: Rejected {Approved=0, Rejected=1010}
|   loan_term ≤ 5
|   |   loan_amount > 26250000
|   |   |   income_annum > 9600000
|   |   |   |   loan_amount > 29050000: Approved {Approved=3, Rejected=0}
|   |   |   |   loan_amount ≤ 29050000: Rejected {Approved=0, Rejected=3}
|   |   |   income_annum ≤ 9600000: Approved {Approved=26, Rejected=0}
|   |   loan_amount ≤ 26250000
|   |   |   income_annum > 6950000
```

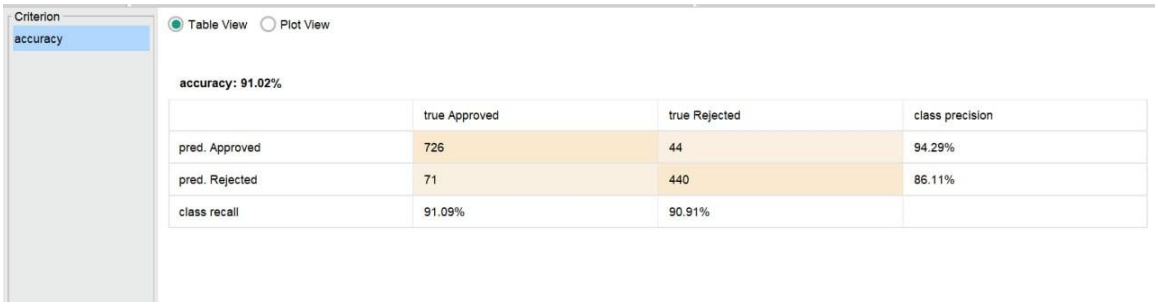
```
|   |   loan_amount ≤ 26250000
|   |   |   income_annum > 6950000
|   |   |   |   cibil_score > 309
|   |   |   |   |   loan_amount > 25850000
|   |   |   |   |   |   self_employed = No: Rejected {Approved=0, Rejected=1}
|   |   |   |   |   |   self_employed = Yes: Approved {Approved=1, Rejected=0}
|   |   |   |   |   loan_amount ≤ 25850000
|   |   |   |   |   |   income_annum > 7250000: Rejected {Approved=0, Rejected=35}
|   |   |   |   |   |   income_annum ≤ 7250000
|   |   |   |   |   |   |   no_of_dependents > 1.500: Rejected {Approved=0, Rejected=3}
|   |   |   |   |   |   |   no_of_dependents ≤ 1.500: Approved {Approved=1, Rejected=0}
|   |   |   |   cibil_score ≤ 309: Approved {Approved=1, Rejected=0}
|   |   |   income_annum ≤ 6950000
|   |   |   |   loan_amount > 16050000
|   |   |   |   |   luxury_assets_value > 22850000
|   |   |   |   |   |   loan_amount > 20450000: Approved {Approved=6, Rejected=0}
|   |   |   |   |   |   loan_amount ≤ 20450000: Rejected {Approved=0, Rejected=4}
|   |   |   |   |   luxury_assets_value ≤ 22850000: Approved {Approved=27, Rejected=0}
|   |   |   |   loan_amount ≤ 16050000
|   |   |   |   |   bank_asset_value > 3450000
|   |   |   |   |   |   income_annum > 5000000: Rejected {Approved=0, Rejected=17}
|   |   |   |   |   |   income_annum ≤ 5000000
|   |   |   |   |   |   |   loan_amount > 12150000
|   |   |   |   |   |   |   |   commercial_assets_value > 6700000: Rejected {Approved=1, Rejected=2}
|   |   |   |   |   |   |   |   commercial_assets_value ≤ 6700000: Approved {Approved=6, Rejected=0}
|   |   |   |   |   |   |   loan_amount ≤ 12150000
|   |   |   |   |   |   |   |   cibil_score > 341.500: Rejected {Approved=0, Rejected=12}
```

```
|   |   |   |   loan_amount > 16050000
|   |   |   |   |   luxury_assets_value > 22850000
|   |   |   |   |   |   loan_amount > 20450000: Approved {Approved=6, Rejected=0}
|   |   |   |   |   |   loan_amount ≤ 20450000: Rejected {Approved=0, Rejected=4}
|   |   |   |   |   luxury_assets_value ≤ 22850000: Approved {Approved=27, Rejected=0}
|   |   |   |   loan_amount ≤ 16050000
|   |   |   |   |   bank_asset_value > 3450000
|   |   |   |   |   |   income_annum > 5000000: Rejected {Approved=0, Rejected=17}
|   |   |   |   |   |   income_annum ≤ 5000000
|   |   |   |   |   |   |   loan_amount > 12150000
|   |   |   |   |   |   |   |   commercial_assets_value > 6700000: Rejected {Approved=1, Rejected=2}
|   |   |   |   |   |   |   |   commercial_assets_value ≤ 6700000: Approved {Approved=6, Rejected=0}
|   |   |   |   |   |   |   loan_amount ≤ 12150000
|   |   |   |   |   |   |   |   cibil_score > 341.500: Rejected {Approved=0, Rejected=12}
|   |   |   |   |   |   |   |   cibil_score ≤ 341.500: Approved {Approved=1, Rejected=1}
|   |   |   |   |   bank_asset_value ≤ 3450000
|   |   |   |   |   |   loan_amount > 10300000
|   |   |   |   |   |   |   cibil_score > 310.500: Approved {Approved=16, Rejected=0}
|   |   |   |   |   |   |   cibil_score ≤ 310.500: Rejected {Approved=0, Rejected=1}
|   |   |   |   |   |   loan_amount ≤ 10300000
|   |   |   |   |   |   |   luxury_assets_value > 4250000
|   |   |   |   |   |   |   |   loan_amount > 6800000: Approved {Approved=9, Rejected=5}
|   |   |   |   |   |   |   |   loan_amount ≤ 6800000: Rejected {Approved=1, Rejected=14}
|   |   |   |   |   |   |   luxury_assets_value ≤ 4250000
|   |   |   |   |   |   |   |   loan_amount > 900000: Approved {Approved=24, Rejected=7}
|   |   |   |   |   |   |   |   loan_amount ≤ 900000: Rejected {Approved=1, Rejected=5}
```

A.2 Logistic Regression



attribute	weight
educatio...	-0.088
educatio...	0
self_emp...	0.082
self_emp...	0
no_of_d...	-0.004
income_...	1.538
loan_am...	-1.239
loan_term	0.904
cibil_score	-4.269
residenti...	-0.025
commerc...	-0.077
luxury_a...	-0.248
bank_as...	-0.075

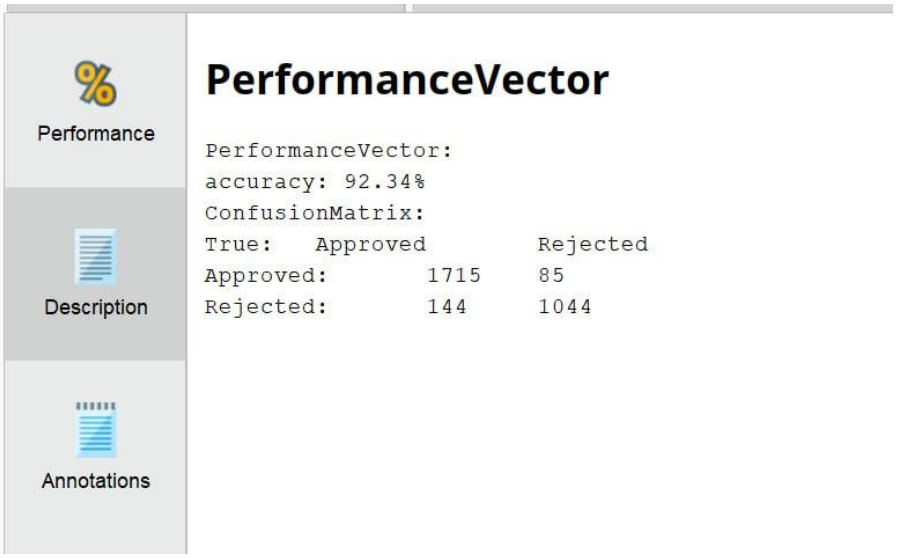


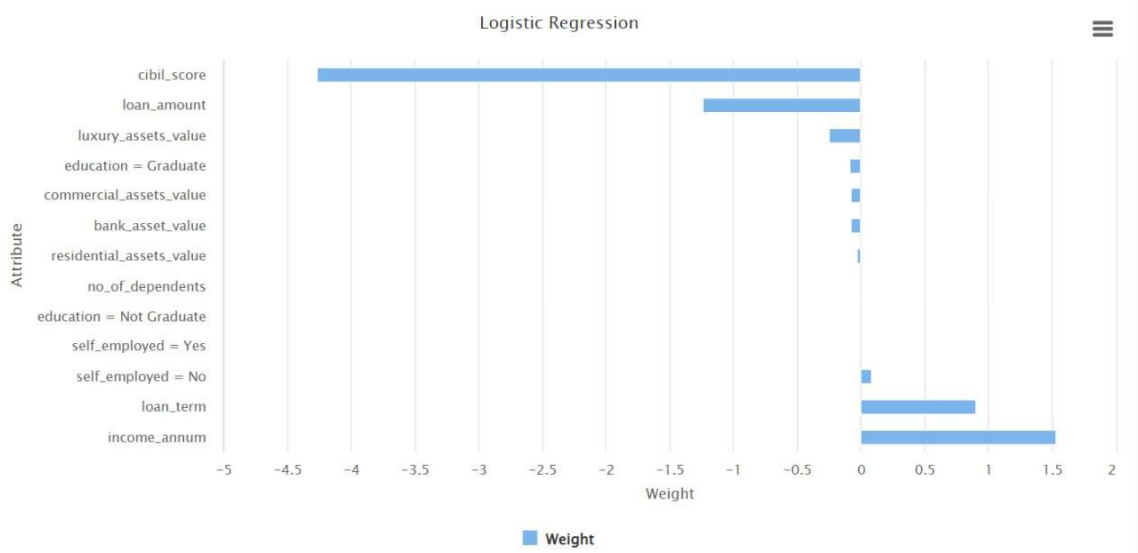
	true Approved	true Rejected	class precision
pred. Approved	726	44	94.29%
pred. Rejected	71	440	86.11%
class recall	91.09%	90.91%	

Figure 1: Test Performance (Classification)

	true Approved	true Rejected	class precision
pred. Approved	1715	85	95.28%
pred. Rejected	144	1044	87.88%
class recall	92.25%	92.47%	

Figure 2: Training Performance (Classification)





Logistic Regression Model

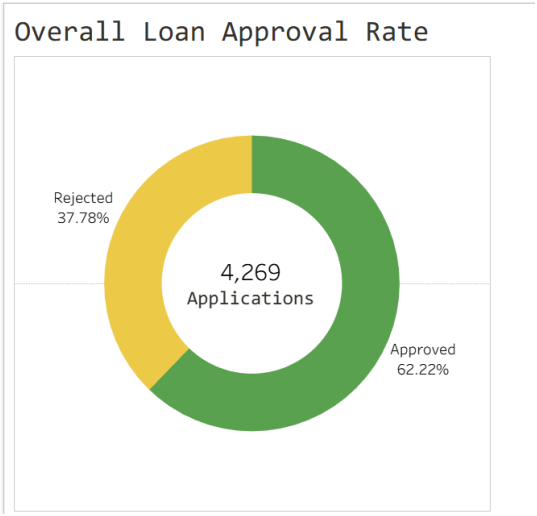
```
Warning:
Removed collinear columns [education = Not Graduate, self_employed = Yes]

Model Metrics Type: BinomialGLM
 Description: N/A
 model id: rm-h2o-model-logistic_regression-1
 frame id: rm-h2o-frame-logistic_regression-1
 MSE: 0.06218542
 RMSE: 0.24937005
 R^2: 0.7354691
 AUC: 0.9685944
 pr_auc: 0.94152075
 logloss: 0.22117335
 mean_per_class_error: 0.07646949
 default threshold: 0.4284253716468811
 CM: Confusion Matrix (Row labels: Actual class; Column labels: Predicted class):
          Approved  Rejected   Error         Rate
Approved      1713       146  0.0785  146 / 1,859
Rejected        84      1045  0.0744   84 / 1,129
  Totals      1797      1191  0.0770  230 / 2,988
Gains/Lift Table (Avg response rate: 37.78 %, avg score: 29.04 %):
  Group  Cumulative Data Fraction  Lower Threshold      Lift  Cumulative Lift  Response Rate     Score  Cumulative Response Rate  Cumulative Sc
      1                0.01004016         0.471446  0.088220         0.088220       0.033333  0.506224                  0.033333          0.506
      2                0.02008032         0.458733  0.264659         0.176439       0.100000  0.466107                  0.066667          0.486
```

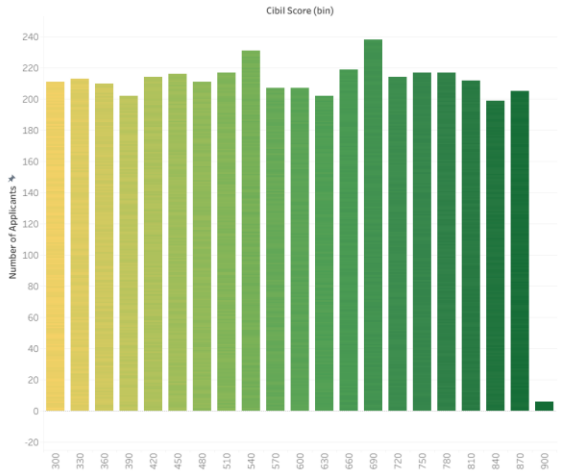
```
CM: Confusion Matrix (Row labels: Actual class; Column labels: Predicted class):
          Approved  Rejected   Error         Rate
Approved      1713       146  0.0785  146 / 1,859
Rejected        84      1045  0.0744   84 / 1,129
  Totals      1797      1191  0.0770  230 / 2,988
Gains/Lift Table (Avg response rate: 37.78 %, avg score: 29.04 %):
   Group  Cumulative Data Fraction  Lower Threshold      Lift  Cumulative Lift  Response Rate     Score  Cumulative Response Rate  Cumulative Sc
       1                0.01004016         0.471446  0.088220         0.088220       0.033333  0.506224                  0.033333          0.506
       2                0.02008032         0.458733  0.264659         0.176439       0.100000  0.466107                  0.066667          0.486
       3                0.03012048         0.446789  0.088220         0.147033       0.033333  0.452042                  0.055556          0.474
       4                0.04016064         0.432988  0.264659         0.176439       0.100000  0.439153                  0.066667          0.465
       5                0.05020080         0.423063  0.529318         0.247015       0.200000  0.427314                  0.093333          0.458
       6                0.10006693         0.396641  0.444059         0.345207       0.167785  0.408885                  0.130435          0.433
       7                0.15026774         0.374527  0.476386         0.389031       0.180000  0.385150                  0.146993          0.417
       8                0.20013387         0.359528  0.568395         0.433722       0.214765  0.366597                  0.163880          0.404
       9                0.30020080         0.329008  0.637306         0.501583       0.240803  0.343170                  0.189521          0.384
      10                0.39993307         0.303351  0.905880         0.602404       0.342282  0.315429                  0.227615          0.367
      11                0.50000000         0.282454  1.017919         0.685562       0.384615  0.292621                  0.259036          0.352
      12                0.60006693         0.261298  1.186097         0.769031       0.448161  0.271742                  0.290574          0.338
      13                0.69979920         0.239913  1.145671         0.822708       0.432886  0.250994                  0.310856          0.326
      14                0.79986613         0.217855  1.292315         0.881458       0.488294  0.228968                  0.333054          0.314
      15                0.89993307         0.198173  1.371978         0.936001       0.518395  0.208120                  0.353663          0.302
      16                1.00000000         0.146892  1.575562         1.000000       0.595318  0.183203                  0.377845          0.290
 null DOF: 2987.0
 residual DOF: 2976.0
 null deviance: 3962.0828
```

```
 null DOF: 2987.0
 residual DOF: 2976.0
 null deviance: 3962.0828
 residual deviance: 1321.7319
Variable Importances:
                Variable Relative Importance Scaled Importance Percentage
             cibil_score            4.269398          1.000000   0.499533
            income_annum            1.537754          0.360180   0.179922
             loan_amount            1.238731          0.290142   0.144935
               loan_term            0.903957          0.211729   0.105766
     luxury_assets_value            0.247726          0.058024   0.028985
    education = Graduate            0.086472          0.020254   0.010118
      self_employed = No            0.081887          0.019180   0.009581
 commercial_assets_value            0.077280          0.018101   0.009042
        bank_asset_value            0.074799          0.017520   0.008752
residential_assets_value            0.024856          0.005822   0.002908
         no_of_dependents            0.003922          0.000919   0.000459
education = Not Graduate            0.000000          0.000000   0.000000
     self_employed = Yes            0.000000          0.000000   0.000000
GLM Model (summary):
   Family  Link Regularization Number of Predictors Total Number of Active Predictors Number of Iterations                    Training Frame
 binomial logit           None                   13                             11                     7 rm-h2o-frame-logistic_regression-1
Scoring History:
```

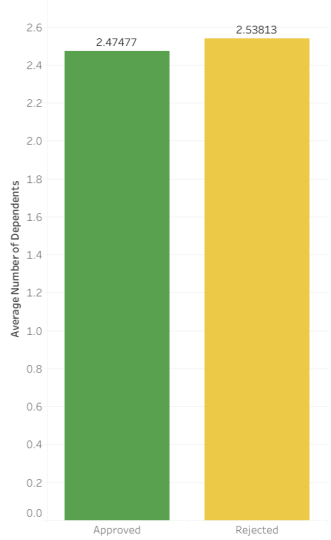
• Appendix C: Extra Tableau views and screenshots



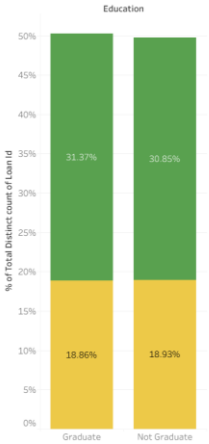
Distribution of Applicant CIBIL Scores



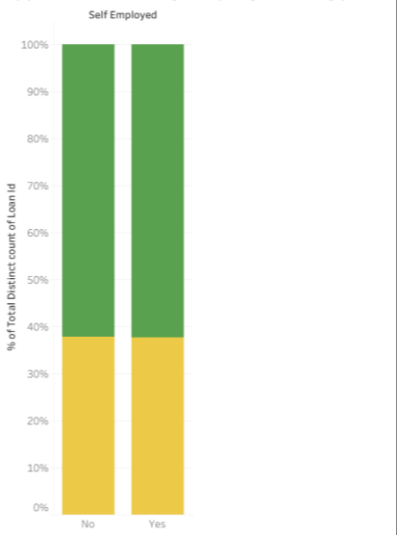
Average Number of Dependents by Loan Status



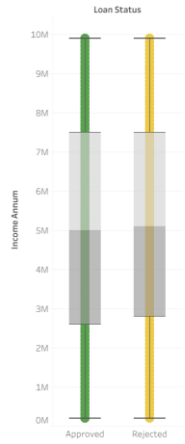
Approval Rate by Education Level



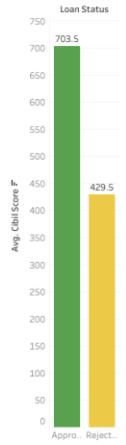
Approval Rate by Employment Type

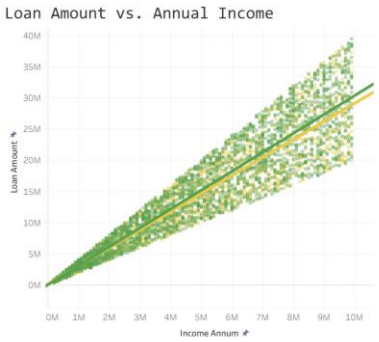


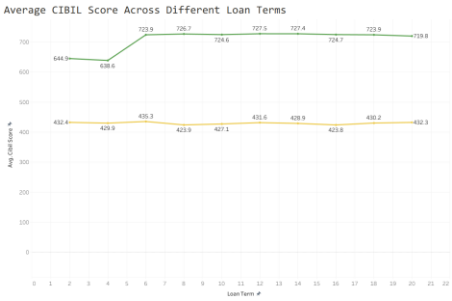
Income Distribution by Loan Status

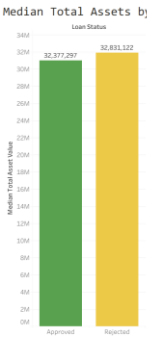


Average CIBIL Score by Loan Status

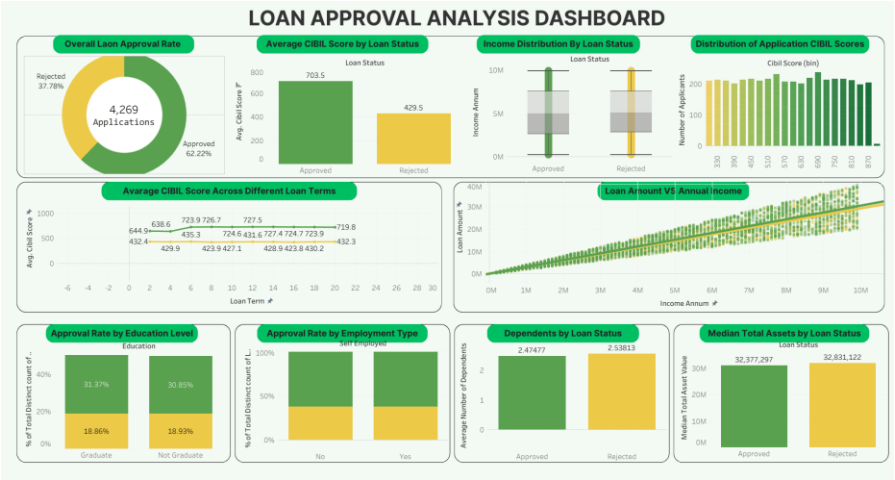



Loan Amount vs. Annual Income
Loan Amount
Income Annum

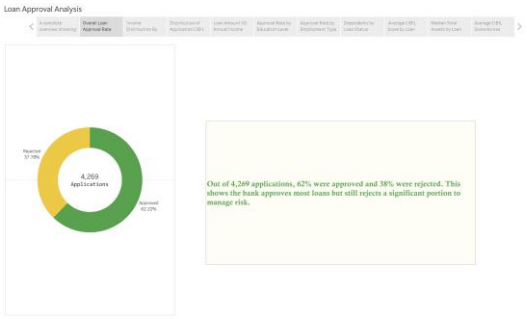

Average CIBIL Score Across Different Loan Terms
Avg Cibil Score
Loan Term

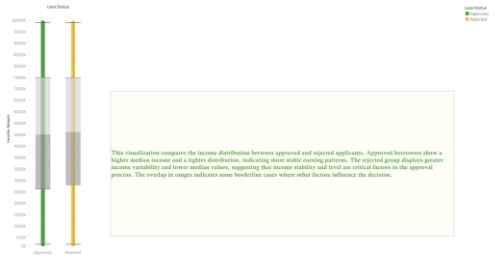

Median Total Assets by Loan Status
Loan Status
32,377,297
32,831,122
Median Total Asset Value
Approved
Rejected

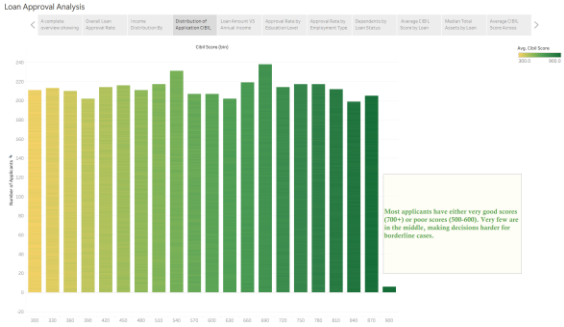
Dashboard

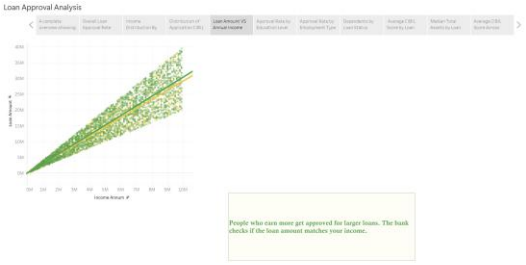

LOAN APPROVAL ANALYSIS DASHBOARD
Overall Laon Approval Rate
Rejected 37.78%
4,269 Applications
Approved 62.22%
Average CIBIL Score by Loan Status
Loan Status
703.5
429.5
Approved
Rejected
Income Distribution By Loan Status
Loan Status
Income Annum
Approved
Rejected
Distribution of Application CIBIL Scores
Cibil Score (bin)
Number of Applicants
Avarage CIBIL Score Across Different Loan Terms
Avg Cibil Score
Loan Term
Loan Amount VS Annual Income
Loan Amount
Income Annum
Approval Rate by Education Level
Education
31.37%
30.85%
18.86%
18.93%
Graduate
Not Graduate
Approval Rate by Employment Type
Self Employed
No
Yes
Dependents by Loan Status
2.47477
2.53813
Average Number of Dependents
Approved
Rejected
Median Total Assets by Loan Status
Loan Status
32,377,297
32,831,122
Median Total Asset Value
Approved
Rejected

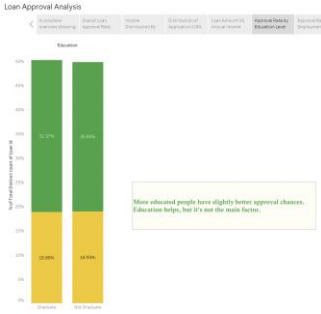
Story


Loan Approval Analysis
4,269
Applications
Out of 4,269 applications, 62% were approved and 38% were rejected. This shows the bank approves most loans but still rejects a significant portion to manage risk.

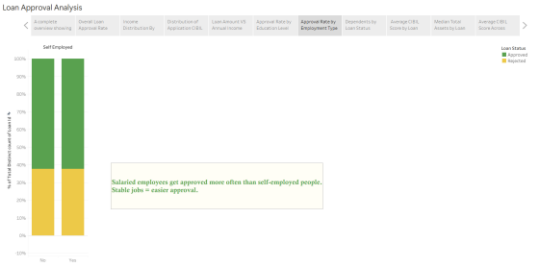

This visualization compares the income distribution between approved and rejected applicants. Approved borrowers show a higher median income and a tighter distribution, indicating more stable earning patterns. The rejected group displays greater income variability and lower median values, suggesting that income stability and level are critical factors in the approval process. The overlap in ranges indicates some borderline cases where other factors influence the decision.


Loan Approval Analysis
Most applicants have either very good scores (700+) or poor scores (300-600). Very few are in the middle, making decisions harder for borderline cases.

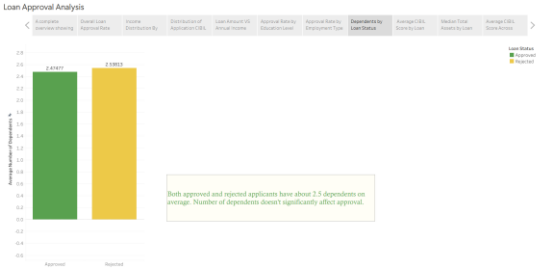

Loan Approval Analysis
People who earn more get approved for larger loans. The bank checks if the loan amount matches your income.


Loan Approval Analysis
More educated people have slightly better approval chances. Education helps, but it's not the main factor.

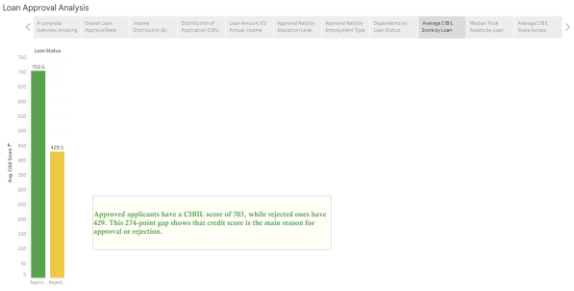
Loan Approval Analysis



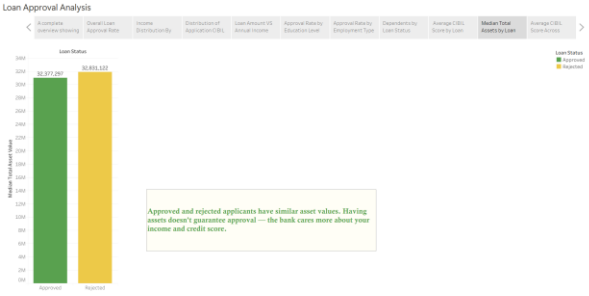
Loan Approval Analysis



Loan Approval Analysis



Loan Approval Analysis



Loan Approval Analysis