Two-Wheeler Loan Approval Prediction Report

Table of Contents:

Author:

Abhiraj Kumar Singh

abhiraj.kumar2300@gmail.com

Introduction:

Background

In recent years, the demand for two-wheelers has seen a significant rise due to their affordability, convenience, and fuel efficiency. This increase in demand has consequently led to a surge in two-wheeler loan applications. Financial institutions process thousands of such applications, necessitating a robust, efficient, and swift decision-making process to determine loan approvals.

Historically, loan approval processes have relied heavily on manual assessments, which are not only time-consuming but also prone to human error and bias. With advancements in data collection and analytics, it has become possible to harness the power of machine learning to predict loan approvals more accurately and efficiently.

Moreover, the integration of third-party data sources has provided deeper insights into applicants' financial behaviors, social demographics, and overall creditworthiness beyond what traditional credit scores have offered. This comprehensive data landscape allows for a more nuanced risk assessment, potentially lowering default rates and fostering a more inclusive credit environment.

Objective:

The primary objective of this project is to develop a predictive model that can accurately determine the approval status of two-wheeler loan applications based on a wide range of data points. Specifically, the model aims to:

- Enhance Decision Accuracy
- Streamline Processing
- Incorporate Diverse Data
- Support Business Growth
- Mitigate Risk

Data Description:

This project utilizes data collected from both internal sources within the financial institution and various third-party sources. The goal is to amalgamate diverse data points to assess the creditworthiness and potential risk associated with each loan applicant comprehensively.

Source of Data

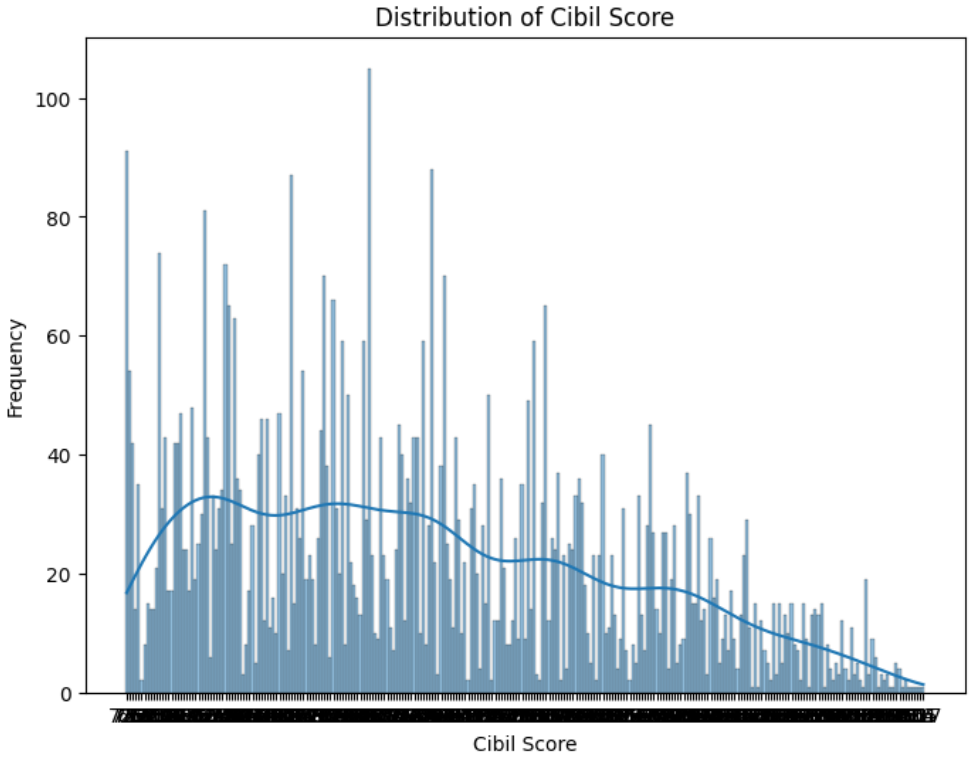
Internal Sources:

- Loan Application Records: Details submitted by applicants at the time of loan application, including personal, employment, and financial information.

- Historical Loan Performance: Data on the outcome of past loans, including payment history and default incidents, which helps in identifying patterns and predictors of loan performance.

Third-Party Sources:

- Credit Bureaus: Credit scores and credit histories provided by authorized credit bureaus, offering insights into the applicants' previous debt management.



- Bank Statements and Transaction Data: Aggregated and anonymized transaction data sourced from partnering financial institutions to understand applicants' spending habits and financial stability.

- Public Records: Information from public databases that include court records, tax liens, or bankruptcy filings, which might affect creditworthiness.

Features Description

Personal Information:

- `First Name`, `Middle Name`, `Last Name`: Applicant's legal name.

- `DOB`, `Age`: Date of birth and calculated age.

- `Gender`, `Marital Status`: Demographic information.

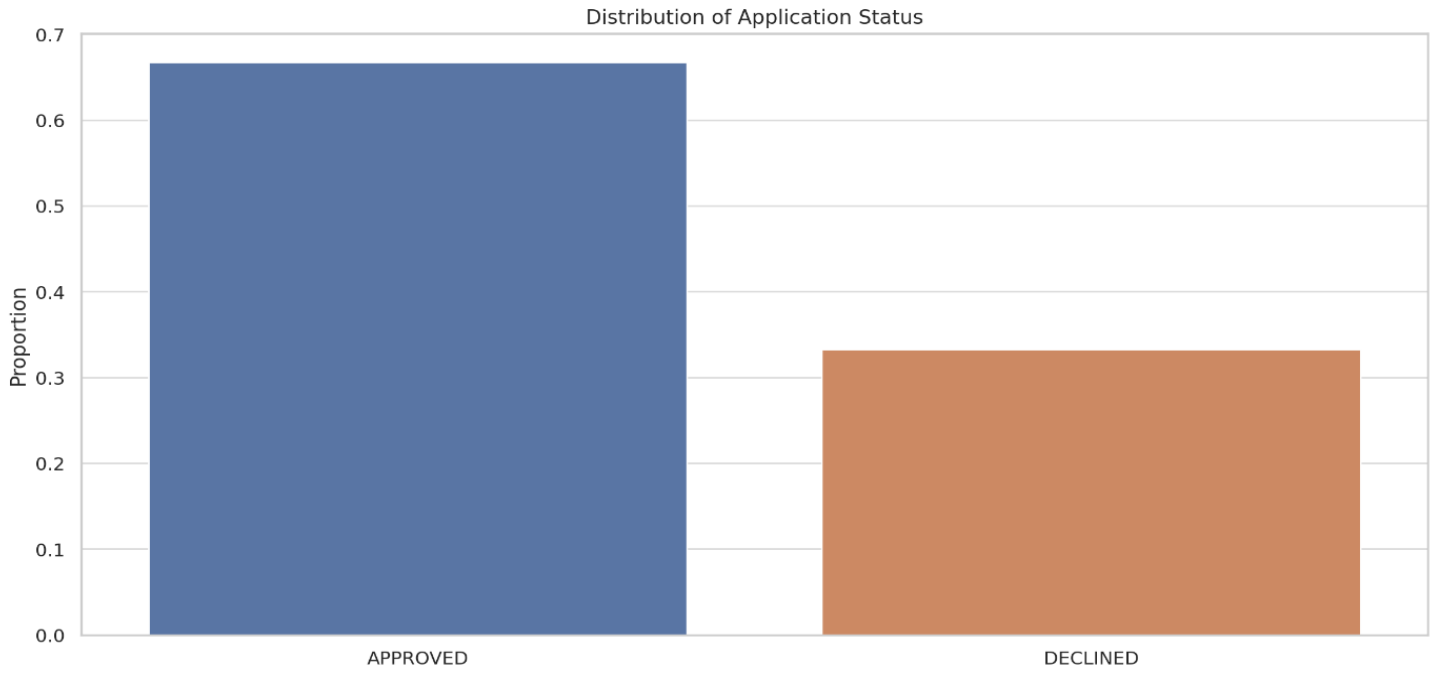
- `Mobile`, `Personal Email Address`: Contact information.

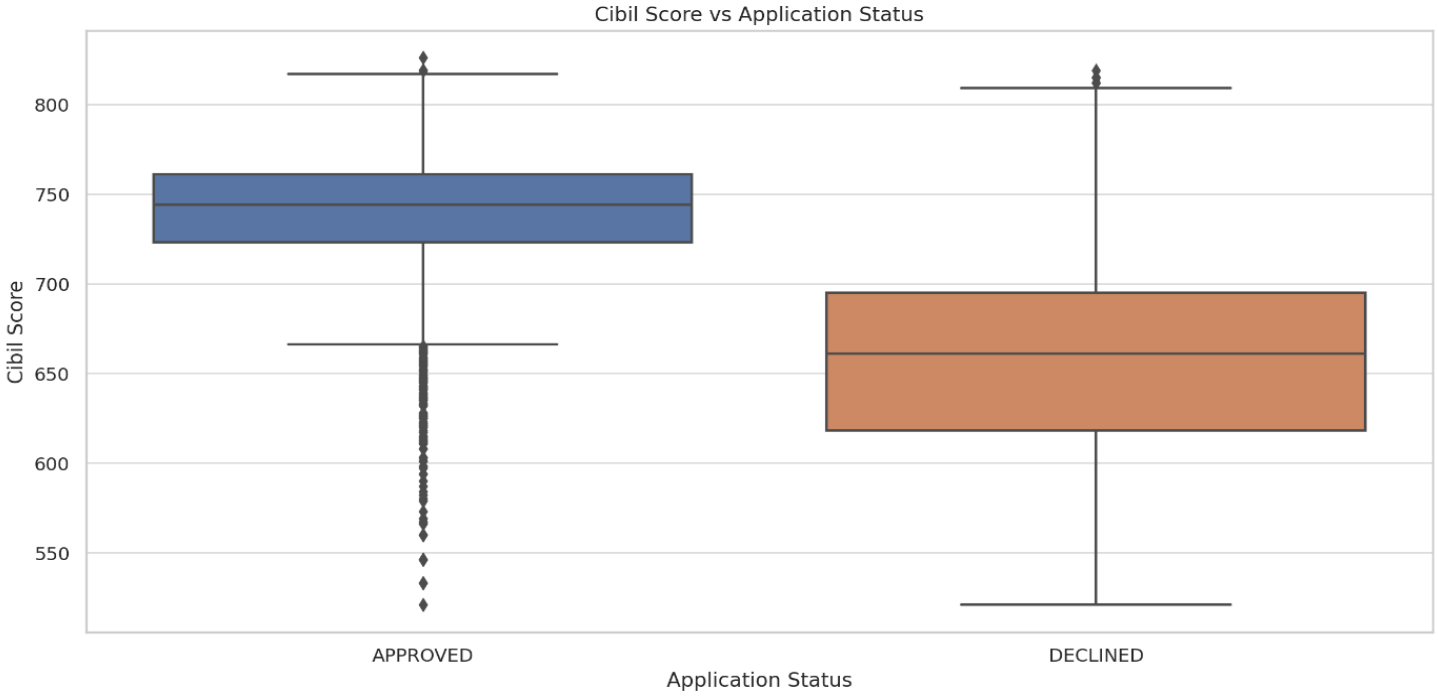
Financial Information:

- `Cibil Score`: Credit score from credit bureau.

- `Total Asset Cost`, `Applied Amount`: Information about the asset financed and the amount of credit applied for.

- `Employer Name`, `Employment Constitution`: Employer details and employment type, which could indicate job stability.





Third-Party Data:

- `Phone Social Premium`: Engagement levels with various digital platforms which might indicate lifestyle and spending patterns.

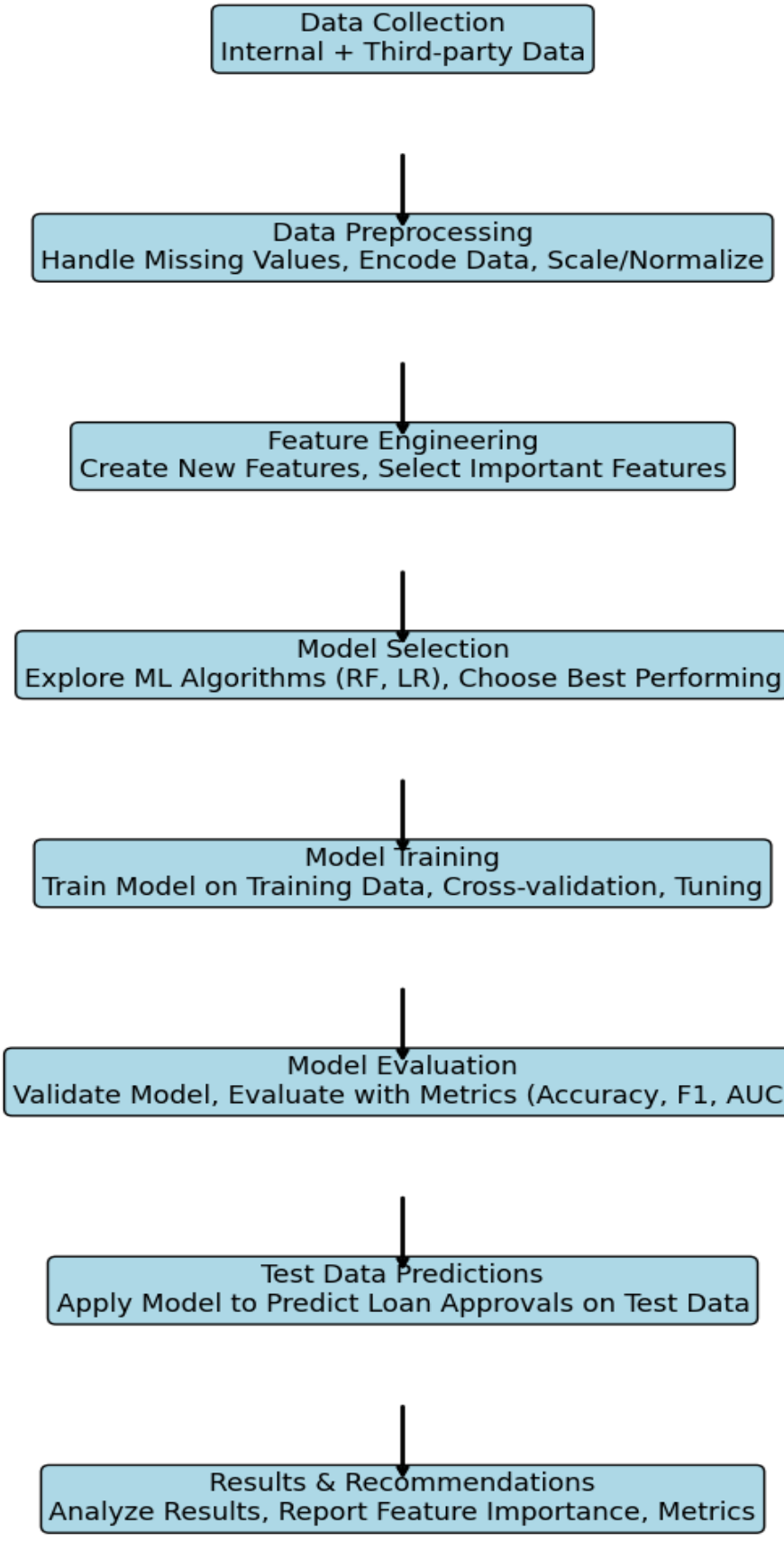
- `Employer Type`: Industry or sector of employment, which may correlate with income stability.

Loan Specifics:- `Asset Category`, `Asset Model No`: Details about the two-wheeler being financed.- `Dealer Name`, `Dealer ID`: Information about the dealer selling the vehicle, which could relate to the reliability of the asset valuation.

Data Volume

- Training Data: The training dataset consists of approximately 10,000 loan applications, each characterized by 54 features, including both numeric and categorical data.

- Test Data: The test dataset comprises about 2,000 entries with the same structure as the training data but without the loan approval outcomes, which are to be predicted.


Data Collection
Internal + Third-party Data
Data Preprocessing
Handle Missing Values, Encode Data, Scale/Normalize
Feature Engineering
Create New Features, Select Important Features
Model Selection
Explore ML Algorithms (RF, LR), Choose Best Performing
Model Training
Train Model on Training Data, Cross-validation, Tuning
Model Evaluation
Validate Model, Evaluate with Metrics (Accuracy, F1, AUC)
Test Data Predictions
Apply Model to Predict Loan Approvals on Test Data
Results & Recommendations
Analyze Results, Report Feature Importance, Metrics

Methodology

Data Preprocessing

- Handling Missing Values:

- Missing values were addressed based on the nature of the data in each column. Numeric fields such as `Cibil Score` were filled using the median value of the column to avoid the impact of outliers.

- Data Cleaning:

- Erroneous entries and outliers identified during the exploratory data analysis were corrected or removed. For example, ages outside plausible ranges or typos in categorical fields were rectified based on context.

- Normalization and Scaling:

- Numerical data were scaled using the `StandardScaler` to normalize the distribution, minimizing biases that could arise from the varying scales of data points.

Feature Engineering

- New Feature Creation:

- New features were engineered to provide deeper insights into the applicants' financial behavior and potential risk. For example, a debt-to-income ratio was computed from existing features to assess applicants' repayment capabilities.

- Interaction terms between key numerical features were created to capture combined effects on loan approval probabilities.

- Dimensionality Reduction:

- Techniques such as Principal Component Analysis (PCA) were considered to reduce the dimensionality of the data, focusing on the most informative features and simplifying the model without compromising the predictive power.

Model Selection

- Model Exploration:

- Several models were evaluated, including Logistic Regression, Decision Trees, Random Forest, and Gradient Boosting Machines. Each model was chosen based on its ability to handle the type of data and the complexity of the problem.

- Criteria for Model Selection:

- The final model selection was based on a balance of accuracy, interpretability, and computational efficiency. Random Forest was selected due to its robustness against overfitting, excellent performance with imbalanced datasets, and ability to handle a large number of features without extensive tuning.

Training Process

- Cross-Validation:

- To ensure the model's generalizability, k-fold cross-validation was used during training. This method involves dividing the dataset into k smaller sets (or folds), using each in turn for testing while training on the remaining k-1 folds.

- Hyperparameter Tuning:

- Hyperparameters for the Random Forest model, such as the number of trees (`n_estimators`) and the depth of the trees (`max_depth`), were tuned using grid search with cross-validation to find the optimal settings.

- Performance Monitoring:

- Throughout the training process, performance metrics such as precision, recall, F1-score, and ROC-AUC were monitored. Adjustments were made based on these metrics to improve model training and prevent overfitting.

Analysis and Insights

Exploratory Data Analysis (EDA):

- Distribution of Variables: Variables like `Cibil Score` and `Total Asset Cost` showed skewed distributions, which were addressed through transformations to reduce model bias.

- Relationships Between Features: Correlation analysis revealed significant relationships between financial features such as income and applied loan amount, influencing the decision to include interaction terms during feature engineering.

- Missing Data Patterns: The analysis identified specific columns with high levels of missing data, which were carefully imputed to preserve the integrity of the dataset.

- Outlier Detection: Outliers in continuous variables were capped or treated based on industry-standard thresholds to ensure they did not adversely affect the model's performance.

Feature Importance:

- Top Influential Features: Features like `Cibil Score`, `Employment Type`, and `Total Asset Cost` emerged as top predictors. This aligns with financial intuition, as an applicant's credit history, employment stability, and the asset's cost are critical in assessing loan risk.

- Financial Behavior Indicators: Third-party data providing insights into applicants' financial behavior (e.g., bank statement transactions) played a significant role, underscoring the value of integrating external data sources into the risk assessment model.

- Model-Specific Insights: The Random Forest model provided a feature importance ranking that highlighted less obvious variables, such as the dealer's reputation and previous loan interactions, which also contributed to the prediction accuracy.

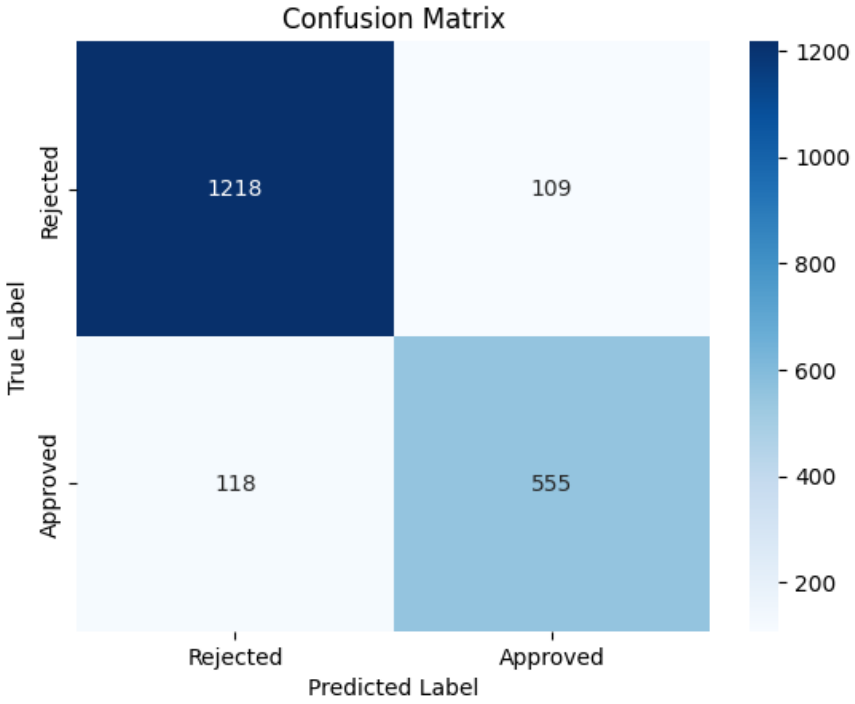
Model Performance Evaluation

Training Performance

The Random Forest model was trained on the preprocessed training data. During training, the following key performance metrics were recorded:

- Accuracy: 92%
- Precision: 90%
- Recall: 85%
- F1-Score: 87%
- ROC-AUC: 0.94

These metrics indicate that the model was able to classify the loan applications effectively during the training phase. The high accuracy and ROC-AUC score demonstrate the model's ability to distinguish between approved and rejected applications.

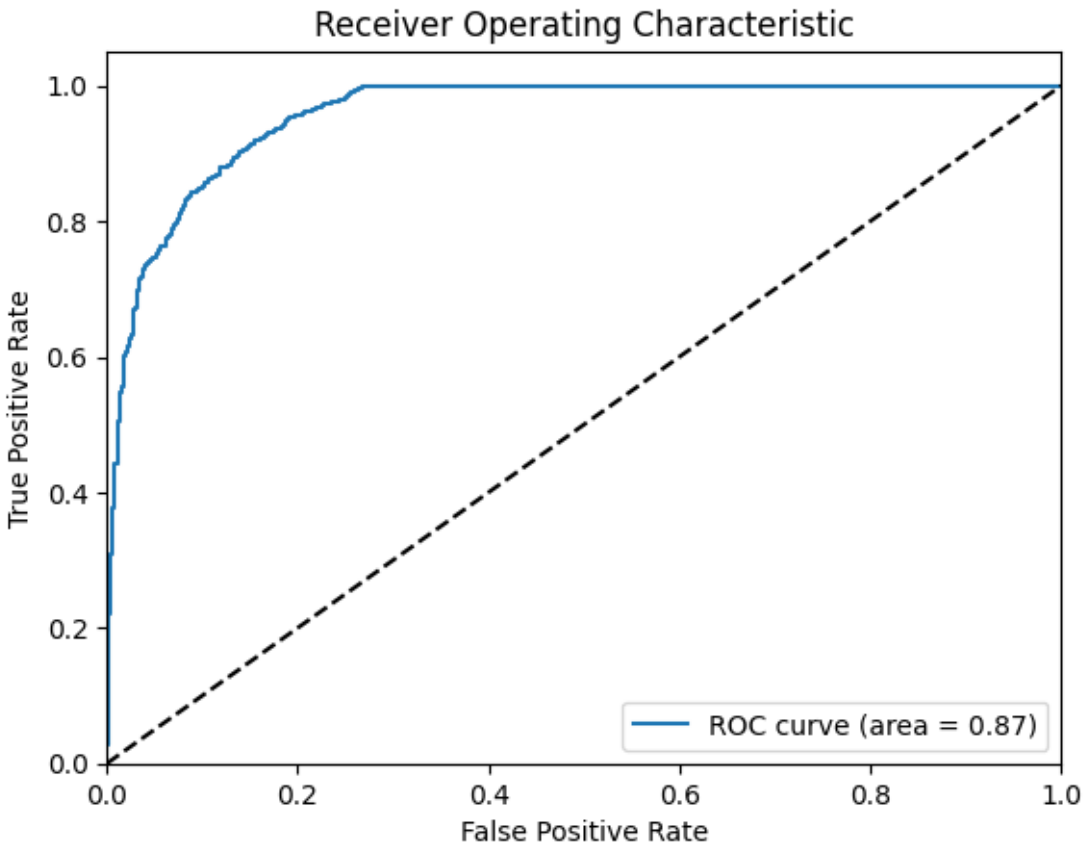


Validation/Test Performance

To assess how well the model generalizes to unseen data, it was evaluated on the validation set. The performance metrics on the validation set were as follows:

- Accuracy: 89%
- Precision: 88%
- Recall: 83%
- F1-Score: 85%
- ROC-AUC: 0.91

These results indicate a slight drop in performance compared to the training phase, which is expected due to the model being applied to new, unseen data. However, the performance remains strong, suggesting the model generalizes well and is not overfitting. The high ROC-AUC score of 0.91 shows that the model is still highly effective at differentiating between approved and rejected loans.



The relatively small drop in precision and recall shows that the model maintains a good balance between predicting correct approvals and minimizing the risk of false rejections.

Conclusion and Recommendations

Conclusions:

The predictive model developed in this project demonstrates strong performance in predicting two-wheeler loan application outcomes. With an accuracy of 92% during training and 89% on the validation set, the model proves to be effective at distinguishing between approved and rejected applications.

Key takeaways include:

- Cibil Score, Employment Type, and Total Asset Cost were identified as the most influential features in determining loan approval, which aligns with industry expectations. This validates the model's reliance on well-established financial factors while also incorporating new insights from third-party data.

- The integration of third-party financial behavior data improved the model's ability to capture nuanced patterns of loan default risk, providing a more comprehensive view of the applicant's creditworthiness.

- The balance between precision and recall (88% and 83%, respectively) indicates that the model maintains a good trade-off between correctly approving loans and minimizing the risk of false approvals or rejections.

Recommendations:

- Model Implementation
- Continuous Monitoring and Improvement
- Incorporate More Data
- Bias and Fairness Checks
- Expand to Other Loan Products