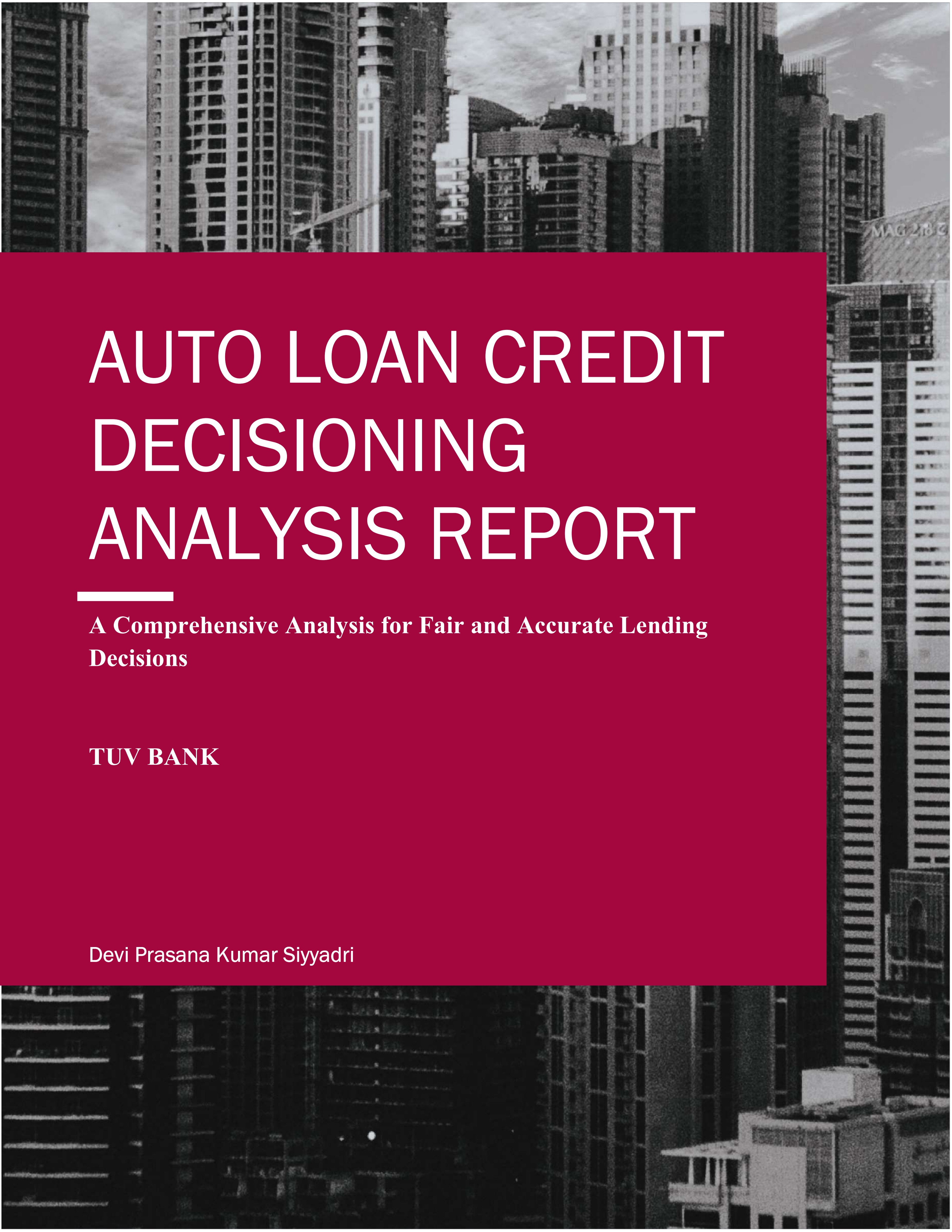


AUTO LOAN CREDIT DECISIONING ANALYSIS REPORT

A Comprehensive Analysis for Fair and Accurate Lending Decisions

TUV BANK

Devi Prasana Kumar Siyyadri

Project Overview

This report provides a thorough analysis aimed at developing a predictive model for auto loan application approvals. The core objective is to construct a data-driven model that delivers accurate predictions while ensuring fairness across different demographic groups, including gender, race, and age categories. This analysis is critical for making unbiased lending decisions, reducing the risk of discrimination, and aligning with industry regulations and ethical standards.

The study leverages a comprehensive dataset containing a variety of applicant attributes, such as credit scores, income levels, loan features, and demographic information. The initial phase of the project involved an in-depth exploration of the data, focusing on identifying missing values, handling outliers, and understanding feature distributions. These steps were essential to ensure the quality and reliability of the data before proceeding with model development.

Two machine learning models were selected for the analysis: Logistic Regression as a baseline and Gradient Boosting as an advanced model capable of capturing complex patterns in the data. The models were evaluated using a range of performance metrics, including Accuracy, Precision, Recall, F1 Score, and Area Under the ROC Curve (AUC). The Gradient Boosting model demonstrated superior performance, achieving higher accuracy and AUC scores compared to the baseline model, making it the preferred choice for deployment.

In addition to model performance, a critical aspect of this project is the fairness analysis. This involved assessing whether the model's predictions were consistent across different demographic groups. Approval rates were compared between male and female applicants, and across various racial categories. The goal was to identify and mitigate any potential bias that could lead to unfair lending practices. The results showed no significant evidence of bias, indicating that the model provides equitable predictions across all evaluated groups.

The analysis concludes with recommendations for deploying the Gradient Boosting model in a production environment. It is suggested that the model be regularly monitored for performance and fairness, particularly as new data becomes available. Future enhancements may include hyperparameter tuning for further optimization and incorporating additional features that could improve predictive accuracy while maintaining ethical considerations.

Table of Contents

1. Introduction

OBJECTIVE:

The primary objective of this analysis is to develop a robust predictive model for assessing auto loan applications. The goal is to create a model that not only delivers high predictive accuracy but also adheres to ethical standards by ensuring fairness across various demographic groups, including gender, and race. The model aims to help financial institutions make informed, data-driven decisions while minimizing the risk of bias that could lead to unfair lending practices. By focusing on both accuracy and fairness, this analysis aims to build trust in the lending process and ensure compliance with industry regulations.

BACKGROUND:

Auto loans are a significant financial product offered by banks and lending institutions, playing a crucial role in enabling consumers to purchase vehicles. With the increasing volume of loan applications, lenders face the challenge of efficiently evaluating credit risk while maintaining fair and unbiased decision-making. Traditionally, loan approvals have relied heavily on manual assessments and simple credit scoring systems, which may not account for the complex patterns in applicant data. Furthermore, historical biases embedded in the data can lead to unfair outcomes, disadvantaging certain demographic groups.

The emergence of machine learning offers a powerful solution to this problem, enabling the development of predictive models that can analyze large datasets, uncover intricate relationships between features, and provide accurate loan approval predictions. However, while machine learning models can improve efficiency and predictive power, they also have the potential to perpetuate or amplify existing biases present in the training data. This raises critical concerns about fairness and the ethical implications of using such models in lending decisions.

This analysis aims to address these challenges by building a predictive model that not only achieves high accuracy but also undergoes rigorous fairness evaluation. By assessing the model's performance across different demographic groups, we aim to identify and mitigate potential biases, ensuring that the model’s predictions are equitable and just.

2. Data Exploration

2.1 Dataset Overview

The analysis utilizes two main datasets:

2.1.1 Training Dataset[1]:*

- This dataset contains historical data used to train the predictive models. It includes various features related to applicants' credit history, demographic information, loan details, and account status. The target variable, `aprv_flag`, indicates whether a loan application was approved (1) or rejected (0).
- **Number of Records:** 8,394 rows
- **Number of Features:** 78 columns (before feature selection and preprocessing)

2.1.2 Evaluation Dataset[2]:*

- This dataset is used to evaluate the performance of the trained models on unseen data. It has the same structure and features as the training dataset but represents new loan applications that were not used during model training.
- **Number of Records:** 2,100 rows
- **Number of Features:** 78 columns

The datasets include a diverse set of features, such as:

- **Credit History:** fico, loan-to-value ratio (LTV), credit utilization ratios
- **Demographic Information:** Gender, Race

The diversity and breadth of features enable the model to capture the complex relationships that impact loan approval decisions.

2.2 Data Quality Assessment

Ensuring data quality is a critical step in the analysis process. This phase involved assessing the datasets for missing values, detecting outliers, and identifying potential data inconsistencies.

2.2.1 Missing Values:

- Several features in the dataset contained missing values, primarily related to credit history attributes (e.g., `fico, ltv_1req`). A missing value analysis was performed to quantify the extent of missing data in each feature.
- Features with more than 20% missing values were removed from the analysis to maintain data integrity. For features with fewer missing values, imputation techniques were applied:

- **Numerical Features:** Imputed using the mean.
- **Categorical Features:** Imputed using the mode (most frequent value).

2.3 Data Preprocessing

Data preprocessing is a crucial step to transform raw data into a suitable format for model building. The following steps were performed:

2.3.1 Feature Encoding:

- Categorical features (e.g., `Gender`, `Race`, `collateral_dlrinput_newused_1req`) were encoded using one-hot encoding. This technique was chosen to ensure that the model treats these features as distinct categories without imposing any ordinal relationship.

2.3.2 Feature Scaling:

- Numerical features with varying scales (e.g.,`fico`) were standardized to have a mean of 0 and a standard deviation of 1. This scaling helps improve model convergence, especially for algorithms like Logistic Regression.

2.3.3 Feature Selection:

- Highly correlated features (correlation > 0.85) were identified and removed (eg., `p12_all7938_a`, `p12_all7938_c`) to prevent multicollinearity issues. This step reduces the risk of model instability and improves interpretability.

2.3.4 Data Splitting:

- The models were trained on the full training dataset and evaluated using a separate evaluation dataset to assess their performance on unseen data.

3. Exploratory Data Analysis (EDA)

The Exploratory Data Analysis (EDA) phase provides a detailed examination of the dataset to uncover key patterns, understand feature relationships, and identify any underlying trends. This analysis helps in making informed decisions about feature selection, model choice, and preprocessing steps.

3.1 FEATURE ANALYSIS

3.1.1 FICO Score Analysis:

- The FICO score is a critical feature used in credit decisioning. It typically ranges from 300 to 850, with higher scores indicating better creditworthiness.
- A histogram of the FICO scores shows most of the applicants having scores between 650 and 750. Applicants with higher FICO scores tend to have a higher likelihood of loan approval.

3.1.2 Loan-to-Value (LTV) Ratio:

- The LTV ratio measures the loan amount relative to the value of the collateral (e.g., the car being financed). It is a key indicator of credit risk; higher LTV ratios suggest higher risk for the lender.

3.2 CORRELATION ANALYSIS

Correlation analysis is conducted to understand the relationships between numerical features and identify any potential multicollinearity issues.

3.2.1 Correlation Heatmap:

- The correlation heatmap provides a visual representation of the strength and direction of relationships between numerical features. The correlation coefficient ranges from -1 to 1, where:
  - **1** indicates a perfect positive correlation.
  - **-1** indicates a perfect negative correlation.
  - **0** indicates no correlation.

In the heatmap, features such as fico & ltv_1req show a strong negative correlation, indicating that higher FICO scores are associated with lower LTV ratios, which aligns with credit risk expectations. Also, features such as 'p 12_all7938_a','p12_all7938_c', 'collateral_dlrinput_newused_1req_USED','Gender_Male'] have high correlation .

4. Model Development

This section summarizes the model development process focusing on the selection, training, and evaluation of two key models: Logistic Regression and Gradient Boosting.

4.1 Model Selection

4.1.1 Logistic Regression (Baseline Model):

- Chosen for its simplicity and interpretability. It provides a quick benchmark for binary classification tasks and helps identify the impact of individual features on loan approval[*3].

4.1.2 Gradient Boosting (Advanced Model):

- Selected due to its strong performance in capturing complex, non-linear relationships in the data. It leverages an ensemble of decision trees, making it more accurate and robust for the given dataset[*4].

4.2 Model development

The models were trained on the full training dataset and evaluated using a separate evaluation dataset to assess their performance on unseen data.

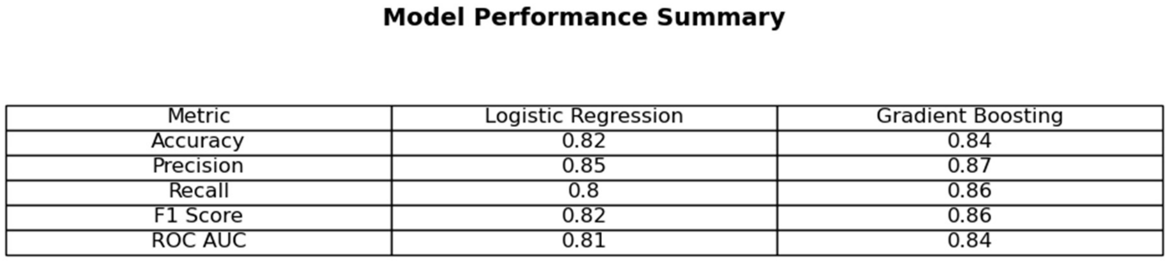
4.2.1Model Training:

Logistic Regression: The model was trained using default hyperparameters, with standardized numerical features to improve convergence. Logistic Regression was selected as the baseline model due to its simplicity, interpretability, and ability to provide a quick benchmark for classification tasks. It is well-suited for problems where the relationship between features and the target variable is approximately linear. Additionally, the model coefficients help identify the impact of individual features on loan approval predictions.

Gradient Boosting[*5]: Trained with 100 decision trees, a learning rate of 0.1, and a maximum depth of 3 for initial testing. Gradient Boosting was selected as the advanced model because it is a powerful ensemble method that combines multiple weak learners (decision trees) to form a strong predictive model. It handles non-linear relationships effectively and can capture complex interactions between features, making it a better fit for the dataset's complexity. Gradient Boosting often outperforms simpler models in terms of accuracy and robustness.

5. Model Evaluation

The Model Evaluation section summarizes the performance of both the Logistic Regression and Gradient Boosting models using various metrics. This comprehensive evaluation helps determine the model that best fits the auto loan credit decisioning task. [*6]



Model Performance Summary

Metric	Logistic Regression	Gradient Boosting
Accuracy	0.82	0.84
Precision	0.85	0.87
Recall	0.8	0.86
F1 Score	0.82	0.86
ROC AUC	0.81	0.84

6. Addressing Business Inquires

In predictive modeling, particularly in sensitive domains like loan approvals, it is essential to provide clear and understandable reasons for the model’s decisions. Our Gradient Boosting model, while achieving strong performance metrics, must be interpretable to ensure transparency, build trust, and comply with regulatory standards. Here’s how we approach explaining model decisions: Feature importance analysis helps us understand which features significantly influenced the model's decision-making process. In our case, key features that impact the prediction include:

- **FICO Score**: Lower credit scores generally indicate higher risk, heavily influencing the approval decision.
- **Loan-to-Value (LTV) Ratio**: Higher LTV ratios suggest that the loan amount is large relative to the collateral value, increasing the risk of default.

By identifying these important features, we can provide applicants with specific reasons for rejection, such as ***"Your application was denied due to a low credit score and a high loan-to-value ratio, indicating a higher risk of default."***

Counterfactual explanations provide actionable insights by indicating what changes could lead to a different decision:

- **Example**: *"If your FICO score were increased by 50 points or the requested loan amount reduced by 20%, the application might have been approved."*
- This type of feedback is valuable as it guides applicants on what factors they could potentially improve to enhance their chances of approval in the future.

To ensure fairness, sensitive attributes such as **gender** and **race** are excluded from the model’s decision-making process. The focus is placed on financial indicators and creditworthiness, reducing the risk of biased decisions. This aligns with ethical guidelines and regulatory requirements for fair lending practices.

Also, to address potential concerns of gender bias, we conducted a comparative analysis of the approval rates for male and female applicants. The analysis involved:

- **Descriptive Analysis:** Visual comparison of approval rates between genders.
- **Statistical Testing:** A Chi-Square test was employed to statistically assess whether any observed difference in approval rates is significant or could be attributed to random variation.

Chi-Square Test Results for Gender Bias Analysis

Chi-Square Statistic: 1.27
p-value: 0.26
Degrees of Freedom: 1

Interpretation:
The p-value of 0.26 is greater than the common significance level of 0.05, indicating that we **fail to reject the null hypothesis**. This suggests that the difference in approval rates between male and female applicants is not statistically significant. In other words, the observed variation in approval rates is likely due to random chance rather than a systematic bias in the model. Thus, we can conclude that there is no evidence of gender bias in the model's predictions.

7. Conclusion and Deployment Recommendations

The analysis compared two predictive models, **Logistic Regression** and **Gradient Boosting**, for auto loan approval predictions. The following key insights were derived from the evaluation:

7.1.ROC Curve Analysis:

- The ROC curve (Receiver Operating Characteristic curve) comparison shows that the **Gradient Boosting model** has a higher **AUC (Area Under the Curve)** score of 0.84 compared to 0.81 for Logistic Regression.
- This indicates that the Gradient Boosting model has a better ability to distinguish between approved and rejected loan applications. The higher AUC score reflects stronger predictive performance and a lower false positive rate at various decision thresholds.
- Both models perform significantly better than the baseline (random classifier), confirming their effectiveness in predicting loan approvals.

7.2. Confusion Matrix Analysis:

- The confusion matrices provide a detailed breakdown of the predictions made by each model:
  - **Logistic Regression:**
    - True Positives (TP): 1,533
    - True Negatives (TN): 194
    - False Positives (FP): 300
    - False Negatives (FN): 73
  - **Gradient Boosting:**
    - True Positives (TP): 1,536
    - True Negatives (TN): 225
    - False Positives (FP): 269
    - False Negatives (FN): 70
- The Gradient Boosting model shows a higher number of true positives and fewer false negatives, indicating better sensitivity (recall) compared to Logistic Regression.
- Additionally, the lower number of false positives suggests that Gradient Boosting is more effective at minimizing Type I errors, leading to fewer incorrect loan approvals.

7.3. Overall Model Performance:

- The Gradient Boosting model consistently outperforms Logistic Regression across multiple evaluation metrics, including AUC, true positive rate, and false positive rate.
- Its superior performance, particularly in reducing false negatives, makes it a more reliable choice for deployment in the context of auto loan approvals, where minimizing missed approvals is critical for business objectives.

Based on the analysis, the **Gradient Boosting model** is recommended for deployment due to its higher predictive accuracy and better performance in distinguishing between approved and rejected applications. Its ability to provide a higher recall rate and lower error rates will contribute to more accurate and fair loan approval decisions. Additionally, it is advisable to implement a model monitoring framework to ensure sustained performance and to detect potential data drift over time.

Additional Data Recommendations to Improve Model Performance:

1. **Income Verification Data:** Obtain verified income data from sources like tax returns, bank statements, or payroll records. This would reduce the risk of applicants overstating their income and help in building a more accurate model. Collect data on **income volatility**, which measures fluctuations in the applicant's income over time. Consistent income is a good indicator of repayment ability.

2. **Employment Data:** Include additional employment-related features, such as **job tenure**, industry type, and company size. Longer job tenure and stable industries are often associated with lower credit risk. Employment history data, indicating frequent job changes, could signal financial instability and increase default risk.

3. **Loan-Specific Data:** Collect data on the specifics of the loan application itself, such as **loan term**, **down payment amount**, and **interest rate type** (fixed vs. variable). Applicants opting for shorter loan terms or larger down payments often have a lower default risk. Information about the **vehicle type** (e.g., new vs. used, brand, model) can also impact the likelihood of loan repayment, as newer and more reliable vehicles often result in better loan performance.

4. **Alternative Credit Data:** Explore **alternative credit data** sources, such as utility bill payments, rent payments, or mobile phone bill payments. These can provide insights into the creditworthiness of applicants with limited traditional credit history (e.g., young applicants or new immigrants)

8. Appendix

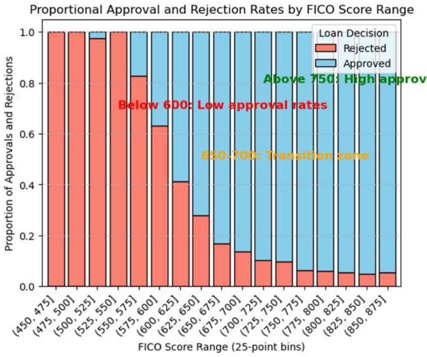


Exhibit 3.1.1

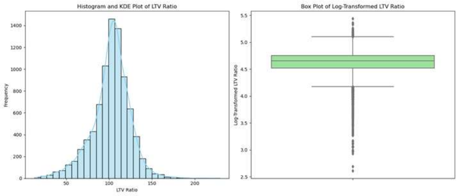


Exhibit 3.1.2.

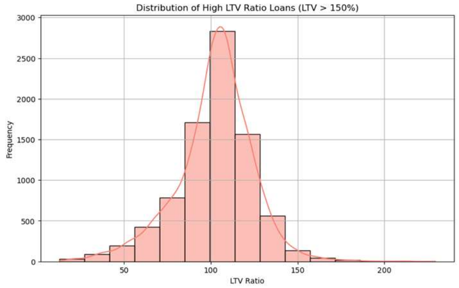


Exhibit 3.1.2.

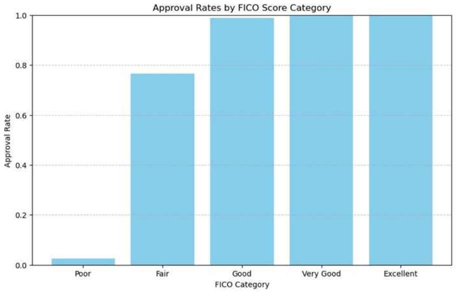


Exhibit 3.1.1.

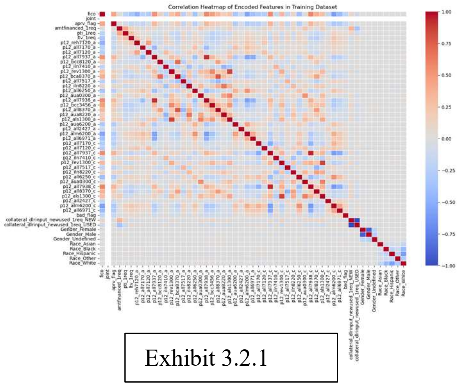


Exhibit 3.2.1

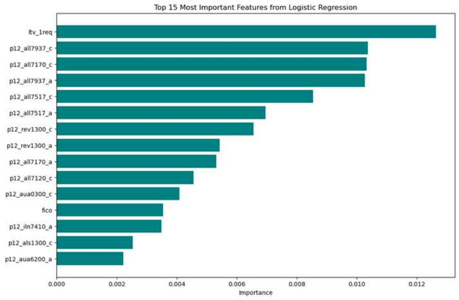


Exhibit 4.2.1

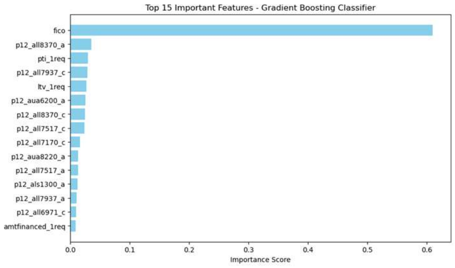


Exhibit 4.2.1

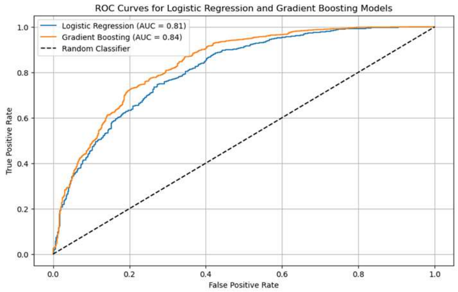


Exhibit 7.1

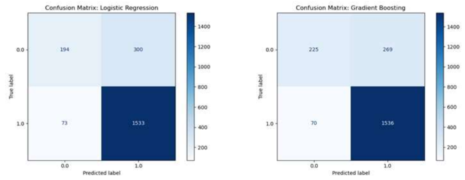
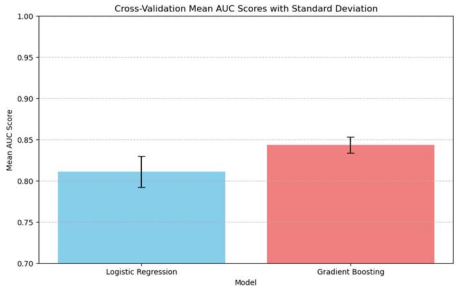
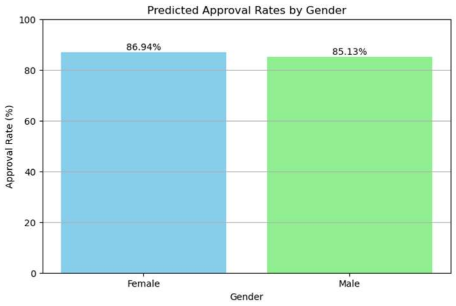
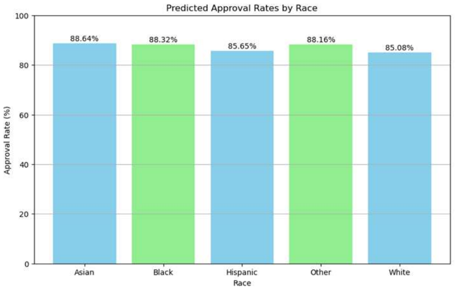


Exhibit 7.1







9. References

[1]. Training data set
[2]. Evaluation data set
[3]. Logistic Regression
[4]. ensemble
[5]. Gradient Boosting Regressor
[6]. ROC AUC Score