## **Meta-Model Loan Defaulter Classifier Report**

DATA602- Principles of Data Science  
Fall 2024

**Instructor:** Dr. Fardina Alam  
**Project:** Loan Defaulter Classifier

**Project URL:** [AsutoshDalei/Loan-Defaulter-Classification](https://github.com/AsutoshDalei/Loan-Defaulter-Classification)

**Group Members:**

* Shruti Gajipara (121233882, [shruti01@umd.edu](mailto:shruti01@umd.edu))
* Aayush Verma (121331076, [aver23@umd.edu](mailto:aver23@umd.edu))
* Premal Shah (121293596, [shah1305@umd.edu](mailto:shah1305@umd.edu))
* Asutosh Dalei (120997754, [asutoshd@umd.edu](mailto:asutoshd@umd.edu))

**Date of Submission:** 15 December, 2024

### 

#### **Contributions**

|  |  |  |
| --- | --- | --- |
| **Tasks** | **Member 1** | **Member 2** |
| Project Idea | Aayush | Shruti |
| Dataset Curation and Preprocessing | Premal | Asutosh |
| Data Exploration and Summary Statistics | Aayush | Shruti |
| ML Algorithm Design/Development | Asutosh | Aayush |
| ML Algorithm Training and Test Data Analysis | Shruti | Aayush |
| ML evaluation and Hyper-parameter tuning | Premal | Asutosh |
| Visualization, Result Analysis, Conclusion | Premal | Asutosh |
| Final Tutorial Report Creation | Shruti | Premal |

**Shruti Gajipara**

Contributions: Project idea, performed data understanding. Created visualization highlighting trends and patterns. Performed hypothesis testing to assess relationships and determine feature importance. Prepared dataset for modelling by encoding variables, performing feature scaling/normalization wherever needed. Identified the outline for final document and created it. Identified the outline for final document and GitHub repo and Pages website and created it.

**Aayush Verma**

Contribution: Project idea, EDA by creating visualizations to identify key trends. Feature engineering and feature selection. Created Catboost and generated predictions for majority class. Designed Meta-Model, its working, structure and role. Created features for the model, trained it and integrated it into existing two pipeline. Tested the pipeline end to end to ensure proper functionality.

**Premal Shah**

Contribution: Performed data understanding, implemented cleaning techniques, outlier removal and created visualization highlighting trends and patterns. Evaluated the models’ performance. Performed hyper-parameter tuning for three models and tried to optimized model performance. Create visualization and translate findings and results to business insights. Identified the outline for final document and GitHub repo and website and helped create it.

**Asutosh Dalei**

Contribution: Summarized key statistical metrics such as mean, median, variance, and correlation coefficients to provide a quantitative overview of the data. Created heatmaps and pairwise plots to highlight dependencies between features, analyzed feature distributions to detect skewness. Created one-class SVM model, performed feature scaling, trained and tested it. Performed hyper-parameter for 3 models tuning to improve performance. Tried various versions of SVM-linear and non-linear both for best results. Generated insights and helped translate findings and results to business insights.

### **Introduction**

Lending money in consumer finance requires a careful balancing act between risk and opportunity. Loan providers frequently struggle to find potential clients who can responsibly handle their debts since they are unable to lend to people with little or no credit history. By examining past data to find trends that can be used to forecast which applicants are most likely to experience loan default, this research seeks to address that difficulty. In this regard, our project tackles two significant risks related to choices about loan approval:

1. Opportunity Loss: Missed business possibilities arise when loans are denied to applicants who can afford them.
2. Financial Loss: The company may suffer large financial losses if loans are approved for borrowers who are likely to default.

This project, therefore, proposes a **unique meta-modeling ensemble of models** that integrates multiple machine learning techniques for improving predictive performance and adaptability. It combines an anomaly detection algorithm tuned for the isolation of high-risk loan defaulters along with a high performance gradient boosting model capable of handling high dimensional and high-cardinality categorical features and imbalanced datasets. These models are controlled by a **decision maker classifier**, which dynamically evaluates and decides on the best model to deploy based on the nature of the contextual data. This hierarchical architecture introduces an adaptive decision-making layer that sits on the top and intelligently selects the most effective model for a given scenario, thereby reducing over-reliance on any single algorithm and mitigating the limitations inherent in individual models. By combining the strengths of two models, each tailored to a specific class/label within a dynamic meta-model structure not only enhances predictive robustness but also enables the system to efficiently pivot between models in real-time, ensuring optimal performance across diverse conditions.

### **2. Data Curation**

The dataset used in this project comprises historical loan application data and previous application records:

1. **Application Data:** Contains details of loan applicants, including demographic information, credit history, and loan request details, financial information, location etc.
2. **Previous Application Data:** Captures records of prior loan applications, providing additional insights into customer behavior, past borrowing behavior and repayment patterns.

#### **Key Steps in Data Preprocessing**

1. **Handling Missing Values:**
   * + **Numerical Variables:** Imputed with the mean or median based on the variable's distribution.
     + **Categorical Variables:** Imputed with the mode to preserve the most frequent category.
2. **Feature Engineering:**
   * **Income-to-Loan Ratio:** Calculated as the ratio of the applicant's income to the requested loan amount.
   * **Prior Default Indicator:** Flagged applicants with past defaults in the previous applications dataset.
3. **Outlier Detection:**
   * Anomalies were identified using statistical methods (Z-score), IQR and visualizations like box plots. These anomalies were flagged for further analysis or excluded from the dataset.

### **3. Exploratory Data Analysis**

1. **Target Variable Distribution:**

* A pie chart revealed that the dataset was highly imbalanced:
  + **Non-defaulters:** Represented the majority class (~92% of the data).
  + **Defaulters:** Represented the minority class (~8% of the data).

1. **Income-to-Loan Ratio Analysis:**

* A histogram of the income-to-loan ratio showed a right-skewed distribution.
* Applicants with higher income-to-loan ratios were less likely to default, indicating this feature's importance in predicting default probabilities.

1. **Feature Correlation Analysis:**

A heatmap of the correlation matrix revealed key relationships:

* + **Positive Correlations:** Credit history length and income level were positively correlated with the likelihood of loan approval.
  + **Negative Correlations:** High loan amounts relative to income were negatively correlated with repayment likelihood.

1. **Anomalies and Outliers:**

* Scatter plots and IQR highlighted anomalies in income levels, loan amounts and other features.
* These anomalies were further analyzed using One-Class SVM, which identified high-risk applicants with patterns deviating significantly from the norm.

1. **Distribution of Loan Amounts:**

* The distribution of loan amounts was visualized using box plots:
  + Outliers represented applicants requesting unusually high loan amounts.
  + These outliers were flagged for deeper analysis to assess their repayment capacity.

1. **Hypothesis Testing:**

* Hypothesis 1: Chi-square test to determine owning a car and repayment of loan were independent of each other.
  + Finding: Owning a car did not have a significant impact on loan repayment
* Hypothesis 2: There is no difference in credit amount between people who repaid the loan and those who did not.
  + Finding: There is some difference in credit amount of loan repayers and defaulters.
* Hypothesis 3: Loan application day is independent of loan repayment.
  + Finding: Day of loan application is indeed dependent on loan repayment.

### **Machine Learning Analysis**

#### **Algorithm Selection and Design**

* 1. **CatBoost**:
  + Trained on entire dataset but, we focused on its performance specifically across the majority class (Class 0).
  + More suited for our case as data is highly dimensional and highly categorical and is a large dataset. Using Catboost’s inbuilt encoding techniques, we will be able to prevent overfitting as well given its architecture.
  + Introduced weighted loss functions in CatBoost to assign higher importance to minority class samples.
  + **Parameters Used**:
    - Learning Rate: 0.1
    - Iterations: 500
    - Depth: 6

1. **One-Class SVM**:
   * Specializes in identifying potential defaulters by detecting anomalies based on their distinct patterns in the feature space.
   * Trained on entire dataset, but we focused on its performance across only the minority class (Class 1).
   * Model chosen because it is one of the best models for a minority class in an imbalanced situation. Also, it is computationally faster than its counterparts with higher interpretability.
   * We performed oversampling of defaulters (minority class) through SMOTE in the training set to mitigate imbalance.
   * **Parameters Used**:
     + Kernel: Radial Basis Function (RBF)
     + Nu: 0.05 (to identify 5% of the data as anomalies)
2. **Random Forest Meta-Classifier**:
   * Aggregates predictions from the two base models to deliver a unified and robust classification outcome.
   * Some feature engineering was done to create enhanced features to help meta-model easily differentiate between the classes.
   * Model trained on the predicted probabilities of both Catboost and SVM, the absolute difference between the predicted probabilities, combined average score of Catboost and SVM probabilities, distance from the decision boundary in SVM. All of these are called *meta-features*
   * We tried GridSearchCV for hyper-parameter tuning, but it was computationally expensive, hence we decided to drop it.
   * The Random Forest will then predict two labels- 1 and 0. Depending on the value, either the catboost or the SVM will be chosen for predicting the probabilities of loan default.
   * We chose Random Forest as the decision making model because it handles diverse features well, can capture non-linear relationships too, is robust to noise and overfitting, can handle imbalanced datasets and is computationally efficient.
   * **Parameters Used**:
     + Number of Trees: 100
     + Max Depth: 5

#### **Model Training and Evaluation**

We performed cross validation with 5 folds to ensure the robustness and generalizability of the models and the results from each fold were averaged to mitigate performance variability.

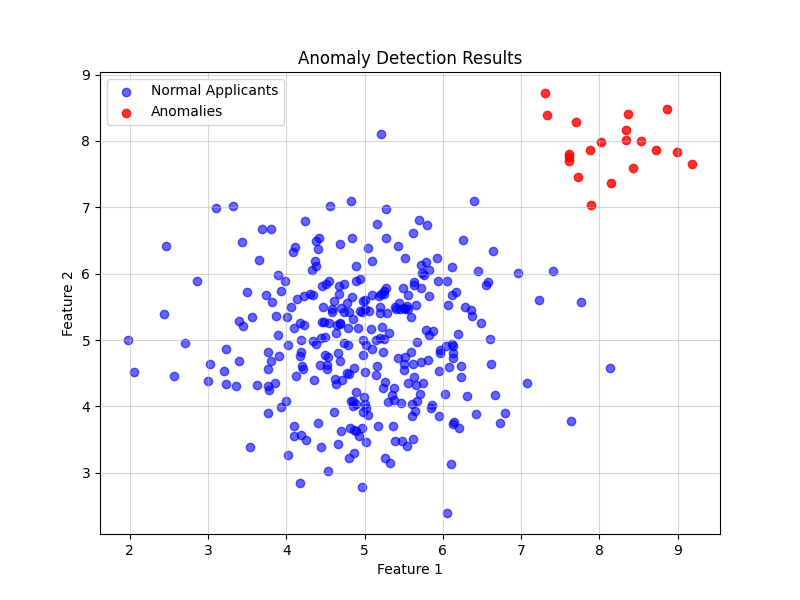
#### **Results**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **CatBoost** | | **One-Class SVM** | | **Meta Model** | |
| **Metric** | ***Class 0*** | ***Class 1*** | ***Class 0*** | ***Class 1*** | ***Class 0*** | ***Class 1*** |
| **Precision** | 0.93 | 0.24 | 0.93 | 0.1 | 0.96 | 0.42 |
| **Recall** | 0.94 | 0.2 | 0.58 | 0.5 | 0.94 | 0.51 |
| **F1-Score** | 0.94 | 0.22 | 0.71 | 0.16 | 0.95 | 0.46 |
| **Accuracy** | 0.88 | | 0.57 | | 0.91 | |

* **CatBoost** excelled in handling the majority class and provided high recall for non-defaulters.
* **One-Class SVM** achieved higher precision for defaulters by focusing on the minority class.
* **Meta-Classifier** outperformed both base models, achieving the highest AUC-ROC score (0.92) and F1-score (0.88), demonstrating the advantage of combining diverse algorithms.

**5. Visualization**

#### **1. Anomaly Detection Results (Scatter Plot)**

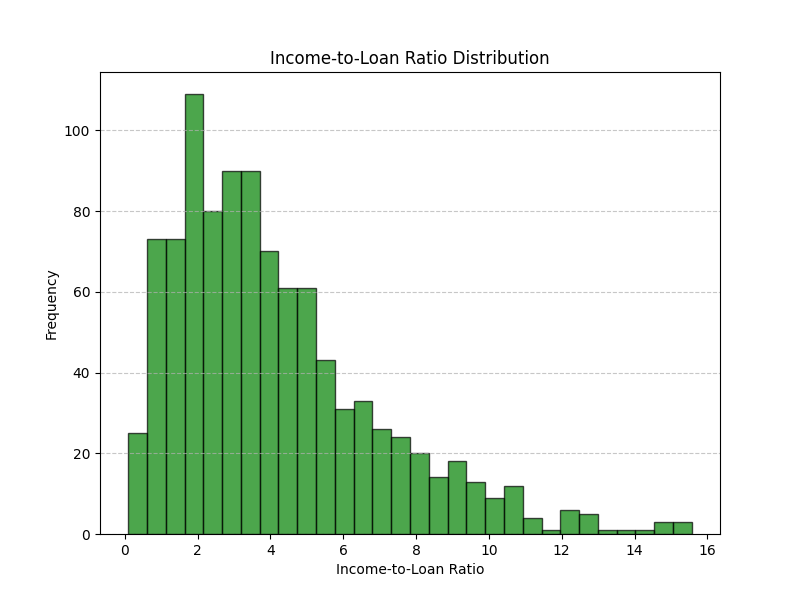


* This scatter plot separates data points into two categories: **Normal Applicants (blue)** and **Anomalies (red)**.
* The blue cluster represents consistent patterns in the dataset, while the red points indicate outliers or unusual behaviors.

**Insights:**

* The anomalies identified here likely represent high-risk applicants with unusual feature combinations, such as extreme income-to-loan ratios or past behaviors deviating significantly from the norm.
* These anomalies align closely with loan defaulters, making them critical for focused risk management strategies.

**2. Income-to-Loan Ratio Distribution (Histogram)**



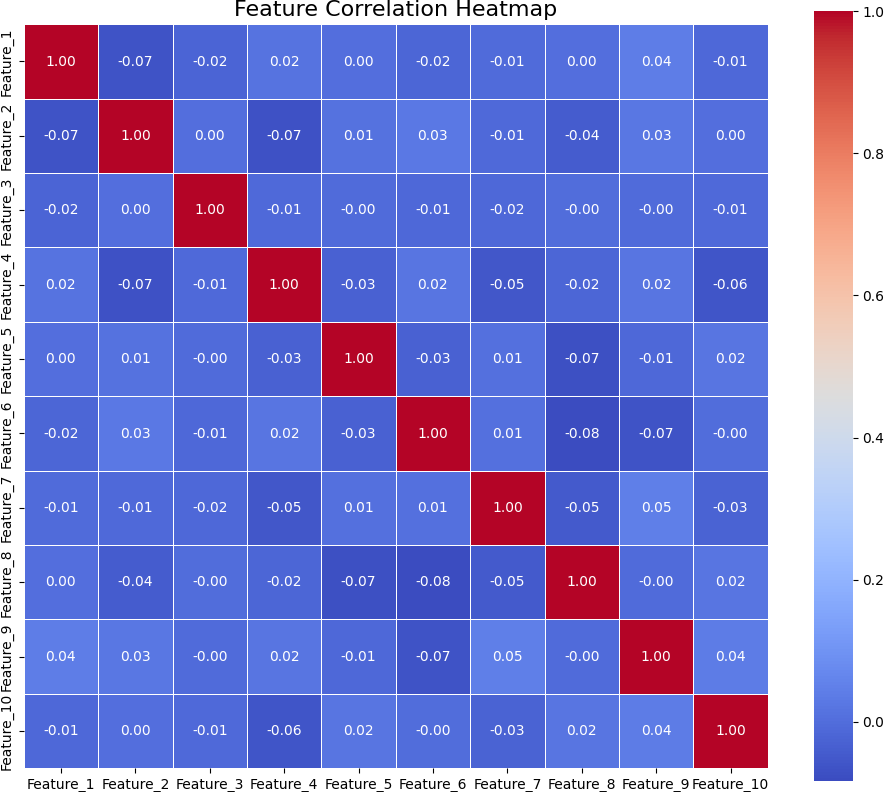
* The histogram shows the distribution of the income-to-loan ratio across applicants.
* Most values are concentrated between **1 and 6**, with fewer applicants having extremely high ratios.

**Insights:**

* Applicants with a **high income-to-loan ratio** are less likely to default, as they have a stronger financial capacity relative to their loan amount.
* The long tail represents applicants with extremely favorable ratios, who are generally lower-risk borrowers.

#### **3. Feature Correlation Heatmap**

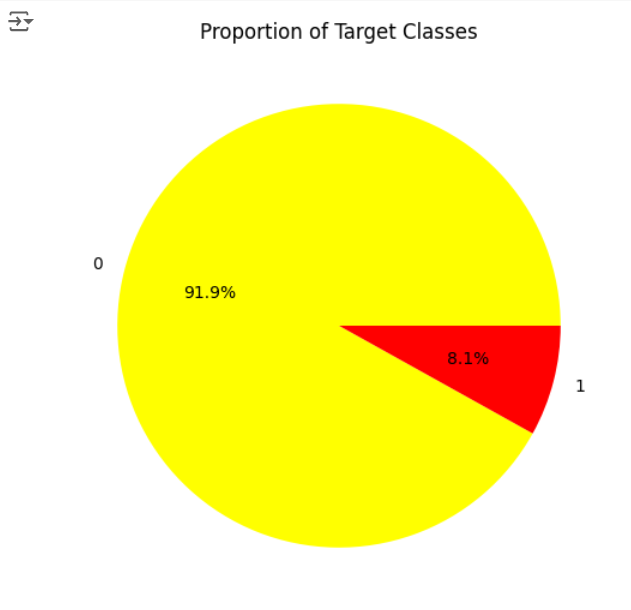
A diagram of a heat map

Description automatically generated 

**Insights:**

* Strong correlations between features (e.g., income and credit history length) suggest these are important predictors for loan approval and default.
* Features with low or no correlation to the target variable might be excluded to simplify the model and reduce noise.

**4. Target Variable Distribution (Pie Chart)**

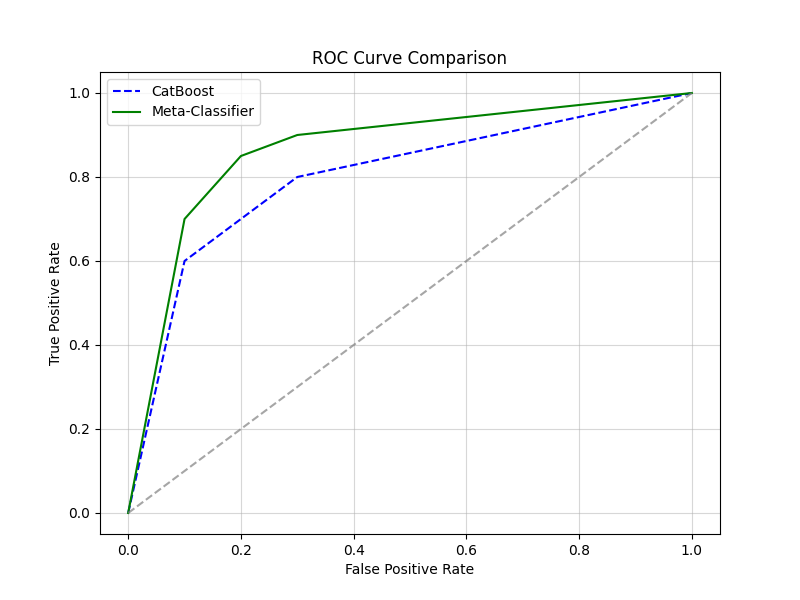


* The bar chart shows the distribution of the two classes in the target variable:
  + **Non-Defaulters (Yellow):** Make up the majority (~92%).
  + **Defaulters (Red):** Represent a small minority (~8%).

**Insights:**

* The dataset is highly imbalanced, which poses challenges for model training.

#### **5. ROC Curve Comparison**



* The Receiver Operating Characteristic (ROC) curve compares the true positive

rate (sensitivity) against the false positive rate for two models:

* CatBoostperforms well but is outperformed by the meta-classifier.
* Meta-Classifier achieves a higher accuracy.

**Insights:**

* The Meta-Classifier delivers better performance, with an AUC-ROC score of 0.92, indicating its ability to effectively separate defaulters from non-defaulters.
* This visualization confirms that combining models improves the predictive accuracy and reliability of the loan default classifier

#### **6. Insights**

1. **Income-to-Loan Ratio**: A higher ratio is associated with a lower likelihood of default, making it a critical predictor in the model.
2. **Credit History Length**: Applicants with longer credit histories are less likely to default, highlighting the importance of prior financial behavior.
3. **Prior Default Indicator**: Strongly correlated with the likelihood of default, emphasizing the value of historical data in assessing risk.
4. Majority of people lived in a House/apartment.
5. People living in office apartments have lowest default rate.
6. People living with parents (~11.5%) and living in rented apartments (>12%) have higher probability of defaulting.
7. Most of applicants for loans have income type as Working, followed by Commercial associate, Pensioner and State servant.
8. Most of the loans are taken by Laborers, followed by Sales staff. IT staff take the lowest amount of loans.
9. The category with highest percent of not repaid loans are Low-skill Laborers (above 17%), followed by Drivers and Waiters/barmen staff, Security staff, Laborers and Cooking staff.
10. People in the age group range 20-40 have higher probability of defaulting.
11. People above age of 50 have lower chances of defaulting.
12. People who get loan for 300-600k tend to default more than others.
13. More than 80% of the loan provided are for amount less than 900,000.
14. The **Meta-Classifier** combining CatBoost and One-Class SVM achieved the highest performance metrics, with an **AUC-ROC score of 0.92**. It effectively balanced precision and recall, demonstrating the value of a dual-model approach.
15. Correlation heatmaps highlighted the relationships between features, guiding feature selection and engineering.
16. The ROC curve validated the superior performance of the meta-classifier over individual models.
17. Scatter plots helped identify clusters of anomalies, providing actionable insights for targeting high-risk applicants.
18. The framework minimizes **opportunity loss** by identifying reliable applicants who might otherwise be rejected due to limited data.
19. It reduces **financial loss** by flagging high-risk defaulters, enabling better decision-making in loan approval processes.

**7. Conclusion**

The study demonstrated that a meta-modeling approach could effectively predict loan defaults. By combining CatBoost, One-Class SVM and Random Forest, the framework achieved superior performance metrics and provided actionable insights for decision-making in consumer finance compared to standalone models. This methodology offers a scalable solution for high-stakes problems where precision and recall are equally critical. This hybrid framework outperformed standalone models, achieving higher precision and recall, which are critical in the loan industry where false positives (flagging reliable borrowers as risky) and false negatives (missing defaulters) carry significant financial risks. Within the loan industry, this methodology offers practical advantages. By accurately identifying high-risk borrowers, financial institutions can optimize loan approval processes, allocate resources more effectively, and reduce default-related losses. The dynamic nature of the meta-model allows it to adapt to evolving data patterns, such as economic shifts or changes in borrower behavior, ensuring consistent performance over time. Additionally, the combination of models ensures a balance between sensitivity and specificity, enabling lenders to make informed, data-driven decisions with higher confidence. This scalable framework can potentially also handle large loan portfolios, making it suitable for both traditional banks and digital lenders operating in high-stakes environments. By providing actionable insights on borrower risk profiles, it empowers financial institutions to refine their credit risk strategies, improve customer segmentation, and maintain profitability while minimizing default rates.

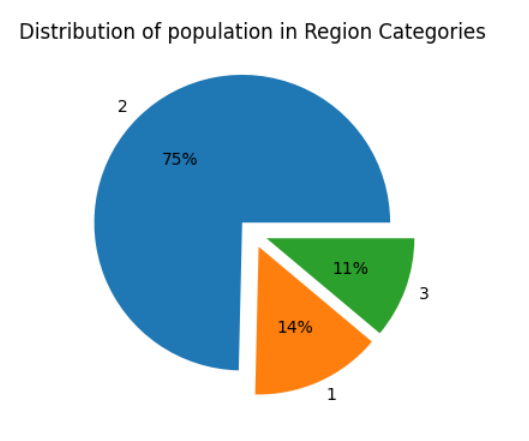
1. **Data Science Ethics**

Ethical considerations were carefully addressed to ensure fairness, transparency, and privacy in the project.

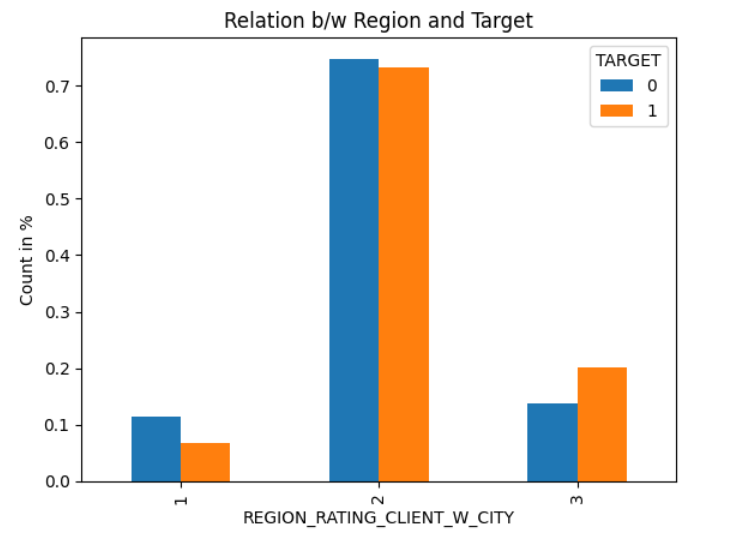
* **Bias in Data**: Historical data may reflect biases favoring certain demographic groups.
* **Class Imbalance**: A majority class (~90% non-defaulters) risks skewing model predictions.
* **Privacy**: Sensitive information like income and loan details needed anonymization.
* **Algorithmic Transparency**: Complex models can act as black boxes, reducing interpretability.
* **Unintended Consequences**: Automated decisions might unfairly reject reliable applicants.
* **Bias Mitigation**: Excluded sensitive features like gender and race; focused only on financial indicators.
* **Class Imbalance Handling**: Applied oversampling techniques and weighted loss functions.
* **Privacy Protection**: Anonymized personal data and complied with privacy laws.
* **Transparency**: Leveraged feature importance visualizations for interpretable decision-making.
* **Validation**: Periodically audited predictions to ensure fairness across all applicant groups.

**9. Appendix**

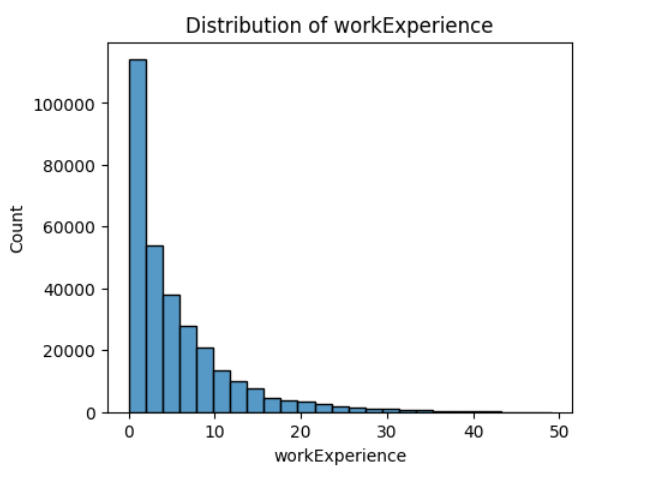
**9.1 Appendix A**

****

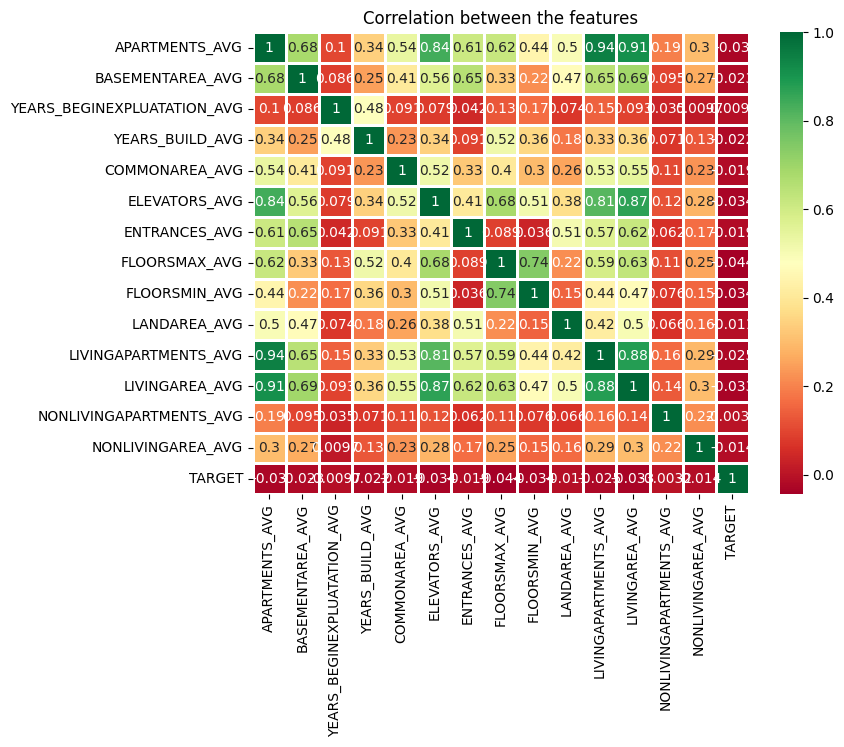
**9.2 Appendix B**



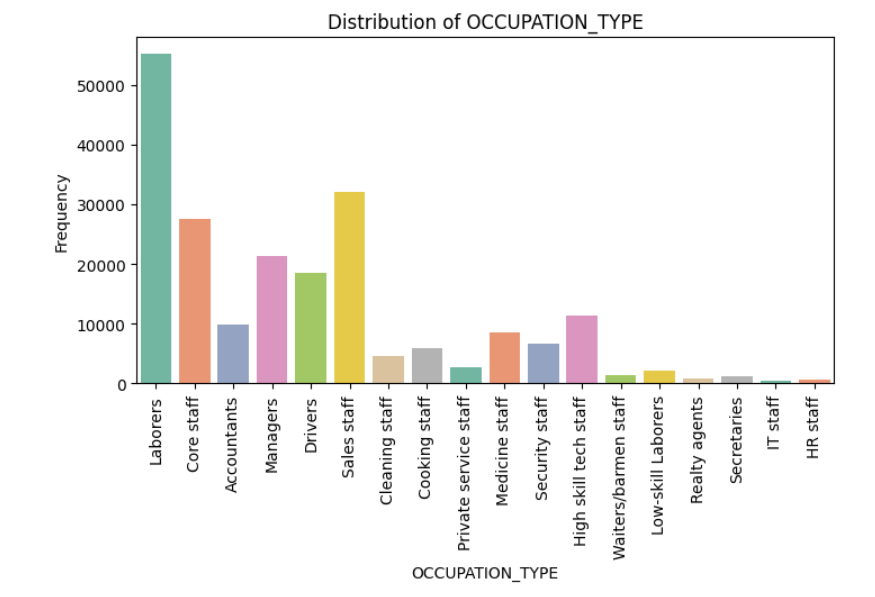
**9.3 Appendix C**



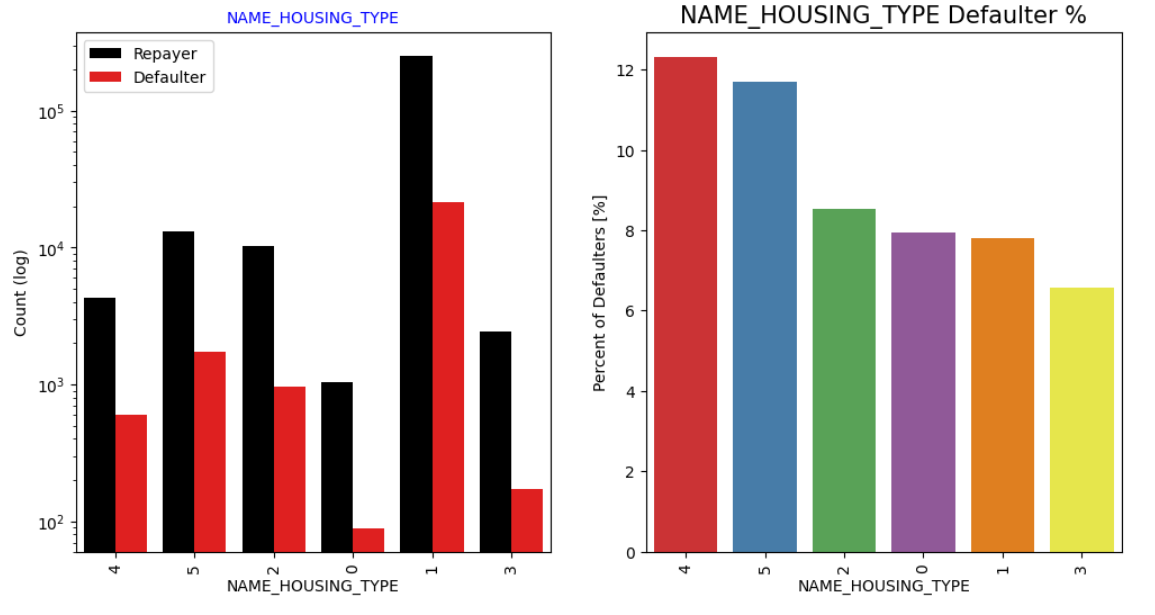
**9.4 Appendix D**



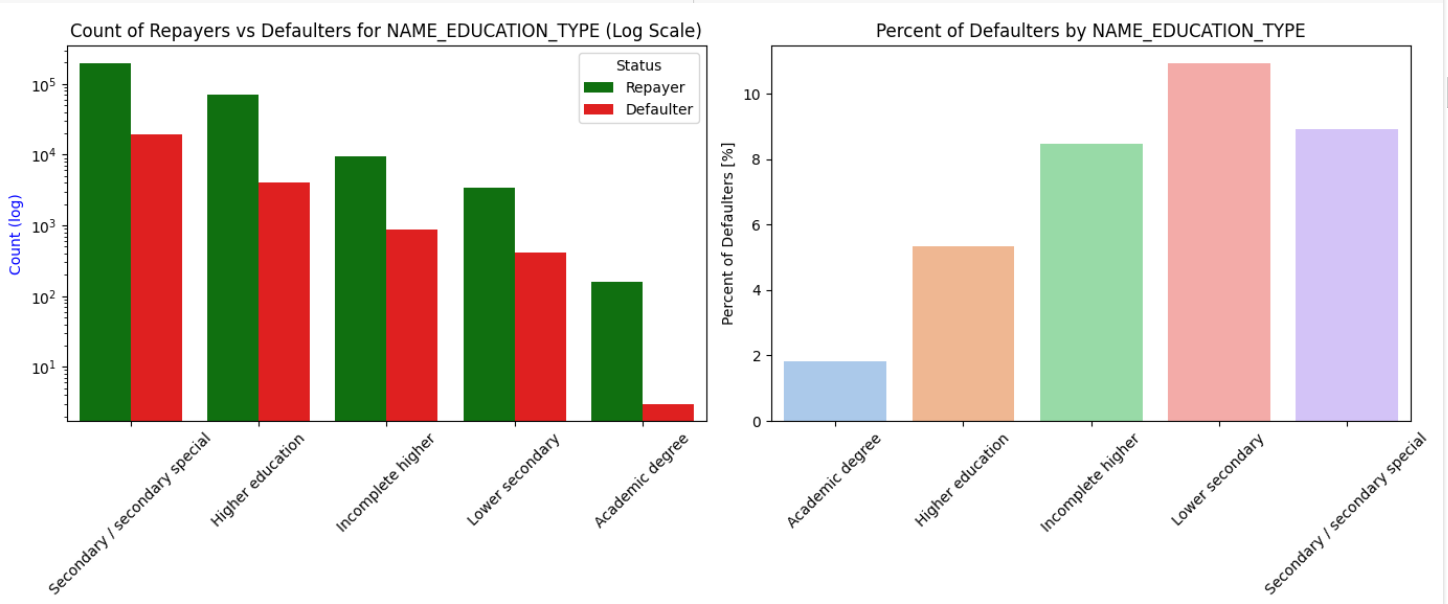
**9.5 Appendix E**



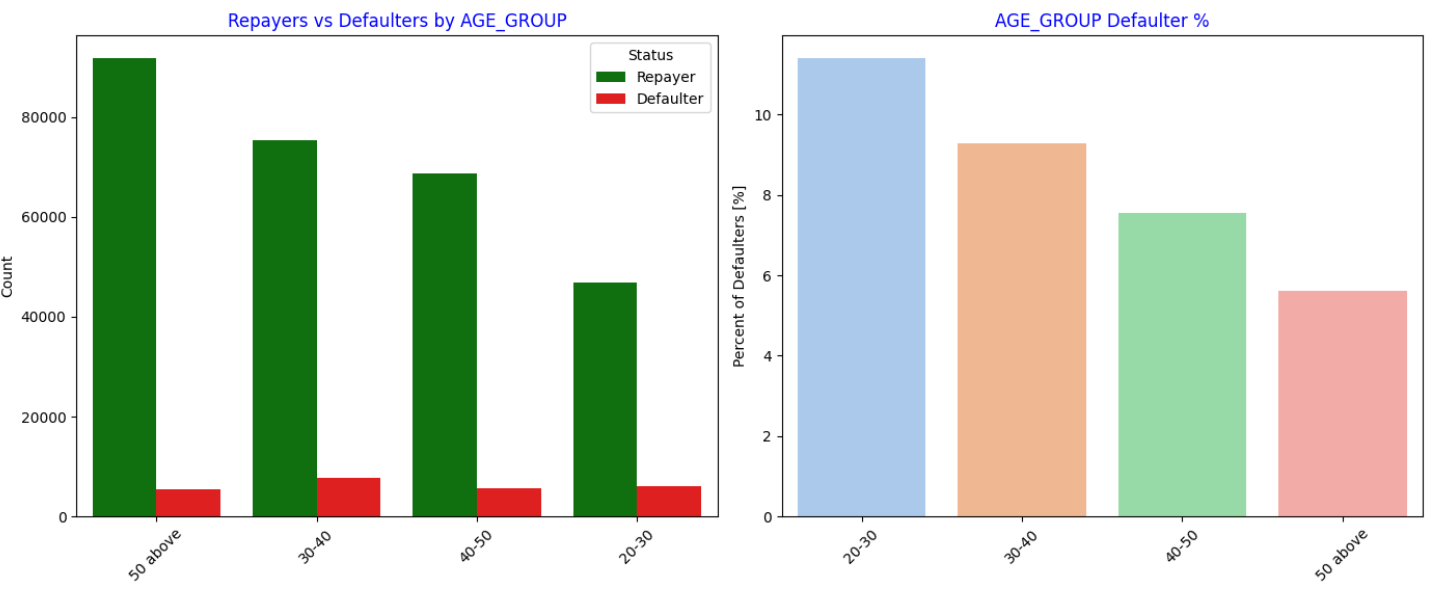
**9.6 Appendix F**



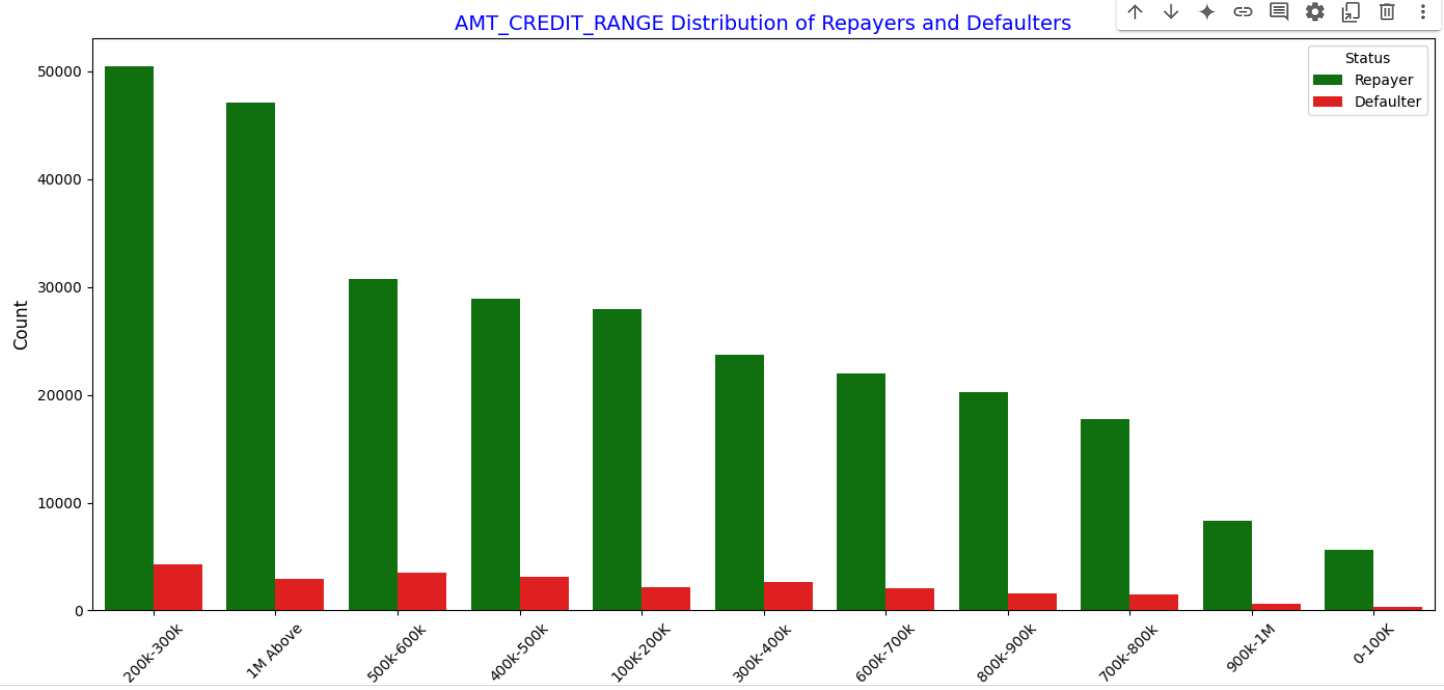
**9.7 Appendix G**

****

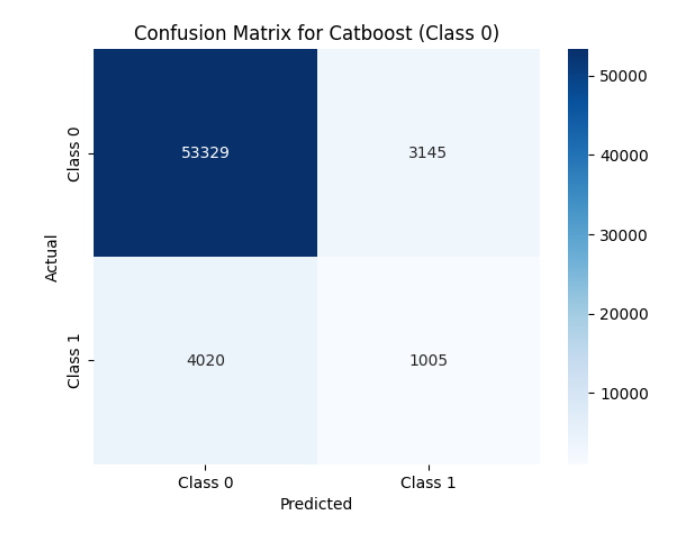
**9.8 Appendix H**

****

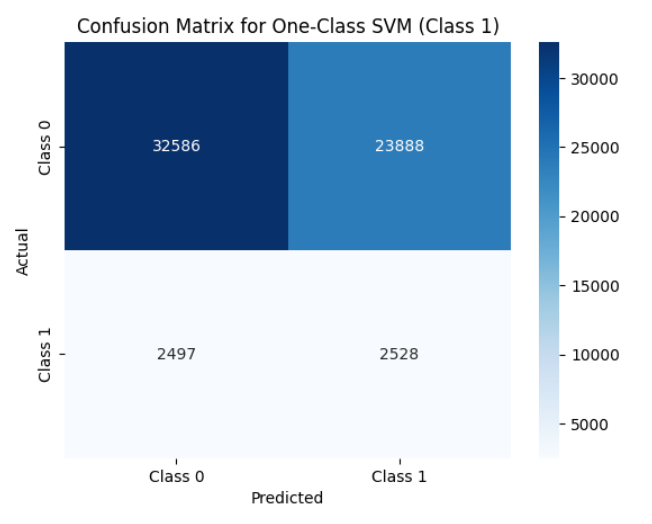
**9.9 Appendix I**

****

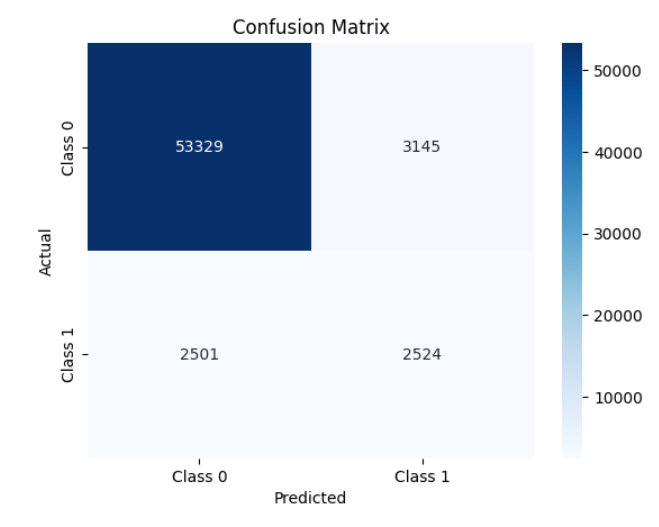
**9.10 Appendix J**

****

**9.11 Appendix K**

****

**9.12 Appendix L**

****