DS-GA 1017, RDS Course Project: Loan Eligibility Prediction

Anna Dominic
amd9200@nyu.edu

Manan Patel
mp6561@nyu.edu

1 Background

1.1 Purpose and Goals

This Automated Decision System (ADS) represents an integration of machine learning with the traditional loan approval process, engineered to enhance the efficiency and user experience within the financial sector. It functions by automating the analysis of a myriad of financial credentials, thus facilitating a swift and user-friendly loan application process for both the clients and the institutions. The ADS's capability to process extensive data rapidly allows it to offer timely predictions of creditworthiness and repayment likelihood, thereby expediting the decision-making process without sacrificing the precision and reliability of its assessments.

The ADS is meticulously crafted not only to quicken loan processing but also to refine client experience and operational fluidity for financial institutions. Simultaneously, it upholds a delicate balance between accelerating loan approvals and adhering to rigorous accuracy and risk management protocols. The system underscores a commitment to optimizing financial services that meet the speed of modern business while ensuring responsible lending practices.

1.2 Trade-offs

Speed vs. Accuracy: The ADS is engineered for expediency in loan processing, but this can potentially compromise the detailed scrutiny of applicants' financial profiles. The trade-off lies in balancing swift application processing with accurate, risk-aware decision making.

Automation vs. Personalized Assessment: The ADS's efficiency through automation might not cater to the unique financial contexts of all applicants, unlike manual assessments. The trade-off here is operational efficiency versus the nuanced understanding a human evaluator provides.

Overall Accuracy vs. Fairness: The ADS may prioritize overall accuracy in loan decisions, potentially leading to disparities in treatment across different demographic or socio-economic groups. The trade-off is between maximizing predictive accuracy and ensuring fairness in lending practices. Balancing these objectives requires careful consideration of model performance metrics, fairness measures, and potential biases in decision-making.

2 Input and Output

2.1 Data Description

This ADS uses data from the Loan Eligible Dataset that is accessible on Kaggle, claimed to be used by the Dream House Financing Company.

The dataset contains Personal information, Employment details, Financial metrics including Applicant and Co-Applicant Incomes, Loan specifics such as Loan Amount, Loan Status, and the applicant's Credit History along with the geographical category of the applicant's residence.

The dataset most likely received preprocessing to clean and format the data before being pushed to Kaggle for public use.

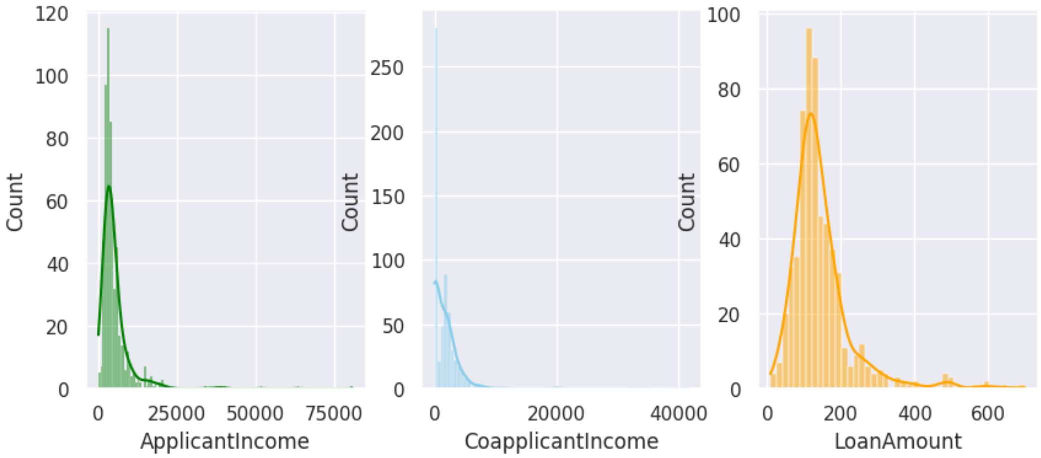

Figure 1: Skew Visualizations of the original data

2.2 Input Features

The dataset utilized by the Automated Decision System (ADS) for predicting loan eligibility comprises various input features with a mix of data types. The heatmap depicted in Figure 3 provides a visual representation of the missing data correlation in the dataset, whereas the pairwise correlation heatmap in Figure 4 illustrates the relationships between numerical features in the dataset. Below is a detailed description of each input feature:

- **Loan_ID**: A unique identifier for the loan application. It is a nominal datatype with no missing values.
- **Gender**: Categorical data indicating male or female applicants, with 13 missing values.
- **Married**: Categorical data indicating marital status (Yes/No), with 3 missing values.
- **Dependents**: Categorical data representing the number of dependents, with 15 missing values. It exhibits a moderate positive correlation with the **Married** status, indicating a pattern in missing data.
- **Education**: Categorical data signifying applicant education level (Graduate/Undergraduate), with no missing values.
- **Self_Employed**: Categorical data indicating self-employment status (Yes/No), with 32 missing values.
- **ApplicantIncome**: Numeric data representing the applicant's income. There are no missing values, and it has a moderate positive correlation (0.57) with **LoanAmount**.
- **CoapplicantIncome**: Numeric data for the coapplicant's income, also with no missing values. It has a weak positive correlation (0.19) with **LoanAmount**.
- **LoanAmount**: Numeric data indicating the loan amount in thousands. There are 22 missing values and a very weak correlation with **Loan_Amount_Term** and **Credit_History**.
- **Loan_Amount_Term**: Numeric data representing the term of the loan in months, with 14 missing values. It shows very little correlation with other features.
- **Credit_History**: Numeric data indicating whether the credit history meets the guidelines, with 50 missing values. There is essentially no correlation with other features.

Value distributions for numerical features like **ApplicantIncome**, **CoApplicant Income** and **LoanAmount** are right-skewed, indicating a concentration of applicants with lower incomes, loan amounts and coapplicant incomes as shown in Figure 1.

The bar chart shown in Figure 2 provides a visual representation of the completeness of each feature in the dataset. The y-axis represents the proportion of non-missing values for each variable, with the total number of observations at the top of each bar. Features such as *Loan_ID*, *Education*, and

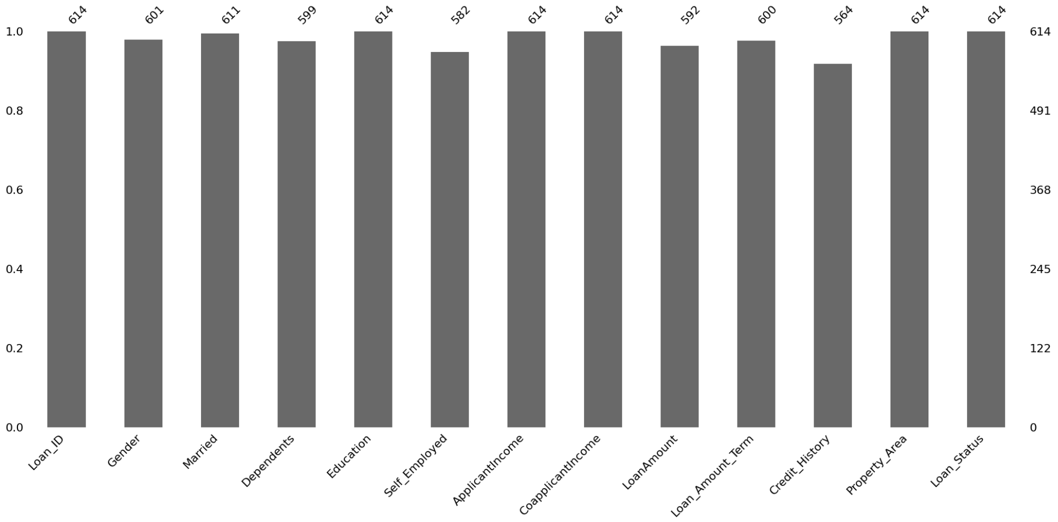

Figure 2: Bar chart illustrating the completeness of each feature in the dataset.

Property_Area have no missing values, indicating complete data capture for these categories. However, variables such as *Gender*, *Married*, and *Credit_History* exhibit some degree of missingness. This visualization is crucial for identifying patterns of missing data, which may inform data preprocessing strategies such as imputation or omission, and from the heatmap in Figure 3, we can also see that there is little to no patterns between missing values in the dataset, which means we can safely remove these rows without skewing or distorting our analysis.

2.3 Data Preparation and Preprocessing

This section details the steps taken to prepare and preprocess the dataset for model training and evaluation. We evaluated three versions of the data preparation process to understand the impact of different preprocessing techniques on model performance and fairness. In this report we will refer to these three implementation versions as follows: the Original Implementation, With Imputation Implementation and Without Imputation Implementation. Further details on these implementations are given below:

2.3.1 Original Implementation

This is the ADS that we are auditing. While going through the code given for this ADS, we noticed some key errors in the preprocessing steps that can potentially lead to bias and distort the model's predictions while inflating accuracy. These errors are given below:

- **SMOTE Application**: SMOTE was applied to the entire dataset, including the test set, leading to potential data leakage. This also leads to artificially created records for the test set, which do not represent the real-world data points.
- **Scaling**:The MinMaxScaler was applied to the entire dataset instead of fitting it on the training set only, which could introduce bias. This leads to data leakage by allowing information from the test set to influence the scaling process.
- **Outlier Removal**:The implementation included outlier removal, which could have resulted in the loss of important data points. Although extreme, these data points cannot be considered false or unimportant and thus must be included.
- **Skewed Distribution Treatment**: A log transformation was applied to treat skewed distributions, potentially altering the data's natural structure. This leads to completely inaccurate data points for the model to train on and thus model cannot be fit properly.

Furthermore, the ADS also used mean and mode imputation to fill in missing values for numeric and categorical columns respectively. To investigate how this choice of imputation might introduce technical bias into our model, we implemented two other versions of the model, with imputation and without, after rectifying the above mistakes. More details to follow in the subsequent sections.

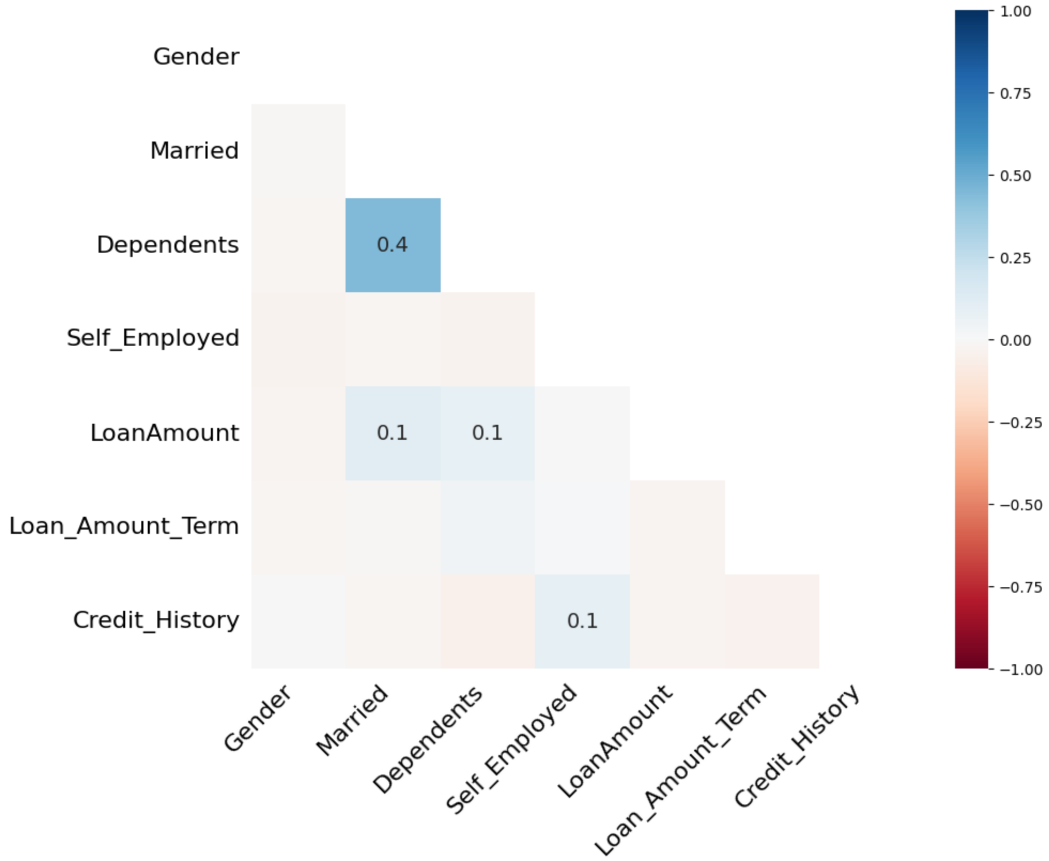

Figure 3: Heatmap illustrating the correlation of missing data between features in the dataset.

2.3.2 With Imputation Implementation

To rectify the identified mistakes, while retaining the choice of mean and mode imputation, this version of the data preparation process was implemented as follows:

- **SMOTE Application**: Applied only to the training set to avoid data leakage.
- **Scaling**: MinMaxScaler was fitted on the training set and then used to transform both the training and test sets.
- **Retention of Outliers**: No outliers were removed to maintain the integrity of the dataset.
- **Natural Skewness Retained**: Skewed distributions were not log-transformed, preserving the original data characteristics.

2.3.3 Without Imputation Implementation

In this version, in addition to the steps taken in the above implementation, instead of imputing missing values using mean or mode, rows with NaN values were dropped, simplifying the preprocessing pipeline:

- **NaN Removal**: Directly removed rows with missing values to avoid potential biases introduced by imputation.

2.4 System Output

The system's prediction of the loan approval status, represented by a binary classification label (Y/N), is its output. The proposed model used in the ADS is Random Forest Classification because it was the model that performed best in the original implementation (Table 1) which operates by aggregating the predictions of multiple decision trees.

- Bootstrap Sampling (Bagging):

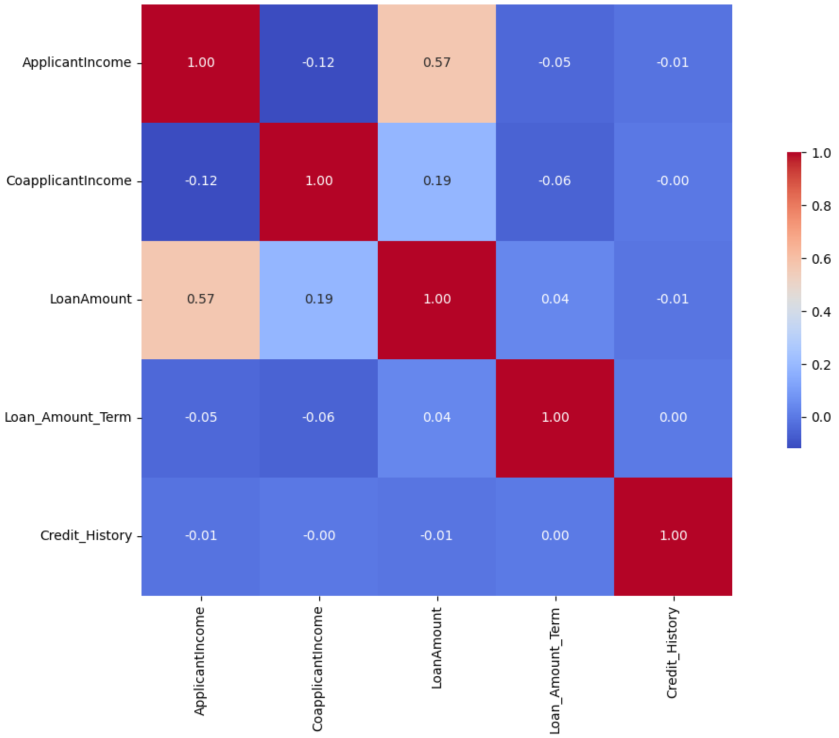

Figure 4: Heatmap showing the pairwise correlations between numerical features in the dataset.

  - Random Forest builds multiple decision trees by sampling the training data with replacement.
  - Each tree is trained on a different subset of the data.
- Decision Trees:
  - Each decision tree in the Random Forest is trained independently using a subset of randomly selected features at each node.
  - The trees grow deep enough to minimize training error but avoid overfitting.
- Voting (Classification):
  - For classification tasks, each decision tree predicts the class label for each observation.
  - The final prediction is made by majority voting among all the decision trees.

$$\hat{y} = \text{mode}\left(\{f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_N(\mathbf{x})\}\right) \tag{1}$$

- Probability Estimation:
  - Random Forest estimates the probability of each class for a given observation by taking the proportion of trees that predict each class.

This prediction is crucial for financial institutions to make informed decisions regarding loan approvals, thereby optimizing their processes and improving customer satisfaction.

3 Implementation and Validation

3.1 Data Cleaning and Pre-processing

The following data cleaning and pre-processing steps were undertaken in the implementation:

Handling Missing Values: The original implementation handled missing values by filling in the most frequent value (mode) for categorical variables and the mean for numerical variables. However, filling missing values with the mode for categorical variables can lead to a homogenization effect, especially if the mode does not fairly represent minority groups. For example, since *Gender* is predominantly male, imputing missing values with the mode could disproportionately increase the representation of males, potentially introducing bias.

Outlier Removal: Outliers were identified and removed using the Interquartile Range (IQR) method. This method considers values outside 1.5 times the IQR from the first and third quartiles as outliers. This step could inadvertently remove valid instances that belong to minority groups, thus potentially reducing the representativeness of these groups in the dataset and introducing bias.

Skewed Distribution Treatment: The implementation applied log transformation to variables with skewed distributions to normalize the data. However, transforming data could obscure the true distribution of sensitive features, potentially affecting the model's ability to fairly assess these features.

One-Hot Encoding of Categorical Variables: Categorical variables were transformed into a numerical format using one-hot encoding, where each category is represented by a separate binary column. This is a standard method and ncoding preserves the distinctions between different categories, helping the model to learn from each category accurately without introducing unintended bias.

Data Splitting and Resampling (SMOTE): The original implementation applied SMOTE (Synthetic Minority Over-sampling Technique) to the entire dataset to balance the dataset by generating synthetic samples for the minority class. However, applying it to the entire dataset introduces bias as previously discussed.

Data Scaling (Normalization): The MinMaxScaler was applied to the entire dataset, scaling all features to a range between 0 and 1. While proper scaling ensures that all features contribute equally to the model and avoids bias due to feature magnitude differences, in this case it distorts the model and inflates accuracy since the scaler was fit on the test set as well instead of just the training set.

3.2 Implementation Overview

Feature Selection: All features except *LoanID* were selected due to their relevance to the task at hand.

Choice of Algorithm (Random Forest): The primary algorithm chosen for the loan approval prediction task is the Random Forest classifier.

3.3 Validation of the ADS

Validation Methods:

- **Holdout Validation**: The dataset was split into training and test sets, with the training set used for model training and the test set reserved for validation.
- **Cross-Validation**: Cross-validation (within GridSearchCV), was employed to further assess the model's performance.

Performance Metrics:

- **Accuracy**: Accuracy provides a general measure of model performance, crucial for understanding the overall correctness of loan approval predictions.
- **Precision**: Precision is essential in loan approval scenarios to minimize the risk of approving bad loans (false positives), ensuring that most loans approved are indeed good.
- **Recall**: Recall is critical in ensuring that all potential good loans are identified, reducing the chance of missing out on approving eligible applicants (false negatives).
- **False Positive Rate (FPR)**: In the context of loan approval, a lower FPR is desirable to minimize the number of bad loans approved, thereby reducing financial risk.
- **False Negative Rate (FNR)**: A lower FNR is crucial to ensure that eligible applicants are not wrongly denied loans, promoting fairness and inclusivity.
- **Equalized Odds Ratio**:Ensuring equalized odds helps in assessing and mitigating bias, ensuring that the loan approval model is fair across sensitive attributes.

Model	Accuracy (%)
Random Forest	86.67
K Neighbors	82.22
SVM	80.00
Logistic Regression	77.78
Categorical NB	77.78
Gradient Boost	77.78
Decision Tree	75.56
Gaussian NB	73.33

Table 1: Comparison of Model Accuracies

- **Selection Rate Difference**: Monitoring selection rate differences is key to ensuring that loan approval rates are equitable across different demographic groups, promoting fairness.

4 Outcome:

Metrics	Original Implementation	With Imputation	Without Imputation
Accuracy	0.84	0.83	0.79
Precision	0.81	0.82	0.75
Recall	0.85	0.98	1.00
Selection Rate	0.50	0.87	0.84
False Positive Rate (FPR)	0.18	0.58	0.57
False Negative Rate (FNR)	0.15	0.02	0.00

Table 2: Comparison of Overall Metrics Across Different Implementations

Table 2 presents a detailed comparison of overall metrics for the loan eligibility prediction ADS across three different implementations: the original implementation, the implementation with imputation, and the implementation without imputation.

4.1 Auditing Methodology

To ensure the robustness, fairness, and transparency of the Automated Decision System (ADS) for loan eligibility prediction, we employed multiple auditing tools and methodologies, including Fairlearn, Aequitas, and SHAP (SHapley Additive exPlanations). Each tool provided unique insights into different aspects of model performance and fairness.

Fairlearn We utilized Fairlearn to measure and understand the fairness impact of our ADS across different demographic groups. The methodology applied involved evaluating group fairness metrics such as selection rate and disparate impact, which helped us understand how the ADS's decisions varied across groups defined by sensitive attributes like gender and marital status. Additionally, Fairlearn's interactive dashboard was employed to visualize and compare the performance and fairness of the ADS across different subgroups, facilitating a detailed examination of trade-offs between accuracy and fairness.

Aequitas Using Aequitas, we generated comprehensive fairness reports and metrics. The methodology involved detecting biases across multiple metrics, including false positive rates (FPR), false negative rates (FNR), and accuracy, evaluated across different groups to identify any disparities. We performed a detailed analysis of group metrics to understand how the ADS's performance differed across various intersections of gender and marital status, providing insights into the model's treatment of different demographic subgroups. Furthermore, Aequitas's visualization tools were leveraged to plot group metrics, enabling a clear comparison of fairness metrics across different groups, which helped identify specific areas where the ADS might require improvements.

SHAP (SHapley Additive exPlanations) We utilized SHAP's visualization tools, such as summary plots, dependence plots, and waterfall plots, to illustrate the effect of each feature on the model's output, providing a clear interpretation of the model's decision-making process. Additionally, SHAP was used to generate local explanations for individual predictions, highlighting how each feature influenced specific decisions, thus helping in diagnosing potential biases and understanding model behavior on a case-by-case basis.

4.1.1 Overall Metrics

The comparative analysis of overall metrics across different data preprocessing implementations reveals distinct trends in model performance. From the bar-plot comparison presented in Figure 5, it is evident that each preprocessing method has unique impacts on the model's metrics. Imputation generally enhances recall and selection rate but also increases the false positive rate, leading to a higher number of bad loans being approved. Conversely, the no imputation method shows similar improvements in recall and selection rate compared to imputation, with a slightly lower false positive rate, making it a more balanced choice. The original method, characterized by the lowest false positive rate, exhibits lower recall and selection rate, indicating a conservative approach to loan approvals, potentially resulting in the denial of more eligible applicants.

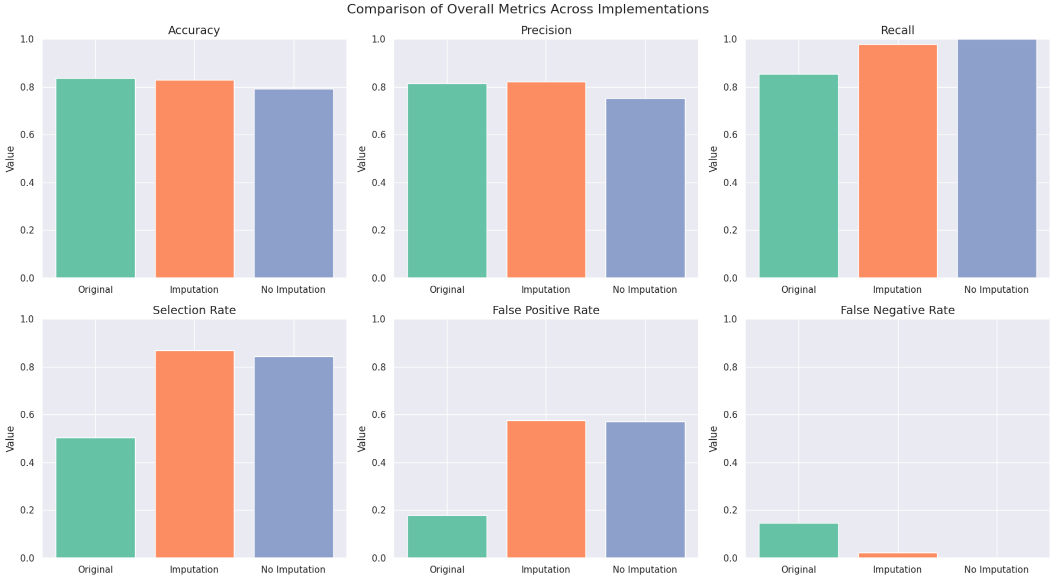

Figure 5: Comparison of Overall Metrics Across Implementations

4.1.2 SubGroup Analysis

Original Implementation: For the Gender 0.0 (Female) group, the predicted counts for both ineligible and eligible loans closely match the actual counts, demonstrating strong prediction performance (Figure 11 and 10). The Random Forest classifier has accurately predicted the eligibility status for most instances in this group. For Gender 1.0 (Male), the predicted counts for ineligible loans are almost identical to the actual counts. However, there is a slight underestimation in the eligible count (~50 predicted vs. ~55 actual), indicating a minor underestimation in loan eligibility predictions for this group.

With Imputation: For the Married = 0.0 (Unmarried) group, there is an underestimation of ineligible loans, with predicted counts lower than actual. The eligible counts are accurately predicted, matching the actual counts. For Married = 1.0 (Married), the predicted counts for both ineligible and eligible loans closely match the actual counts, indicating accurate prediction performance for the married group.

Without Imputation: In the Gender and Marital Status intersectional analysis, the classifier performs well across all categories, with predictions generally aligning closely with actual values. Notable discrepancies include a slight overestimation of eligible loans in the Female_Unmarried group and a slight underestimation of ineligible loans in the Male_Unmarried group. The other groups have predictions that closely match the actual values, suggesting minimal bias. The Random Forest classifier demonstrates strong performance, particularly for Male_Married and Female_Married groups, while the Female_Unmarried and Male_Unmarried groups show slight discrepancies that could be areas for further model refinement.

Summary: Across different subgroups, the classifier exhibits high accuracy and balanced performance, with minor improvements needed in specific areas. The Male group shows a slight underestimation in eligible loan predictions, while the Unmarried group shows an underestimation in ineligible loan predictions. The intersectional analysis reveals that while the classifier generally predicts eligibility accurately, minor improvements could reduce overestimation for Female_Unmarried and underestimation for Male_Unmarried.

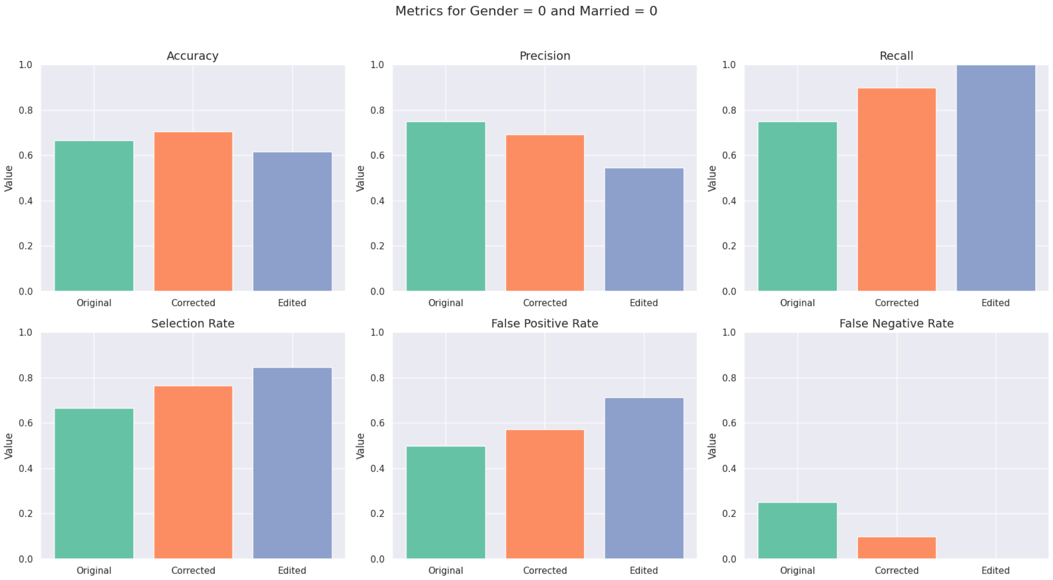

Figure 6: Comparison of Metrics for Gender = 0 and Married = 0

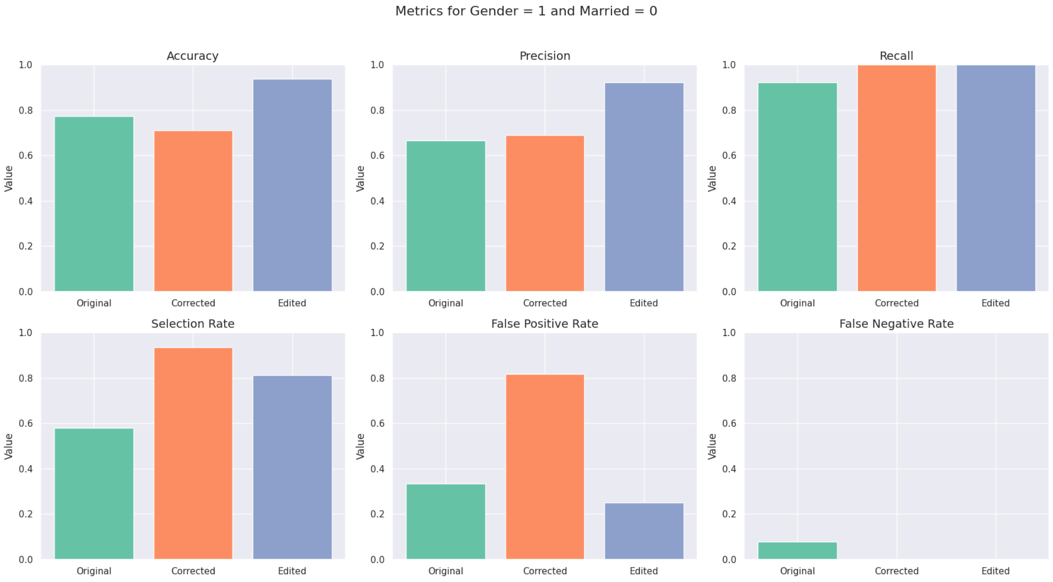

Figure 7: Comparison of Metrics for Gender = 1 and Married = 0

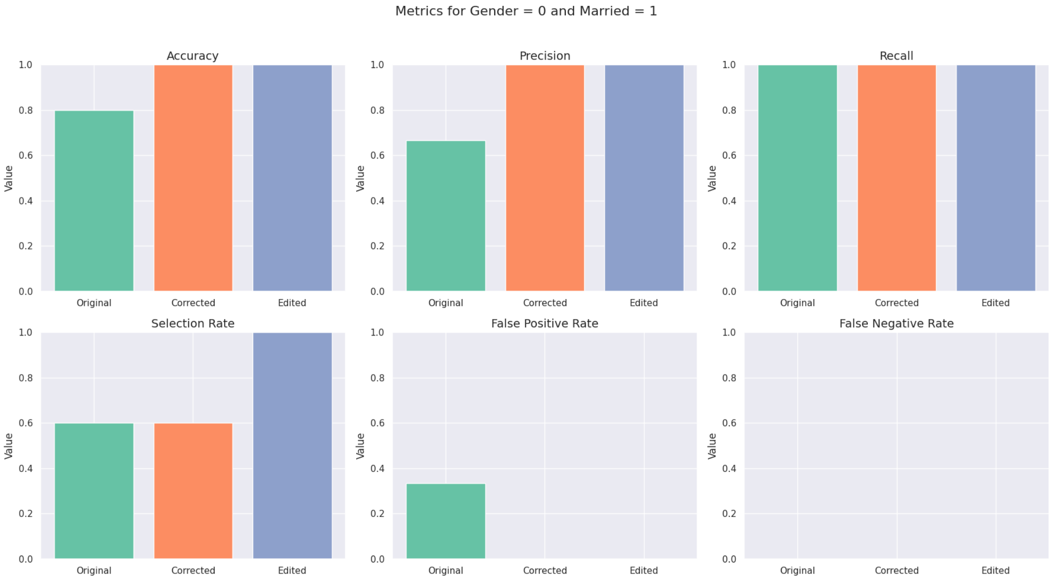

Figure 8: Comparison of Metrics for Gender = 0 and Married = 1

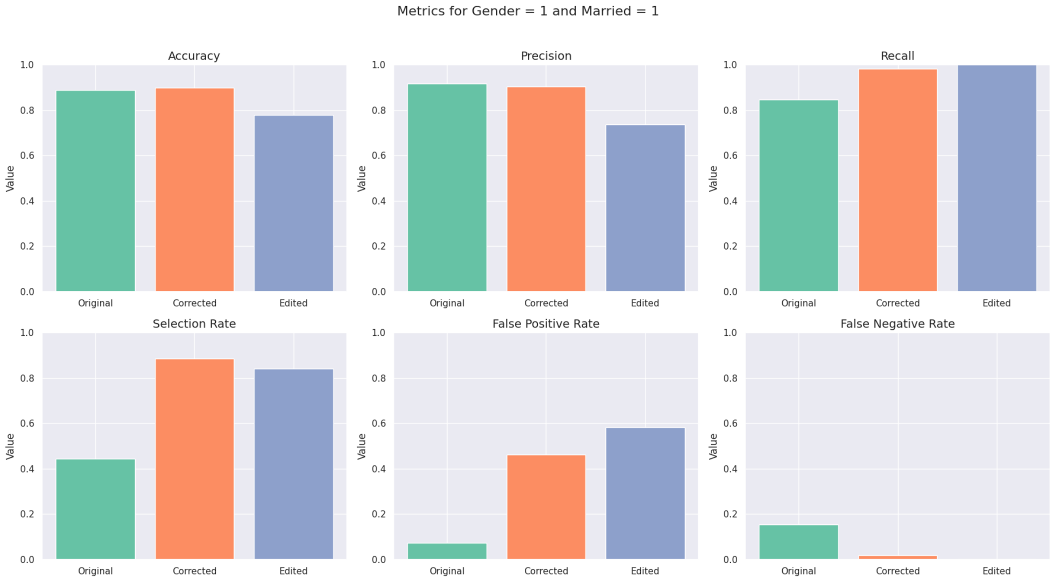

Figure 9: Comparison of Metrics for Gender = 1 and Married = 1

4.2 Overall Conclusion

In conclusion, the analysis of overall metrics and subgroup performance highlights the strengths and areas for improvement in the Random Forest classifier's performance. The classifier demonstrates commendable accuracy and balanced prediction performance across different demographic groups, with specific biases observed in gender and marital status categories. Males and married individuals, particularly males who are married, enjoy higher accuracy, recall, and precision, while females and unmarried individuals face higher false positive and false discovery rates, leading to more prediction errors. The intersectional analysis further reveals that the model performs best for Male_Married and worst for Female_Married, underscoring the need for more equitable performance across all demographic intersections. Overall, while the classifier's performance is strong, targeted refinements could enhance fairness and accuracy, particularly for underrepresented and minority groups.

5 Summary

This section reflects on the overall findings and implications of the Automated Decision System (ADS) for loan eligibility prediction.

5.1 Data Appropriateness

The dataset used for this ADS was deemed appropriate given its comprehensive coverage of relevant personal, financial, and loan-specific attributes necessary for predicting loan eligibility. The inclusion of demographic details such as Gender and Marital Status, along with financial metrics like Applicant Income and Credit History, provided a robust foundation for training the model. However, the presence of missing values and skewed distributions necessitated careful preprocessing to ensure data quality and integrity.

5.2 Implementation Robustness, Accuracy, and Fairness

The Random Forest classifier, chosen for its robustness and strong performance in handling diverse feature sets, demonstrated high accuracy and balanced performance across various demographic groups. The choice of accuracy and fairness measures, including Precision, Recall, False Positive Rate (FPR), False Negative Rate (FNR), and Selection Rate Difference, was critical in evaluating the model's effectiveness and equity. These metrics are particularly relevant for stakeholders such as financial institutions and regulatory bodies focused on fair lending practices and risk management.

The classifier showed commendable accuracy, particularly in predicting loan eligibility for male and married applicants, with minor discrepancies noted in the Female and Unmarried categories. This suggests a slight bias that needs to be addressed to ensure fairness across all demographic intersections. The high recall and selection rate with imputation, contrasted by the increased false positive rate, underscore the trade-offs between recall and precision in the decision-making process.

5.3 Deployment Considerations

Deploying this ADS in the public sector or industry could be considered with some caution. While the model exhibits strong performance and general robustness, the observed biases—especially those favoring married individuals and males—warrant further refinement. The fairness measures indicate that while the model performs well overall, ensuring equitable treatment across all demographic groups is crucial. Therefore, continuous monitoring and adjustment of the model would be necessary to maintain fairness and accuracy over time.

5.4 Recommendations for Improvement

Several recommendations are suggested to enhance the data collection, processing, and analysis methodology of the ADS:

- **Data Collection:** Enhance data collection practices to reduce missing values and ensure a more balanced representation of all demographic groups. This can help mitigate biases introduced during data preprocessing.
- **Data Processing:** Implement more sophisticated imputation techniques or consider retaining natural data distributions to preserve the inherent characteristics of the dataset.
- **Model Analysis:** Continuously assess and refine the model to address any emergent biases, particularly those affecting minority or underrepresented groups. This includes periodic recalibration of fairness measures to ensure ongoing equity in loan approvals.
- **Bias Mitigation:** Introduce bias mitigation strategies such as reweighting or adversarial debiasing to further enhance the fairness of the model, particularly in handling sensitive attributes like Gender and Marital Status.

5.5 Conclusion

In conclusion, the Random Forest classifier shows strong potential for effective loan eligibility prediction, with robust accuracy and balanced performance across most demographic groups. However, to ensure the ADS is truly fair and equitable, continuous improvements in data processing and model refinement are essential. By addressing the identified biases and enhancing the model's fairness, this ADS can be confidently deployed to support fair and efficient loan approval processes in both the public and private sectors.

References

[1] Caesar, Mario. *Loan Eligibility Prediction Machine Learning*. Kaggle, [Online]. Available: `https://www.kaggle.com/code/caesarmario/loan-prediction-w-various-ml-models`

[2] Loan Eligible Dataset. *Available on Kaggle*. [Online]. Available: `https://www.kaggle.com/loan-eligible-dataset`

A Appendix

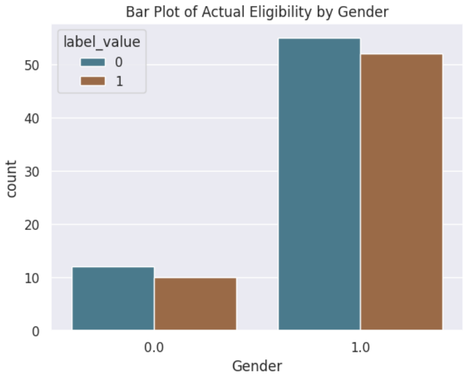

Figure 10: Bar Plot of Actual Eligibility by Gender

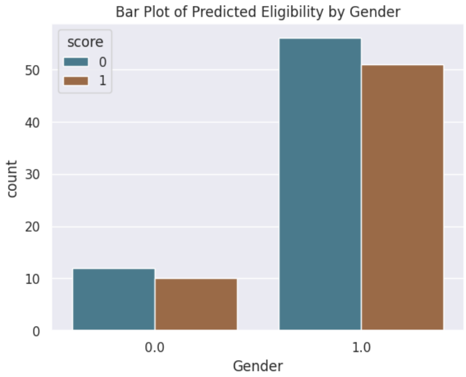

Figure 11: Bar Plot of Predicted Eligibility by Gender

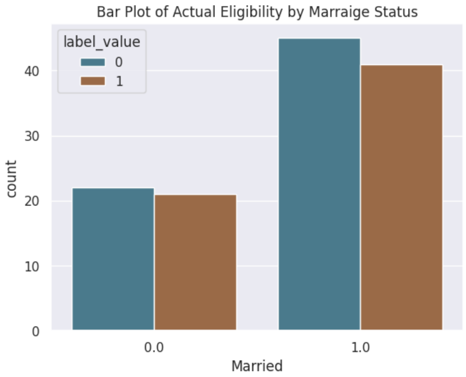

Figure 12: Bar Plot of Actual Eligibility by Marital Status

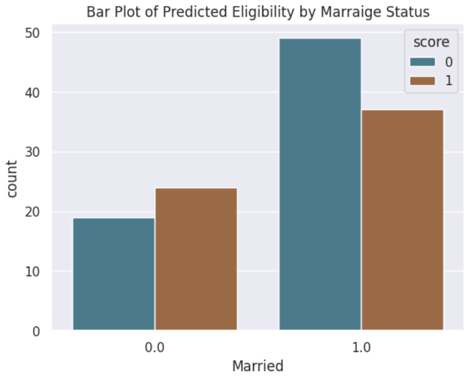

Figure 13: Bar Plot of Predicted Eligibility by Marital Status

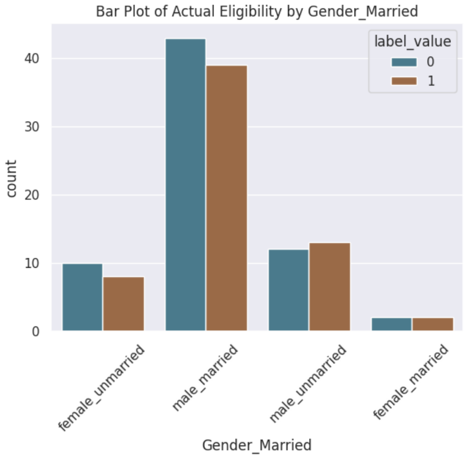

Figure 14: Bar Plot of Actual Eligibility by Gender and Marital Status

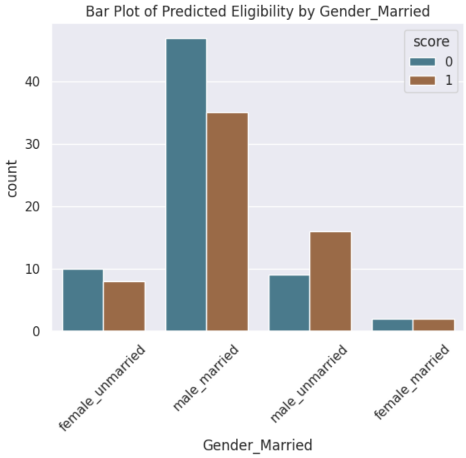

Figure 15: Bar Plot of Predicted Eligibility by Gender and Marital Status