Improving Transparency and Fairness in Loan Approval Algorithms Through Explainable AI

Final Project – Rutgers University (Two-person team)

Abstract

This project focuses on addressing biases and enhancing transparency in loan approval algorithms using Explainable AI (XAI) techniques. Loan approval systems often lack transparency, potentially leading to discriminatory decisions that disproportionately affect marginalized groups. By analyzing the Home Mortgage Disclosure Act (HMDA) 2022 dataset for New Jersey, we identify biases in factors such as gender, race, and income. Using XAI, we interpret model decisions to ensure accountability and fairness. The findings aim to promote equitable loan approval practices, fostering trust among stakeholders and compliance with regulatory standards.

Introduction

Loan approval processes frequently rely on machine learning models that function as "black boxes," where decisions are made without clear explanations. This lack of transparency raises ethical concerns, particularly regarding biases that may unfairly disadvantage specific demographics. This project explores how Explainable AI can be leveraged to analyze and mitigate such biases, ensuring fairness in loan approvals.

This problem is particularly significant because access to fair and unbiased financial opportunities is a cornerstone of economic justice. Existing studies have explored fairness in machine learning, but few have focused on applying XAI techniques to financial datasets like HMDA 2022. Our work contributes to the field by proposing methods to enhance fairness and interpretability in these systems, offering stakeholders actionable insights.

Our objectives include identifying and analyzing potential biases in loan approval processes, leveraging fairness algorithms and explainable AI (XAI) techniques to enhance model transparency, and delivering interpretable insights to stakeholders to promote trust and accountability.

Background

The research by [5] Genovesi et al. (2023) in their article *"Standardizing fairness-evaluation procedures: interdisciplinary insights on machine learning algorithms in creditworthiness assessments for small personal loans"* addresses the challenges of standardizing fairness-evaluation procedures in AI systems used for creditworthiness assessments. The paper highlights three significant fairness concerns: unequal distribution of predictive outcomes, perpetuation of existing biases, and a lack of transparency in algorithmic decision-making. To address these issues, the authors propose minimal ethical standards, such as regular checks for demographic parity, the exclusion of discriminatory parameters, and ensuring transparency and explainability in algorithmic decisions. This framework aims to prevent unfair outcomes and practices in AI-driven credit assessments, contributing to a more equitable and transparent financial system.

In [6] *"Algorithmic Fairness"*, Das, Stanton, and Wallace (2023) review the challenges and advances in algorithmic fairness, particularly in credit scoring and mortgage lending. Despite the elimination of face-to-face interactions in lending decisions, they note that biases related to gender, age, and ethnicity persist. The authors apply fairness metrics to mortgage data, revealing imbalances in loan approval for women and minority groups. Although modern machine learning (ML) techniques outperform traditional methods in prediction accuracy, their complexity makes them harder to explain to applicants and regulators. This article stresses the need for further research into debiasing datasets and improving regulatory standards to ensure fairness in AI/ML decision-making processes.

Ayad et al. (2023), in their work [1] *"A Proposed Model for Loan Approval Prediction Using XAI"*, explore the use of Explainable AI (XAI) to address biases in loan approval decisions. By utilizing the "Give Me Some Credit" dataset, they propose a model using machine learning algorithms like Random Forest and Support Vector Machines, with fairness metrics such as demographic parity and equal opportunity to evaluate and reduce bias. Their study integrates explanation techniques, including SHAP, Local and Global Surrogates, and Partial Dependence Plots, which offer clear, interpretable insights into model predictions. This approach improves transparency, helping stakeholders understand AI decisions and fostering trust in the loan approval process.

The research by Azith, Sushma, and Swathi (2024) in [2] *"Revolutionizing Loan Approvals with Explainable AI"* proposes an innovative approach to loan underwriting through the integration of belief-rule-based (BRB) systems with Explainable AI techniques. Their system, which incorporates both factual and heuristic rules, provides clear explanations for loan decisions while improving scalability and transparency. By demonstrating how the BRB system balances predictive accuracy and explainability, they showcase its potential for enhancing loan approval processes.

Govindu et al. (2024) in [3] *"Loan Prediction System with Exploring Explainable AI Transfer Learning with SHAP"* highlight the application of SHAP to loan prediction models, emphasizing its role in improving transparency and interpretability. By visualizing feature

contributions, this study makes the decision-making process clearer and more understandable for financial institutions, fostering fairness in loan approval decisions.

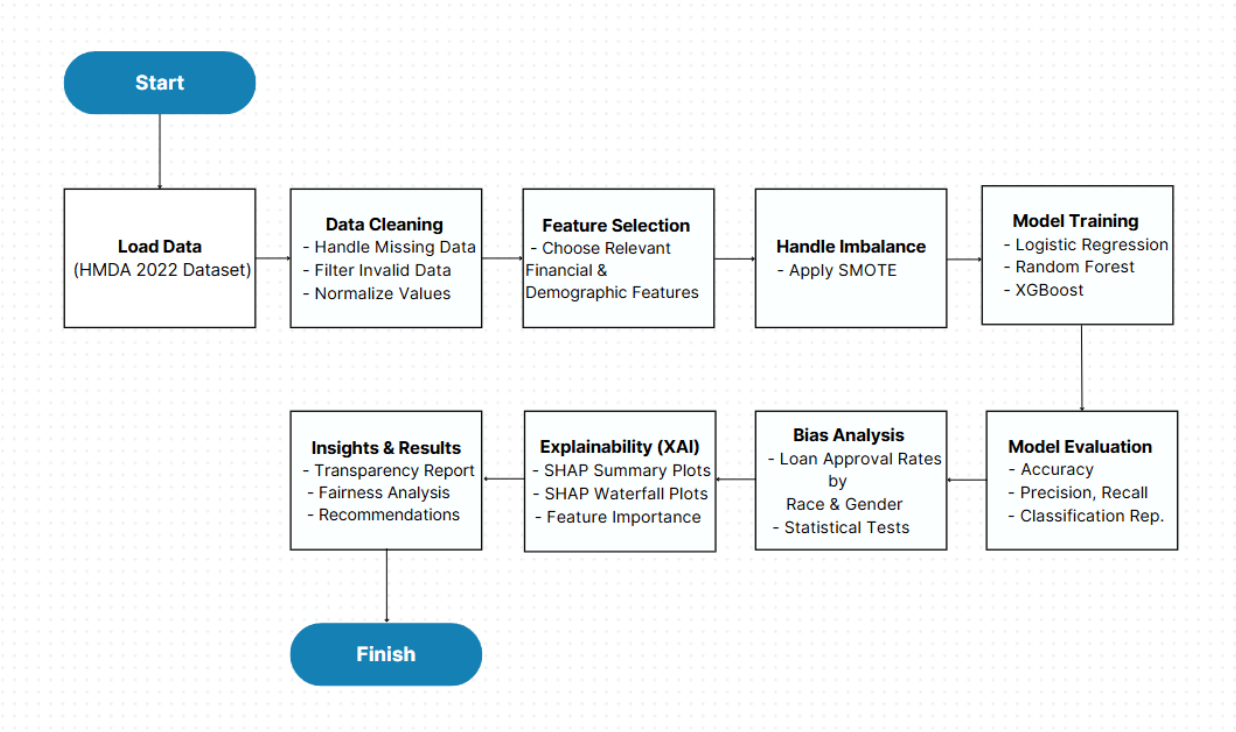
Nallakaruppan et al. (2024), in their paper [4] *"An Explainable AI Framework for Credit Evaluation and Analysis"*, propose a Random Forest-based XAI framework to address transparency issues in loan decision-making. By using explanation methods like LIME and SHAP, their framework provides detailed insights into the factors influencing loan approvals, ensuring that the decision-making process remains fair and transparent.

Through these studies, the application of XAI in loan approval systems is shown to not only improve transparency but also mitigate biases, thereby promoting fairness and trust in the financial system.

Methodology

The dataset was sourced from Kaggle and includes variables like race, gender, and income. Data cleaning involved removing duplicates, race filtering and mapping, sex filtering and mapping, income handling, loan-to-value Ratio, debt-to-income ratio, and credit score handling. For this analysis, the focus is on the primary applicant only to simplify the process while still achieving fairness and transparency objectives.

The data cleaning process used algorithms for duplicate removal, categorical variable mapping, and outlier handling. Duplicates were eliminated using drop_duplicates, while categorical variables like race and sex were filtered for validity and mapped to meaningful labels. Continuous variables, such as income, loan-to-value ratio, and debt-to-income ratio, were handled using clipping to remove outliers (based on the 99th percentile) and imputation techniques like median or mode replacement for missing values. Numeric conversion was applied to ensure consistency, and invalid entries were filtered out. These algorithms focused on ensuring data integrity, handling missing values, and standardizing the dataset for further analysis.



Data

The dataset underwent comprehensive cleaning and preprocessing to ensure the integrity and usability of the data. Initially, duplicate rows were removed to eliminate redundancy. Invalid values for race and gender were filtered out, with valid categories being mapped to descriptive labels such as “American Indian or Alaska Native” and “Male” or “Female,” respectively. Financial metrics like income, loan-to-value ratio, and debt-to-income ratio were scrutinized for anomalies. Negative values were replaced with NaN, outliers were capped at the 99th percentile, and missing values were imputed using appropriate statistical measures (e.g., median or mode). Each column was converted to a suitable numeric type when required, ensuring consistency for further analysis.

The selection of features was driven by their relevance to loan approval decisions. Applicant demographics like race and gender were included to analyze potential biases. Financial metrics, such as income, loan amount, loan-to-value ratio, and debt-to-income ratio, were prioritized for their direct impact on risk assessment. The target variable, “action_taken,” represented approval or denial decisions, forming the basis for predictive modeling. This curated set of features balanced fairness analysis and predictive accuracy.

Experiments

The analysis was conducted using two different systems: a MacBook with an M1 processor, 8GB RAM, running macOS, and a Windows 10 Pro Desktop with an AMD Ryzen 7 2700X Eight-Core Processor @ 3.70 GHz, 16GB RAM, and a 64-bit operating system. The experiments utilized Python libraries such as pandas and numpy for data processing, scikit-learn for modeling and evaluation, matplotlib for visualization, and SHAP for explainability analysis. The HMDA 2022 dataset was split into 80% for training and 20% for testing to build and evaluate the predictive models effectively.

Results

Visualizations revealed significant insights into loan approval trends. Bar charts highlighted disparities among racial groups, with approval rates ranging from 60.63% for Native Hawaiian or Other Pacific Islander to 79.99% for Asian applicants (See Figure 1). Chi-square tests confirmed statistically significant relationships between race, gender, and income brackets with loan approvals. Higher income brackets exhibited consistently higher approval rates, emphasizing financial stability as a key determinant (See Figure 2). The plots showed minimal differences between genders (See Figure 3).

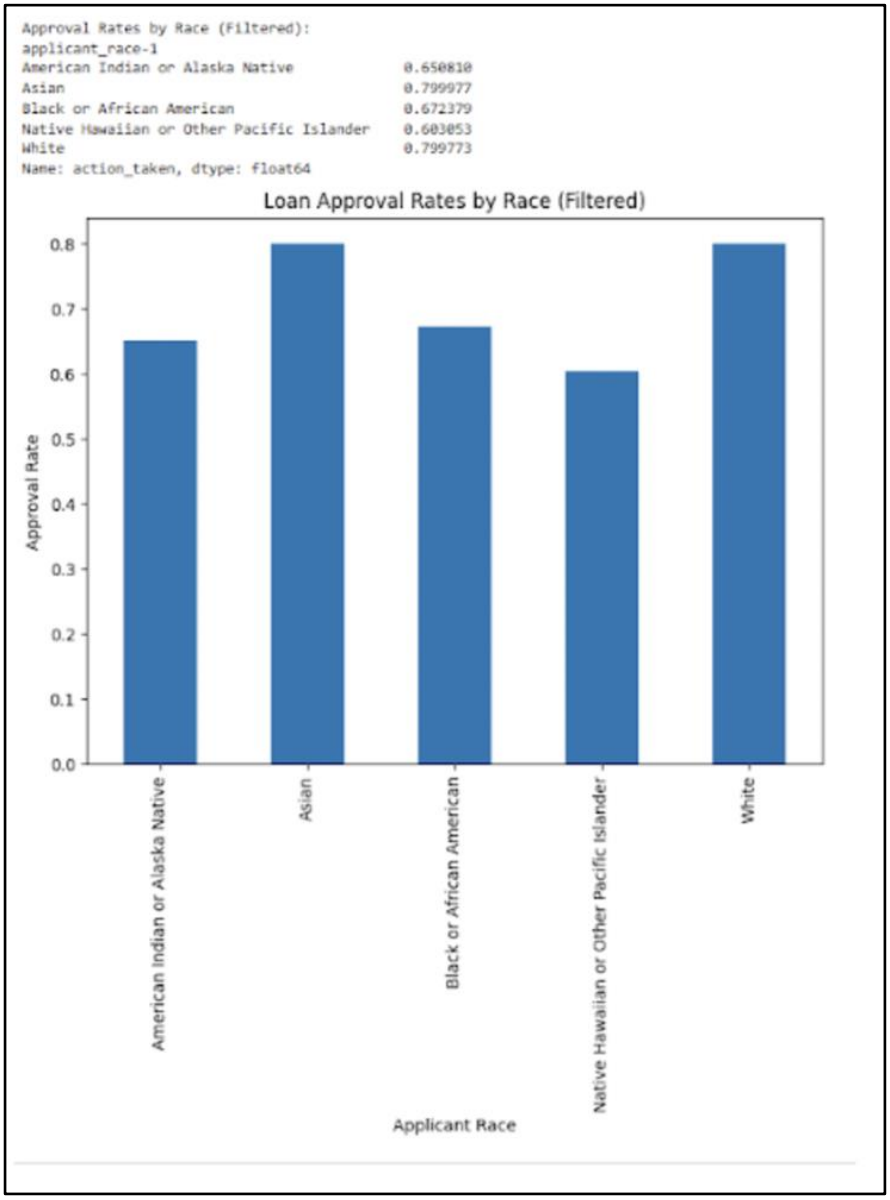


Figure 1: Loan Approval Rates by Race

The analysis reveals significant disparities in loan approval rates among racial groups, with White (79.97%) and Asian (79.99%) applicants having higher approval rates compared to Black or African American (67.63%), American Indian or Alaska Native (65.08%), and Native Hawaiian or Other Pacific Islander (60.63%) applicants. A Chi-Square test produced a Chi2 value of 1729.857 and a P-value of 0.0, indicating a statistically significant relationship between race and loan approval rates. This suggests strong evidence of potential bias or disparity in loan approvals based on race.

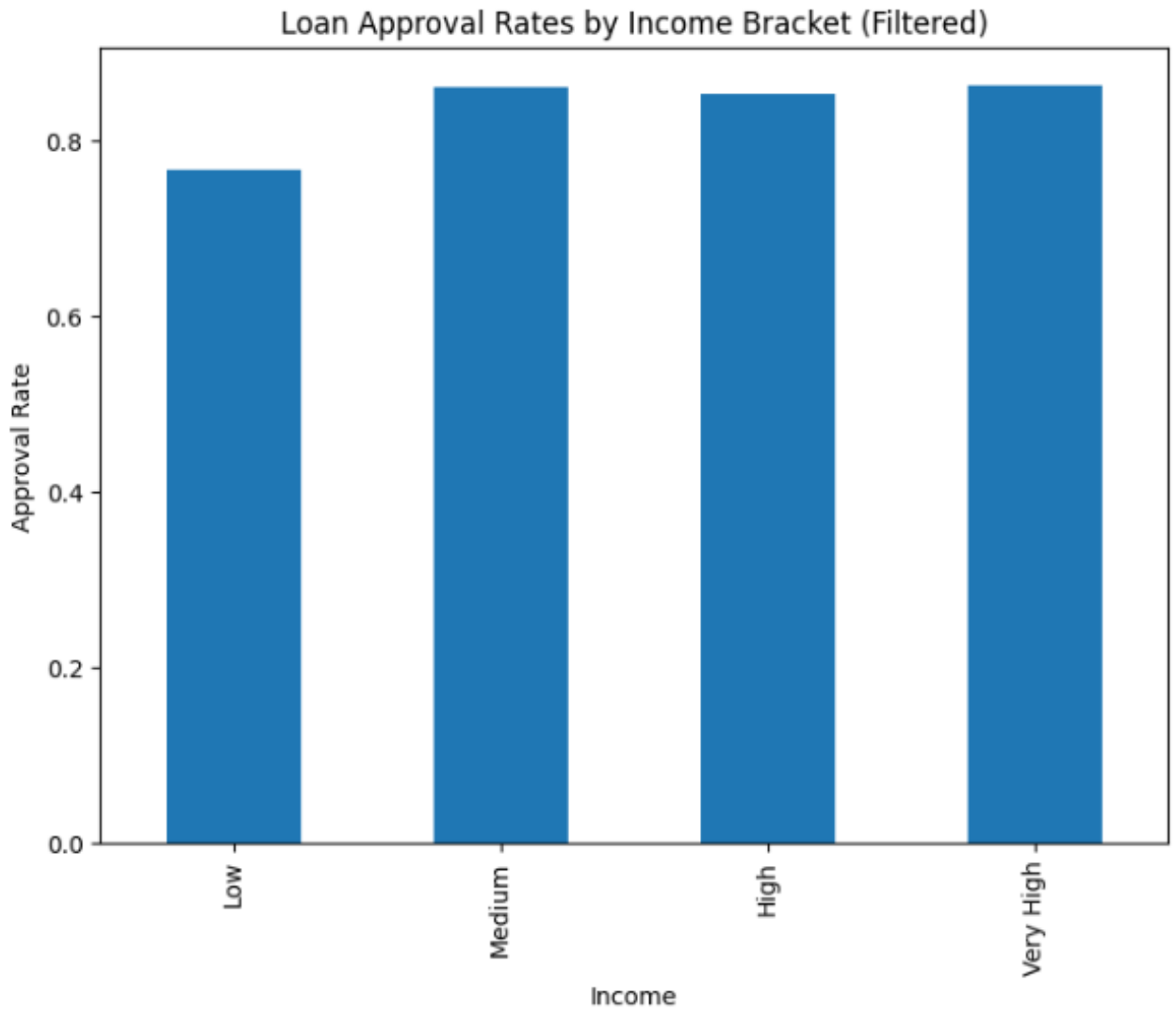


Figure 2: Loan Approval Rates by Income Bracket

The analysis shows a clear correlation between income levels and loan approval rates, with approval increasing as income rises: Low (76.64%), Medium (86.18%), High (83.01%), and Very High (86.21%). A Chi-Square test yielded a Chi2 value of 1389.802 and an extremely low P-value (4.81e-301), indicating a statistically significant relationship. This suggests that higher-income applicants are more likely to be approved, aligning with expected patterns of creditworthiness.

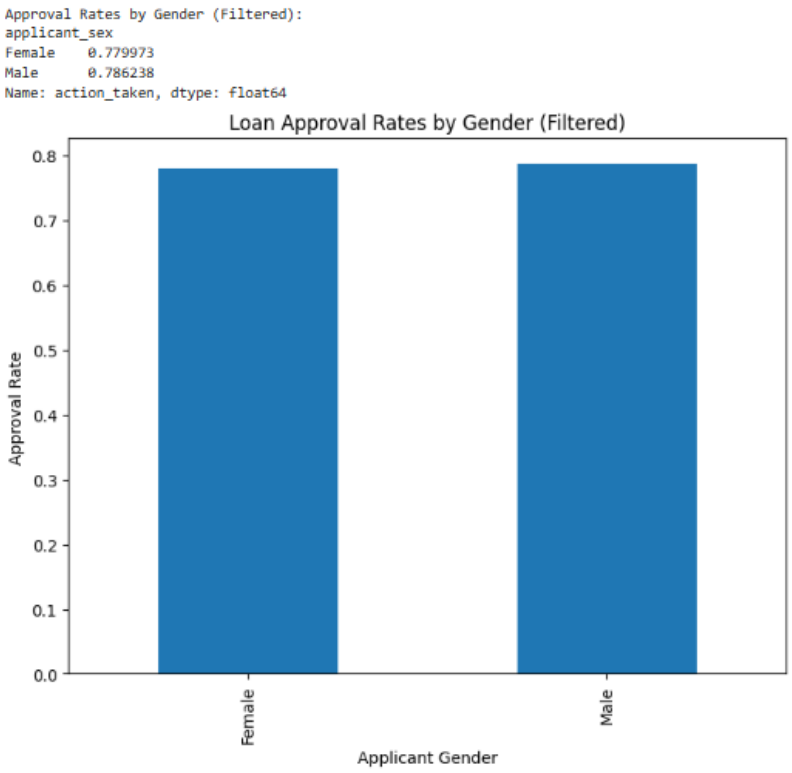
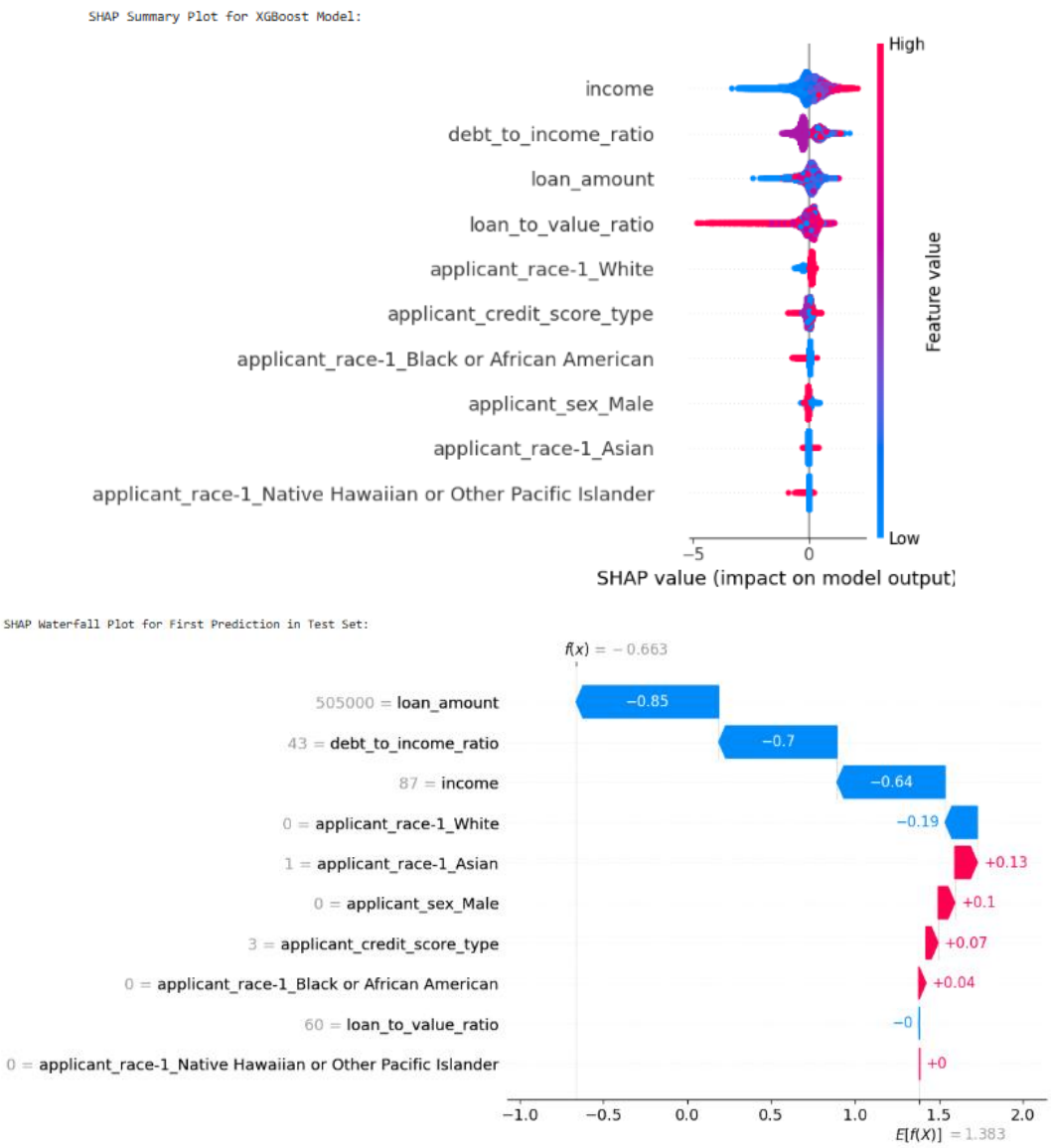


Figure 3: Loan Approval Rates by Gender

The analysis reveals minimal differences in loan approval rates between genders, with females at 78.99% and males at 76.83%. A Chi-Square test resulted in a Chi2 value of 8.561 and a P-value of 0.0034, indicating a statistically significant relationship. While the difference is statistically measurable, it is relatively small and unlikely to indicate large disparities based on gender.

Revealed is notable disparities in approval rates based on race, with significant differences suggesting potential bias. Gender also influences approval rates, though the disparity is relatively small, while approval rates increase significantly with income. In terms of model performance, Logistic Regression achieved 69% accuracy but struggled with identifying non-approvals, showing low precision (28%) and recall (30%) for Class 0, though it performed well with approvals (Class 1) with 81% precision and 79% recall. Random Forest, with 81% accuracy, showed improvements for non-approvals (59% precision) but still had low recall (30%) for Class 0, while excelling in predicting approvals (83% precision, 94% recall). The SHAP-integrated XGBoost model outperformed both with 82% accuracy, significantly improving precision for non-approvals (70%) but still facing challenges with recall (27%). For approvals, XGBoost demonstrated excellent precision (83%) and recall (97%), performing better overall, especially in handling the approval class, while also improving non-approval identification (See Figures 4-6)).



Figures 4-5: SHAP Waterfall Plot Analysis

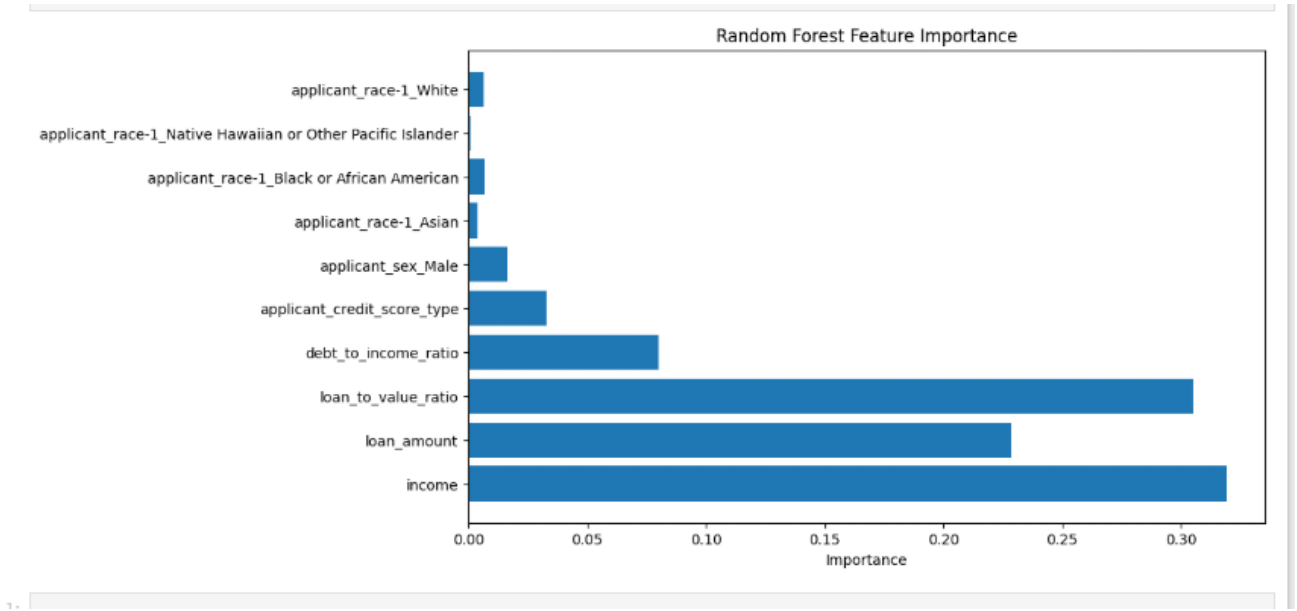


Figure 6: Random Forest Feature Importance Analysis

The findings highlight a significant class imbalance, with all models showing a bias toward predicting approvals (Class 1) despite the use of SMOTE. This is reflected in the consistently low recall for non-approvals (Class 0) across all models. In terms of model performance, Logistic Regression struggled the most with non-approvals and had lower overall accuracy compared to the tree-based models. Random Forest showed a marked improvement over Logistic Regression, particularly in predicting approvals, but still had difficulty predicting non-approvals. XGBoost outperformed both Logistic Regression and Random Forest, offering better precision for non-approvals and overall better performance, especially for predicting approvals. The conclusion is that while tree-based models like XGBoost perform better in handling the class imbalance, further adjustments are needed to improve recall for non-approvals.

Conclusion

From these findings, it can be concluded that biases exist in loan approval processes, particularly in relation to race and income. The use of Explainable AI (XAI) and fairness metrics has provided valuable insights into the factors influencing loan decisions. While machine learning models can improve transparency and fairness, they still struggle with class imbalance, especially in predicting loan rejections. Furthermore, the application of XAI techniques such as

SHAP has allowed for deeper interpretability, shedding light on how different features, like income and race, influence decision-making processes.

Despite attempts to address this with techniques like SMOTE, all models showed a bias toward predicting loan approvals. Improving recall for non-approvals (Class 0) is a key challenge that needs further exploration, particularly through techniques like oversampling or balancing class distributions. While the analysis identified disparities based on race and income, further research is needed to develop more robust methods for mitigating these biases. This includes exploring additional fairness algorithms and metrics to ensure that models are not unintentionally perpetuating existing disparities. Many of the challenges identified, such as ensuring fairness and mitigating bias in a live system, remain unresolved. Future work could focus on how to effectively deploy these techniques in real-time banking environments, where privacy, data protection, and regulatory compliance are critical. While SHAP and other XAI methods provided insights, there is a need for more user-friendly explanations for stakeholders without technical expertise. Making the results actionable for regulators and customers will be key to ensuring the system's effectiveness.

This project encountered several challenges that required careful problem-solving and iterative refinement. Data cleaning was particularly time-consuming due to numerous invalid or missing values for race, gender, and financial attributes like income and debt-to-income ratio. Feature selection posed difficulties, as determining relevant columns required multiple iterations and led to the exclusion of features with poor data quality. Initially, the project aimed to analyze both main applicant and co-applicant data, but the complexity of handling these relationships led to a narrowed focus on the main applicant. Feature preparation, including scaling numerical values and encoding categorical variables, was critical for performance but challenging to balance with interpretability.

This study highlights the importance of using XAI to meet regulatory standards for fairness in loan approvals, which could help institutions comply with anti-discrimination laws and build trust among customers. By improving the transparency of machine learning models, financial institutions can adopt more ethical practices, ensuring that loan decisions are both fair and explainable. This research has implications for the banking sector, where improving fairness and transparency in automated decision-making can lead to more equitable financial services. The integration of XAI techniques in loan approval systems can help reduce biases and enhance customer satisfaction. In summary, this project provides a step forward in addressing biases in loan approval algorithms, though several challenges remain for future researchers to explore, particularly around class imbalance and real-time implementation.

References

[1] Ayad, O., Hegazy, A., & Dahroug, A. (2023). A proposed model for loan approval prediction using XAI. *Nile Journal of Communication and Computer Science, 6*(1), 1-11. https://doi.org/10.21608/njccs.2024.215974.1013

[2] Azith, Y., Sushma, D., & Swathi, A. (2024). Revolutionizing loan approvals with explainable AI. https://www.apgcu.edu.in/pdf/mba-publications/2024-18-2.pdf

[3] Govindu, S., Manohar, N., Kumar, J. M. S. S., & Reddy, T. S. (2024). Loan prediction system with exploring explainable AI transfer learning with SHAP. In *2024 5th International Conference for Emerging Technology (INCET)* (pp. 1-7). IEEE. https://doi.org/10.1109/INCET61516.2024.10593029

[4] Nallakaruppan, M. K., Balusamy, B., Shri, M. L., Malathi, V., & Bhattacharyya, S. (2024). An explainable AI framework for credit evaluation and analysis. *Applied Soft Computing, 153*, 111307. https://doi.org/10.1016/j.asoc.2024.111307

[5] Genovesi, S., Mönig, J. M., Schmitz, A., Poretschkin, M., Akila, M., Kahdan, M., Kleiner, R., Krieger, L., & Zimmermann, A. (2024). Standardizing fairness-evaluation procedures: Interdisciplinary insights on machine learning algorithms in creditworthiness assessments for small personal loans. *AI and Ethics, 4*, 537-553. https://doi.org/10.1007/s43681-023-00291-8

[6] Das, S., Stanton, R., & Wallace, N. (2023). Algorithmic fairness. *Annual Review of Financial Economics, 15*(1), 565–593. https://www.annualreviews.org/content/journals/10.1146/annurev-financial-110921-125930