A Comparative Study of Machine Learning Approaches for Home Mortgage Approval

Ankit Mistry
Department of Electrical Enginnering and Computer Science
University of Tennessee
Knoxville, USA
amistry2@vols.utk.edu

Shivam Mistry
Department of Electrical Enginnering and Computer Science
University of Tennessee
Knoxville, USA
smistry1@vols.utk.edu

***Abstract*—This study delves into the realm of home mortgage approval leveraging a dataset extracted from the Home Mortgage Disclosure Act (HMDA) for the year 2022, specifically focused on Tennessee. The primary objective was to construct predictive models aimed at foreseeing home mortgage approval based on crucial features encompassing income levels, loan amounts, interest rates, debt-to-income ratios, loan-to-value ratios, among others.**

To address potential imbalances in the dataset, oversampling techniques were employed to ensure a more representative and equitable training environment. Subsequently, three distinct machine learning approaches—namely, MLP Classifier, Random Forests, and Ada-Boost—were applied and evaluated their efficacy in predicting mortgage approval outcomes. Throughout the experimentation process, it became evident that the Random Forest Classifier exhibited superior generalization capabilities, yielding the most promising and consistent results across various metrics. Its adeptness in capturing the nuanced relationships within the dataset translated into a robust predictive performance for mortgage approval prediction. Moreover, Ada-Boost and MLP Classifier, despite not outperforming Random Forest in overall accuracy, showcased remarkable proficiency in effectively classifying the minority class associated with loan denials. This underlines their specialized strength in categorizing instances where mortgage applications are likely to face rejection.

In conclusion, the study highlights the efficacy of machine learning methodologies in predicting home mortgage approval, with the Random Forest Classifier emerging as the preferred model for its well-rounded performance, while also acknowledging the niche strength of MLP Classifier and Ada-Boost in handling the challenging task of classifying loan denials.

***Index Terms*—Home Mortgage, Machine Learning, Approval, MLP Classifier, Ensemble Methods**

I. Introduction and Motivation

The realm of home mortgage approval stands as a critical juncture in the housing finance landscape, influencing individuals' access to home ownership and financial stability. In this pursuit, leveraging comprehensive and reliable datasets becomes pivotal for understanding the dynamics governing credit decisions. For our investigation, we turned our focus to the Home Mortgage Disclosure Act (HMDA) dataset for the year 2022, specifically encapsulating the financial landscape of Tennessee. This dataset, being government-provided and encompassing information from a myriad of financial institutions nationwide, offers a robust and diverse foundation for our explorations into mortgage approval dynamics.

The choice of the HMDA 2022 dataset was motivated by its credibility and inclusivity. Originating from a government source, this dataset assures a level of reliability crucial for our analyses. Moreover, its comprehensive coverage of financial institutions, both major and minor, across the nation, contributes to its richness, capturing diverse aspects of mortgage applications and approvals.

Our overarching goal in this study was two-fold. Firstly, we aimed to decipher the important features influencing credit decisions related to home mortgages. Acknowledging the prevalent use of machine learning techniques in credit-related decisions, we sought to explore the viability of these methodologies within the context of mortgage approvals. Secondly, recognizing the existence of inherent biases in datasets, we aimed to unravel and assess any biases present within the HMDA dataset, particularly in the domain of home mortgage approvals.

To tackle this multifaceted goal, we structured the problem as a classification task. We employed both neural network-based approaches and ensemble methods. Specifically, the Multi-layer Perceptron (MLP) Classifier was chosen as our neural network approach, while Random Forests and Ada-Boost were selected as representative ensemble methods. The rationale behind this selection lay in the varied strengths these methodologies offer, allowing for a comprehensive evaluation of their effectiveness in predicting mortgage approval outcomes.

To evaluate the success of our pursuits, we utilized the intrinsic property of ensemble methods—feature importance—to find the most influential factors steering mortgage approval decisions. Concurrently, our investigation into dataset biasness involved careful data analysis aimed at unveiling and comprehending any inherent biases embedded within the dataset.

In summary, our motivation for this study stemmed from the quest to uncover the influential factors driving home mortgage approvals, to evaluate the efficacy of machine learning techniques in this domain, and to scrutinize and comprehend potential biases within the dataset, all of which have significant implications for the financial well-being and accessibility of home ownership for individuals.

II. Dataset

A. Dataset Selection and Preprocessing

The dataset utilized for this study was sourced from the Home Mortgage Disclosure Act (HMDA) website, aiming to capture the landscape of home mortgage applications and approvals. Initially, during the dataset selection process, we encountered a dataset dating back to 2017. However, a careful examination revealed that this dataset lacked crucial attributes, notably many important features such as debt-to-income ratio and loan-to-value ratio. Recognizing the significance of these parameters in credit evaluation, we sought a more comprehensive and current dataset.

Upon further investigation, we discovered the availability of the 2022 dataset, which notably encompassed these essential features that play a pivotal role in determining credit eligibility. A comparative analysis between the 2017 and 2022 datasets urged us to incorporate these additional attributes for a more nuanced and accurate predictive model in our study.

Initially, our attempt to utilize the nationwide dataset was met with computational constraints due to its vast size, rendering it impractical for efficient processing given our computing resources. Consequently, we narrowed our focus to Tennessee, a region-specific dataset for its manageable size, enabling computational feasibility without compromising the integrity of the analysis.

However, the Tennessee dataset contained diverse action taken, describing the status of loan application, labels that were not directly pertinent to our study objectives. To streamline the dataset and align it with our research focus on mortgage approval decisions, we utilized HMDA's filtering feature. This feature facilitated the extraction of a refined dataset, exclusively comprising instances associated with approved and denied labels, eliminating extraneous categories and enhancing the dataset's relevance to our study.

The utilization of this tailored dataset ensured computational efficiency while preserving the dataset's integrity, enabled us to concentrate specifically on instances relevant to our investigation into home mortgage approval.

B. Feature Selection and Exploratory Analysis

In our pursuit of constructing predictive models for home mortgage approval, we curated a subset of features to be employed across our three machine learning models: Multilayer Perceptron (MLP) Classifier, Random Forests, and AdaBoost. The selected features encompassed vital aspects of mortgage applications, such as:

1) action_taken
2) purchaser_type
3) preapproval
4) loan_type
5) loan_purpose
6) business_or_commercial_purpose
7) loan_amount
8) loan_to_value_ratio
9) interest_rate
10) hoepa_status
11) property_value
12) occupancy_type
13) income
14) debt_to_income_ratio
15) applicant_credit_score_type

The rationale behind this selection primarily rested on the objectivity and relevance of these factors in assessing creditworthiness without incorporating subjective or potentially biased traits such as gender, race, or ethnicity.

Notably, the raw dataset contained features such as "derived_sex," "derived_race," and "derived_ethnicity," which were subjective and potentially influential in introducing biases into the machine learning models. Therefore, we deliberately excluded these traits to ensure the models' fairness and mitigate any inadvertent bias that might affect the predictive outcomes.

In addition to the features utilized for modeling, a supplementary set of attributes was employed specifically for exploratory data analysis and visualizations. The features are as follows:

1) loan_amount
2) interest_rate
3) income
4) applicant_sex
5) applicant_race-1
6) applicant_ethnicity-1
7) denial_reason-1
8) denial_reason-2
9) denial_reason-3
10) denial_reason-4

These attributes were selected to delve deeper into the dataset's composition, which enabled us to do a comprehensive examination of existing biases and denial reasons associated with mortgage applications.

Through exploratory data analysis, our objective was twofold: firstly, to scrutinize and identify any inherent biases already present within the dataset, and secondly, to comprehend the denial reasons associated with mortgage applications. This comprehensive analysis facilitated a nuanced understanding of the dataset's intricacies, enabling informed decisions regarding feature inclusion in predictive modeling and insights into potential biases present in the mortgage approval process.

C. Data Transformation and Preprocessing

For data transformation and preprocessing, initially, our code uniformly handles missing or ambiguous values denoted by 'Exempt' across critical columns ('loan_to_value_ratio', 'interest_rate', 'property_value', 'debt_to_income_ratio') by substituting them with the numeric placeholder 8888. This standardization allows for consistent treatment of missing data points.

Moreover, the 'debt_to_income_ratio' undergoes an intricate transformation, converting categorical ranges ('20%–<30%', '30%–<36%', '<20%', '50%–60%', '>60%') into

corresponding numeric representations (25, 33, 15, 55, 65 respectively). This conversion streamlines the data, enabling numerical analysis and interpretation of debt-to-income ratios.

To ensure data uniformity, we proceed to cast specific columns ('loan_to_value_ratio', 'interest_rate', 'property_value', 'debt_to_income_ratio') to floating-point numeric types which facilitated mathematical operations and standardizing data types across these features.

Additionally, for modeling clarity, the 'action_taken' column undergoes binary encoding. This process maps values 1, 2, and 3 to 1, 1, and 0 respectively, transforming the multi-class classification into a binary classification, simplifying the target variable for subsequent modeling tasks.

Through these careful transformations and preprocessing steps, we prepare the dataset by addressing missing values, standardizing data types, and modifying the target variable, laying a robust foundation for subsequent modeling and analytical endeavors.

D. Data Visualization

1) The Relationship between Income, Loan Amount and Action Taken: To explore the intricate relationship between income, loan amount, and their influence on the action taken regarding mortgage applications, a scatter plot was generated and analyzed. The plot aimed to elucidate any discernible trends or dependencies between these pivotal factors and their impact on loan approval outcomes. The scatter plot depicted income on the x-axis and loan amount on the y-axis, with distinct visual cues differentiating between approved and denied loans.

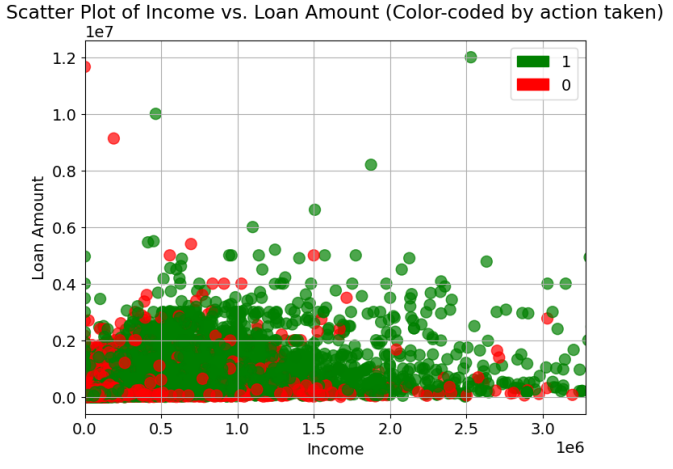

Fig. 1. Scatter Plot of Income vs. Loan Amount.

In this visualization, approved loan instances were denoted by green dots, while denied loans were represented by red dots. The distinct color coding facilitated a clear distinction between the different action taken categories for a comprehensive analysis of their distribution concerning income levels and loan amounts.

However, upon examining the scatter plot, we found no such discernible trends which suggested an absence of explicit correlations or thresholds indicative of specific income prerequisites for the approval of distinct loan amounts. The absence of discernible patterns implies that the approval or denial of loans may not be contingent upon specific income thresholds or predefined criteria within the examined dataset.

This visualization underscores the complexity and potential variability in the factors influencing loan approval decisions, emphasizing the need for further multifaceted analyses and considerations beyond income and loan amount when predicting mortgage approval outcomes.

2) The Denial Reasons in Mortgage Applications: In an effort to dissect and comprehend the grounds for denial within mortgage applications, an insightful exploration was conducted on the dataset's denial reason columns. These columns encapsulated eight unique denial reasons: Debt-to-income ratio, Employment history, Credit history, Collateral, Insufficient cash (down payment/closing costs), Unverifiable information, Credit application incomplete, and Mortgage insurance denial. Through the generation of a grouped bar chart, the frequency distribution of these denial reasons across the four columns was visually presented and analyzed.

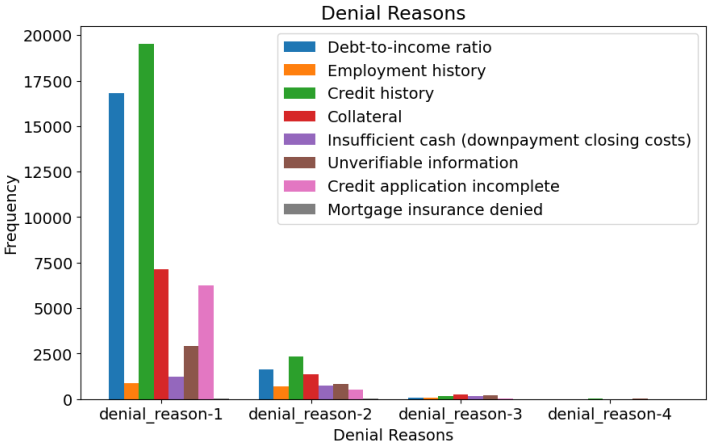

Fig. 2. Grouped Bar for various denial reasons.

The visualization unveiled compelling insights into the predominant factors influencing the denial of mortgage applications. Remarkably, the primary reason for denial emerged as credit history, followed closely by debt-to-income ratio. This observation underscores the critical importance of a sound credit history and favorable debt-to-income ratios in determining the fate of mortgage applications within the dataset.

Furthermore, a notable pattern surfaced during the analysis: a substantial portion of applications exhibited dual denial reasons. This pattern suggests that the presence of two distinct red flags, particularly related to credit history and debt-to-income ratios, significantly increased the likelihood of application rejection. Moreover, a minority of applications faced the hurdle of four denial reasons, highlighting the stringent criteria or multiple deficiencies leading to the outright rejection of these cases.

This visualization sheds light on the pivotal role played by credit history and debt-to-income ratios in the mortgage approval process, elucidating a tendency towards multiple denial reasons contributing to application rejection. These findings underscore the significance of addressing these specific factors to enhance the likelihood of mortgage application approval.

3) Exploration of Demographic Bias in Mortgage Applications: In a comprehensive endeavor to scrutinize potential biases within the dataset, two distinct analyses were conducted, shedding light on demographic disparities across various facets of mortgage applications.

Groups	Count	Income	Loan Amount	Interest Rate
Male	126102	178477.1	280438.38	4.84
Female	64466	101403.08	232323.4	4.92
White	154297	160196.45	263447.67	4.87
Black	16448	95395.91	220867.58	4.84
Hispanic	7567	109010.31	262646.36	5.14
Non Hispanic	169163	155873.35	261915.28	4.85

Fig. 3. Table shows Mean Income, Loan Amount and Interest Rates for different demographics.

The initial examination involved the creation of a comprehensive table encompassing mean income, loan amount, and interest rates for diverse demographic groups, including Male, Female, White, Black, Hispanic, and Non-Hispanic. This analysis revealed substantial disparities in the number of applications generated across demographics, notably observing a significant dominance in application numbers from Male, White, and Non-Hispanic groups compared to others. Interestingly, while interest rates remained relatively consistent across most groups, a notable exception was observed in the Hispanic demographic, where the mean interest rate exceeded 5%.

Another investigation compared the ratios of approval and denial decisions across the same demographic categories. Intriguingly, a disparity emerged wherein Male, White, and Non-Hispanic groups demonstrated lower denial rates, with less than 25% of applications being denied compared to those approved. However, a notable contrast was identified in the Black demographic, where the denial rate surpassed 50%.

Upon synthesis of these findings, a pronounced pattern of demographic bias in mortgage applications became evident. Specifically, the Black and Hispanic groups exhibited lower application numbers and approval rates, indicative of an inherent bias within the dataset. Remarkably, this bias did not manifest in metrics such as income, loan amount, or even interest rates, which remained relatively consistent across demographic segments.

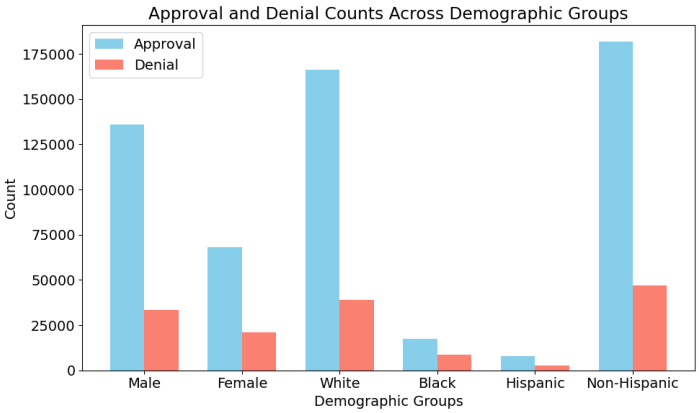

Fig. 4. Bar Chart of Approval and Denial Frequencies for different demographics.

In conclusion, while disparities exist in application and approval rates for different demographics, the absence of biases in metrics such as income, loan amount, and interest rates underscores the nuanced nature of biases within the dataset, highlighting the need for further investigation and mitigation strategies to ensure equitable treatment in mortgage application evaluations.

III. Machine Learning Approaches and Methodology

To address the inherent challenges posed by imbalanced datasets, particularly the scarcity of instances in the minority class, we employed strategic methodologies such as oversampling to enhance the robustness of our predictive models. In response to the class imbalance, we used the RandomOverSampler module from the imbalanced-learn library, effectively augmenting the representation of the minority class and fostering a more equitable distribution within the dataset.

Subsequently, our focus shifted to framing the problem as a binary classification task, where the objective was to predict the approval or denial of applications based on relevant feature values. To achieve this, we implemented three distinct machine learning approaches: the Multi-Layer Perceptron (MLP) Classifier, Random Forest, and Ada-Boost. Each of these algorithms brings unique strengths to the predictive modeling process, collectively contributing to a comprehensive evaluation of the dataset and its underlying patterns. In this section, we elucidate the intricacies of our chosen methodologies, delineating the rationale behind their selection and detailing the key parameters configured for optimal performance.

A. MLP Classifier

As we investigate the Multi-layer Perceptron (MLP) Classifier, our first move was to get a handle on the dataset's variability. This model tends to be quite particular about data nuances, so we started by checking how much each feature liked to vary. Turns out, there was a whole lot of variation going on among the features, so we employed StandardScaler

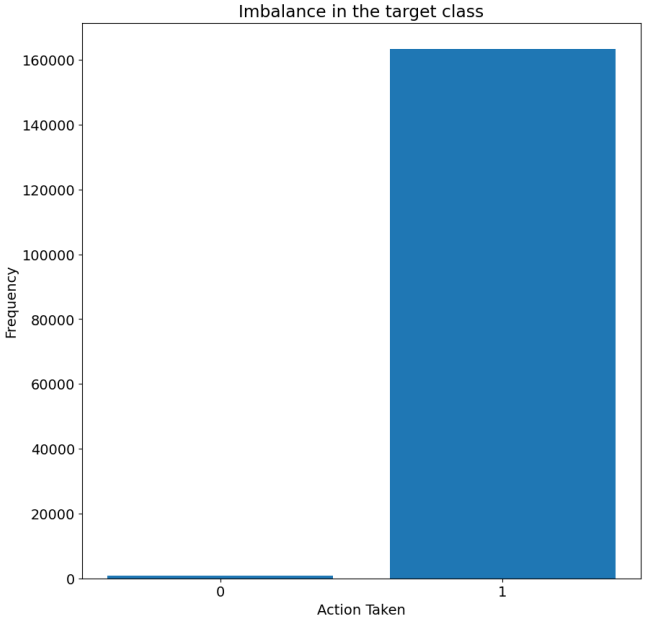

Fig. 5. This graph shows the imbalanced data before oversampling.

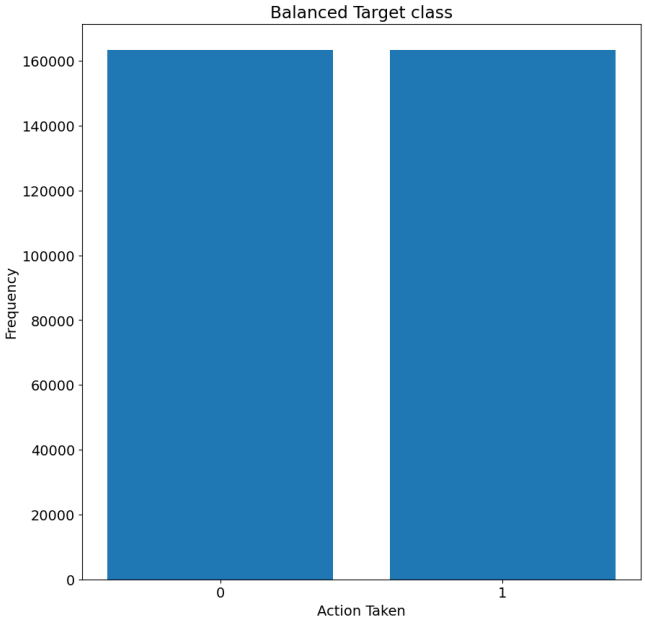

Fig. 6. This graph shows the balanced data after oversampling.

from sklearn's preprocessing toolkit. This tool helped us to standardize and scale the data.

When setting up the MLP Classifier, the spotlight was on tweaking those crucial hyper-parameters. We took a deep dive into optimizing the learning rate initialization and fine-tuning the hidden layer sizes. First, we used single hidden layer values such as (50, 100, 200, 500) to test different learning rates. We found that the best learning rate was 1. Furthermore, we tested different hidden layer sizes for this learning rate and we found that the best structure to use is a single hidden layer consisting

Columns	Mean	Standard Deviation
purchaser_type	4.071966171031036	15.154154799799327
preapproval	1.9866956127673125	0.1145747813135602
loan_type	1.1784165788893517	0.5484287950091999
loan_purpose	10.439794968522646	13.578134788598767
business_or_commercial_purpose	567.3886901991524	554.4081821261333
loan_amount	215498.21791637066	242134.27101966194
loan_to_value_ratio	4679.700638649633	4403.180566061311
interest_rate	4647.30752840923	4437.048215754524
hoepa_status	2.560958283320677	0.4965909137593629
property_value	218384.96019400828	375485.136361288
occupancy_type	1.1799475785709037	0.554240336259699
income	138580.09577934007	8821356.431451393
debt_to_income_ratio	4662.058790387771	4421.61977234445
applicant_credit_score_type	582.2228308796512	553.2647827851681

Fig. 7. This table shows the dataset before preprocessing.

Columns	Mean	Standard Deviation
purchaser_type	0.0	1.0
preapproval	-0.0	1.0
loan_type	-0.0	1.0
loan_purpose	0.0	1.0
business_or_commercial_purpose	0.0	1.0
loan_amount	0.0	1.0
loan_to_value_ratio	-0.0	1.0
interest_rate	-0.0	1.0
hoepa_status	0.0	1.0
property_value	-0.0	1.0
occupancy_type	-0.0	1.0
income	0.0	1.0
debt_to_income_ratio	0.0	1.0
applicant_credit_score_type	-0.0	1.0

Fig. 8. This table shows the dataset after preprocessing.

of 100 neurons. The goal was to strike that perfect balance to empower the model to navigate the complexity inherent in the data and make informed predictions.

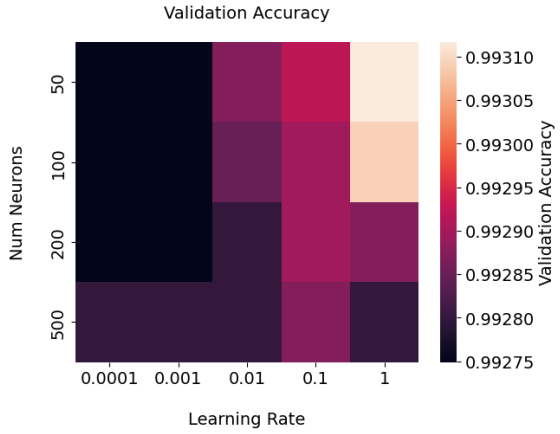

Fig. 9. This is the heat map of the different validation accuracies generated by the MLP classifier.

B. Random Forest

In our pursuit of an effective classification solution, the second machine learning approach employed was the Random

Forest Classifier. To ensure a robust evaluation of the model's performance, we initiated the process by partitioning the dataset into distinct sets for training, testing, and validation. To optimize the classifier's hyper-parameters, we conducted a systematic exploration on the validation set. Specifically, we focused on the crucial parameter of max_depth, ranging from 2 to 40 with intervals of 2, to comprehensively assess the classifier's sensitivity to this parameter. Our tuning efforts revealed that the Random Forest Classifier achieved its peak accuracy when configured with a max_depth of 34. This hyper-parameter configuration was identified as the optimal setting, showcasing the model's ability to discern intricate patterns within the dataset and attaining heightened predictive accuracy. In the subsequent sections, we delve into the nuanced outcomes of this experimentation and elucidate the significance of the selected hyper-parameter values in refining the Random Forest Classifier's performance for our classification task.

C. AdaBoost

For our third machine learning approach, we turned to AdaBoost, a powerful ensemble learning algorithm renowned for its ability to adapt and improve model performance without an extensive reliance on hyperparameter tuning. In the case of AdaBoost, the primary parameter of interest is the number of base learners, denoted as n_estimators. Given the algorithm's inherent simplicity regarding hyperparameter configuration, we initially instantiated an AdaBoost model with 100 base learners, establishing a baseline for our experimentation.

Nevertheless, in our commitment to comprehensive model evaluation, we conducted a secondary analysis to test the impact of varying n_estimators values. To facilitate this exploration, a validation set was crafted, and the AdaBoost model was retrained with different values for n_estimators. Intriguingly, our findings echoed the initial configuration, as the highest accuracy was consistently attained when the number of base learners was set to 100. This result underscores the robustness of AdaBoost in achieving optimal performance with minimal hyperparameter fine-tuning, emphasizing its efficacy as a valuable component in our ensemble of classification models. In the subsequent sections, we expound upon the implications of these findings and elucidate the distinctive characteristics that contribute to AdaBoost's success in our classification task.

IV. Results

In the evaluation of our machine learning models, we commenced with the Multi-Layer Perceptron (MLP) classifier, implementing it with the best hyperparameters derived from our earlier tuning efforts. The visualization of the loss curve demonstrated a promising trend, showcasing a consistent decrease and indicating the model's ability to learn from the data effectively.

Furthermore, we examined the training and testing accuracy, along with the F1 score, revealing the classifier's proficiency in generalization. The corresponding confusion matrices for both the training and testing sets underscored the model's adeptness in handling false negatives and false positives, highlighting its robustness in credit decision-making.

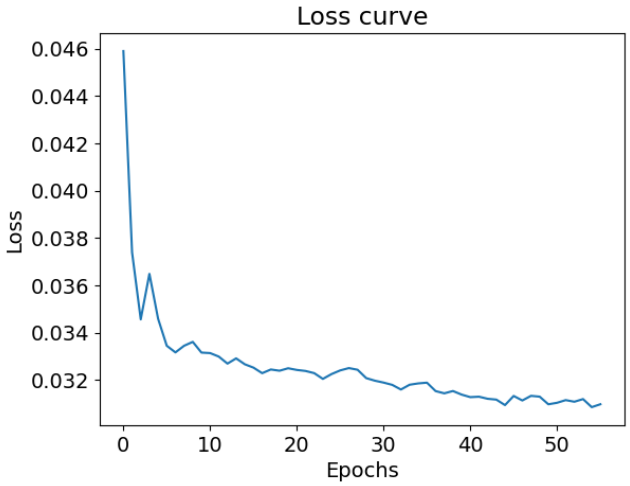

Fig. 10. This graph shows the loss curve of the MLP Classifier.

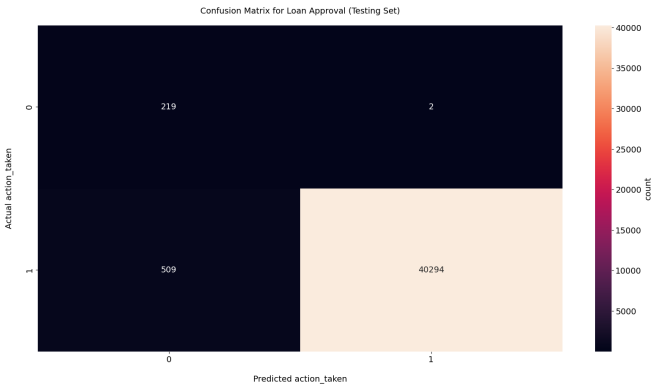

Fig. 11. This heat map shows the confusion matrix for the Testing Set for the MLP Classifier.

Transitioning to the Random Forest classifier, our results echoed the efficacy observed with the MLP model. Despite exhibiting slightly more false positives and false negatives, this classifier demonstrated exceptional accuracy across the entire dataset. Delving into the feature importance, we identified key determinants influencing credit decisions, with interest rate and debt-to-income ratio topping the list. Notably, credit score and loan-to-value ratio also emerged as significant contributors to the model's decision-making process.

The AdaBoost classifier, while displaying comparable results, exhibited marginally lower training and testing accuracy compared to its counterparts. Intriguingly, its prowess in classifying the minority class was noteworthy, as evidenced by the confusion matrix analysis. Feature importance unveiled income as the most influential factor, followed by loan amount, loan purpose, and loan type. This outcome was contradicting since when we explored the dataset, it was clear that the relationship between income, loan amount, and approval decision was almost negligible.

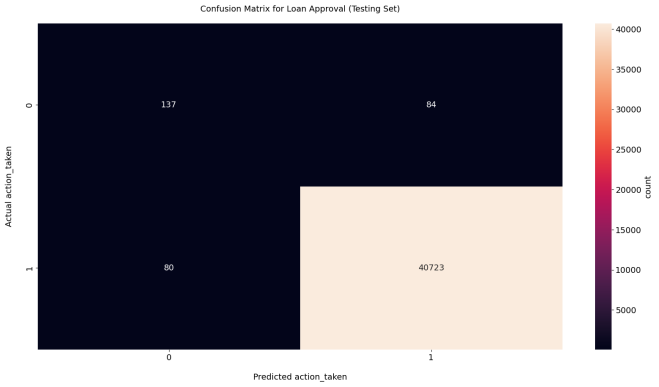

Fig. 12. This heat map shows the confusion matrix for the Testing Set for the Random Forest Classifier.

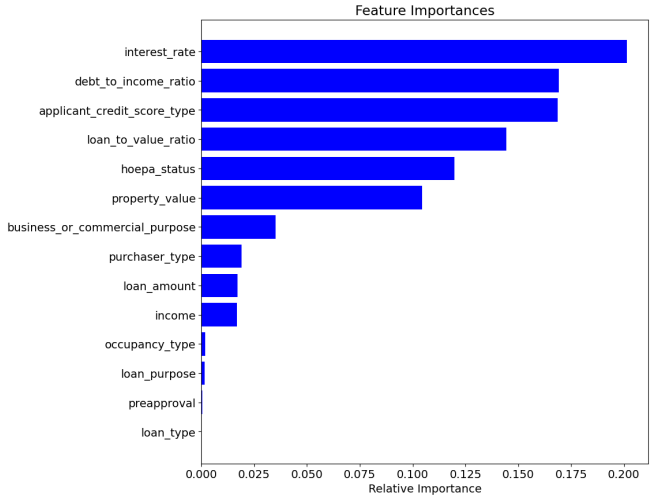

Fig. 13. This bar graph shows the feature importance for the Random Forest Classifier.

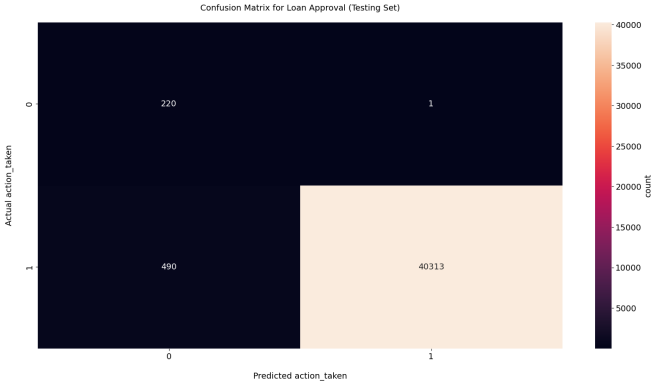

Fig. 14. This heat map shows the confusion matrix for the Testing Set for the AdaBoost Classifier.

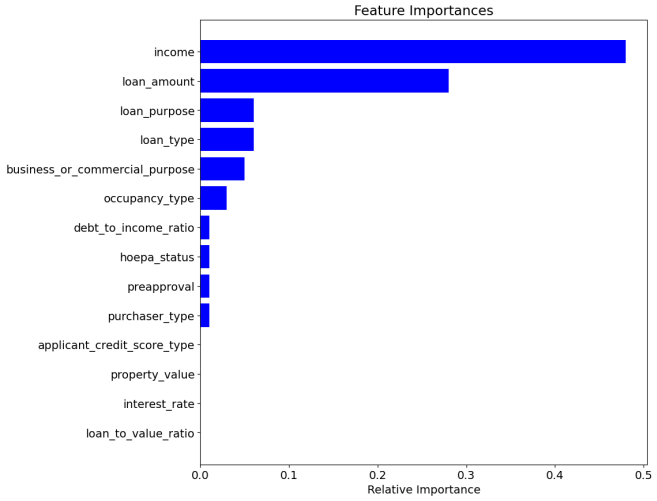

Fig. 15. This bar graph shows the feature importance for the AdaBoost Classifier.

In our comparative analysis, the Random Forest emerged as the most versatile, showcasing superior generalization across classes. Among the other classifiers, the MLP Classifier model was deemed more appropriate, aligning well with our earlier data analyses and demonstrating robust performance. AdaBoost, while proficient in handling the minority class, exhibited a potential bias towards it. In conclusion, our findings affirm the Random Forest as the optimal choice for credit decision-making, with the MLP classifier standing out as a reliable alternative.

V. Discussion, Conclusion and Future Work

In the realm of machine learning for home mortgage approval, our investigation revealed the superior performance of the MLP Classifier and Random Forest Classifier among the three models employed. Notably, the Random Forest Classifier showcased remarkable accuracy, achieving 99% accuracy on testing data, closely followed by the MLP Classifier at 98% accuracy. A detailed analysis highlighted the pivotal role of interest rates and debt-to-income ratios, particularly underscored by the Random Forest's efficacy in determining mortgage application outcomes.

In scrutinizing potential biases within the dataset, our analysis revealed intriguing insights. While interest rates displayed no significant biases, a notable imbalance surfaced across demographic groups concerning approved versus denied loans. This disparity raises pertinent questions about equitable treatment across diverse demographic segments within the mortgage approval process.

Looking ahead, our future endeavors aim to refine and broaden the scope of this study. Firstly, procuring a preprocessed dataset directly from authoritative sources like HMDA or specialized institutions gathering Home Mortgage Application data would enhance data integrity and accuracy. Moreover, future work will prioritize mitigating biases ingrained within the dataset. One approach involves striving for

balanced samples of approved and denied applications across all demographic groups, fostering a more equitable dataset for robust analysis and modeling.

VI. Contribution of Team Members

- Ankit Mistry
  - Responsible for doing data analysis and visualizations.
  - Formatting and writing the final report.
  - Formatting and making the presentation.
- Shivam Mistry
  - Responsible for finding the dataset and cleaning it.
  - Applying machine learning to the cleaned dataset.
  - Writing the final report