National College of Ireland

Comprehensive Analysis and Prediction of Loan Amount Distributions in the United States Mortgage Market

MSc Research Project
MSc in Data Analytics

Bolormaa Mendbayar
Student ID: X23176725

School of Computing
National College of Ireland

Supervisor: Naushad Alam

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Bolormaa Mendbayar
Student ID:	X23176725
Programme:	MSc in Data Analytics
Year:	2024
Module:	MSc Research Project
Supervisor:	Naushad Alam
Submission Due Date:	12/08/2024
Project Title:	Comprehensive Analysis and Prediction of Loan Amount Distributions in the United States Mortgage Market
Word Count:	XXX
Page Count:	25

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	11th August 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Comprehensive Analysis and Prediction of Loan Amount Distributions in the United States Mortgage Market

Bolormaa Mendbayar
X23176725

Abstract

The 2022 Federal Financial Institutions Examination Council (FFIEC) Home Mortgage Disclosure Act (HMDA) dataset was analyzed to enhance financial risk assessment and decision-making in mortgage lending. Employed advanced data processing techniques such as shuffling and data chunking alongside exploratory data analysis (EDA) to uncover insights. Traditional methods, based on basic statistical techniques, offered limited insights into lending practices. Conducted a state-level analysis highlighted regional lending patterns and trends, identifying disparities and unique insights for localized strategies while demographic and financial factors visualized using Sankey diagrams revealed slight but not significant demographic bias. Random Forest was used for feature importance analysis, leading to the selection of the 20 most important variables, for developing ten predictive models including an ensemble approach. The ensemble method, which combined the three most accurate models—Random Forest, XGBoost, and CatBoost—proved to be the most effective, delivering the highest R-squared value and the lowest error metrics.

1 Introduction

The banking industry is one of the industries that has seen a lot of technological improvements. There are more loan approval rates going up on a daily basis. When reviewing a loan application banks usually have a strict set of guidelines and polices which might include things such as a person's repaying capabilities as well as some personal criteria including credit score, income level, employment history, age, debt-to-income ratio etc. It will not be easy to verify every individual before suggesting them for loans as it may prove dangerous. Based on this personal information will use machine learning to determine any demographic bias at the national or state level, and to predict loan amounts.

Artificial Intelligence (AI) is increasingly in demand and is being used in several domains to solve some very niche problems. In the financial, it certainly could as well be leveraged to analysis loan and mortgage applications, and help the lending institution get a comprehensive report on the applications, especially in scenarios where the applicant does not a good enough credit score, or income or not any ability to get a loan. Given the huge surge in loan applications, banks are having a large number of applications daily, it is challenging to analyze each one individually and accurately. Despite technological advances, and availability of huge amount, there is a lack of detailed analysis regarding the

distribution of loan amounts across different ranges. Additionally, the factors influencing these distributions have not been thoroughly explored.

The Home Mortgage Disclosure Act (HMDA) is a U.S. law designed to promote fairness in lending and make it easier for more people to obtain mortgages. Since the HMDA data started getting published in 1975, a lot of studies/analysis have been done on it previously such as (Guy et al.; 1982) found higher black people populations received fewer mortgages . Many researchers identified applicants are directly treated differently based on race, and minorities have consistently been disadvantaged (Ross and Yinger; 2002).

Recent analyses confirm that these disparities persist. For instance, Black and Hispanic applicants face higher denial rates compared to White applicants, even when accounting for factors like debt-to-income ratio and credit history (Bhuyan et al.; 2023).Additionally, geographic disparities also exist, with certain metropolitan areas exhibiting higher denial rates for minority applicants. These findings highlight the necessity for localized strategies to address these issues.

This study aims to conduct an in-depth analysis of loan amounts within specified ranges and identify relevant patterns using the 2022 Federal Financial Institutions Examination Council (FFIEC) Home Mortgage Disclosure Act (HMDA) dataset. By employing advanced techniques such as shuffling, data chunking, exploratory data analysis (EDA), and data cleaning, seek to uncover the factors influencing loan approval rates and and develop models to predict loan amounts. The primary goal is to enhance the process of financial risk assessment and aid in decision-making for mortgage lending, building on previous research to better understand the statistical characteristics and distributions of loan amounts across different ranges.

In this work, will address the following research question,

How do financial and demographic factors, including state-level differences, influence loan approval rates in the US, and which machine learning model best predicts loan amounts?

This paper proposes using machine learning to identify biases in national and state-level analyses based on applicants' demographic and financial factors. It will then perform feature analysis to determine the best model for predicting loan amounts.By addressing these questions, this study aims to contribute to the development of fairer, more inclusive financial systems.

2 Related Work

Loan distribution is one of the significant functions in the financial industry which is very competitive and constantly evolving as it plays a crucial role in the stabilization and development of the economy. Machine learning and big data analytics have brought a big change in the credit risk assessment and loan amount distribution.

2.1 Current Research on Discrimination in Mortgage Lending

Current research reveals that the discrimination of minorities in mortgage lending in the United States remains present. In HMDA data, Bhutta and Ringo (2024) evaluated the racial discrimination in mortgage lending. They found that minority applicants, particularly Black and Hispanic individuals, face higher denial rates primarily due to lower credit scores and higher leverage rather than overt discrimination by lenders. However,

the study admits that there could be gaps such as discouraging applications from the minority, which might hide the level of discrimination fully (Reserve; 2024; Journal; 2024).

A research that used 2006 HMDA data (Avery et al.; 2007) used multiple regression analysis and logistic regression analysis in analyzing the mortgage denial rates among the different racial groups. The research also excluded other characteristics of applicants and loans such as credit scores, debt-to-income ratios, and loan-to-value ratios in order to establish the effect of race on the lending decisions. However, the controls showed that the denial rates of the minority applicants remained higher than the white applicants. The study highlighted newer statistical methods have less of direct racism, the possibilities of bias and discrimination are still present and probable if the fair lending laws are not strictly implemented.

Future studies have consistently established demographic differences in mortgage lending. The Fuster et al. (2020) utilized machine learning techniques, such as Random Forests and other advanced algorithms, to control for observable applicant characteristics like credit scores and loan-to-value ratios. Despite these efforts, the study still identified persistent unexplained gaps in denial rates, which suggest that discrimination may be at play. Although the use of sophisticated algorithms enhances the reliability of the analysis, the potential for algorithmic bias cannot be entirely ruled out.

The of Minneapolis (2023) examined unexplained Black-White disparities in denial rates among the largest mortgage firms using HMDA data, credit scores, Loan to Value (LTV) ratios, and Debt to Income (DTI) ratios. They employed regression models such as Decision Trees to identify the aspects that most influenced loan denials. They found differences in denial rates across institutions and found that Black and Hispanic borrowers had higher negative credit decisions even after controlling for risk factors. Decision Trees were useful in pinpointing the critical attributes in decision-making procedures.

The credit score has been confirmed as the main factor to mortgage among the minorities and low income earners. According to the Institute (2023), millions of mortgage opportunities have been missed because of high credit standards particularly for the black people. In a study by Bocian et al. (2008), the researchers used logistic regression to establish that the minority consumers are more likely to be given high-interest subprime loans that only serve to perpetuate existing disparities. However, their study was criticized for not effectively dealing with the selection bias issue in subprime loan applications.

2.2 Machine Learning Techniques

Delis and Papadopoulos (2019) used the Random Forest and Gradient Boosting models and concluded that discrimination in lending exists as Black and Hispanic applicants get lower credit than White applicants even when credit risk is controlled for. Their study's strength is its ability to simulate interactions, but they did not explore the possible reasons for such patterns.

Hanson and Hawley (2016) provided correspondence tests that gave an evidence of racism in mortgage loan lending. They employed logistic regression to establish the likelihood ratio for loan approval based on race and it gave direct estimates of racial impacts but may have disguised the nature of discrimination.

2.3 State Analysis in Mortgage Lending

An analysis of HMDA data from 2018-2021 shows that Black and Hispanic borrowers face more challenges than White borrowers, this indicates systemic problems with lending Reserve (2022). The analysis of seven counties in Kentucky, Ohio, and Pennsylvania underscores the need for state-level research to address regional disparities in mortgage credit.

Similarly, the California Reinvestment Coalition Coalition (2016) examined seven metropolitan areas 2010 HMDA data and discovered that Black and Latino borrowers were steered to more costly FHA/VA loans despite having similar income and credit scores. This practice exacerbates segregation and economic inequality. However, extensive studies at the state level are limited, and the existing literature does not include all cities.

Based on these results, there is a significant need for more extensive state-level research to enhance the current knowledge of the regional lending discrimination.

3 Methodology

The methodology consists of six steps throughout the project, as outlined below and illustrated in Figure 1.

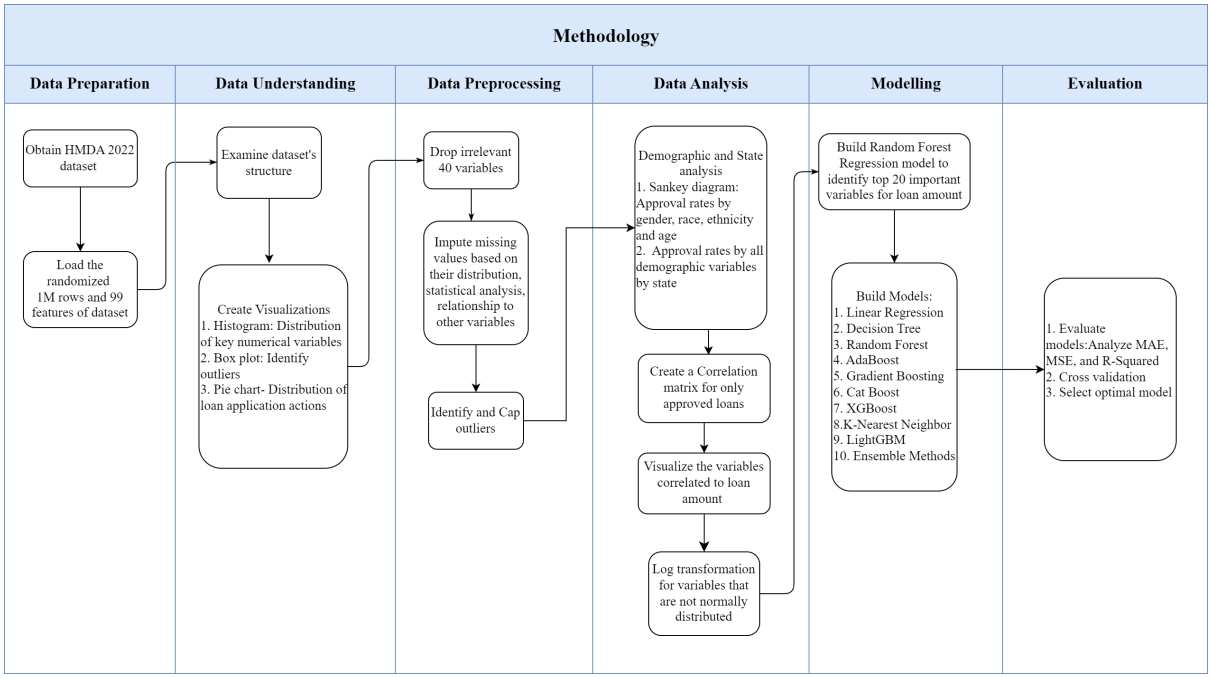


Figure 1: Research Methodology

3.1 Data Preparation

The first step in data preparation involved obtaining a randomized sample from the HMDA 2022 dataset [1]. The complete dataset, originally in CSV format, was read into

[1]Dataset: `https://ffiec.cfpb.gov/data-publication/snapshot-national-loan-level-dataset/2022`

a pandas DataFrame. To ensure the sample was representative and free from ordering biases, the DataFrame was shuffled using the pandas sample method with frac=1, which randomly permutes all rows. Subsequently, the reset_index(drop=True) function was applied to reset the index, creating a new, randomized order of rows without retaining the original indices. From this randomized DataFrame, a sample of 1 million rows was extracted. This sample size was selected to balance computational efficiency with the need for a sufficiently robust dataset for analysis. Note: To ensure consistency in the shuffling process, a fixed random seed was used.

The dataset contains 99 features detailed information on mortgage applications, including variables related to the loan, applicant demographics, and property characteristics.

Key variables include:

- Loan Amount and Type: Details about the loan amount, type, and purpose
- Applicant and co-applicant Information: Includes demographic details such as ethnicity, race, sex, age, and credit scores.
- Property Information: Details about the property such as value, location (state, county, census tract), and type of dwelling.

This analysis is done in Python using a Jupyter notebook leveraging libraries such as matplotlib, pandas, seaborn, numpy etc.

3.2 Data Understanding

The second step in data understanding involved conducting a comprehensive examination of the dataset's structure, attributes, and initial patterns. Created visualizations such as histograms, box plots, scatter plots as well as Sankey diagrams to explore the distribution and relationships of loan amounts and other variables. Initially segregated the loan amounts into 3 buckets such as $5k$ to $100k$, $100k$ to $1M$, $1M$ to above.

Figures A1-A5 are illustrated in Appendix A section. Figure A.1 shows that $5k$ to $100k$ is common loan amounts, $100k$ to $1M$ has a decreasing trend, more loans at lower amounts, above $1M$ has a significant peak at lower end, fewer larger loans.

Figure A.2 shows the distribution of several key variables in the dataset. Their descriptions are explained below:

- Combined-Loan-To-Value (CLV) Ratio: Represents the total loan amount relative to the property value.
- Interest Rate: The annual percentage rate charged on the mortgage.
- Total Loan Costs: The sum of all costs associated with the loan over its lifetime.
- Total Points and Fees: The total amount of points and fees paid upfront at loan origination.
- Origination Charges: Fees paid by the borrower to cover the loan's processing costs.
- Lender Credits: Amounts provided by the lender to reduce the borrower's closing costs.

- Property Value: The appraised value of the property being financed.
- Income: The gross annual income of the borrower.
- Debt To Income (DTI) Ratio: The ratio of the borrower's total monthly debt payments to their gross monthly income.

The interest rate is normally distributed, showing almost a symmetric spread around the mean. Other variables, such as CLV ratio, total loan costs, total points and fees, and property value, exhibit skewness towards lower values. The DTI ratio shows notable peaks in the 20%-30% and 30%-36% ranges, indicating that these ranges are more common among borrowers compared to other DTI ratios.

Figure 2 illustrates that the majority of applications are originated, notable percentage of denied and withdrawn applications indicates areas where applicants may face challenges or reconsider their borrowing needs. For the purpose of this analysis, actions 2 through 8 will be grouped together and categorized as denied applications.

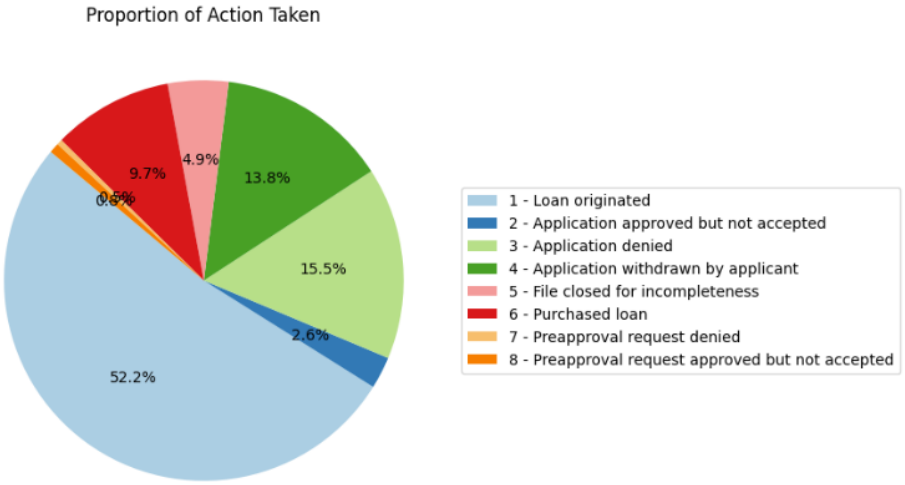


Figure 2: Distribution of Loan Application Actions

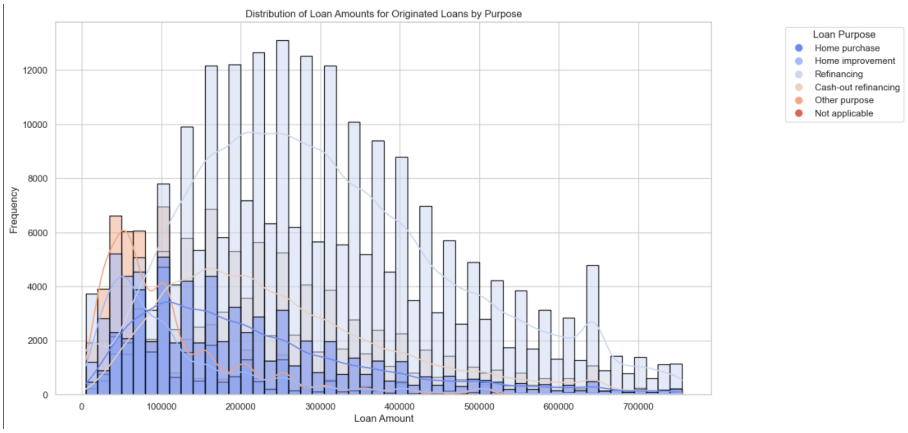


Figure 3: Distribution of loan amounts for originated loans by purpose

Figure 3 presents the distribution of loan amounts for originated loans, segmented by their purpose. The most common loan amounts are for home purchases and refinancing, with peaks around 200,000 to 300,000. On the other hand, loans for home improvement and other purposes are generally smaller, indicating these types of loans are often for less expensive projects.

3.3 Data Preprocessing

The third step in data preprocessing involved handling missing values, log transformation, and outliers. Before identifying missing values, hot encoding for categorical variables was used to convert them to numeric variables.

Handling missing values: Initially, 40 variables were identified and then dropped due to their irrelevance to the analysis. For instance, submitted year, all denial reason, all co applicant race and ethnicity etc. A comprehensive list of these dropped variables is provided separately in the Appendix section of this report.

For the remaining variables, imputation methods were selected based on the distribution, statistical analysis, and relationships to other variables:

- Mean imputation: Used for variables with a normal distribution, e.g., `interest_rate`.
- Mode imputation: Used for variables with a clear peak or common value, e.g., `loan_term`.
- Median imputation: Used for skewed distributions, e.g., `combined_loan_to_value_ratio`, `rate_spread`, `lender_credits`, `intro_rate_period`, `debt_to_income_ratio_numeric`.
- Regression imputation: Used for highly correlated variables:
  - Regression imputation 1: `total_loan_costs`, `origination_charges`, `discount_points` to each other vice versa.
  - Regression imputation 2: Imputing `property_value` using `loan_amount`, `income`, and imputing income using `property_value` and `loan_amount`.
  - Regression imputation 3: Imputing `applicant_age_numeric` using `co_applicant_age_numeric`, and vice versa.

Handling outliers: Outliers were identified and capped based on the 5th and 95th percentiles of each numeric column. Values below the lower bound were replaced with the lower bound value, and values above the upper bound were replaced with the upper bound value.

3.4 Data Analysis

The fourth step in data analysis involved creating Sankey diagrams to identify biases using demographic and financial factors at both national and state levels. Additionally, a correlation matrix was created using only approved loans to examine the relationships between the loan amount and other variables, as well as the relationships among the variables themselves.

Demographic analysis Relationship between ethnicity, race, gender, age, loan amounts, and all originated loans graphs are shown below. It helps in understanding how different demographic groups are distributed and how these distributions contribute to the total number of loans, and identify any bias in the dataset.

Figure 4, shows loan approval and denial rates across different ethnic groups. It highlights that Hispanic or Latino applicants have a slightly higher approval rate (52.7%) compared to the overall approval rate (52.9%). However, within this group, certain subcategories like Mexican and Puerto Rican applicants have notably higher denial rates of 53.8% and 53.5% respectively. Applicants whose ethnicity was Not Provided have a balanced approval rate (51.3%) and denial rate (48.7%), similar to the overall trend. Interestingly, applicants categorized as Not Hispanic or Latino also have a comparable approval rate (59.6%) to the overall rate. The category marked as Not Applicable shows an overwhelmingly high denial rate (89.5%), suggesting that there might be a significant issue with the applications in this category, possibly due to incomplete or incorrect information

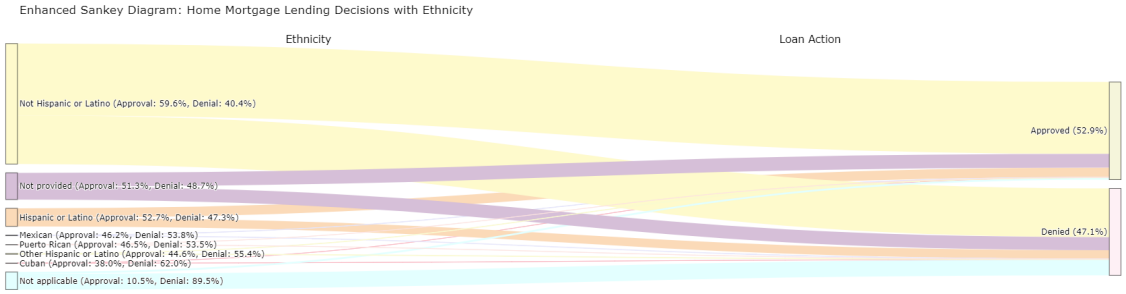


Figure 4: Ethnicity and Loan Amount

Figure 5 depicts loan approval and denial rates across different racial groups. White applicants have the highest approval rate at 60.9%, which is significantly higher than the overall approval rate of 52.9%. This suggests a favorable lending outcome for White applicants compared to other racial groups. Applicants who did not provide their race show an approval rate of 50.4% and a denial rate of 49.6%. This balanced outcome closely matches the overall trend, indicating no apparent bias for this group. Asian applicants, as a whole, have an approval rate of 57.8%. However, there are notable variations within the subcategories: Chinese (55.6%) and Native Hawaiian (51.1%) subcategories have higher approval rates. Japanese (61.1%) and Korean (53.4%) subcategories have lower approval rates but are still above the overall average.

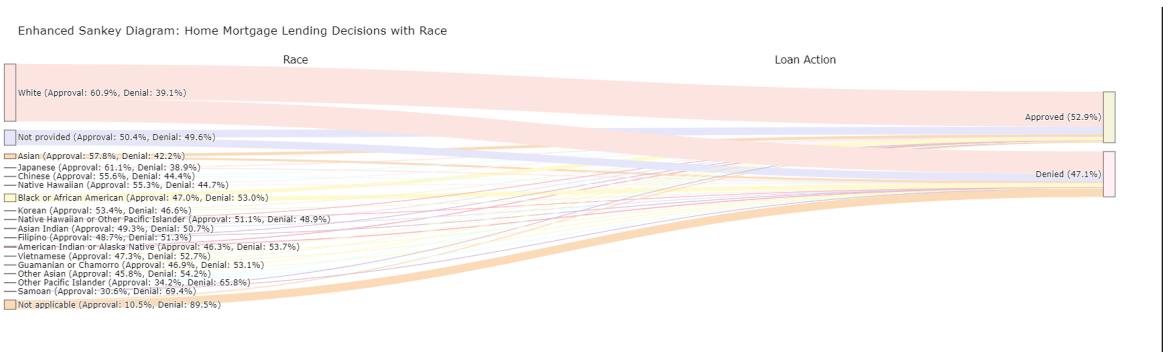


Figure 5: Race and Loan Amount

Black or African American applicants have a lower approval rate of 47.0% and a higher denial rate of 53.0%. This indicates a disparity in lending outcomes for this

group, suggesting potential biases or systemic issues that need addressing. Other groups have slightly less than 50% of approval rates, but other Pacific Islander and Samoan groups have lowest approval rates 34.2% and 30.6% respectively, for those racial groups have slight bias.

Figure 6, shows loan approval and denial rates across different gender groups. Male Applicants have the highest approval rate at 58.8%, suggesting a favorable trend towards loan approvals compared to other groups. Female Applicants With an approval rate of 56.9%, female applicants also have a relatively high approval rate, though slightly lower than their male counterparts. Not Provided applicants who did not provide their gender have an almost balanced approval (49.5%) and denial (50.5%) rate, indicating no significant bias but a need for further investigation into why gender was not provided. Applicants selected both male and female group has an approval rate of 50.3%, similar to the overall trend, but it highlights the complexity and potential challenges faced by individuals who do not identify strictly as male or female. Not Applicable category has a very low approval rate (10.4%) and a high denial rate (89.6%), suggesting significant issues with these applications, possibly due to incomplete or incorrect information.

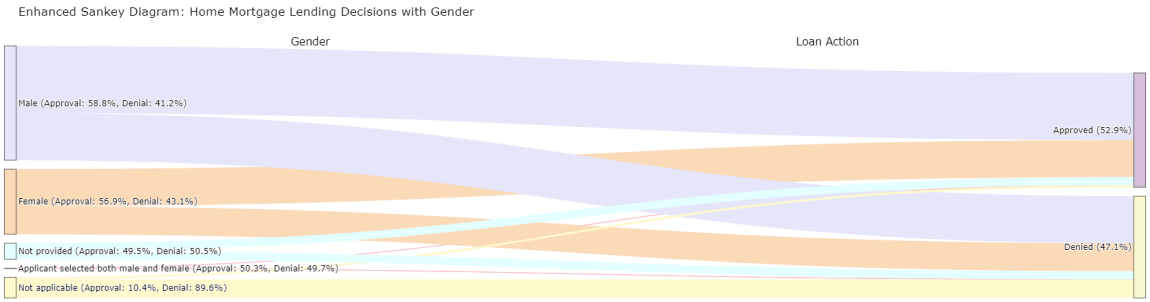


Figure 6: Gender and Loan Amount

Figure 7 illustrates loan approval and denial rates across different age groups. Young Adults (< 25 Years and 25-34 Years) groups have the highest approval rates, both have 61.4%. Middle-Aged Adults (35-64 Years) also shows relatively high approval rates, with those aged 35-44 years having a 58.6% approval rate, 45-54 years having a 57.3% approval rate, and 55-64 years having a 55.7% approval rate. Older Adults (65+ Years) approval rates decrease for older age groups. Applicants aged 65-74 years have an approval rate of 53.9%, while those over 74 years have an approval rate of 49.7%. Not Applicable category shows a very low approval rate (10.2%) and a high denial rate (89.8%).

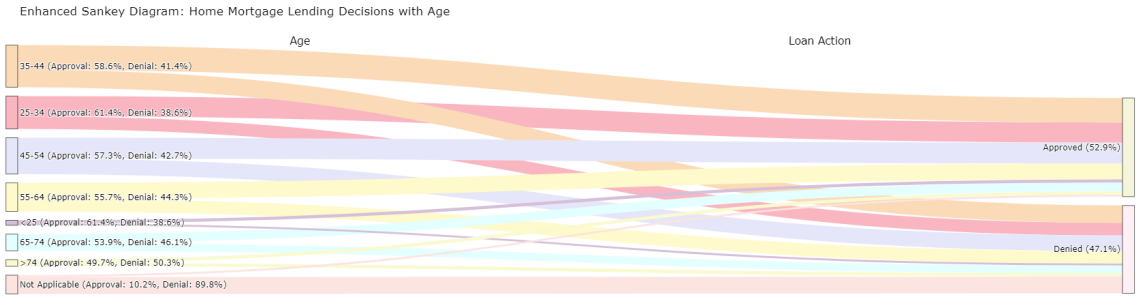


Figure 7: Age and Loan Amount

State Analysis A state-level analysis was conducted to explore potential biases in loan approval rates and loan amounts.

Figure 8 shows that the approval rates range from 0.475 to 0.65 by state. To create this plot, an extra dataset from the Cartographic Boundary Files - KML/GeoJSON dataset from the United States Census Bureau was used[2]. Northern states like North Dakota (ND), Iowa (IA), and Wisconsin (WI) exhibit higher home loan approval rates, indicated by the darker red shades. In contrast, Southern states, including Texas (TX), Louisiana (LA), Mississippi (MS), and Florida (FL), show the lowest approval rates, represented by the lighter shades.

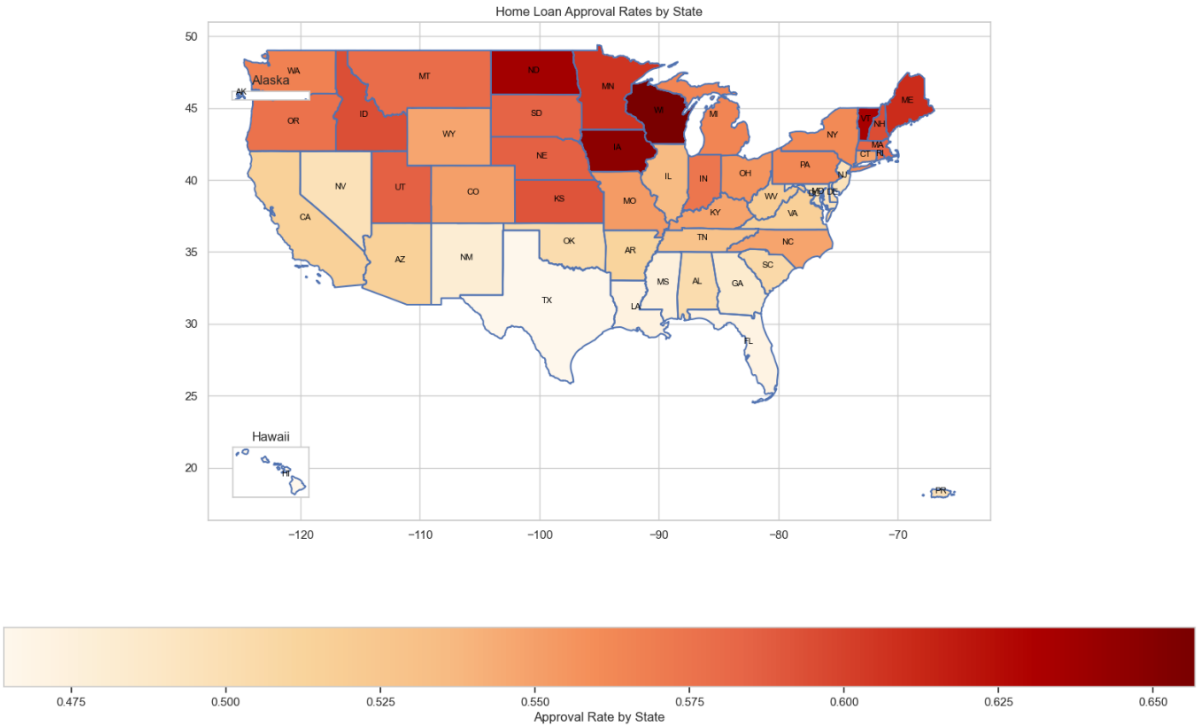


Figure 8: Loan approval rate by state

Figure 9 illustrates loan amount by state. The bar chart shows that states such as California (CA) and Hawaii (HI) have the highest average loan amounts, often exceeding 400k$, while states like West Virginia (WV), Arkansas (AR), and Mississippi (MS) typically fall below 200k$. This dual visualization highlights significant regional differences, suggesting that coastal states, particularly on the West Coast and the Northeast, have higher loan amounts.

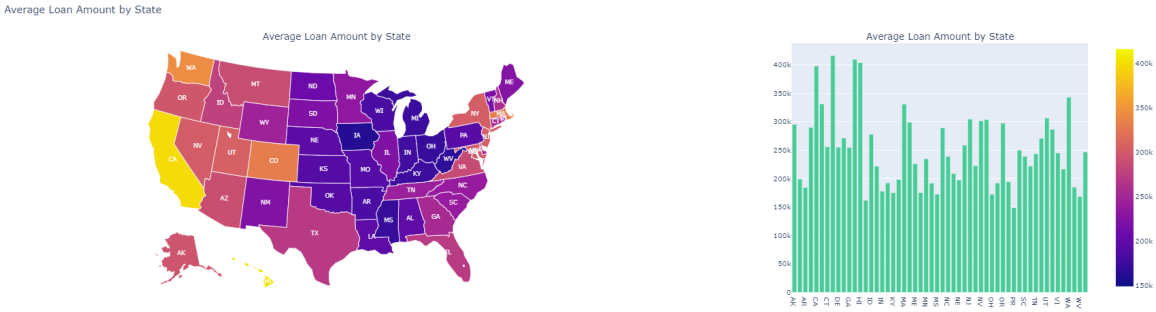


Figure 9: Loan amount by state

[2]Cartographic Boundary Files: `https://www.census.gov/geographies/mapping-files/time-series/geo/kml-cartographic-boundary-files.html`

Figure 10-13, heatmaps illustrate the approval rate by state and race, ethnicity age and gender. Figure 10 finds significant differences in approval rates between the states and races, which may indicate some bias. The Virgin Islands (VI) is at significant extremes, with **0%** approval for Asian, Black or African American, Native Hawaiian or Other Pacific Islander, and Not applicable categories, and **100%** for White, which indicates towards a possibility of racial. On the other hand, California (CA) and Ohio (OH) have almost equal approval rates of all the races, meaning that there is less racism. Also, the approval rates of the applicants vary depending on the city, but White applicants have a higher approval rate than the other racial categories, while Native Hawaiian or Other Pacific Islander has a lower approval rate.

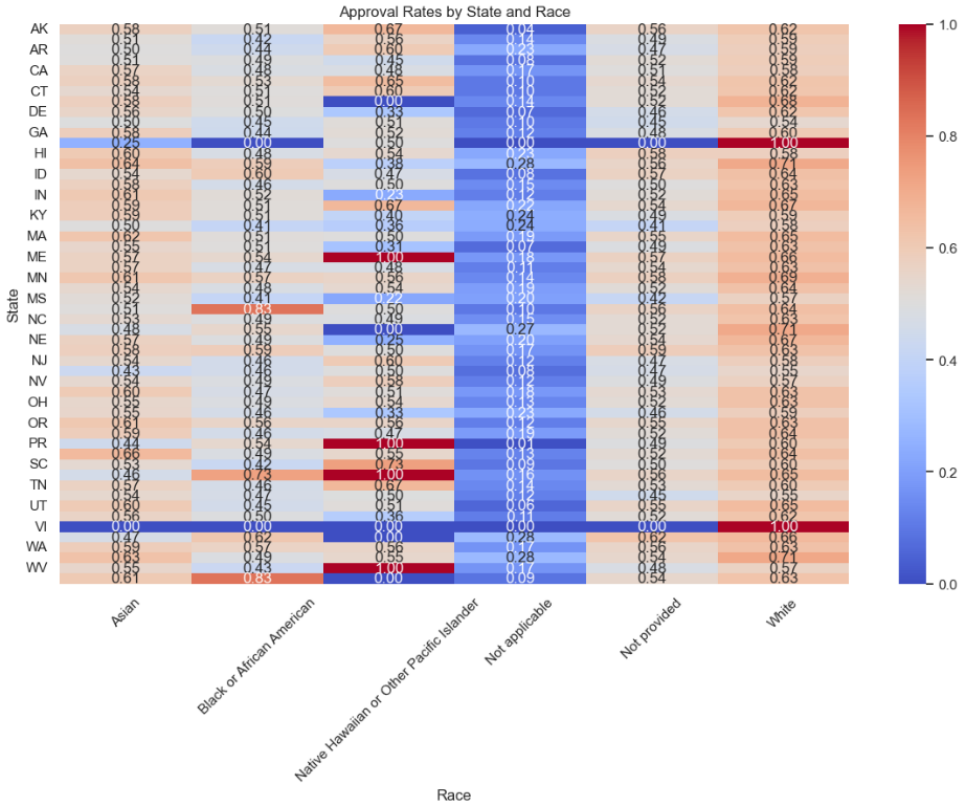


Figure 10: Approval rates by race and state

Figure 11 Hawaii (HI) shows extreme values with a **100%** approval rate for Hispanic or Latino and **0%** approval for Not provided. Virgin Islands (VI) also displays significant extremes, with a **100%** approval rate for Not Hispanic or Latino, and **0%** for Not Hispanic or Latino, Not Applicable, and Not provided groups. While states like Georgia and Massachusetts appear more balanced. The "Not applicable" and "Not provided" categories tend to have lower approval rates (indicated by blue cells)

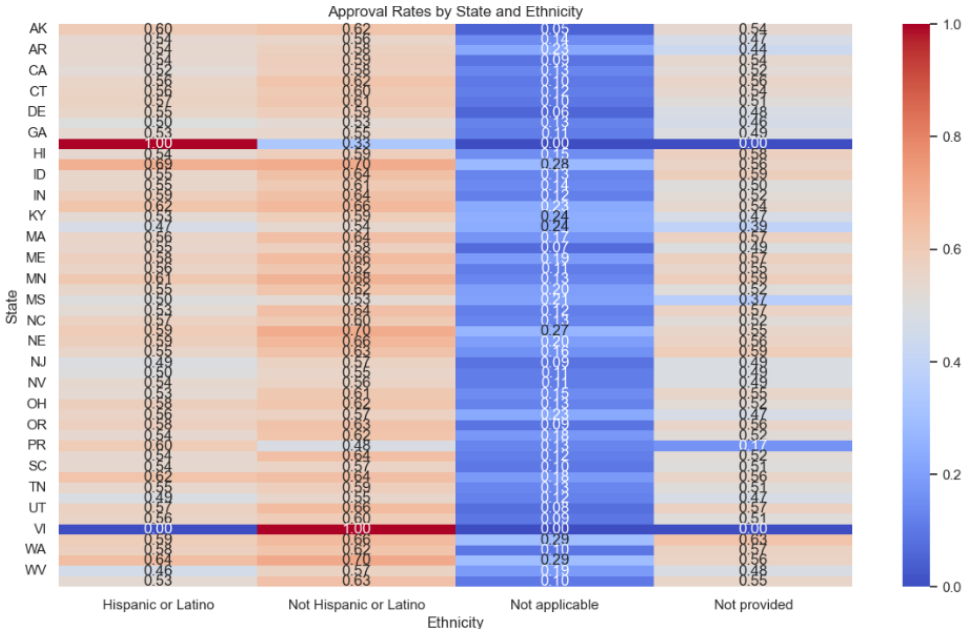


Figure 11: Approval rates by ethnicity and state

Figure 12 reveals potential gender biases in certain states, with the most significant extremes seen in the Virgin Islands (VI) display extreme values with **0%** approval for Female, Not applicable, and Not provided categories, and **100%** approval for Male. Additionally, most cities exhibit higher approval rates for men compared to women, with male approval rates having a minimum of **43%**. This pattern highlights a potential systemic bias favoring male applicants across various states.

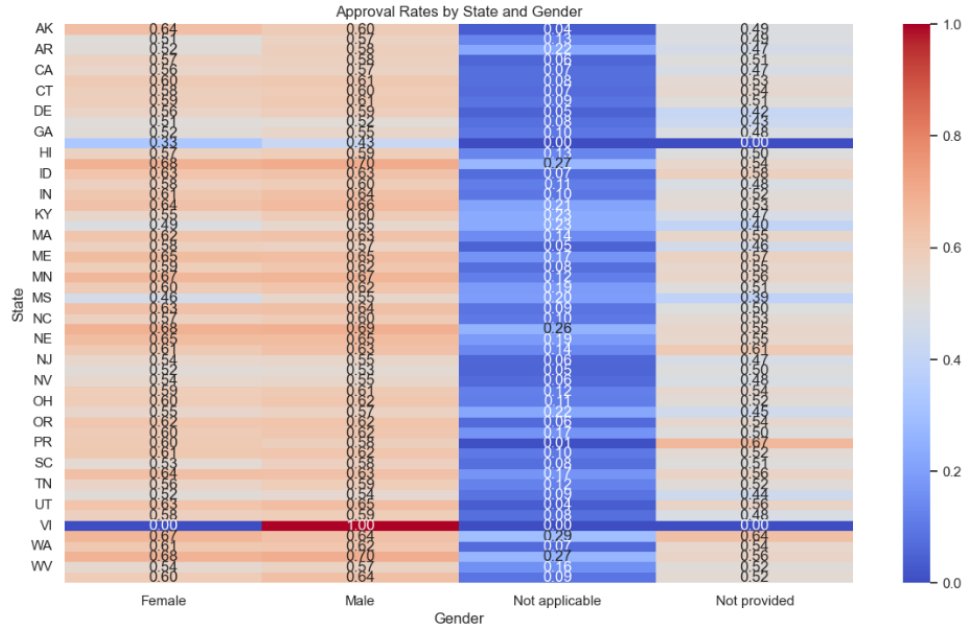


Figure 12: Approval rates by gender and state

Figure 13 visualizes approval rates across different states and age groups, ranging from 25 to 74 years old. Notably, Hawaii (HI) and the Virgin Islands (VI) exhibit extreme values, with HI showing **0%** approval for ages 35-44 and 45-54, and **100%** for ages 65-74,

while the Virgin Islands display **0%** approval for all age groups except 65-74, which is at **100%**.

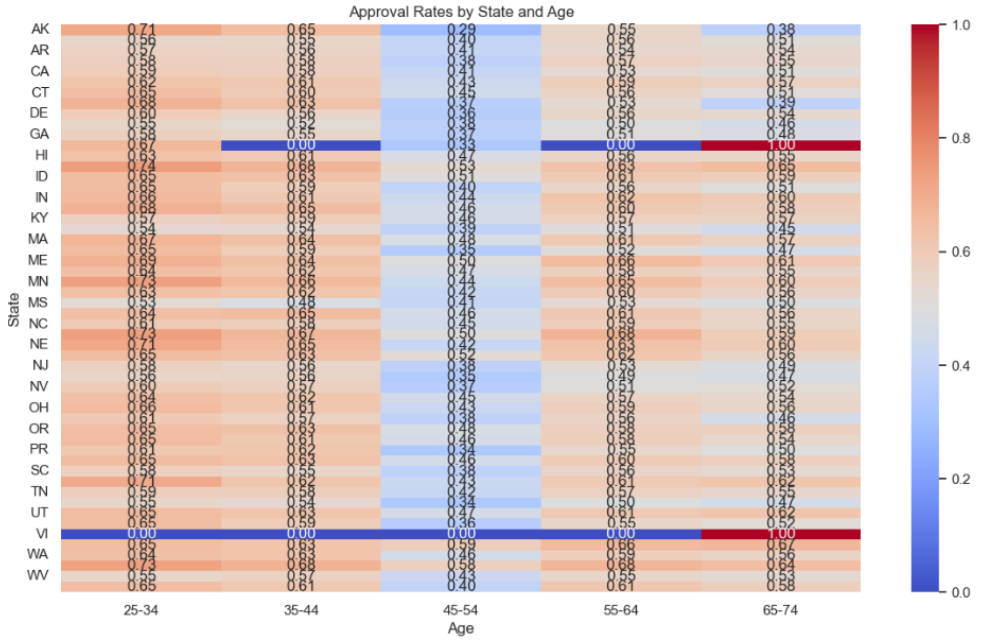


Figure 13: Approval rates by age and state

Other states show varied approval rates, with younger age groups (25-34) generally having higher approval rates across most states, transitioning to more mixed and often lower rates in middle age groups (35-54), and then fluctuating in older age groups (55-74).

Figure A3 offers a comparative analysis of property values, interest rates, credit scores, income levels, lien statuses, and loan terms across all states.

Figure A.3a shows that property values in Hawaii (HI) exceed **$700k**, with California (CA) and Washington, D.C. (DC) also exceeding **$700k**. Most states have property values around **$400k**, indicating that their financial abilities are not significantly different.

Figure A.3b demonstrates that interest rates are relatively similar across states, ranging from 4% to 5%.

Figure A.3c illustrates notable variation in average credit scores. For instance, states like Arkansas (AR) have high scores around **6**, indicating strong financial health, while the Virgin Islands (VI) show lower scores around **3**. Most other states have similar credit scores, close to **5**.

In Figure A.3d, observe that DC and VI stand out with an exceptionally high average income of approximately **160** units, compared to the more consistent range of **100-200** units in other states, such as California (CA) with **150** units.

The lien status distributions (Figure A.3e and Figure A.3f) highlight that states such as California (CA), Florida (FL), and Texas (TX) have high counts of first liens, exceeding **70k**. Subordinate lien statuses in these states are also significant but comparable in number. The frequency of first lien statuses is three times higher than that of subordinate liens, indicating that people tend to have first liens for mortgage loans.

Lastly, in Figure A.3g, loan terms typically range between **300-350** months across all states. The analysis reveals key relationships: higher incomes generally correlate with higher property values and larger loan amounts, as seen in DC and CA.

Figure A.4 shows the correlation matrix for the most related 10 variables to loan amount. As expected, `property_value` is the most correlated variable with the loan amount, indicating its significant impact on the loan amount. Given its strong correlation, `property_value` will be removed from the modeling process to prevent multicollinearity issues. Similarly, `total_loan_costs` and `origination_charges` are also removed due to their strong correlation with each other.

Transformation: After imputing missing values, the Shapiro-Wilk test was conducted for each continuous numeric variable to determine whether the data follows a normal distribution. A p-value greater than 0.05 indicates a normal distribution. For features that were not normally distributed, a log transformation was applied.

3.5 Modelling

The fifth step in modelling involved determining feature importance and selecting the model with the lowest Mean Absolute Error (MAE), Mean Squared Error (MSE), and highest R-Squared accuracy. Feature importance was assessed using Random Forest, which identified the top 20 most significant variables for predicting loan amounts. Following this, several models were built and evaluated, including Linear Regression, Decision Tree, Random Forest, AdaBoost, Gradient Boosting, and an Ensemble method that combines Random Forest, Gradient Boosting, and AdaBoost. Each model's performance was evaluated using a validation set and cross-validated to ensure robustness. Hyperparameters were tuned to optimize model performance.

3.6 Evaluation

The sixth step in evaluation involved analyzing metrics from each model to determine the optimal model. Compared models based on their MAE, MSE, and R-Squared values. Additionally, visualizations such as performance curves and comparison plot was used to assess model performance. Cross-validation was employed to ensure that the models' performance was consistent and not due to overfitting. The model with the lowest MAE, MSE, and the highest R-Squared value was selected as the optimal model. These steps ensure that the dataset is preprocessed effectively and that the best model is selected based on rigorous evaluation criteria.

4 Design Specification

This section describes the methods, structure, and foundation of the implementation that are useful in the solution for the detailed analysis of loan amount distributions.

Linear Regression is one of the simplest and most commonly used algorithms for supervised learning of continuous data. It models the association between a dependent variable and one or more independent variables by estimating a regression equation on the collected data Seber and Lee (2012). Due to its ease of interpretation and understanding, Linear Regression is a good starting point for loan amount prediction. The algorithm achieves the least square between the actual and predicted values so as to fit the model properly. Linear Regression was selected for this study because of its simplicity and nature of offering information on the importance of various features.

Decision Tree employs a tree structure of decisions where the internal nodes are the input features, branches are the decision rules, and the end nodes are the outcome Song and Lu (2015). The model for the target variable is the simple decision rules that are learned from the features of the data. Decision Trees are most effective when it comes to interpretability and comprehensibility since they replicate human decision-making. They are ideal for use with categorical data, which is the type of data that is present in the features of the dataset used in this study.

Random Forest is a technique of ensemble learning for classification as well as regression. This creates many decision trees and combines their results to increase accuracy and decrease overfitting Breiman (2001). Every tree is created from a random sample of the training data and the final decision is reached by taking the mean of all the trees. It is also not sensitive to outliers and is efficient when applied to big data and large feature spaces. It was chosen for this study because of its capability of dealing with the interactions between the features and high accuracy without fine-tuning of the hyperparameters.

Gradient Boosting creates models one after another, and the new model tries to minimize the errors of the previous model Friedman (2001). It employs a gradient descent algorithm in order to minimize the loss function, and therefore is very suitable for predictive tasks. Gradient Boosting is said to be very accurate but it can easily overfit the data if the right hyperparameters are not set. In this study, Gradient Boosting was used as it has the ability to enhance the prediction performance especially by handling the difficult to predict instances Natekin and Knoll (2013).

AdaBoost or Adaptive Boosting, is an ensemble method that uses the outputs of weak learners to build a strong learner Freund and Schapire (1997). It alters the weights of misclassified instances in such a way that other models pay more attention to the difficult ones. AdaBoost is also useful in the reduction of bias and variance hence improving the overall accuracy of the prediction. This algorithm was selected since it is simple and has been proven to improve the performance of the base learners employed in this research.

K-Nearest Neighbors is a simple and intuitive algorithm that is used for both classification and regression tasks. It predicts the target variable based on the K most similar instances in the feature space Cover and Hart (1967). The simplicity and efficiency of KNN make it a good choice for smaller datasets and scenarios where interpretability is important.

XGBoost is an optimized gradient boosting algorithm designed for speed and performance Chen and Guestrin (2016). It handles sparse data efficiently and is highly flexible, making it a popular choice for many machine learning competitions and real-world applications. XGBoost was selected for its robustness and ability to handle large datasets effectively.

LightGBM is a gradient boosting framework that uses tree-based learning algorithms Ke et al. (2017). It is designed to be efficient and scalable, making it suitable for large datasets. LightGBM was chosen for its ability to handle high-dimensional data and deliver high performance with reduced training time.

CatBoost handles categorical features automatically and efficiently Prokhorenkova et al. (2018). It is particularly effective in dealing with datasets containing a mixture of categorical and numerical features, which is common in financial data. CatBoost was selected for its high accuracy and ease of use with categorical data.

Ensemble Method combines the results from the three best models, namely, Random Forest, XGBoost, and Catboost. This approach takes advantage of each of the models and produces the best final prediction as compared to the other approach. As

it is with most machine learning techniques, the ensemble methods are used in order to boost the performance of a model by using several models where each of the models brings to the table a different strength in the prediction Zhou (2012).

5 Implementation

This section aims to identify the model with the highest accuracy and lowest errors for loan amount prediction, utilizing all models described in Section 3.5.

Initially, employed Random Forest Regression (n_estimators=50, random_state=42) on the approved loan dataset to determine the top 20 most important variables for predicting loan amounts. The resulting feature importances are illustrated in Figure A.5.

Based on the analysis, 20 variables such as `lien_status`, `income`, `loan_term`, `combined_loan_to_value_ratio`, `debt_to_income_ratio_numeric`, `ffiec_msa_md_median_family_income`, `tract_to_msa_income_percentage`, `state_code_encoded`, `rate_spread`, `occupancy_type`, `tract_median_age_of_housing_units`, `discount_points`, `tract_minority_population_percent`, `interest_rate`, `derived_msa_md`, `tract_population`, `tract_one_to_four_family_homes`, `tract_owner_occupied_units`, `loan_purpose` and `co_applicant_age_numeric` were selected for further modeling.

To evaluate the models, employed a standard data split approach, dividing the data into training, validation, and test sets.

The data was first split into a training set and a test set, with 20% of the data reserved for testing. The remaining 80% of the data was then further split into training and validation sets, with 25% of the remaining data allocated to validation, resulting in a final split where the validation set constitutes 20% of the original data.

This approach ensures that the models are trained and validated on different subsets of the data, and the final performance metrics are evaluated on an independent test set. The hyperparameters used for each model are shown in Table 1.

Model	**Set 1 (n_est=10)**	**Set 2 (n_est=20)**	**Set 3 (n_est=40)**
Linear Regression	-	-	-
Decision Tree	r_state=30	r_state=50	r_state=80
Random Forest	r_state=30, n_est=10	r_state=50, n_est=20	r_state=80, n_est=40
Gradient Boosting	r_state=30, n_est=10	r_state=50, n_est=20	r_state=80, n_est=40
AdaBoost	r_state=30, n_est=10	r_state=50, n_est=20	r_state=80, n_est=40
K-Nearest Neighbors	n_neighbors=10	n_neighbors=20	n_neighbors=40
XGBoost	r_state=30, n_est=10	r_state=50, n_est=20	r_state=80, n_est=40
LightGBM	r_state=30, n_est=10	r_state=50, n_est=20	r_state=80, n_est=40
CatBoost	r_seed=30, iter=10	r_seed=50, iter=20	r_seed=80, iter=40

Table 1: Model Hyperparameters

To evaluate each model, used standard metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R^2) score. Each model was trained and tested on the same dataset to ensure comparability of results.

- **Linear Regression**: Used as a baseline model without hyperparameters.
- **Decision Tree**: Tuned with different random states to test the impact of initial splits.

- **Random Forest**: Evaluated with varying numbers of estimators and random states to balance bias and variance.
- **Gradient Boosting, AdaBoost**: Tuned similarly to Random Forest for consistency in boosting performance.
- **K-Nearest Neighbors**: Number of neighbors set to test the effect on the model's smoothness.
- **XGBoost, LightGBM**: Tested with different random states and estimators to optimize boosting efficiency.
- **CatBoost**: Iterations and random seed adjusted to explore the effect on convergence and stability.

6 Evaluation

In this section, will discuss the results from modeling. As illustrated in Table 2, MAE, MSE, and R-squared metrics from test set for all models are presented. Validation and test set metrics were almost the same, indicating that the model's performance generalizes well to unseen data. Notably, the models in Set 3, Random Forest, XGBoost, and CatBoost exhibited the lowest error metrics and highest R-squared values. For the ensemble method, these three models were combined with the same hyperparameters as in Set 3.

Model	**Set 1**			**Set 2**			**Set 3**		
	MSE	**MAE**	**R^2**	**MSE**	**MAE**	**R^2**	**MSE**	**MAE**	**R^2**
Linear Regression	0.273	0.375	0.571	0.192	0.328	0.699	0.192	0.328	0.699
Decision Tree	0.273	0.375	0.572	0.274	0.376	0.569	0.272	0.375	0.574
Random Forest	0.145	0.279	0.773	0.136	0.271	0.787	0.133	0.266	0.792
Gradient Boosting	0.318	0.455	0.503	0.239	0.388	0.626	0.181	0.328	0.717
AdaBoost	0.262	0.417	0.589	0.275	0.430	0.570	0.287	0.442	0.551
K-Nearest Neighbors	0.290	0.397	0.545	0.310	0.413	0.512	0.343	0.437	0.463
XGBoost	0.152	0.295	0.761	0.136	0.274	0.787	0.126	0.262	0.802
LightGBM	0.270	0.414	0.577	0.192	0.341	0.699	0.156	0.297	0.756
CatBoost	0.158	0.299	0.753	0.143	0.282	0.776	0.132	0.270	0.793

Table 2: Performance metrics of different models across three datasets. Best performing values are highlighted in yellow.

Figure 14 is showing best 3 models and ensemble method all metrics. As shown, the ensemble method achieved the highest R-squared value and the lowest error metrics, with XGBoost following closely.

Although the imputation of missing values and handling of outliers were thorough, incorporating additional domain-specific data and more sophisticated feature engineering could further enhance model performance. Optimized hyperparameter tuning and cross-validation are recommended for further model improvement.

The results are consistent with prior studies indicating that ensemble methods tend to be more accurate than individual models due to their ability to learn various aspects of the data.

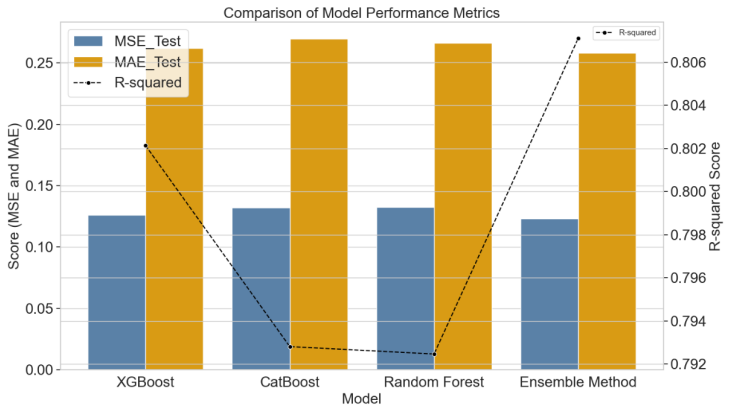


Figure 14: XGBoost, CatBoost, Random Forest and Ensemble method metrics comparison

7 Conclusion and Future Work

The study's purpose was to examine loan amount distributions in the context of the 2022 FFIEC HMDA dataset to improve the assessment and decision-making of financial risks in mortgage lending. The study also used shuffling, data chunking, exploratory data analysis (EDA), and data cleaning as some of the techniques used in the study. Different models were created, and it was discovered that ensemble method outperformed than others.

Key findings of the study include:

- Gender Bias: Out of all the applicants, the male applicants had the highest approval rate at 58.8% while the female applicants were slightly lower at 56.9%.

- Age Bias: The approval rates were the highest among the young people under 25 years and 25-34 years, and the rates decreased as the age increased.

- Ethnicity and Race Bias: White candidates received the highest approval of 60.9%. However, Black or African American applicants had a lower approval rate of 47.0% which shows that there is a disparity in the lending results of this group. The Virgin Islands showed a clear racial prejudice with 0% approval rates for Asian, Black and other minority groups.

- State-Level Analysis:The northern states such as North Dakota, Iowa, and Wisconsin had a higher approval of home loans. The disapproval rate was the lowest in the southern states of Texas, Louisiana, Mississippi, and Florida. California and Hawaii had the highest average loan amount while the majority of the other states had average loan amount below 0. 4 million.

- Model Performance: The Ensemble method gave the highest R-squared value and the lowest error metrics which shows that the Ensemble method was the most accurate in predicting the loan amounts.

Future work: Future efforts should focus on analysing factors and evaluating the models across the entire dataset, rather than using random subsets, to ensure comprehensive performance assessment. And further improvement of the presented methodology

of selecting an appropriate set of hyperparameters will be a crucial step. Incorporating domain-specific data preprocessing—such as handling unique data inconsistencies or anomalies identified through expert insights—will enhance the models' accuracy and relevance. In addition, extending the understanding of how to evaluate the feature importance, for instance by the permutation feature importance, and principal component analysis (PCA) could generate a better understanding of data characteristics and improving models' performance.

References

Avery, R. B., Brevoort, K. P. and Canner, G. B. (2007). The 2006 hmda data, *Federal Reserve Bulletin* .

Bhutta, N., H. A. and Ringo, D. (2024). How much does racial bias affect mortgage lending? evidence from human and algorithmic credit decisions, *Federal Reserve* .

Bhuyan, B. P., Tomar, R. and Cherif, A. R. (2023). Statistical Machine Learning in Loan Analysis from Financial Institutions, *in* A. Ramdane-Cherif, T. P. Singh, R. Tomar, T. Choudhury and J.-S. Um (eds), *Machine Intelligence and Data Science Applications*, Springer Nature, Singapore, pp. 405–418.

Bocian, D., Ernst, K. and Li, W. (2008). Race, ethnicity, and subprime home loan pricing, *Journal of Economics and Business* **60**(1-2): 110–124.

Breiman, L. (2001). Random forests, *Machine Learning* **45**(1): 5–32.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794.

Coalition, C. R. (2016). Paying more for the american dream vi: Racial disparities in fha/va lending.

Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification, *IEEE Transactions on Information Theory* **13**(1): 21–27.

Delis, M. and Papadopoulos, P. (2019). Mortgage lending discrimination across the us: New methodology and new evidence, *Journal of Financial Services Research* **56**(3): 341–368.

Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting, *Journal of Computer and System Sciences* **55**(1): 119–139.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine, *Annals of Statistics* **29**(5): 1189–1232.

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T. and Walther, A. (2020). Predictably unequal? the effects of machine learning on credit markets, *Federal Reserve Bank of New York* .

Guy, R. F., Pol, L. G. and Ryker, R. E. (1982). Discrimination in mortgage lending: The Home Mortgage Disclosure Act, *Population Research and Policy Review* **1**(3): 283–296.

Hanson, A. and Hawley, Z. (2016). Discrimination in mortgage lending: Evidence from a correspondence experiment, *The Journal of Urban Economics* .

Institute, U. (2023). Home mortgage disclosure act data. Available at: `https://www.urban.org`.

Journal, A. B. (2024). Fed study finds 'limited role' for racial bias in mortgage lending, *ABA Banking Journal* .

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree, *Advances in Neural Information Processing Systems*, pp. 3146–3154.

Natekin, A. and Knoll, A. (2013). Gradient boosting machines, a tutorial, *Frontiers in Neurorobotics* **7**: 21.

of Minneapolis, F. R. B. (2023). Higher mortgage denials for solo applicants feed racial disparities in lending. Available at: `https://www.minneapolisfed.org/article/2023/higher-mortgage-denials-for-solo-applicants-feed-racial-disparities-in-lending`.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. and Gulin, A. (2018). Catboost: Unbiased boosting with categorical features, *Advances in Neural Information Processing Systems*, pp. 6638–6648.

Reserve, C. F. (2022). Home mortgage lending by race and income in a time of low interest rates. Accessed: 2024-08-09.
URL: *https://www.clevelandfed.org/publications/cd-reports/albtn-20221129-home-mortgage-lending-by-race-and-income-in-a-time-of-low-interest-rates*

Reserve, F. (2024). How much does racial bias affect mortgage lending?, *Federal Reserve Report* .

Ross, S. L. and Yinger, J. (2002). *The Color of Credit: Mortgage Discrimination, Research Methodology, and Fair-Lending Enforcement*, MIT Press.

Seber, G. A. F. and Lee, A. J. (2012). *Linear Regression Analysis*, John Wiley & Sons.

Song, Y. Y. and Lu, Y. (2015). Decision tree methods: Applications for classification and prediction, *Shanghai Archives of Psychiatry* **27**(2): 130–135.

Zhou, Z. H. (2012). *Ensemble Methods: Foundations and Algorithms*, CRC Press. Available at: `https://www.crcpress.com/Ensemble-Methods-Foundations-and-Algorithms/Zhou/p/book/9781439830031`.

A Appendix

A.1 Distribution of variables

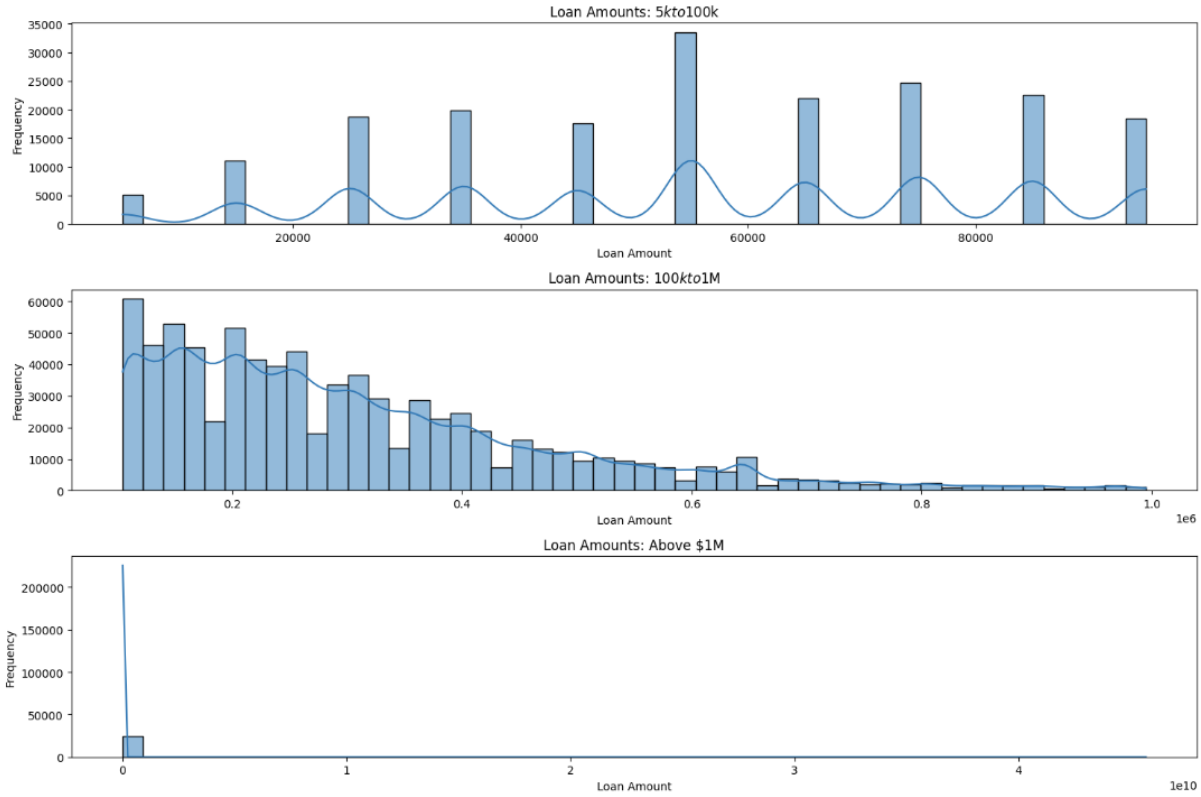


Figure A.1: Loan distribution

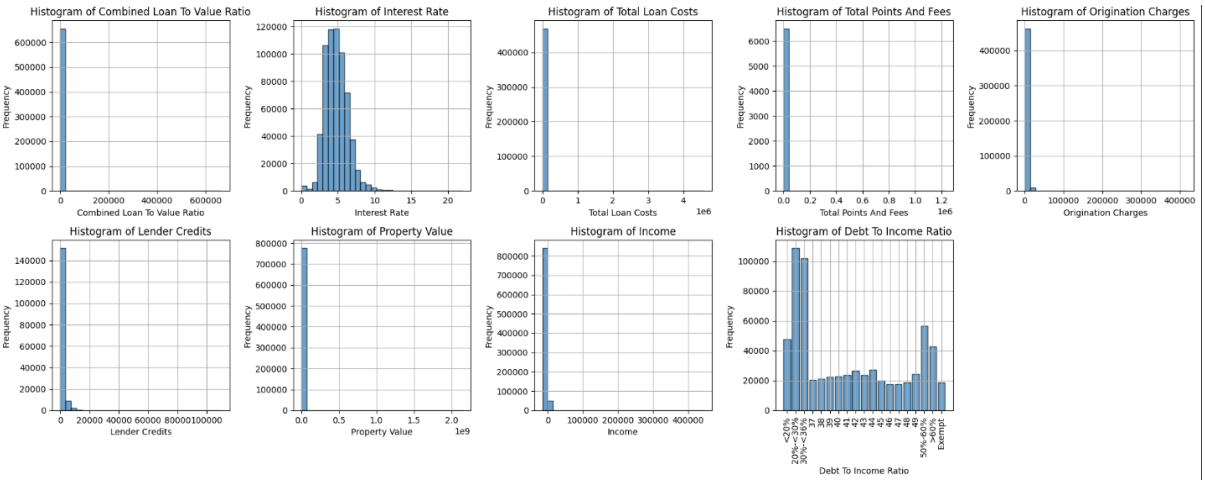


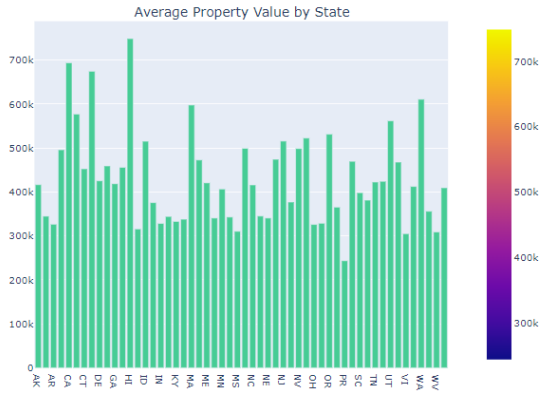
Figure A.2: Distribution of CLV Ratio, Interest Rate, Total Loan Costs, Total Points and Fees, Origination Charges, Lender Credits, Property Value, Income, and DTI Ratio

A.2 Dropped Variables

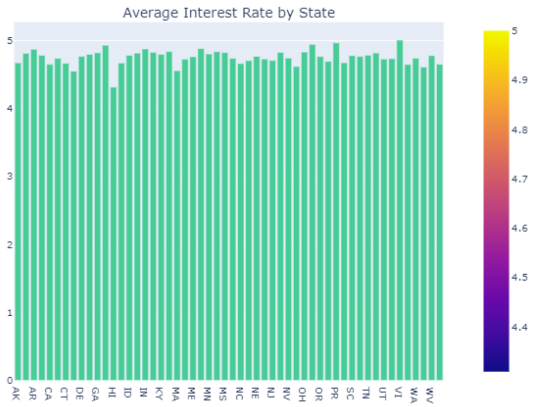
In the process of data analysis, several variables were identified but ultimately dropped due to their irrelevance to the analysis. The dropped variables are listed below:

- activity_year
- interest_only_payment
- balloon_payment
- other_nonamortizing_features
- co_applicant_credit_score_type
- applicant_ethnicity_2
- applicant_ethnicity_3
- applicant_ethnicity_4
- applicant_ethnicity_5
- co_applicant_ethnicity_2
- co_applicant_ethnicity_3
- co_applicant_ethnicity_4
- co_applicant_ethnicity_5
- applicant_race_2
- applicant_race_3
- applicant_race_4
- applicant_race_5
- co_applicant_race_2
- co_applicant_race_3
- co_applicant_race_4
- co_applicant_race_5
- aus_1
- aus_2
- aus_3
- aus_4
- aus_5
- denial_reason_2
- denial_reason_3
- denial_reason_4
- census_tract
- lei
- open_end_line_of_credit
- multifamily_affordable_units
- applicant_age_above_62
- co_applicant_age_above_62
- prepayment_penalty_term
- total_points_and_fees
- derived_race
- derived_sex
- derived_ethnicity

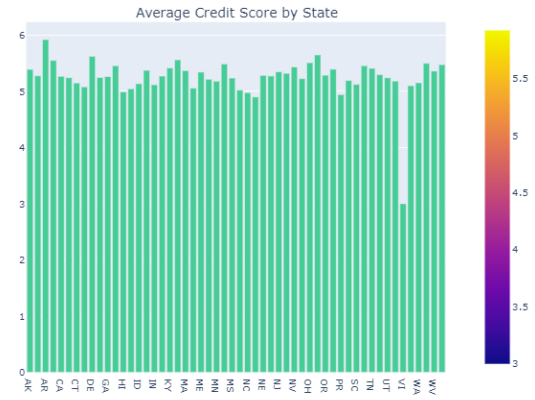
A.3 State Analysis



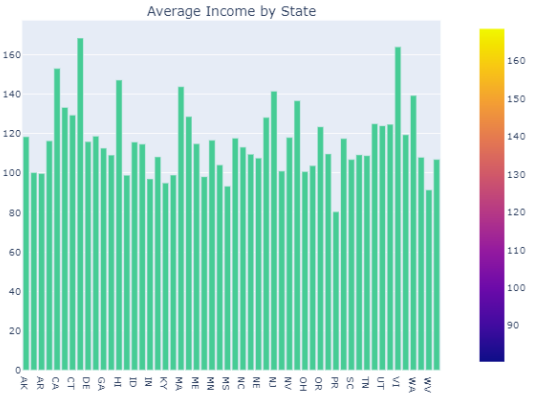
(a) Property value by state



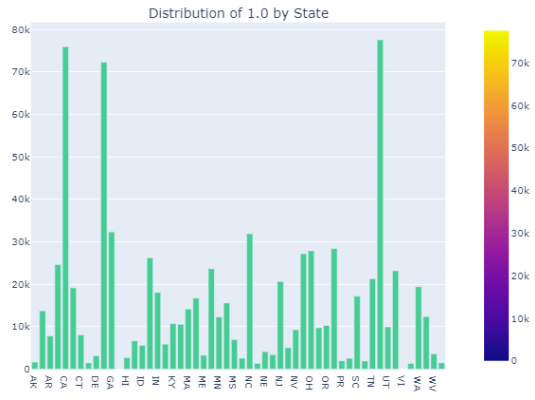
(b) Interest rate by state



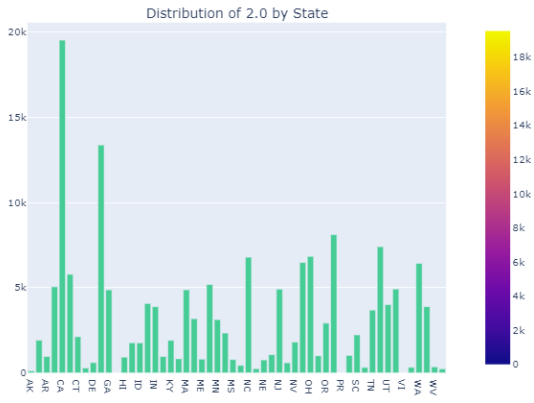
(c) Credit score by state



(d) Income by state

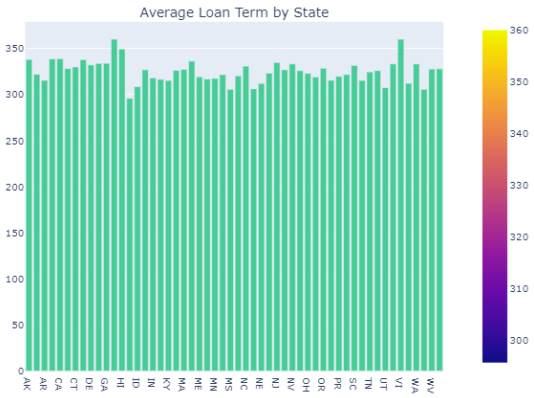


(e) A first lien by state



(f) A subordinate lien by state

Figure A.3: Property value, Interest rate, Credit score, Income, and Lien status by State



(g) Loan term by state

Figure A.3: Loan term by State

A.4 Correlation Matrix

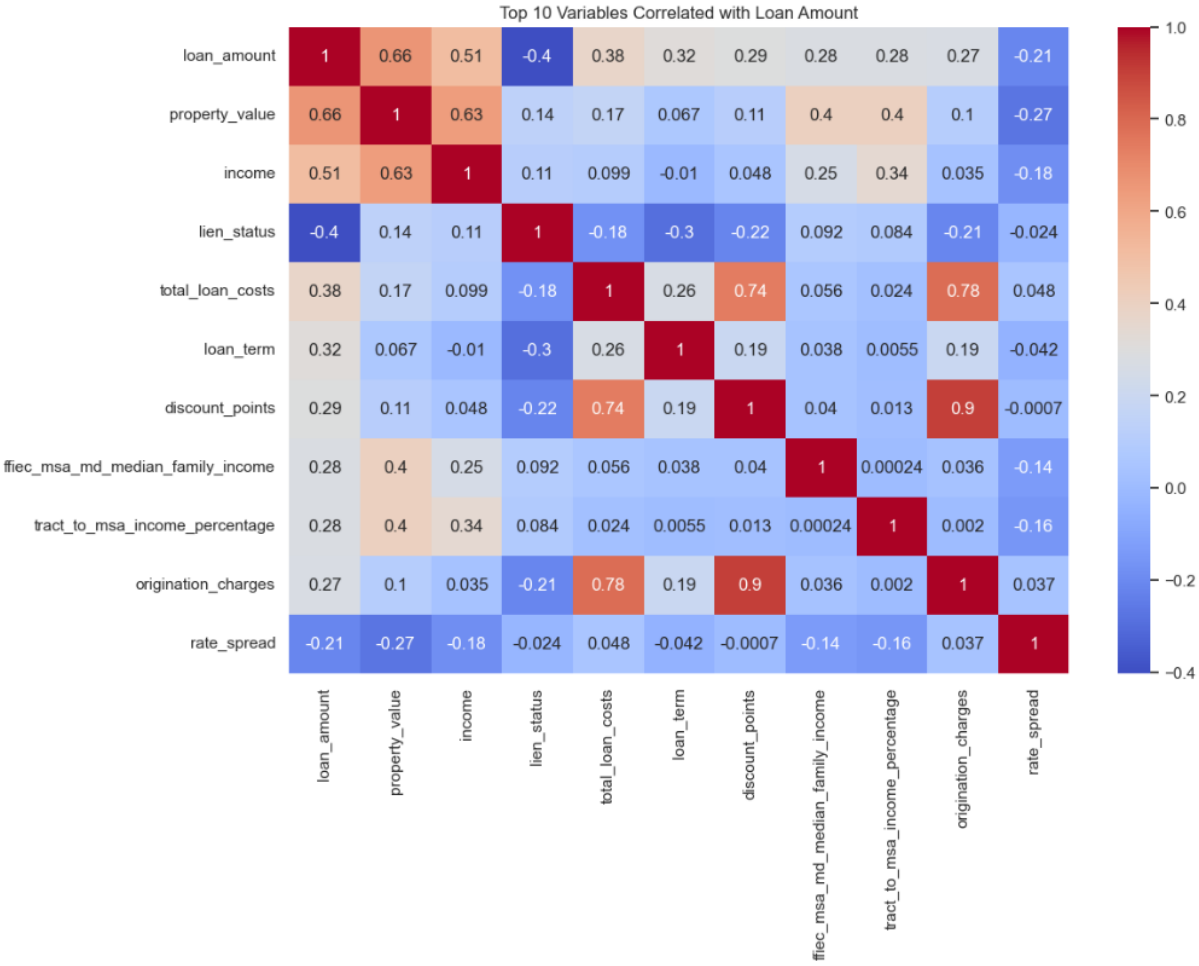


Figure A.4: Top 10 correlated variables to loan amount

A.5 Feature Importance

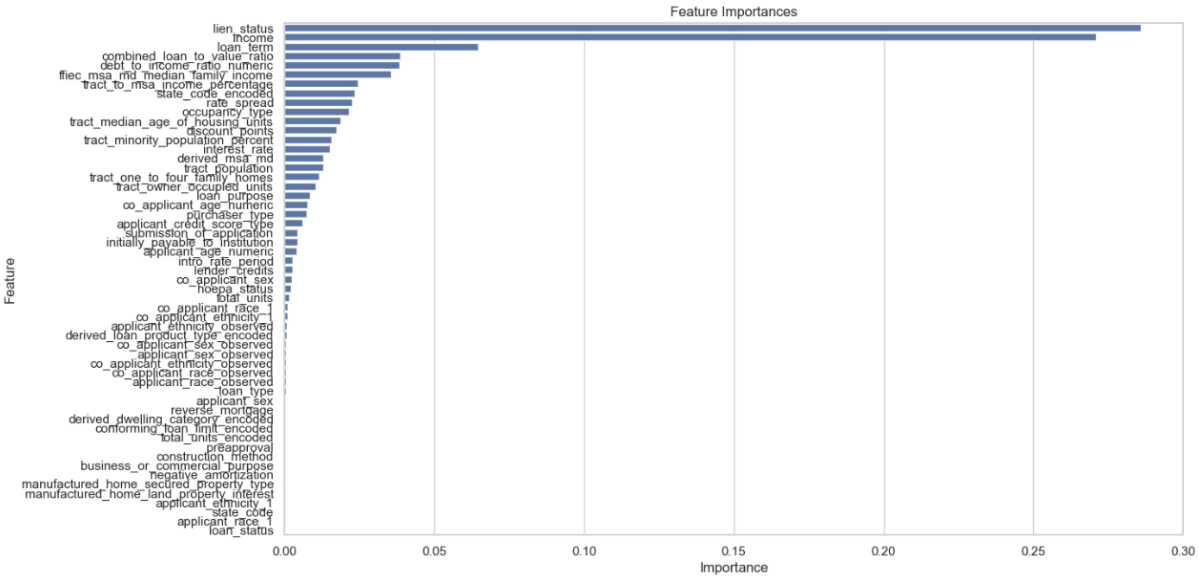


Figure A.5: Top 20 Most Important Variables for Predicting Loan Amount