**Methodology**

**Data Preprocessing**

The data used for this thesis is from a Public Use Database by the Federal Home Loan Bank (FHLB) System, which is a network of government-sponsored banks in the United States. It was established as part of the Federal Home Loan Bank Act in 1932 to provide stability and liquidity to the housing and community lending markets.

It consists of 11 regional banks, each serving a specific geographic region of the United States. These regional banks are cooperatives owned by their member financial institutions, which include banks, credit unions, thrifts, insurance companies, and community development financial institutions. The FHLB’s mission is to support the housing industry and community development and to ensure access to affordable housing for Americans through low-cost funding and other financial services. It acts both directly, by investing in affordable housing and community development projects, and indirectly through its member financial institutions, by purchasing mortgage loans from them and by making advances that are typically used to fund residential mortgage loans. The ultimate goal is to provide additional liquidity to the market.

The FHLBanks do not receive federal funds; instead, they fund themselves mainly by issuing consolidated obligations in the public capital markets. They are able to keep their rates only slightly above the comparable U.S. Treasury rates thanks to their status as a government-sponsored enterprise, and all banks are jointly liable for the System’s debt.

The FHLB System is regulated and overseen by the Federal Housing Finance Agency (FHFA), an independent federal agency. By ensuring that the System remains financially stable, well-capitalized, capable of raising funds and committed to its housing finance mission, the FHFA is able to guarantee the stability and effectiveness of the system itself.

(fonte:<https://www.fhfa.gov/SupervisionRegulation/FederalHomeLoanBanks/Pages/About-FHL-Banks.aspx>)

For this analysis, we have utilized data pertaining to mortgages acquired by each Federal Home Loan Bank during the years 2021, 2020, and 2019, which includes census tract-level information. This data has been published in an unaudited form, as reported by the Federal Home Loan Banks to the FHFA.

(INSERIRE LINK DATI XXXXXX <https://www.fhfa.gov/DataTools/Downloads/Pages/Public-Use-Databases.aspx> )

In total, the dataset considered contains XXXXXX observations, with XXXXXX variables relating to the personal characteristic of the primary and secondary borrower, if present, to their financial situation and area of residence (for a more comprehensive list of the variables used, please refer to the Table 1 in the appendix).

The variable of interest in our analysis is the interest rate, which has a distribution spreading between 1.625% and 6.25%, as shown in Figure 1. While it can be roughly approximated with a normal distribution, it presents a spike on the left side around 3%.

A blue and black line graph

Description automatically generated

Figure . InterestRate distribution and comparison with fitted normal distribution

All variables have undergone a preprocessing phase, which involved cleaning numerical codes and eliminating redundant or uninformative columns. Additionally, distributional checks have been made to ensure that categorical variables are coded in meaningful ways and that outliers will not skew our results. Then, we assessed the proportion of missing values on the dataset and decided to drop them in cases where they constituted a small percentage of our dataset. For instance, the percentage of missing values in the variable *Age1* was 0.002%.

For what concerns the sensitive variables needed for this analysis, we consider the age and gender of the applicants as originally reported in the data, while we use the available ethnicity variables to code a new variable *Minority* indicating the minority status of the applicant. This new variable is obtained using the variable *Borrower1Race1Type*, which presents the following levels:

1. American Indian or Alaska Native,
2. Asian,
3. Black or African American,
4. Native Hawaiian or other Pacific Islander,
5. White,
6. Information not provided by Borrower,
7. Not Applicable (First or primary borrower is an institution, corporation or partnership).

From Figure 2(a) we can see that the distribution is not balanced, with a vast majority of loan applicants belonging to the category “5 – White”, and much fewer applicants in the other categories.

Therefore, we will group categories 1, 2, 3 and 4 into a single minority category.

The resulting variable *Minority* has three levels, as shown in Figure 2(b):

1. Minority,
2. Missing,
3. White.

A blue and white bar graph

Description automatically generated

Figure . Distribution of variables Borrower1Race1Type and Minority, the latter resulting from the aggregation of categories 1, 2, 3 and 4 of the first

It must be noted that the dataset is clearly unbalanced since we have a vast majority of observations that belong to the non-minority category, and this will need to be considered during the analysis.

The original dataset presents four additional variables indicating other ethnic traits of borrowers. However, they will be excluded from our analysis, since most observations fall under the following categories: 5=White; 6=Information not provided; 7=Not Applicable (borrower is an institution, corporation or partnership).

Considering the variables about the second borrower’s ethnicity, we find once again a majority of white applicants, and a consistent number of loans for which there is no co-borrower. However, we can find the same patterns as before, hence we group all minority categories into the same one and exclude the other race variables which do not provide any additional information.

**Data Analysis**

Exploratory data analysis is a crucial preliminary step before constructing a regression model. It allows us to unveil hidden data patterns and is a first step towards the objective of identifying potential instances of discrimination.

FARE LO STESSO PER GENDER E AGE??? COME POSSO FARE CON AGE?

Firstly, we analyze the relationship between *InterestRate* and the newly created *Minority* variable. ~~Figure 3 presents a Kernel Density Estimation (KDE) plot of~~ *~~InterestRate~~* ~~by~~ *~~Minority~~*~~, which is a graphical representation of the probability density of this continuous variable.~~ From the KDE plot, we can observe that on the lower end of the interest rate distribution minority applicants tend to be assigned lower interest rates. This suggests that there may be a trend of favorable interest rates for minority applicants at the lower range of rates. On the contrary, on the upper end of the interest rate distribution the opposite is true, where minority applicants tend to be assigned higher interest rates. This indicates that at the higher range of interest rates, when other risk factors are likely to be present simultaneously, the minority status of an applicant may be perceived as a higher risk by bankers. Consequently, this results in a trend of less favorable interest rates for minority applicants, potentially raising concerns about discriminatory lending practices if this difference cannot be traced back to other factors.

A diagram of a person and person

Description automatically generated

Figure 3

After these preliminary steps, numerical and categorical variables were subject to two separate flows of cleaning and exploration, consisting respectively of outlier detection, winsorization and correlation analysis for numerical variables and balance checks, recategorization and encoding for categorical variables.

Upon inspecting the numerical variables available in the dataset, we observed significant skewness in their distributions and the presence of outliers in several cases. Therefore, we decide to apply winsorization to the most extreme values, in order to obtain a more balanced dataset while keeping the informational value of these data points. The threshold used was determined using the interquartile range, a measure of the spread of the middle 50% of the data. In particular, the lower threshold was established by subtracting three times the IQR from the first quartile and the upper threshold by adding three times the IQR to the third quartile. Any data points falling below the lower threshold or above the upper threshold were considered outliers.

Two exceptions to this process are the variables *PaymentCount* and *PMI.* The variable *PaymentCount* represents the Term of the Mortgage in Months, and we chose not to apply winsorization due to its limited meaningfulness and the highly skewed distribution. Applying winsorization in this case would result in a loss of all information content within the variable. Similarly, the variable *PMI* represents the percent of mortgage balance at origination covered by loan level PMI (Private Mortgage Insurance), and it is also highly skewed towards 0. By the same reasoning, modifying this variable would lead to loss of information therefore winsorization was not applied.

After a standardization step, we ran a correlation analysis to study the relations between the variables. From this analysis we identified some very strong correlations between pairs of variables, which are due to the fact that they present very similar information.

For example, the variables *Year* and *NoteDate* presented a correlation of 0.96. We decided to exclude the variable *Year*, indicating the year in which the mortgage was acquired by the Federal Home Loan Bank, and we kept *NoteDate*, the variable indicating the year in which the mortgage was originated. This is due to the higher correlation that the year of origination of a mortgage has with the interest rate associated to it. In fact, the target variable, *InterestRate*, presents a very strong negative correlation of -0.72 with *NoteDate*, which suggests that in more recent years interest rates have gone down. This is coherent with the macroeconomic trends over the period: as a matter of fact, the FED funds rate has been decreasing over the period considered. This can be clearly seen in Figure 4, which presents a Kernel Density Estimation (KDE) plot of *InterestRate* by *NoteDate*, which is a graphical representation of the probability density of this continuous variable by year of origination.

A graph of different colored lines

Description automatically generated

Figure : KDE plot of InterestRate by NoteDate

The matrix in Figure 5 shows pairwise correlations among the numeric variables considered in our analysis.

The target variable *InterestRate* shows strong correlation with *NoteDate*, as mentioned before, and with *PaymentCount*, which can be traced back to the fact that longer-term mortgages are normally associated with higher interest rates.

Notably, the third highest value of the correlation coefficient between *InterestRate* and the other variables is with *MinorityRatio*, with a value of 0.2. The value is small, however it indicates that mortgages originated in areas with higher minority ratios present slightly higher interest rates.

Regarding the other variables, we can identify strong correlation between the age of borrower 1 and the age of borrower 2, which we can explain if we assume that often the second borrower is borrower 1's partner.

*HousingExpense* and *DebtExpense* present a correlation of 0.6, as we can expect, meaning that higher housing expenses are associated with higher debt expenses.

Additionally, *MonthlyIncome* presents a positive correlation of 0.58 with UPB, and negative correlations with both *HousingExpense* and *DebtExpense*, suggesting that people with lower income tend to borrow less.

High correlation can also be seen for *LTV* and *PMI*, indicating that for higher Loan-to-Value ratios we can find higher percentages of mortgage balance at origination covered by loan level PMI.

Lastly, we highlight that *CreditScore1* and *CreditScore2* do not present high correlations with the other variables, not even between the two.

A graph with numbers and symbols

Description automatically generated with medium confidence

Figure 5

From a graphical inspection, it is clear that some of the variables needed to be recoded or excluded for the analysis since they are extremely unbalanced. In particular, variables with a too low variance were removed because they don't have any relevant informational content, with nearly all observations falling in the same category.

Additionally, gender variables are recoded, grouping all options other than male and female in the same category, as well as the number of borrowers for each loan, which is transformed into a dummy indicating whether there is more than one borrower or not.

Lastly, some geographical variables were excluded from the analysis because they presented too much variance, such as the county and area code, while the *State* variable was kept.

The final set of categorical features is therefore made up of 14 features, which were encoded for the subsequent analysis.