1. Introduction

This project involves the analysis and prediction of defaults on already granted mortgages. This point is very important; this analysis and model cannot be extrapolated to the granting of mortgages. There is a survivor bias since the institutions that have granted the loans have already made a selection from the general population.

2. Methods and Discussion

2.1. Data Processing and análisis

2.1.1 Data Cleaning

Different approaches to data cleaning have been carried out. Rows with artifacts have been deleted, for example, in ID, and strange data has been corrected, such as in IDs and dates. In other cases, such as TotalIncome, due to its significant impact and the large number of issues, statistical techniques have been used to replace erroneous data.

2.1.2 Handling of missing values

The following methodologies were followed:

1) When there was a very high percentage (+60%) of missing data and it was understood from a business perspective that these variables would not be very determinant, the columns were deleted.
2) When there were not many missing values and the categories were small, the data was filled with the most frequent value.
3) Using heuristics. Assumptions were made about how the relationships between variables work (e.g., interest rates and credit scores).
4) Using Random Forest, interpolating from the rest of the dataset.

2.1.3 Descriptive and temporal análisis

It is evident that dates exert an influence, as macroeconomic fluctuations impact the mortgage market. To capture this influence, moving averages of the percentage of mortgage defaults were calculated and added to the dataset. The author acknowledges that this approach introduces data leakage, which is further justified later in the text.

For future research, the modeling of interest rates through time series analysis has been left as a potential development avenue.

2.1.4 Target analysis

There has been no direct analysis of the target, however, it has been verified that each column had statistical influence on the target using Chi2 or Anova.

2.2. Predictive Model

2.2.1. Feature Engineering

New variables have been created. Categories were extracted from IDs, moving averages were calculated to represent the number of temporary defaults, and variables such as the total loan duration and the age at which the borrower was granted the loan were derived.

To some extent, values like *BorrowerCreditQuality* and *BureauScoreValue* have been extensively reconstructed and can be considered practically new features.

Additionally, binary variables were constructed to indicate missing values for continuous data, such as the borrower's age at the time of applying for a mortgage.

2.2.2. Feature Selection

An effort was made to keep the model as simple as possible. Categorical variables with low-frequency categories were grouped under "Other" to reduce the model's complexity.

Variables with more than 60% missing data were deleted, as well as variables where all values were identical, such as *BorrowerType*.

2.2.3. Machine Learning algorithm

A PCA (Principal Component Analysis) was chosen to address collinearity issues, followed by a logistic regression. This combination allows the model to retain explainability, as the PCA transformation can be reversed.

An XGBoost model was also implemented to provide a comparison of the model's performance. While XGBoost tends to deliver better results than logistic regressions, it comes with the trade-off of reduced explainability.

2.2.4. Scoring

The results of both models have been poor, indicating the need for more thorough data cleaning. As expected, logistic regression performed worse in terms of both F1 and ROC-AUC scores.

A critical metric in the construction of credit models is recall, as minimizing false negatives is intuitively a priority. XGBoost also outperformed logistic regression in this metric.

3.Limitacions

Pipeline Limitations: Due to the project's scope, the necessary training and prediction pipelines for data processing were not implemented. The current method of data transformation introduces a data leak, particularly through the assignment of moving averages. Furthermore, the absence of these pipelines resulted in the lack of a validation set to verify whether the training performance reflects actual performance. This constitutes another form of data leakage, as decisions made during data modeling were influenced by the outcomes.

Data Deficiency: The most critical dataset, *credit score*, was nearly empty. From a business perspective, the credit score is designed to predict the probability of default. Efforts were made to reconstruct this feature as accurately as possible within the project's scope. However, access to more reliable credit score data would likely lead to significantly improved default predictions.