**Feature selection: the main drivers of interest rates**

To insert :

* Tabella da *results\_fs\_STANDARDIZED.xlsx. – TESTO*
* Png di MRI (in fondo al doc) – APPENDIX?
* *MRI\_scores.xlsx – APPENDIX SICURO*

Struttura:

1. Spiegare in linea generale cos’è FS, perché viene applicata e lo scopo (cioè identificare quali sono le variabili più importanti per determinare interest rate)
2. Metodi usati
3. Excluding second borrower
4. INSERIRE TABELLA XLSX
5. results – solo riformulare se serve
6. commento ai risultati

Testo:

1.

Different feature selection algorithms are applied to the data to identify the most relevant and informative from all available features in the dataset. The goal of this analysis is to understand whether any sensitive variables play a significant role in determining loan interest rates.

2. metodi usati

Three feature selection methods were employed in this analysis. First, Mutual Information Regression was applied, which measures the dependency between two variables. When the value is zero, it indicates independence, whereas higher values signify a stronger dependency. (SEE FIGURE XXXXX? AND TABLE XLSX XXXXX?)

Second, a K-fold analysis was conducted with Select K Best, a technique that selects a given number k of features based on the k highest scores. In this case, the scoring function used was f\_regression, which sequentially evaluates the impact of individual regressors using a linear model. This method helps identify the most influential features in predicting the target variable.

The third approach employed is Recursive Feature Elimination (RFE). RFE utilizes an external estimator, in this case, a Support Vector Regressor (SVR) with a linear kernel, which assigns weights to the features. The primary objective of RFE is to iteratively select features by progressively considering smaller subsets of features. Initially, the estimator is trained on the entire set of features, and the importance of each feature is assessed. Subsequently, RFE systematically eliminates the least important features from the current set, repeating this process until the desired number of features is retained.

All methods are applied twice, to select the 5 and 10 most relevant features, to have a deeper understanding of the magnitude of their influence on the variable of interest.

These methods are helpful in identifying the most relevant features for predictive modeling, and as they follow different approaches, we combine their outputs to leverage their differences.

3. excluding second borrower

From the results of the Mutual Information regressor (FIGURA XXXXX), it seems clear that those variables related to the second borrower on a given loan are not extremely correlated with interest rates.

Therefore, we exclude variables related to the second borrower for 2 main reasons. Firstly, we believe it is unlikely that the additional borrower might have such an important impact of the determination of an individual's credit score. Nonetheless, we will keep the dummy indicating whether a second borrower is present or not (BorrowerCount) since this might have some impact. Secondly, only 60% of observations have an additional borrower, therefore the informational content is missing for a large part of the observations.

**4-RESULTS: ok**

INSERIRE TABELLA XLSX – CAPIRE FORMATTAZIONE

As shown in TABLE XXXXXX, the variables which are selected by most algorithms, and hence have the most influence on the target variable, are:

* *PaymentCount:* term of the Mortgage in Months.
* *NoteDate:* year the mortgage was originated.
* *FirstTime:* numeric code indicating whether borrower is a first time homebuyer (0=No, 1=Yes).
* *SelfEmployed:* numeric code indicating whether the borrower is self- employed (0=No, 1=Yes).
* *MinorityRatio:* the percentage of the property's census tract population that is minority.
* *HUDMedIncome:* current median income for a family of four for the area as established by HUD.
* *LoanPurpose\_1:* dummy indicating whether the purpose of the loan belongs to class 1 i.e. Purchase mortgages, as opposed to refinancing loans (code 2 for no cash out and code 6 for cash out). Around 30% of the observations fall in this category, since the classes in the variable *LoanPurpose* are almost balanced.

**5. COMMENT ON THE RESULTS**

Variables directly related to individuals’ protected information were not selected. However, it is important to note that the variable MinorityRatio, which indicates the percentage of individuals belonging to a minority group in the area, is among those considered most influential on the interest rates associated with individual mortgages. This highlights indirect and more subtle discrimination, either in the form of a proxy for the ethnicity of the individual or driven by economic factors. For instance, this may be linked to the fact that, on average, areas predominantly inhabited by people of color correspond to areas with lower incomes, increased instability, and higher crime rates. Nevertheless, this constitutes a very strong signal with significant implications for policymakers and legislators. It also raises opportunities for research and ethical considerations regarding how an individual's environment can influence their economic success or determine their access to services, as in this case, access to credit.

Additionally, among the most influential variables is *NoteDate* because of the macroeconomic correlation between interest rates on mortgages and the FED funds rate, which has been decreasing consistently over the period considered. Therefore, the origination year of the loan plays a big role in predicting its interest rate.

Lastly, we can note that PaymentCount, FirstTime, SelfEmployed, HUDMedIncome are all valid financial variables that are considered when evaluating loan applications. In particular, HUDMedIncome might be so strongly correlated to the interest rate because of the high correlation that it has with residents’ incomes in the given area.

Fine

**Appunti miei**

**Methods used:**

1. **K-fold analysis with Select K Best:** Select features according to the k highest scores, using f\_regression as score function, which uses a linear model for testing the effect of a single regressor, sequentially for many regressors.
2. **Mutual Information Regression:**

Mutual information (MI) between two random variables is a non-negative value, which measures the dependency between the variables. It is equal to zero if and only if two random variables are independent, and higher values mean higher dependency. (<https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.mutual_info_regression.html>)

10 highest mutual information scores (*MRI\_scores.xlsx*):

NoteDate 0.429103

HUDMedIncome 0.364674

CensusTractMedFamIncome 0.180638

PaymentCount 0.141290

MinorityRatio 0.113670

SelfEmployed 0.098010

FirstTime 0.092436

DebtExpense 0.037864

LTV 0.034194

NoteAmount 0.033553

1. **Recursive Feature Elimination:** We use an external estimator that assigns weights to features which is an SVR with linear kernel, and the goal of recursive feature elimination (RFE) is to select features by recursively considering smaller and smaller sets of features. First, the estimator is trained on the initial set of features and the importance of each feature is obtained. Then, the least important features are pruned from current set of features until the desirred number of features is left.

## Excluding 2nd borrower

From the results of the Mutual Information regressor (FIGURA XXXXX), it seems clear that those variables related to the second borrower on a given loan are not extremely correlated with interest rates.

Therefore, we exclude variables related to the second borrower for 2 main reasons. Firstly, we believe it is unlikely that the additional borrower might have such an important impact of the determination of an individual's credit score. Nonetheless, we will keep the dummy indicating whether a second borrower is present or not since this might have some impact. Secondly, only 60% of observations have an additional borrower, therefore the informational content is missing for a large part of the observations.