Model selection

For model selection, the dataset was divided into a training set and a test set, following a 4:1 ratio, using the *scikit-learn* package in Python. (XXXXXX citare da Wang).

The models explored are:

* LinearRegression
* RANSAC
* NeuralNetwork
* RegressionTree
* RandomForest
* XGBoost
* LightGBM
* HybridModel
* Stacking

All of these models exhibit distinct features and performance attributes, with some adapting better to the specific data we are dealing with. They are all regression models used to predict the values of our target variable I*nterestRate*, using as independent variables all the other variables available, as presented in Table 1 in the Appendix.

The models chosen for this comparative analysis were selected by drawing from the existing literature and with the intent to address unexplored gaps, by selecting different options that were not applied in this field before. In the subsequent sections, we will explore their characteristics more in detail and evaluate their results.

The **metrics** used for the evaluation and comparison of the models are:

* Mean Squared Error (MSE), which calculates the mean of the squared difference between actual and predicted values. It is commonly used because it is differentiable and commonly understood, with the ideal model having a MSE of 0 and higher values as the fit worsens. However, it is sensible to outliers.
* Mean Absolute Error (MAE), which calculates the mean absolute difference between actual and predicted values. It maintains the same unit as the target variable and is robust to outliers.
* Mean Residual (MR), which calculates the mean difference between actual and predicted values, taking into account the direction of the error. In particular, a positive mean residual suggests that the model tends to overpredict, while a negative mean residual suggests underprediction.
* R-squared, or coefficient of determination, which measures the proportion of variance in the dependent variable that can be explained by the independent variables.

Each of them highlights a different aspect of the performance of the model, and considering them together we can have a better understanding of the goodness of fit.

FINO A QUI OK

1. **Linear Regression:**

Linear regression is one of the simplest and most applied methods for regression. It is highly interpretable, and it aims to establish a linear relationship between the dependent variable and one or more independent variables.

The metrics are reported in table XXXXXX, and they seem to show quite a good performance, especially in terms of the R-squared, which indicates that around 71% of the variance of *InterestRate* can be explained by the other variables in the analysis.

Upon inspecting the coefficients and p-values for each individual predictor variable, we can see that the most standard financial indicators are all significant at a 5% significance level, including *CreditScore1, CreditScore2, MonthlyIncome, HousingExpense, DebtExpense, NoteAmount, LTV, PMI.* This is reasonable given that these are the factors that should be considered when determining the interest rate for a loan.

Additionally, we can find coefficients statically different from zero for some variables that are considered sensible and shouldn’t be playing a role in determining the *InterestRate*.

In particular, some geographical variables seem to have an impact, and we can infer that the average interest rate varies from state to state. Moreover, the variable *MinorityRatio* has a statistically significant positive coefficient, which the interest rate assigned to a given mortgage will be higher in areas which have higher ratios of the minority residents. This is not necessarily an indicator of discrimination, since it might be a proxy for other risk factors that could correlate with it, but it AAAAAAAAAA.

Similarly, we find that *Age1* has a positive correlation with *InterestRate*, suggesting that older individuals will be assigned lower rates, and this could be again traced to some other economic factor such as a better credit history or higher incomes.

However, we find no evidence that other sensitive variables, such as the minority status or the gender, have a significant impact on mortgage rates.

**NoteDate**

Lastly, the linear regression model run presents a high number of insignificant coefficients, especially when it comes to the encoded categorical variables, which will be better analyzed with non-linear models.

1. **RANSAC (RANdom SAmple Consensus):**

The performance of the classical Linear Regression highly depends on the normality assumption of the residuals, which is violated in this case, leading to the poor results discussed above. Therefore, we explore a different approach to the regression problem by using the robust regression algorithm RANSAC. (Fischler & Bolles, 1981 XXXXXXXX)

SPIEGARE DA WANG

Escluso, non migliora la performance della LinearRegression, anzi fa peggio. È immaginabile, dato che i dati non sono adatti a questo tipo di modello

* + RANSAC is a robust regression algorithm used to fit a model to a dataset with outliers. It iteratively selects random subsets of data points to build models, effectively ignoring outliers and improving the overall model's robustness 🡪 scritto da chatgpt, rifare prendendo da wang.

**AGGIUNGERE RSULTATI**

1. **Neural Network:**

Given the apparent non-linearity in the data, we switch to analyse machine learning models that can be more suited to the relationships within our dataset, starting from neural networks.

A deep feedforward neural network, as described by Schmidhuber in 2015 XXXXXXXXX CITARE !!!!, comprises multiple layers within the model, and the connections between nodes in different layers are complete. When employed for regression analysis, it operates without the necessity of making prior assumptions, such as assuming the normality of residuals. Consequently, this feature renders it a suitable option for predicting mortgage note rates.

Cv

Config wang

Risultati

1. **Regression Tree:**
   * A regression tree is a decision tree used for regression tasks. It recursively splits the data into subsets based on the values of independent variables, and each leaf node provides a regression prediction.
2. **Random Forest:**
   * Random Forest is an ensemble learning method that combines multiple decision trees to improve predictive accuracy and reduce overfitting. It is widely used for classification and regression tasks.
3. **XGBoost (Extreme Gradient Boosting):**
   * XGBoost is a gradient boosting algorithm known for its high performance. It builds an ensemble of decision trees sequentially, optimizing for both predictive accuracy and computational efficiency.
4. **LightGBM:**
   * LightGBM is another gradient boosting framework that focuses on speed and efficiency. It uses a histogram-based learning method to achieve faster training and improved accuracy.
5. **Hybrid Model:**
   * A hybrid model combines multiple machine learning techniques or models to address specific complex problems. These models can consist of a mix of regression, clustering, neural networks, and other methods.
6. **Stacking:**
   * Stacking, or stacked generalization, is an ensemble learning technique that combines multiple models by using their predictions as input to a meta-model. This approach can improve overall predictive performance by leveraging the strengths of individual models.