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 prendere cit 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 (Wang, Truong) 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 (RANdom SAmple Consensus), which is used to fit models to data that may contain outliers or noise. (Fischler & Bolles, 1981 XXXXXXXX)

The algorithms begins by randomly selecting a subset of data points, fits a model using this subset and identifies inliers (data points with residuals smaller than a predefined tolerance level) and outliers (data points that deviate significantly). Then it checks if the model has enough inliers to meet a predefined threshold and this process is repeated iteratively until a stopping criterion is met to determine the best model.

After applying the RANSAC linear regression, we verify that it yielded even poorer results than the classic Linear Regression model. This outcome can be attributed to the nature of the dataset, which is not well-suited for Linear Regression, because of the presence of various encoded categorical variables.

Therefore, the next models examined have a clear non-linearity focus to ensure that these variables’ relationship with *InterestRate* can be extrapolated.

* Ok!

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. Neural networks, powered by the advances in hardware, excel at handling large datasets, which makes them extremely appreciated in several fields, such as for scientific, business, and AI applications. When employed for regression analysis, they operates without the necessity of making prior assumptions, such as assuming the normality of residuals. Consequently, this feature renders neural networks a more suitable option for predicting mortgage note rates.

For this application, we experimented with several model configurations ranging from very simple and shallow to more complex and deep ones, and we finally pick the best performing one on the validation set. The final neural network is therefore composed of 3 HIDDEN LAYERS WITH

three hidden layers are used with 10 nodes in the first hidden layer, 8 nodes in the second hidden layer, and 6 nodes in the third hidden layer. RISCRIVERE QUANDO FINISCE DI RUNNARE !!!!

All layers use the *tanh* activation function, one of the most popular choices for the activation function and the most appropriate for this application. Mean square error is used as the loss function, and the weights are trained using stochastic gradient descent algorithm with a batch size of 254 and epoch of 100.

The results obtained from the neural network are notably superior to those of the linear regression (Table XXXXXXX). In fact, all the considered metrics show improvements, indicating that this model is better suited for the data at hand, particularly due to its ability to capture non-linear relationships. DIRE PEGGIORAMENTO MR CHE PERò NON è ROBUSTO

1. **Regression Tree:**

While neural networks can capture complex relationships, regression trees offer a clear, rule-based framework since they recursively partition the data into subsets, based on the values of input features, and assign a constant value (typically a mean or median) to each subset as the predicted output.

In this case, the transparency and interpretability guaranteed by this type of model do not translate to an improvement of the metrics.

???? At the same time, the results provided by this model are not extremely worse than those of a neural network, which offers a promising opportunity for further research into which VARIABLES ARE IN THE FIRST SPLITS, HENCE WHICH ARE THE MOST IMPORTANT TO DETERMINE THE INTEREST RATE. ???? VEDERE SE POSSO VEDERLO IO

* + 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.

1. **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.
2. **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.
3. **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.
4. **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.
5. **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.