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 four HIDDEN LAYERS WITH respectively 54, 36, 24, 10 nodes. 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 (tendency to underpredict about 1%)

1. **Regression Tree:**

While neural networks can capture complex relationships, regression trees offer a clear, rule-based framework, which improves model interpretability.

The regression tree algorithm starts by examining the entire dataset and identifying the feature that provides the best split, optimizing the chosen metric, in this case the mean squared error. The dataset is then divided into two subsets based on this feature's value, forming two child nodes. This process repeats recursively for each node, splitting them further until a stopping criterion is met, such as a maximum tree depth or a minimum number of samples in a node. In the end, the algorithm assigns a constant value, usually the mean or median of the target variable, to each leaf node, and predictions are obtained by following the sequence of decision splits until a leaf node is reached and returning the assigned constant value as the prediction. The primary aim of the regression tree algorithm is to create a hierarchical set of rules based on the features that best capture the relationships in the data, making it a versatile and interpretable tool for regression tasks. Similarly to what described above for the neural network, the model parameters are chosen using cross-validation to minimize the loss function MSE, and the final tree has a maximum depth of 11 nodes.

In this case, the transparency and interpretability guaranteed by this type of model do not lead to an improvement in the metrics compared to the neural network, except for a better value of the Mean Residual. This suggests that, overall, the Regression Tree does not exhibit a bias towards overprediction or underprediction.

Simultaneously, the results provided by this model are not significantly worse than those of a neural network, offering a promising opportunity for further research into which variables play a role in the initial splits, and thus, which variables are the most important in determining the output variable.

* **ok**

1. **Random Forest:**

Random Forest is an ensemble learning method that combines multiple regression trees to improve predictive power and reduce overfitting.

The random forest algorithm begins by generating a random bootstrap sample of size n, allowing for replacement. Subsequently, it constructs a decision tree using this sample. At each node of the tree, it randomly selects d features without replacement and divides the node using the feature that provides the best split, typically optimizing information gain. This process repeats k times, and the final outcome is determined by aggregating predictions from each tree using majority vote (Raschka and Mirjalili, 2019).

In this case, the cross-validated model has a maximum depth of 30 and uses 128 estimators, obtaining much better results than the single regression tree presented above.

Controllare I numeri e inserire il grafico

Risultati

Rifare grafico con csv salvato

1. **XGBoost (Extreme Gradient Boosting):**

XGBoost is a gradient boosting algorithm known for its high performance. It builds an ensemble of gradient boosting decision trees sequentially, optimizing for both predictive accuracy and computational efficiency (Chen and Guestrin, 2016). It has gained a lot of popularity as the winning solution to many machine learning challenges, and it is extensively used by practitioners to create state of art data science solutions (DMLC, 2019).

In this application, the XGBoost model with the cross-validated parameters provides a further improvement upon the Random Forest model described above, thus conferming its adaptability and predictive power.

1. **LightGBM:**

The last model considered is LightGBM, another gradient boosting decision tree framework that focuses on speed and efficiency for high dimensionality and and larger datasets. It uses a histogram-based learning method that bundles features together and excludes data points with small gradients to achieve faster training and maintain comparable accuracy (Ke et al., 2017).

As a matter of fact, in our application, LightGBM obtains the same performance as XGBoost, with a significant reduction in training time. XGBoost only exhibits a better metrics when it comes to MR, where the slightly higher value of the mean residuals of LightGBM indicates a less biased model with a tendency to overpredict.

1. **Hybrid Model:**

An ensemble of regressors is a collection of regressors whose individual predictions are aggregated in some manner, often through weighted or unweighted averaging or other combination methods, to predict outcomes for new data points. In the realm of supervised learning, there has been considerable ongoing research into devising effective techniques for building robust ensembles of regressors. A fundamental finding in this field is that ensembles frequently yield superior predictive performance compared to the individual regressors that constitute them.

A crucial requirement for an ensemble of regressors to outperform any of its individual members is that the constituent regressors exhibit both accuracy and diversity in their predictions. This means that each regressor in the ensemble should provide accurate predictions, and the regressors should also differ in some meaningful way, allowing them to capture diverse aspects of the underlying data relationships (Dietterich, 2000).

For these reasons, we chose to analyze two different ensemble models based on the last three models presented, which have the highest predictive power and different characteristics, aiming to obtain better results by combining their strengths.

The Hybrid Regression Model is an approach that averages the individual predictions of two or more models to generate a final prediction. This process aims to mitigate the trade-off between bias and variance in the individual models, resulting in enhanced accuracy within the final composite predictions. This method has been applied successfully in various settings, as exemplified by Lu et al. (2017) in the domain of house price predictions. In their study, they combined Lasso and XGBoost for a Kaggle challenge, achieving a top-ranking position within the upper 1% among all participants.

In our application, we applied different sets of weights to the three chosen models and obtained the best results by averaging 20% of the Random Forest, 40% of XGBoost and 40% of LightGBM predictions. The results achieved with this approach are significantly better than those of each model considered separately, with a reduction of MAE of more than 0.01 compared to the Random Forest and of 0.005 compared to XGBoost and LightGBM. In the context of interest rate prediction, such small errors can lead to significant differences in the installments to be paid, hence reaching better results, even though with a slight difference, can have a real impact on the borrwers.

1. **Stacking:**

Stacking, or stacked generalization, is a machine learning ensemble technique that combines multiple models by using the predictions of previous models as input to a second-stage model that learns how to best combine them (Wolpert, 1992). Similarly to the hybrid regression framework, this approach can improve overall predictive performance by leveraging the strengths of individual models and correcting for their weaknesses.

In this study, we employed the prevalent 2-level stacking framework to predict *InterestRate*. The initial stacking level comprised the utilization of Random Forest and LightGBM, while the subsequent stacking level involved the use of XGBoost.

We observed enhancements in predictive performance through stacking when compared to using individual models. Nevertheless, it's important to note that while stacking yielded improved results, these improvements did not surpass the performance of the Hybrid Regression model, although coming very close. The same pattern was observed in the study conducted by Truong et al., who explored very similar architectures for Hybrid Regression and Stacking (2020).

POI GENERICO METTERE QUESTO:

Many iterations of performance tuning were done to find the optimal solution of each model. Random Forest Re- gression, XGBoost, and LightGBM were intensively tuned by function GridSearchCV provided by scikit-learn [9] to achieve the results listed in Table 2. For Hybrid Regression and Stacking methods, performance tuning was not required since both methods were combinations of the best regressions. Instead, architecture implementation could be considered to further enhance the prediction.