Fundamental of Data Science – Final Project

Kaggle Competition House Prices: Advanced Regression Techniques

Team name: *Gandalf Sax* **Score:** *0.11166*

Members:

- Valerio Antonini *1611556*
- Maria Luisa Croci *1597797*
- Daniele Sanna *1257436*

Language: *Python*

Data Tyding:

1. Use the logarithm in the column *SalePrice;*
2. Look at correlations on *train.csv* data frame;
3. Merge the *test* and *train* in a new data frame called *Union_data* in order to clean quickly the data frames;
4. Remove outliers;
5. Drop columns with high % of *NA* values and with low correlation with *SalePrice* (or if exists multicollinearity);
6. Change type of *MoSold*, *YrSold*, *MSSubClass* and *OverallQual* as string;
7. Impute *NA* values replacing with *0* or the mode of the feature or with *None* for categorical variables;
8. Apply *get_dummies* function on *Union_data* in order to have numeric values for categorical features. The function creates new binary variables;
9. Split *Union_data* in *train* and *test*.

Feature Engineering:

1. Create new feature *TotalSf* as sum of *TotalBsmtSF* and *GrLivArea.*

Modelization – Brute force optimization of a regression equation:

1. Lasso: a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces;
2. Ridge: a regression model where the loss function is the linear least squares function and regularization is given by the l2-norm;
3. Lgbm: is a gradient boosting framework that uses tree based learning algorithm;
4. Elastic net: is a regularized regression method that linearly combines the L1 and L2 penalties of the lasso and ridge methods;
5. XGBoost: provides a gradient boosting framework , a technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees;
6. Stacking (models choosen for stacking: lasso, elastic net, ridge and xgboost), is a model ensembling technique used to combine information from multiple predictive models to generate a new model;
7. Final model combining stacking and other models using average weighted.