Report

Machine Learning project

Houses prices predictions

Paul Peyssard
paul.peyssard@etu.univ-cotedazur.fr
Msc Data Science & Artificial Intelligence

Supervised by :
Michele Rivell
Diane Lingrand

Introduction

In this report, I will explain how, in my machine learning project, I selected all the features needed in order to predict the price of residential homes in Ames, Iowa.
In order to predict the price, I used python and more particularly the sklearn library. I first cleaned the data by analyzing it, removing the outliers, the non needed values, the surplus or null values or the value with too many occurences or even the value with bad correlations to the target feature.

After the cleaning of the data, I created some models like decision trees, K-neighbors algorithm, random forest and SVM and I compared the RMSE metrics. The one that performed better was the SVM. Finally, I used the SVM to predict the price of houses and created a new result dataset with the ID of the houses and their predicted prices.

Feature selection

In order to explain which feature I selected, I think it more relevant to say why I deleted the others. In the beginning, I had 80 features.

I) Null values :

I wanted to check the features with the most null values in the train set and I saw that there was some feature with more than 50% null values and up to 99%. I deleted all the 5 features with the most null values, the one that contained more than 50%.
It was the features : FireplaceQu, Fence, Alley, MiscFeature, PoolQC.
And for the rest of the features (numerical) containing null values, I replaced them with the mean of the feature.

II) Outliers and same values :

I saw that some features had more than 90% of the same values and some had very big outliers, I decided to delete these one :
Street, Utilities, LandSlope, Condition2, RoofMatl, Heatin, CentralAir, Electricalj, LowQualFinSF, BsmHalfBath, KitchenAbvGr, Functional, GarageQual, GarageCond, PavedDriv, 3SsnPorch,'ScreenPorch, PoolArea, MiscVal.

III) Correlations :

In order to keep only the features needed, I decided to check the correlations between the features and the correlations between the features and SalePrice.
I deleted the features with a correlation coefficient (absolute value) between -0.2 and 0.2 with the SalePrice :
EnclosedPorch, OverallCond, MSSubClass, YrSold, BsmtFinSF2, MoSold, BedroomAbvGr.
And to delete the features with too much correlation together with more than 0.75 :
OverallQual, YearBuilt, GrLivArea.

IV) Train set and test set differences :

I noticed that there was some features in the test set that was not in the train set and vice versa, I decided to also delete those :
RoofStyle_Shed, Exterior1st_AsphShn, Exterior1st_CBlock, Exterior1st_ImStucc, Exterior1st_Stone, Exterior2nd_CBlock, BsmtCond_Po.

V) Incorrect unique values :

In some of the text features, the unique values were not the same in the train et test set. I decided to replace, in the test set, the unique value that we did not have in the train test by the unique value with the most occurences.

Conclusion

At the end of the data processing, I tried to analyze the meaning of the rest of the features but it did not seem relevant to delete anything with just a description because I am not an expert in real estate. After that, the data cleaning and the selection of features was done with 45 features left. The rest is the model selecting and processing that you can see in the notebook. I compared some models and used the SVM in order to predict the price of houses that you can see in the result.csv file.