|  |  |
| --- | --- |
| Data & Web Mining Lab Project |  |

Project organization

* Cleaning
* Feature Engineering
* Feature Transformations
* Encoding
* Scaling
* Target transformation
* Model Selection
* Hyperparameter Optimization
* Ensembling

# Cleaning

Regarding Data cleaning, we made sure to look out for any missing values, null values or anything that would disrupt the process.

We also separated the target feature from the dataset.

# Feature Engineering

For Feature Engineering we created 3 new features, on for the square foot per room,

another for the total number of bathrooms, giving different weights to half baths, and one feature for the total effective square foot of the house, focusing on the 1st floor, 2nd floor and the basement, giving different weight to the basement, since we don’t value it as highly as the normal floors.

We also change the month sold feature, this feature is interesting because by itself don’t represent nothing because the value being 1 or 6 does not give the total context between winter and summer, so we decided to use the cosine function, when on the winter months the value its closer to -1, and in the summer months its closer to 1, we think that this represents better this feature.

These new features do improve the models score, but its not significant.

Feature Transformation

We started by analyzing all the numerical features, to find any skewed features, on the features found we applied a logaritm function to make the features have a better distribution, which will help when training the models.

# Encoding

For encoding we used the function from pandas to apply the Hot One Encoding to our dataset, transforming the categorical features to usefull numerical features.

Scaling

We applied the StandardScaler to our dataset to center the values to 0 which tends to be better for the methods we are using.

Target Transformation

Since we transformed our skewed features on our dataset, we also checked our target to see if it was skewed, and it was. To change that we used a logarithmic transformation to have a better distributed data range on our target.

Model Selection

For our model selection we used the pycaret library, with this library we were able to insert our training data on a setup function, and with that function we can then test our dataset on various of models, and gives us a list of the best ones, and we picked out the best.

Hyperparameter Optimizations

Instead of testing our models manually with different parameters over and over, we used the optuna library, which allowed us to find the best parameters from each model for our dataset, thus giving our models the best parameters to test our dataset on.

Ensembling

For this project we focused mainly on ensembling predicting models, given that this a regression problem, and ensembling models being fast to run and computationally efficient, making it a very strong prediction method for us.

Final thoughts

During the realization of this project, we have learned immensely about the vast world on Data Web Mining, these are so many models, techniques, libraries that can change the whole workflow of our project, and each of them have different strengths and weaknesses. We felt that the longer we worked on this project the less we thoughts we knew, and we felt like we were just scratching the surface, and we got more and more interested in it the more we worked on it.

We definitely learned a lot on this journey on the amazing world of data.