Final Project
Predicting Housing Prices

Serikzhan Sadakbayev
Derrick

Description of the problem

Processing (cleaning) data

- Removing columns, where more than 2000 rows have the same value
- Removing columns, where number of unique values is less than 6
- Removing columns, where less than 200 rows have value
- Removing rows, where SalePrice has no value
- Adding dublicate columns with numerical values for columns with text values (this would help for machine learning functions)
- Replacing all Nan values with 0

Correlations and Heatmap

In order to make our data even cleaner we would like to proceed with correlation of SalePrice with other columns of data frame. After we got correlation results, we would like to get rid of columns, which have correlation less than 30% on positive side and more than -40% on the negative side. Heatmap is a way to visualize statistical data about a site using a color palette. This analysis is one of the techniques used to determine which features affect the target variable the most, and in turn is used in predicting that target variable. In other words, it is a widely used feature selection method in machine learning.

Dividing dataset to test and training

The data (dataset) is separated into training and test samples to measure the quality of model predictions. To keep some of the instances concealed from the algorithm, we chose an 80/20 ratio. The model gains knowledge from the first sample by comparing house parameters and pricing. Then we feed her the apartment characteristics from the test sample (but we do not show prices from the test sample, she will try to guess them herself). The next step is to determine the error, which is the difference between the expected and actual cost. Additionaly, since theoretically some of the machine learning models such as Random forests prefer uncorrelated of low correlated dataset, we created dataframe with columns, which are low correlated with SalePrice.

Methods and Techniques

1. Linear regression is a linear approach to modeling the relationship between a scalar response (or dependent variable) and one or more independent variables (or independent variables). This approach was chosen, because it's one of the best and simplest approaches to have a prediction model. The result of predictions for our linear regression model is 81.15%.
2. Before applying random forest we wanted to see how decision tree would work on our dataset. So we applied 2 of our datasets (correlated and uncorrelated). For now it's the only model, which has higher predictions results for uncorrelated dataset, but comparatively low to others. Uncorrelated one has score from 1 to 2%, when correlated one has 0-1% result. As we can see decision tree for our type of data set. So in order to get better results with this model, we need to change dataset in order to avoid overfitting, high variance and other noise data.
3. The fundamental concept behind the random forest is simple but powerful - it is the wisdom of the crowd. The reason the random forest model works so well for predictions is because many relatively uncorrelated trees working together will outnumber any of their individual constituents. Random forests train each tree independently, using a random sample of the data. This randomness helps to make the model more robust than a single decision tree, and less likely to overfit on the training data The result of predictions for Random Forest is 89% using correlated dataframe, where columns have high correlation with SalePrice. At the same time the prediction score of random forest for the dataset with columns with lower correlations is 80%. This is quite interesting fact for us.
4. Gradient Boosting Regression is a machine learning technique, which is helpful in regression and classification of tasks. When a decision tree is the weak learner, the resulting algorithm is called gradient-boosted trees; it usually outperforms random forest. The result of predictions for Gradient Boosting Regression is 88.51% using our correlated dataframe. Prediction score with uncorrelated dataframe is 5% less and is 83%.

Conclusion

In conclusion, we processed our dataframe with different methods, visualized some parts of dataframe, which we find interesting, found correlation between SalePrice and other columns, plotted heatmap based on our correlation data, splitted our data into 2 dataframes: with higher correlation with SalePrice and low correlation. Also we splitted those dataframes into training ones and testing ones. Then we applied several machine learning techniques on both datasets we had such as: Linear Regression, Decision Tree, Random Forest, Gradient Boosting Regression. As for the results the best and highest accuracy score has Random Forest with high correlated dataframe. So, finally, we successfully developed and assessed different types of machine learning techniques and models in Python, selecting the best model for our supplied dataset after going through a series of steps. But this isn't the end of it. Every model we created has its own set of statistical and mathematical principles behind it.