**House Prices - Advanced Regression Techniques**

To buy or not to buy, a questioned asked by many prospective house buyers. In this section we explore how machine learning techniques can help one in making a decision. We specifically study the Ames housing dataset available in Kaggle. This dataset set consists of 80 features that describe various characteristics of a house and thus are exploited by machine learning techniques to predict the house selling price.

*Feature types*

The table below displays the various type of features available in the dataset.

|  |  |  |  |
| --- | --- | --- | --- |
| Variables | Number of | Purpose | Examples |
| Continuous | 20 | Various area dimensions | BsmtFinSF1, 1stFlrSF, 2ndFlrSF |
| Discrete | 14 | Number of occurences | Bedroom, Kitchen, TotRmsAbvGrd |
| Categorical - Nominal | 23 | Identify various types of items | Neighourhood, HouseStyle, BldgType |
| Categorical - Ordinal | 23 | Rate various items within property | OverallQual, PoolQC, Fence |

*Exploratory data analysis*

We explore the top features as predicted by SelectKBest that appear to contribute the most in making the predictions about the sale price. SelectKBest supports two scoring functions for regression: ‘f\_regression’ and ‘mutual\_info\_regression’. We use both these functions to get the top 10 most important features as shown below:

|  |  |  |  |
| --- | --- | --- | --- |
| F\_regression | | Mutual\_Info\_Regression | |
| Features | Score | Features | Score |
| OverallQual | 2271.07804 | *Neighborhood* | 0.56149 |
| YearBuilt | 1365.01095 | OverallQual | 0.508816 |
| ExterQual | 1088.56943 | YearBuilt | 0.493831 |
| BsmtQual | 906.479922 | ExterQual | 0.370663 |
| TotalBsmtSF | 858.722738 | BsmtQual | 0.36877 |
| *1stFlrSF* | 826.363181 | TotalBsmtSF | 0.363677 |
| GrLivArea | 771.646539 | GrLivArea | 0.363104 |
| KitchenQual | 726.321286 | KitchenQual | 0.325255 |
| GarageCars | 713.963853 | GarageCars | 0.322886 |
| GarageArea | 668.406649 | GarageArea | 0.313251 |

Interesting both the functions select the same 9 out of 10 features though their ranking is slightly different. ‘*Neighborhood*’ is the highest scoring feature with scoring function ‘mutual\_info\_regression’ but is not present in the list of the other scoring function.

*Data cleaning*

Both the training and test dataset consisted features with relatively large number of NA’s, but on closer inspection of most of the categorical variables this turned out to be valid values. In such cases the NA’s were replaced by ‘None’ value. For the other variables these were replaced by either the mean or the mode, depending on the feature.

*Outliers*

The dataset contains at least 19 features in which data lies at more than 3 standard deviations away. Due to the relatively small size of the dataset, removing records with outliers could potentially lead to serious loss of data and lead to weaker machine learning model. Hence, there was no outliers removed in this study.

*Data normalization*

For normalization of data, I employed the usage of ‘PowerTransformer’, this was preferred choice due to its superior ability of handling of outliers present in the data. I decided to keep the default transformation method of ‘Yeo-Johnson’.

*Feature selection*

Feature selection in this exercise was performed by various combinations of SelectKBest and PCA on ElasticNet model.

*Machine learning models*

Following were the machine learning models used for training:

* ElasticNet
* XGBoost
* Neural network
* SVR
* SGD

*GitHub repo*

https://github.com/kalyanramm/HousePricesAdvancedRegressionTechniques.git