**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.

*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. Following are the features:

* YearBuilt - Original construction date
* TotalBsmtSF - Total square feet of basement area
* 1stFlrSF - First Floor square feet
* GrLivArea - Above grade (ground) living area square feet
* FullBath - Full bathrooms above grade
* TotRmsAbvGrd - Total rooms above grade (does not include bathrooms)
* GarageCars - Size of garage in car capacity
* GarageArea - Size of garage in square feet
* ExterQual - Evaluates the quality of the material on the exterior
* BsmtQual - Evaluates the height of the basement
* Neighborhood - Physical locations within Ames city limits

*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.

*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.

*Feature selection*

Feature selection in this exercise was performed by various combinations of SelectKBest and PCA and model performance was evaluated for each of the combination.

*Machine learning models*

Due to the non-linear nature of the dataset, following were the machine learning models used for training:

* XGBoost
* Neural networks

*Results -*

The following section displays the performance, i.e., Kaggle score, of the various machine learning models on feature set selected using SelectKBest and PCA.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Total no.of features | XGB | Neural Network |
| All | 331 | 0.14199 | 0.17159 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| SelectKBest |  | | | PCA (n\_components=0.5) | | | PCA (n\_components=0.90) | | |
| Total no.of features | XGB | Neural Network | Total no.of features | XGB | Neural Network | Total no.of features | XGB | Neural Network |
| K=50 | 50 | 0.15527 | 0.17962 | 68 | 0.14763 | 0.20869 | 132 | 0.14392 | 0.1882 |
| K=100 | 100 | 0.14503 | 0.20495 | 118 | 0.14506 | 0.15935 | 175 | 0.13956 | 0.25393 |
| K=200 | 200 | 0.14568 | 0.33725 | 209 | 0.14189 | 0.42776 | 245 | 0.14235 | 0.3308 |