**House Prices: Advanced Regression Techniques**

**Problem Statement**

How to give a price range to homebuyer, by predicting the price, based on the different features of a house, including no of bedrooms, size of the living area, etc.…

**Dataset Source**

<https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data>

**Dataset** **Details**

The data has 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, USA.

Records details

**Total No of Rows: 1460**

**Total No of Columns: 80**

**Data Description**

Sale Price: - The property's sale price in dollars. This is the target variable

The remaining data are related to other properties of a normal house like

No of bedrooms

No of garage

**Missing Data Analysis**

There are more than four properties are having more than 80% of Null values. Those features, has been removed, as Part of the Data wrangling.

**Initial findings**

* Total No of Quantitative Features: 37
* Total No of Qualitative Features: 39
* The Sale price distribution details are
  + Min Selling Price is: 34900
  + Average Selling Price is: 180921.20
  + Median Selling Price is: 163000.0
  + Max Selling Price is: 755000
  + Skewness: 1.882876
  + Kurtosis: 6.536282

**Sale Price Analysis**

* + It becomes hard to analyze, uni-variate fashion and commenting, on values of a variable which is continuous in nature and having too many values, becomes tough. So, here categorizing the SalePrice(s) to help us to understand the behavior of SalePrice in a much convenient way.[¶](http://localhost:8888/notebooks/SpringBoard/Capstone-01/Analysis-EDA.ipynb#It-becomes-hard-to-analyze,uni-variate-fashion-and-commenting,-on-values-of-a-variable-which-is-continuous-in-nature-and-having-too-many-values,-becomes-tough.-So,-here-categorising-the-SalePrice(s)-to-help-us-to-understand-the-behaviour-of-SalePrice-in-a-much-convinient-way.)
  + Breakpoints of 100000, 200000 and 300000 is considered to be low, mid and high range, which is also clearly seen from the above histogram plot of SalePrice.
  + We have divided the Sale Price in to Multiple Category to do some further analysis. The details are as below
    - SalePrice <= 100000 == 0
    - SalePrice > 100000 and SalePrice <=200000 == 1
    - SalePrice > 300000 == 2
  + The total no of data with the above category
    - 0 -> 435
    - 1 -> 910
    - 2 -> 115
  + Looks like we have more houses falling in the midrange as compared to low and high
    - Along with above categorization as low, medium and high price, we can also look for quantile-based segregation with 20 bins. We found out that the max no of houses are between the price of 135500 and 141000

**MSSubClass: The building class - Analysis**

* + (20,30,40) - 1 story  
    (45,50) - 1 1/2 story  
    (60,70) - 2 story
  + It looks like MSSubClass >= 70 have higher chances of having medium selling price within the range of 1l to 2l.

**YearRemodAdd – Analysis**

* + We have categorized < 1950 as 0 and every 10 years as a category. We found that
  + Only Houses that have been re-modelled recently have higher prices

### **Categorical data**

* + Some categories seem to more diverse with respect to SalePrice than others. Neighborhood has big impact on house prices.
  + Most expensive seems to be Partial SaleCondition. Having pool on property seems to improve price substantially. There are also differences in variabilities between category values.

**Machine Learning Model**

* + 'SalePrice' is right-skewed normal distribution. So, all features with similar distribution along with high correlation coefficient with 'SalePrice' are strong features for this exercise.
  + Strong features - '1stFlrSF', 'BsmtUnfSF', 'GarageArea', 'GrLivArea', 'LotFrontage', 'MoSold', 'OverallCond', 'OverallQual', 'TotalBsmtSF', 'TotRmsAbvGrd', 'YearRemodAdd'
  + On the other hand, a feature with counts in value significantly higher will not be a good feature, and can be safely dropped.
  + Additional features that can be dropped - '3SsnPorch', 'BsmtFinSF2', 'BsmtHalfBath', 'EnclosedPorch', 'KitchenAbvGr', 'LowQualFinSF', 'PoolArea', 'ScreenPorch'\*

We have tried Multiple Algorithms, as mentioned below, we found that Cross Validated Linear Regression is able to give less RMSE(**Best RMSE: 0.0995717579958**), as requested in Kaggle for the validity of the Model

|  |  |
| --- | --- |
| **Model** | **Best RMSE** |
|  |  |
| **Linear Regression** | **0.114** |
| **Ridge Regularization** | **0.114** |
| **Nearest Neighbors** | **0.199** |
| **Gaussian Process** | **0.359** |
| **Decision Tree** | **0.183** |
| **Neural Net** | **2.863** |
| **Random Forest** | **0.136** |
| **XGBoost** | **2.134** |
| **Gradient Boost** | **0.134** |
| **Extra Tree** | **0.139** |
| **AdaBoost** | **0.138** |
| **Bayesian Ridge** | **0.114** |
| **Bagging** | **0.135** |