**Capstone Project I Final Report: House Prices: Advanced Regression Techniques**

**Baskaran Viswanathan**

**05/15/2018**

**Abstract**

This project seeks to give a price range to home buyer, by predicting the price, based on the different features of a house, including no of bedrooms, size of the living area, etc.… The given datasets analyzed to understand, cleaned and feature selection is performed to select the most relevant features. Detailed exploratory methods are deployed to get idea of the relationship between features and sales price. Several learning methods have been tested to get the best result.

1. **Introduction**

This is the final report for the capstone project – 01 in Springboard. The dataset used for this project downloaded from Kaggle. Details of the Data set and Analysis are available in GitHub, which link has been given at the end of this document.

1. **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.…

1. **Data Description**

**Train.csv -> a training dataset with all features details and the saleprice**

**Test.csv -> a testing data set to test the predictive algorithm for its accuracy in predicting the price**

* SalePrice - the property's sale price in dollars. This is the target variable that you're trying to predict.
* MSSubClass: The building class
* MSZoning: The general zoning classification
* LotFrontage: Linear feet of street connected to property
* LotArea: Lot size in square feet
* Street: Type of road access
* Alley: Type of alley access
* LotShape: General shape of property
* LandContour: Flatness of the property
* Utilities: Type of utilities available
* LotConfig: Lot configuration
* LandSlope: Slope of property
* Neighborhood: Physical locations within Ames city limits
* Condition1: Proximity to main road or railroad
* Condition2: Proximity to main road or railroad (if a second is present)
* BldgType: Type of dwelling
* HouseStyle: Style of dwelling
* OverallQual: Overall material and finish quality
* OverallCond: Overall condition rating
* YearBuilt: Original construction date
* YearRemodAdd: Remodel date
* RoofStyle: Type of roof
* RoofMatl: Roof material
* Exterior1st: Exterior covering on house
* Exterior2nd: Exterior covering on house (if more than one material)
* MasVnrType: Masonry veneer type
* MasVnrArea: Masonry veneer area in square feet
* ExterQual: Exterior material quality
* ExterCond: Present condition of the material on the exterior
* Foundation: Type of foundation
* BsmtQual: Height of the basement
* BsmtCond: General condition of the basement
* BsmtExposure: Walkout or garden level basement walls
* BsmtFinType1: Quality of basement finished area
* BsmtFinSF1: Type 1 finished square feet
* BsmtFinType2: Quality of second finished area (if present)
* BsmtFinSF2: Type 2 finished square feet
* BsmtUnfSF: Unfinished square feet of basement area
* TotalBsmtSF: Total square feet of basement area
* Heating: Type of heating
* HeatingQC: Heating quality and condition
* CentralAir: Central air conditioning
* Electrical: Electrical system
* 1stFlrSF: First Floor square feet
* 2ndFlrSF: Second floor square feet
* LowQualFinSF: Low quality finished square feet (all floors)
* GrLivArea: Above grade (ground) living area square feet
* BsmtFullBath: Basement full bathrooms
* BsmtHalfBath: Basement half bathrooms
* FullBath: Full bathrooms above grade
* HalfBath: Half baths above grade
* Bedroom: Number of bedrooms above basement level
* Kitchen: Number of kitchens
* KitchenQual: Kitchen quality
* TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)
* Functional: Home functionality rating
* Fireplaces: Number of fireplaces
* FireplaceQu: Fireplace quality
* GarageType: Garage location
* GarageYrBlt: Year garage was built
* GarageFinish: Interior finish of the garage
* GarageCars: Size of garage in car capacity
* GarageArea: Size of garage in square feet
* GarageQual: Garage quality
* GarageCond: Garage condition
* PavedDrive: Paved driveway
* WoodDeckSF: Wood deck area in square feet
* OpenPorchSF: Open porch area in square feet
* EnclosedPorch: Enclosed porch area in square feet
* 3SsnPorch: Three season porch area in square feet
* ScreenPorch: Screen porch area in square feet
* PoolArea: Pool area in square feet
* PoolQC: Pool quality
* Fence: Fence quality
* MiscFeature: Miscellaneous feature not covered in other categories
* MiscVal: $Value of miscellaneous feature
* MoSold: Month Sold
* YrSold: Year Sold
* SaleType: Type of sale
* SaleCondition: Condition of sale

1. **Dataset Features Details**

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

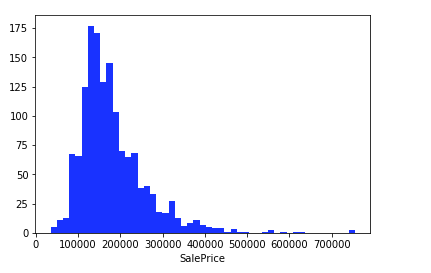
Records details

* Total No of Observations: 1460
* Total No of Features: 79
* Target variable – Sale price – 1
* Total No of Quantitative Features: 40
* Total No of Qualitative Features: 39

1. **Data Wrangling Details** 
   1. **Initial findings**

The Sale price distribution details are

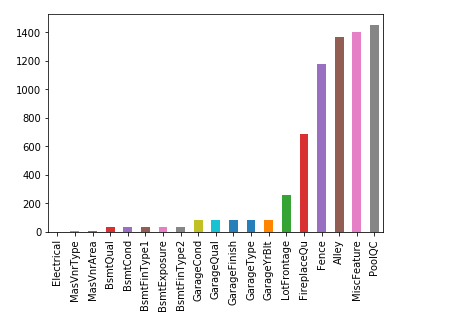
* + - Min Selling Price is: 34900
    - Average Selling Price is: 180921.20
    - Median Selling Price is: 163000.0
    - Max Selling Price is: 755000
    - Skewness: 1.882876
    - Kurtosis: 6.536282



* 1. **Missing Data Analysis**

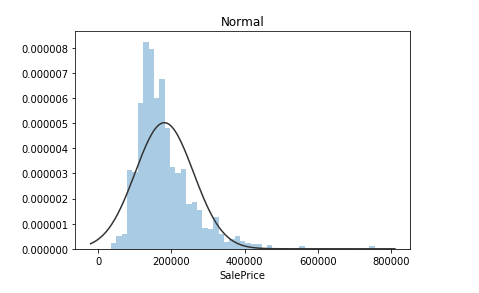
Some of the features, like Alley and PoolQC have no observations data, and due to many features to be analyzed and considered for the prediction, we forced to remove the features with 30% or less NaN values.

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.

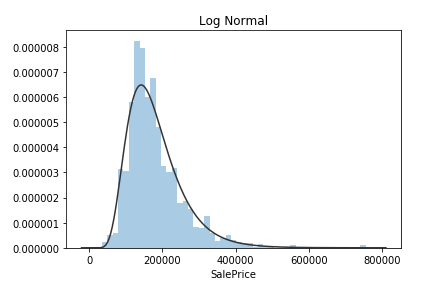


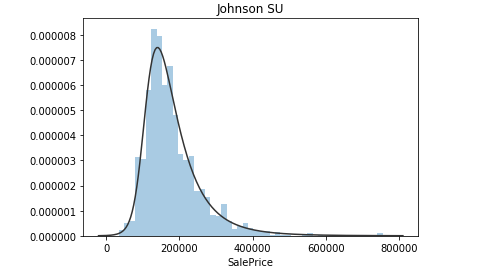
* 1. **Sale Price Analysis**

**Distribution Analysis**

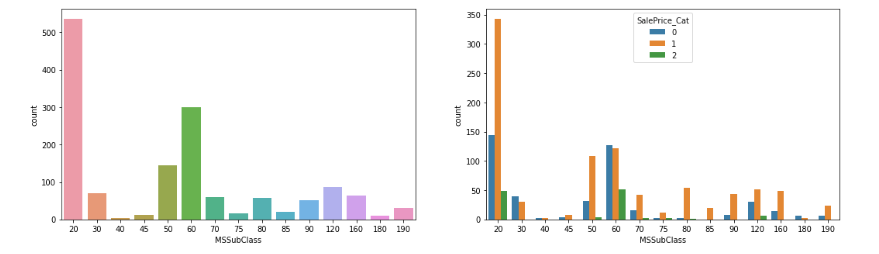


* + It is apparent that Sale Price doesn't follow normal distribution, so before performing regression it has to be transformed. While log transformation does pretty good job, best fit is unbounded Johnson distribution.

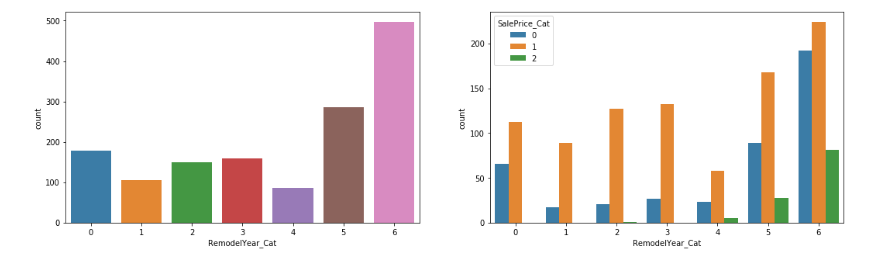




* + It becomes hard to analyze, uni-variate fashion and commenting, on values of a variable which is continuous in nature and having too many values, becomes tough. So, here categorizing the SalePrice(s) to help us to understand the behavior of SalePrice in a much convenient way.
  + Breakpoints of 100000, 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
  1. **MSSubClass: The building class – Analysis**
     + (20,30,40) - 1 story  
       (45,50) - 1 1/2 story  
       (60,70) - 2 story



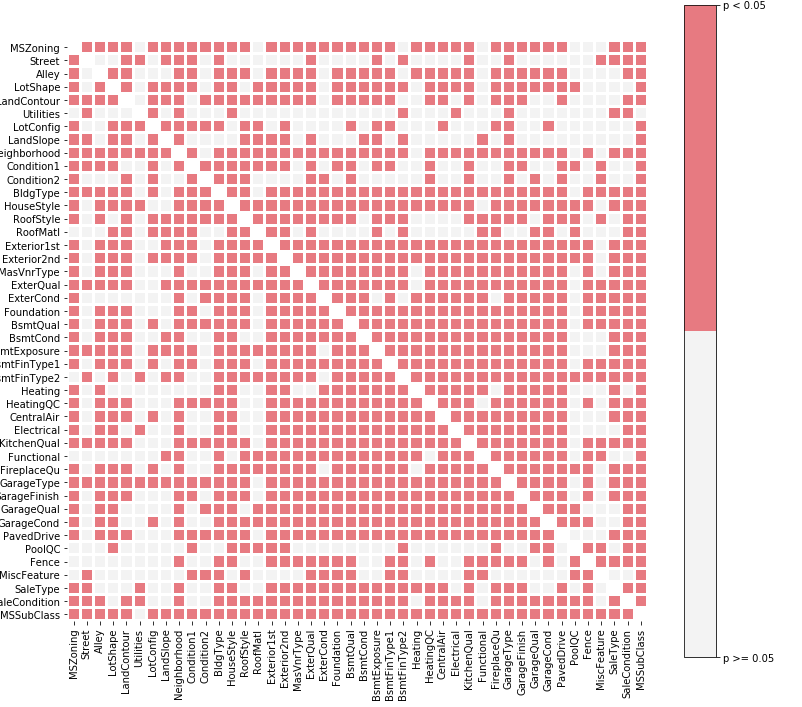
* + - It looks like MSSubClass >= 70 have higher chances of having medium selling price within the range of 1l to 2l.
  1. **YearRemodAdd – Analysis** 
     + We have categorized < 1950 as 0 and every 10 years as a category. We found that
     + Only Houses that have been re-modelled recently have higher prices



* 1. **Categorical data Analysis** 
     + 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.
  2. **Relationships Significant Analysis**

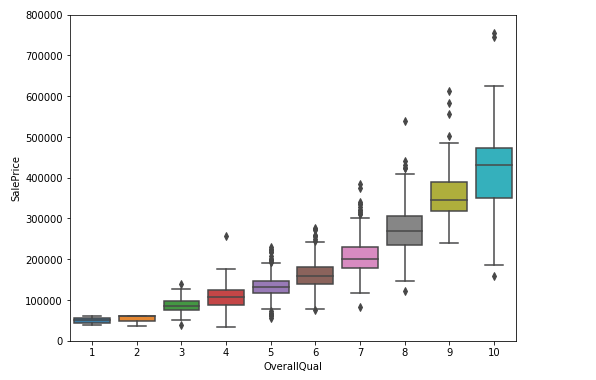
|  |  |
| --- | --- |
| **Feature Name** | **Percentage of Relationships Significant** |
| **MSSubClass** | 97.70% |
| **GarageType** | 93.00% |
| **Neighborhood** | 90.70% |
| **ExterQual** | 90.70% |
| **BldgType** | 90.70% |
| **Exterior2nd** | 90.70% |
| **BsmtQual** | 88.40% |
| **Exterior1st** | 88.40% |
| **KitchenQual** | 88.40% |
| **HouseStyle** | 88.40% |
| **BsmtFinType1** | 86.00% |
| **BsmtExposure** | 86.00% |
| **MSZoning** | 83.70% |
| **Foundation** | 83.70% |
| **BsmtFinType2** | 83.70% |
| **FireplaceQu** | 81.40% |
| **GarageCond** | 81.40% |
| **GarageFinish** | 81.40% |
| **PavedDrive** | 81.40% |
| **SaleCondition** | 81.40% |
| **HeatingQC** | 79.10% |
| **CentralAir** | 76.70% |
| **MasVnrType** | 76.70% |
| **GarageQual** | 76.70% |
| **BsmtCond** | 76.70% |
| **LotShape** | 74.40% |
| **RoofStyle** | 74.40% |
| **Electrical** | 72.10% |
| **Alley** | 69.80% |
| **LandContour** | 69.80% |
| **ExterCond** | 69.80% |
| **Functional** | 67.40% |
| **SaleType** | 65.10% |
| **Condition1** | 60.50% |
| **Heating** | 58.10% |
| **Fence** | 55.80% |
| **LotConfig** | 46.50% |
| **RoofMatl** | 46.50% |
| **LandSlope** | 39.50% |
| **Condition2** | 39.50% |
| **MiscFeature** | 37.20% |
| **Street** | 37.20% |
| **PoolQC** | 34.90% |
| **Utilities** | 20.90% |

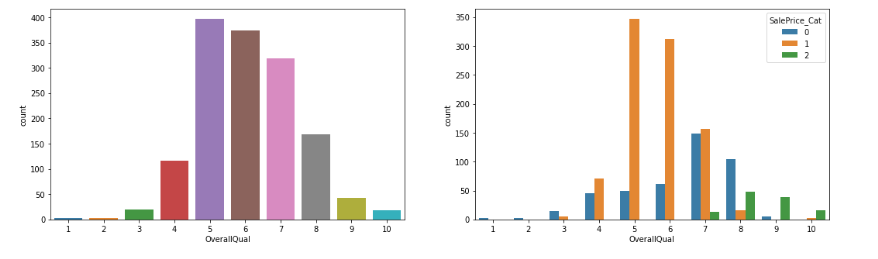
* 1. **Heat map of Statistically Significant vs Non-significant Categorical x Categorical Relationships**



* 1. **Overall qualification Analysis**

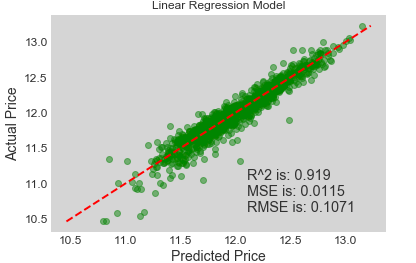
The Overall Qualification is one of main features in the Sale price



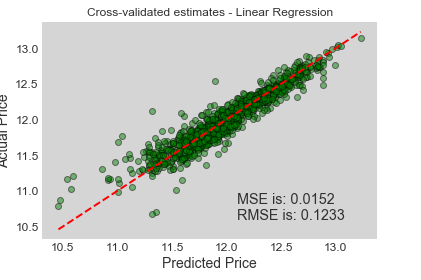


1. **Machine Learning Model**

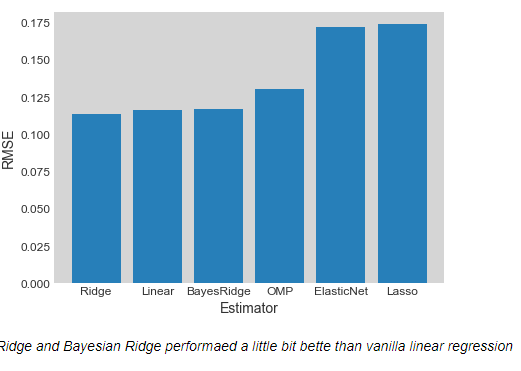
* 'SalePrice' is right-skewed normal distribution. So, all features with similar distribution along with high correlation coefficient with 'SalePrice' are strong features for this exercise.
* Strong features - '1stFlrSF', 'BsmtUnfSF', 'GarageArea', 'GrLivArea', 'LotFrontage', 'MoSold', 'OverallCond', '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'\*
  1. **Linear Regression**



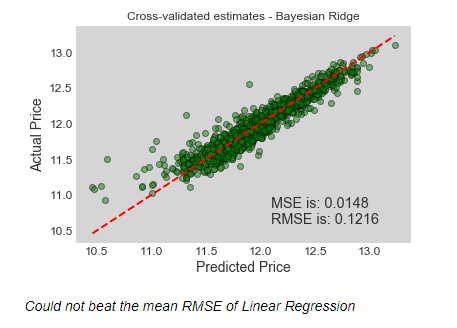
* 1. **Cross Validation - Linear Regression**



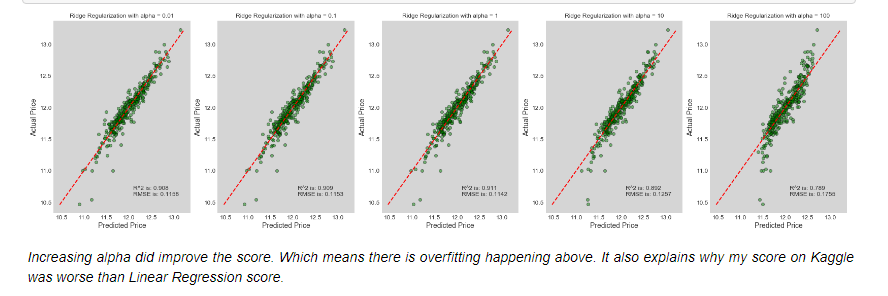
* 1. **Explore other models and learn other aspects of Machine Learning regression models.**



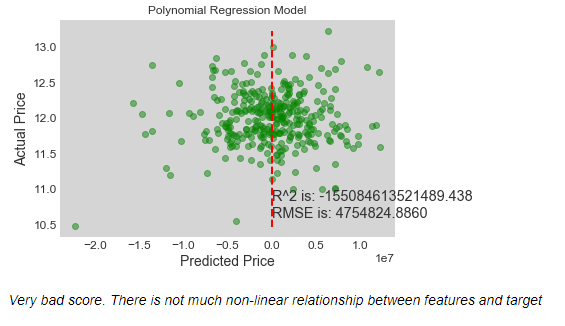
* 1. **Cross Validation - Linear Regression (Bayesian Ridge)**



* 1. **Ridge Regression or Tikhonov regularization**



* 1. **Polynomial Regression**



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** |

**Dataset Source**

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