House Price Prediction

Submitted by:
Shyamli Rao(153050009)
Amit Khandelwal(153050012)
Achala Bhati(153050056)

Outline

- Objective
- Problem Description
- Data Description
- Literature Survey
- Steps Used for prediction
- Understanding the data
- Data Cleansing and preprocessing
- Feature Engineering
- Machine Learning Techniques Applied
- Results
- Inferences

Objective:

- Predicting the sale price of a house using machine learning advanced regression techniques.

Problem Description:

- It is often necessary to accurately predict the price of a house between sales. One method of predicting house values is to use data on the characteristics of the area's housing stock.
- However, how to select the most appropriate the training parameter value is the important problem before applying regression technique. Also data set should be clear i.e. It should not have missing values and outliers
- Then regression techniques can be applied to predict the values.

Data Description:

- We have following files containing the data:
    - Train.csv : this file contain 1460 instance of data. Each instance have value of 79 features. This data file is used to train our model using machine learning techniques. Last feature 'SalePrice' have the output value y.
    - Test.csv : this file contain 1459 rows and 78 feature. The data model generated by the train.csv is used to predict the output value of test.csv data.
    -
- Data type :
    - Numerical : have real continuous values
    - Categorical : have text or label values

We have to convert categorical values to the dummy variable before using them in regression techniques.

Literature Survey

- Dubin, Robin A. "Predicting house prices using multiple listings data." *The Journal of Real Estate Finance and Economics* 17.1 (1998): 35-59.
    - Predict house sale price using data on the characteristics of the area's house stock to estimate hedonic regression(**hedonic regression** or **hedonic demand theory** is a revealed preference method of estimating demand or value) using ordinary least squares (OLS) as the statistical technique.
- Gu, Jirong, Mingcang Zhu, and Liuguangyan Jiang. "Housing price forecasting based on genetic algorithm and support vector machine." *Expert Systems with Applications* 38.4 (2011): 3383-3386.
    - In this study, a hybrid of genetic algorithm and support vector machines (G-SVM) approach is presented in housing price forecasting. Support vector machine (SVM) has been proven to be a robust and competent algorithm for both classification and regression in many applications.
    - Compared to Grid algorithm, genetic algorithm (GA) method consumes less time and performs well.Thus, GA is applied to optimize the parameters of SVM simultaneously.

Literature Survey

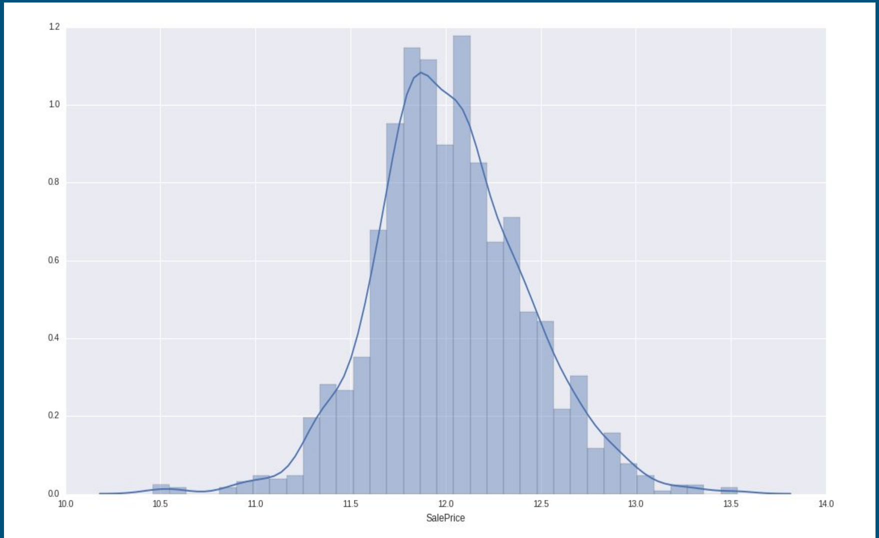
- Mu, Jingyi, Fang Wu, and Aihua Zhang. "Housing value forecasting based on machine learning methods." *Abstract and Applied Analysis*. Vol. 2014. Hindawi Publishing Corporation, 2014.
  - In this paper support vector machine (SVM), least squares support vector machine (LSSVM), and partial least squares (PLS) methods are used to forecast the home values. They have also compared algorithms according to the predicted results.
  - Experiment done in this paper shows that although the data set exists serious nonlinearity, the result also show SVM and LSSVM methods are superior to PLS on dealing with the problem of nonlinearity.
  - The global optimal solution can be found and best forecasting effect can be achieved by SVM because of solving a quadratic programming problem.
- Wang, Xibin, et al. "Real estate price forecasting based on SVM optimized by PSO." *Optik-International Journal for Light and Electron Optics* 125.3 (2014): 1439-1443.

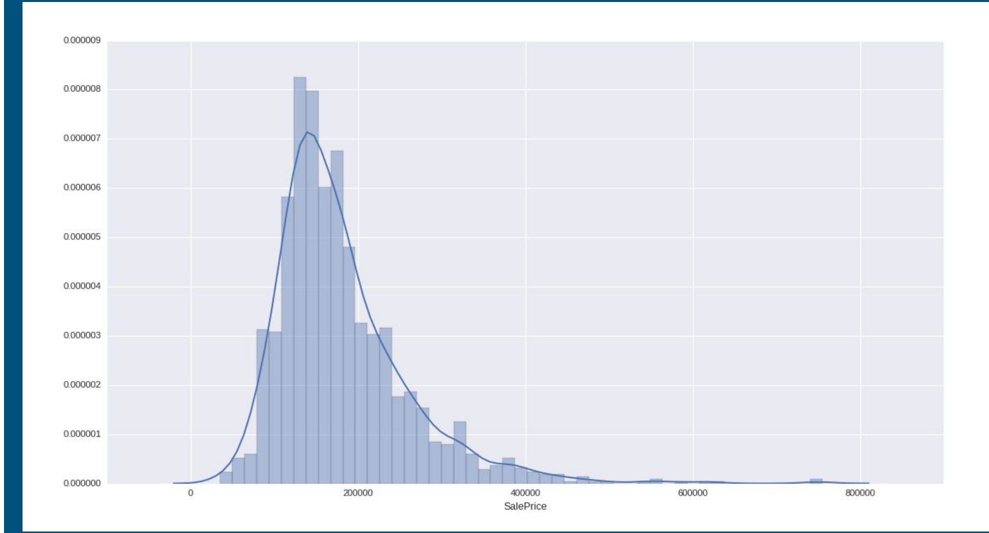
Steps Used for Prediction

- Understanding the data
- Data Cleansing
- Data Preprocessing
- Feature Engineering
- Advanced Regression Techniques
- Output Prediction

Understanding the data:

- First we studied about the features given in the test.csv file and then try to understand their relation with the 'SalePrice' i.e. our required output values.
- We plot the distribution of the 'SalePrice' Value. Min : 34900 Max: 755000

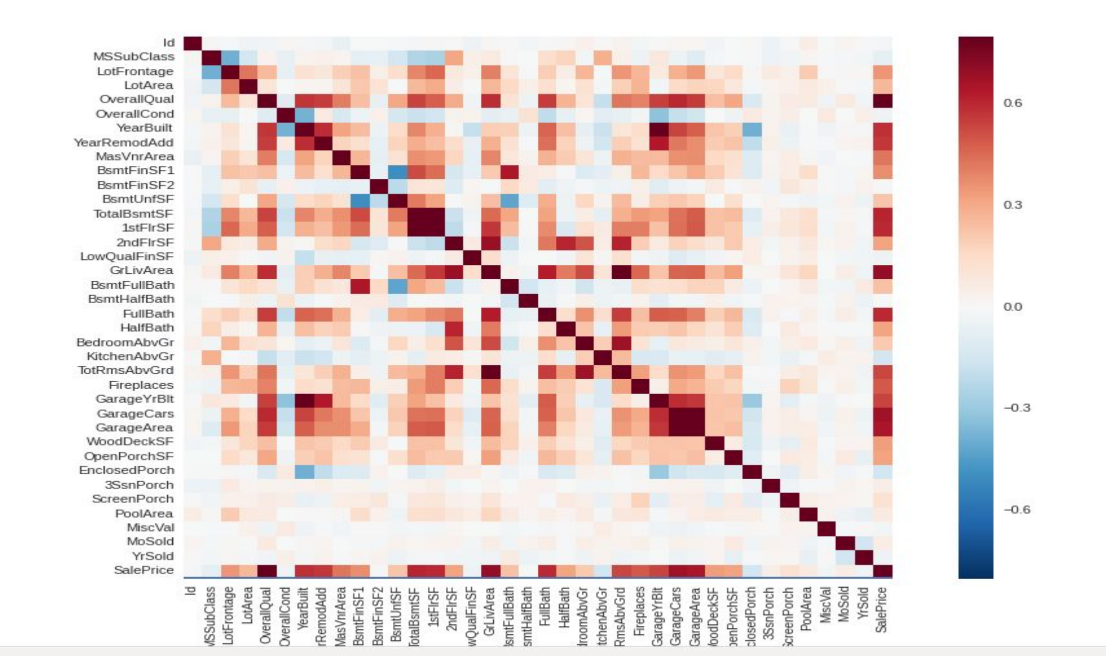




Data Cleansing and preprocessing:

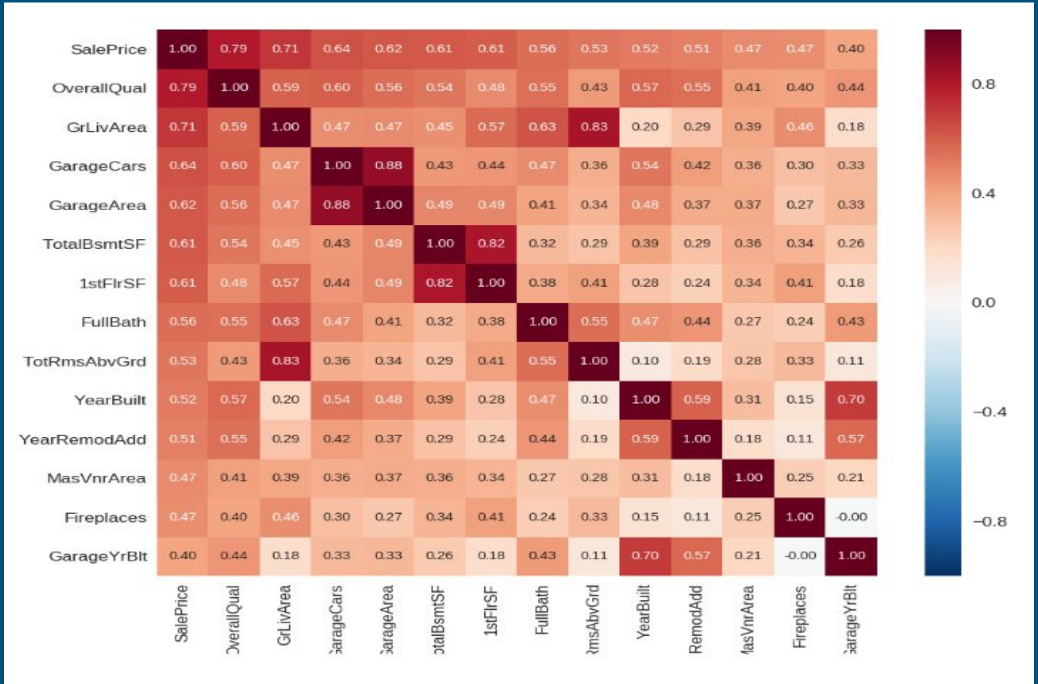
- In the train.csv and test.csv we have lots of missing values. In order to make it suitable for applying further machine techniques, we replaced this missing values with mean value
- Data files contain both types of the data i.e. categorical and numerical data. We can not apply regression techniques directly on the categorical data. So we label the categorical data with integer values.

Feature Engineering:

- We have selected 13 features that have considerable effect on the output 'SalePrice' value. In order to find these features we have used co-relation matrix between the features.
- 

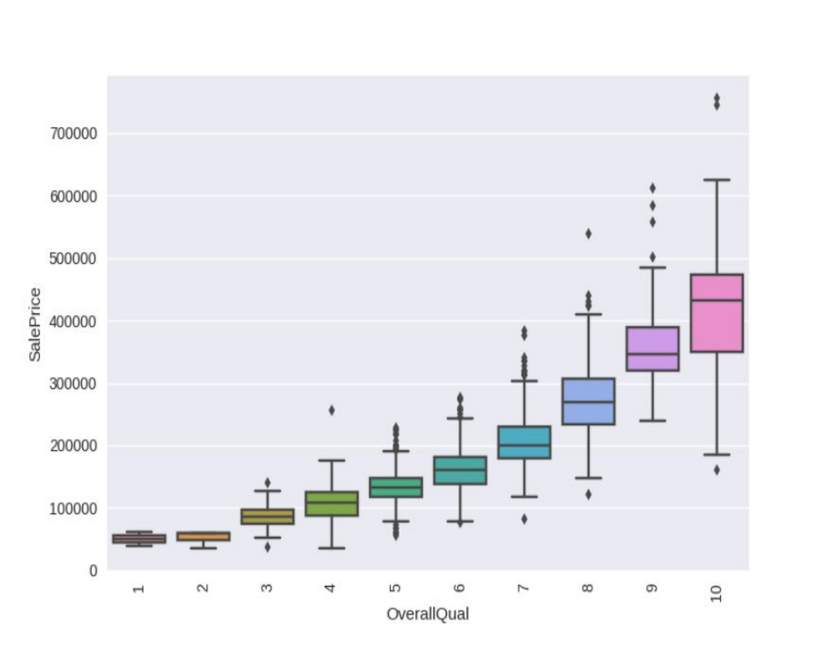
Feature Engineering

- Again we find the correlation among the 13 features and remove one of correlated pairs. From this we selected 9 Final Features :
- OverallQual, YearBuild,FullBath,Fireplaces,YearRemodAdd,MasVnrArea, TotalBsmtSF, GrLivArea, GarageArea



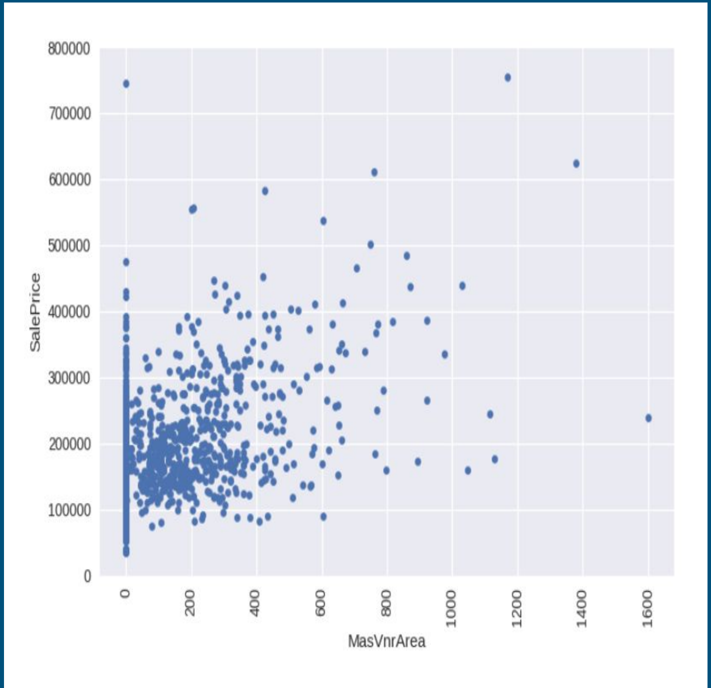
Feature Engineering:

- OverallQual:Sales price is quadratically increasing as the
- overall Quality is increasing.

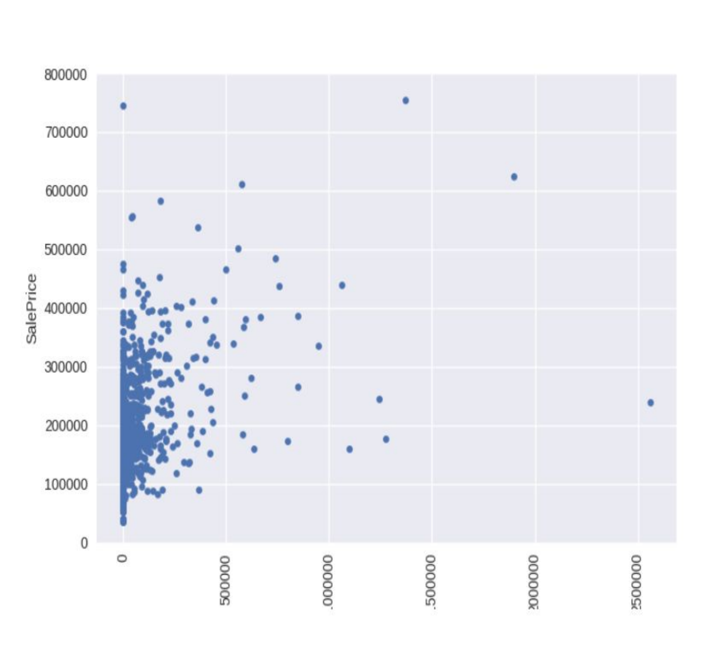


Feature Engineering:

- MasVnrArea:Data is unevenly distributed. After applying quadratic function it seems concentrated and linear.



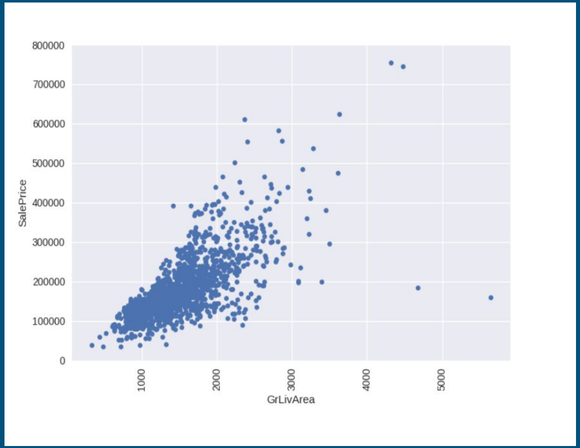
MasVnrArea



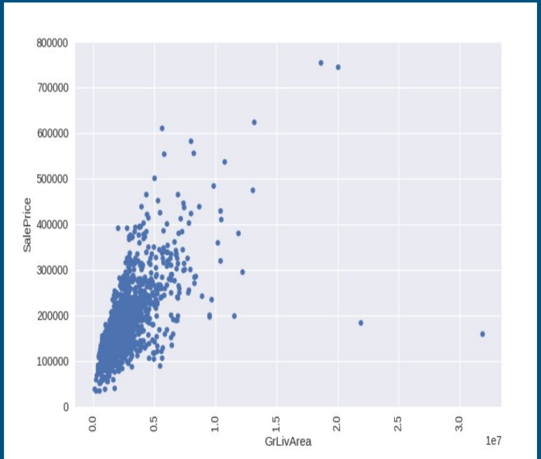
quad(MasVnrArea)

Feature Engineering:

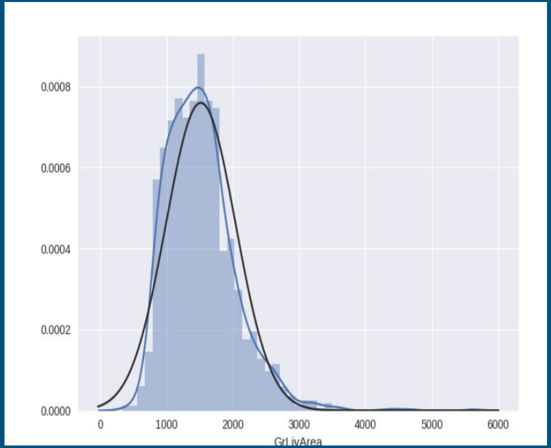
- GrLivAread :Data is unevenly distributed. After applying quadratic function it seems concentrated and linear.Also it is positive skewed., hence applied log before hand.



MasVnrArea



quad(MasVnrArea)



distribution

Machine Learning Techniques Applied:

1. Ridge Regression
2. Kernelized Ridge Regression
3. Lasso Regression
4. Feed Forward Neural Network
5. Support Vector Machines
6. Partial Least Square
7. Stochastic gradient decent
8. Bayesian regression
9. Boosting
10. Applied cross validation to all the above machine learning models.

Results:

Without Feature Engineering:

Model	Error (Without Cross Validation)	Error (With Cross Validation)
Ridge Regression	0.93136	0.55967
Kernelized Ridge Regression	0.83350	0.43901
Lasso Regression	0.94388	0.57851
Feed Forward Neural Network	0.21314	0.21956
Support Vector Machines	0.41890	0.41881
Partial Least Square	1.79965	NA
Stochastic gradient decent	NA	NA
Bayesian regression	0.15672	0.17316
Elastic Net	0.15298	0.55967
Boosting	0.12087	NA

Results:

With Feature Engineering:

Model	Error (Without Cross Validation)	Error (With Cross Validation)
Ridge Regression	0.21572	0.20987
Kernelized Ridge Regression	0.20620	0.43901
Lasso Regression	0.21572	0.20987
Feed Forward Neural Network	0.30532	0.33858
Support Vector Machines	0.41890	0.41881
Partial Least Square	0.35663	0.35726
Stochastic gradient decent	NA	NA
Bayesian regression	0.21633	0.30827
Elastic Net	0.21636	0.20859

Inferences

- Boosting is giving best result without applying feature selection as it uses random forest that already implements the feature selection techniques.
- The Best Result we are getting is 0.12087. Our results are among top 25% on kaggleout of 2280 participants.
- Both Ridge and Lasso results gets better when we applied Cross Validation Technique on them as shown in above table. Here We are using K-fold Cross Validation.
- Both Ridge and Lasso results gets better as we applied feature engineering on the data set. Error reduced by 76.83% in case of ridge and 74.11% in case of Lasso .
- There is no effect of Feature Engineering on Neural network. As it does so internally.
- But the result is better in case when no feature engineering is applied.SVM model 's result is same in all cases. Also its result are better than Ridge and
- Lasso when there is no feature engineering is applied.
- Bayesian give best result when no feature engineering and cross validation is applied.

Thank you