# **MID TERM PROJECT REPORT**

# **Predict Sales Price for Ames Housing Dataset**

Navya Sai Paladugu (nxp153030)

Raghunadh Metta (rxm153530)

**Purpose:** The main aim of this project is to build a model that will predict the possible sales price of a house based upon the other predictors. The Cross-Validation techniques should be used to test the models.

**Dataset(s):**  We used the following data set for this assignment:

1. **Ames Housing Dataset:** This dataset was provided by professor on eLearning.

**Approach:**

## **Data Cleaning and Preprocessing:**

Below is the procedure followed for data cleaning and pre-processing of data.

1. We converted the char variables to factors using the myData=as.data.frame(unclass(AmesHousing)) trick.
2. We removed Order and PID right away
3. Instead of removing unwanted predictors, we went through the summary and selected the predictors that seemed appropriate.
4. The predictors were selected based on the following criteria which:
   * Do not have larger numbers of NAs
   * Have a good variation (i.e., most of the values are not one category for factors, or are not mostly zero for numerical).
5. As recommended (both by professor and in the dataset description), we removed any houses with more than 4000 sq. ft in size, since these are outliers (there are only 5 of them).
6. After all this, and also based on instinct we selected the following 18 features: Lot.Frontage, Lot.Area, Neighborhood, House.Style, Overall.Qual, Overall.Cond, Year.Built, Year.Remod.Add, Foundation, Bsmt.Qual, Total.Bsmt.SF, X1st.Flr.SF, Gr.Liv.Area, Garage.Finish, Garage.Qual, Full.Bath, TotRms.AbvGrd, SalePrice.
7. After finalizing the features, we removed NA’s from the newly formed dataset.
8. We applied logarithm to the response variable i.e., SalePrice as the log would help make the SalePrice distribution more variant than that of the original. And this also helps in getting the MSE small.
9. Other Continuous factors such as Lot.Area, Total Basement Square Feet were scaled.

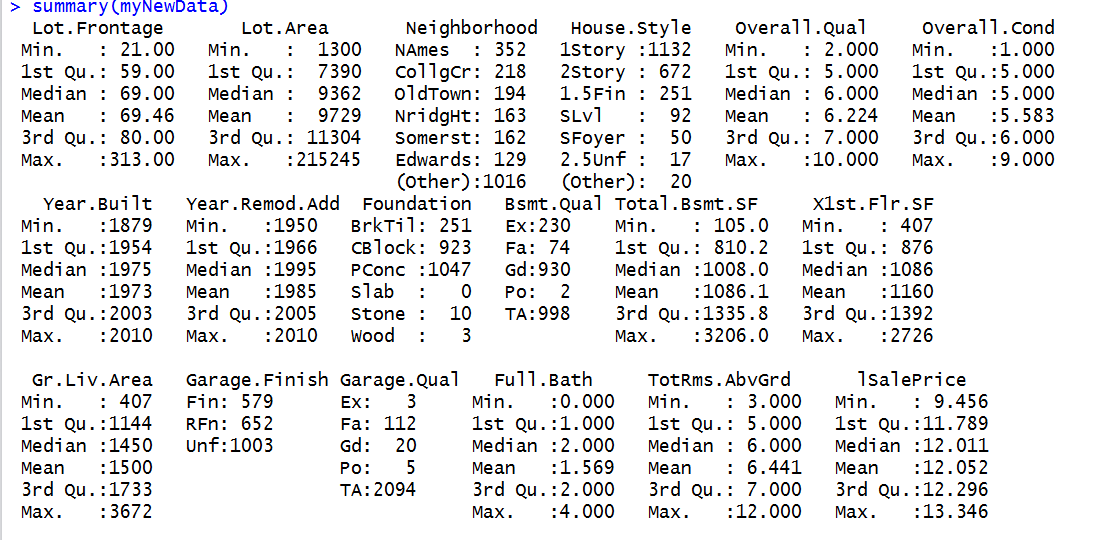


Figure-01: Summary of selected features of Dataset – New Dataset

1. The dataset consists of features in various formats. It has numerical data such as area, prices and numbers of bathrooms/bedrooms/living rooms, as well as categorical features such as Garage Quality, Neighborhood. To make this data with different format usable for our algorithms, categorical data was converted into separated indicator data, which expands the number of features in this dataset.
2. The final dataset is created using model.matrix() function and this final dataset has 61 features with 2234 rows.
3. Out of the 2234 observations in the final dataset, we considered 1787 as training data and remaining 447 as test data.

## **Modelling:**

We performed regression type of supervised learning algorithm on the finalized dataset.

### **Forward Subset Selection:**

We applied forward subset selection on the dataset and obtained the following graphs regarding RSS, AIC, BIC, Cp. And their predictions. We used Leaps package for this. Selecting the AIC Step-forward subset selection we cross-validated the log of Sales Price and the MSE was found to be 0.01528 (RMSE – 0.124)

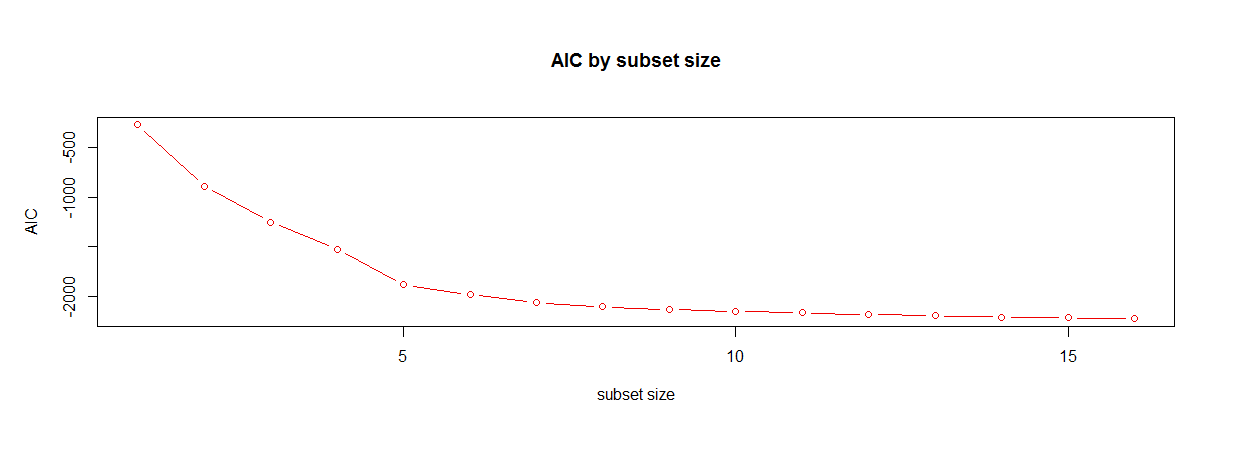
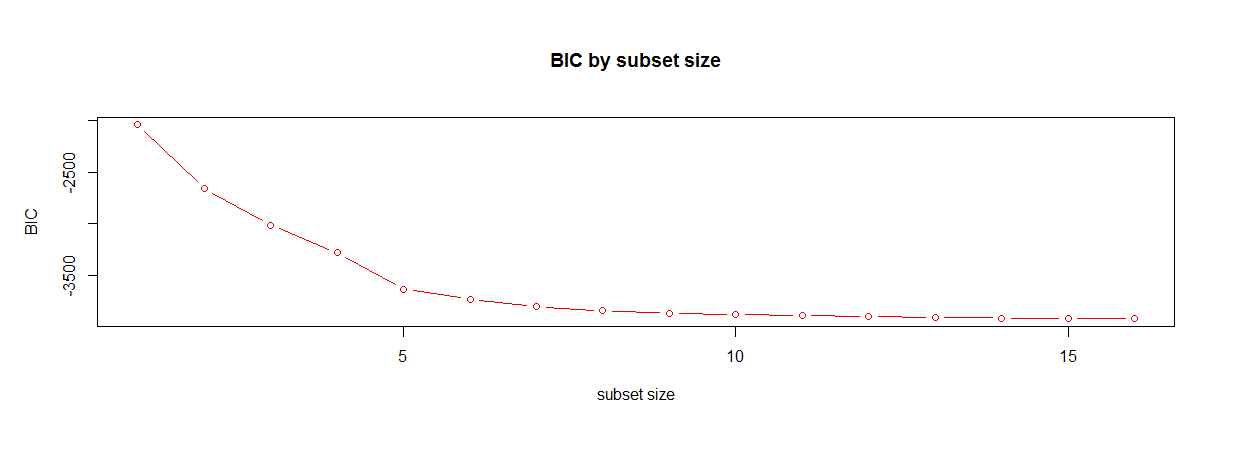
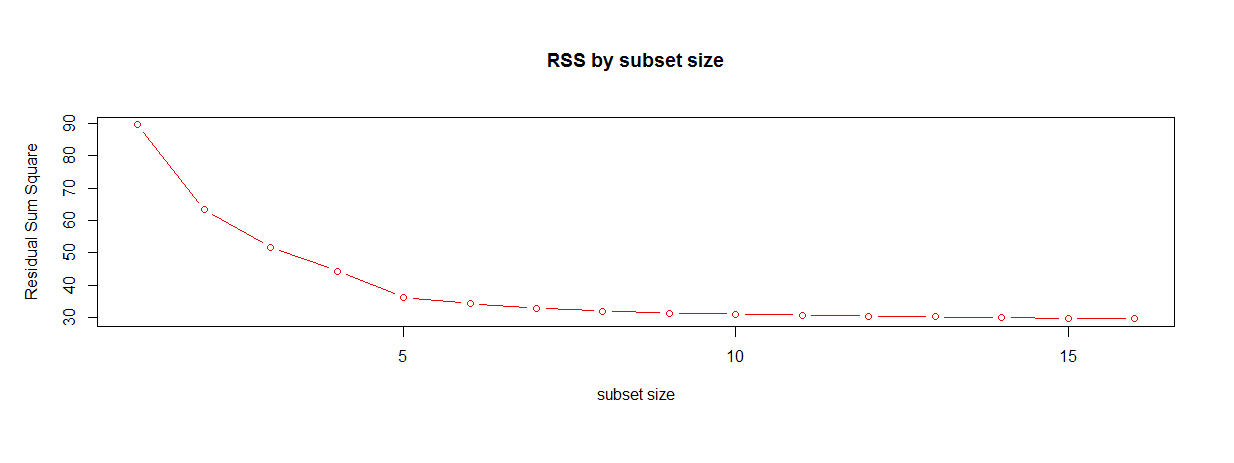
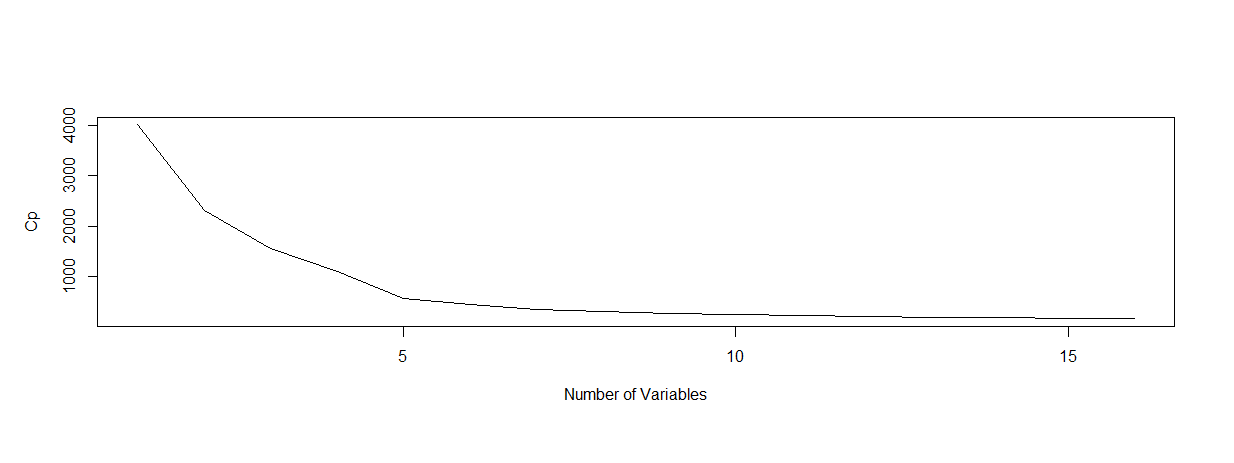


Figure-02: Cp, BIC, AIC and RSS models of Dataset

### **LASSO:**

We applied LASSO regression modelling on the dataset and obtained the following figures 03 and 04. The MSE obtained from cross-validating in this case is 0.0133 (with lambda.min) and 0.01519 (with lambda.lse) i.e RMSE 0.115 and 0.126 repectively.

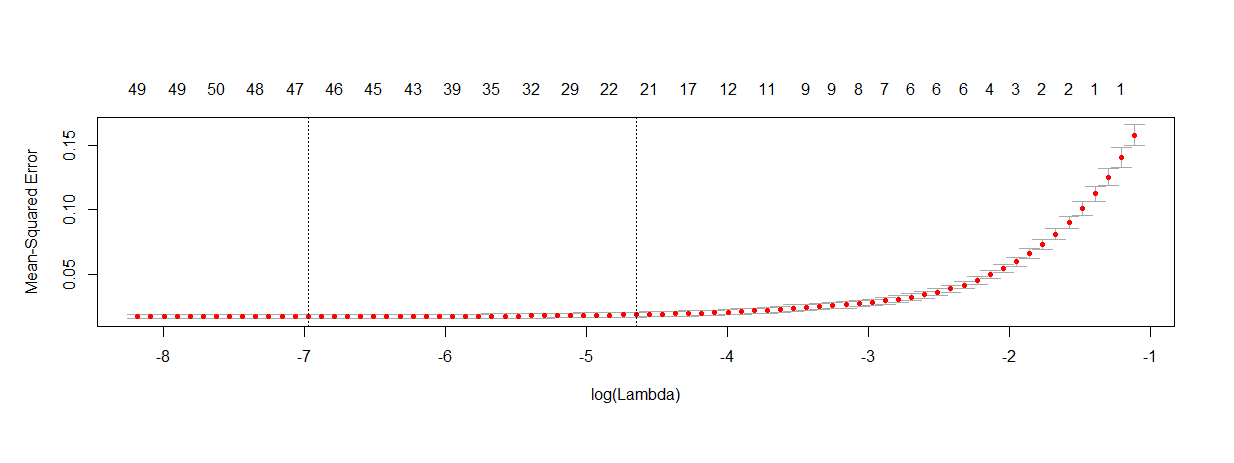


Figure-03: Mean-Squared Error versus Log(Lambda) for Lasso

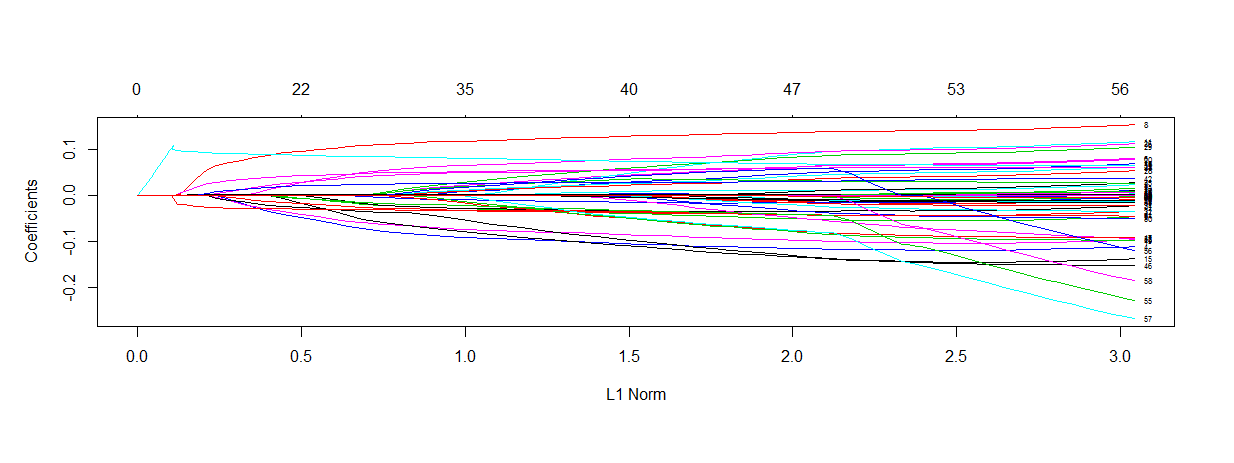


Figure-04: Coefficients Versus L1 Norm for Lasso

### **RIDGE:**

We applied Ridge regression modelling on the dataset and obtained the following figures 05 and 06. The MSE obtained from cross-validating in this case is 0.0137 (with ridge.min) and 0.01519 (with ridge.lse) i.e RMSE 0.117 and 0.126 repectively.

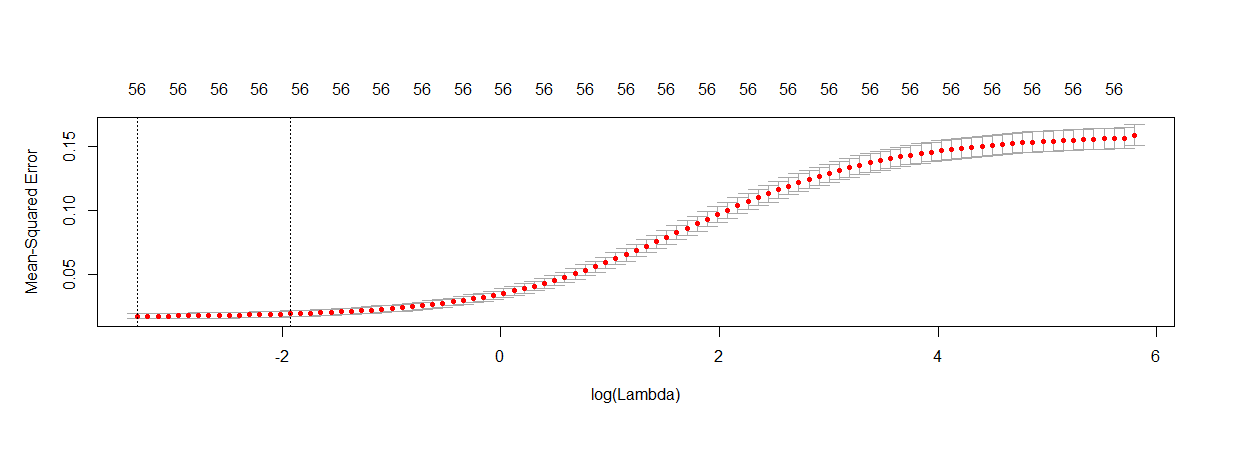


Figure -05: Mean-Squared Error versus Log(Lambda) for Ridge

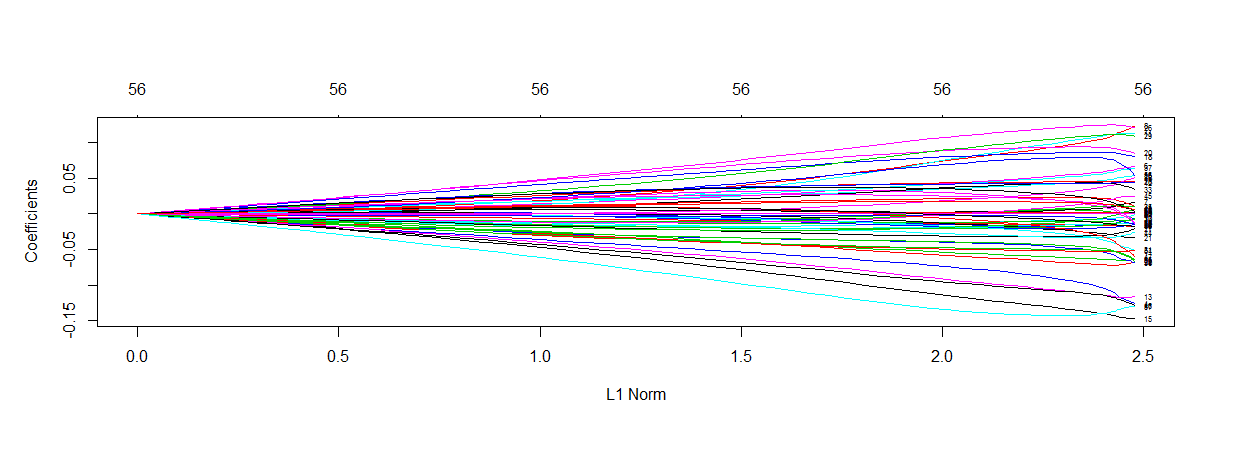


Figure -06: Coefficients Versus L1 Norm for Ridge

### **Linear Regression Model:**

We applied Linear regression modelling on the dataset. The MSE obtained from cross-validating in this case is 0.0154 i.e. RMSE 0.124.

### **PCR:**

We applied PCR modelling on the dataset (see Figure-07). The MSE obtained from cross validating in this case is 0.0888 i.e. RMSE 0.298.

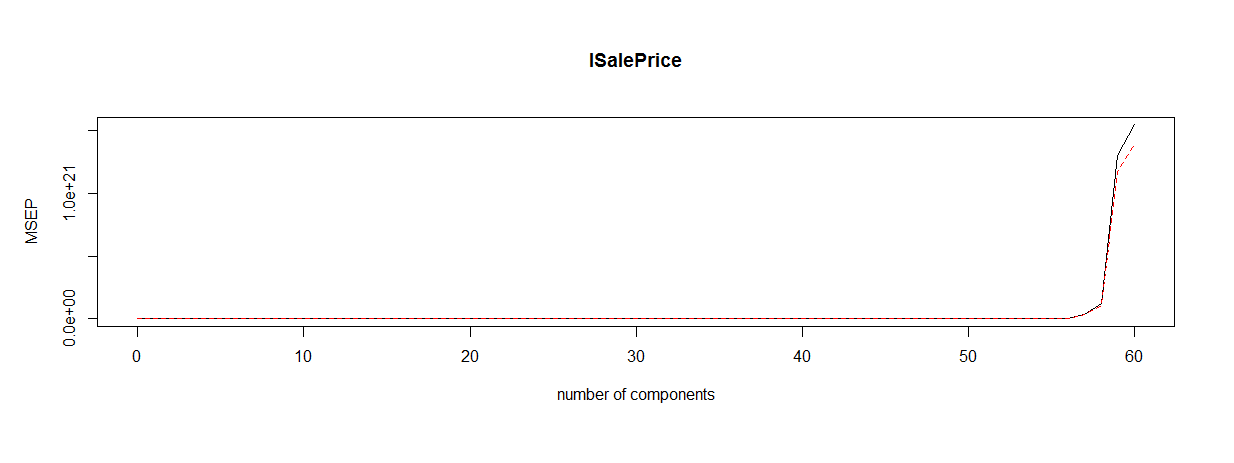


Figure -07: MSEP vs number of components PCR

### **KNN:**

We applied KNN modelling on the dataset. We ran the KNN for k=1 to 15 and by observing we determined that k=4 has lowest MSE which is 0.0307 i.e. RMSE 0.175.

## **Conclusion:**

By comparing the RMSEs obtained from different models examined for our finalized dataset, we can conclude that any of LASSO, Ridge, forward subset selection can be used for modelling the dataset.

*The observations made from this project are:*

* *We can use both classification and regression methods for predicting the housing prices. For classification, the prices should be divided into some number of buckets as to help cluster the data.*
* *Though the best approach for these kind of datasets is using a regression model.*
* *"Best subset" methods can be unstable with multiple regression, especially when there are a lot of variables. We might want to try a random forest approach in future.*
* *We might also want to try Neural Nets on the dataset and compare it with other models.*
* *In our project, we removed the NA’s. This might affect the overall prediction of the model. Hence, it is sometimes better to either assign ‘0’ or ‘mean’ value to the missing values. This helps in a better modelling of data.*
* *We observed that sometimes the features which does not have good variance (skewed features) may significantly affect the response variable. Hence, it is advised to consider the significance of each attribute by testing its p-value rather than by looking at a Histogram.*
* *In our project, we applied PCR for only the features we have considered. Usually, we can apply PCR for all the predictors. The condition that all the predictors must be numeric, when applied to our original dataset gives us around 288 predictors and this may increase the dimensionality of the problem.*
* *Ksmooth and Loess methods can also be applied to the dataset.*