CS6301.004 – R for Data Scientists
Mid Term Project (Housing Data)

Khushbu Patil – KXP153130

Vatsalkumar Patel – VRP140230

Purpose:

- The purpose of this project is to use techniques studied in the class to develop a regression model that can be used to predict a value of one attribute based on the values of other attributes.
- We have tried different approaches and presented our findings for each of these attributes.

Dataset(s):

- Data set used in this project is Housing data. The source for the data is Ames, Iowa Assessor's Office.
- This Data set contains information from the Ames Assessor's Office used in computing assessed values for individual residential properties sold in Ames, IA from 2006 to 2010.
- In the data set, tab characters are used to separate variables in the data file.
- The data has 82 columns and is a mixed data set which include 23 nominal, 23 ordinal, 14 discrete, and 20 continuous variables (and 2 additional observation identifiers).
- The purpose here is to build a regression model which can be used to predict the SalePrice attribute based on the value of other attributes.
- This dataset is raw for most part, thus it needs to be cleaned before it can be used for analysis.

Data Cleaning:

- First, we loaded the data in R using read.csv() function.
- That converted some columns (mainly the ones which are factors) to characters. That is not acceptable for the data so we used as.data.frame() function along with unclass() to convert those columns into factors.
- Even after this step, some of the attributes are represented as numeric even though they are factors. We converted them to factors explicitly.

- There are also some outliers in the data. They can lead towards biased and wrong analysis results if not removed before the analysis part.
- The dataset description clearly mentions that the outliers can be spotted by checking the value of attribute 'Gr.Liv.Area'. Any value above 4000 for this attribute suggests that this point is an outlier. Thus, we removed the data points that have value more than 4000 for this attribute.

- The first two attributes in the data set are Order and PID. Both attributes are just bookkeeping attributes and they serve no purpose towards the regression model. Thus, we removed both attributes from the data.

- The data set has 82 attributes. Using all these attributes to build regression model can result in very inefficient prediction model. Thus, we need to eliminate some attributes.
- There are several methods that we can use i.e. Subset Selection, Lasso and Ridge Regression, but performing these methods on full data is very inefficient and does not guarantee good results.
- Thus, we used data summary to select the attributes. Useful attribute with respect to regression model are attributes which have,
    - Numeric attributes which have higher variation
    - Numeric attributes for which most values are not just in one range
    - Factor attributes for which most of the values aren't just one category
- After analyzing the data summary, we selected attributes which doesn't have any of the above properties. We experimented with following two group of attributes:

Group -1	**Group-2**	
"MS.SubClass	MS.SubClass	Garage.Area
MS.Zoning	MS.Zoning	SalePrice
Lot.Frontage	Lot.Frontage	
Lot.Area	Lot.Area	
Lot.Shape	Lot.Shape	
Land.Contour	Land.Contour	
Lot.Config	Lot.Config	
Neighborhood	Neighborhood	
Bldg.Type	Bldg.Type	
House.Style	House.Style	
Overall.Qual	Overall.Qual	
Overall.Cond	Overall.Cond	
Roof.Style	Roof.Style	
BsmtFin.Type.1	Exterior.1st	
Bsmt.Unf.SF	Bsmt.Qual	
X1st.Flr.SF	Bsmt.Exposure	
Gr.Liv.Area	BsmtFin.Type.1	
Full.Bath	BsmtFin.SF.1	
TotRms.AbvGrd	Bsmt.Unf.SF	
SalePrice	Central.Air	
	X1st.Flr.SF	
	Gr.Liv.Area	
	Full.Bath	
	TotRms.AbvGrd	
	Garage.Finish	

Approach:

- We used same approaches and for both groups of attributes. They are as follows:

    - Multiple Linear Regression model
    - K-fold cross validation
    - Forward Subset Selection
    - Ridge Regression model
    - Lasso Regression model

- Here, all the results are square root of the mean squared error presented by the model.
- First, we used multiple linear regression model. Here, we represented SalePrice as linear combination of other attributes. We developed the model using training dataset and tested the accuracy of the model on test dataset. Here are the results for both groups:
    - Group-1: 25619.07
    - Group-2: 8132405
- It is very clear that Group-2 gives very bad result. Interestingly, model built out of Group-2 has higher Adjusted R-squared value compared to Group-1.
- This is an indication that R-squared value alone can't be used to assess the efficiency of the model.

- Next, we used simple linear regression model with k-fold cross validation. We used k = Number of folds = 10. Here, we represented SalePrice as linear combination of other attributes.
- We divided data into 10 folds and took one fold as test set one at a time and rest of them as training set. MSE then will be average over all the models. Here are the results for both groups:
    - Group-1: 25713.94
    - Group-2: 25713.94
- One interesting thing to notice here that, for Group-1, there is not much change in the results. But for group-2, there is a huge change in the result. That shows the robustness of k-fold cross validation approach.

- Next, we used simple linear regression model with forward subset selection.
- Even though we are still using linear regression model, for some reason, FSS seems to provide very bad results for both group of attributes.

- Next, we used ridge regression model. Here, we represented SalePrice as linear combination of other attributes and added some penalty term. This term will tone down the values of coefficients for attributes which are not that significant. Here are the results for both groups:
    - Group-1: 23989.28
    - Group-2: 23199.02
- Next, we used Lasso regression model. Here, we represented SalePrice as linear combination of other attributes and added some penalty term.
- Unlike Ridge regression, this term will tone down the values of coefficients to zero for attributes which are not that significant. Here are the results for both groups:

  - Group-1: 24713.41
  - Group-2: 23070.71

- Both Ridge and Lasso regression model seem to work well for both group of attributes. They seem to be good improvements on the multiple linear model.
- Also, where forward subset selection fails in reducing dimensionality, regression models with added penalty term seem to work with good efficiency.

Summary

- We need to clean the data properly before using it for any kind of analysis purpose. Not cleaning the data properly will have adverse effects on the results.
- Also, from all the model tried, we conclude that k-fold cross validation with multiple linear regression model seem to be the best.
- Lasso and ridge regression models also seem to be very efficient choice for the given data.
- Lastly, Group-1 seem to be most optimized subset of attributes for the provided data per our experimentation. Removing any attribute from Group-1 seem to affect the perform negatively.