**KAGGLE PROJECT (MSDS6371)**

**Predictive Model for Sales Prices of Homes (Ames, Iowa)**

**By Ronak Hamzehlou**

**April 9th, 2022**

# **Introduction**

## **Restatement of Problem**

On this project, we study the Ames, Iowa housing dataset to determine how home's sales prices are influenced by home's living area square footage along with predicting the home's sales prices based of other features. In order to build a model that can serve the community, we using 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, to predict the final price of each home by using different methods in order to choose the best model. At the end, we will challenge this prediction dataset on online Kaggle's competition and record our dataset score.

# **II. Data Description**

This incredible Ames, Iowa dataset was compiled by Dean De Cock, describing the home's sales price from years 2006 to 2010. We're having two data sets (Train & Test). On the Train dataset there is explanatory variables and it has the responses where on the Test set it just has the explanatory variables without the responses. The idea is to fit the best model through Cross-Validation method and other techniques to estimate the model's parameters on Train set and to take that model and make predictions of those response variables in the Test set. Then we will submit the final predicted sale prices on Kaggle and record our score. For the analysis, we focus on the relationship between the square footage of the living area of the house and sale price in 3 neighborhoods.

# **III. Analysis for Question 1**

## **Build and Fit the Model**

The Century 21 Ames real estate company wants to analyze the relationship of square footage of living area of house (GrLivArea) to its sales price in the three neighborhoods in NAmes, BrkSide, and Edward. In this order, we will build a model with the provided dataset and use our conclusion that quantifies the relationship between living area and sale price with respect to the three neighborhoods.

According to the Scatter plot SalePrice vs. GrLIvArea (Appendix, PLOT 1), it looks like there is a linear relationship between the living area and the sale price. However, there are some outliers. We need to further check with some tentative models and decide how to deal with these outliers. We assuming that the observations are independent.

## **Checking Assumptions**

1. **Model 1**

*MuSalesPrice = ß0 + ß1 \*(GrLivArea)*

Linearity: According to scatter plot, QQ plot and histogram of the residuals (Appendix, PLOT 2), the general trend follows linear relationship between Square Foot of Living Area and Sale Price, however there are few outliers that may work against this trend which as we can see, four outliers with studentized residuals larger than ± 2.5 and one outlier with highest Cook’s D > 5. founded. Also, there is no evidence that the residuals do no follow a normal distribution with constant variance. Because we only have to deal with 4 outliers from 383 observations, we can remove the outliers and continue with our testing.

1. **Model 2**

All assumptions are met with this model; however, we are interested in adding in the neighborhood factors. So, we will go on building out Model 3 and see if the neighborhood is significant in our analysis. (Adjusted R-Squared= 0.449). We rerun the same model (1) without these four outliers:

We can see that Scatter plots indicate random distributed residuals.

Cook’s D values lower than 0.10.

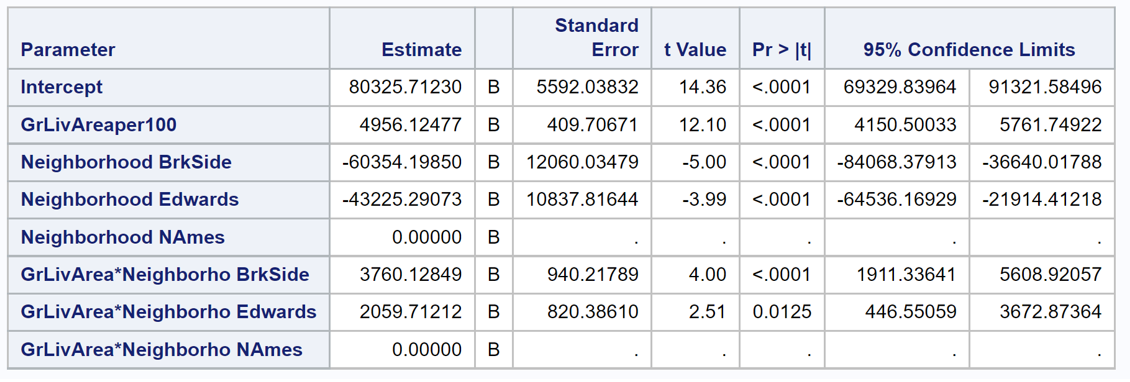
Straight line in QQ plot and symmetric shape of histogram indicate the normal distribution.

Adj R-Square = 0.449.

1. **Model 3**

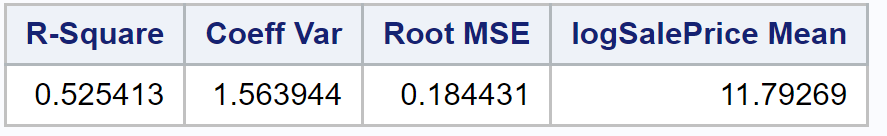
The model assigns SalePrice as the response variable and GrLivArea as the explanatory variable. Based on the scatterplots, there is strong evidence the model will have different slopes and intercepts for each significant parameter.

PredictedSalesPrice = ß0 + ß1 \* (BrkSide) + ß2 \* (Edwards) + ß3 \* (GrLivAreaper100) + ß4 \* (BrkSide \* GrLivAreaper100) + ß5 \* (Edwards \* GrLivAreaper100)

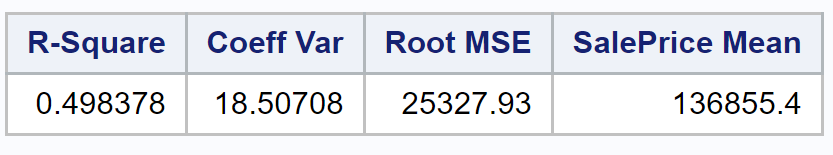


## **Comparing Competing Models**

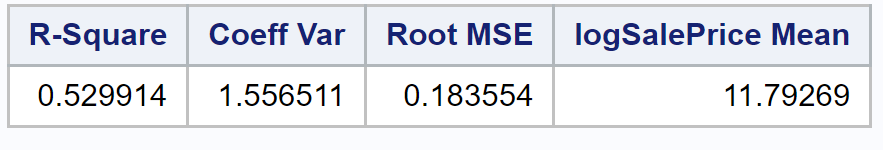
1-Model with log transformed SalePrice



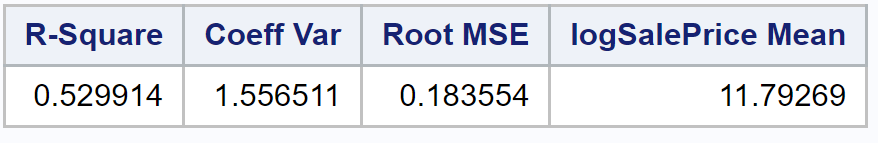
2-Model with log transformed GrLivAreaper100



3-Model with log-log transformed SalePrice-GrLivAreaper100\*



4-Model with log-log transformed SalePrice-GrLivAreaper100 on homes that are less than $300,000 and 40 100 sq. ft.



## **Parameters**

## **Fitted Model**

PredictedSalesPrice = $80,325 – $60,354 \* (BrkSide) – $43,225 \* (Edwards) + $4,956 \* (GrLivAreaper100) + $3,760 \* (BrkSide \* GrLivAreaper100) + $2,059 \* (Edwards \* GrLivAreaper100)

## **Three Regression Equations:**

NAmes Neighborhood: PredictedSalesPrice= $80,325 + $4,956 \* GrLivAreaper100

BrkSide Neighborhood: PredictedSalesPrice= $19,971 + $8,716 \* GrLivAreaper100

Edwards Neighborhood: PredictedSalesPrice= $37,100 + $7,015 \* GrLivAreaper100

## **Interpretation**

The estimated average sale price for houses in the NAmes neighborhood with zero square feet of living area is $80,327.

In the NAmes neighborhood for every 100 square feet increase in living area the estimated average sale price will increase by $4,956.

The estimated average sale price for houses in the BrkSide neighborhood with zero square feet of living area is $19,971.

In the BrkSide neighborhood for every 100 square feet increase in living area the estimated average sale price will increase by 8,716.

The estimated average sale price for houses in the Edwards neighborhood with zero square feet of living area is $37,100.

In the Edwards neighborhood for every 100 square feet increase in living area the estimated average sale price will increase by $7,015.

## **Confidence Intervals**

95% CI of NAmes’ intercept: ($69330.31, $91321.11)

95% CI of difference between NAmes and Brkside’s intercept: ($-84067.36, $-36641.03)

95% CI of difference between NAmes and Edwards’ intercept: ($-64535.25, $-21915.32)

## **Conclusion**

After using the vanilla regression and addressing the outliers of a living area greater than 4,000 sq. ft. and a sales price of over $300,000, there is significant evidence that both the neighborhood and square footage of the living area influence the estimated sale price of the home in Ames. The intercept in this model provides an estimated average sale price of $80,327 for a house with a living area of zero square feet in the NAmes neighborhood, with a 95% CI of ($69330.31, $91321.11). For a house with a living area of zero square feet, the model gives us an estimated average sales price decrease of $60,354 with a 95% CI of ($36641.03, $84067.36) from houses in the NAmes neighborhood for the BrkSide neighborhood and an estimated average sales price decrease of $43,225 with a 95% CI of ($21915.32, $64535.25) from the NAmes neighborhood for the Edwards neighborhood.

For every increased living area of one hundred square feet the estimated average sales price increases $4,956 for a home in the NAmes neighborhood, the estimated average sales price in the BrkSide neighborhood increases $3,780 more than the houses in the NAmes neighborhood and the estimated average sale price in the Edwards neighborhood increases $2,059 more than the NAmes neighborhood. Since this is an observational study, only associations can be made to those three neighborhoods.

# **IV. Analysis for Question 2**

## **Restatement of Problem**

We want to build the most predictive model for sales prices of homes in all of Ames Iowa by using all the provided variables Century 21 company. Our method is to use multiple linear regression to evaluate and analyze all variables in the dataset in order to get a good model.

## **Model Selection**

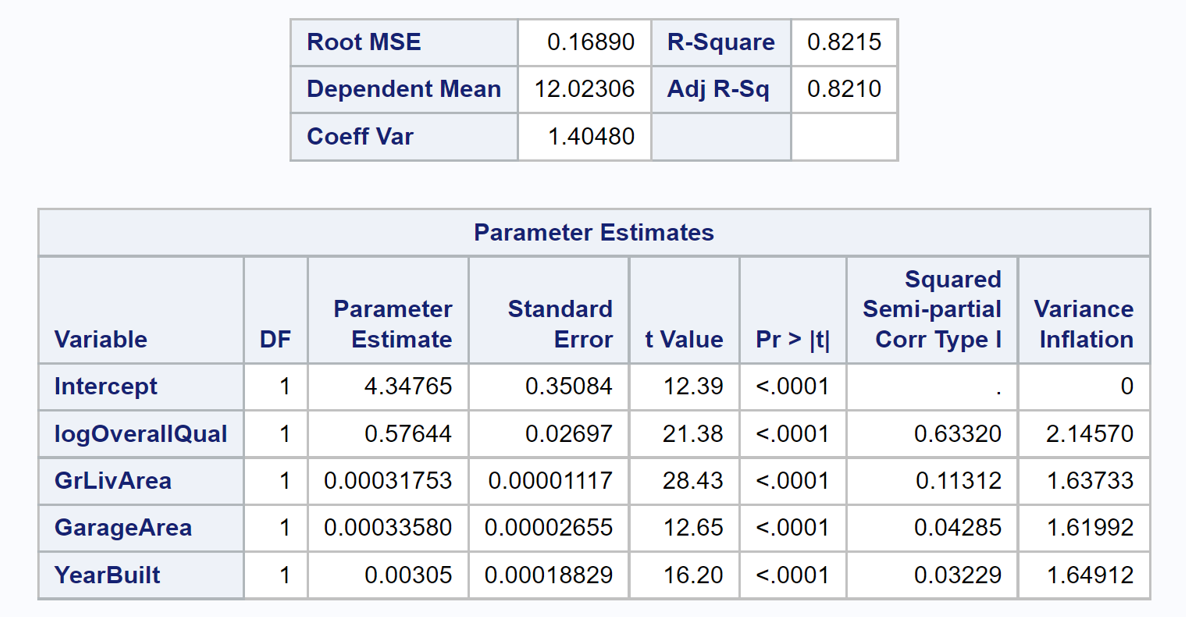
We will use Stepwise, Forward, Backward and Custom process selection in our analysis. We will also compare the parameters of the different models over the adjusted R2, internal CV Press and Kaggle Score.

**Type of Selection**

First, we assume all observations in the dataset are independent. We will prepare the data by selecting numerical variables as predictor. By removing variables with NA’s (eg. LotFrontage, MasVnrArea, GarageYrBlt) and variables with wrong name (eg. 1stFlrSF, 2ndFlrSF, 3SsnPorch), we have 31 numerical variables as predictors. The Custom was used where we selected all variables that resulted in a correlation when plotted against sale price. For categorical variables if a category showed an increase or decrease in sale price when compared to its counterparts this variable was included into the model.

**Stepwise**

In this part, we build a model using Stepwise method with 12 Numeric Predictors. Next, we build a model with Top 4 Numeric Predictors and we do a log transformation on SalePrice and OverallQual. By running SAS program, the analysis shows an Adj R-Square = 0.821. Residual scatter plots looks much better, QQ plot is linear and the Histogram indicates that normality holds. Cook’s D values lower than 0.125.

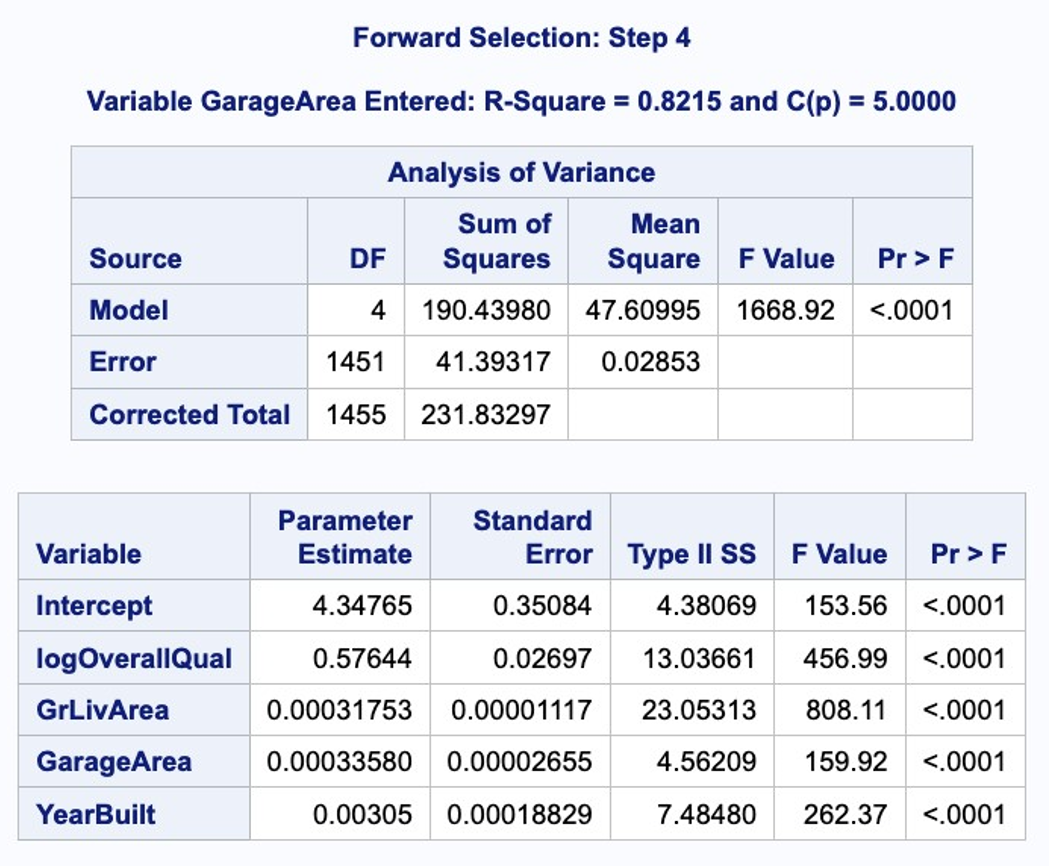


Our regression model is

μ(logSalePrice) = b0 + b1(logOverallQual) + b2(GrLivArea) + b3(YearBuilt) + b4(GarageArea) With b0 = 4.34765, b1 = 0.57644, b2 = 0.00031753, b3 = 0.00033580 and b4 = 0.00305.

**Forward**

First, we build this Model with 12 numerical predictions. Then choosing the top 4 with highest Partial R2 squared. In order to removing the curve relation between SalePrice and OverallQuall, we build the log transformation with final Adj R-Squared of 0.8215 after four steps and as we can see the scatter plots has straight line in QQ plot.



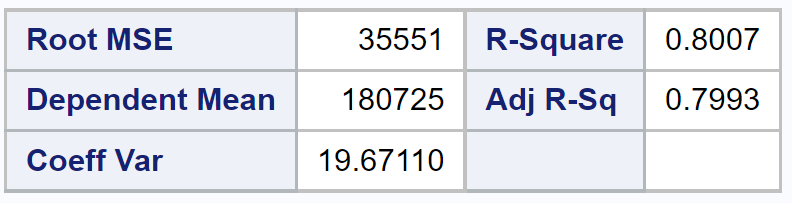
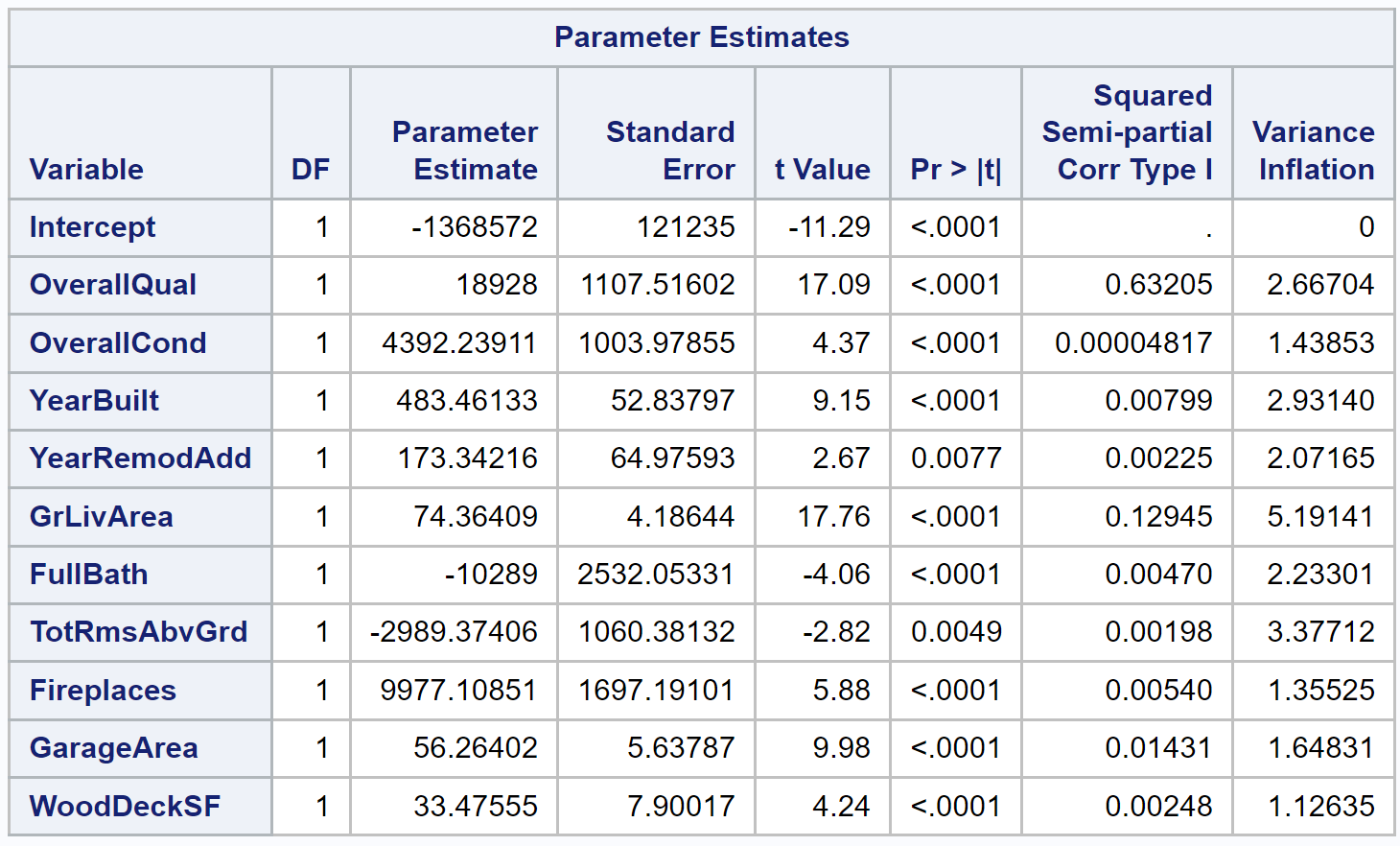
Our regression model is:

μ(logSalePrice) = b0 + b1(logOverallQual) + b2(GrLivArea)+ b3(YearBuilt) + b4(GarageArea)

with b0 = 4.34765, b1 = 0.57644, b2 = 0.00031753, b3 = 0.00033580 and b4 = 0.00305.

**Backward**

We will apply the Backward Selection in this model using same data from previous. We will use 12 numerical variables as Predictors for this selection type. Please refer to Table 2.29, Plots 2.30-2.32 [Appendix], we have similar observation results as the Forward Selection Method with 12 Numeric Predictors (Cook’s D, Constant Variance, Normality...). By running SAS program, we eliminate 2 variables and the Adj R-Square = 0.7993.

Then our regression model is:

μ(SalePrice)=b0+b1(OverallQual)+b2(OverallCond)+b3(YearBuilt)+b4(YearRemodAdd)+b5(GrLivArea)+b6(FullBath)+b7(TotRmsAbvGrd)+b8(Fireplaces)+b9(GarageArea)+b10(WoodDeckSF)

with b0 = -1368572, b1 = 18928, b2 = 4392.2, b3 = 483.5, b4 = 173.34, b5= 74.36, b6= -10289 , b7 = -2989.37, b8 = 9977.1, b9 = 56.26 and b10 = 33.48.

**Custom**

Next, we will customize a model with categorical variables added. (see Appendix for SAS codes)

## **Checking Assumptions**

Although all the assumptions were not entirely satisfied, we assumed central limit theorem would take effect with a sizable dataset. Our approach during the model selection influenced our conclusions on the following assumptions:

1. Normality: There is no strong evidence against normality when considering the histogram of residual plots. With regards to all the explanatory variables, it is not ideal to have normality for every predictor considered in each model.
2. Linearity: There is no strong evidence against linearity when considering the scatterplots of residual Again, although not ideal to have linearity for every selected predictor, we proceeded with accepting our selected variables based of a reasonable adj. R^2 and low Kaggle score.
3. Equal Variance: There is no strong evidence against equal variance when considering the QQPlot of residuals. On the top-end of the tail, there is evidence of some points not having equal variance but due to the size of the data set, we allowed for the observations to stay in model. After proceeding with the various models, it was concluded that the observations were not influential enough to be removed as it did not greatly affect the prediction score.
4. Independence: We will assume independence, although the data gathering process was not explained.

## **Comparing Competing Models**

|  |  |  |  |
| --- | --- | --- | --- |
| **Predictive Models** | **Adjusted R2** | **CV PRESS** | **Kaggle Score** |
| Forward | 0.8782 | 27.28878 | 0.14707 |
| Backward | 0.8773 | 27.20917 | 0.14688 |
| Stepwise | 0.8759 | 27.30510 | 0.14771 |
| CUSTOM | 0.8769 | AIC-3465.75254  ICC -3464.75834  SBC -4582.07548 | 0.14710 |

## **Conclusion**

After running the three models in the last step (Build models), we upload our files to Kaggle to get our Kaggle score. Finally, we chose our custom model as Backward Selection with Cross Validation. In running the various models, we noticed that each run of the same model could produce different output and produce Kaggle scores It is also worth exploring that a higher adjusted R^2 does not mean the Kaggle score will be better than another with a lower adjusted R^2. Therefore, we conclude that assumptions as well as multiple linear regression are not always ideal in predictive models, we suspect algorithms like Random Forest or Logistics Regression could be a better fit.

**Kaggle Competition: Resulted in Place# Resulted in Place# 3260 and Kaggle Score 0.17575(Appendix, Page 7)**