System Used:
MacBook Pro - 2.8 GHz Intel Core i7, 16 GB 2133 MHz LPDDR3

For Project 1 Part 1, I evaluated 4 models with Ames Housing dataset.

I tried 1) Linear Regression Model, 2) Lasso Regression 3) Random Forest and 4) Gradient Boost (xgboost)

Submitted code has Lasso, RF and GBM implementation present.

I evaluated my code in 2 ways:

- Random 70% training and 30% test splits
- 10 splits provided in **Project1_test_id.txt**

In this report, I am submitting RMSE values from some of the splits from Project1_test_id.txt

Accuracy Evaluation:1
Test Data: 5th column of Project1_test_id.txt
Training Data: Rest of the dataset

glmnet RMSE: 0.12388 – Elapsed time: 1.957
Random Forest RMSE: 0.11728 – Elapsed time: 7.475
Xgboost RMSE: 0.11782 – Elapsed time: 11.202

Accuracy Evaluation:2
Test Data: 7th column of Project1_test_id.txt

Training Data: Rest of the dataset

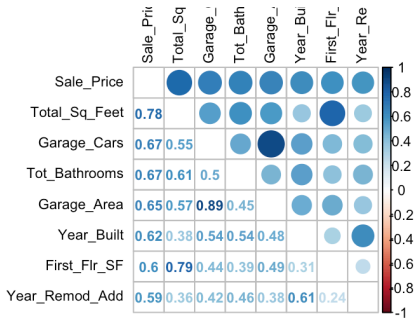
glmnet RMSE: 0.10831 – Elapsed time: 1.725
Random Forest RMSE: 0.12732 – Elapsed time: 6.963
Xgboost RMSE: 0.115471 – Elapsed time: 10.733

Accuracy Evaluation:3
Test Data: 10th column of Project1_test_id.txt
Training Data: Rest of the dataset

glmnet RMSE: 0.11439 – Elapsed time: 1.903
Random Forest RMSE: 0.12968 – Elapsed time: 7.421
Xgboost RMSE: 0.118100 – Elapsed time: 11.726

Feature Engineering:

1. Used correlation matrix as shown below to figure out correlation relation between the predictors and drop the variable accordingly.



2. Outliers: Evaluated the numerical parameters against the sale price and removed few outliers like living area > 4500, lot frontage > 200 etc
3. Normalized the response variable in log.

4. Converted few numerical variables into factors and many categorical variables into numeric based on exploratory analysis and some common sense.
5. Merged few variables together which showed strong correlation with the sale price afterwards. For example, converted ground floor living area and basement area into total square feet area.
6. Handling of NULL values by trying imputing but eventually replacing then with 0 worked well for me.
7. Categorical level handling between training and test dataset by converting less frequent levels into ‘other’ category.
8. Experimented with Winsorization technique
9. Spend time tuning the parameters of Random Forest and Xgboost, which helped in improving the prediction further.
10. Employed one hot encoding technique for handling the categorical variables.
11. Tried transforming few independent variables and eventually transformed Lot_Area to logarithm, as it looked more normal and helped in better prediction.
12. Tried many more techniques along the way to improve the accuracy. Listing the techniques in this report which I eventually utilized to build the model.