Smith
SCHOOL OF BUSINESS
Queen's University

Master of
Management Analytics
Toronto

MMA 867
Predictive Modeling

Professor Jue Wang

Team Assignment 1
Due Date: April 30, 2022

Team Adelaide

Student Name	Student Number
Ishaan Sinha	20290167
Jiali Bai	20316342
Juefei Chen	20311141
Kripa Shanker Nayak	20315710
Oluwaseun Adelana	20310041
Yichen Yuan	20315046

Order of files:

Filename	Pages	Comments and/or Instructions
MMA867_TeamAssignment_1	10	

Additional Comments:

MMA 860 Assignment 2

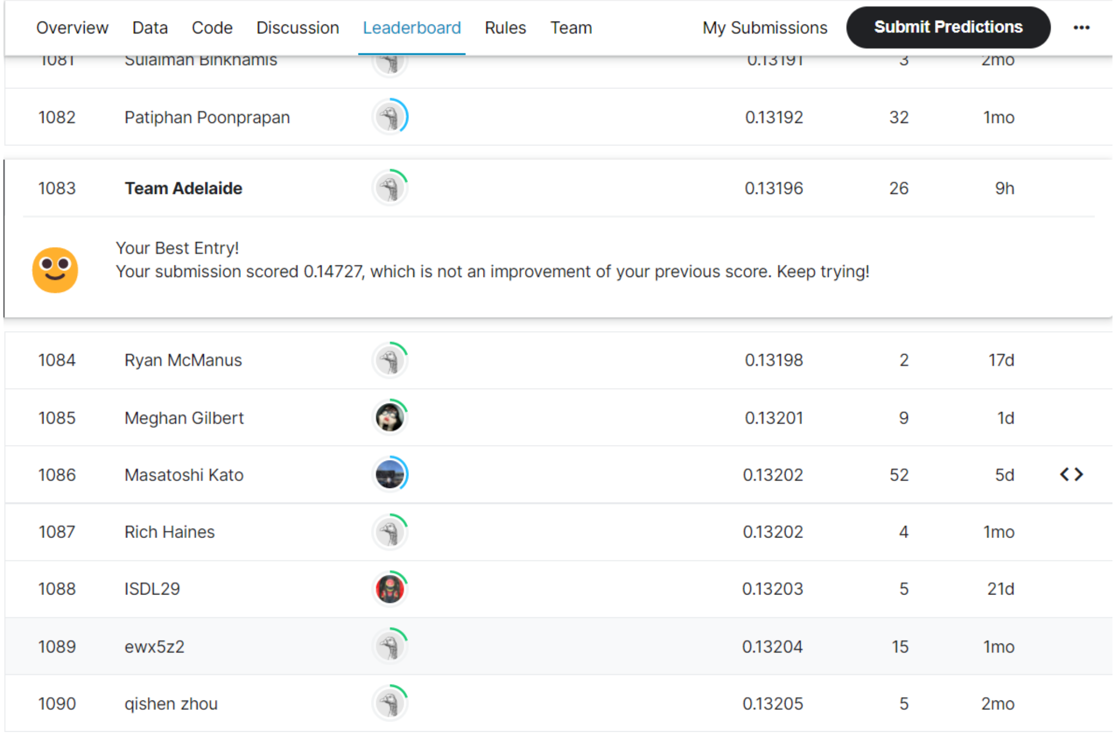
Go to the following competition on house price prediction:
https://www.kaggle.com/c/house-prices-advanced-regression-techniques/overview

1. **[20 pts]:** Read the instructions on Kaggle. Learn how to join a Kaggle competition and submit your results. You will find that some predictors contain "missing data", NA. Figure out how to handle missing data in regression.

2. **[60 pts]:** Build a regression model for house price prediction and write a report explaining how you approached the task, the steps you took, and how you revised your model (**must explore both LASSO and Ridge regression**) as your analyses progressed, etc. Comment on the quality of your predictions.
Include your model as an Appendix in your report. Submit the PDF of your report, and your model file(s) (code, spreadsheet, etc.)

3. **[20 pts]:** On the front page of your report, include your position on the leaderboard at the time of your last submission. Please also include the screenshot showing your team's position on the leaderboard in the Appendix.
a. If ranked top 20% on the leaderboard, each team member will receive a full 20 pts.
b. If ranked top 21-30% on the leaderboard, each team member will receive 15 pts.
c. If ranked top 31-40% on the leaderboard, each team member will receive 10 pts.
d. If ranked top 41-50% on the leaderboard, each team member will receive 5 pts.
e. If ranked below 50% on the leaderboard, each team member will receive 0 pt.

Leaderboard Position: -



1081	Sulaiman Binkhamis	0.13191	3	2mo
1082	Patiphan Poonprapan	0.13192	32	1mo
1083	**Team Adelaide**	0.13196	26	9h

Your Best Entry!
Your submission scored 0.14727, which is not an improvement of your previous score. Keep trying!

1084	Ryan McManus	0.13198	2	17d
1085	Meghan Gilbert	0.13201	9	1d
1086	Masatoshi Kato	0.13202	52	5d
1087	Rich Haines	0.13202	4	1mo
1088	ISDL29	0.13203	5	21d
1089	ewx5z2	0.13204	15	1mo
1090	qishen zhou	0.13205	5	2mo

Introduction

This report explains how we build a regression model for house price prediction. Several factors such as type of neighbourhood, school district zoning, number of rooms, access to transportation, and amenities influence how much a house is sold. Buyers typically have a checklist of features they look for in a property, and these features determine the value and final sale price of the property.

We have been provided with a dataset that includes 1 response variable; Sale price, and 80 predictor variables that describe various features of residential homes in Ames, Iowa. The predictor variables are a combination of numeric and categorical data. They describe features and factors that influence the final sale price of a property, some of which include property zoning classification, property size, location, type of house being sold, condition of the property, the year it was built or remodeled, property exterior features, the existence of a basement and its conditions, kitchen features, quality, and its conditions, etc.

We used techniques in exploratory data analysis to transform the dataset in preparation for modeling. Categorical variables such as basement quality (BaseQual) which had 'NA' as responses only indicated that the feature does not exist for the property. for example, NA for basement quality meant the property does not have a basement.

The dataset is split into two categories; the training dataset which is used to fit the best model with the least MSE value while the validation/test dataset is used to validate the model by comparing actual prices to the predicted property sale prices.

Data Cleaning

After reviewing and summarizing the dataset, the key observations are the following:

- The training dataset comprises 1460 rows and 81 columns.
- The testing dataset comprises 1459 rows and 80 columns.
- Based on the description of the dataset, not all NA values represent missing data, some NA values do have meanings.
- Most of the columns do not have any missing value, however, the missing values are concentrated in the variables of Alley, PoolQC, Fence, MiscFeature, FireplaceQu, LotFrontage.

The key action for our data cleaning process is to get rid of the missing value cells. Those variables with too many missing values are directly removed from the dataset, for the rest, the NAs values are replaced by None to avoid confusion. We also used mode for categorical data and median/mean for numerical data, to replace all the NAs values. Finally, we converted some of the numerical value which represented categorical data from numeric object to a string data type.

Model Development

Once the data had been prepared, we leveraged it to build a linear regression model with Sale Price as the response variable and all other variables were included as predictor variables. Summary of our regression model provided insights into which of the predictor variables are more significant compared to others in the model. Moreover, the Normal Q-Q plot shows us that the error terms were not normally distributed, and the data was heteroskedastic.

To deal with these issues, we tried to normalize some of the numerical data by using data transformations such as log transformation. Whether a variable should be normalized or not was decided by looking at their respective histograms. Variables which showed skewness in distribution were then normalized. As a result, the response variable SalePrice along with other variables were transformed. We chose not to transform numerical predictors such as GarageCars and Bedrooms as these are discrete variables. Finally, we looked at all the categorical predictors and tried to understand what other predictors can be related to these. Based on this, we tried to come up with interactions for all the categorical predictors.

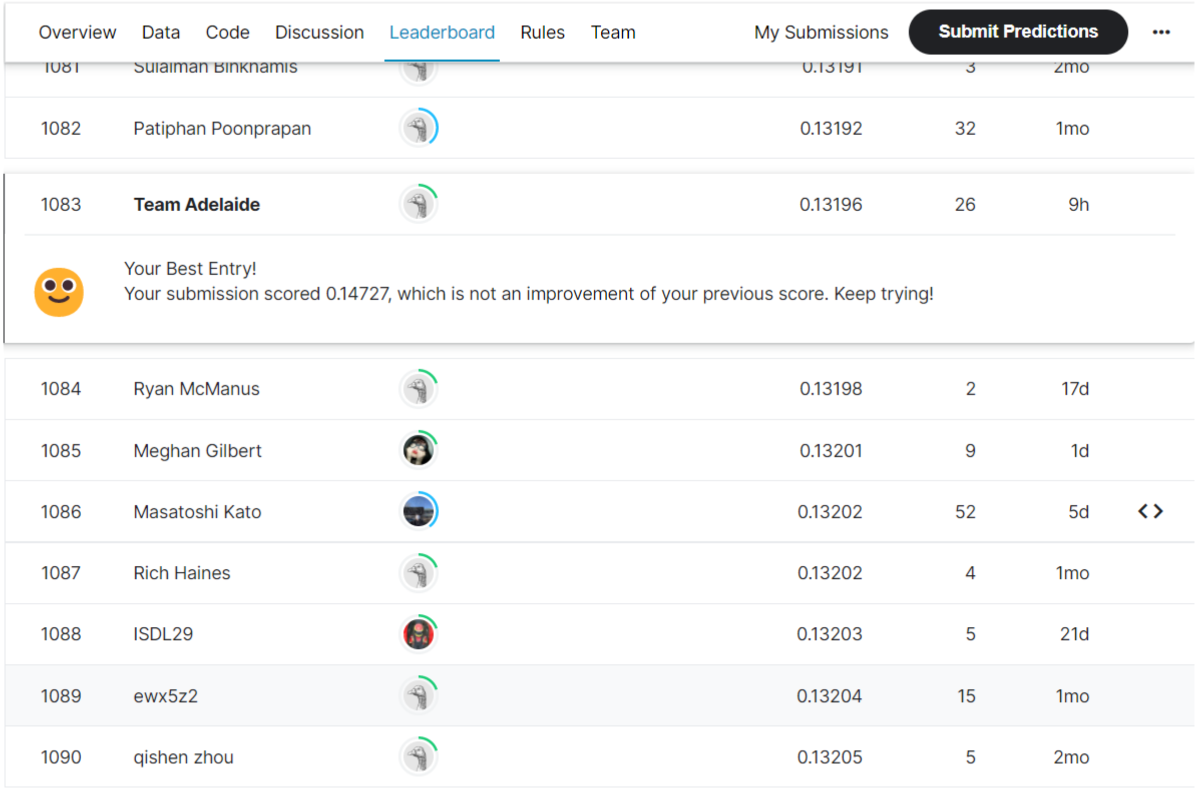
In our regression model, we added interaction variables to our model matrix to seek if there would be any relationships among different variables that bring significant impact on a combined basis to the model rather than independently. We grouped variables with similar attributes together; for example, variables related to basement situations (BsmtQual, BsmtCond, and TotalBsmtSF) were considered as a set of interaction variables and were tested to see how these variables are jointly impacting the house sales price. After that, we applied transformation, such as log and square root calculation on all continuous variables to improve model fit and stabilize the impact from large value. All these variables were then added in the model matrix and using the 'glmnet' package, we ran both a LASSO and a Ridge Regression. We ran the model using multiple combinations of predictor variables and selected the one with the lowest average RMSE.

Conclusion

After data cleaning, feature engineering, and modeling, we found this dataset performs better on a Lasso regression with the optimal penalty parameter, lambda 0.0085003. Lasso tends to perform better when there are a small number of significant parameters, and the others are close to zero. Among 80 variables we fit in the model, only 27 variables have coefficients greater than 0.01. In the result of the coefficient from the model, we can see factors such as overall condition, house quality rating, housing area square feet, and the number of rooms/baths would make a significant impact on the sale price. Overall, this is a great performance model and received a score of 0.13 from Kaggle, which evaluated the model on Root-Mean-Squared-Error between the logarithm of the predicted value and the logarithm of the observed sales price.

Appendix

Leaderboard: -



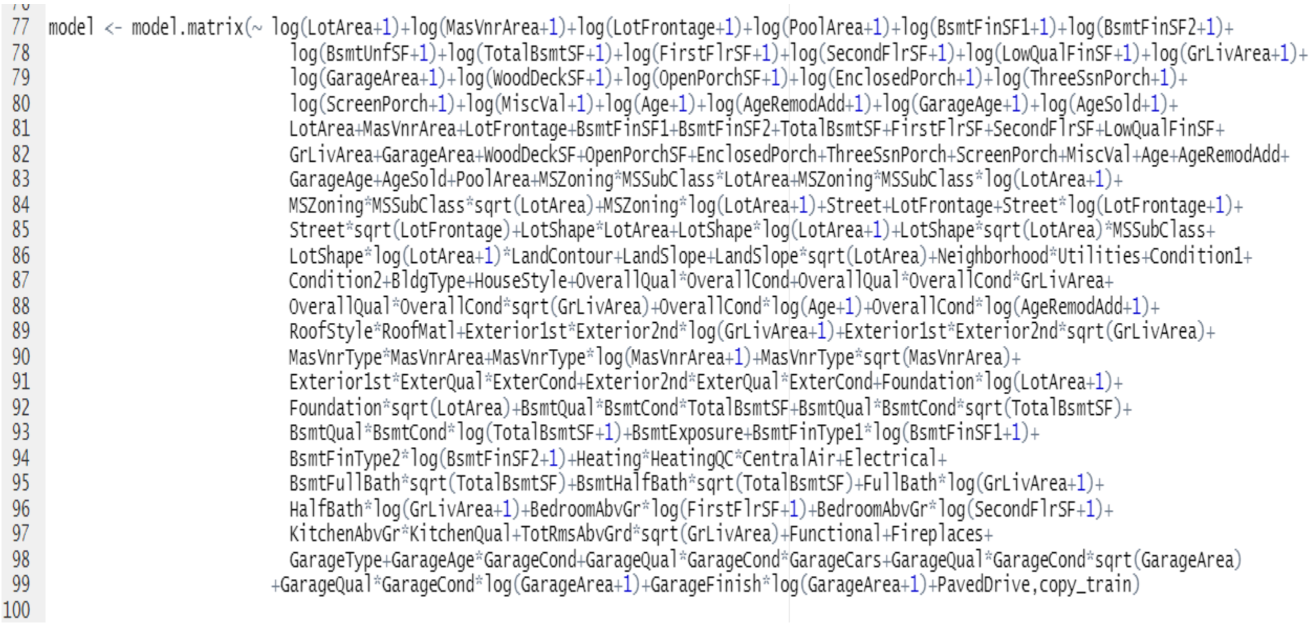
1081	Sulaiman Binkhamis	0.13191	3	2mo	
1082	Patiphan Poonprapan	0.13192	32	1mo	
1083	**Team Adelaide**	0.13196	26	9h	

Your Best Entry!
Your submission scored 0.14727, which is not an improvement of your previous score. Keep trying!

1084	Ryan McManus	0.13198	2	17d	
1085	Meghan Gilbert	0.13201	9	1d	
1086	Masatoshi Kato	0.13202	52	5d	<>
1087	Rich Haines	0.13202	4	1mo	
1088	ISDL29	0.13203	5	21d	
1089	ewx5z2	0.13204	15	1mo	
1090	qishen zhou	0.13205	5	2mo	

Our final rank is 1083 out of 4164, which is in the top 26%.

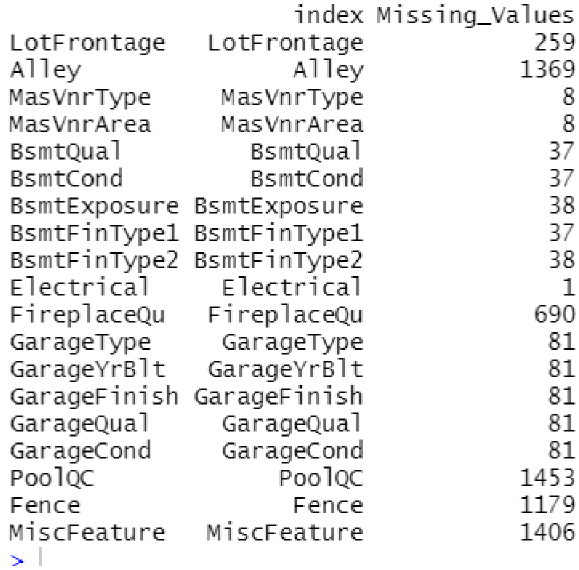
Final Model: -



```r
model <- model.matrix(~ log(LotArea+1)+log(MasVnrArea+1)+log(LotFrontage+1)+log(PoolArea+1)+log(BsmtFinSF1+1)+log(BsmtFinSF2+1)+
                          log(BsmtUnfSF+1)+log(TotalBsmtSF+1)+log(FirstFlrSF+1)+log(SecondFlrSF+1)+log(LowQualFinSF+1)+log(GrLivArea+1)+
                          log(GarageArea+1)+log(WoodDeckSF+1)+log(OpenPorchSF+1)+log(EnclosedPorch+1)+log(ThreeSsnPorch+1)+
                          log(ScreenPorch+1)+log(MiscVal+1)+log(Age+1)+log(AgeRemodAdd+1)+log(GarageAge+1)+log(AgeSold+1)+
                          LotArea+MasVnrArea+LotFrontage+BsmtFinSF1+BsmtFinSF2+TotalBsmtSF+FirstFlrSF+SecondFlrSF+LowQualFinSF+
                          GrLivArea+GarageArea+WoodDeckSF+OpenPorchSF+EnclosedPorch+ThreeSsnPorch+ScreenPorch+MiscVal+Age+AgeRemodAdd+
                          GarageAge+AgeSold+PoolArea+MSZoning*MSSubClass*LotArea+MSZoning*MSSubClass*log(LotArea+1)+
                          MSZoning*MSSubClass*sqrt(LotArea)+MSZoning*log(LotArea+1)+Street+LotFrontage+Street*log(LotFrontage+1)+
                          Street*sqrt(LotFrontage)+LotShape*LotArea+LotShape*log(LotArea+1)+LotShape*sqrt(LotArea)*MSSubClass+
                          LotShape*log(LotArea+1)*LandContour+LandSlope+LandSlope*sqrt(LotArea)+Neighborhood*Utilities+Condition1+
                          Condition2+BldgType+HouseStyle+OverallQual*OverallCond+OverallQual*OverallCond*GrLivArea+
                          OverallQual*OverallCond*sqrt(GrLivArea)+OverallCond*log(Age+1)+OverallCond*log(AgeRemodAdd+1)+
                          RoofStyle*RoofMatl+Exterior1st*Exterior2nd*log(GrLivArea+1)+Exterior1st*Exterior2nd*sqrt(GrLivArea)+
                          MasVnrType*MasVnrArea+MasVnrType*log(MasVnrArea+1)+MasVnrType*sqrt(MasVnrArea)+
                          Exterior1st*ExterQual*ExterCond+Exterior2nd*ExterQual*ExterCond+Foundation*log(LotArea+1)+
                          Foundation*sqrt(LotArea)+BsmtQual*BsmtCond*TotalBsmtSF+BsmtQual*BsmtCond*sqrt(TotalBsmtSF)+
                          BsmtQual*BsmtCond*log(TotalBsmtSF+1)+BsmtExposure+BsmtFinType1*log(BsmtFinSF1+1)+
                          BsmtFinType2*log(BsmtFinSF2+1)+Heating*HeatingQC*CentralAir+Electrical+
                          BsmtFullBath*sqrt(TotalBsmtSF)+BsmtHalfBath*sqrt(TotalBsmtSF)+FullBath*log(GrLivArea+1)+
                          HalfBath*log(GrLivArea+1)+BedroomAbvGr*log(FirstFlrSF+1)+BedroomAbvGr*log(SecondFlrSF+1)+
                          KitchenAbvGr*KitchenQual+TotRmsAbvGrd*sqrt(GrLivArea)+Functional+Fireplaces+
                          GarageType+GarageAge*GarageCond+GarageQual*GarageCond*GarageCars+GarageQual*GarageCond*sqrt(GarageArea)
                         +GarageQual*GarageCond*log(GarageArea+1)+GarageFinish*log(GarageArea+1)+PavedDrive,copy_train)
```

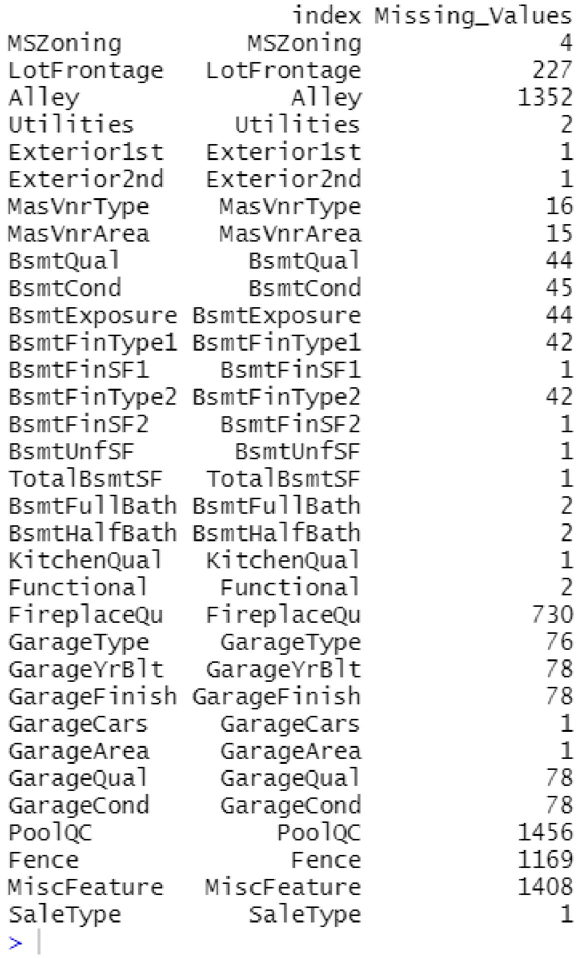
Missing Data: -

Train Set



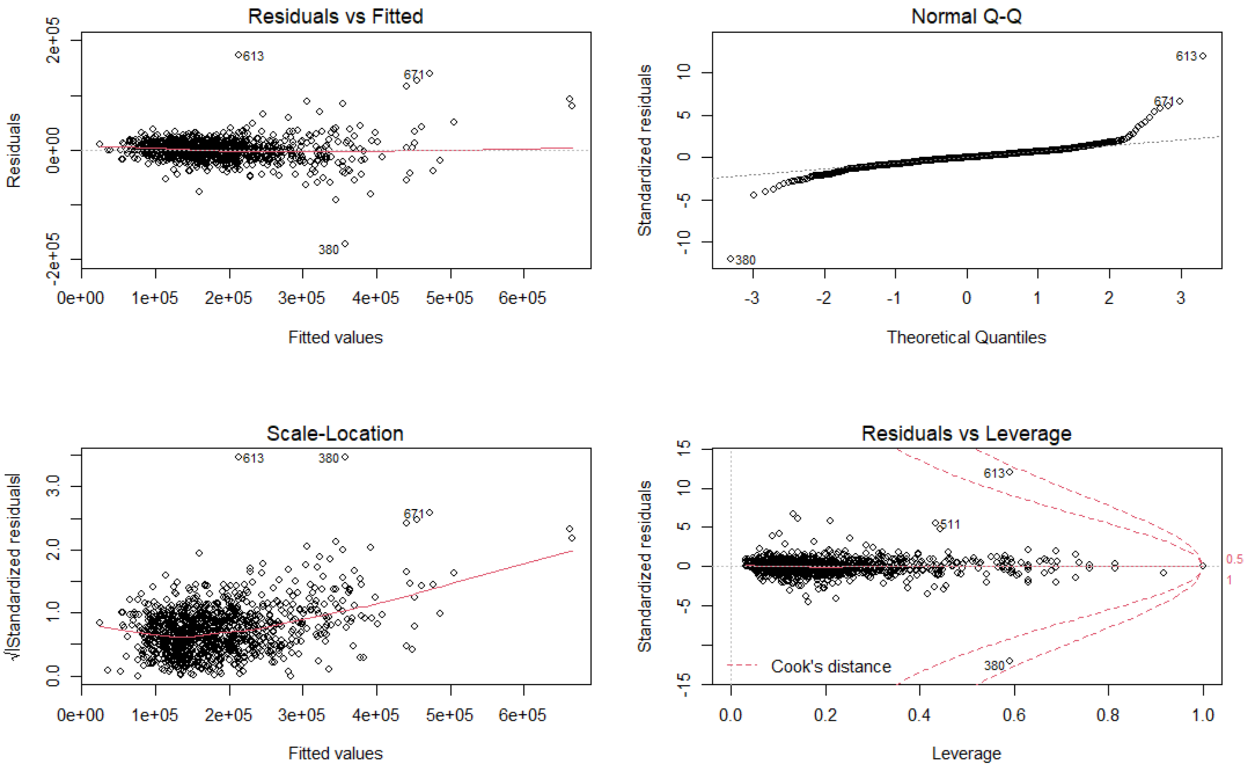
```
                   index Missing_Values
LotFrontage   LotFrontage            259
Alley               Alley           1369
MasVnrType     MasVnrType              8
MasVnrArea     MasVnrArea              8
BsmtQual         BsmtQual             37
BsmtCond         BsmtCond             37
BsmtExposure BsmtExposure             38
BsmtFinType1 BsmtFinType1             37
BsmtFinType2 BsmtFinType2             38
Electrical     Electrical              1
FireplaceQu   FireplaceQu            690
GarageType     GarageType             81
GarageYrBlt   GarageYrBlt             81
GarageFinish GarageFinish             81
GarageQual     GarageQual             81
GarageCond     GarageCond             81
PoolQC             PoolQC           1453
Fence               Fence           1179
MiscFeature   MiscFeature           1406
> 
```

Test Set: -

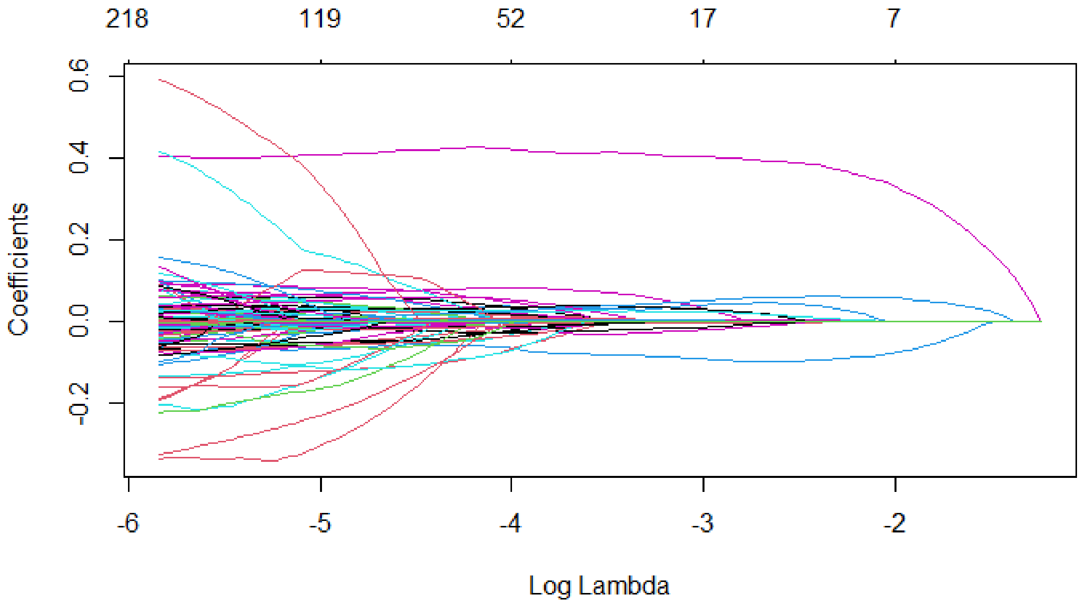


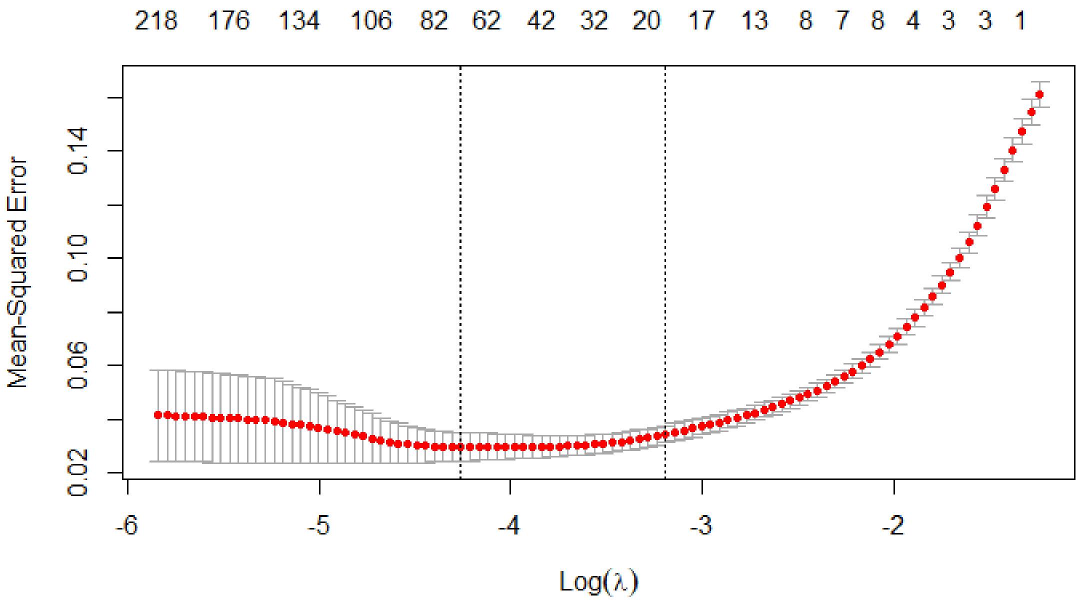
```
                    index Missing_Values
MSZoning         MSZoning              4
LotFrontage   LotFrontage            227
Alley               Alley           1352
Utilities       Utilities              2
Exterior1st   Exterior1st              1
Exterior2nd   Exterior2nd              1
MasVnrType     MasVnrType             16
MasVnrArea     MasVnrArea             15
BsmtQual         BsmtQual             44
BsmtCond         BsmtCond             45
BsmtExposure BsmtExposure             44
BsmtFinType1 BsmtFinType1             42
BsmtFinSF1     BsmtFinSF1              1
BsmtFinType2 BsmtFinType2             42
BsmtFinSF2     BsmtFinSF2              1
BsmtUnfSF       BsmtUnfSF              1
TotalBsmtSF   TotalBsmtSF              1
BsmtFullBath BsmtFullBath              2
BsmtHalfBath BsmtHalfBath              2
KitchenQual   KitchenQual              1
Functional     Functional              2
FireplaceQu   FireplaceQu            730
GarageType     GarageType             76
GarageYrBlt   GarageYrBlt             78
GarageFinish GarageFinish             78
GarageCars     GarageCars              1
GarageArea     GarageArea              1
GarageQual     GarageQual             78
GarageCond     GarageCond             78
PoolQC             PoolQC           1456
Fence               Fence           1169
MiscFeature   MiscFeature           1408
SaleType         SaleType              1
> 
```

Regression plots: -

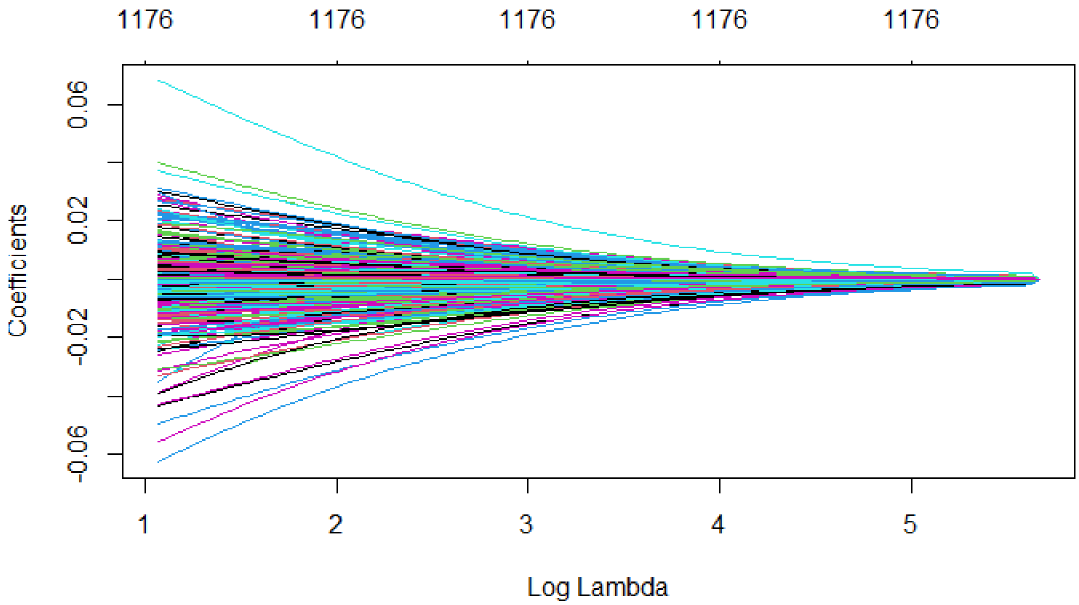


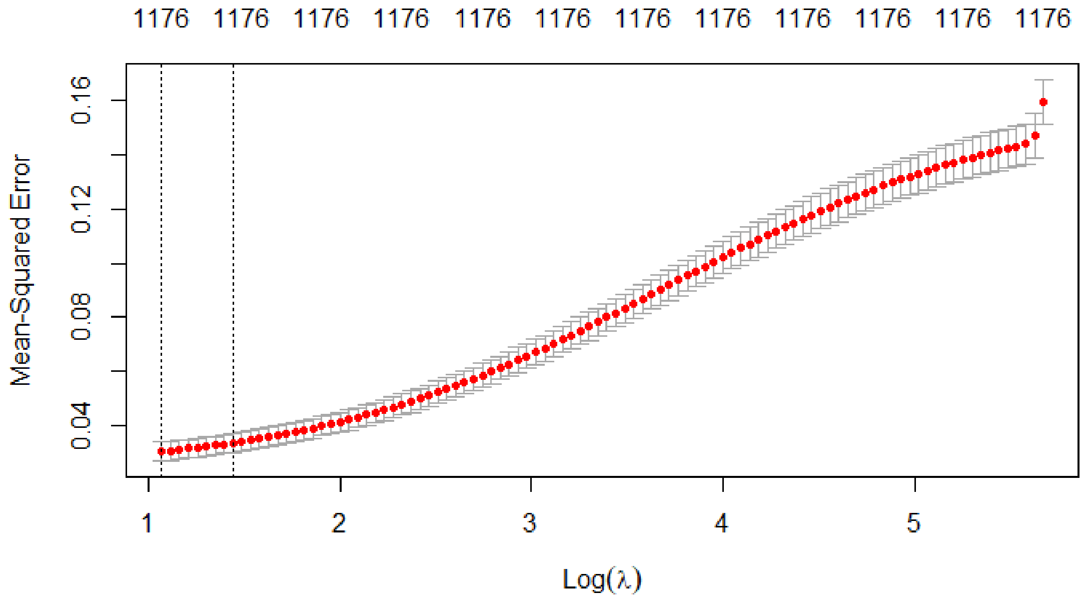
LASSO:


218
119
52
17
7
Coefficients
0.6
0.4
0.2
0.0
-0.2
-6
-5
-4
-3
-2
Log Lambda


218
176
134
106
82
62
42
32
20
17
13
8
7
8
4
3
3
1
Mean-Squared Error
0.14
0.10
0.06
0.02
-6
-5
-4
-3
-2
Log(λ)

Ridge Regression: -


Coefficients
Log Lambda


Mean-Squared Error
Log(λ)