Executive Summary

The objective of this project was to find the best model with the least test prediction error for a test data based on a model developed on the training set of 550 observations and 8 predictors. Multiple models were considered including the modification of inputs using arithmetic operations like natural logs and squares. Three best models were chosen by splitting the initial training data set into a training and test data set into multiple splits of 90-10, 75-25, 50-50 and evaluating the prediction error.

The technical details of the three best models chosen are mentioned in the technical analysis report below(pg. 2-8). The best model among the three was reported and the details are also attached(pg. 8). Further, when the actual test data was made available, this model was used to predict the responses based on the values of the predictors in the test data, and the error between the predicted responses and actual responses was evaluated(pg. 9). Further, various combinations of the parameters of the best model were evaluated to improve the prediction accuracy of the best model, technical details of which are available on pages 10-11. There was a significant improvement of about **67.75%** in the prediction error of the previous best model and the new best model obtained after evaluating the parameters.

The project team details and contributions are summarized below:

Participants			
Sr. No.	**Name**	**UIN**	**Email**
1	Ajinkya Mahesh Zalkikar	530005943	ajinkya.zalkikar@tamu.edu
2	Mohit Deepak Chhaparia	431000925	mohit_chhaparia@tamu.edu
3	Eashwar Venkitesan Iyer	931002077	eashwar07@tamu.edu
4	Rahul Ramesh	430000147	rahul.ramesh@tamu.edu

Sr. No.	**Name**	**Contribution**
1	Ajinkya	Report, Evaluating Ridge, Lasso Models
2	Mohit	Report, Evaluating Tree Models
3	Eashwar	Report, Evaluating Linear Regression Models
4	Rahul	Report, Bootstrapping on Linear Regression Models

Technical Analysis – ISEN 613 – Project Report (Spring 2021)

Model 1: Linear regression with interaction terms

The first step before applying models to the data was to determine the correlation of different predictors. De-correlation was carried out to remove highly correlated variables to improve the model. Based on the output obtained from the correlation matrix, it was observed that predictor X4 was highly correlated with X1, X2 & X5, and thus, to avoid masking of significant predictors, X4 was eliminated from the dataset.

Next, we applied the normal linear regression model on the entire data, wherein we concluded that X6 was the only non-significant predictor, and thus X6 was further eliminated from the dataset for which various models were tried out.

Thus, X4 and X6 have been eliminated from all the ensuing models.

Multiple linear regression was done extensively considering various interaction terms and transformations to obtain the best possible model based on least MSE. The model with the least test MSE involves the output (Y) as a function of linear combination of predictors X1, X2, X3, X5,X7,X8, X1, X2^2, X3^2, X7^2, X8^2 and some important interactions among them. The complete model is as below:

$$Y = \beta 0 + \beta 1 * X1 + \beta 2 * X1^2 + \beta 3 * X2 + \beta 4 * X2^2 \ + \beta 5 * X3 \ + \beta 6 * X3^2 + \beta 7 * X5 + \beta 8 * X7 \\ + \beta 9 * X7^2 + \ \beta 10 * X8 + \beta 11 * X8^2 + \beta 12 * X1 * X2 \ + \beta 13 * X1 * X3 + \beta 14 * X2 * X3 + \\ \beta 15 * X1 * X5$$

The model selection was carried out using two splits of 90:10 and 75:25 as well as cross validation using LOOCV and K-fold Cross validation with k=5 and k= 10.

The coefficients are as follows:

$\beta 0$	$\beta 1$	$\beta 2$	$\beta 3$	$\beta 4$	$\beta 5$	$\beta 6$	$\beta 7$
-4.094e+05	5.426e+05	-1.794e+05	5.799e+02	-1.217e-01	-3.394e+01	1.398e-01	1.303e+03

$\beta 8$	$\beta 9$	$\beta 10$	$\beta 11$	$\beta 12$	$\beta 13$
2.763e+01	-1.509e+01	1.392e+00	-2.218e-01	-4.302e+02	7.143e+01

$\beta 14$	$\beta 15$
-1.967e-01	-1.824e+02

The test MSE obtained for the splits, LOOCV and K fold CV are tabulated below.

	Degree 1 + interactions	**Degree 2 + interactions**
90:10 split	1.833122	1.62239
75:25 split	1.891256	1.4024
LOOCV	1.7921632	1.4662305
K fold (k=5)	1.8384132	1.4639924
K fold (k=10)	1.8233764	1.5104314

Table 1 : Summary of test errors using different validation approaches.

Based on the results obtained, it is evident that LOOCV, K=5 & K=10 Cross validation error estimates all pointed towards the model with Degree 2 + interactions as the best model. This was further validated using the validation set approach by using two different splits on the given data at 90:10 & 75:25 split, where former value represents the proportion of the training data set and latter represents the test data. In each of the approach used, Degree 1 + interactions model gave higher test MSE compared to degree 2 + interactions model.

Model 2: Boosting with interaction depth = 2

X4 and X6 have been eliminated from the model as mentioned in model 1.

The model was obtained using boosting technique with an interaction depth of 2. 20000 trees were generated for the purpose of boosting, with the learning level (i.e., shrinkage parameter) set at 0.005. This model was chosen based on least test mean squared error. Three splits were considered to evaluate models based on tree-based methods involving recursive binary splitting, tree pruning, bagging, random forests and boosting. Splits of 90:10 percent, 75:25 percent, and 50:50 percent were considered with the former half of split indicating the percentage of dataset used for training the model. The results for the tree-based models are tabulated below with the selected model results highlighted.

Train to test Split Percentage	Making tree	Pruning Tree	Boosting (interaction depth = 2)	Boosting (interaction depth = 4)
90:10	199.475	5.409257	0.2303339	0.1847711
75:25	163.9445	4.885503	0.1751920	0.1639379
50:50	153.5769	7.041852	0.2443600	0.260074

Table 2 : Summary of errors for tree-based models

From Table 1, we can see that for all the splits considered, boosting results in the **least test MSE.**

Boosting as contrary to random forests and bagging, grows trees sequentially based on the information from previously grown trees. The high performance of this method can be attributed towards its ability to learn slowly in each step(learning speed can be defined, which was set at 0.005). The complexity of the tree here is determined by the interaction depth which controls the number of splits in each tree(i.e., level of variable interactions considered). We can see that the interaction depth of d =2, provides the best model. Although, the error is slightly higher for d=2 in comparison with d=4 for 2 out of the 3 splits of training data, since the error difference is very small, we still go ahead with the d=2 model owing to its better interpretability and reduced chances of overfitting, which is proved when the same model is fit for 50-50 splitting, where the error for depth = 2 is lower than for depth = 4.

From the relative influence plot (Figure 1), we can see that the X1 and X2 are the most influential predictors followed by X5 and X3. These four variables together explain about 90 % of the influence of the predictors on the output. The plots for the output vs the variables X1 and X2 are also provided for reference.

Based on the above results, the best model obtained is:

Boosting (6 predictors, Number of trees = 20000, Shrinkage Parameter = 0.005, Interaction Depth = 2)

This model was then fit on the entire training data set provided in Project Phase I, and upon implementing K-fold cross validation with K=10 folds, the training mean squared error estimate obtained was 0.2217018.

Plots

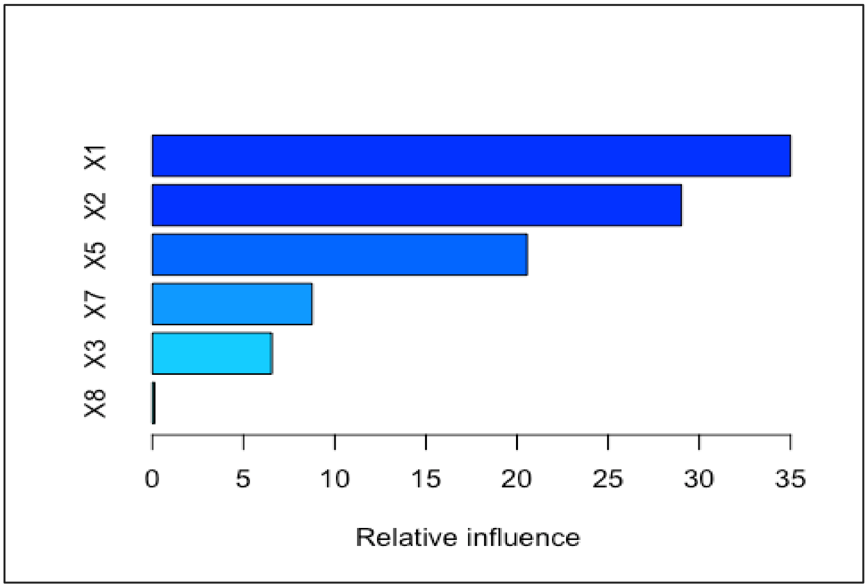

Figure 1 : Relative influence of predictors on the output

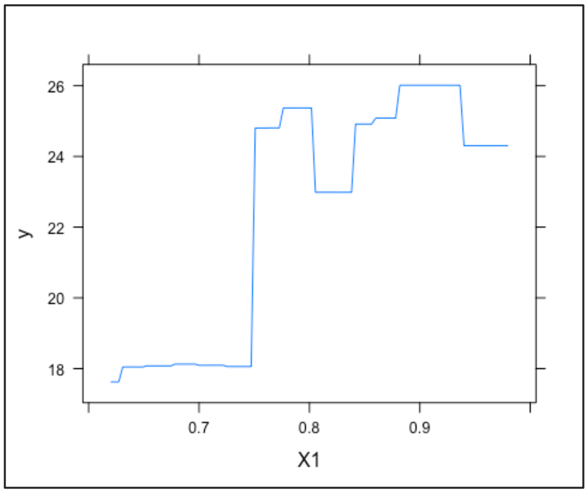

Figure 2 : Y vs X1

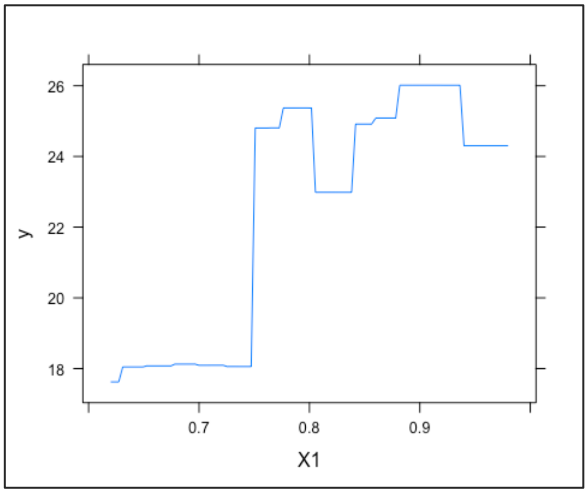

Figure 3 : Y vs X2

Model 3: Bagging (B = 500)

X4 and X6 have been eliminated from the model as mentioned in model 1.

The model was obtained using bagging technique on tree-based models. Model selections were done based on **least test MSE**. Three splits were considered to evaluate models based on tree-based methods involving recursive binary splitting, tree pruning, bagging, random forests and boosting. Splits of 90-10 percent, 75-25 percent, and 50-50 percent were considered with the former half of split indicating the percentage of dataset used for training the model.

The results for the test MSE for the model are highlighted below.

Train to test split percentage	**Bagging(mtry = p = 6)**	**Random Forest(mtry = p/3 = 2)**
90 – 10	0.3022740	0.90098988
75 -25	0.2883616	0.682674
50 -50	0.3836165	0.8560854

Table 3 : Summary of test errors for tree-based models

The bagging method is based on the principle of bootstrap aggregation. Regression trees are generated simultaneously with the predictions being the average of the tree predictions. Since these trees are not pruned, the averaged resulting model has both low bias and variance. Bagging is a special case of random forest method where the number of predictors used at each split is equal to the actual available predictors in the model. The best model is the model obtained at 75:25 split with the lowest test MSE of 0.2883616.

For the model under consideration, 500 trees were generated considering all the predictors. The predicted response and the test response closes match each other as supported by Figure 4. From the importance table, we can see that removing X2, X7 and X8 contribute the most towards the increase in MSE.

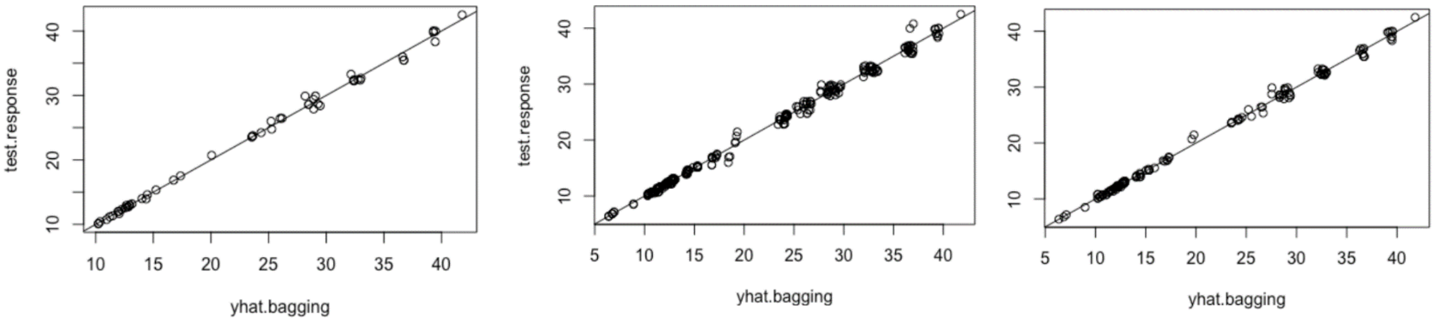

Figure 4 : Test response vs predicted response for splits of 90-10,75-25, 50-50 respectively.

The importance of each variable for the 75:25 split data for the bagging model is:

```
##        %IncMSE      IncNodePurity
## X1   9.465174     3627.871
## X2   59.265103    29020.284
## X3   21.286449    1848.121
## X5   4.499977     1247.325
## X7   59.285411    2044.610
## X8   30.554848    1031.279
```

Comparison of the 3 models:

The 3 best models that were chosen were as follows:

i. **Linear Regression** (Predictors – 6, Interactions – 4, Degree - 2)
ii. **Boosting** (Predictors – 6, Number Of Trees – 20000, Interaction Depth – 2, Shrinkage Parameter – 0.005, K-fold CV – 10 folds)
iii. **Bagging** (Predictors – 6, Number of Trees – 500)

Based on the results of Mean Squared Error estimate obtained by splitting dataset with 75:25 training to test ratio, we can summarize the results as below:

Model	**K=10 K-fold Cross Validation**
Linear Regression	1.4024000
Boosting	0.1751920
Bagging	0.2883616

Table 4 : Summary of test errors for the three best models

Predictors used – X1, X2, X3, X5, X7 & X8 (for all models).

The minimum test MSE was observed for the boosting model mentioned above.

Thus, we have chosen the boosting model as our best model.

The R code for all the boosting model is attached separately along with the submission as a .R file.

Evaluating Test Data on the best model:

The best model mentioned above (Boosting with Predictors – 6, Number Of Trees – 20000, Interaction Depth – 2, Shrinkage Parameter – 0.005, K-fold CV – 10 folds) was trained on the entire train dataset as previously mentioned on page 4 with training error of the model as 0.2217018002.

The new test data provided on April 24th , 2021 was then used to predict the test error rate using the above model.

For the best model, the test error rate that was obtained was 0.6453843.

Summary:

Model	**Boosting**
Predictors	6 (X1, X2, X3, X5, X7, X8)
Interaction Depth	2
Number of trees	20000
Shrinkage Parameter	0.005
K-fold Cross-validation	K=10
Test Error (75-25 Split)	0.1751920
Training Error(Entire train dataset)	0.2217018
Test Error (Test data)	0.6453843

Table 5 : Summary of the best model after evaluating the test data.

Modification of the best model:

The test error obtained on the new data was not satisfactory enough, so we decided to try different variations of the parameters to obtain a model that was better at predicting the test data.

The different parameters tested were as below:

Number of Predictors, Number of Trees, Shrinkage Parameter, Interaction Depth, No. of Folds for K-fold Cross Validation.

The following variations of these parameters were tried:

Number of Predictors	Number of Trees	Shrinkage Parameter	Interaction Depth	K-Fold CV No. of Folds
6	10000	0.1	1	0
7	15000	0.01	2	2
8	20000	0.007	3	5
		0.005	4	10
		0.003	5	
		0.001	6	

Table 6 : Summary of the various parameter settings evaluated.

The best model obtained was with following parameters:

Number of Predictors	Number of Trees	Shrinkage Parameter	Interaction Depth	K-Fold CV No. of Folds
7	10000	0.08397	5	5

Table 7 : Summary of the parameters for the new best model obtained.

The shrinkage parameter was optimized by interpolation of the results obtained from the above tryouts for different test MSEs.

The original best model had only 6 predictors (X1,X2,X3,X5,X7,X8) as X4 was correlated to 3 other predictors (as mentioned on pg.2) and X6 was found to be insignificant on performing linear regression (as mentioned on pg.2). However, for the hyper-grid evaluation of the best model, we tried including X4 as well X6 along the previous 6 predictors (all 6 + X4, X6 and X4&X6) in all different combinations to reach the best model. We found out the best model was obtained when only X6 was included with the previous 6 predictors.

The Mean Squared Error obtained after fitting this model on the entire train data set and then predicting the test error is 0.2081315.

As can be seen from the plot below, X6 has the least influence in the model, i.e., it has the least influence on the predicted value of Y based on the test data, however, the best prediction results were obtained after including X6 in the original model.

The R code for this best model has been attached along with the submission as a .R file. It contains various plots like Error vs #trees, Error vs Interaction Depth, Error vs CV Folds, and Error vs Shrinkage, looking at which we can decipher the best set of parameters to generate the lowest test error.

Warning: Executing the hyper grid code can take significant amount of time and computational power as it computes hundreds of boosting models one after another. It is recommended to only run the code for best improved model.

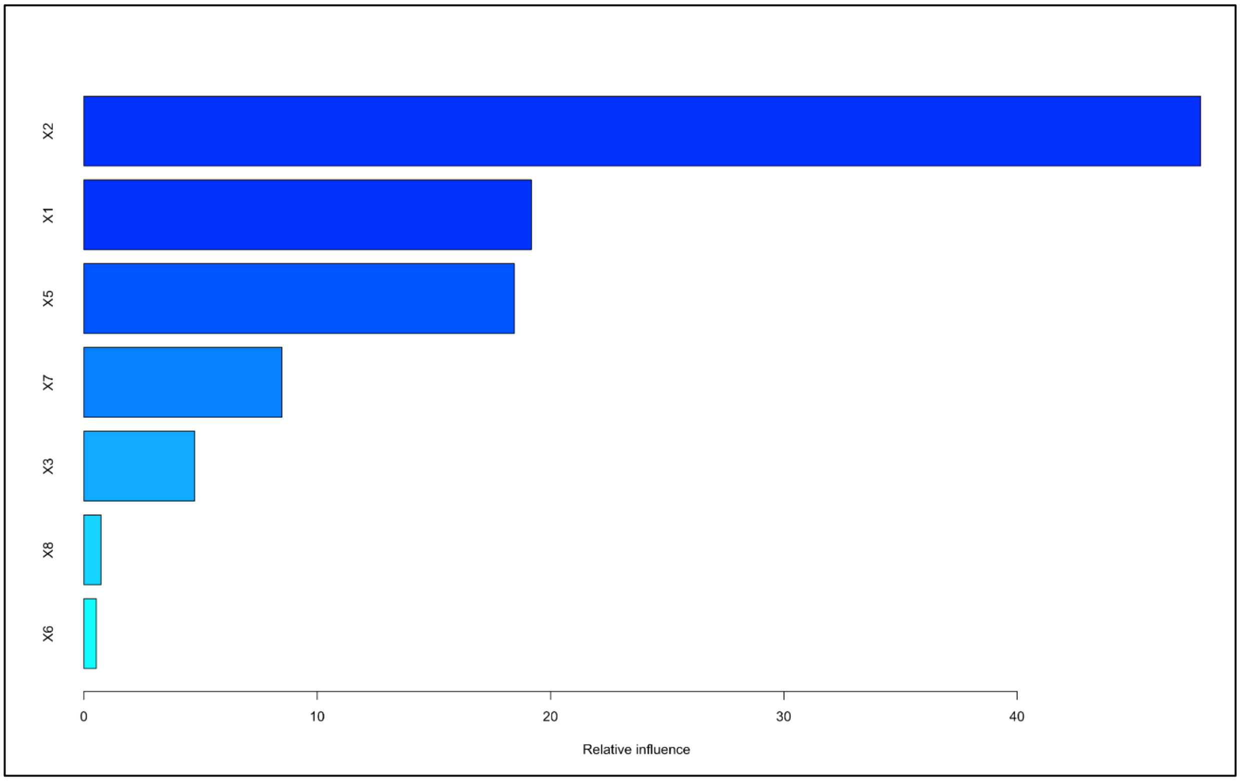

Figure 5 : Relative influence of predictors on the output.

Parameter	**Previous Best Model**	**(Modified) New Best Model**
Number of Trees	20000	10000
Interaction depth	2	5
CV fold	10	5
Shrinkage Parameter	0.005	0.08397
Test MSE	0.6453843	0.2081315
Percentage decrease in Test MSE	67.75076%	

Table 8 : Comparative summary of the best model and improved model.