IE 7280 Spring 2023 Course Project

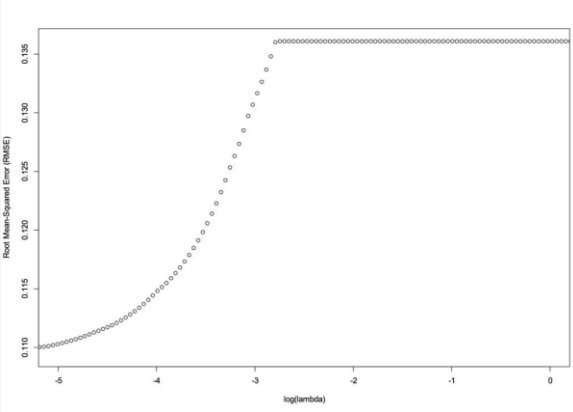
Anjali Patil, Archit Raj, Manwell Hanna, Shriya Kenkre

EXECUTIVE SUMMARY: We reported our best two models as: MLE with all variables and a lasso regression model.

MLE: We implemented the MLE model with training and testing dataset and we got RMSE of 0.122 and adjusted r-square of 0.29 which is similar to our report 1 model where we got RMSE of 0.115 and adjusted r-square of 0.26.

```
> #build model
> model<- lm( log(Y) ~ X1+X2+X3+X4+X5+X6+X7+X8+X9+X10+X11, data = df_TRAIN)
> pred_full <- predict(model, newdata=df_TEST)
> rmse_full <- rmse(log(df_TEST$Y), pred_full)
> rmse_full
[1] 0.1221278
```

Lasso: Next we implemented our final model, lasso regression. First, we loaded the training dataset and testing dataset, placing a specific seed to avoid random data generation on each iteration of the model, while preparing “cross validation” as our hyperparameter tuning technique. Like before, we used a built-in library and applied the cross-validation function to tune the lambda.



When we ran a testing and training dataset on a previously reported model with 10 folds for cross-validation, we got RMSE of 0.17and R-square value as 0.32.

```r
# Predict outcome using model from training data based on testing data
predictions1 <- predict(model1, newdata=test_df)
# Model performance/accuracy
mod1perf <- data.frame(RMSE=RMSE(predictions1, log(test_df$Y)),
                       Rsquared=R2(predictions1, log(test_df$Y)))

# Print model performance/accuracy results
print(mod1perf)
```

Table | Console

"RMSE"	"Rsquared"
0.179841203892169	0.327557090497357

Improvements: For improving our model, we thought of playing around with a number of folds for cross-validation. For 200 k fold cross validation we got RMSE of 0.109 and r square of 0.499.

```
[1] 0.1094276
```

```r
# Create function to identify Rsquared for best lambda,
# where x = Rsquared vector, y = lambda vector, &  z = optimal lambda value
Rsquared_lasso <- function(x, y, z){
  temp <- data.frame(x, y)
  colnames(temp) <- c("Rsquared", "lambda_val")
  rownum <- which(temp$lambda_val==z)
  print(temp[rownum,]$Rsquared) ^Rsquared_lasso
}

# Apply newly created Rsquared_lasso function
Rsquared_lasso(x=model1$results$Rsquared, # x = Rsquared vector
               y=model1$results$lambda,   # y = lambda vector
               z=model1$bestTune$lambda)  # z = optimal lambda value
```

```
[1] 0.4994113
```

After an evaluation of a multiple linear regression model and Lasso regression, we concluded that the Lasso regression model with 200 k folds gives us the best results and is the optimal choice for the given dataset.