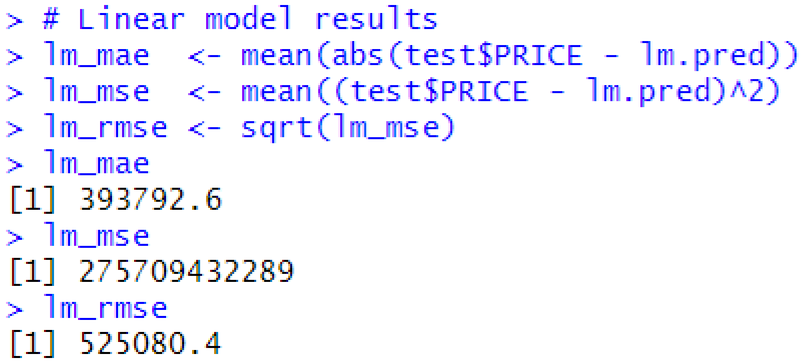
Matthew Voynovich
Data Analytics Lab 6
Professor Eleish

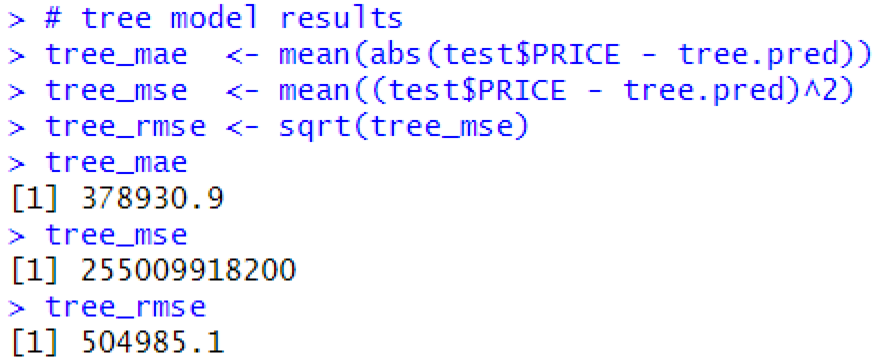
Train 3 regression models each using a different algorithm to predict price from square footage. Evaluate model performance using the MAE, MSE and RMSE metrics.

Linear Model Results



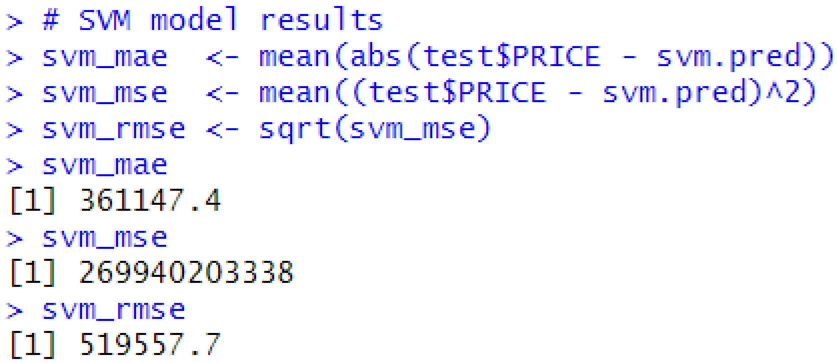
```
> # Linear model results
> lm_mae  <- mean(abs(test$PRICE - lm.pred))
> lm_mse  <- mean((test$PRICE - lm.pred)^2)
> lm_rmse <- sqrt(lm_mse)
> lm_mae
[1] 393792.6
> lm_mse
[1] 275709432289
> lm_rmse
[1] 525080.4
```

Decision Tree Results



```
> # tree model results
> tree_mae  <- mean(abs(test$PRICE - tree.pred))
> tree_mse  <- mean((test$PRICE - tree.pred)^2)
> tree_rmse <- sqrt(tree_mse)
> tree_mae
[1] 378930.9
> tree_mse
[1] 255009918200
> tree_rmse
[1] 504985.1
```

SVM model results



```
> # SVM model results
> svm_mae  <- mean(abs(test$PRICE - svm.pred))
> svm_mse  <- mean((test$PRICE - svm.pred)^2)
> svm_rmse <- sqrt(svm_mse)
> svm_mae
[1] 361147.4
> svm_mse
[1] 269940203338
> svm_rmse
[1] 519557.7
```

Model Comparison

I first started by removing price outliers from the data as I knew that that would heavily skew the models and affect their performance. To prove this I ran the models without doing this outlier removal and saw that my mean average error was in the millions instead of hundreds of thousands. None of these models are particularly good because they are predicting with a singular feature, but the best model is the Decision Tree Regressor. It has the lowest mean absolute error and the lowest mean squared error. The second best model was the SVM with a polynomial kernel while the plain linear model performs the worst. All of the mean absolute errors were in the 300 thousand range with all of the mean squared errors in the 200 billion range and the root mean squared error in the 500 thousand range. This goes to show that despite the decision tree being the best model of the three it should not be used in production and we likely need to add more features to create an accurate predictor.