Lab 3 – Group 1

Training and Testing: (Linear) Predictions

Exercise #1

Working with dataset "Ames":

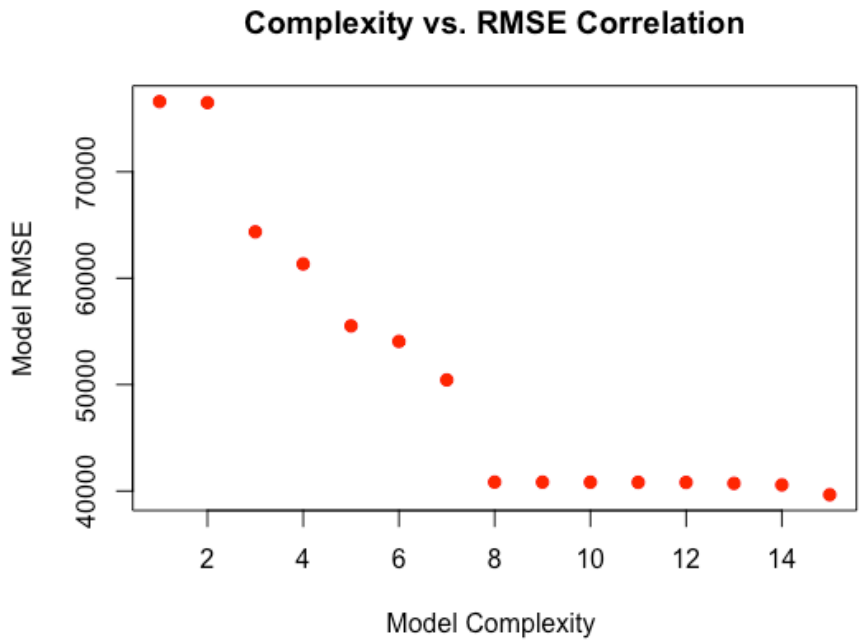
```
Ames <- read.table("https://msudataanalytics.github.io/SSC442/Labs/data/ames.csv",
                   header = TRUE,
                   sep = ",")

NewAmes1 = select(Ames,-18:-19)
```

Utilizing forward selection – a form of stepwise regression, which variables are added one after the other to improve the model:

1. Variables included in the forward selection process are LotArea, MSSubClass, YearBuilt, YearRemodAdd, Fireplaces, BsmtFinSF1, TotalBsmtSF, GrLivArea, FullBath, MoSold, YrSold, MiscVal, PoolArea, ScreenPorch, and GarageArea.

Modeling complexity versus the root-mean-square-error (RMSE):

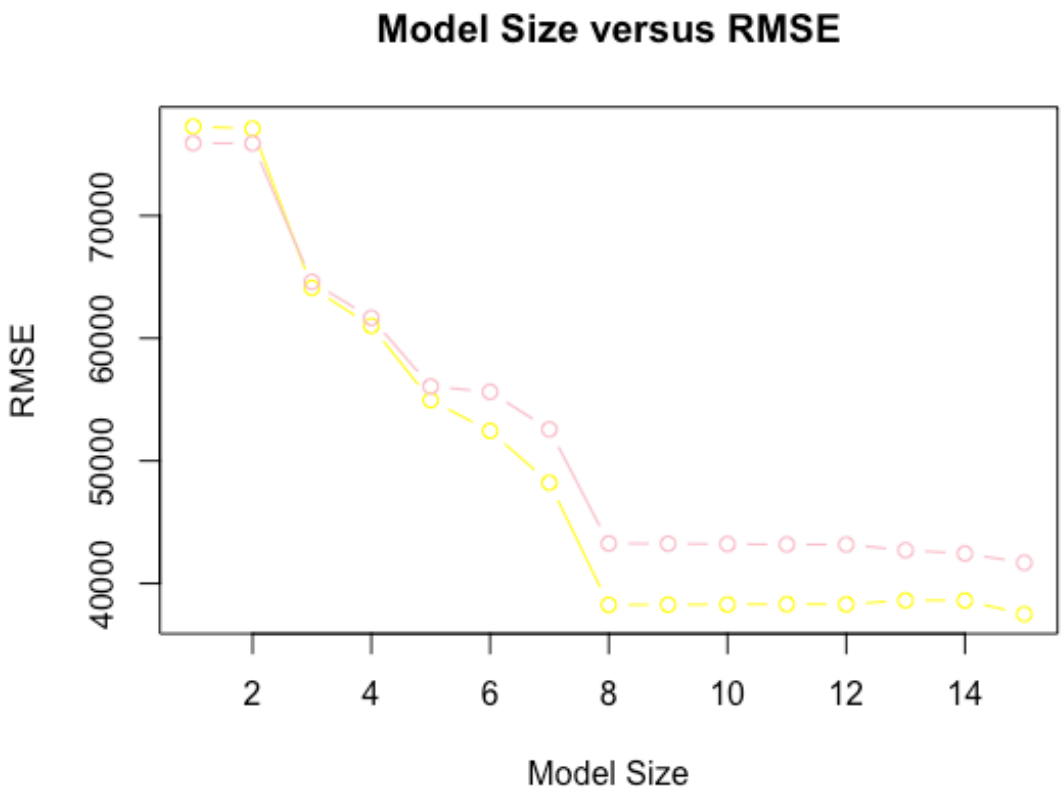


Observations and comments on the above model:

1. As the complexity of the model increases, the root-mean-squared-error steadily decreases.
2. A full model should be used so you are able to see more data and get a more accurate depiction of the relationship between complexity and RMSE. The criterion used to make this statement are the downward trends in Model Complexity, as well as the lack of variables included in the model above. We believe that an even lower RMSE value could be achieved as more variables are used, again making our predictions and analysis more accurate.

Exercise #2

Plot the Train and Test RMSE for the 15 models:



Legend: pink representing the test data and yellow representing the train data.

Predict SalePrice:

1. Using train data:

	fit	lwr	upr
324	147418.62	71969.04	222868.2
36	290783.97	215372.84	366195.1
302	267001.95	191658.91	342345.0
315	161473.16	85414.78	237531.5
647	93557.48	17830.64	169284.3
196	163092.89	87581.73	238604.1

   Train RMSE = 35,861.02

2. Using test data:

	fit	lwr	upr
2	186145.9	110409.456	261882.3
3	222100.8	146738.139	297463.4
5	281786.2	206424.311	357148.1
6	204255.7	125791.313	282720.1
9	167664.9	91111.554	244218.2
10	75292.4	-1017.678	151602.5

   Test RMSE = 39,654.51

Describe the model:

As depicted above, in order to predict SalePrice, we utilized the predict function. First on the whole Ames dataset and then on each individual train and test dataset. We used “prediction” as an interval in the prediction function in order to formulate an upper and lower bounder for the predicted sale price, where an accurate prediction would fall somewhere in between. As well, our fit model included the maximum number of numerical variables that didn’t have NAs as values. That is, an additional 16 regressors (BsmtFinSF2, BsmtUnfSF, X1stFlrSF, X2ndFlrSF, LowQualFinSF, BsmtFullBath, BsmtHalfBath, HalfBath, BedroomAbvGr, KitchenAbvGr, TotRmsAbvGrd, GarageCars, WoodDeckSF, OpenPorchSF, EnclosedPorch, and X3SsnPorch). We started with only 15 regressors tested (fit15) and calculated a train RMSE of 37,500.75 and test RMSE of 41,691.14. However, with the additions making 31 regressors to predict Sale Price our train RMSE went down to 35,861.02 and our test RMSE went down to 39,654.51. As previously stated in our analysis of exercise 1, we believe that due to modifications increasing the complexity (now 31), there now is a lower predicted RMSE. Furthermore, there were no interactions present in our model. Thus, due to the inclusion of more numerical regressors, we believe that our groups prediction will perform well, relative to other groups. We only hope.