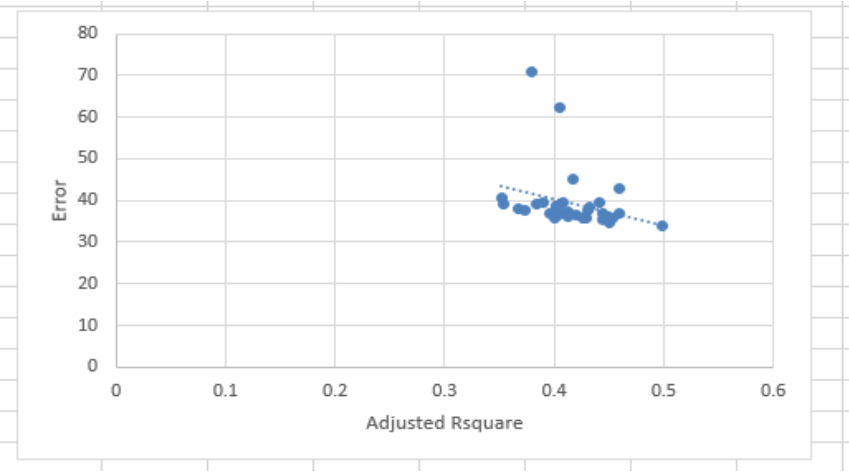
2019SP_MSDS_411-DL_SEC56

Unit 1 - Bonus Assignment - Adjusted R2

Prabhat Thakur

I began by looking at Adjusted R2 and Error data by generating scatter plot to see how both values are related. Adjusted R2 values reported by students are very similar to each other and ranges between 0.3514 and 0.4978. From model's goodness of fit perspective, the variance explained by model in dependent variable by the independent variables are not very impressive and can produce inaccurate predictions (large errors) for new data. Except 2 Error values, all error values are very close to each other. I wonder how two students got very high errors on their test dataset. A possible reason could be not taking care of missing and outlier values or imputing them differently than the training dataset. This gives us some indication that Adjusted R2 value cannot be directly translated to the model accuracy in test dataset or new data.

From the trend line we can see that in general as model Adjusted R2 value increase toward 1, model accuracy also increase (decrease in error) but this is not always true. A model can give a perfect Adjusted R2 value for training data but perform poorly on test data, for example overfitting training data.



- **I want you to tell me this: "IN YOUR OPINION, IS SELF REPORTED ADJUSTED R-SQUARE A GOOD PREDICTOR OF MODEL PERFORMANCE ON NEW DATA?" Let's take it one step further and generalize the question. "DOES PERFORMANCE ON TRAINING DATA DO A GOOD JOB PREDICTING PERFORMANCE ON NEW DATA?".**

  Adjusted R2 should be used as a matric to evaluate how well our model fits the training data and also evaluating different models to select the best model. How well a model will perform on new data should be validated using test dataset derived from the same population sample. Model validation is done by checking how accurate model prediction is when compared to actual value of target (dependent) variable. Comparing error (root mean square error or mean square error) between training dataset and test dataset can reveal if model is under fitting or overfitting training dataset. Based on this, model can be adjusted to achieve better prediction accuracy.

- **Will the insight from this analysis on self reported performance on TRAINING DATA affect or not affect your technique in building predictive models?**

  It certainly does. Objective of any predictively model is to predict target value for new unseen data. How confident we can be about those predictions can be assessed based on model’s prediction accuracy on the test dataset. Hence it is very important to validate the model on test dataset.