Daniel Prusinski
Predict 412

Programming Assignment 2: Evaluating Regression Models in R

In 2011, I successfully proposed to my future wife, which involved an extensive ring/diamond research process. The "Two Month's Salary" data set written by Brian Pope has 425 observations and 7 variables, and would have been very helpful to me during my own diamond exploration period. Price is one of, if not, the key factor(s) when buying a diamond. Through this exploratory data analysis (EDA), I will be predicting price based on the other variables. I will be utilizing multiple regression and tree-structured regression as comparative models. The root-mean squared error will be the metric for evaluating model performance. Carat, Color, Price, and Clarity are all number/integers in the dataset and modeled as continuous. Cut, Channel, and Store are discrete variables, which will work fine for this EDA. All outputs referenced in this document refer to Appendix 1. Output 3 shows a density plot for price and one can see that the data distribution is positively skewed and a data transformation in the form of a logarithmic transformation would be helpful. After the log transformation, one can see that price, the response variable, now follows a more normal distribution.

Linear regression has certain assumptions that must be satisfied in order for the results to be reliable. Linearity is one assumption and is best assessed through a scatter plot showing the response and predictor variable(s). Outputs 5, 6, and 7 show the scatter plots of Carat, Color, and Clarity, and it can be seen that carat has a strong correlative, linear relationship, carat is medium, and clarity is hard to see. Color and clarity may not be strong single predictor variables, but when added to other variables there predictive importance should be magnified. The class variables cut, channel, and store are best expressed as additional variable(s) compared with one of the continuous variables. Moving forward, I have established linearity and will fit the model with linear regression.

The data have been divided into training (2/3) and testing (1/3), which can be seen in output 8.

The principle of parsimony is essential to a cohesive EDA, and stepwise accomplishes this by reviewing each variables strength of correlation with the response variable by analyzing the t-score. A threshold is predetermined such that any given variable must add more value than noise to the model. The training data produced a root mean square error (RMSE) of .2328 (output 9 shows the other model parameters for the regression model). The residuals do not look random, but the Variance Inflation Factor are less than 2, which leads me to believe this model does not suffer from multicollinearity. The variables have statistically significant p-values. After training the data, I tested the data on the test data and the RMSE was .2763, and this is a 19% loss of accuracy.

Tree Regression is a modeling technique that is normally used for datasets that have many predictor variables, but in this instance the model is helpful in comparing two different modeling techniques. Tree regression splits the data with a variable that best divides the data into two polar groups, and then another variable is used to split the subset data. The pruning of variables in Rcart, the method I used, is done using the geometric means for the cp values, which contain the mean, standard deviation of errors of the functions (T. Therneau, Beth Atkinson, Brian Ripley). This process is similar to stepwise, but uses cross-validation. The visual tree and metrics can be seen in output 9. The RMSE for the testing data was 0.2626 and the training data was .2484. As a percent, the difference between training and testing was a 6% loss in accuracy.

Between the two models, the regression tree had a better RMSE on the testing data. From a management standpoint, I would go with the multiple regression model for the following reasons: in my opinion, it is far easier to explain multiple regression with both an R-square value (I would use the adjusted R-squared) as well as the RMSE, management would question why going with a regression tree is necessary when there are few variables, and I feel more comfortable with the multiple regression model because I have in-depth personal experience. Reflecting back on this project, time continues to be a scare resource as I manage learning, installing, and utilizing a new programing language.