**Classifying Diamond Clarity**

**Introduction:**

Natural diamonds are result of carbon exposed to tremendous heat and pressure deep in the earth. This process can result in variety of internal characteristics called “inclusions” and external characteristics called “blemishes”. Evaluating the diamond clarity involves determining the number, size, relief, nature and position of these characteristics, as well as how these effect the appearance of this stone. While no diamond is perfectly pure, the closer it comes, higher the value.

Many inclusions and blemishes are too tiny to be seen by anyone other than a trained grader. To naked eye VS1 and SI2 may exactly look the same, but they quite different in terms of quality. Expensive equipment and trained professionals are required to grade the clarity of the diamond. Due to this the buyer is dependent on the certificate given by the trader and no other means of verifying the diamond clarity. This leads to buyer getting duped or accidently got sold the wrong clarity diamond. This model tries to classify the clarity of the diamond based on other features of the diamond, so the buyer can get a sense of confidence that he is being sold the correct diamond. This model classifies the groups of clarity of the diamond and but does not include several level within each group.

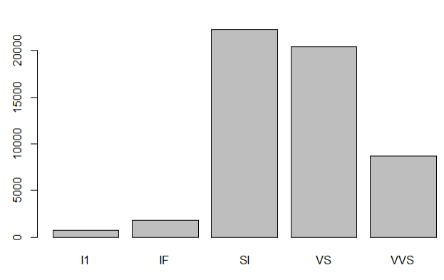
**Dataset:**

The data set contains 10 variables, 3-categorical variables (Cut, Color and clarity) and 7 –continuous variables (Price, Carat, X, Y, Z, Depth and Table). Since the measurement of each continuous variable is a different measurement and different scale, data normalization is required for any methods which uses distance based algorithms. Since the scope of the project to classify the Clarity of diamond based on other features, Clarity is our response variable and rest of the variables are our independent variables.

Below is the explanation of each variable in the data set:

* **Price**: price in US dollars (\$326--\$18,823), **Carat:**  weight of the diamond (0.2--5.01), **Cut:** quality of the cut (Fair, Good, Very Good, Premium, Ideal), **Color:**  diamond color, from J (worst) to D (best)
* **Clarity:** measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best)), **X:**  length in mm (0--10.74), **Y:**  width in mm (0--58.9), **Z**: depth in mm (0--31.8), **Depth:**  total depth percentage = z / mean(x, y) = 2 \* z / (x + y) (43--79) and **Table:** width of top of diamond relative to widest point (43--95)

**Data Exploration:**

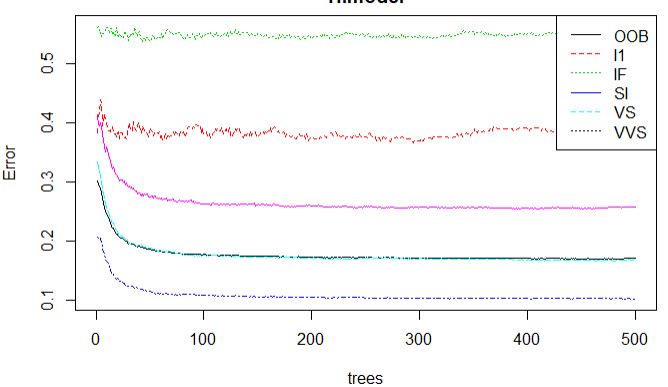
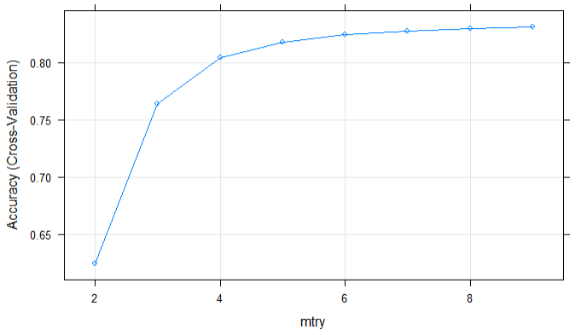
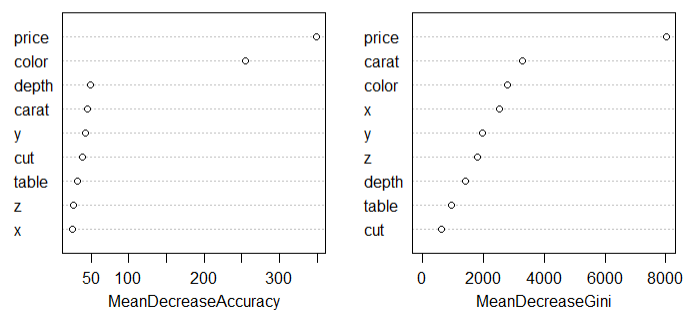
Since the scope of the project is to classify the clarity of the diamond and not classifying levels within each clarity, we will make a new variable where VVS1&2 as VVS, VS1&2 as VS, SI1&2 as SI, L1 and IF. Based on this data set the number of samples are very low for I1 and IF compared to others. There is a class imbalance within the different classes of the sample. Other than that the dataset looks good. No NA’s or missing values in dataset. Based on correlation test there is high correlation between predictor variables x,y,z, carat and price. Since we have a data set which ahs huge number samples (~55k samples) we use divide the data into training set and validation set. We will use training set the train and build the model and validation set to verify how well our model works.

**Data mining:**

In this project we will use to two classification methods to classify the diamond clarity

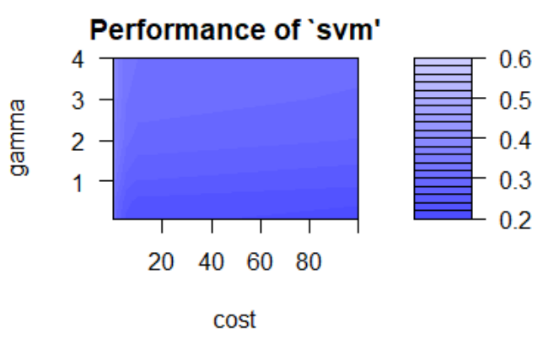
1. **Random Forest:**

Random forest is a supervised non-parametric method which is based on decision tress, which is well suited for multi-class classification problems (response variable which has more than 2 factors). Random forest works well with a mixture of numerical and categorical variables. Since Random forest algorithm is not based on distance, variables with different scales should work well, so scaling is not required. Random forests also work well with large data sets and can efficiently handle high number of variables. It also gives estimates for what variables are important in the classification problem. Random forest performs better when there are strong correlations between predictor variables. Since our predictors have high correlation and our dataset meets all the above feature requirements, Random forest has been chosen as one of the methods to classify Clarity of the diamond in our diamonds data set.

To optimize Random forest to give better quality classification we need to optimize number of trees and number of predictor variables. We use 10 fold cross validation to come up with an optimum number of predictor variables and look at the error plot to get optimum number of trees. Based on the error plot we can see the error saturated right after 100 tress and error for predictor variables saturated after 6, so we chose 100 trees and 6 predictor variables for our model. Based on the optimum model we got an accuracy rate of 84.12%. Color and price has the highest contribution in terms of accuracy and price has the highest contribution to node purity which is given by Gini index. The sensitivity of IF clarity is lower compared to others because of lower samples and most of misclassification was to VVS which is the clarity level one below IF.

1. **Support Vector Machine (SVM):**

SVM is a supervised learning method which can be used for classification problems. It uses a technique called the kernel trick to transform the data and based on these transformations it finds an optimal boundary between the possible outputs. SVM is very close to logistic regression except that, logistic regression performs better for binary classifiers, whereas SVM performs better in multi class problems. SVM also supports non-linear kernel which means that the boundary that the algorithm calculates doesn’t have to be straight line. This helps capture more complex relationships between he data points. The downside of SVM is the training time is much longer as it is computationally more intensive. For the size of diamonds data set to perform cross-validation to get ideal cost and gamma parameters the computation time was >48hrs. Also SVM results are hard to interpret. Since SVM algorithm classifies the variables based on the distance, it is required that we scale the data since the diamonds data set has variables which has different measures and scales.

Both linear and radial kernels were used to classify the clarity of diamonds and 10 fold cross validation was used to get the optimum value for cost function and gamma value used in the model. Radial kernel with cost=100 and Gamma=0.1 provided the best results.

**Conclusion:**

Based on results from random forest and SVM, random forest performed the best. Also per class sensitivity levels were better for random forest.

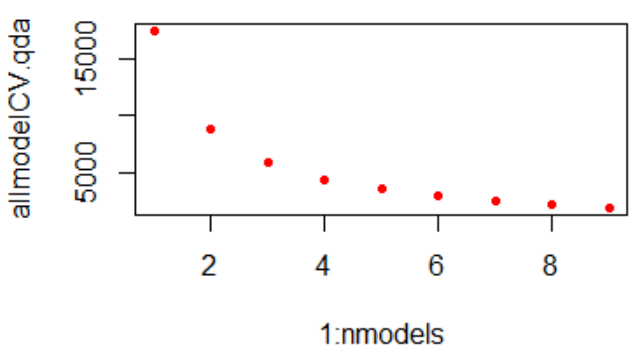
I1 and IF has the least sensitivity among all the other classes. To improve model accuracy collect more samples of I1 and IF since there is huge imbalance in samples. Our Random forest model classifies the diamond clarity with ~84% accuracy. SI accuracy is the best at 90% and IF is the least at 56%.

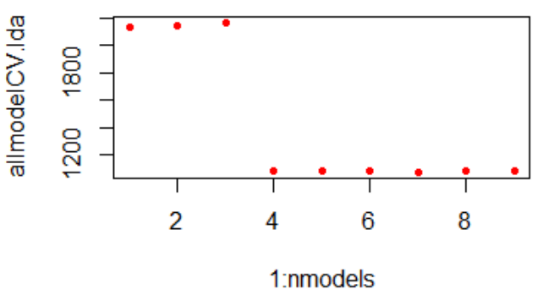
**APPENDIX:**

**LDA and QDA methods were also used, since the scope of the project was to use 2 methods, this has been included in the appendix.**

**Discriminant analysis (LDA and QDA):**

Logistic regression is a classification algorithm traditionally limited to only two class classification problems. If we have more than 2 classes like in diamonds data set, LDA/QDA is the preferred linear classification technique. LDA/QDA is parametric method and standard implementation assumes Gaussian distribution of input variables. Since some of our variables are not normally distributed we will have to transform the variables. LDA also assumes that each input variables has the same variance, so it’s good to scale the data before using LDA so it has a mean of 0 and Standard deviation of 1. QDA uses its own estimate of variance.

 Based on the variable importance from random forest set of 9 models combinations were included to get the best model for LDA and QDA using 10 fold CV. For LDA model with 4 predictors was chosen and for QDA model with 9 predictor variables was chosen.



LDA/QDA performed worse compared to Random forest and SVM with radial kernel.