MET CS 555 Term Project

Student name: Pragya Sen

PROBLEM STATEMENT: This project aims to determine which chemical features are the best indicators of red wine quality and using the most appropriate method to model the same.

DATASET: The Red Wine Quality Dataset(https://archive.ics.uci.edu/dataset/186/wine+quality) was used. While the original dataset was larger, 1000 samples have been used for this project. It contains a total of 12 variables where 11 are chemical factors potentially affecting the dependent variable, i.e, the quality.

Some of the other data preparation steps included:

1. Checking the datatype of all the features

```
> str(df)
'data.frame':   1000 obs. of  12 variables:
 $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
 $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
 $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
 $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
 $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071 ...
 $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
 $ total.sulfur.dioxide: int  34 67 54 60 34 40 59 21 18 102 ...
 $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
 $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...
 $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...
 $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
 $ quality             : int  5 5 5 6 5 5 5 7 7 5 ...
```

Figure1: Datatype of all variables in the dataset

2. Checking for the presence of missing values in the dataset

```
Variables sorted by number of missings:
             Variable Count
        fixed.acidity     0
     volatile.acidity     0
          citric.acid     0
       residual.sugar     0
            chlorides     0
  free.sulfur.dioxide     0
 total.sulfur.dioxide     0
              density     0
                   pH     0
            sulphates     0
              alcohol     0
              quality     0
```

Figure2: Missing value check

Hence, there are no missing values.

© 2021 Trustees of Boston University. Materials contained within this course are subject to copyright protection.

3. Summary of the dataset to understand the distribution of data

```
> summary(df)
 fixed.acidity    volatile.acidity  citric.acid     residual.sugar     chlorides
 Min.   : 4.600   Min.   :0.1200   Min.   :0.0000   Min.   : 1.200   Min.   :0.01200
 1st Qu.: 7.400   1st Qu.:0.4000   1st Qu.:0.1200   1st Qu.: 2.000   1st Qu.:0.07200
 Median : 8.300   Median :0.5200   Median :0.2800   Median : 2.300   Median :0.08100
 Mean   : 8.729   Mean   :0.5283   Mean   :0.2946   Mean   : 2.579   Mean   :0.09037
 3rd Qu.: 9.800   3rd Qu.:0.6350   3rd Qu.:0.4700   3rd Qu.: 2.700   3rd Qu.:0.09300
 Max.   :15.900   Max.   :1.3300   Max.   :1.0000   Max.   :15.500   Max.   :0.61100
 free.sulfur.dioxide total.sulfur.dioxide    density             pH          sulphates
 Min.   : 1.00       Min.   :  6.00       Min.   :0.9906   Min.   :2.740   Min.   :0.3300
 1st Qu.: 7.00       1st Qu.: 23.00       1st Qu.:0.9964   1st Qu.:3.190   1st Qu.:0.5600
 Median :13.00       Median : 39.00       Median :0.9973   Median :3.300   Median :0.6200
 Mean   :15.17       Mean   : 48.33       Mean   :0.9973   Mean   :3.299   Mean   :0.6685
 3rd Qu.:20.25       3rd Qu.: 64.25       3rd Qu.:0.9984   3rd Qu.:3.400   3rd Qu.:0.7400
 Max.   :68.00       Max.   :165.00       Max.   :1.0032   Max.   :3.900   Max.   :2.0000
    alcohol         quality
 Min.   : 8.40   Min.   :3.000
 1st Qu.: 9.50   1st Qu.:5.000
 Median : 9.90   Median :5.000
 Mean   :10.24   Mean   :5.594
 3rd Qu.:10.80   3rd Qu.:6.000
 Max.   :14.90   Max.   :8.000
```

Figure3: Summary of available dataset

4. Checking if data is normally distributed

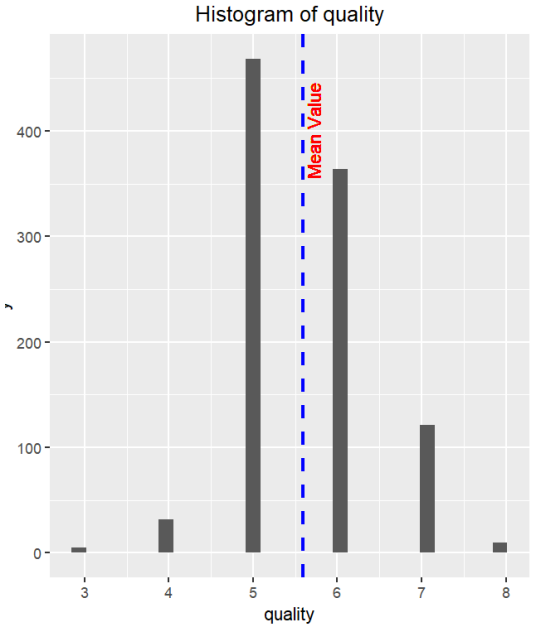

Figure4: Histogram of quality variable

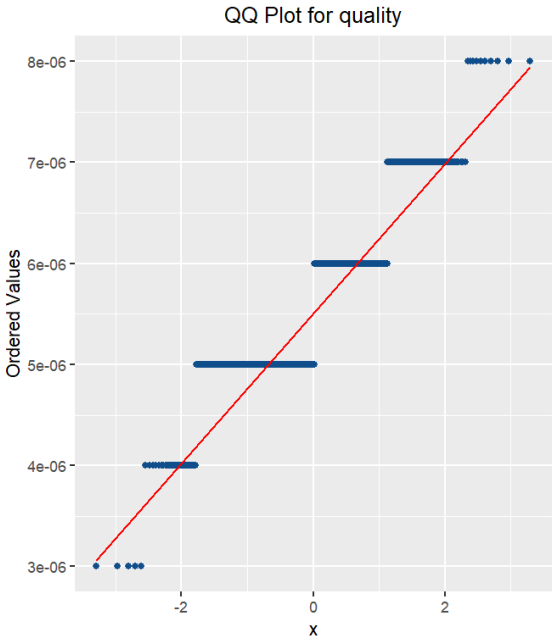

Figure5: QQ Plot for quality Variable

The histogram and the QQ-plot indicate that the data is approximately normally distributed but there seems to be some degree of imbalance (certain classes seem to be more frequent than others).

5. Splitting Data

Before modelling the data, the dataset was split into **Training(70%), Testing(15%) and Validation(15%)** subsets.

STATISTICAL METHODS USED: Correlation Matrix, Histograms, QQ-plots, Simple Linear Regression, Multiple Linear Regression, Lasso Method (L1 regularization), Random Forest

© 2021 Trustees of Boston University. Materials contained within this course are subject to copyright protection.

RESULTS:

Correlation of each factor to quality (in decreasing order) :

```
> head(absoutcome_cor[order(absoutcome_cor, decreasing = TRUE)],12)
             quality              alcohol     volatile.acidity total.sulfur.dioxide
          1.00000000           0.48420787           0.34670861           0.24657487
         citric.acid            sulphates        fixed.acidity              density
           0.21961379           0.21631821           0.17234368           0.11357515
           chlorides  free.sulfur.dioxide       residual.sugar                   pH
           0.10435097           0.09269870           0.06974019           0.05733762
```

Figure6: Correlation between all predictors and the quality variable

It can be seen that the top-4 factors with the highest correlation to the **quality** are: **alcohol, volatile.acidity, total.sulfur.dioxide, citric.acid**

Correlation Matrix of all factors with each other :

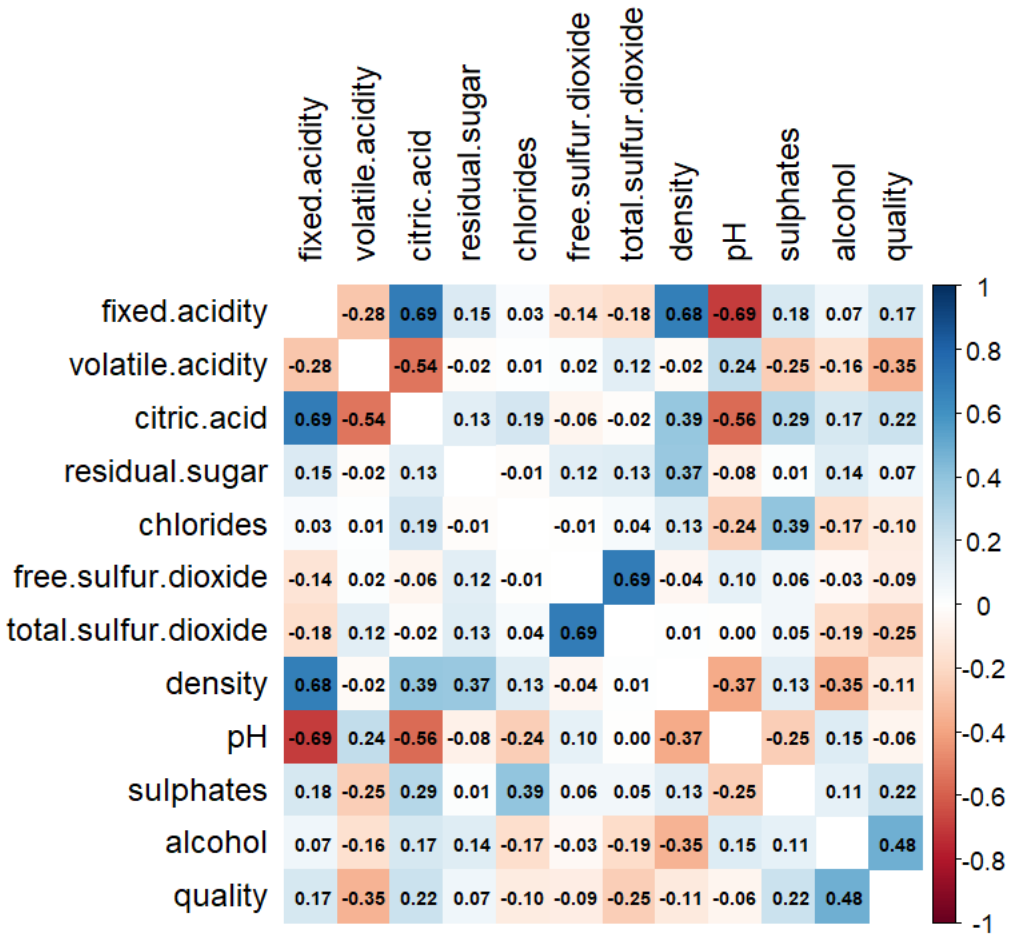

Figure7: Correlation Matrix of all predictors

From the correlation matrix, we note the correlation between the top-4 factors:

- alcohol & volatile.acidity = -0.16

© 2021 Trustees of Boston University. Materials contained within this course are subject to copyright protection.

- alcohol & total.sulfur.dioxide = -0.19
- alcohol & citric.acid = 0.17
- volatil.acidity & total.sulfur.dioxide = 0.12
- volatile.acidity & citric.acid = -0.54
- total.sulfur.dioxide & citric.acid = -0.02

It can be seen amongst the 4 variables, only **volatile.acidity** & **citric.acid** have a significant magnitude of correlation.

Model 1: <u>Simple Linear Regression (SLR)</u>:

Using only 1 predictor variable, i.e, **alcohol** (highest correlation with **quality**).

```
> summary(linear_model)

Call:
lm(formula = quality ~ alcohol, data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8757 -0.3716 -0.1777  0.5426  2.1243

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.61006    0.26210   6.143 1.36e-09 ***
alcohol      0.38779    0.02539  15.275  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6987 on 699 degrees of freedom
Multiple R-squared:  0.2503,    Adjusted R-squared:  0.2492
F-statistic: 233.3 on 1 and 699 DF,  p-value: < 2.2e-16
```

Figure8: Summary of Simple Linear Regression model

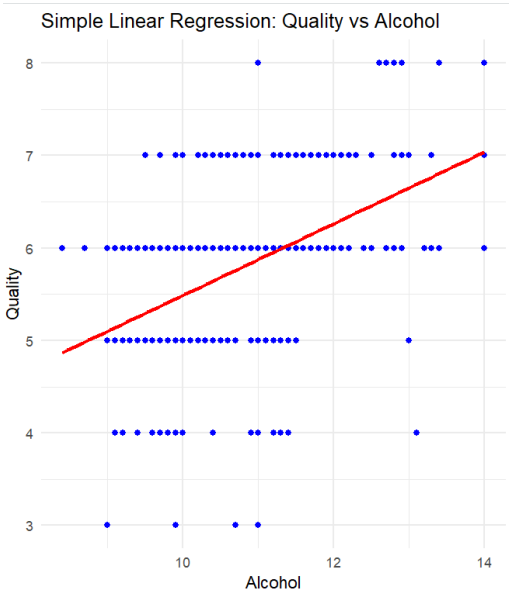

Figure9: Prediction vs Actual plot of SLR model

© 2021 Trustees of Boston University. Materials contained within this course are subject to copyright protection.

RMSE values:

```
> cat("RMSE for Simple Linear Regression (Training Data):", slr_rmse, "\n")
RMSE for Simple Linear Regression (Training Data): 0.6976873

RMSE for SLR on Test Data:  0.7176445
> # Calculate RMSE for Validation Data
> slr_validation_rmse <- sqrt(mean((slr_validation_pred - validation_data$quality)^2))
> cat("RMSE for SLR on Validation Data: ", slr_validation_rmse, "\n")
RMSE for SLR on Validation Data:  0.6841933
```

Figure10: RMSE values of training, testing and validation data for SLR

Model 2: Multiple Linear Regression (MLR):

Using the top 4 variables with highest correlation (i.e., **alcohol, volatile.acidity, total.sulfur.dioxide, citric.acid**) to the quality variable.

```
> summary(multiple_linear_model)

Call:
lm(formula = quality ~ alcohol + volatile.acidity + total.sulfur.dioxide +
    citric.acid, data = train_data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.57586 -0.38879 -0.06715  0.47723  2.02233

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          2.8399431  0.2881354   9.856  < 2e-16 ***
alcohol              0.3392839  0.0246367  13.771  < 2e-16 ***
volatile.acidity    -1.1827220  0.1672977  -7.070 3.79e-12 ***
total.sulfur.dioxide -0.0026919  0.0007741  -3.478 0.000538 ***
citric.acid          0.0633633  0.1466061   0.432 0.665729
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6575 on 696 degrees of freedom
Multiple R-squared:  0.3389,    Adjusted R-squared:  0.3351
F-statistic: 89.19 on 4 and 696 DF,  p-value: < 2.2e-16
```

Figure11: Summary of Multiple Linear Regression model

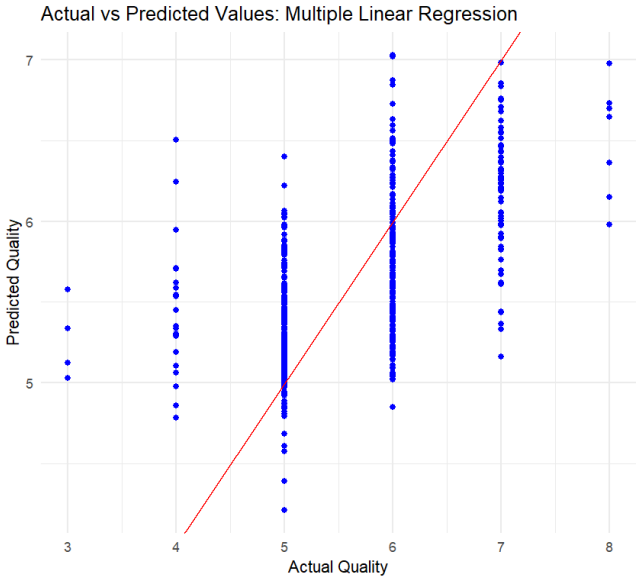

Figure12: Prediction vs Actual plot of MLR model

© 2021 Trustees of Boston University. Materials contained within this course are subject to copyright protection.

RMSE values:

```
> cat("RMSE for Multiple Linear Regression (MLR):", mlr_rmse, "\n")
RMSE for Multiple Linear Regression (MLR): 0.65516

RMSE for MLR on Test Data:  0.6950011
> # Calculate RMSE for Validation Data
> mlr_validation_rmse <- sqrt(mean((mlr_validation_pred - validation_data$quality)^2))
> cat("RMSE for MLR on Validation Data: ", mlr_validation_rmse, "\n")
RMSE for MLR on Validation Data:  0.6192866
```

Figure13: RMSE values of training, testing and validation data for MLR

Model 3: Multiple Linear Regression with Lasso Technique:

Using the Lasso technique to eliminate the effect of correlation within variables. It aims to shrink coefficients of some features to zero, effectively performing feature selection.

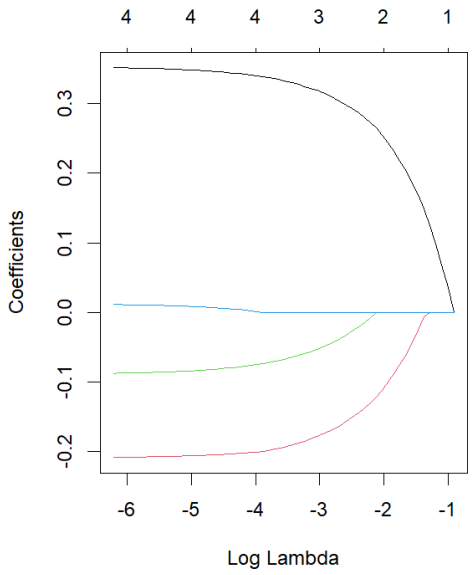

Figure14: Lasso path for MLR

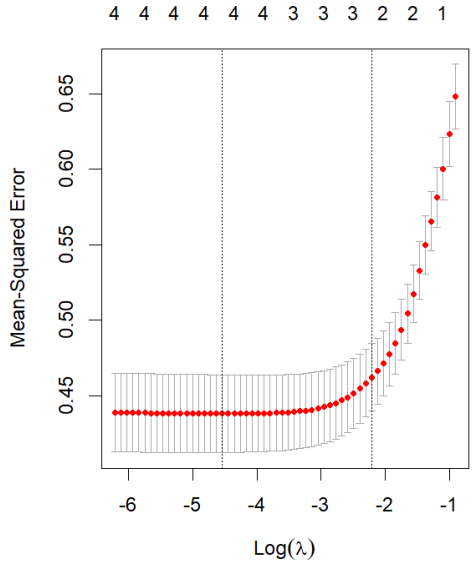

Figure15: CV-plot

```
> cat("Best lambda: ", best_lambda, "\n")
Best lambda:  0.01070671

> print(lasso_coefficients)
5 x 1 sparse Matrix of class "dgCMatrix"
                              s1
(Intercept)           5.593437946
alcohol               0.345490943
volatile.acidity     -0.203880634
total.sulfur.dioxide -0.080734998
citric.acid           0.006177788
```

Figure16: Best Lambda value & Lasso Coefficients

© 2021 Trustees of Boston University. Materials contained within this course are subject to copyright protection.

It can be seen that the magnitude of the **coefficient for citric.acid is extremely small** indicating that it will have **minimal impact** on the model.

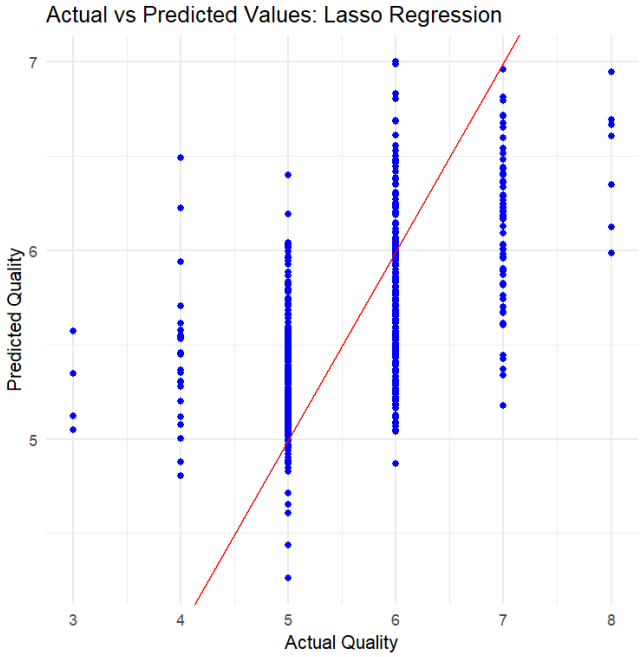

Figure17: Prediction vs Actual plot with Lasso method

```
Adjusted R-squared for Lasso (Training Data): 0.3346372

> cat("RMSE for Lasso Regression (Training Data):", lasso_rmse, "\n")
RMSE for Lasso Regression (Training Data): 0.6553805

RMSE for Lasso on Test Data:  0.6947213
> # Calculate RMSE for Validation Data
> lasso_validation_rmse <- sqrt(mean((lasso_validation_pred - validation_data$quality)^2))
> cat("RMSE for Lasso on Validation Data: ", lasso_validation_rmse, "\n")
RMSE for Lasso on Validation Data:  0.6186442
```

Figure18: RMSE values of training, testing and validation data for MLR (with L1 regularization)

Model 4: <u>Random Forest</u>:

Since there is some degree of imbalance in the data, the Random Forest model could be a good fit. It also helps with feature selection which could handle the problem of correlation between 2 predictors.

```
> print(rf_model)

Call:
 randomForest(x = X_train, y = y_train, ntree = 100, importance = TRUE)
               Type of random forest: regression
                     Number of trees: 100
No. of variables tried at each split: 1

          Mean of squared residuals: 0.3751017
                    % Var explained: 42.23
```

Figure19: Summary of Random Forest model

© 2021 Trustees of Boston University. Materials contained within this course are subject to copyright protection.

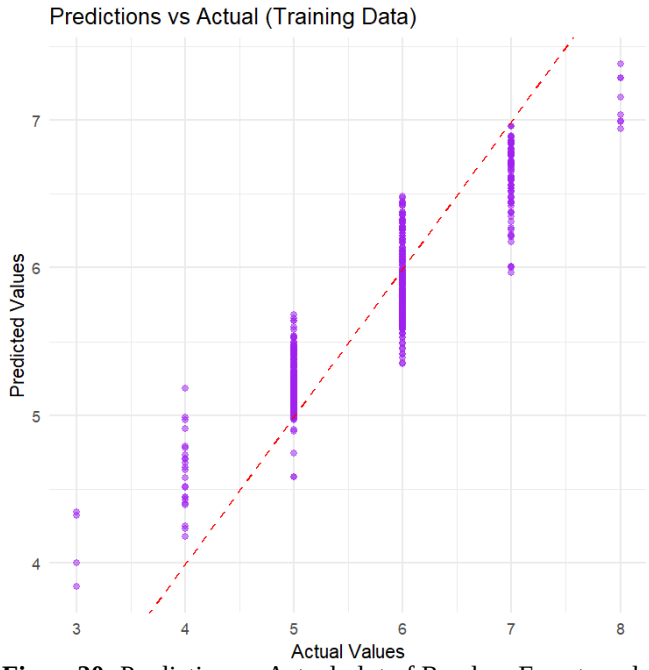

Figure20: Prediction vs Actual plot of Random Forest model

```
> cat("R-squared for the Random Forest model (training data):", r_squared, "\n")
R-squared for the Random Forest model (training data): 0.835848

> cat("RMSE for Random Forest (Training Data):", rf_rmse, "\n")
RMSE for Random Forest (Training Data): 0.3264613
RMSE for Random Forest on Test Data:  0.6444687
> # Calculate RMSE for Validation Data
> rf_validation_rmse <- sqrt(mean((rf_validation_pred - y_validation)^2))
> cat("RMSE for Random Forest on Validation Data: ", rf_validation_rmse, "\n")
RMSE for Random Forest on Validation Data:  0.5915996
```

Figure18: RMSE values of training, testing and validation data for Random Forest model

```
> print(importance(rf_model))
                      %IncMSE IncNodePurity
alcohol              29.80084     135.77314
volatile.acidity     18.49965     102.98832
total.sulfur.dioxide 13.31841      83.12335
citric.acid          13.12491      75.38987
```

Figure19: Feature importance in Random Forest model

Comparing all 4 models using certain evaluation metrics:

	Model 1	**Model 2**	**Model 3**	**Model 4**
R^2 (training)	0.2492	0.3351	0.3346	0.8358
RMSE (training)	0.6976	0.6551	0.6553	0.3264
RMSE (testing)	0.7176	0.6950	0.6947	0.6444
RMSE (validation)	0.6841	0.6192	0.6186	0.5915

Table1: Model comparison using various metrics

© 2021 Trustees of Boston University. Materials contained within this course are subject to copyright protection.

CONCLUSIONS & LIMITATIONS:

As can be seen from Table1, Model 4 (i.e., **Random Forest**) has the **highest R^2** value and the **lowest RMSE** (Root Mean Squared Error) values for training, testing and validation data. A higher R^2 generally implies that the model has good predictive power and explains most of the variability in the target variable. RMSE provides an indication of how well the model's predictions align with the actual observed values. Hence, we can say that the **Random Forest model is the best fit for the Red-wine dataset**.

It can also be concluded that the most influential features in predicting red-wine quality are: **alcohol, volatile.acidity, total.sulfur.dioxide.** It can be seen in Figure16 and Figure19 that amongst the top-4 features, **citric.acid** does not contribute much to the model. This is likely due to the high correlation between volatile.acidity & citric.acid. **Multicollinearity** occurs when two or more independent variables in the model are highly correlated with each other. This implies that the variables contain redundant information and are not providing unique contributions to the model. Multicollinearity can exacerbate **overfitting**.

One of the limitations of this analysis is that there is some degree of imbalance in the dataset. A majority of the quality values were “regular” (5 and 6), which made no significant contribution to finding an optimal model. These values made it slightly harder to identify each factor's different influence on a "high" or "low" quality of the wine.

Another limitation could be the possibility of overfitting. The Random Forest model was chosen to be the best fit for this dataset. But if looked closely in Table1, it can be seen that the RMSE value of Model4 for the training data is significantly smaller than the testing data. This could be an indicator of overfitting. Regardless, the Random Forest model appears to be the best choice since the RMSE values, even for testing and validation data, for Model 4 were significantly better than the other models.

© 2021 Trustees of Boston University. Materials contained within this course are subject to copyright protection.