HW-4

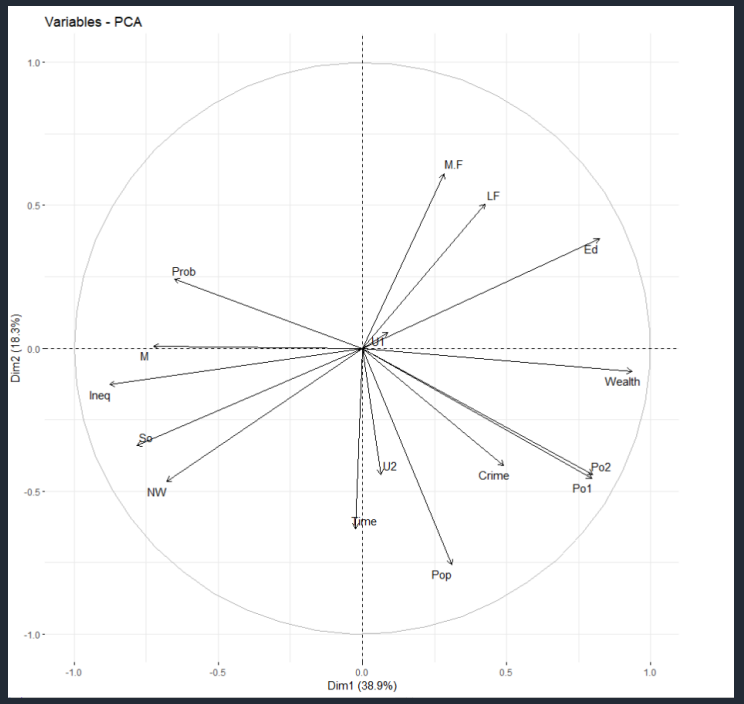
Responses & Summary

Question 9.1

- Calculated the PCA using prcomp library, and further used factoextra to visualize the PCA's.
- Created a linear model based on scaled data.

Complete Repository available on GitHub:

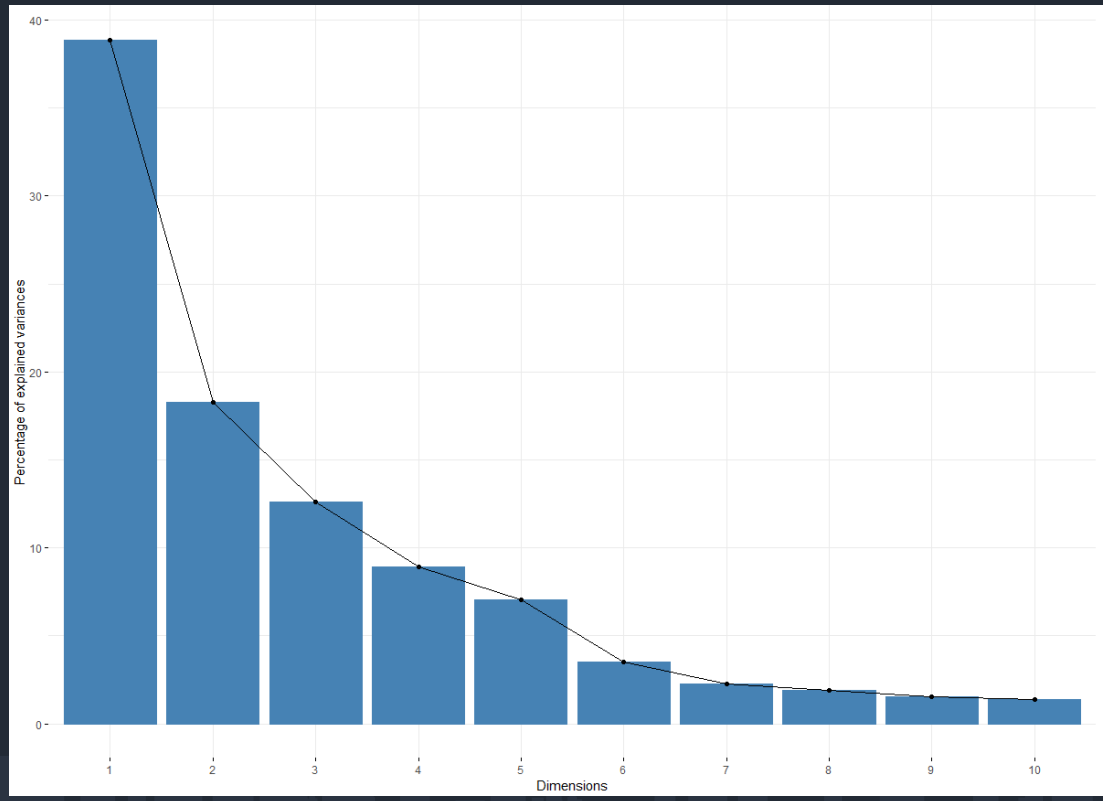
https://github.com/Hizzyth/GTX_Introduction-to-Analytics-Modelling




```
> summary(crime.pca.lm)

Call:
lm(formula = crime$Crime ~ crime.pca$x[, 1])

Residuals:
    Min      1Q  Median      3Q     Max
-603.40 -269.20  -30.52  224.37  741.55

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)        905.09      49.74  18.197  < 2e-16 ***
crime.pca$x[, 1]    75.89      20.16   3.765  0.00048 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 341 on 45 degrees of freedom
Multiple R-squared:  0.2396,    Adjusted R-squared:  0.2227
F-statistic: 14.18 on 1 and 45 DF,  p-value: 0.0004802
```

Question 9.1

- Based on eigen values it became clear that 6th Order of PCA's accounts for majority of expected variance. No major change after that.
- Based on explained variance, picking up PCA from 1 to 6.
- Unclear on Backtransforming data and hence couldn't proceed with predicting the new city.




```
> crime.pca.lm_multiple <- lm(crime$Crime~crime.pca$x[,1:6])
> summary(crime.pca.lm_multiple)

Call:
lm(formula = crime$Crime ~ crime.pca$x[, 1:6])

Residuals:
    Min      1Q  Median      3Q     Max
-284.59  -86.88   12.08   74.59  293.30

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           905.085     20.141  44.938  < 2e-16 ***
crime.pca$x[, 1:6]PC1  75.891      8.162   9.299 1.51e-11 ***
crime.pca$x[, 1:6]PC2 -92.650     11.898  -7.787 1.54e-09 ***
crime.pca$x[, 1:6]PC3  40.535     14.329   2.829  0.00727 **
crime.pca$x[, 1:6]PC4 -212.374    17.024 -12.475 2.28e-15 ***
crime.pca$x[, 1:6]PC5  51.545     19.145   2.692  0.01031 *
crime.pca$x[, 1:6]PC6 -46.415     27.113  -1.712  0.09466 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 138.1 on 40 degrees of freedom
Multiple R-squared:  0.8892,    Adjusted R-squared:  0.8725
F-statistic: 53.48 on 6 and 40 DF,  p-value: < 2.2e-16
```

Complete Repository available on GitHub:

https://github.com/Hizzyth/GTX_Introduction-to-Analytics-Modelling

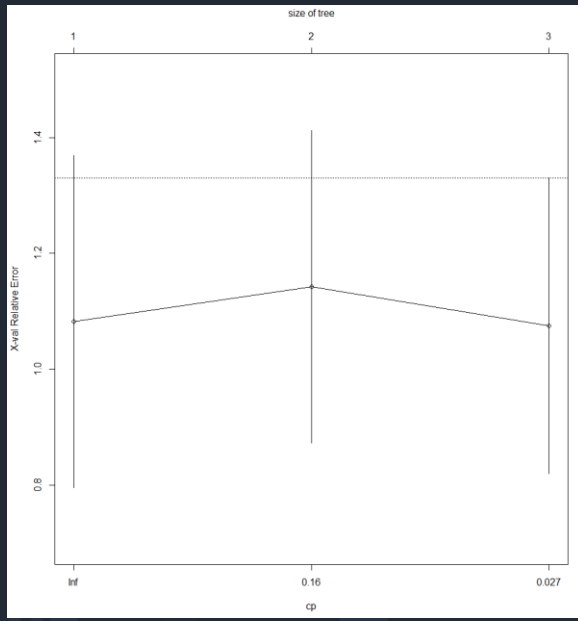
Question 10.1

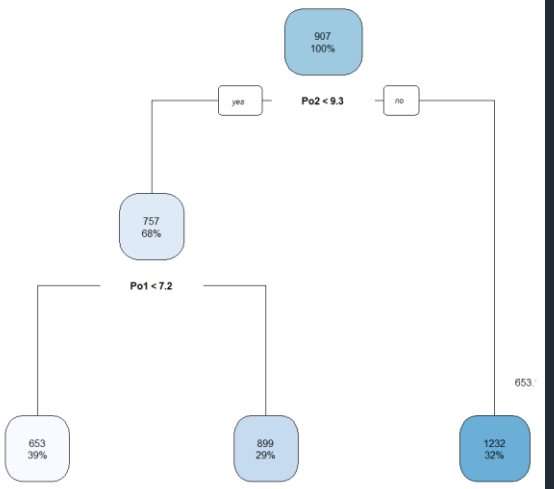
- Data split into training and testing dataset
- Using Rpart library, created the regression tree model (on training set). At this stage the tree is not pruned.

- Based on this 1st Prediction model was created and RMSE calculated against training set
  - RMSE _ 1st Model : 172.811

- Function created to optimize the number of branches
- Based on that Tree was pruned.
- Prediction was run and found that the RMSE went up.
  - RMSE _ 2nd Model : 337.7369

- Hence out of these 2 will continue to use model 1.

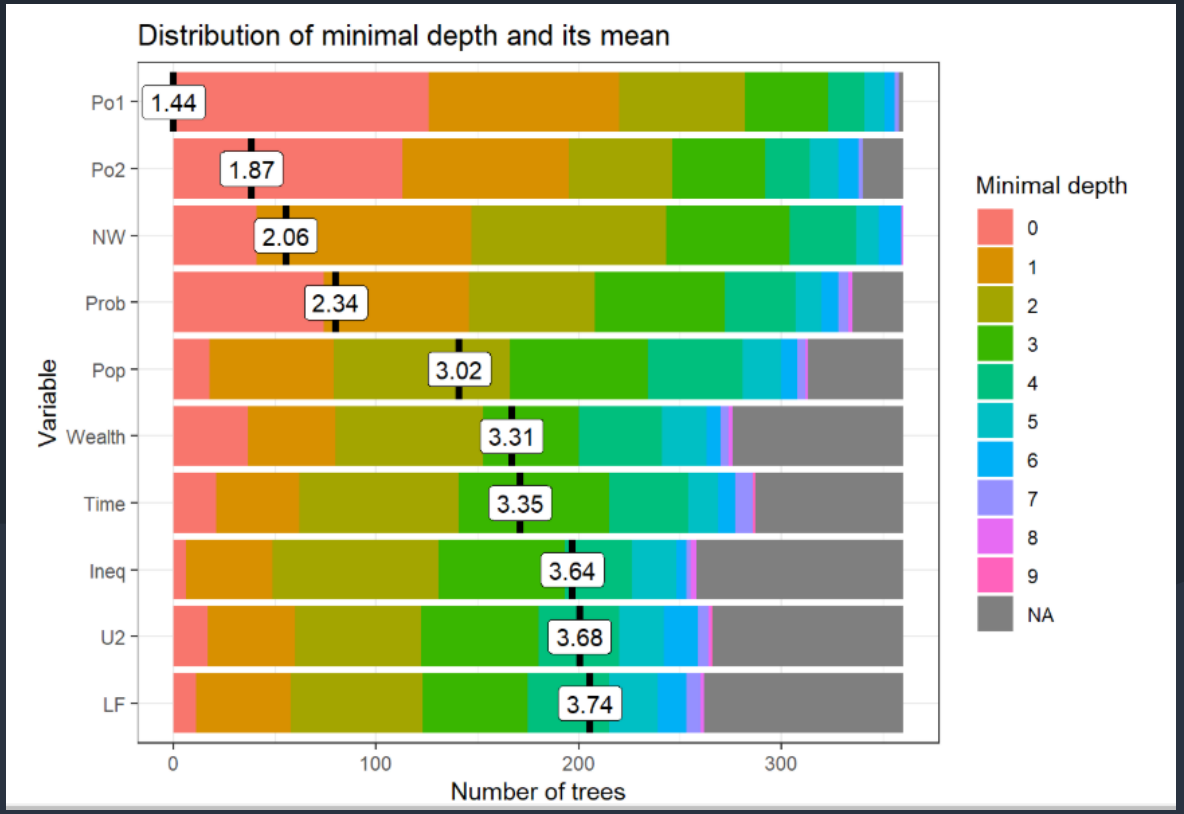





```
get_cp <- function(x) {
  min    <- which.min(x$cptable[, "xerror"])
  cp <- x$cptable[min, "CP"]
}
```

Question 10.1 (Random Forest Method)

- Created the Random Forest Model and visualized the distribution of minimal depth and mean.
- From this visualization it was clear that up to 200 trees are enough and we don't have to consider all 500

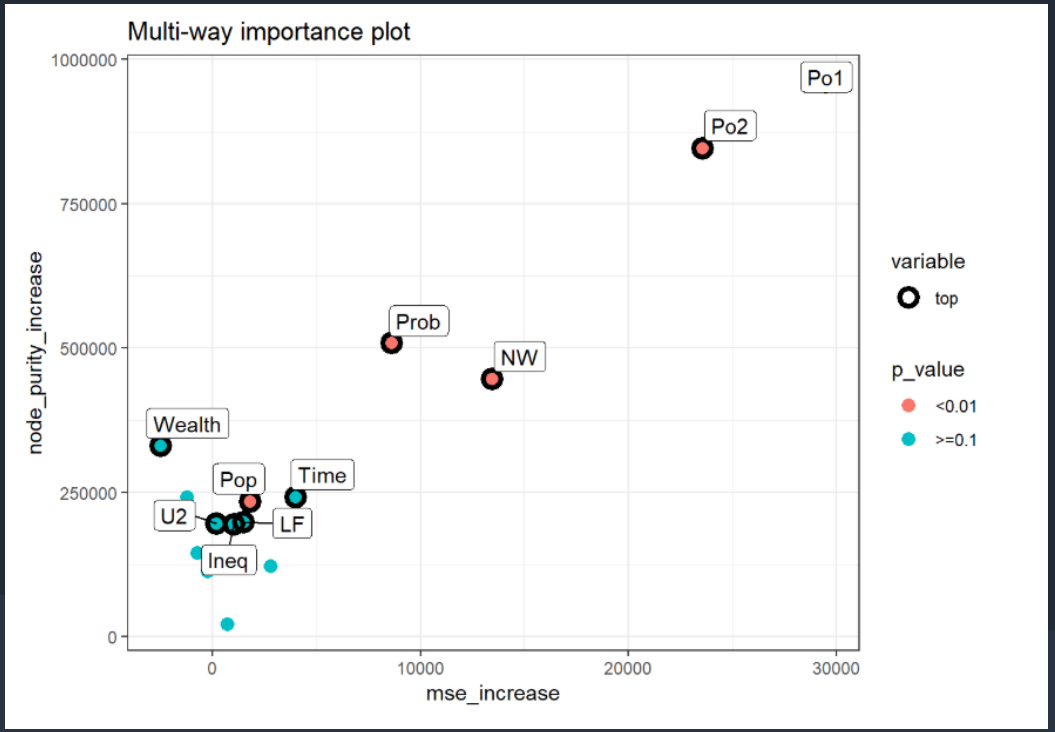


https://github.com/Hizzyth/GTX_Introduction-to-Analytics-Modelling

Question 10.1 (Random Forest Method)

- Multi way importance chart highlighted which parameters were most important and color coded with P-value.

- Based on analyzing both Random forest and Regression Tree model, qualitative takeaways:
  - Police Expenditure is highly correlated to crime rate
  - Also higher probability of imprisonment lead to higher crime rates



https://github.com/Hizzyth/GTX_Introduction-to-Analytics-Modelling

QUESTION 10.2 : Logistic Regression

Question: Describe a situation or problem from your job, everyday life, current events, etc., for which a logistic regression model would be appropriate. List some (up to 5) predictors that you might use.

Answer:

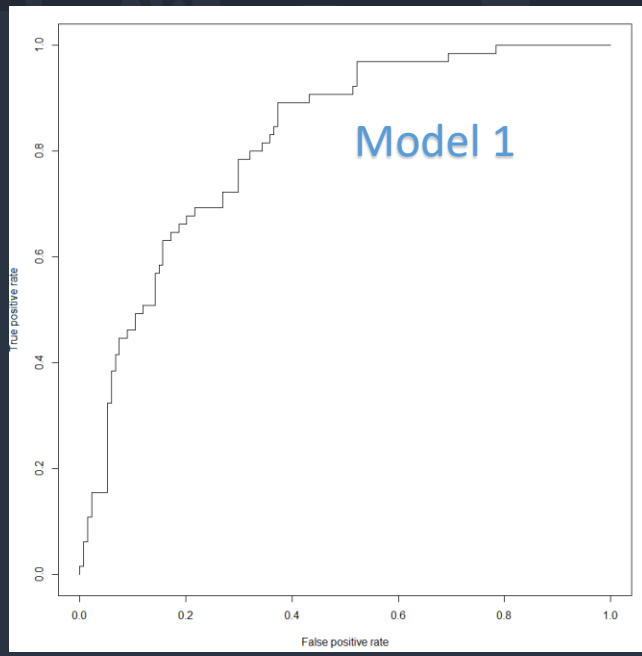
Prognostic health monitoring of equipment's, if they will continue to work fine or fail after being used. Predictors:

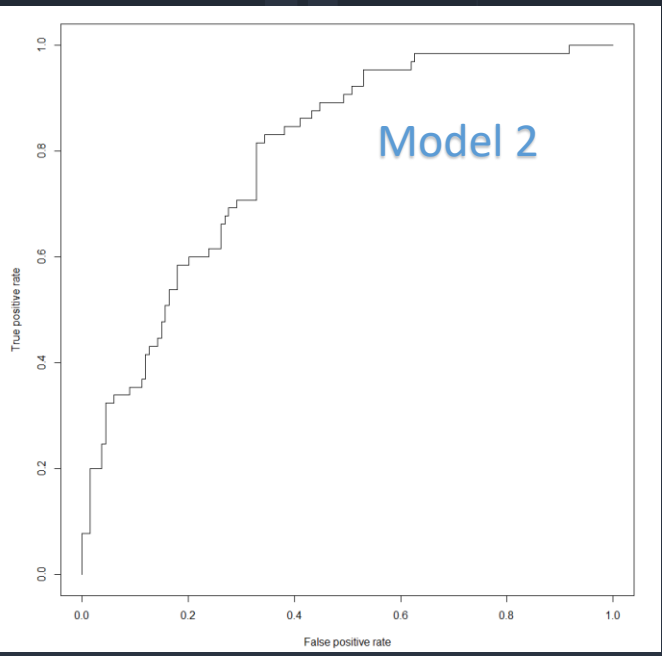
1. Operating Temperature
2. Operating Pressure
3. Voltage consumption
4. Physical deformation
5. Electronic Circuit status

QUESTION 10.3 : Logistic Regression (German Credit data)

- Split the data into training and testing
- Tried creating a logit model right away but error popped up for y values must be 0<=y<=1
- Converted the response (V21) to 1 & 0, successfully created 1st model.
- High degree of deviance from 1st model observed, decided to refine the model
- After refining compared both models and found deterioration in deviance.

```
> anova(credit.card.reg_model, credit.card.reg_model.refined, test = "Chisq")
Analysis of Deviance Table

Model 1: V21 ~ V1 + V2 + V3 + V4 + V5 + V6 + V7 + V8 + V9 + V10 + V11 +
    V12 + V13 + V14 + V15 + V16 + V17 + V18 + V19 + V20
Model 2: V21 ~ V1 + V2 + V3 + V4 + V5 + V6 + V8 + V14 + V20
  Resid. Df Resid. Dev  Df Deviance Pr(>Chi)
1       752     710.43
2       774     746.97 -22  -36.545  0.02652 *
---
```



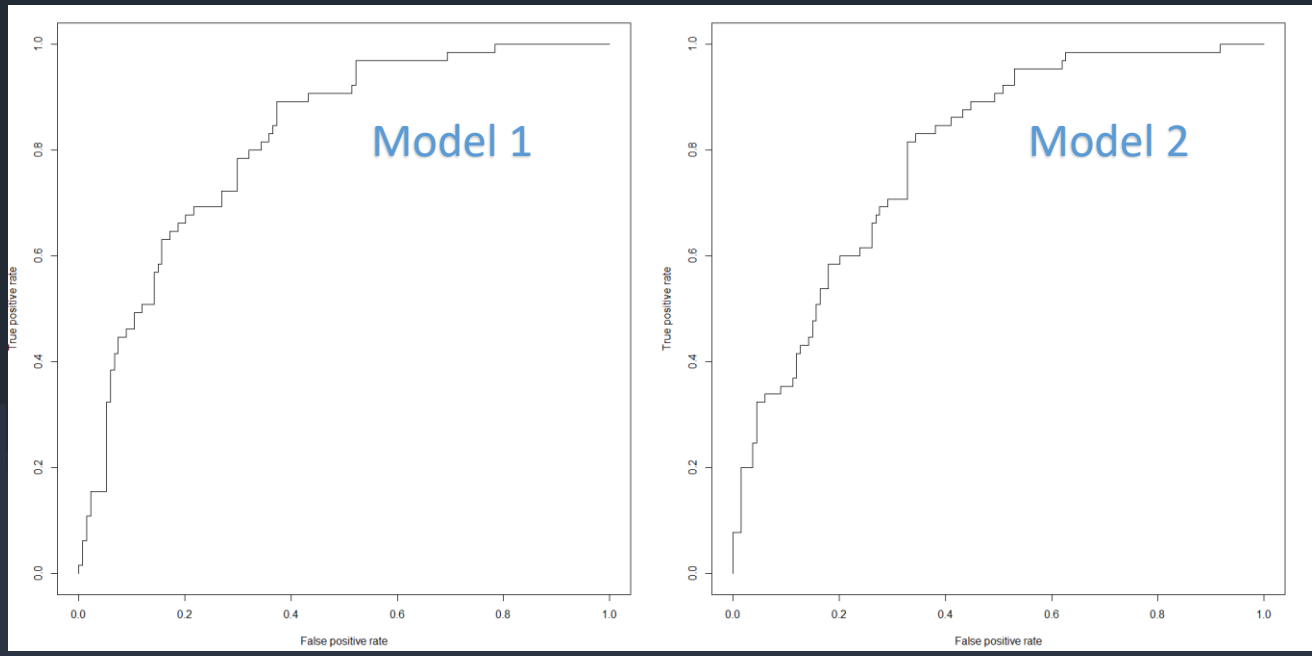


https://github.com/Hizzyth/GTX_Introduction-to-Analytics-Modelling

QUESTION 10.3 : Logistic Regression (German Credit data)

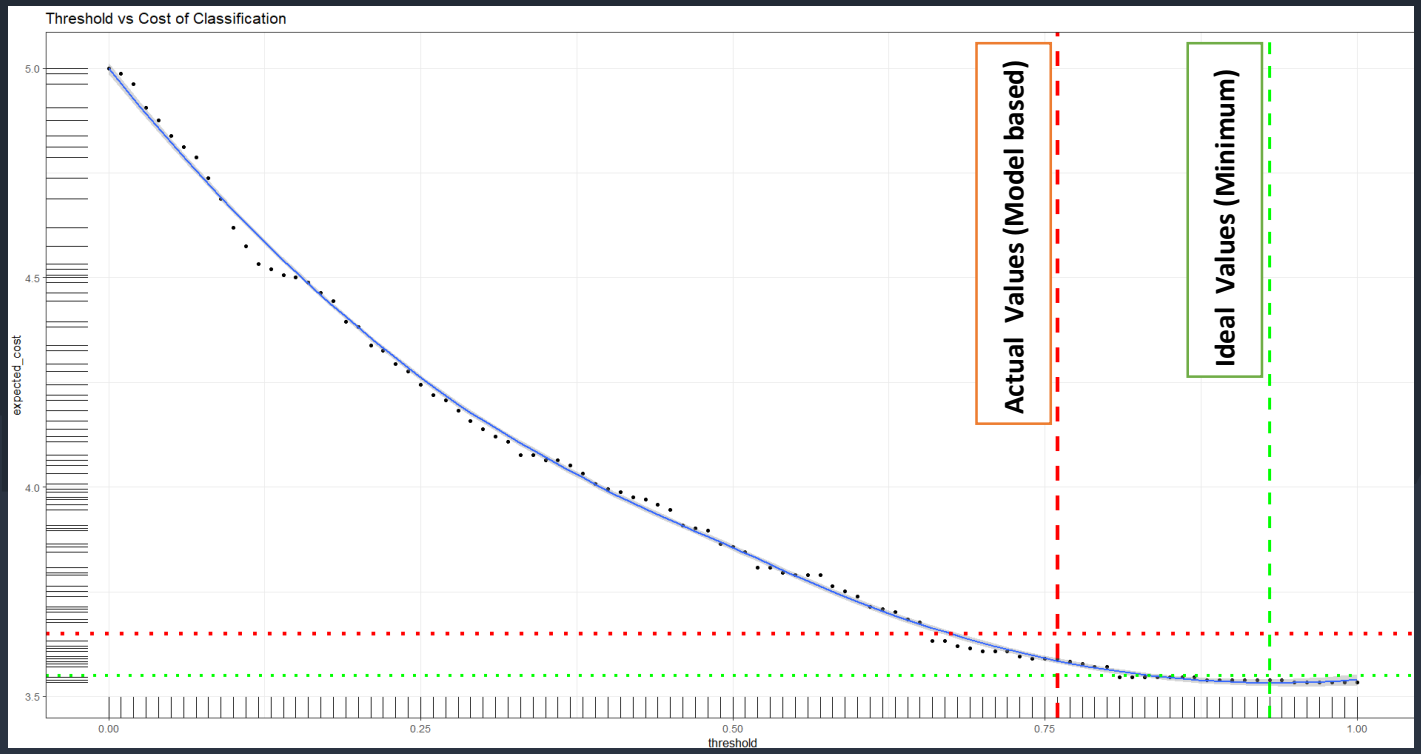
- AUC for Model 1: **81.7%**
- AUC for Model 2: **79.3%**

- **Area under curve is slightly better in Model 1 , and also visually we can see that Model 1 does better job at identifying True positives, compared to Model 2 which is more biased towards false positives.**



https://github.com/Hizzyth/GTX_Introduction-to-Analytics-Modelling

QUESTION 10.3 : Logistic Regression (Cost of Classification)



- Created Visualization for Threshold results vs Cost of Classification. I am using the values from 1st and end model for Threshold and based on that Dashed and dotted lines are drawn.
- Based on model and cost calculations, .95 is the ideal threshold.

```
> # Calculating the minimum cost of Classification
> result_df[which(result_df$expected_cost == min(result_df$expected_cost)), ]
    threshold expected_cost
96       0.95      3.533084
97       0.96      3.533084
98       0.97      3.533084
99       0.98      3.533084
100      0.99      3.533084
101      1.00      3.533084
>
```

https://github.com/Hizzyth/GTX_Introduction-to-Analytics-Modelling