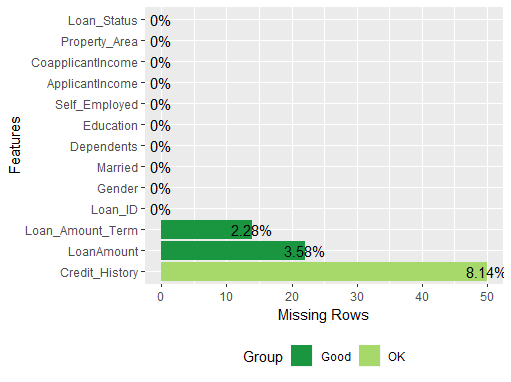
Giorgi Lomia Exam 1.

1. Done
2. There are 614 samples. There are 5 numeric predictors and There are 7 non-numeric predictors
3. 
   * Here is the chart of missing values. I will impute all of them
4. Done
5. Done. It was already a factor
6. $ Loan\_Status *<dbl>* 1, 0, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 0, 1, 1, 1, 0, 0, 1, 0, 1, ...
   * $ Gender *<fct>* Male, Male, Male, Male, Male, Male, Male, Male, Male, Male, Male, ...
   * $ Married *<fct>* No, Yes, Yes, Yes, No, Yes, Yes, Yes, Yes, Yes, Yes, Yes, Yes, No,...
   * $ Dependents *<fct>* 0, 1, 0, 0, 0, 2, 0, 3+, 2, 1, 2, 2, 2, 0, 2, 0, 1, 0, 0, 0, 0, 1,...
   * $ Education *<fct>* Graduate, Graduate, Graduate, Not Graduate, Graduate, Graduate, No...
   * $ Self\_Employed *<fct>* No, No, Yes, No, No, Yes, No, No, No, No, No, , No, No, No, No, No...
   * $ ApplicantIncome *<int>* 5849, 4583, 3000, 2583, 6000, 5417, 2333, 3036, 4006, 12841, 3200,...
   * $ CoapplicantIncome *<dbl>* 0, 1508, 0, 2358, 0, 4196, 1516, 2504, 1526, 10968, 700, 1840, 810...
   * $ LoanAmount *<int>* NA, 128, 66, 120, 141, 267, 95, 158, 168, 349, 70, 109, 200, 114, ...
   * $ Loan\_Amount\_Term *<int>* 360, 360, 360, 360, 360, 360, 360, 360, 360, 360, 360, 360, 360, 3...
   * $ Credit\_History *<int>* 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, NA, 0, 1, 1, 0, 1,...
   * $ Property\_Area *<fct>* Urban, Rural, Urban, Urban, Urban, Urban, Urban, Semiurban, Urban,..
7. Done
   * Train Dimensions
   * > dim(train)
   * [1] 492 12
   * Test Dimension
   * > dim(test)
   * [1] 122 12

1. Data Recipe

Inputs:

role #variables

outcome 1

predictor 11

Training data contained 492 data points and 65 incomplete rows.

Operations:

Dummy variables from Gender, Married, Dependents, Education, Self\_Employed, Property\_Area [trained]

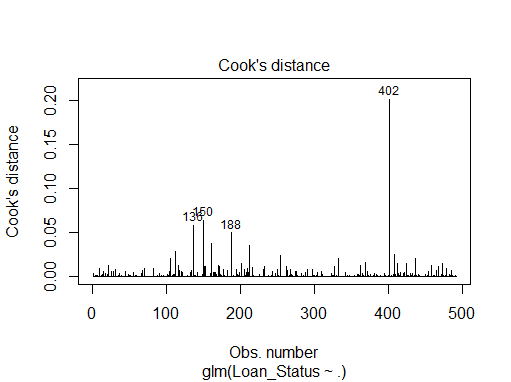
Bagged tree imputation for ApplicantIncome, CoapplicantIncome, ... [trained]

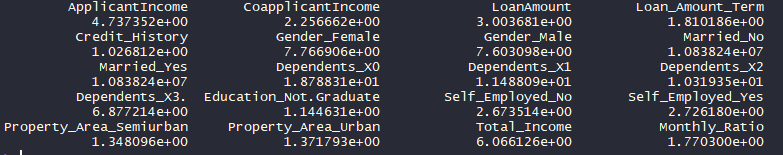
Variable mutation for Total\_Income [trained]

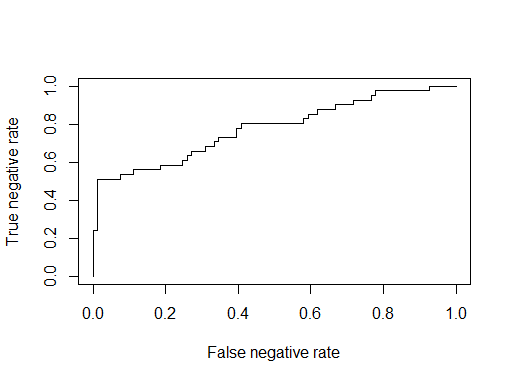
Variable mutation for Monthly\_Ratio [trained]

Log transformation on Total\_Income [trained]

We took a log of the Total\_Income in order to decorelate the total income from other variables. And we specifically chose the TOTAL income because it is the bigger variable that encompasses all the other related variables. Thus this step would give us the most beneficial results.

1. * > dim(train\_clean)
   * [1] 492 21
   * > dim(test\_clean)
   * [1] 122 21
2. Top 2 variables in predicting the loan status is: 1. Credit History and 2. Property Area. Important to note that coapplicant income is next in line of importance
3. 
   * No Studentized residuals with Bonferonni p < 0.05
   * Largest |rstudent|:
   * rstudent unadjusted p-value Bonferonni p
   * 474 -2.539467 0.011102 NA
   * There are no outliers. Nothing to remove.
4. Nothing needs to be removed. All of them are bellow 10.



1. * predicted.classes N Y
   * N 21 4
   * Y 20 77
   * Accuracy:
   * [1] 0.8032787
   * There is an imbalance. Our model gives more false positive values than false negative values. That means our model is too lenient and it would tell us to give loans to people who should not be getting them which is not good.
2. 
   * Area: 0.7777778
3. LDA Table and Acc:
   * lda.class N Y
   * N 21 3
   * Y 20 78
   * Accuracy
   * 0.8114754

QDA Table and Acc:

QDA was not able to run because it had rank deficiency. Because the classes do not have unique covariance matrices.