ALY 6015

Introduction:

The dataset includes secondary school student achievement data from two Portuguese schools. The data was imported from Kaggle and includes student grades, demographic, social, and school-related variables. This dataset includes numerical information about the students' first-period (G1), second-period (G2), and final grades (G3). We have included a new column where we determine if a student has passed with a grade more than or equal to 10 based on the final grade (G3). In order to determine whether a student has passed, we use Logistic Regression on this data in this project.

Analysis:

Data Cleaning:

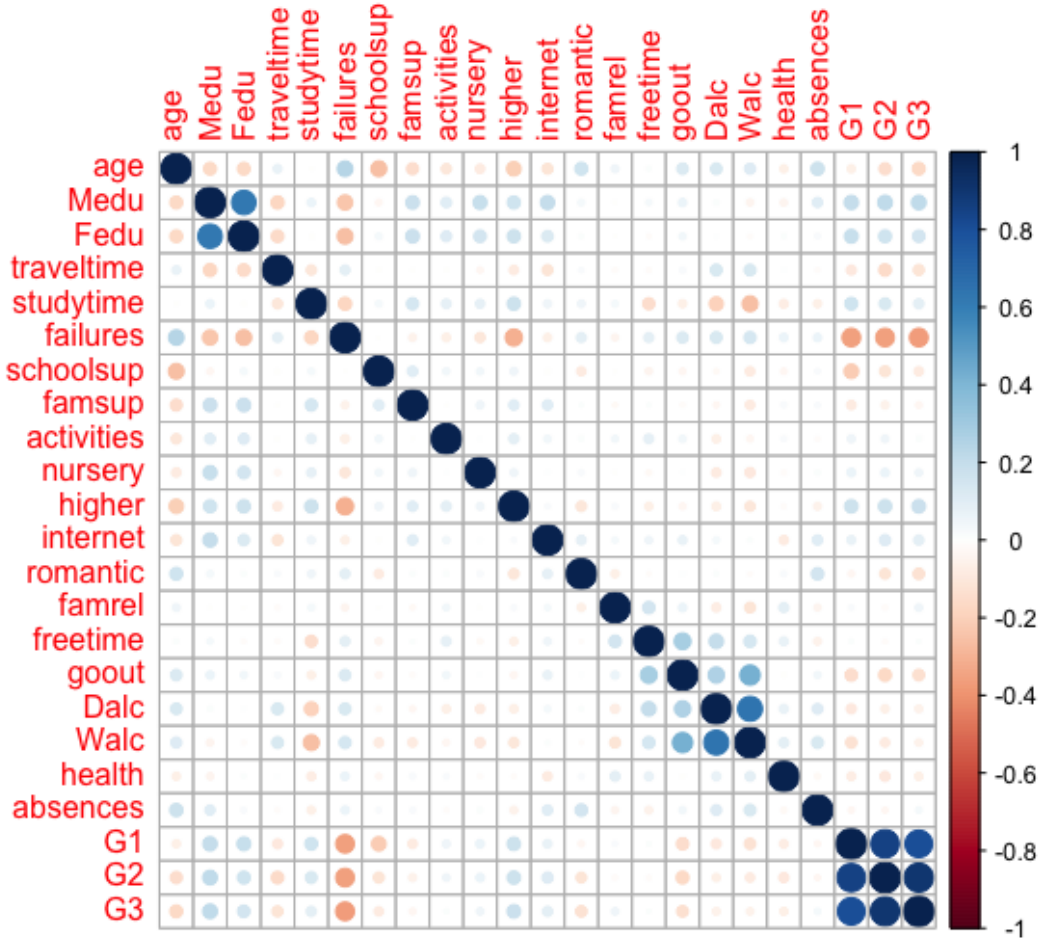
There were no NAs in the dataset. The categorical variables having yes/no values are converted to binary variables, where 1 denotes a yes and 0 denotes a no. FinalGrade, an output variable, is configured as a factor type.

Descriptive Statistics:

```
> summary(StudentDF)
    school              sex                 age         address            famsize
 Length:395         Length:395         Min.   :15.0   Length:395         Length:395
 Class :character   Class :character   1st Qu.:16.0   Class :character   Class :character
 Mode  :character   Mode  :character   Median :17.0   Mode  :character   Mode  :character
                                       Mean   :16.7
                                       3rd Qu.:18.0
                                       Max.   :22.0
   Pstatus              Medu            Fedu           Mjob               Fjob
 Length:395         Min.   :0.000   Min.   :0.000   Length:395         Length:395
 Class :character   1st Qu.:2.000   1st Qu.:2.000   Class :character   Class :character
 Mode  :character   Median :3.000   Median :2.000   Mode  :character   Mode  :character
                    Mean   :2.749   Mean   :2.522
                    3rd Qu.:4.000   3rd Qu.:3.000
                    Max.   :4.000   Max.   :4.000
    reason            guardian          traveltime      studytime        failures        schoolsup
 Length:395         Length:395         Min.   :1.000   Min.   :1.000   Min.   :0.0000   Min.   :0.0000
 Class :character   Class :character   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:0.0000
 Mode  :character   Mode  :character   Median :1.000   Median :2.000   Median :0.0000   Median :0.0000
                                       Mean   :1.448   Mean   :2.035   Mean   :0.3342   Mean   :0.1291
                                       3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:0.0000   3rd Qu.:0.0000
                                       Max.   :4.000   Max.   :4.000   Max.   :3.0000   Max.   :1.0000
     famsup         activities        nursery          higher          internet         romantic
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:1.0000   1st Qu.:1.0000   1st Qu.:1.0000   1st Qu.:0.0000
 Median :1.0000   Median :1.0000   Median :1.0000   Median :1.0000   Median :1.0000   Median :0.0000
 Mean   :0.6127   Mean   :0.5089   Mean   :0.7949   Mean   :0.9494   Mean   :0.8329   Mean   :0.3342
 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
     famrel         freetime         goout            Dalc            Walc           health
 Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
 1st Qu.:4.000   1st Qu.:3.000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:3.000
 Median :4.000   Median :3.000   Median :3.000   Median :1.000   Median :2.000   Median :4.000
 Mean   :3.944   Mean   :3.235   Mean   :3.109   Mean   :1.481   Mean   :2.291   Mean   :3.554
 3rd Qu.:5.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:5.000
 Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000
    absences           G1              G2              G3         FinalGrade
 Min.   : 0.000   Min.   : 3.00   Min.   : 0.00   Min.   : 0.00   0:130
 1st Qu.: 0.000   1st Qu.: 8.00   1st Qu.: 9.00   1st Qu.: 8.00   1:265
 Median : 4.000   Median :11.00   Median :11.00   Median :11.00
 Mean   : 5.709   Mean   :10.91   Mean   :10.71   Mean   :10.42
 3rd Qu.: 8.000   3rd Qu.:13.00   3rd Qu.:13.00   3rd Qu.:14.00
 Max.   :75.000   Max.   :19.00   Max.   :19.00   Max.   :20.00
```

Exploratory Data Analysis:

To find the variables with highest correlation, we have created a correlation matrix and correlation plot.

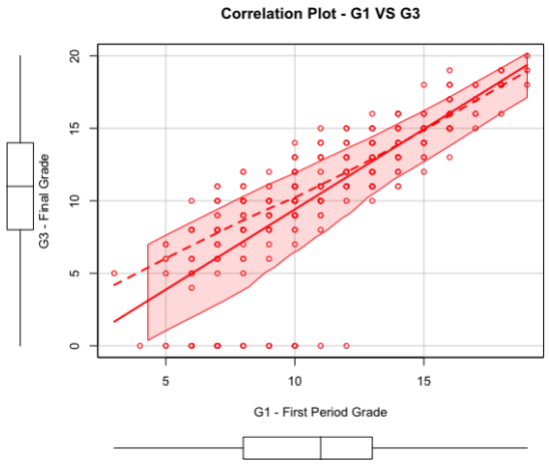


- From the above correlation plot, we can see the variables G1and G2 have the highest correlation with G3 with a correlation values of 0.8 and 0.9 respectively.
- Apart from these, the variable ‘failures’ has the next highest correlation of -0.36.
- These variables would be used as the predictors in our Logistic Regression Model.

Correlation Analysis:

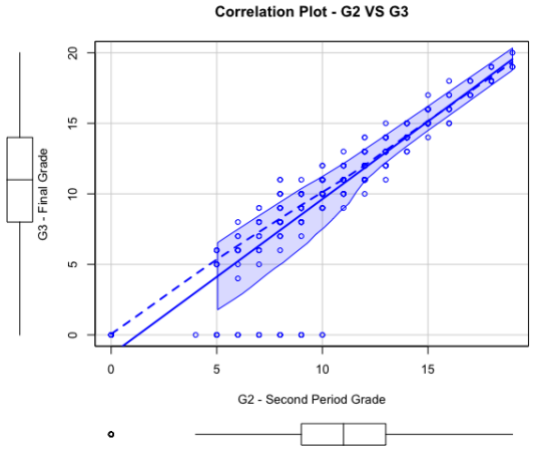
Correlation between G1 and G3:

- From the below scatterplot, we can see there is a positive correlation of G1 with G3.



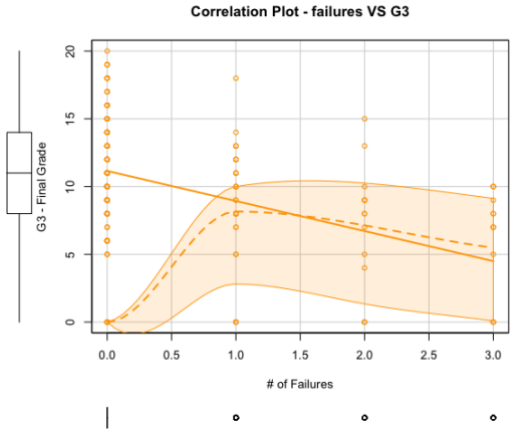
Correlation between G1 and G3:

- From the below scatterplot, we can see there is a positive correlation of G2 with G3.



Correlation between failures and G3:

- From the below scatterplot, we can see the correlation between failures and G3 is low when compared to failures and G3. There is a weak negative correlation.



Splitting the dataset to train and test sets:

- We split the dataset to train and test set by using the createDataPartition method in R.
- The training data that we created accounts for 70% of the dataset, with the test data account for 30% of the dataset.

```
> # Splitting the dataset into train and test set
> set.seed(123)
> trainIndex <- createDataPartition(StudentDF$FinalGrade, p=0.7, list = FALSE)
> train <- StudentDF[trainIndex,]
> test <- StudentDF[-trainIndex,]
```

Logistic Regression Model 1 (With 3 Predictors):

```
> LogRegModel1 <- glm(FinalGrade ~ G1+G2+failures ,data=train, family = binomial(link = logit))
> summary(LogRegModel1)

Call:
glm(formula = FinalGrade ~ G1 + G2 + failures, family = binomial(link = logit),
    data = train)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.84867  -0.03312   0.01628   0.12254   2.14908

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -16.85994    2.87352  -5.867 4.43e-09 ***
G1            0.14241    0.19042   0.748    0.455
G2            1.68950    0.31392   5.382 7.37e-08 ***
failures      0.04986    0.32158   0.155    0.877
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 350.75  on 276  degrees of freedom
Residual deviance: 106.77  on 273  degrees of freedom
AIC: 114.77

Number of Fisher Scoring iterations: 8
```

- We have created a Logistic Regression Model with the three independent variables G1, G2 and failures which have an impact on G3.

```
> # Regression Coef (Log Odds)
> coef(LogRegModel1)
 (Intercept)           G1           G2     failures
-16.85994150   0.14241441   1.68949560   0.04985899
>
> # Regression Coef (Odds)
> exp(coef(LogRegModel1))
 (Intercept)           G1           G2     failures
4.762340e-08 1.153054e+00 5.416748e+00 1.051123e+00
```

Confusion Matrix for Train Set:

```
> confusionMatrix(predicted.classes.min, train$FinalGrade)
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0  84  18
         1   7 168

               Accuracy : 0.9097
                 95% CI : (0.8697, 0.9407)
    No Information Rate : 0.6715
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.8016

 Mcnemar's Test P-Value : 0.0455

            Sensitivity : 0.9231
            Specificity : 0.9032
         Pos Pred Value : 0.8235
         Neg Pred Value : 0.9600
             Prevalence : 0.3285
         Detection Rate : 0.3032
   Detection Prevalence : 0.3682
      Balanced Accuracy : 0.9132

       'Positive' Class : 0
```

- Accuracy: From the above confusion matrix, we can see the Accuracy of the model is 0.9097 which means the model is correct 90.97% of times.

- Sensitivity: The sensitivity of the model is 0.9231, which means when the final grade is ‘pass’, how many times the model will predict it as ‘pass’. 92.31% is a good sensitivity score.

- Specificity: The specificity of the model is 0.9032 which means when the final grade is ‘fail’, how many times the model will predict it as ‘fail’. The model predicts it 90.32% of the times which is a good score.

- Precision: The precision shows out of all predicted values, how many times we get the true values. The precision is 0.8235 or 82.35% which is a good score.

Confusion Matrix for a Test Set:

```
> # Confusion matrix for Test Set
>
> probabilities.test <- predict(LogRegModel1, newdata=test, type="response")
>
> predicted.classes.min <- as.factor(ifelse(probabilities.test>=0.5, 1, 0))
>
> confusionMatrix(predicted.classes.min, test$FinalGrade)
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 38  6
         1  1 73
```

```
               Accuracy : 0.9407
                 95% CI : (0.8816, 0.9758)
    No Information Rate : 0.6695
    P-Value [Acc > NIR] : 1.179e-12

                  Kappa : 0.8702

 Mcnemar's Test P-Value : 0.1306

            Sensitivity : 0.9744
            Specificity : 0.9241
         Pos Pred Value : 0.8636
         Neg Pred Value : 0.9865
             Prevalence : 0.3305
         Detection Rate : 0.3220
   Detection Prevalence : 0.3729
      Balanced Accuracy : 0.9492

       'Positive' Class : 0
```

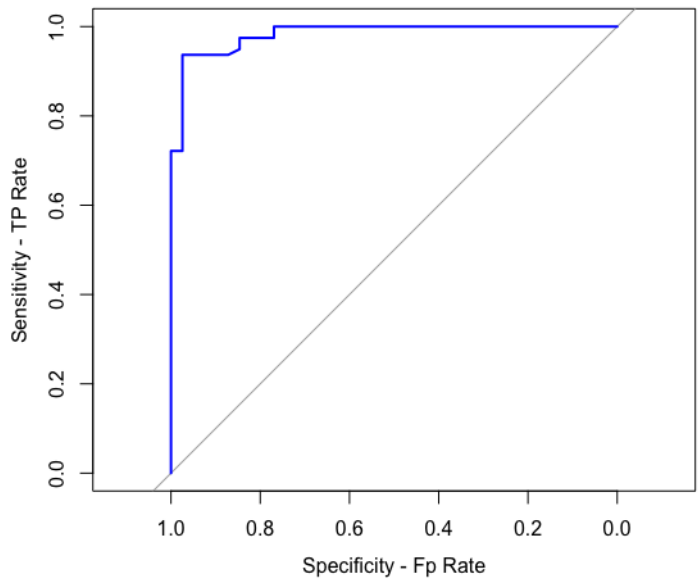
- Accuracy: From the above confusion matrix, we can see the Accuracy of the model is 0.9407 which means the model is correct 94.07% of times.

- Sensitivity: The sensitivity of the model is 0.9744, which means when the final grade is ‘pass’, how many times the model will predict it as ‘pass’. 97.44% is a good sensitivity score.

- Specificity: The specificity of the model is 0.9241 which means when the final grade is ‘fail’, how many times the model will predict it as ‘fail’. The model predicts it 92.41% of the times which is a good score.

- Precision: The precision shows out of all predicted values, how many times we get the true values. The precision is 0.8636 or 86.36% which is a good score.

ROC Curve:



- The ROC curve shows the trade-off between Sensitivity and Specificity of the model.

- From the above ROC curve, we can see that the curve is not closer to the 45 degree line, which means that the test is accurate.

AUC Score:

```
> # 8. AUC
> AUC1 = auc(ROC1)
> AUC1
Area under the curve: 0.983
```

- The AUC score is an indicator of the degree to which a measure of separability can be determined. It reveals the model's ability to discriminate between various types of data.
- From the above figure, we can see the Area Under Curve value is 0.983 which is a near perfect result.

Logistic Regression Model 2 (With all variables as Predictors):

```
Call:
glm(formula = FinalGrade ~ ., family = binomial(link = logit),
    data = train)

Deviance Residuals:
       Min          1Q      Median          3Q         Max
-3.722e-05  -2.100e-08   2.100e-08   2.100e-08   4.803e-05

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -1.552e+02  3.791e+05   0.000    1.000
schoolMS         1.475e+01  7.844e+04   0.000    1.000
sexM             1.271e-01  4.480e+04   0.000    1.000
age             -6.257e+00  2.441e+04   0.000    1.000
addressU        -7.105e+00  5.651e+04   0.000    1.000
famsizeLE3       4.651e-01  5.602e+04   0.000    1.000
PstatusT        -6.122e+00  8.804e+04   0.000    1.000
Medu             5.483e+00  2.371e+04   0.000    1.000
Fedu            -1.847e+00  3.310e+04   0.000    1.000
Mjobhealth      -2.164e+01  8.706e+04   0.000    1.000
Mjobother       -9.960e+00  6.291e+04   0.000    1.000
Mjobservices    -1.411e+01  9.891e+04   0.000    1.000
Mjobteacher     -2.074e+01  8.057e+04   0.000    1.000
Fjobhealth      -2.643e+01  8.044e+04   0.000    1.000
Fjobother       -2.051e+00  7.249e+04   0.000    1.000
Fjobservices    -1.090e+01  6.303e+04   0.000    1.000
Fjobteacher     -3.151e+00  9.627e+04   0.000    1.000
reasonhome       2.928e+00  6.306e+04   0.000    1.000
```

```
reasonother       8.818e+00  1.263e+05   0.000    1.000
reasonreputation  9.674e+00  5.902e+04   0.000    1.000
guardianmother   -5.167e+00  3.385e+04   0.000    1.000
guardianother     6.582e+00  6.563e+04   0.000    1.000
traveltime        4.720e+00  2.670e+04   0.000    1.000
studytime        -5.814e+00  3.384e+04   0.000    1.000
failures          1.492e+00  3.550e+04   0.000    1.000
schoolsup        -7.078e+00  6.640e+04   0.000    1.000
famsup           -5.673e+00  4.377e+04   0.000    1.000
paid              1.632e+01  5.851e+04   0.000    1.000
activities       -1.964e+00  3.275e+04   0.000    1.000
nursery          -1.041e+01  4.775e+04   0.000    1.000
higher           -4.896e+00  1.792e+05   0.000    1.000
internet          2.673e+00  2.750e+04   0.000    1.000
romantic         -3.203e+00  4.086e+04   0.000    1.000
famrel            6.074e+00  2.835e+04   0.000    1.000
freetime         -2.698e+00  2.512e+04   0.000    1.000
goout             2.303e+00  2.451e+04   0.000    1.000
Dalc             -4.668e+00  3.349e+04   0.000    1.000
Walc              1.353e+00  1.836e+04   0.000    1.000
health           -1.348e+00  1.530e+04   0.000    1.000
absences         -5.814e-01  4.620e+03   0.000    1.000
G1               -3.425e-01  2.283e+04   0.000    1.000
G2                2.293e+00  2.091e+04   0.000    1.000
G3                2.780e+01  2.138e+04   0.001    0.999

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3.5075e+02  on 276  degrees of freedom
Residual deviance: 2.6418e-08  on 234  degrees of freedom
AIC: 86

Number of Fisher Scoring iterations: 25
```

- The dependent variable is "FinalGrade" and the independent variables are all the other variables in the dataset.

- The coefficients for each independent variable are shown, along with their standard errors, z-scores, and p-values. The intercept is also shown.

- The deviance residuals, which measure the difference between the observed and predicted values, are shown. The minimum and maximum values are close to zero, indicating a good fit.

- However, all of the p-values for the independent variables are 1, indicating that none of the independent variables are significantly associated with the dependent variable. This may be due to multicollinearity or other issues with the data or model.

Log Odds and Odds:

```
> # Regression Coef (Log Odds)
> coef(LogRegModel2)
     (Intercept)        schoolMS            sexM             age        addressU       famsizeLE3
    -155.1967992       14.7484656        0.1271201       -6.2574158       -7.1051568        0.4650629
        PstatusT             Medu             Fedu       Mjobhealth       Mjobother    Mjobservices
      -6.1220586        5.4826640       -1.8471530      -21.6379562       -9.9603173      -14.1149084
     Mjobteacher       Fjobhealth        Fjobother    Fjobservices      Fjobteacher       reasonhome
     -20.7420575      -26.4302911       -2.0508580      -10.9027442       -3.1515058        2.9275361
     reasonother reasonreputation   guardianmother   guardianother       traveltime        studytime
       8.8179608        9.6739481       -5.1671195        6.5815190        4.7203726       -5.8135925
        failures        schoolsup           famsup             paid       activities          nursery
       1.4915953       -7.0780152       -5.6729244       16.3219094       -1.9641711      -10.4077528
          higher         internet         romantic           famrel         freetime            goout
      -4.8958988        2.6726740       -3.2026947        6.0738104       -2.6979173        2.3034499
           Dalc             Walc           health         absences               G1              G2
      -4.6676284        1.3534098       -1.3483358       -0.5814114       -0.3424883        2.2929453
            G3
      27.8017602
> # Regression Coef (Odds)
> exp(coef(LogRegModel2))
     (Intercept)        schoolMS            sexM             age        addressU       famsizeLE3
    3.970878e-68     2.542010e+06     1.135553e+00     1.916191e-03     8.208610e-04     1.592114e+00
        PstatusT             Medu             Fedu       Mjobhealth       Mjobother    Mjobservices
    2.193935e-03     2.404865e+02     1.576855e-01     4.006407e-10     4.723774e-05     7.412644e-07
     Mjobteacher       Fjobhealth        Fjobother    Fjobservices      Fjobteacher       reasonhome
    9.813838e-10     3.322541e-12     1.286245e-01     1.840765e-05     4.278765e-02     1.868155e+01
     reasonother reasonreputation   guardianmother   guardianother       traveltime        studytime
    6.754477e+03     1.589799e+04     5.700967e-03     7.216347e+02     1.122100e+02     2.986681e-03
        failures        schoolsup           famsup             paid       activities          nursery
    4.444180e+00     8.434456e-04     3.437797e-03     1.226070e+07     1.402721e-01     3.019746e-05
          higher         internet         romantic           famrel         freetime            goout
    7.477186e-03     1.447863e+01     4.065251e-02     4.343325e+02     6.734562e-02     1.000865e+01
           Dalc             Walc           health         absences               G1              G2
    9.394523e-03     3.870601e+00     2.596720e-01     5.591087e-01     7.100014e-01     9.904065e+00
            G3
    1.186181e+12
```

- The intercept is -155.197, and the coefficients for the predictor variables range from -26.43 to 27.802. The odds ratios range from 0.0000000000004 to 1.186181e+12. The odds ratio for a predictor variable represents the increase in the odds of the outcome for a one-unit increase in the predictor variable.
- a one-unit increase in age is associated with an odds ratio of 0.002, meaning that the odds of passing the final exam decrease by a factor of 0.002 for each one-year increase in age.

Confusion Matrix for Train Set:

```
> # Confusion matrix for Train Set
> probabilities.train <- predict(LogRegModel2, newdata = train, type = "response")
> predicted.classes.min <- as.factor(ifelse(probabilities.train >= 0.5, 1, 0))
> confusionMatrix(predicted.classes.min, train$FinalGrade)
```

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0  91   0
         1   0 186

               Accuracy : 1
                 95% CI : (0.9868, 1)
    No Information Rate : 0.6715
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1

 Mcnemar's Test P-Value : NA

            Sensitivity : 1.0000
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 1.0000
             Prevalence : 0.3285
         Detection Rate : 0.3285
   Detection Prevalence : 0.3285
      Balanced Accuracy : 1.0000

       'Positive' Class : 0
```

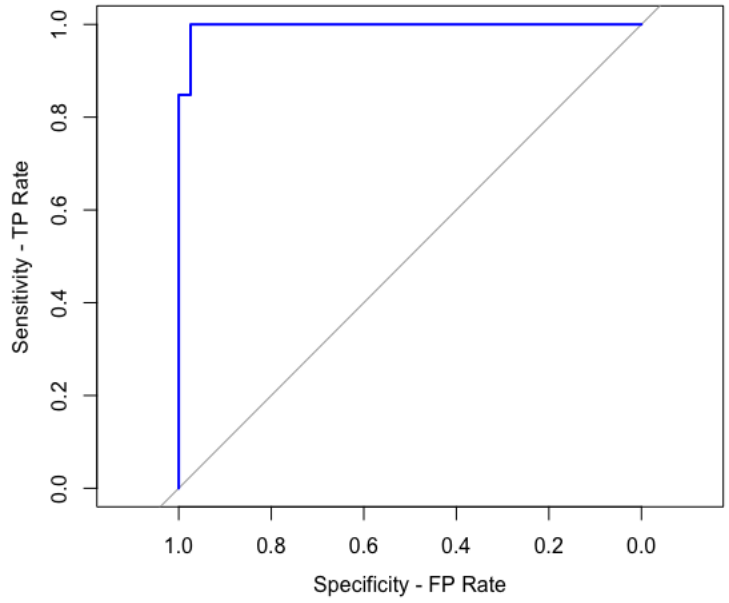
- The confusion matrix shows the number of correct and incorrect predictions for each class (0 or 1). The accuracy of the model is 100%, meaning that it correctly classified all the observations in the training set. The sensitivity and specificity are also 100%, indicating that the model is very good at identifying both positive and negative cases. The code is not predicting any positive cases, as indicated by the 'Positive' class label being 0.

Confusion Matrix for Test Set:

```
> # Confusion matrix for Test Set
> probabilities.test <- predict(LogRegModel2, newdata = test, type = "response")
> predicted.classes.min <- as.factor(ifelse(probabilities.test >= 0.5, 1, 0))
> confusionMatrix(predicted.classes.min, test$FinalGrade)
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 38  3
         1  1 76

               Accuracy : 0.9661
                 95% CI : (0.9155, 0.9907)
    No Information Rate : 0.6695
    P-Value [Acc > NIR] : 1.342e-15

                  Kappa : 0.9244

 Mcnemar's Test P-Value : 0.6171

            Sensitivity : 0.9744
            Specificity : 0.9620
         Pos Pred Value : 0.9268
         Neg Pred Value : 0.9870
             Prevalence : 0.3305
         Detection Rate : 0.3220
   Detection Prevalence : 0.3475
      Balanced Accuracy : 0.9682

       'Positive' Class : 0
```

- The model shows a high accuracy of 0.9661, indicating that the model was able to correctly predict the final grade for 96.61% of the test cases. The specificity is 0.9620, which means that the model correctly identifies 96.20% of students who will not pass the exam. The sensitivity is 0.9744, which means that the model correctly identifies 97.44% of students who will pass the exam. The precision (positive predictive value) is 0.9268, which means that when the model predicts a student to pass the exam, there is a 92.68% chance that the student will actually pass the exam.

ROC Curve:



- The ROC curve shows the trade-off between Sensitivity and Specificity of the model.
- From the above ROC curve, we can see that the curve is not closer to the 45 degree line, which means that the test is accurate.

AOC Value:

```
> # AUC
> AUC2 <- auc(ROC2)
> AUC2
Area under the curve: 0.9961
```

- An AUC of 1 represents perfect discrimination, while an AUC of 0.5 represents a random classifier. In this case, the AUC value is 0.9961, indicating that the classifier has a very good performance in distinguishing between the two classes.

Conclusion:

The above logistic regression models were created to determine whether a student has passed or failed. This model can be used in predicting if a student with certain grades in G1 and G2 with number of failures can pass in G3 or not. By comparing the values in the four sections of the matrix, the regression model can be easily verified. The Second model using all variables as predictors appears to be performing better as it has a higher accuracy, sensitivity, specificity, and AUC compared to the model using only 3 variables as predictors. It is important to note that other factors, such as the size and representativeness of the dataset, the business context, and the computational complexity of the models, should also be taken into consideration when comparing models.

References:

1. *Student Grade Prediction*. (2018, September 14). Kaggle. https://www.kaggle.com/datasets/dipam7/student-grade-prediction?select=student-mat.csv
2. *Logit Regression | R Data Analysis Examples*. (n.d.). https://stats.oarc.ucla.edu/r/dae/logit-regression/
3. C. (2021b, December 7). *Simply Explained Logistic Regression with Example in R*. Medium. https://towardsdatascience.com/simply-explained-logistic-regression-with-example-in-r-b919acb1d6b3