2022

ST3189 Coursework 190526753

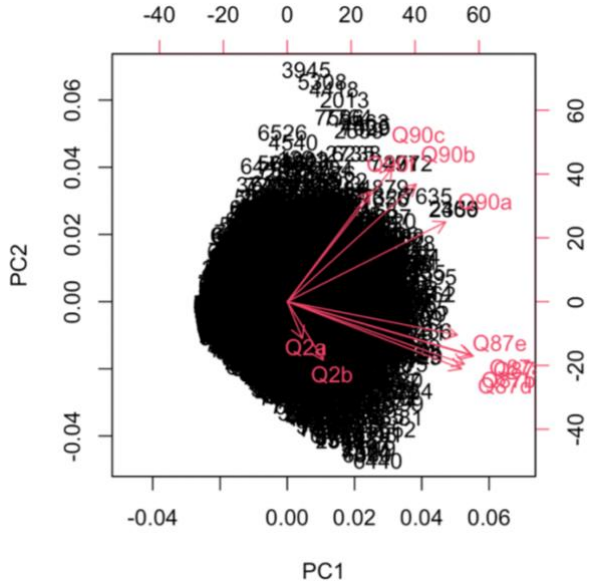
Table of Contents

1 Part 1

1.1 Question 1

Analyze and summarize the information in the data with unsupervised learning techniques.

1.2 EWCS Biplot



As we look at the biplot, Q87a ~ Q87e contributes most to the first principal component and Q90a~Q90c and Q90f contributes most to the second principal component. Furthermore, the closer the distance and direction, the higher the correlation of the variables. Thus, we can assume that Q87a ~ Q87e are correlated and Q90a~Q90c and Q90f are correlated. Hence, we will label and group them into two different classification, “Life” and “Work”. “L” for life and “W” for work.

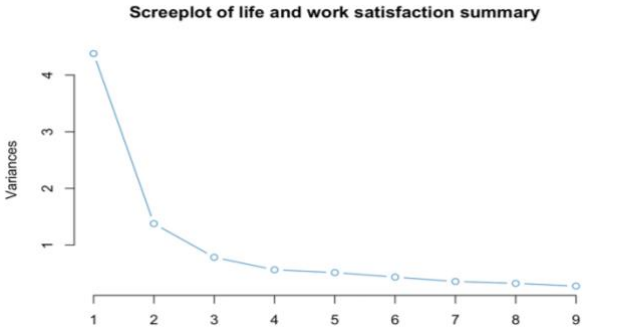
1.3 Various aspects of the data

To further scrutinize the data, we extracted variables to look at the various aspects of the data. We dived them into 3 different group, life-work satisfaction, life satisfaction and work satisfaction, then conducted PCA. Among life, “I woke up feeling fresh and rested” and among work, “I am enthusiastic about my job” had the biggest variance which then tell us that European workers vary the most on those factors.

```
L_Cheerful       L_Calm     L_Active L_WakeUpFreshed  L_Interesting
  1.228888     1.494328     1.311350        1.636771       1.411519

W_Energetic W_Enthusiastic    W_TimeFlies       W_Expert
  0.7167108      1.0269436      0.9390320      0.4536516
```

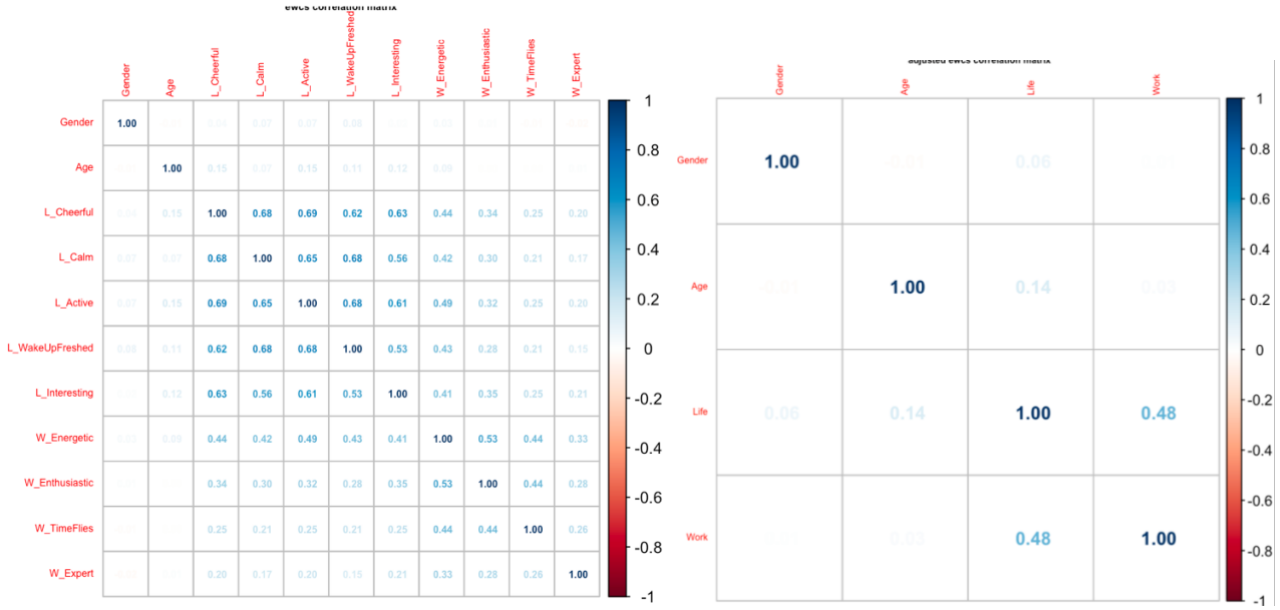
For life-work satisfaction which includes both “life” and “work” variables, 5 principal component (PC) seems to be the elbow point, a point at which the slope changes. With the summary and plot, we can identify that PC5 has cumulative proportion of 85% which means that with 5 PC, we can explain 85% of the variation. Hence, 5 PC could be a simpler substitute for all 9 factors as it could explain most of the variability without losing copious amount of its initial variability.




```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.0928 1.1737 0.88442 0.74975 0.71598 0.65853 0.59602
## Proportion of Variance 0.4867 0.1531 0.08691 0.06246 0.05696 0.04819 0.03947
## Cumulative Proportion  0.4867 0.6397 0.72664 0.78909 0.84605 0.89424 0.93371
##                           PC8     PC9
## Standard deviation     0.56770 0.52378
## Proportion of Variance 0.03581 0.03048
## Cumulative Proportion  0.96952 1.00000
```

```
## Standard deviations (1, .., p=9):
## [1] 2.0928204 1.1737194 0.8844224 0.7497501 0.7159772 0.6585330 0.5960200
## [8] 0.5676965 0.5237831
##
## Rotation (n x k) = (9 x 9):
##                         PC1        PC2         PC3         PC4         PC5
## L_Cheerful      -0.3913935  0.2145874 -0.03657888  0.03076865 -0.18884134
## L_Calm          -0.3791006  0.2659415 -0.02985531  0.08182211  0.16291474
## L_Active        -0.3965664  0.2153921 -0.01930540  0.05021775  0.09626028
## L_WakeUpFreshed -0.3718983  0.2719571  0.01071816  0.14547235  0.36815180
## L_Interesting   -0.3634299  0.1412722 -0.04224498 -0.13683679 -0.72104140
## W_Energetic     -0.3392943 -0.3131456  0.11803397 -0.21376803  0.48446971
## W_Enthusiastic  -0.2777600 -0.4467453  0.27766299 -0.61876652 -0.09789550
## W_TimeFlies     -0.2263318 -0.5125126  0.37107470  0.71737559 -0.16969792
## W_Expert        -0.1787233 -0.4273225 -0.87565589  0.08218019  0.00185347
##                         PC6         PC7         PC8         PC9
## L_Cheerful       0.04881360  0.63571903  0.266498682  0.534111082
## L_Calm           0.40807759  0.27941113 -0.578403554 -0.415068521
## L_Active        -0.18414840 -0.06535586  0.644941326 -0.574499260
## L_WakeUpFreshed  0.20550712 -0.61946594  0.013531916  0.452508421
## L_Interesting   -0.34425995 -0.31473490 -0.299450773 -0.017535605
## W_Energetic     -0.63313474  0.14717915 -0.252825622  0.077701110
## W_Enthusiastic   0.47109683 -0.08517536  0.154921357 -0.022979738
## W_TimeFlies      0.06334863 -0.01398893  0.004831339 -0.027406265
## W_Expert         0.09811384 -0.04013732  0.026073459  0.002463957
```

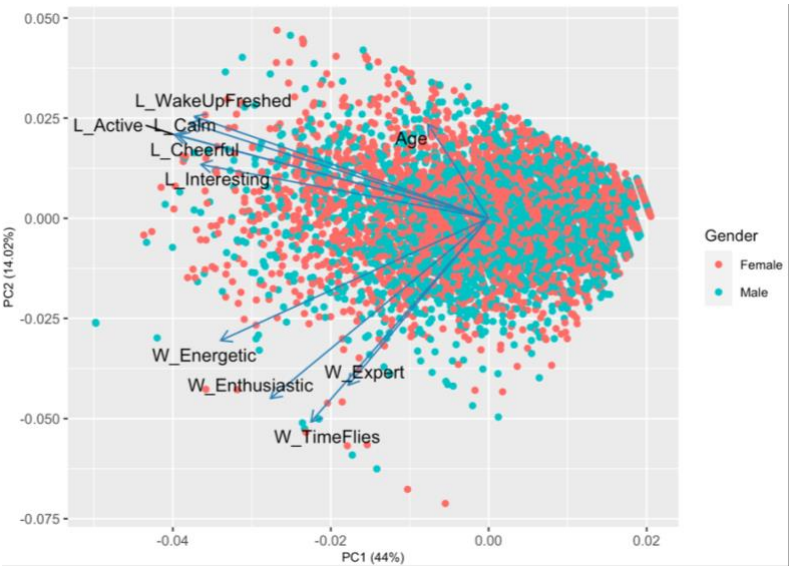
1.4 Correlation matrix



From the EWCS correlation matrix, we identified that gender and age have very low or no correlation to each other and also to life and work satisfaction variables. Furthermore, as seen from the previous biplot, we once again identified that life satisfaction variables (cheerful, calm, active, wakeupfreshed, interesting) and work satisfaction variables (energetic, enthusiastic, timeflies, expert) have correlation within each of the variables. Hence, to see the whole picture, we summed up the scores of each life and work satisfaction and then plotted correlation matrix. From the adjusted EWCS correlation matrix, we identified that there is correlation of 0.48 between “life” and “work”.

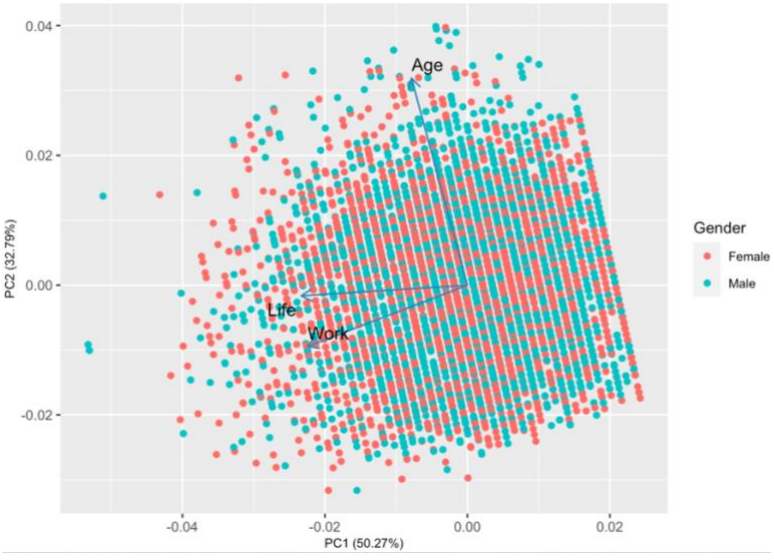
1.5 Adjusted biplot

From the contributions of each variable to the principal components, PC1 has large negative associations with “Life” variables and PC2 has large negative associations with “Work” variables. These could mean that the principal component initially measures the workers life satisfaction first, then measures working satisfaction next. Furthermore, as cosine of the angle between age and work variables except “W_Energetic” are beyond perpendicular to each other, we can assume that working satisfaction doesn’t vary much on its age. However, as age increases, “Life” variables tend to vary more as they seem correlated. Considering “L_Active” and “L_Cheerful” which have wider variance compared to other variables, may tend to differ more between workers as their age increases. Thus, worker may feel less energized at work but may have sufficient life satisfaction as they age.




```
##                          PC1         PC2
## Age              -0.07673934  0.2341908
## L_Cheerful       -0.39125911  0.2059654
## L_Calm           -0.37753362  0.2372082
## L_Active         -0.39650599  0.2079502
## L_WakeUpFreshed  -0.37122407  0.2530487
## L_Interesting    -0.36298615  0.1336218
## W_Energetic      -0.33809406 -0.3038659
## W_Enthusiastic   -0.27520815 -0.4478020
## W_TimeFlies      -0.22412564 -0.5062590
## W_Expert         -0.17729376 -0.4149388
```

Moreover, from adjusted EWCS biplot, we can see that the first PC has large negative associations of “Life” and “Work” while second PC has positive associations of age. Each PC then explains 50% and 33% of the total variance. Hence, we can once again identify with a clear picture that “Life” and




```
## Importance of components:
##                           PC1    PC2    PC3
## Standard deviation     1.2281 0.9919 0.7127
## Proportion of Variance 0.5027 0.3280 0.1693
## Cumulative Proportion  0.5027 0.8307 1.0000
```

```
## Standard deviations (1, .., p=3):
## [1] 1.2280820 0.9918849 0.7127264
##
## Rotation (n x k) = (3 x 3):
##              PC1          PC2         PC3
## Age  -0.2357538  0.95767696 -0.1651514
## Life -0.7005319 -0.04968643  0.7118892
## Work -0.6735541 -0.28352442 -0.6825971
```

“Work” have positive correlation to each other but not to gender and age.

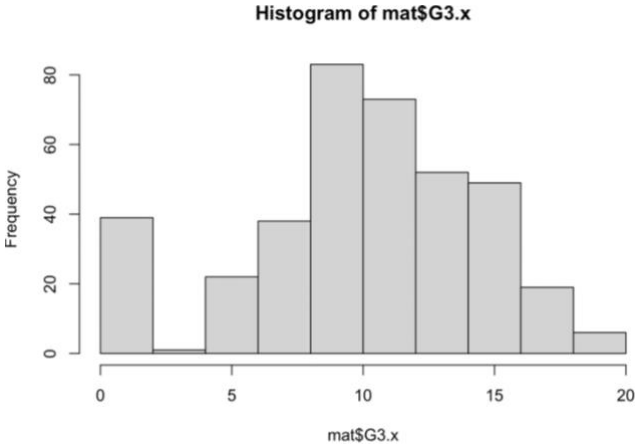
2 Part 2

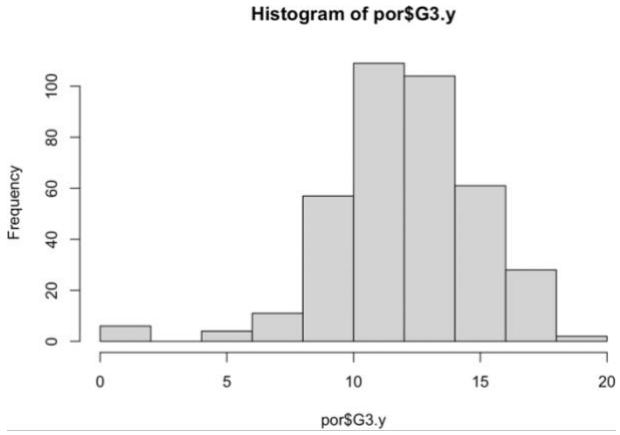
2.1 Question 2

Compare regression models by interpreting and assessing its performance.

2.2 Exploratory data analysis (EDA)

Before we deep dive into predicting model, we went through EDA to analyze the data. From the correlation matrix of mat and por, we observed that G3 is slightly correlated to medu, fedu, and studytime for both. However, from the EDA, besides identifying that histogram of G3 being left skewed, nothing interesting was observed.





2.3 Multiple linear regression model

First model we went through was multiple linear regression model. We first assumed that the data meets the assumptions for linear regression. Then later, we adjusted the model to meet its assumptions. Furthermore, we divided our dataset into two datasets to train and validate the model.

```
set.seed(1)
mat_training_sample <- createDataPartition(mat$G3.x,p = 0.7, list = FALSE)
por_training_sample <- createDataPartition(por$G3.y,p = 0.7, list = FALSE)
```

```
mat_dataset<- mat[mat_training_sample,]
por_dataset<- por[por_training_sample,]
```

- Use the 70% of data to train the model

```
mat_validation <- mat[-mat_training_sample,]
por_validation <- por[-por_training_sample,]
```

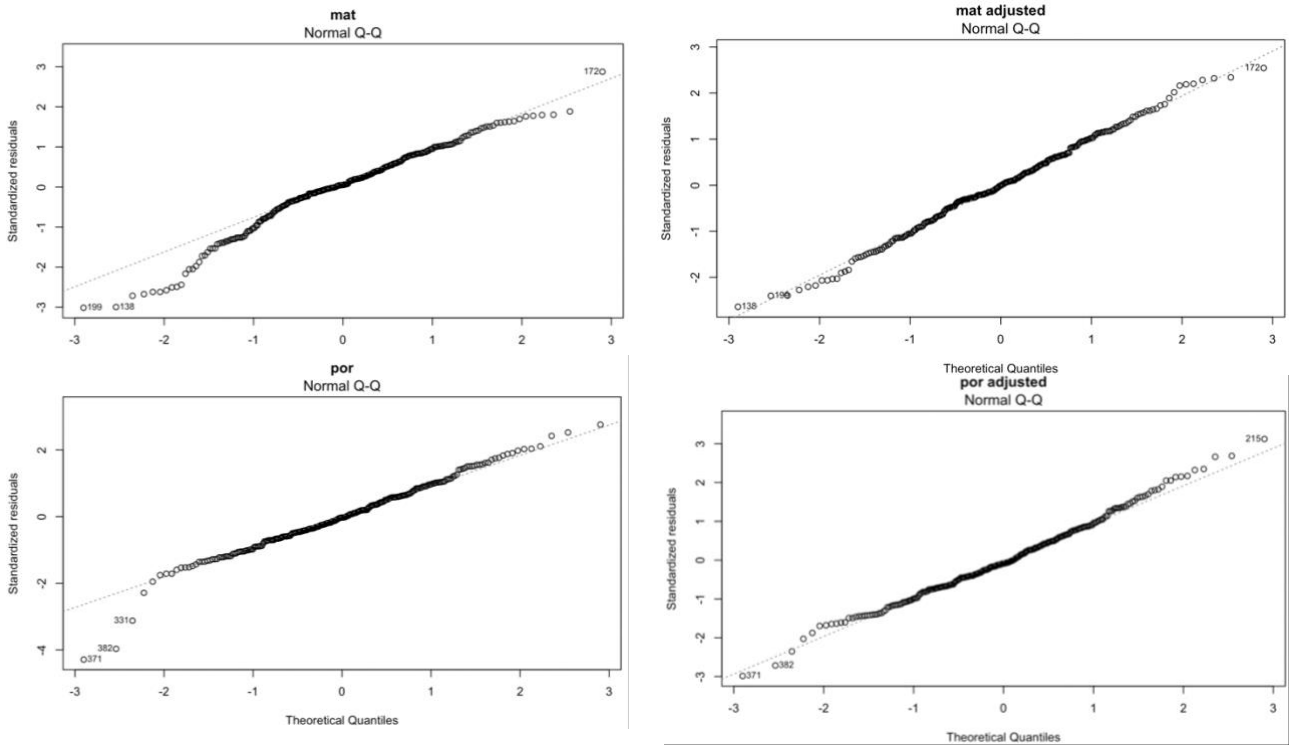
- Use the 30% of the data to validate the model.

Multiple linear regression RMSE and R^2 (assumed noramlly distributed)

- Train
- RMSE: 3.730919(mat) 2.110835(por)
- R^2: 0.3454868(mat) 0.4543227(por)
- Test
- RMSE: 3.341797(mat) 2.00442(por)
- R^2: 0.5227731(mat) 0.5901086(por)

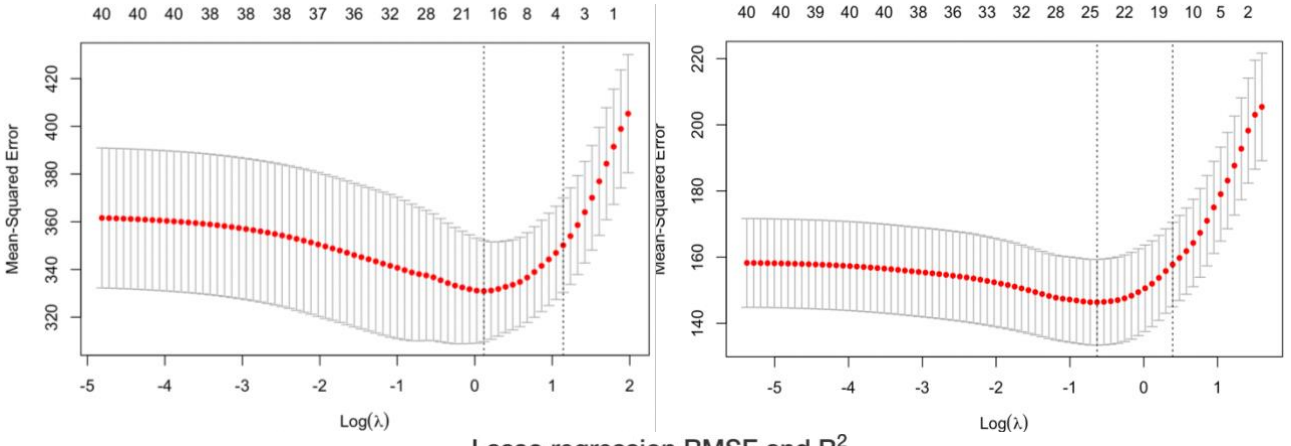
Since RMSE and R^2 evaluate the fit of dataset into the model, we identified them for both train and test sets. Next, we conducted a normality test to test the assumptions of the regression model by ploting diagnostic plots and conducting Shapiro-Wilk Normality Test. Looking at the normal Q-Q plot from diagnostic plots, distribution of residuals seems heavy on the tails for both mat and por. Also, the Shapiro-Wilk Normality Test shows that both are not normally distributed as P-values are less

than 0.05. Hence to normalize the distribution, we raised the response variable by 3/2 since logarithmic way doesn't seem to normalize the model.



2.4 Regularization with Lasso Regression

To regularize the model with lasso regression, we found the optimal lambda value of 1.13 for "mat" and 0.53 for "por" with the use of k-fold cross validation method then conducted the model. We then identified the coefficients of the best model with non-influential coefficients dropped. Furthermore, with the optimal model to be used, we trained and validated the dataset to get the RMSE and R^2 of the model.



Lasso regression RMSE and R^2

- Train
- RMSE: 4.0240604(mat) 2.2005416(por)
- R^2: 0.2673442(mat) 0.4270536(por)
- Test
- RMSE: 4.7071885(mat) 2.8858860(por)
- R^2: 0.1065235(mat) 0.1598111(por)

2.5 Random Forest

Furthermore, for tree-based method, two different methods of random forest model were used. For the first model, we conducted 30-fold cross validation to test the model. When testing and validating the model, mtry, the number of variables sampled randomly to represent the model, of 5.5 was held constant. It gave out the result of the following:

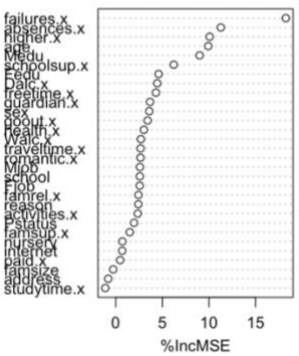
Random Forest RMSE and R^2

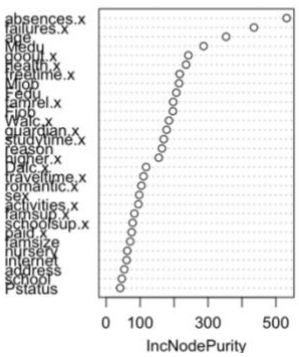
- Train
- RMSE: 3.936823(mat) 2.345072(por)
- R^2: 0.2973943(mat) 0.3521341(por)
- Test
- RMSE: 4.417505 (mat) 2.76248(por)
- R^2: 0.2494508(mat) 0.2834576(por)

While for the second method, regression type of random forest was conducted with 500 decision trees that examined 5 variables at each split. From the second method, we also identified each of the variable's contribution to the model. For mathematics, number of past class failures, number of school absences, and one's desire to take higher education showed the best predictive influence on the model while for Portuguese language, one's desire to take higher education, extra educational support and one's weekly study time showed importance.

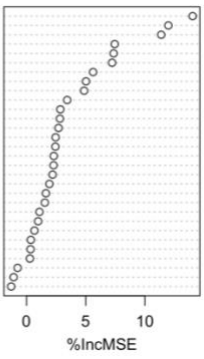
```
## Call:
##  randomForest(formula = G3.x ~ ., data = mat_dataset, mtry = sqrt(ncol(mat_dataset_inpu
t)),      importance = TRUE)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 5
##
##           Mean of squared residuals: 15.2736
##                     % Var explained: 28.18
```

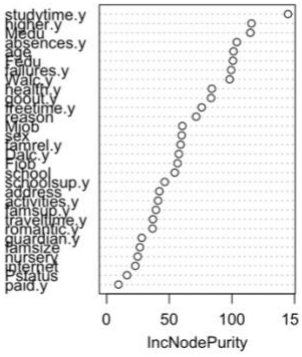
rf_test_mat





rf_test_por





To conclude, Portuguese language showed better linearity and fit for all the models conducted. It showed lesser Root Mean Square Error and higher R^2 compared to mathematics' prediction model results. Furthermore, students desire to take higher education was the variable that had the most influence on both subject's grade. Hence to increase the grade for both, teachers could emphasize the importance of getting higher education to the students to enhance their learnings.

3 Part 3

3.1 Question 3

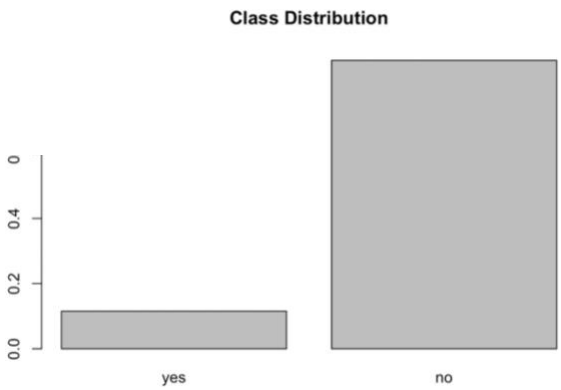
Compare classification models by interpreting and assessing its performance.

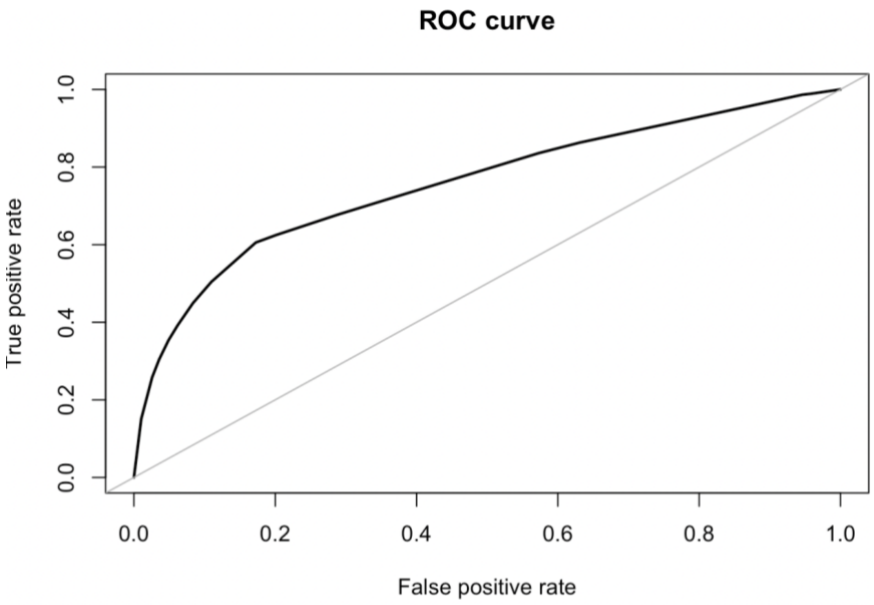
3.2 Logistic Regression

Using 30-fold cross validation, we conducted a logistic regression to train and test the model. From the confusion matrix, we identified the accuracy of 89%. However, the true negative value seemed biased, hence we deep dived into the dataset to see the reason why. As we looked at the proportion of the response dataset, we observed that "yes" (12%) was much less compared to "no" (88%). This may cause class imbalance which may then result to bias. Since our interest was on the client's status of subscription after the phone call, we needed to revise the model because "not subscribed" status was dominated over "subscribed" status. Hence, to make class balanced, we tested undersampling, oversampling, and both (under & over) to adjust the model. With the 3 techniques, we predicted the model using each method's data and evaluated its accuracy then built a decision tree models to identify their scores.

```
## Cross-Validated (10 fold, repeated 3 times) Confusion Matrix
##
## (entries are percentual average cell counts across resamples)
##
##            Reference
## Prediction  yes   no
##        yes  1.9  1.1
##        no   9.7 87.4
##
##  Accuracy (average) : 0.8924
```

```
## bank1_dataset_output
##       yes        no
## 0.1153239 0.8846761
```






```
## Area under the curve (AUC): 0.756
```

Furthermore, Bothsampling method showed the highest AUC score. Using Bothsampling method to conduct the logistic regression, it resulted with accuracy of 73%, sensitivity level of 63% and specificity of 75%. This could then be interpreted as having out of all the positive classes and predictive positive classes, 63% and 75% were correctly predicted with the model.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction yes  no
##        yes  98 304
##        no   58 896
##
##               Accuracy : 0.733
##                 95% CI : (0.7086, 0.7564)
##     No Information Rate : 0.885
##     P-Value [Acc > NIR] : 1
##
##                  Kappa : 0.2223
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.62821
##             Specificity : 0.74667
##          Pos Pred Value : 0.24378
##          Neg Pred Value : 0.93920
##              Prevalence : 0.11504
##          Detection Rate : 0.07227
##    Detection Prevalence : 0.29646
##       Balanced Accuracy : 0.68744
##
##        'Positive' Class : yes
```

3.3 Random Forest

For Random Forest model, mtry of 21 resulted the highest accuracy score. However, as we conduct the prediction with the unadjusted dataset, it seemed to predict majority of the negative(“No”) class and result low sensitivity rate. The model seemed to poorly measure the amount of actual positive cases that was predicted as positive. Hence, we conducted the model again with the use of Bothsampling method.

```
## Random Forest
##
## 3165 samples
##   15 predictor
##    2 classes: 'yes', 'no'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 2849, 2849, 2849, 2849, 2848, 2848, ...
## Resampling results across tuning parameters:
##
##   mtry  Accuracy   Kappa
##    2    0.8872050  0.05306595
##   21    0.8897297  0.20918673
##   41    0.8894112  0.23048059
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was mtry = 21.
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  yes   no
##        yes   25   22
##        no   131 1178
##
##                Accuracy : 0.8872
##                  95% CI : (0.8691, 0.9035)
##     No Information Rate : 0.885
##     P-Value [Acc > NIR] : 0.4198
##
##                   Kappa : 0.2039
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.16026
##             Specificity : 0.98167
##          Pos Pred Value : 0.53191
##          Neg Pred Value : 0.89992
##              Prevalence : 0.11504
##          Detection Rate : 0.01844
##    Detection Prevalence : 0.03466
##       Balanced Accuracy : 0.57096
##
##        'Positive' Class : yes
```

With the adjusted random forest model, we identified that the sensitivity rate has increased from 16% to 35% although its accuracy and specificity have reduced by small amount.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  yes   no
##        yes   55  103
##        no   101 1097
##
##                Accuracy : 0.8496
##                  95% CI : (0.8294, 0.8682)
##     No Information Rate : 0.885
##     P-Value [Acc > NIR] : 1.0000
##
##                   Kappa : 0.2653
##
##  Mcnemar's Test P-Value : 0.9442
##
##             Sensitivity : 0.35256
##             Specificity : 0.91417
##          Pos Pred Value : 0.34810
##          Neg Pred Value : 0.91569
##              Prevalence : 0.11504
##          Detection Rate : 0.04056
##    Detection Prevalence : 0.11652
##       Balanced Accuracy : 0.63337
##
##        'Positive' Class : yes
```

To conclude, as we look at the accuracy rate of both Logistic and Random Forest model, Random Forest model seemed to show better accuracy as it resulted higher accuracy rate. However, since our interest was on correctly identifying the clients who would subscribe to a term deposit, logistic regression model would be preferred as it showed 63% of sensitivity rate compared to Random Forest model which gave 35%. Furthermore, we should keep in mind that accuracy is not the only evaluation to consider when evaluating the model. One should also consider the flexibility and robustness of the model when facing new datasets.