Machine Learning

Coursework
(ST3189)

Name: Gillian Chiow Yoong Shin

Student ID: 180335848/1

Table of Contents

1. COURSEWORK PART 1 3
1.1 Scatterplot Analysis 3
1.2 Principle Components Analysis (PCA) 3
1.3 K-mean Clustering 4
2. COURSEWORK PART 2 6
2.1 Exploratory Data Analysis 6
2.2 Clustering Technique 7
2.3 Model 1: Linear Regression Model 8
2.4 Model 2: Ridge Regression Model 8
2.5 Model 3: Tree-Based Model 9
2.6 Model 4: Random Forest Model 10
2.7 Model Selection 10
3. COURSEWORK PART 3 10
3.1 Exploratory Data Analysis 10
3.2 Model 1: Logistics Regression 11
3.3 Model 2: Random Forest Model 11
3.4 Model 3: Linear Discriminant Analysis Model 11
3.5 Model 4: Classification Tree Model 12
3.6 Model Selection 12

1. COURSEWORK PART 1

1.1 Scatterplot Analysis

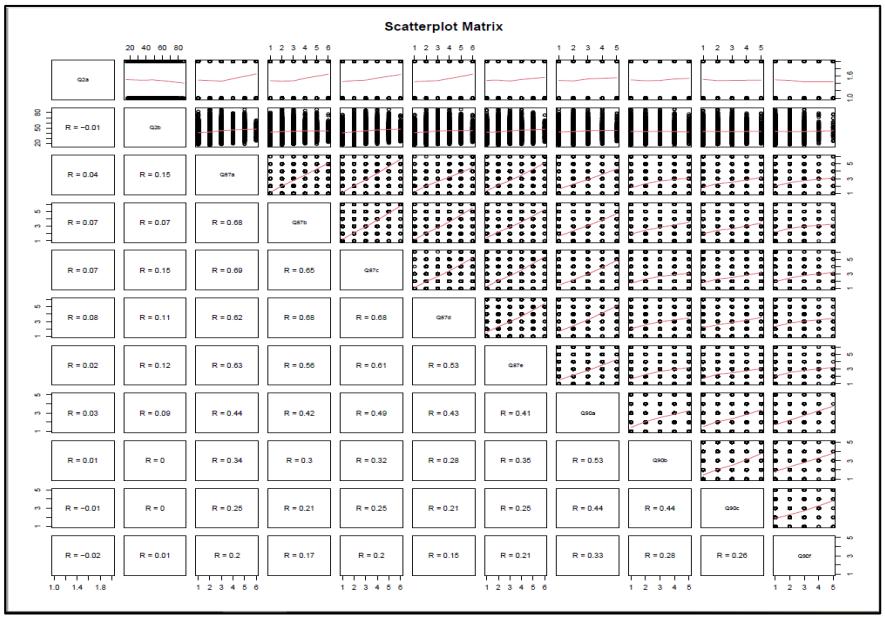

Figure1: Overall scatterplot matrix analysis

Scatterplot matrix above, item analysis can be used to describe how well a group of questions measure one characteristic and helps to analyze questions that are problematic. Q87a and Q87b are moderately correlated in positive way. We deduce that the questions compute different forms of condition on a Likert scale. For example, we can visualize the respondents gave a high rating in Q87a, are also giving Q87b a high rating. Those who rated Q87a low are likely to rate Q87b low. The same observations imply to the Q90a to Q90f.

1.2 Principle Components Analysis (PCA)

```
> pc <- prcomp(data, scale=T)
> summary(pc)
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9    PC10    PC11
Standard deviation     2.0985 1.1874 1.01058 0.97061 0.88319 0.74969 0.71299 0.65201 0.59566 0.56499 0.52321
Proportion of Variance 0.4003 0.1282 0.09284 0.08564 0.07091 0.05109 0.04621 0.03865 0.03226 0.02902 0.02489
Cumulative Proportion  0.4003 0.5285 0.62133 0.70697 0.77788 0.82898 0.87519 0.91384 0.94609 0.97511 1.00000
> ###We take PC1 to PC4 into consideration as they covered 70% of variances
> ##rotate to better understand each component's contribution
> pc$rotation
             PC1         PC2          PC3          PC4          PC5          PC6          PC7          PC8
Q2a  0.03203956 -0.1386327  0.796373784  0.57638908 -6.171166e-02  0.01266193 -0.092896333  0.01060833
Q2b  0.07652230 -0.2204528 -0.584133839  0.76073419  7.105450e-02  0.00515712  0.005490225 -0.12077633
Q87a 0.39103574 -0.1996019 -0.038763673 -0.07849823 -3.148653e-02  0.02786038 -0.179630910 -0.07249015
Q87b 0.37759153 -0.2359578  0.077079602 -0.16741716 -4.488656e-02  0.08133873  0.158131884 -0.36972267
Q87c 0.39652146 -0.2056496 -0.004550283 -0.03679735 -1.796326e-02  0.05172394  0.097909301  0.16394260
Q87d 0.37141006 -0.2534245  0.062704331 -0.09378305 -5.747547e-05  0.14878121  0.360503139 -0.19974893
Q87e 0.36263461 -0.1259478 -0.059239797 -0.08174241 -3.299479e-02 -0.14466435 -0.712223779  0.37711846
Q90a 0.33784962  0.3007859  0.002609147  0.12630062  1.210966e-01 -0.20735048  0.495518337  0.62792742
Q90b 0.27485090  0.4436706  0.054692725  0.05645151  2.715430e-01 -0.62004729 -0.100309447 -0.47561321
Q90c 0.22363116  0.5038874  0.015633656  0.08498139  3.729994e-01  0.71576928 -0.177731352 -0.06749933
Q90f 0.17680118  0.4160141 -0.080175825  0.10806045 -8.713190e-01  0.08308652 -0.008177733 -0.09643027
             PC9          PC10         PC11
```

Figure2: Table of Principle components analysis in over

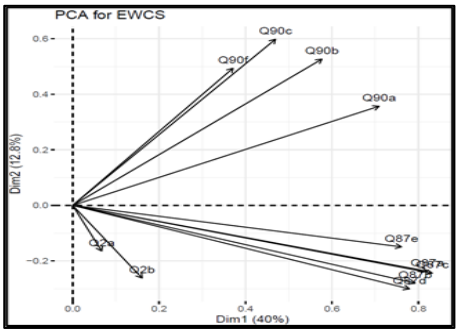

Figure3: Loadings plot for PCA

In Figure 2, we observe that PC1 to PC4 capture about 70.697% of variance. We investigate detail of PC1 to PC4 by using pc$rotation. PC1 suggest that gender and age do not contribute much to the questionnaire, but they are relatively important component in PC2, PC3 and PC4. PC2 results of only Q90a to Q90f return positive value heavy loading, shows they are measuring a same characteristic, and Figure 3 shows Q90a to Q90f have large positive loadings supports this statement. On the other side, PC1 indicates Q87a to Q87e contribute heavily, we can say that the set measures another characteristic, and so made up a reliable questionnaire, which is measuring on two characteristics.

We refer to the report (Eurofound, 2017) and understand that Q87a to Q87e are questions fall under “3.3 Maintaining and promoting health and well-being, sub-section: General health and subjective well-being”. Correlation value (Figure3) shows minimal correlation of gender and Q87a to Q87e, as all correlations are positive, we can say that female has slightly higher score than male regarding the questions about “subjective well-being measured through the World Health Organization’s well-being index –5” (Psykiatric Center North Zealand, n.d.). The same

result can be obtained from PCA1 and PCA2 in Figure2, Q2a shows small contribution to PC1 and PC2, same direction shown for Q2a and Q87a to Q87e. Next, we assess the set of questions separately to see if there are any different findings.

```
> pc1 <- prcomp(data1, scale=T)
> summary(pc1)
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7
Standard deviation     1.888 1.0093 0.9797 0.70465 0.59905 0.57278 0.52447
Proportion of Variance 0.509 0.1455 0.1371 0.07093 0.05127 0.04687 0.03929
Cumulative Proportion  0.509 0.6545 0.7916 0.86257 0.91384 0.96071 1.00000
> ###We analyze PC1 and PC2 as they cover 66% of variances
> ##rotate to better understand each component's contribution
> pc1$rotation
             PC1         PC2         PC3         PC4         PC5         PC6         PC7
Q2a  0.0463111 -0.813293014  0.572149399  0.09115846 -0.02496702  0.003687395 -0.01039259
Q2b  0.1045925  0.568544243  0.808286065 -0.05981633 -0.02016853  0.088016925  0.02756349
Q87a 0.4559757  0.036721607 -0.038440568  0.11980414 -0.60856900 -0.310174532 -0.55528596
Q87b 0.4478097 -0.078708085 -0.111160328 -0.31655459 -0.41969219  0.542011821  0.45911838
Q87c 0.4597845  0.001121161 -0.002294195 -0.06422286  0.21333168 -0.688496530  0.51471280
Q87d 0.4414829 -0.064790225 -0.035748293 -0.48156385  0.56971585  0.170358549 -0.46272587
Q87e 0.4149209  0.059744857 -0.064816694  0.79845099  0.28702885  0.314968034  0.03091234
```

Figure4: PCA on set of questions (Q2a-Q87e), gender (Q2a) and age (Q2b)

We assess only the set of questions (Q87a-Q87e) with the Q2a and Q2b. The result is shown in Figure4. From the summary, we observe that PC1 and PC2 capture about 65.5% of variance, and we use rotation to have clearer view on different component in PC1. Q2a shows only 0.046 contribution to PC1, with positive direction, which both observations suggest that female has slightly higher score than male. With the given value labels, we deduce that a low rating conveys a better level of psychological well-being, whereas high rating suggests that the person has to be aware of their mental health. A low correlation (Figure1) associate between age and Q87a to Q87e. By viewing the loadings plot from Figure3, the positive correlation indicates that by increasing the age, the rating of respective question from Q87a to Q87e will increase too. PC1 and PC2 in Figure2 has providing support to this statement as the value of Q2b and Q87a to Q87e return same direction. The correlation value of age and Q87a to Q87e are slightly higher than the correlation coefficient of gender and Q87a to Q87e (Figure3). A female should be more aware of their psychological well when their age increases.

```
Importance of components:
                          PC1    PC2    PC3    PC4     PC5     PC6
Standard deviation     1.4733 1.0075 0.9993 0.8854 0.76002 0.67405
Proportion of Variance 0.3618 0.1692 0.1664 0.1306 0.09627 0.07572
Cumulative Proportion  0.3618 0.5309 0.6974 0.8280 0.92428 1.00000
> #PC1, PC2 and PC3 capture about 70% of variance
> pc1$rotation
             PC1         PC2         PC3          PC4         PC5         PC6
Q2a  0.007869773 -0.62694501  0.76826964 -0.100032356  0.07157154  0.03886937
Q2b  0.042846960  0.77247842  0.61702699  0.003148056  0.07699203  0.12157811
Q90a 0.549739520  0.03403863  0.10806605  0.121952533 -0.34995992 -0.74000351
Q90b 0.536987943 -0.07119717 -0.02043908  0.254094121 -0.46310743  0.65354789
Q90c 0.500422950 -0.04892014 -0.08096959  0.309599315  0.80244775  0.02921460
Q90f 0.396370388  0.03995244 -0.10191839 -0.902604918  0.08992728  0.09013549
```

Figure5: PCA on set of questions (Q90a-Q90f), gender (Q2a) and age (Q2b)

In the report, it shows Q90a to Q90f are probably related to dimensions of work engagement questionnaires. Referring to Figure2 and perform another set of PCA by excluding Q87a to Q87e from the questionnaire and see the result of how important the age and gender to Q90a to Q90f are. We observe that age and gender to Q90a to Q90f have very low and negative correlation value in Figure3. However, PC1 of Figure5, gender and age are positive components with low value. PCA concludes that gender and age have a relationship to both sets of questions. Some study has proven that subjective well-being and work engagement are correlated. In Figure2 of PC1, we observe that high rating on subjective well-being is positively correlated to high rating on work engagement.

1.3 K-mean Clustering

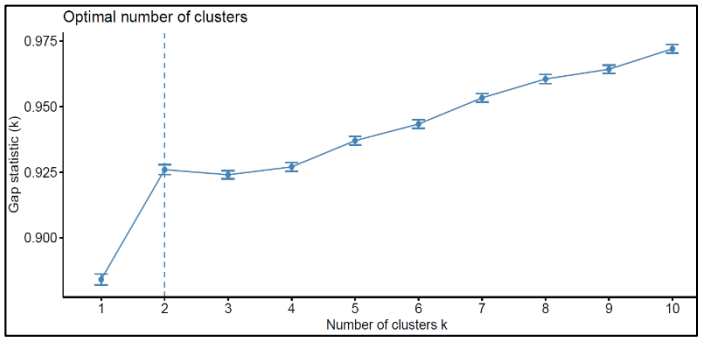

Figure6: Determine optimal number of clusters using gap statistics

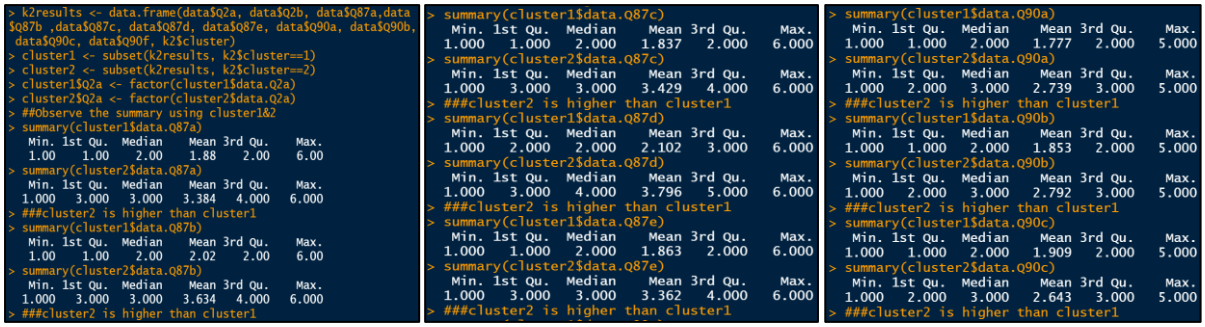

Figure7: K-means clustering (gender)

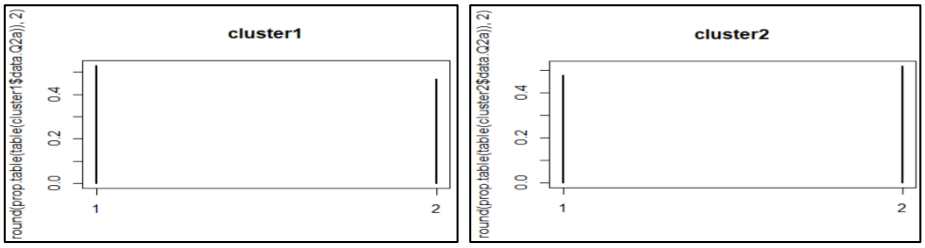

Figure8: Portion of gender in cluster

Next, we use gap statistics to check appropriate k for k-mean clustering. Figure 6 determined k=2. Figure7 shows summary of each question and all questions show cluster2 has higher mean and median score than cluster1 in Q87a to Q90f. We check the proportion of gender in cluster, it results cluster 1 has 0.53 male, 0.47 female; cluster 2 has 0.48 male, 0.52 female. Cluster 2 has more female than cluster 1.

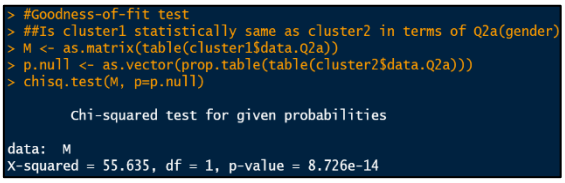

Figure9: Goodness-of-fit test to check significancy (gender)

Goodness-of-fit test is performed to check is cluster1 statistically identical as cluster2 in terms of Q2a. As p-value suggested, with 95% of confidence level, we conclude that gender is a differentiator of the dataset. We can perform the same steps to check whether the age is a differentiator by grouping age into categories (15-24, 25-34, 35-44, 45-54, 55-64, 65-74, 75+).

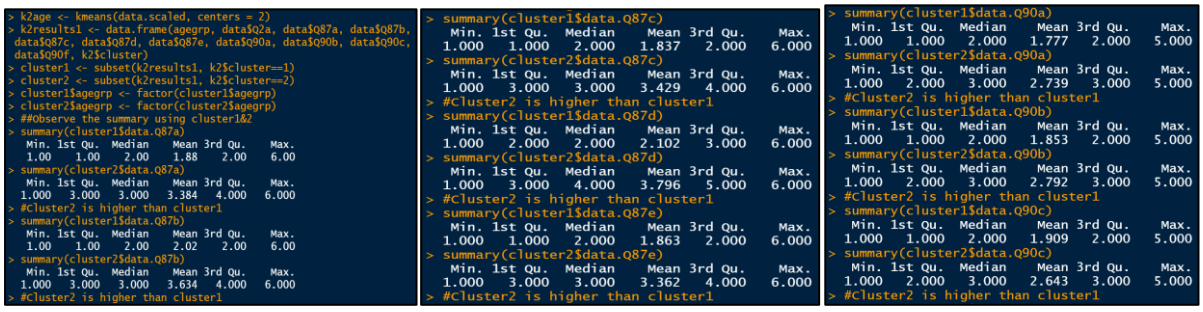

Figure10: K-means clustering (age)

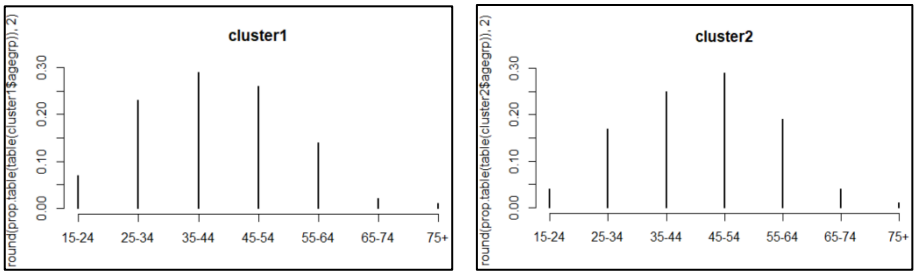

Figure11: Portion of age group in each cluster

Figure10 summarized questions Q87a to Q90f in cluster 1 and 2. All questions show cluster2 has higher mean and median score than cluster1 in Q87a to Q90f. We check the proportion of age in cluster. Figure11 below shows 59% of the individuals in cluster 1 and 46% of the individuals in cluster 2 are comprised with age below 44. From Figure1, we observe that most of the dots in value 6 of Q87a-Q87e and value 5 of Q90a-Q90f are gathered at the middle, it shows the person that has rated value 6 are mostly come from mid-age range respondent. It is likely to say that by getting older, the score will be getting higher.

```
> #Goodness-of-fit test
> #Is cluster 1 statistically same as cluster 2 in terms
 Gender
> M <- as.matrix(table(cluster1$agegrp))
> p.null <- as.vector(prop.table(table(cluster2$agegrp)))
> chisq.test(M, p=p.null)

        Chi-squared test for given probabilities

data:  M
X-squared = 370.1, df = 6, p-value < 2.2e-16
```

Figure12: Goodness-of -fit test to check significancy (age)

As the p-value suggested, with 95% of confidence level, we conclude that age is a differentiator of the dataset as the cluster1 age group proportion is different from cluster2 age group proportion.

2. COURSEWORK PART 2

2.1 Exploratory Data Analysis

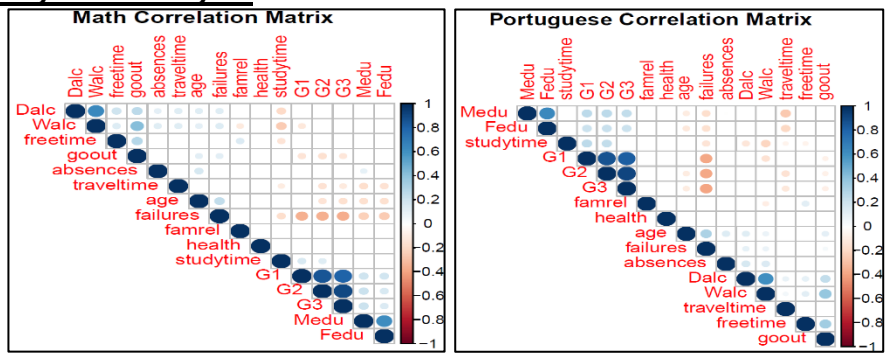

Figure13: Correlation Matrix of Math and Portuguese grade

We try a range of models and prediction methods such as logistics regression, decision trees, random forests to explore task 2. We have 33 variables each in the two datasets. The information including personal details, family information about the student, study, and lifestyle habits, and three period grades. Some additional information provided such as reason of choosing the school, attended nursery or not, internet access, travel time to school, absences, and number of past class failures. Given the information that the G1 and G2 are very well predictors for G3, and this is because G1 and G2 correspond to the first two period grades whereas G3 is the final period grade. The correlation matrix supports this point as G1 and G2 have strong correlation with G3. One of the significant findings is Dalc, Walc, freetime and goout have a strong positive correlation, and a negative correlation with studytime. As those who have more free time will have lesser time to study and higher level of going out with friends, thus the level of alcohol consumption will increase.

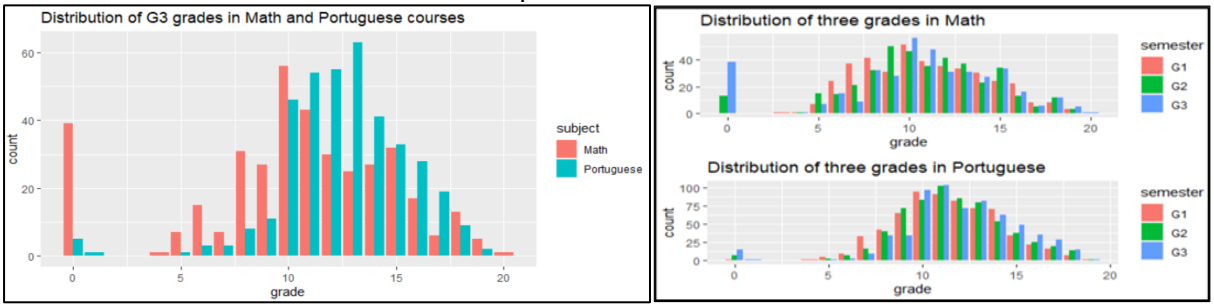

Figure14: Distribution of G3 grades in Math and Portuguese course

To understand the relationship of the two datasets and the merged dataset, a few graphs are generated with the explanation using mean. From Figure14, most of the students scored 13 for Portuguese, scored 10 for Math. We compared the mean value of two datasets and the merged dataset. The mean G3 of the students in merged dataset, Portuguese course are higher than the math course. Mean of G3 for subset of students in both math and Portuguese (10.39) has almost the same value to the dataset that consist all the students in math (10.42). For Portuguese, the mean G3 of the subset of students in both math and Portuguese (12.52) is slightly different to the datasets that consist all the students in Portuguese (11.91). The gap could be due to all the students in the math course are included in merged dataset, but there are more students in Portuguese dataset, that are not in the merged dataset. By visualizing the math and Portuguese grade distribution, math grade has a decreasing trend with a very high increasing number of students with a grade of 0, whereas an increasing trend is spotted in Portuguese grade. These results can be obtained from mean value, all students in math as well as the subset of math students the mean grade slightly decreases as the semester progress, while the students in Portuguese are another way round. We can summarize that there are not too many differences in grades between the merged dataset and two datasets including all students in math and Portuguese.

Males tend to score better in math, whereas females score better in Portuguese. Math grades between the two schools is similar and students from Gabriel Pereira outperform those from Mousinho da Silveira in Portuguese. For Math subject, students whose mothers have jobs in health outperform other students. For Portuguese subject, students whose mothers stay at home seem to underperform compared to other students. In both math and Portuguese, students whose fathers are teachers outperform other students. An interesting insight, the average grade of students with the worst family relationships is scoring higher than students with better relationships for math. Lower travel time, more study time, less past class failure, less alcohol consumption, less number of class absences, having intention to pursue higher education, having internet access, and not comitted into romantic relationship are important elements to achieve higher grade for both Math and Portuguese.

2.2 Clustering Technique

Clustering method is used to observe the relationship across the variables.

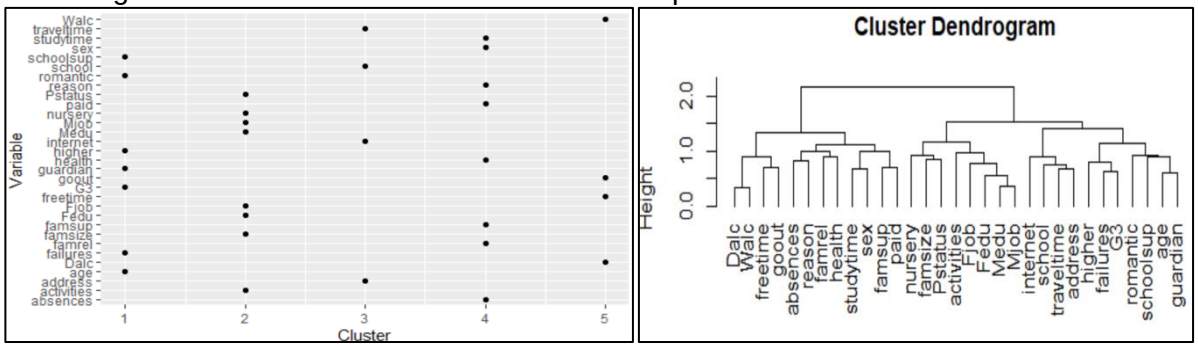

Figure15: Hierarchical clustering analysis for Math

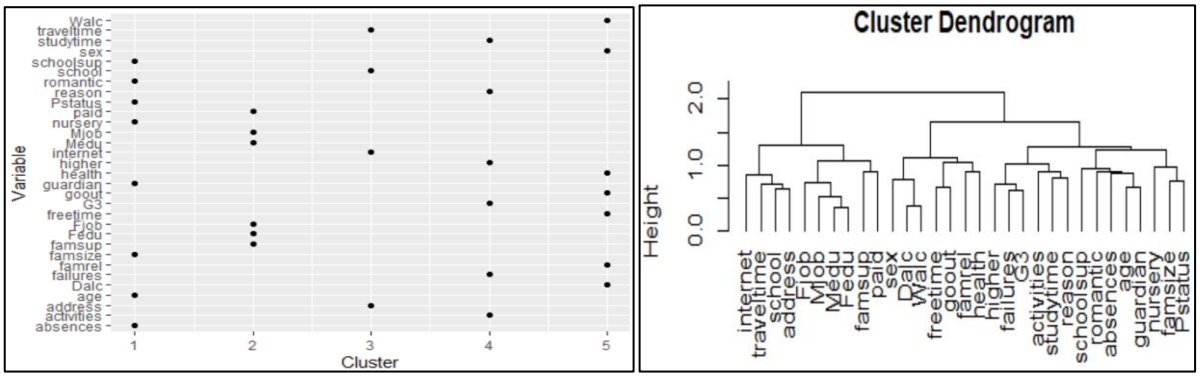

Figure16: Hierarchical clustering analysis for Portuguese

We make k=5 to see the clusters, it makes sense to our first interpretation previously that the relationship between Dalc, Walc, free time and go out. For math G3, failures, age, student's guardian, intention to take higher education, with a romantic relationship, and extra educational support are clustered together. Cluster3 shows the relationship of student living in urban area may have internet access and the distance between school and address could affect the travel time. For Portuguese G3, reason to choose the school, study time, intention to purse higher education, extra-curricular activities, failures are clustered together. Cluster2 explain relationship between Fjob, Mjob, Fedu, Medu, family educational support, extra paid classes. As parents' education level may affect the parents' job, higher education level of parents realizes the important of education to children, hence they will provide family educational support and consider extra paid classes to aid their child's studies. Whether the parents have a job to pay for the extra classes is a factor too.

2.3 Model 1: Linear Regression Model

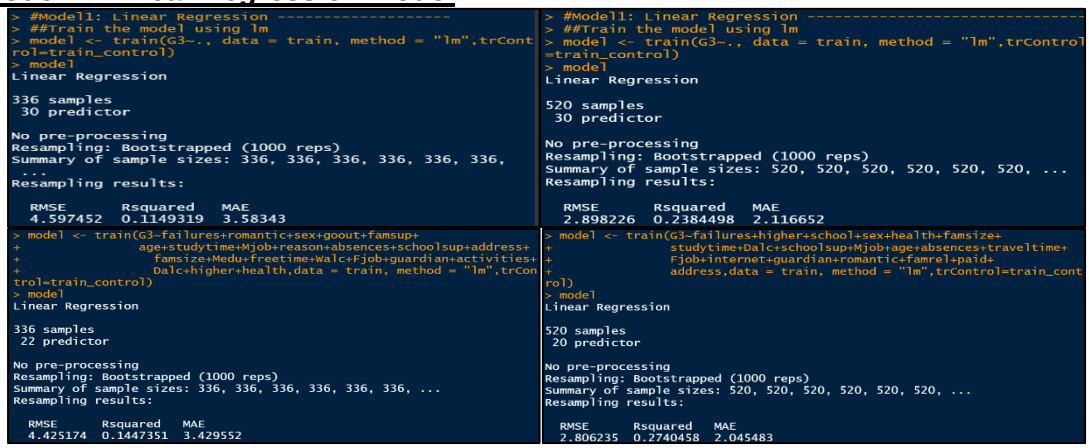

Figure17: Find the lower MAE and highest R^2 for our training model (left:Math right:Por)

In this section, we perform linear regression on math dataset using all variables and selected significant variables by using varImp(). We train our model using the significant variables, and it returns a higher R^2. Prediction was made using the test set. The R^2 and Root Mean Squared Error (RMSE) were calculated. For Math, the R^2 is 10.77% around its mean with 4.44 of RMSE. For Portuguese, R^2 is 27.40 around its mean with 2.04 of RMSE. It is reasonable to have low R^2but it may not be a good predictor. "When the interest is in the relationship between variables, not in prediction, the R^2 is less important." (Frost, 2019), R^2suggests us to analze how does the variables affects G3 outcome. We need get a trustable relationship between the variables and G3, we cannot expect all the variables provide good explanation to the G3. For example, health and gender cannot directly explain the prediction as there are lots of variables that are not taken into consideration that will affect health and gender.

2.4 Model 2: Ridge Regression Model

Ridge regression is used as the predictive model. Choosing λ using cross-validation and fitting ridge regression model on train set.

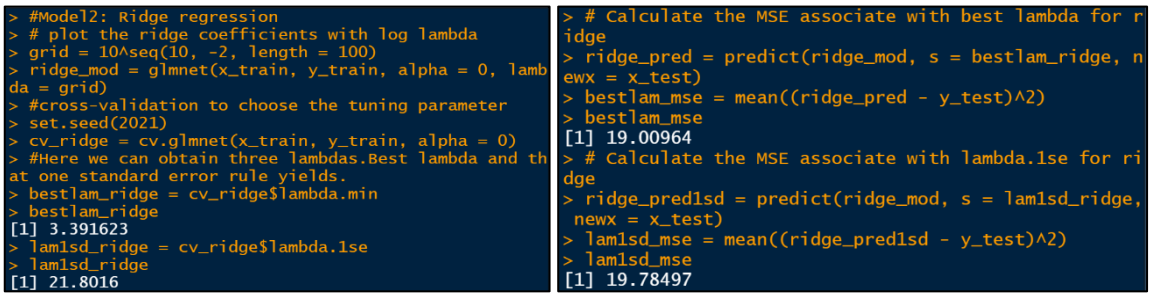

Figure18: Ridge regression and obtain best lambda (math)

We obtain the full model with the best tunning parameter and check the R^2 and RMSE.

```
> # Now obatin the full model with best tunning parameter.
> set.seed(2021)
> x = model.matrix(school1$G3~., school1)[,-1]
> y = as.matrix(school1$G3)
> out = glmnet(x, y, alpha = 0, lambda = grid)
> ridge_coef = predict(out, type = "coefficients", s = bestlam_ridge)[1:40,]
> R2_math_ridge<-R2(ridge_pred, y_test)
> R2_math_ridge
          1
[1,] 0.09981431
> RMSE_math_ridge<-RMSE(ridge_pred, y_test)
> RMSE_math_ridge
[1] 4.360005
```

```
> # Now obatin the full model with best tunning parameter.
> set.seed(2021)
> x = model.matrix(school2$G3~., school2)[,-1]
> y = as.matrix(school2$G3)
> out = glmnet(x, y, alpha = 0, lambda = grid)
> ridge_coef = predict(out, type = "coefficients", s = bestlam_ridge)[1:40,]
> R2_por_ridge<-R2(ridge_pred, y_test)
> R2_por_ridge
          1
[1,] 0.288972
> RMSE_por_ridge<-RMSE(ridge_pred, y_test)
> RMSE_por_ridge
[1] 2.694546
```

Figure19: Full model with best tunning parameter (left:Math,right:Por)

2.5 Model 3: Tree-Based Model

```
> #Model3: Tree Based----------------------------------
> library(rpart)
> #Cross validation of the model
> set.seed(2021)
> tc <- trainControl(method = "cv",
+                    number = 10)
> cp.grid <- expand.grid(cp = seq(0, 0.03, 0.001))
> data.tree.cv <- train(G3~.,
+                       data = train,
+                       method = "rpart",
+                       trControl = tc,
+                       tuneGrid = cp.grid)
> data.tree.cv
RMSE was used to select the optimal
 model using the smallest value.
The final value used for the model was
 cp = 0.024.
```

```
> data.tree.cv <- rpart(G3~., data=train, cp=.024)
> data.tree.cv
n= 318

node), split, n, deviance, yval
      * denotes terminal node

1) root 318 6646.4810 10.443400
  2) failures>=0.5 64 1445.4380  6.906250
    4) absences< 1 20  220.5500  1.350000 *
    5) absences>=1 44  326.7955  9.431818 *
  3) failures< 0.5 254 4198.5550 11.334650
    6) Mjob=at_home,other 132 2057.9700 10.348480 *
    7) Mjob=health,services,teacher 122 1873.3200 12.40
1640 *
> fancyRpartPlot(data.tree.cv, main="Decision Tree",cex
=0.8,cex.main=1)
```

Figure26: Tree-based model, finding the best cp (math)

We use cross-validation to get the best tune grip. The result shows the final value used for the model was cp=0.024, we will proceed using cp=0.024 for our train model.

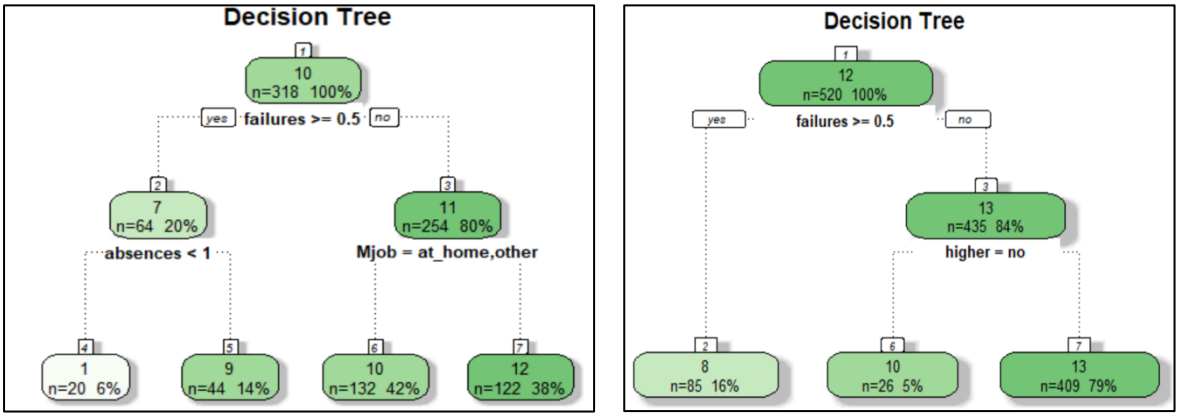

Figure21: Decision tree model (left:Math,right:Por)

We found the top three contributors to this model are number of past class failures, absences, and mother's job. We can have a clearer understanding by using rpart plot to show the decision tree. For Math, 100% of the data would achieve an average of 10 marks. For those who have past class failure scored an average of 7; 11 marks in average for students who do not have any failure in record. For students' mother work in health care related, civil services and teacher, student's average mark increased to 12. An interesting finding is students having less than 1 day of school absences (1mark), are getting lower average score than those who have more than 1 day of school absences (9marks). We have 12 marks in average for Portuguese and student fails once in class will get 8 marks in average. Students who do not have failure record and have no intention to pursue higher education will get 10 marks in average and students who wish to pursue higher education have 13 marks in average.

```
> predict_tree <- predict(data.tree.cv, newdata = test)
> R2_math_tree<-R2(predict_tree, test$G3)
> R2_math_tree
[1] 0.08574884
> RMSE_math_tree<-RMSE(predict_tree, test$G3)
> RMSE_math_tree
[1] 4.657071
```

```
> predict_tree <- predict(data.tree.cv, newdata
 = test)
> R2_math_tree<-R2(predict_tree, test$G3)
> R2_math_tree
[1] 0.1344559
> RMSE_math_tree<-RMSE(predict_tree, test$G3)
> RMSE_math_tree
[1] 3.034683
```

Figure22: Compute RMSE and R^2 (left:Math,right:Por)

We yield a $R^2$of 8.58% and RMSE of 4.66 for math, 13.45% and 3.03 for Portuguese.

2.6 Model 4: Random Forest Model

```
> #Model4: Random Forest-------------------------------------
-------
> library(randomForest)
> set.seed(2021)
> ##We choose the number of trees to increase % var explained
> data.random.forest <- randomForest(G3~.,
+                                    data = train)
> print(data.random.forest)

Call:
 randomForest(formula = G3 ~ ., data = train)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 10

          Mean of squared residuals: 14.65873
                    % Var explained: 29.87
```

```
> ##We use ntree=100
> data.random.forest2 <- randomForest(G3~.,
+                                     data = train,
+                                     ntree = 100,
+                                     importance = TRUE)
> print(data.random.forest2)

Call:
 randomForest(formula = G3 ~ ., data = train, ntree = 100, im
portance = TRUE)
               Type of random forest: regression
                     Number of trees: 100
No. of variables tried at each split: 10

          Mean of squared residuals: 14.5306
                    % Var explained: 30.48
```

Figure23: Random forest model evaluation (math)

We want to choose the best number of trees to increase % variables explained for our random forest model prediction. After visualizing the plot, we plant 100 of trees, it returns an increased % variables explained. Using varImp(), we realize the top contributors for math are “absences”, “failures” and “Medu”; for Portuguese are “failures”, “higher”,”school”. Then, we found the best mtry is 18. We proceed to get R^2 and RMSE.

```
> #Aims to find the best mtry
> data.random.forest3_training <-
+   train(G3~., data = train, method = "rf",
+   ntree = 100,tuneGrid = parameters_rf,
+   trControl = tc,importance = TRUE)

RMSE was used to select the optimal model using the
 smallest value.
The final value used for the model was mtry = 18.
```

```
> #Aims to find the best mtry
> data.random.forest3_training <-
+   train(G3~., data = train, method = "rf",
+   ntree = 300,tuneGrid = parameters_rf,
+   trControl = tc,importance = TRUE)

RMSE was used to select the optimal model
 using the smallest value.
The final value used for the model was mtry = 10.
```

Figure24: Get the best mtry for our training model (left:Math,right:Por)

```
> data.random.forest3 <- randomForest(G3~.,data = train,ntree = 100,
+                                     mtry = 18,importance = TRUE)
> predictrf <- predict(data.random.forest3, test)
> R2_math_rf<-R2(predictrf, test$G3)
> R2_math_rf
[1] 0.1090108
> RMSE_math_rf<-RMSE(predictrf, test$G3)
> RMSE_math_rf
[1] 4.42033
```

```
> #Make prediction
> predictrf <- predict(data.random.forest3, test)
> R2_math_rf<-R2(predictrf, test$G3)
> R2_math_rf
[1] 0.2540453
> RMSE_math_rf<-RMSE(predictrf, test$G3)
> RMSE_math_rf
[1] 2.765505
```

Figure25: Compute RMSE and $R^2$of random forest model (left:Math,right:Por)

2.7 Model Selection

Lastly, we compare each model accuracy using a table. The results show “failures”, “sex”, “schoolsup”, “higher”, “Mjob”, “absences”, and “Medu” are top predictors to math grade, while “failures” and “higher” are two of the most important predictors to Portuguese grade. Random forest could be a better predict model for math grade as it has the highest R^2 and lowest RMSE, while ridge regression and linear regression could be useful to predict Portuguese grade.

model	R2	RMSE	Top_Contributor_school
linear_regression	10.77%	4.44	failures,sex,schoolsup
ridge_regression	10%	4.36	higher,failures,schoolsup
tree_based	8.57%	4.66	failures,absences,Mjob
random_forest	10.90%	4.42	absences,failures,Medu

model	R2	RMSE	Top_Contributor_school
linear_regression	27.40%	2.05	failures,higher,school
ridge_regression	28.90%	2.7	higher,failures,schoolsup
tree_based	13.45%	3.05	failures,higher
random_forest	25.40%	2.77	failures,higher,school

Figure26: Model comparison table (left:Math,right:Por)

3. COURSEWORK PART 3

3.1 Exploratory Data Analysis

Next, we were presented with a classification problem. The dataset is about bank marketing, and we have 4521 observations and 17 variables. We perform simple exploratory data analysis on the people subscribed to a term deposit. Minimum age of subscriber is 19 and maximum age is 87, median age is 39. 37.8% of subscribers are over 60 years old, followed by 14.2% of subscribers at the age group of lower than 30 years old and the lowest subscription age group is those who aged between 30 to 60 (10.2%). We observed that a higher subscription among retired people (23.5%) and student (22.6%). Divorced individuals (14.6%) slightly more likely to subscribe than single (14%) or married (9.9%). In terms of education level, those who received tertiary education tend to subscribe a term deposit (14.3%). Most of the individuals in dataset have settled their dues on time (98.3%), only 76 of

them have not do so, and among 76 of them, 9 subscribed to term deposit. This information does not provide much information to us, chi-squared test is performed, and the returned p-value shows it is highly insignificant predictor (p-value=1).

Without housing loan and personal loan, customers are more likely to subscribe. The record about last contact of the year shows May has the highest contacts made, and Friday is the day of the weeks has the most contact made. A higher rate of successful subscriptions made were made in March, September, October, and in December. Sunday has the lowest successful subscription rate. It takes at least 30 seconds and a median of 442 seconds (approximately 7 minutes) for customers subscribe to the term deposit. Higher median (710) and mean (1572) for those who subscribe to the term deposit, compared to those who do not subscribe (median:419.5, mean:1403.2). We summarized that “poutcome”, “pday”, “campaign”, and “previous” are associated to the campaigns previously and it does not have significant contribution on the current target feature (y) because almost all people contacted during this campaign should be new. We omitted “contact” because it does not show any significant contribution to the outcome.

3.2 Model 1: Logistics Regression

We use summary(mod1) to extract the important contributor for this model and it returns “duration”, “month” and “day” are the top contributors. Most of the variables are significant by having less than 0.05 p-value, so we use all the variables to make prediction.

```
> #Model1: Logistics Regression------------------------------
> #Split the Training / Testing data
> library(caTools)
> bank=read.table("bank.csv",stringsAsFactors = TRUE, sep=";",header=T
RUE)
> bank <- bank[, -which(colnames(bank) %in% c("poutcome", "pdays","pre
vious","campaign","contact"))]
> index<-sample(nrow(bank),0.70*nrow(bank),replace = F)
> train<-bank[index,]
> test<-bank[-index,]
> #Building model
> mod1<-glm(y~.,data = train,family = "binomial")
```

```
> conf_mat1<-confusionMatrix(as.factor(predicti
on),test$y,positive="yes")
> conf_mat1
Confusion Matrix and Statistics

          Reference
Prediction   no  yes
       no  1180  117
       yes   27   33

               Accuracy : 0.8939
                 95% CI : (0.8763, 0.9098)
    No Information Rate : 0.8895
    P-Value [Acc > NIR] : 0.3201
```

Figure27: Logistics regressions and its outcome

The confusion matrix shows that the classifier prediction for “yes” is 600 and “no” is 1297. In reference case, we have 1207 not subscribe to the term deposit and 150 subscribe to the term deposit. For this model, we achieve 89.39% of accuracy.

3.3 Model 2: Random Forest Model

```
> #Model2: Random Forest------------------------------
----------
> library(randomForest)
> rf<-randomForest(y~.,data = train,ntree = 2000,
+                  importance = TRUE)
> importance    <- importance(rf)
```

```
> rankImportance
          Variables Importance Rank
age             age      74.60   #4
job             job      60.32   #5
marital     marital      19.98   #8
education education      23.27   #7
default     default       2.33  #11
balance     balance      84.46   #3
housing     housing      12.55   #9
loan           loan       6.57  #10
month         month      95.34   #2
duration   duration     206.52   #1
day             day      55.80   #6
```

Figure28: Important variables for random forest model

There are a slightly different top three contributors in random forest model, we have “duration”, “month” and “balance”. Random forest model prediction has 89.54% of accuracy rate. It shows that the classifier prediction for “yes” is 94 and “no” is 1263. In reference case, we have 1207 not subscribe to the term deposit and 150 subscribe to the term deposit.

3.4 Model 3: Linear Discriminant Analysis Model

```
> #Model3: Linear Discriminant analysis----------------------
> library(MASS)
> # Linear discriminant regression
> train_lda <- lda(y~duration+balance+job, data = train)
> train_lda
Call:
lda(y ~ duration + balance + job, data = train)

Prior probabilities of groups:
      no      yes
0.886536 0.113464

Group means:
    duration  balance jobblue-collar jobentrepreneur jobhousemaid
no  226.6542 1435.314       0.2099822       0.03814617   0.02352941
yes 546.3733 1640.618       0.1364903       0.02228412   0.03064067
    jobmanagement jobretired jobself-employed jobservices
no      0.2156863 0.04491979       0.04171123  0.09376114
yes     0.2506964 0.10584958       0.04178273  0.06406685
    jobstudent jobtechnician jobunemployed  jobunknown
no  0.01532977     0.1718360     0.0285205 0.007843137
yes 0.03899721     0.1504178     0.0362117 0.013927577
```

```
Coefficients of linear discriminants:
                        LD1
duration       3.954708e-03
balance        1.577066e-05
jobblue-collar -5.707557e-01
jobentrepreneur -3.554448e-01
jobhousemaid   -2.416057e-01
jobmanagement   8.772423e-02
jobretired      7.856283e-01
jobself-employed -4.196915e-02
jobservices    -2.448135e-01
jobstudent      1.023337e+00
jobtechnician  -1.434714e-01
jobunemployed  -2.133639e-01
jobunknown      1.860412e-01
```

Figure29: Linear discriminant analysis model

From the ranking above, we realize duration, job and balance have significant contribution, so we use linear discriminant model output to check what are some of the findings we could get. We realized that the prior probabilities "no" = 0.8865 and "yes" = 0.1135, which means 88.65% of the training observations are individuals who did not subscribe to the term deposit, and 11.35% subscribe to the term deposit. The chosen variables "duration", "job", and "balance" provide group means, are the average of each predictor within each class. The result shows customers that more likely to subscribe, on average, a credit balance of $1640.618 and 546.373 seconds of contact duration. They are more likely to be housemaid, management, retired people, self-employed, student, unemployed. For non-subscriber, they are likely working as blue-collar, entrepreneur, services and technician, with a credit balance $1435.314 with 226.654 seconds of contact duration, which is lower than those who subscribed to the term deposit. Coefficients of linear discriminants result indicated the linear combination of credit balance and job that used to construct the linear discriminant analysis. Therefore, we will have an equation $Y = 3.96e^{-3} \times duration + 1.58e^{-6} \times balance - 5.71e^{-1} \times jobbluecollar + \cdots - 2.13e^{-1} \times jobunemployed$. If the return value is large, the classifier will predict that individual will subscribe to the term deposit. The classifier predicted 69 subscribe to the term deposit and 1288 do not subscribe to the term deposit, with 89.09% of accuracy.

3.5 Model 4: Classification Tree Model

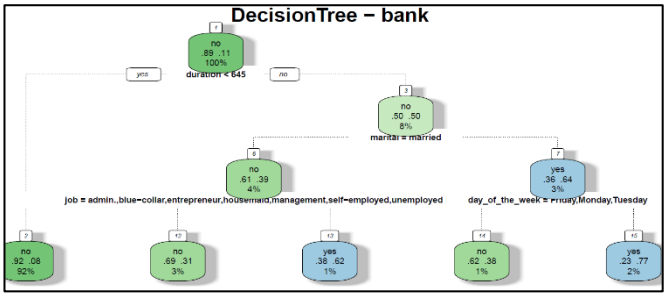

Figure30: Classification Tree Model

From the diagram above, we know that the contact duration less than 645 seconds will have 92% of not subscription chance. If the duration is more than 645 seconds, customer's marital status = married, and occupation is related to admin, blue-collar, entrepreneur, housemaid, management, self-employed or unemployed, the chances of not subscribing dropped to 69%. The contact duration last longer than 645, but not married, and day of the week that contact the customer is Friday, Monday or Tuesday, the not subscription rate dropped to 62%. Classification tree model prediction has 87.91% of accuracy rate. It shows that the classifier prediction for "yes" is 50 and "no" is 1307. In reference case, we have 1195 not subscribe to the term deposit and 162 subscribe to the term deposit.

3.6 Model Selection

	model	accuracy	Top_Contributor
1	logistics_regression	89.39%	day_of_the_week,duration,month
2	random_forest	89.54%	duration,month,balance
3	LDA	89.09%	-
4	Decision Tree	88.06%	duration,marital,job,day_of_the_week
5	SVM	87.91	-

Figure31: Model Comparison Table

Lastly, we compared the model using the accuracy. Random forest is the best classification model to this problem. Significant variables are duration, day of the week, month, balance, marital status and job. However, if we are predicting the data for next campaign, it is better to remove duration, as we will not known the contact duration before the calls. As shown above, the months with higher possibility of successful subscription are March, September, October, and in December. We could probably hold the next campaign in these months and avoid contact the customer on Sunday. We can do a screening on the customers' details such as job, marital status and credit balance to target the customers before we contact them.

References

1. Eurofound. (2017). *Sixth European Working Conditions Survey – Overview report*. Publications Office of the European Union, Luxembourg. https://doi.org/10.2806/422172

2. Frost, J. (2019, October 24). *How to Interpret Adjusted R-Squared and Predicted R-Squared in Regression Analysis*. Statistics By Jim. https://statisticsbyjim.com/regression/interpret-adjusted-r-squared-predicted-r-squared-regression/

3. Psykiatric Center North Zealand. (n.d.). *WHO-5 Questionnaires*. Https://Www.Psykiatri-Regionh.Dk/Who-5/Who-5-Questionnaires/Pages/Default.Aspx. Retrieved March 3, 2021, from https://www.psykiatri-regionh.dk/who-5/who-5-questionnaires/Pages/default.aspx