ST3189 Coursework report

Name: See Yan Yu

Student Number: 190399010

Table of Contents

[Part 1 2](#_Toc67949801)

[1.1: Overview 2](#_Toc67949802)

[1.2: Plot Matrix 2](#_Toc67949803)

[1.3: Principal Component Analysis (PCA) 2](#_Toc67949804)

[1.4: K-means Clustering 3](#_Toc67949805)

[1.4.1: K-means Clustering (Q87d ~ Q90b) 3](#_Toc67949806)

[1.4.2: K-means Clustering (Q87d ~ Q90a) 3](#_Toc67949807)

[1.4.3: Conclusion 3](#_Toc67949808)

[1.5: Hierarchical Clustering 3](#_Toc67949809)

[1.5.1: Hierarchical Clustering (Q87d ~ Q90b) 4](#_Toc67949810)

[1.5.2: Hierarchical Clustering (Q87d ~ Q90a) 4](#_Toc67949811)

[1.5.3: Goodness of fit test 4](#_Toc67949812)

[1.5.4: Conclusion 4](#_Toc67949813)

[Part 2 4](#_Toc67949814)

[2.1: Overview 4](#_Toc67949815)

[2.2: Models 5](#_Toc67949816)

[2.2.1: Full model 5](#_Toc67949817)

[2.2.2: Forward Selection model 5](#_Toc67949818)

[2.2.3: Backwards Elimination 6](#_Toc67949819)

[2.3: Results 6](#_Toc67949820)

[2.4: Analysis 7](#_Toc67949821)

[2.5: Conclusion 7](#_Toc67949822)

[Part 3 8](#_Toc67949823)

[3.1: Overview 8](#_Toc67949824)

[3.2: Methodology 8](#_Toc67949825)

[3.2.1: Logistic Regression 8](#_Toc67949826)

[3.2.2: LDA + QDA 8](#_Toc67949827)

[3.2.3: KNN 9](#_Toc67949828)

[3.3: Results 9](#_Toc67949829)

[3.4: Analysis 9](#_Toc67949830)

[3.5: Conclusion 10](#_Toc67949831)

# Part 1

## 1.1: Overview

In this project, we have a data set containing responses from the European Working Conditions Survey 2016 and we aim to summaries the information in the following report. Unsupervised learning techniques will be utilized. These techniques include:

1. Plot Matrix
2. Principal Component Analysis (PCA)
3. K-Means clustering
4. Hierarchical clustering

The responses listed in the data set can be categorized into 2 types:

1. Employees’ feelings and emotions over the last 2 weeks (Q87a – Q87e)
2. Employees’ attitude towards their job (Q90a – Q90f)

## 1.2: Plot Matrix

A plot matrix contains a scatterplot of all possible combinations of variables in the data frame such that each individual variable is plotted against each other to determine a trend. From the plot matrix generated in R, we have identified the following observations,

1. Gender (Q2a) and Age (Q2b) have no effect on both employees’ emotions over the last 2 weeks and their attitude towards their job.
2. Employees generally feel that they are good at their job (Q90f), regardless of how they feel over the last 2weeks as depicted by the relatively horizontal red line in each plot of the last row.
3. Employees’ feelings and emotions (Q87a – Q87e) seems to have an impact on their attitudes (Q90a – Q90c) towards their job. An upward trend is observed.

## 1.3: Principal Component Analysis (PCA)

Looking at the principal components (PC) derived in R, the first 4 PC capture about 70% of the variation. We have identified the significant variables in each of the PC through the coefficients of the respective variables. Significances of variable is determined by the value of its coefficient; a large coefficient implies a high level of significance. These variables are,

1. PC1: 87a, 87b, 87c, 87d, 87e
2. PC2: 90b, 90c, 90f
3. PC3: 2a
4. PC4: 2a, 2b

As Q2a and Q2b are significant only in the 3rd and 4th PC, we can conclude that they are not important differentiator with regards to employee’s feelings and emotions and their attitude towards their job.

## 1.4: K-means Clustering

We have decided to explore the effects Q87d has on Q90a and Q90b using K-means Clustering as we believe that being well rested and fresh has a huge impact on a person’s positivity and productivity. Likewise, we believe that a positive working environment entails a workforce filled with energy and enthusiasm towards their job.

We propose setting k = 2 to separate the data into natural clusters of 2 whereby employees are separated based on whether they are happy or unhappy with their working conditions. In the following learning technique, we will classify the responses in the following manner:

|  |  |  |
| --- | --- | --- |
| Questions | Response Values | Indications |
| Q87d | 1 - 3 | Woke up feeling fresh and rested majority of the time. |
| Q87d | 4 - 6 | Do not wake up feeling fresh and rested majority of the time. |
| Q90a | 1 - 2 | Feel full of energy at work majority of the time. |
| Q90a | 3 - 5 | Do not feel full of energy at work majority of the time. |
| Q90b | 1 - 2 | Enthusiastic about their job majority of the time. |
| Q90b | 3 - 5 | Not enthusiastic about their job majority of the time. |

Table 1: Response Classification

### 1.4.1: K-means Clustering (Q87d ~ Q90b)

From the results, we have derived that 95% of those in cluster 1 woke up feeling fresh and rested majority of time and hence 81% of those in cluster 1 are enthusiastic about their job majority of the time. Similarly, 53% of those in cluster 2 do not wake up feeling fresh and rested majority of time and hence 60% of those in cluster 2 are not enthusiastic about their job majority of the time.

### 1.4.2: K-means Clustering (Q87d ~ Q90a)

From the results, we have derived that 95% of those in cluster 1 woke up feeling fresh and rested majority of time and hence 90% of those in cluster 1 feel full of energy at work majority of the time. Similarly, 53% of those in cluster 2 do not wake up feeling fresh and rested majority of time and hence 61% of those in cluster 2 do not feel full of energy at work majority of the time.

### 1.4.3: Conclusion

Therefore, using K-means Clustering allows us to conclude that feeling fresh and rested is a significant differentiator with regards to employees’ attitude towards their job.

## 1.5: Hierarchical Clustering

We will still be exploring the effects Q87d has on Q90a and Q90b for the same reason as mentioned in 1.4, however using Hierarchical Clustering instead. Table 1 will used as well to classify the survey’s responses for the following clustering technique. We propose cutting the Dendrogram whereby only 2 clusters are remining for the same reason as mentioned in 1.4. Through various experimentation, we decided to use the Complete Linkage method instead of the Average Linkage one due to the better spread/classification of data.

### 1.5.1: Hierarchical Clustering (Q87d ~ Q90b)

From the results, we have derived that 76% of those in cluster 1 do not wake up feeling fresh and rested majority of the time and hence 68% of those in cluster are not enthusiastic about their job majority of the time. Similarly, 88% of those in cluster 2 wake up feeling fresh and well rested majority of the time and hence 74% of those in cluster 2 are enthusiastic about their job majority of the time.

### 1.5.2: Hierarchical Clustering (Q87d ~ Q90a)

From the results, we have derived that 76% of those in cluster 1 do not wake up feeling fresh and rested majority of the time and hence 70% of those in cluster do not feel full of energy at work majority of the time. Similarly, 88% of those in cluster 2 wake up feeling fresh and well rested majority of the time and hence 79% of those in cluster 2 feel full of energy at work majority of the time.

### 1.5.3: Goodness of fit test

Since P-value < 0.05, we are able to reject the null hypothesis and conclude that the proportion of those who do not feel well rested in cluster 1 is significantly different than that of cluster 2.

### 1.5.4: Conclusion

Using Hierarchical Clustering allows us to concludes that feeling fresh and well rested is a significant differentiator with regards to employees’ attitude towards their job.

# Part 2

## 2.1: Overview

In this project, we aim to compare the various linear regression models separately for each of the 2 datasets provided and hence, determine which of these models are more accurate in terms of its predictive accuracy. The models which contain different input variables are used to predict the outcome variable, G3 also known as the student’s final grade. As tasked, variables G1 and G2 were excluded in the following linear regression exploration. A train-test split procedure will be conducted for each model whereby,

1. Train set data (70% of the data): Used to fit the model
2. Test set data (30% of the data): Used to evaluate the fit of the model

Thereafter, the predictive accuracy will be evaluated by calculating the Root Mean Square Error (RMSE). RMSE is a measure of the differences between the predicted values and the observed values provided in the datasets. A lower RMSE value indicates a model with higher predictive accuracy, whereas a higher RMSE value indicates a model with lower predictive accuracy. Our aim here is to determine the model with the lowest RMSE value.

K-fold cross validation will be conducted as well to determine if there is any further improvements made to the predictive accuracy of the model. In this procedure, the dataset will be divided into K groups evely. Each group will rotate to be the test set data against the remaining groups which are acting as the train set data. The model’s RMSE will be calculated by averaging out all the RMSE derived from each train-test split procedures.

Therefore, we will then conclude the model with the highest predictive accuracy for each of the datasets provided.

## 2.2: Models

### 2.2.1: Full model

The full model applied to each of the dataset contain all of the input variables provided. These variables are included in the Linear Regression to predict the final grade, G3. As such, variables that are not significant in predicting G3 may affect the model’s predictive accuracy.

### 2.2.2: Forward Selection model

This method will begin with an empty model and input variables are added individually into the model. Input variables that provide the best improvement to the model is added during each of the forward step process. This process will conclude once the model no longer improves upon the addition of more input variables. The entire selection process will be automated in R. The variables included in each of the Forward Selection models are included in the table 2.

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset 1: student-mat.csv | | Dataset 2: student-por.csv | |
| Column no. | **Input Variables** | **Column no.** | **Input Variables** |
| 2 | sex | 1 | school |
| 5 | famsize | 2 | sex |
| 7 | Medu | 3 | age |
| 9 | Mjob | 4 | address |
| 14 | studytime | 8 | Fedu |
| 15 | failures | 14 | studytime |
| 17 | famsup | 15 | failures |
| 21 | higher | 16 | schoolsup |
| 23 | romantic | 21 | higher |
| 25 | freetime | 23 | romantic |
| 26 | goout | 24 | famrel |
| 30 | absences | 26 | goout |
|  |  | 27 | Dalc |
|  |  | 29 | health |

Table 2: Forward Selection input variables

### 2.2.3: Backwards Elimination

This method will begin with a model whereby all input variables are included. Variables that do not contribute to the Linear Regression model in predicting the final grade, G3, are removed from the model one at the time. This process will conclude once the model receive no further improvements upon the removal of any input variables. The entire selection process will be automated in R as well. The variables included in each of the Backwards Elimination models are included in the table 3.

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset 1: student-mat.csv | | Dataset 2: student-por.csv | |
| Column no. | **Input Variables** | **Column no.** | **Input Variables** |
| 2 | sex | 1 | school |
| 3 | age | 2 | sex |
| 5 | famsize | 3 | age |
| 7 | Medu | 4 | address |
| 9 | Mjob | 8 | Fedu |
| 14 | studytime | 14 | studytime |
| 15 | failures | 15 | failures |
| 16 | schoolsup | 16 | schoolsup |
| 17 | famsup | 21 | higher |
| 22 | romantic | 23 | romantic |
| 25 | freetime | 24 | famrel |
| 26 | goout | 26 | goout |
| 30 | absences | 27 | Dalc |
|  |  | 29 | health |

Table 3: Backwards Elimination input variables

## 2.3: Results

The results summarised in the following tables are rounded to 5 decimal places.

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset 1: | student-mat.csv | | |
| Model | **Model’s R-code** | **Train set RMSE** | **Test set RMSE** |
| Full model | m.full | 3.41654 | 5.10041 |
| Forward Selection model | m.for | 3.68985 | 4.54301 |
| 5 – fold cross validation on m.for | mfold1 | 3.68985 | 3.95774 |
| Backwards Elimination model | m.back | 3.68133 | 4.52708 |
| 5 – fold cross validation on m.back | mfold2 | 3.68133 | 3.94611 |

Table 4: Mathematics RMSE results

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset 2: | student-por.csv | | |
| Model | **Model’s R-code** | **Train set RMSE** | **Test set RMSE** |
| Full model | m.full | 2.29039 | 3.06114 |
| Forward Selection model | m.for | 2.42398 | 2.82974 |
| 11 – fold cross validation on m.for | mfold1 | 2.42398 | 2.53305 |
| Backwards Elimination model | m.back | 2.42398 | 2.82974 |
| 11 – fold cross validation on m.back | mfold2 | 2.42398 | 2.53305 |

Table 5: Portuguese Language RMSE results

## 2.4: Analysis

From table 4, we can see difference in the test set RMSE among the models. The full model (m.full) has then highest test set RMSE of 5.10041. Both the forward selection model (m.for) and backwards elimination model (m.back) have a lower RMSE than the m.full model. There is a slight difference in test set RMSE between the m.for and m.back model. Hence we decided to conduct a 5 – fold cross validation to determine which is a more accurate model to predict the final grade, G3. A decrease in test set RMSE is observed for both mfold1 and mfold2 models indicating a more accurate predictive performance. The decrease in RMSE for both model are approximately 0.58, with mfold2 having a lower RMSE.

Similarly from table 5, we can see a difference as well in the test set RMSE among the models. The m.full model has the highest testset RMSE of 3.06114. This implies that the m.full model has the least predictive accuracy. Both the m.for and m.back models have the same testset RMSE of 2.82974. Hence we decided to conduct a 11 – fold cross validation to determine which is a more accurate model to predict the final grade, G3. Likewise, a decrease in test set RMSE is observed for both mfold1 and mfold2 models indicating a more accurate predictive performance. However, both models concluded with the same RMSE error as well.

## 2.5: Conclusion

We can conclude that for both of the dataset, the m.full models have the lowest predictive performance in predicting the students’ final grade for Mathematics and Portuguese Language. In predicting the student’s final grade, G3 for Mathematics, we recommend using the m.back model due to its better predictive performance than the m.for model. As mentioned above, both the m.for and m.back models have the same RMSE. As such, either one of the model can be used when predicting the student’s final grade for Portuguese Language.

# Part 3

## 3.1: Overview

In this project, we aim to compare various Classification models if a client will subscribe a term deposit (y), based on the customer’s input variables provided in the dataset. The types of Classification techniques used includes, Logistic Regression, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA) and K Nearest Neighbour (KNN). As the outcome variable, y is a binary categorical variable (yes/no), the predictive performances of each model will be determined using a confusion matrix as shown in table 5.

|  |  |  |
| --- | --- | --- |
| Total observations: n | Predicted: Yes | Predicted: No |
| Actual: Yes | True Positive (TP) |  |
| Actual: No |  | True Negative (TN) |

Table 6: Confusion Matrix

The models’ predictive accuracy will be calculated using the following formula:

## 3.2: Methodology

### 3.2.1: Logistic Regression

Logistic Regression is applied to the full model whereby all the input variables present in the dataset are included to predict if a client will subscribe to a term deposit. Train-test split is then conducted using the full model to examine any improvements made to the model’s predictive accuracy. At the end of each regression, the results will be recorded in 2 separate confusion matrix to determine the models’ predictive accuracy.

### 3.2.2: LDA + QDA

From the Logistic Regression model, we have determined the following input variables to be significant using the model’s summary results.

1. Month
2. Duration
3. Poutcome

Hence LDA and QDA will be applied to various models containing different combination of the above input variables to predict if the client will subscribe a term deposit. The results will be recorded into individual confusion matrix to determine the models’ predictive accuracy.

### 3.2.3: KNN

KNN will also be applied to the various models containing different combination of the significant variables listed in 3.3.2 to predict if the client will subscribe to a term deposit. A value of K = 29 will be used for each of the model whereby the prediction is based on the majority of the 29 closest neighbour. The value 29 is derived by square rooting the total number of clients available in the dataset. In addition, different values of K will be tested on the model with highest accuracy to examine if there is any improvements made to the model’s predictive accuracy. The results will be recorded into individual confusion matrix as well.

## 3.3: Results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Technique used | Model’s R-code | Input Variables | Outcome Variables | Predictive accuracy |
| Logistic Regression | m.full | all | y | 0.84436 |
| Logistic Regression + train-test split | m1 | all | y | 0.79592 |
| LDA | m2 | month | y | 0.76327 |
| LDA | m3 | duration | y | 0.79184 |
| LDA | m4 | month + duration | y | 0.8 |
| QDA | m5 | month | y | 0.75510 |
| QDA | m6 | duration | y | 0.78776 |
| QDA | m7 | month + duration | y | 0.74286 |
| KNN (K = 29) | knn.1 | month | y | 0.75102 |
| KNN (K = 29) | knn.2 | duration | y | 0.76735 |
| KNN (K = 10) | knn.3 | duration | y | 0.75918 |
| KNN (K = 50) | knn.4 | duration | y | 0.78367 |
| KNN (K = 29) | knn.5 | poutcome | y | 0.75102 |

Table 7: Classification model results

## 3.4: Analysis

From table 6, we can see that the Logistic Regression model has a higher model accuracy than the one with train-test split conducted. The model, m.full has an accuracy of 0.84436 indicating that close to 84.4% of the prediction is true.

The model, m4 has the highest accuracy of 0.8, among the LDA models which indicates that the variable “month + duration” is more significant in predicting if a client will subscribe to a term deposit than individual variables alone. This means that 80% of the prediction from this model is true. However, the model with the highest predictive accuracy among the QDA models is m6, whereby only “duration” is used as the input variable. Nonetheless, its accuracy of 0.78776 is lower than that of m4.

Among the KNN models, knn.1 which only has “duration” as the input variable, has the highest accuracy of 0.76735. This implies that nearly 76.7% of the prediction is true. Different values of K are then applied to this model. We can conclude that setting a lower K value will decrease the model’s predictive accuracy and setting a higher K value will increase the model’s predictive accuracy. Setting K=50 will only improve the the accuracy by 2.4%.

## 3.5: Conclusion

Despite improvements made to the initial KNN model, the m.full model has the highest predictive accuracy among all the models explored. Therefore, we recommend using this model in order to better predict if a client will subscribe to a term deposit.