ST3189 Machine Learning

Coursework

Name: Victor Min Thura Shwe

Student Number: 190398965

Table of Contents

[1. Coursework Part 1 2](#_Toc67933922)

[1.1 Background Information 2](#_Toc67933923)

[1.2 Methodology 2](#_Toc67933924)

[1.3 Implementation 2](#_Toc67933925)

[1.4 Summary 4](#_Toc67933926)

[2. Coursework Part 2 5](#_Toc67933927)

[2.1 Background Information 5](#_Toc67933928)

[2.2 Methodology 5](#_Toc67933929)

[2.3 Implementation 5](#_Toc67933930)

[2.4 Summary 8](#_Toc67933931)

[3. Coursework Part 3 9](#_Toc67933932)

[3.1 Background Information 9](#_Toc67933933)

[3.2 Methodology 9](#_Toc67933934)

[3.3 Implementation 9](#_Toc67933935)

[3.4 Summary 11](#_Toc67933936)

[4. Pages used 12](#_Toc67933937)

[5. References 12](#_Toc67933938)

# **1. Coursework Part 1**

## **1.1** **Background Information**

For part 1, I will be summarising the data of EWCS 2016 by finding out which variables have the most significance and how the variables are correlated with each other.

## **1.2** **Methodology**

The various models and techniques that will be used to summarise the data is as follows,

1. Principal Component’s analysis
2. Hierarchical clustering
3. K - means clustering

## **1.3** **Implementation**

The explanation portion of the various models and techniques that will be used.

1. **Principal Component’s analysis**

By performing principal component’s analysis, it helps reduce the dimension of the data by showing us a general correlation of the variables as can be seen on figure 1.

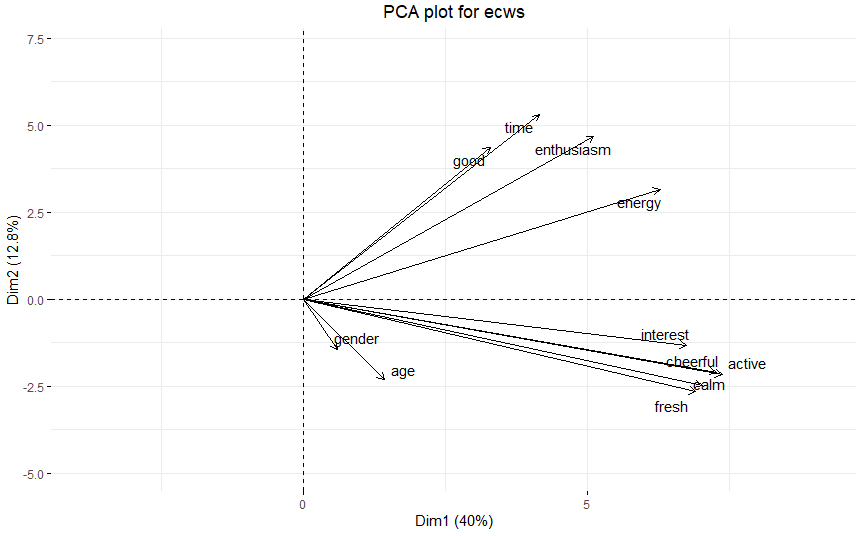
[[1]](#footnote-1)

Figure 1: PCA plot for ECWS

With the PCA plot, I can see judging by the angle in which each variable’s vector pinned at the origin shows how well they correlate with each other.

I can see that the questions from Q87 correlate very closely with each other as is also the case for the questions from Q90. Overall, considering the angle in which all the variables are separated I can see that all of the variables are positively correlated. However since the first two principal components only take up 52.8% of the variance proportion, this plot may not be the most accurate in getting the variable’s correlation with each other.

1. **Hierarchical clustering**

To perform hierarchical clustering, I will first scale the data in order to put equal weight on each of the variables.

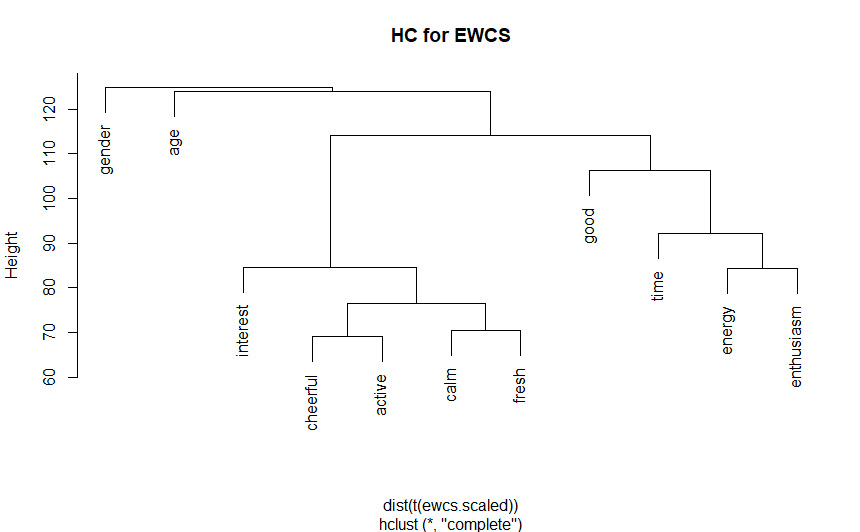


Figure 2: HC plot for EWCS

I can see from the plot that there are 3 main clusters created. The 3 main clusters being Gender, Age and the Questions. I can also see the cluster under questions later on divides into two more clusters, one being a cluster of the questions from Q87 and the other cluster being the questions from Q90. To further summarise the data, I will performing K-means clustering to find out how significant Gender and Age are as differentiators for the questions.

1. **K – Means Clustering**

With reference to the plot derived from hierarchical clustering, I will be using K – means clustering to further summarise the data by finding out if Gender and Age are significant differentiators in terms their ratings by using the most significant variable or question from each questions cluster.

In this case, Interest/Q87e and Good/Q90f as all the questions are positively correlated with each other so a high rating for either question would most likely correlate with a higher rating for the rest of the questions. Naturally, this is an assumption and may not be the best representation of each person in the data as it is a general view.

* **How gender affects their ratings**

I first create a data frame with all the variables excluding age and create two clusters, cluster1 and cluster2 that were derived from k-means clustering. We then factorise the variables Good and Interest as representatives of their respective question. I then find that in both cluster, the average rating of cluster1 is lower than that of cluster2. For Good, the average rating is **1.38** compared to **1.79**. For Interest, it’s **1.86** against **3.36**.

If I were to check the gender proportions in both clusters, it would be 53% are males in cluster1 and 48% in cluster2. By checking the p-value of the clusters, I can also see that since it’s less than 0.05 gender is a significant differentiator between cluster 1 and 2. This means that in general for the survey, males tend to have a more pessimistic or lower rating as compared to women.

* **How age affects their ratings**

For easier interpretability, I will be dividing age into two groups by using median[[2]](#footnote-2) as the cut-off point. The first group will be dubbed as “Younger” and the second group will be dubbed as “Older”.

The results obtained are the same as is for Gender. However the proportions for each cluster vary in terms of there being 45% are “Older” in cluster1 and 56% are “Older” in cluster2. This could mean that those who are “Younger” tend to rate much lower on the survey.

## **1.4** **Summary**

Males tend to rate lower on the survey as compared to Females. Age has an effect on the ratings in the way that those who are older tend to rate higher. I can also see that if a person were to rate highly on Q87e, their ratings for the other questions under Q87 would also be higher. Such is also the case for Q90f for the other questions under Q90. There is also a positive correlation between all the variables. As is for Q87 and Q90 further suggesting if a person were to rate highly on Q87, the person would likely rate highly on Q90 although the positive correlation is weak.

# **2. Coursework Part 2**

## **2.1** **Background Information**

For part 2, I will be comparing various ways in which I will be predicting the final year grade, G3 score of students for two different subjects. I will be titling both respectively as Math for Math courses and Port for Portuguese courses.

For both datasets of Math and Port, I will be taking out the 1st and 2nd period grades as it has a strong correlation with the final year grade which may affect results as I are more concerned with final year grade on its own.

## **2.2** **Methodology**

The various models and techniques that will be used to debate the predictive performance for G3 score will be as follows

1. Linear Regression Model followed up with Regularisation
2. CART model

## **2.3** **Implementation**

The explanation portion of the various models and techniques that will be used.

1. **Linear Regression Model**

In order to derive the linear regression model for predicting the variable G3, I first have to create a linear model for the variable G3 that excludes the G1 and G2 variables. Afterwards, in order to further improve the model I would have to find out which variables are useful in predicting G3 for both subjects. In order to do this I will be using Ridge and Lasso regression.

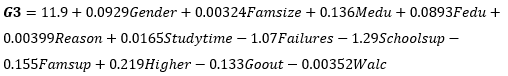
I will then decide whether I would use the regression model derived from ridge regression or lasso regression depending on the results obtained from the root mean square error[[3]](#footnote-3) of both models. The lower one being the better regression model.

* **Math Course**

Firstly, I created a train-test split in a 70-30 ratio where the training data is used to a fit a model and the testing data is used for testing. Then I create a linear regression model with all the variables intact with a RMSE result of **16.78541**. In order to further improve the model in terms of how well it fits the data. I then performed a ridge regression on the dataset of Math and getting a best lambda of **3.72951**. The lambda acts as the tuning parameter in which the higher it is the more the model variance reduces. I then derive the RMSE of the ridge regression which is **4.222624**. Secondly, I will be performing lasso regression on the dataset in order to act as a comparison of which regression model would be a better suit for predicting G3 score. Following the 70-30 train test split, I get a best lambda of **0.2693008**. I then derive the root mean square error of the lasso regression to get **4.287612**.

Once I have derived both RMSE, I can see which model serves as a better estimate for G3 score which in the case for school 1 is the ridge regression model. However, since the difference in root mean square error is not a very significant amount. I will be using the lasso regression model instead as having less variables to calculate will make it easier for prediction although less accurate.

The lasso regression equation is as follows, all variables coefficient were subject to 3 significant figures for easier interpretability.



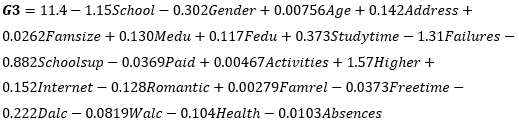
Judging by the first 6 predictions which are 9.27, 12.3, 11.9, 11.5, 5.52, and 12.5 respectively. It does not align very closely to the true values with the closest being the third value, 11.9 and 10. Hence using a linear regression model for predicting G3 scores for the math course may not be the most accurate model to use.

* **Portuguese Course**

For school 2, I will be performing the same regression methods as is for school 1. The initial linear regression model garnered a RMSE of **6.8644** which can be once again improved via ridge/lasso methods. For ridge regression, I obtained a best lambda of **1.011083**and a root mean square error of **2.809792**.

For the lasso regression, I obtained a best lambda of **0.1162499**and a root mean square error of **2.810955**. Once again just like for school 1, I can see that the root mean square error for ridge regression is lower than that of lasso regression however the difference is not that significant hence I will be using the lasso regression model for easier interpretability.

The lasso regression equation is as follows, all variables coefficient were subject to 3 significant figures for easier interpretability.



For school 2, the first 6 predicted values are 13.1, 13.1, 13.2, 12.8, 14.0 and 12.4. For school 2 though, using lasso regression might be a good model for predicting G3 where the first 6 predicted values are very close to the true value with the closest being the 6th variable, 12.4 as compared to 13.

1. **CART**

A tree based method that I can use to predict the G3 score is through the creation of a regression tree. Seeing the order in which variables play the most importance in terms of deciding whether the students achieve a high score or a low score.

* **Math Course**

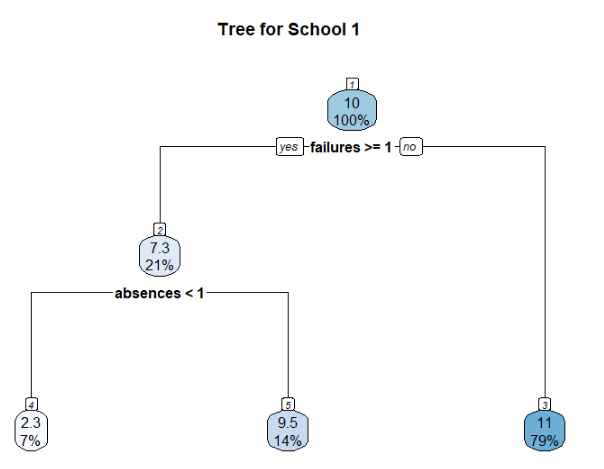
For the math course, I first link G3 as the Y variable in which I are trying to equate the rest of the variables as independent variables after which I would arrive a huge tree that was created which can be quite difficult to interpret. In that case, I would have to prune the tree to make it much more easily interpreted. After finding the optimal tree to use, I can prune it and achieve a much more easily interpreted diagram. 

Figure 3: CART Model for Math

The diagram shows that failures and absences have the most importance in

terms of affecting G3 score with number of failures being more important.

* **Portuguese Course**

I will be performing the same methods for the Portuguese Course. The diagram resulted from this is as follows.

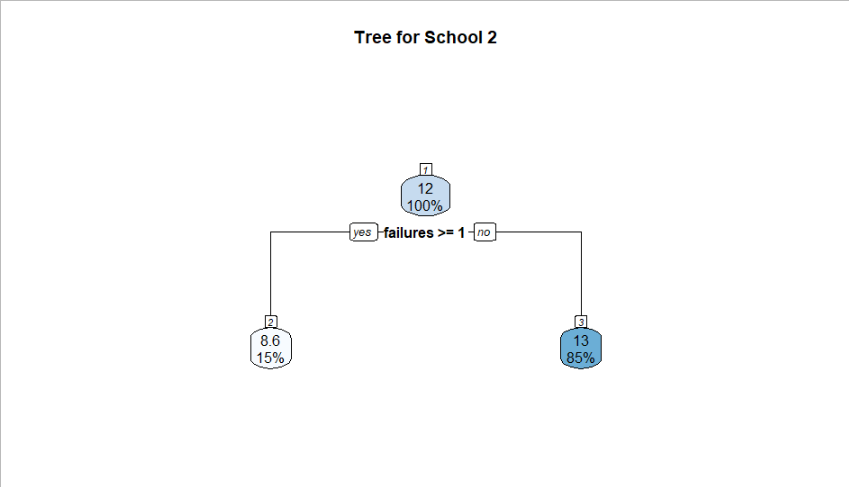


Figure 4: CART Model for Portuguese

From the diagram, once again I can see the number of failures they had play a big role in terms of importance in affecting G3 score.

## **2.4** **Summary**

In summary, with the idea of calculating G3 score based on whether a student was taking a Math course or the Portuguese course. I’ve come to the conclusion that for both courses, it would be easier to just use the CART method to find out whether a student passes or not on their final year grade. However for a result that shows how much they will predictably score which can be easily utilised, they can use the linear model regularised through lasso regression.

# **3. Coursework Part 3**

## **3.1** **Background Information**

For part 3, I will be will be comparing various ways of building a classification model in which is able to predict a binary outcome in this case whether a client will subscribe a term deposit or not.

## **3.2** **Methodology**

The various models and techniques that will be used to debate the predictive performance for subscribing a term deposit, variable y is as follows. The main way I will be debating predictive performance is to test the accuracy of the predictive model against the original results from the dataset.

1. Logistic Regression
2. Linear Discriminant Analysis and Quadratic Discriminant Analysis
3. K – Nearest Neighbour
4. Support Vector Machine

## **3.3** **Implementation**

The explanation portion of the various models and techniques that will be used.

1. **Logistic Regression**

To create a logistic regression model, I will use glmnet which creates a generalised linear model that will allow for the dependent variable to be binary. In this case whether a customer subscribes a term deposit or not, variable y.

After we create a generalised linear model with variable y as the dependent variable. We create a test set of generated predicted numbers derived from the model where we will be comparing the predicted numbers to the original test set.

The derived outcomes can be summarised and placed in a confusion matrix table which helps compare model predicted values against the actual values for the test set. Note that the test sets both have 1356 values.

|  |  |  |
| --- | --- | --- |
|  | Predicted Y: No | Predicted Y: Yes |
| Observed Y: No | **994** | 32 |
| Observed Y: Yes | 206 | **124** |

Figure 5: Confusion Matrix for GLM

In order to derive the accuracy of the model used, I will be adding the true positive and the true negative together divided by the total amount of values which is 1356. So the accuracy of this model is .

1. **Linear Discriminant Analysis and Quadratic Discriminant Analysis**

Both LDA and QDA work on the assumption that the observations from each variable is drawn from a normal distribution. The difference between them however is that LDA assumes a common covariance matrix between all the variables whereas QDA assumes that each variable has its own covariance matrix. Hence we will be testing which model will yield a better accuracy result.

For LDA and QDA, we will be performing very similar methods to what we did for GLM by placing y as the dependant variable. For LDA, the confusion matrix table derived is,

|  |  |  |
| --- | --- | --- |
|  | Predicted Y: No | Predicted Y: Yes |
| Observed Y: No | **1159** | 86 |
| Observed Y: Yes | 41 | **70** |

Figure 6: Confusion Matrix for LDA

The accuracy rate of the LDA model is .

For QDA, the confusion matrix table derived is,

|  |  |  |
| --- | --- | --- |
|  | Predicted Y: No | Predicted Y: Yes |
| Observed Y: No | **1095** | 87 |
| Observed Y: Yes | 105 | **69** |

Figure 7: Confusion Matrix for QDA

The accuracy rate of the QDA model is .

1. **K - Nearest Neighbour**

K – Nearest Neighbours is a model that will help classify data points based off similarity using Euclidean distance hence the scaling of several variables will be required. Those variables being ‘age, balance, day, duration, campaign, pdays and previous’. For K – Nearest Neighbours, the confusion matrix derived is,

|  |  |  |
| --- | --- | --- |
|  | Predicted Y: No | Predicted Y: Yes |
| Observed Y: No | **1201** | 144 |
| Observed Y: Yes | 3 | **8** |

Figure 8: Confusion Matrix for KNN

The accuracy rate of the KNN model is .

1. **Support Vector Machine**

The last model that will be used as a test for predictive performance is SVM. SVM is a model that creates a hyperplane that best divides the original dataset into two partitions. The two partitions in this case can be aptly nicknamed Yes or No under predicted Y. The confusion matrix derived is,

|  |  |  |
| --- | --- | --- |
|  | Predicted Y: No | Predicted Y: Yes |
| Observed Y: No | **1166** | 155 |
| Observed Y: Yes | 34 | **1** |

Figure 8: Confusion Matrix for KNN

The accuracy rate of the SVM model is .

## **3.4** **Summary**

To summarise part 3, the best classification model to use for the prediction of whether a client will subscribe a term deposit or not would be to use a linear discriminant analysis model. This is as the linear discriminant model scored the highest accuracy rate of 90.6% as compared to the other 4 models used meaning that the amount of false negatives and false positives it predicted is the lowest.

# **4. Pages used**

Excluding title page, table of contents and references. The amount of pages used is 10 pages.

# **5. References**

PCA analysis,

<https://www.datacamp.com/community/tutorials/pca-analysis-r>

<https://www.rdocumentation.org/packages/factoextra/versions/1.0.7/topics/fviz_pca>

Hierarchical Clustering

<https://www.datacamp.com/community/tutorials/hierarchical-clustering-R>

KNN,

<https://quantdev.ssri.psu.edu/sites/qdev/files/kNN_tutorial.html>

K-Means Clustering

<https://www.datanovia.com/en/lessons/k-means-clustering-in-r-algorith-and-practical-examples/>

LDA and QDA

<https://datascienceplus.com/how-to-perform-logistic-regression-lda-qda-in-r/>

1. Dots/Observations were removed from the diagram for better viewing and easier interpretability [↑](#footnote-ref-1)
2. Median age is 43. [↑](#footnote-ref-2)
3. Root mean square error will be listed as RMSE from this point on [↑](#footnote-ref-3)