ST 3189

Machine Learning Coursework Report 2021

Chen Xueting

S/N: 180404605

Table of Contents

[Part 1: Visualisation and unsupervised learning on EWCS 2016 2](#_Toc67965371)

[1. Data Cleaning 2](#_Toc67965372)

[2. Data Visualization 2](#_Toc67965373)

[3. Principal Components Analysis 3](#_Toc67965374)

[4. K-means Clustering 3](#_Toc67965375)

[5. Hierarchical Clustering 4](#_Toc67965376)

[Part 2: Regression Model for Student Performance 6](#_Toc67965377)

[1. Exploratory Data Analysis 6](#_Toc67965378)

[2. Linear Regression (Original Model) 7](#_Toc67965379)

[3. Backward Stepwise Regression 7](#_Toc67965380)

[4. Ridge Regression 8](#_Toc67965381)

[5. Random Forest Regression 8](#_Toc67965382)

[6. Conclusion 8](#_Toc67965383)

[Part 3: Classification Model for Bank Marketing 9](#_Toc67965384)

[1. Exploratory Data Analysis 9](#_Toc67965385)

[2. Classification and Regression Trees (CART) 9](#_Toc67965386)

[3. K-Nearest Neighbour (KNN) 11](#_Toc67965387)

[4. Random Forest Classification 11](#_Toc67965388)

[5. Conclusion 11](#_Toc67965389)

[Appendices 12](#_Toc67965390)

[Reference 12](#_Toc67965391)

# **Part 1: Visualisation and unsupervised learning on EWCS 2016**

The EWCS 2016 original dataset has 7813 observations and 11 variables. Q2a is a categorical variable, all of the remaining variables are numerical variable. The objective is to describe the dataset through visualization and unsupervised learning methods such as PCA, K-means clustering and hierarchical clustering.

## Data Cleaning

There are 329 observations recorded as -999 and will not be useful in our analysis hence removed. Now, the dataset is left with 7647 observations.

## Data Visualization

Chart, histogram

Description automatically generatedChart, bar chart, waterfall chart

Description automatically generatedAfter filtering the data, it is reported that 3899 males and 3748 females participated in the survey. We would like to find out whether there is correlation between workers’ mental wellness (Q87a to Q87e) and their working performance (Q90a to Q90c). The diagram below, *Figure 1* is the correlation plot and it shows some correlation between the mental wellness and working performance according to the intensity of colour. This could mean that worker who is feeling positive in terms of mental wellbeing will do better at work. However, the correlation seems to be decreasing in strength towards the last few survey questions answered.

Figure 1

Figure 2

A screenshot of a computer

Description automatically generated with low confidenceNext, we would like to understand the distribution of age among the workers*. Figure 2* shows the age of the workers ranged from early 20s to late 70s, with the median age at approximately 45 years old. Lastly, we want to find out what is the distribution of responses for each question. From *Figure 3,* the most frequent response from the respondents is “ 2: Most of the time ” except for Q90f whereby “ 1: Always” is chosen instead. This shows that respondents are generally very confident about their performance at work. The responses given for each question in the survey are majority also positive to neutral.

Figure 3

## Principal Components Analysis

*Chart

Description automatically generated*Chart, scatter chart

Description automatically generatedPCA is a dimensionality-reduction method which can transform a large set of data or correlated variables into a smaller number of representative variables that still preserve most of the information about the original dataset. *Figure 4* below shows a PCA biplot of the scaled dataset, variables Q2b and Q87 will strongly influence PC1 as they hold the heaviest loading weights. Vectors of Q87a to Q87e are relatively close to one another, hence the variables they represent are positively correlated. Similar observations are made for vectors of Q90a to Q90f, also age and gender. The strength of correlation between questions can be determined by the degree of angle, there will be correlation as long as the vectors are not orthogonal.

Figure 4

Figure 5

*Figure 5* is a scree plot depicting the proportion of variance explained (PVE) and cumulative PVE. According to the scree plots, the first principal component (PC) explains 40% of the variance in the data, the next PC explains 12.8% of the variance. Cumulatively, these two PCs explain only about 52% of the total data which is not accurate enough as a summary hence we might consider including more PCs to increase the accuracy.

## K-means Clustering

The main objective of the K-Means algorithm is to partition the dataset into a certain number of non-overlapping clusters and minimizes the distance between points and their respective cluster centroid. It is important to specify the desired number of clusters before clustering. Then, the K-means algorithm will assign each observation to the closest cluster centroid. The centre centroids will adjust accordingly with each assignment until values of centroid stabilises. Hence the number of clusters to begin is crucial, we will use Elbow method to determine the optimal number of clusters. This can be done by eyeballing *Figure 6* plot and look for the point at which the sum of squares distance drops off. Since the plot looks like an “arm”, the elbow on on the “arm” is the optimal number of clusters (James et. al.,2017). After running K-Means using the optimal K, the observations are divided into 2 clusters with 4869 in the first cluster and 2778 in the second cluster.

Text

Description automatically generated with medium confidenceChart, line chart

Description automatically generated*Figure 7* shows the mean response for each question and the two clusters are represented by “positive” and “neutral to negative” responses. All of the mean values in Cluster 1 are less than two which tells us that most responses in this cluster are positive. Whereas, mean values in Cluster 2 are generally more than two suggesting that there is mixture of neutral and negative responses. Furthermore, there are a few variables such as Q2a (Gender), Q2b (Age) and Q90f that are similar in values between the two clusters. Hence we will conduct Goodness of Fit Test (test) to determine any significant difference between these two clusters.

Figure 6

Figure 7

* Q2a – Gender

The mean values for Cluster 1 and 2 are 1.47 and 1.52 respectively. In proportions, there are 53% males in Cluster 1 and 48% males in Cluster 2. To find out whether the values for gender are significantly different between the two clusters, we will reject for being similar in values if p-value < 0.05. The p-value given by test is 0.0004998 < 0.05 hence conclude that gender is statistically different between the 2 clusters.

* Q2b – Age

The mean age for Cluster 1 and 2 are 41.5 and 45.4 respectively. Age ranges from 15 to 87 in Cluster 1 and 18 to 87 in Cluster 2, we will reject for being similar in values if p-value < 0.05. The p-value given by test is 0.0004998 < 0.05 hence conclude that age is statistically different between the 2 clusters.

* Q90f - In my opinion, I am good at my job [Please tell me how often you feel this way...]

The mean response for Cluster 1 and 2 are 1.38 and 1.78 respectively and both are positive response. This suggests most of the workers are confident in their abilities to perform well in their jobs. We will reject for being similar in values if p-value < 0.05. The p-value given by test is 0.0004998 < 0.05 hence conclude that Q90f responses are statistically different between the 2 clusters.

## Hierarchical Clustering

Hierarchical clustering has many advantages over K-means clustering such as not requiring to specify the number of clusters K beforehand and producing a tree-based representation of the observations knowns as dendrogram which is easy to interpret. The algorithm works in a manner that each data point is assigned to a cluster, the closest two clusters will be identified and combined into one cluster, the process will repeat until all the data points are in one cluster. This structure will be represented by a dendrogram and there are a few ways to determine the proximity of two clusters, namely; complete linkage clustering *(Fig. 7),* average linkage clustering (*Fig. 8*) and single linkage clustering (*Fig. 9*).

Diagram

Description automatically generatedA picture containing text, receipt

Description automatically generatedA picture containing text, receipt

Description automatically generatedDiagram

Description automatically generatedDiagram

Description automatically generated

Figure 10

Figure 9

Figure 8

Figure 7

1. Complete linkage clustering measures the maximum pairwise differences between observations in two clusters and record the largest difference. After cutting the tree, there are 1220 cases in cluster 1 and 6427 cases in cluster 2.
2. Single linkage clustering measures the minimum intercluster pairwise differences between observations in two clusters and record the lowest difference. After cutting the tree, there are 7647 cases in cluster 1 and only 1 case in cluster 2. Single linkage fails to provide sufficient sample size for cluster 2 hence it is not suitable to use.
3. Average linkage clustering measures the mean intercluster pairwise differences between observations in two clusters and record the average of these difference. After cutting the tree, there are 7618 cases in cluster 1 and only 29 cases in cluster 2. Average linkage fails to provide sufficient sample size for cluster 2 hence it is not suitable to use.

Hence, complete linkage clustering is preferred over single linkage and average linkage due to sufficient sample sizes for both clusters. Figure 10 shows the results of the mean values from both clusters under complete linkage clustering. Since Q2a and Q90f are similar in values, we will conduct Goodness of Fit test to find out whether they are statistically different.

* Q2a – Gender

The mean values for Cluster 1 and 2 are 1.527 and 1.483 respectively. In proportions, there are 47% males in Cluster 1 and 52% males in Cluster 2. To find out whether the values for gender are significantly different between the two clusters, we will reject for being similar in values if p-value < 0.05. The p-value given by test is 0.001696 < 0.05 thus conclude that gender is statistically different between the 2 clusters.

* Q90f - In my opinion, I am good at my job [Please tell me how often you feel this way...]

The mean responses for Cluster 1 and 2 are 1.854 and 1.469 respectively and both are positive response. This suggests most of the workers are confident in their abilities to perform well in their jobs. We will reject for being similar in values if p-value < 0.05. The p-value given by test is 0.0004998 < 0.05 thus conclude that Q90f responses are statistically different between the 2 clusters.

# **Part 2: Regression Model for Student Performance**

There are 395 and 649 observations for Mathematics and Portuguese course respectively, after removing G1 and G2 each dataset is left with 31 variables. All of the categorical input variables will be transformed into factors for to fit the models better. The target output variable is G3 (student’s final grade) and objective is to predict G3 using regression model of the lowest RMSE. No missing data was found hence it is a clean dataset.

## Timeline Description automatically generatedChart, histogram Description automatically generatedChart, histogram Description automatically generatedTimeline Description automatically generatedExploratory Data Analysis

Figure 12. Portuguese

Figure 11. Mathematics

Correlation plot is useful for data exploration and finding interaction terms. However, this is applicable only for numerical data, therefore categorical data is excluded from the plot. There are not much interaction terms for G3, the plots only show that “failure” is negatively correlated to G3. Other variables such as “age”, “Medu” and “goout” have weak correlation with G3 that are negligible. It is expected that Mother’s (Medu) and Father’s (Fedu) education levels are positively correlated, same for Workday (Dalc) and Weekend (Walc) alcohol consumption. In terms of score, students scored higher for Portuguese course than Mathematics. This can be seen from histogram in *Figure 12* whereby the mode score for Portuguese course is 11 and it is higher than the mode for Mathematics course.

## Linear Regression (Original Model)

Diagram, schematic

Description automatically generatedDiagram

Description automatically generatedTo build a linear regression model, dummies are created for all categorical variables then, split the data into 70% train set and 30% test set. To analyse the data, we can run diagnostic plots as follows:

Figure 13. Portuguese

Figure 12. Mathematics

1. In the Residuals vs Fitted plots, residuals from both courses do not have any distinct pattern hence it is a good indication that the data does not have non-linear relationships.
2. In both Normal Q-Q plots, the residuals are lined well along the dashed line indicating residuals are normally distributed.
3. In both Scale-Location plots, we can see residuals spread equally and randomly along horizontal line which is a good indication of equal variance (homoscedasticity).
4. Lastly in Residuals vs Leverage plots, all of the cases are within Cook’s distance suggesting that there is no influential case or outliers identified. (Kim, 2015)

We can proceed to run the linear regression model and train set. Setting G3 as the Y variable, some of the significant variables found in the train set model include: “reason\_reputation”, “failures”, “romantic\_yes”, “free\_time” and “go\_out”. Next, run predictions on the test set and the predicted score for Mathematics course ranges from 0.8929 to 16.7312, and for Portuguese course prediction score ranges from 6.769 to 15.665. The Root Mean Squared Error (RMSE) of Linear Regression model for Mathematics course is 4.22 and for Portuguese course is 2.82.

## Backward Stepwise Regression

As seen from the linear regression models, out of the 40 variables (including dummies) there are only some variables which are statistically significant. If all of the X variables are fit into the model, it might cause overfitting hence Backward Stepwise Selection is used in attempt to eliminate those insignificant variables. First, 10-fold cross validation is conducted then, model will be trained to return the maximum number of 30 variables. For Mathematics G3 model, the best model obtained consists of 4 predictors and RMSE of 4.26, model is given as: .

For Portuguese G3 model, the best model consists of 6 predictors and RMSE of 2.74, model is given as: Notice that RMSE for Mathematics model did not reduce, thus we also need to take into consideration the issues with fitting only statistically significant X variables. Some of these predictors may be redundant in predicting G3 while in the presence of other statistically significant variables.

## Ridge Regression

Ridge regression can regularize the regression model and thus balance model complexity with predictive errors. It adds penalty term (L2-norm) equivalent to the sum of the squared coefficient for variables with minor contribution to the outcome by shrinking the coefficient close to zero. It is a good solution for keeping all the predictors that contribute to predicting the dependent variable. It minimizes the sums of square of coefficients to reduce the impact of correlated predictors. The amount of the penalty can be fine-tuned using lambda (λ). Choosing a good value for λ is crucial and for our model, λ will range from to which can cover a full range of scenarios from null model with intercept only to the least squares fit.

1. First, indicate the predictor variables and dependent variable (G3) from the train set and fit into a penalized linear regression function, specify “alpha = 0” for Ridge regression.
2. Next, conduct cross-validation to find the best λ (given λ = 4.35 for Mathematics, λ = 1.17 for Portuguese) and the optimal value for λ can minimise the cross-validation prediction error rate.
3. Then, fit the final model into the training set and make predictions on the test set.

The model performance metrics, RMSE, is reported to be 4.17 for Mathematics and 2.81 for Portuguese model.

## Random Forest Regression

Random Forest is supervised learning algorithm, it is an ensemble of decision trees by using bootstrap samples on the training data set. Instead of using all the variables in a single tree, random forest uses a subset of the predictor variables for each tree. Prediction of the model will be produced by aggregating averaged predictions from all trees. Each tree is trained on two-third of the total training data and cases are chosen randomly with replacement from the original dataset. This process known as random record selection forms the train set for growing the tree. Random forest only uses a subset of the predictors chosen randomly, and best split is selected from all predictors. The optimum number of predictors for regression model is the total number of variables divided by 3 (Bhalla, 2014). For our random forest model, we employed ntrees = 500 and mtry = 30/3 which means growing up to 500 trees and 10 predictors will be randomly sampled at each split. Predictions made on each tree are averaged to produce the final prediction for the model. It is reported the model performance metrics, RMSE for Mathematics and Portuguese model are 3.68 and 2.76 respectively. We can see that RMSE for Mathematics model dropped the most prominent as compared to other techniques.

## Conclusion

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Subjects  (RMSE) | Linear Regression (Original) | Backward Stepwise Regression | Ridge Regression | Random Forest Regression |
| Mathematics | 4.22 | 4.26 | 4.17 | **3.68** |
| Portuguese | 2.82 | **2.74** | 2.81 | 2.76 |

From the table above, the best technique to predict student performance for Mathematics is Random Forest and for Portuguese is Backward Stepwise Regression. However, since the RMSE value between Backward Stepwise Regression and Random Forest are relatively close for the Portuguese model, it is still recommended to use Random Forest to predict student performance instead of Backward Stepwise. There are some limitations to use Backward Stepwise as it is subjected to have more predictor variables than the number of observations. It will also be challenging if the predictor variables are too many. Whereas, Random Forest will not have these restrictions and it also reduces overfitting. The merging of averaged predictions makes Random Forest better than other techniques in terms of accuracy and predictive performance. Hence, Random Forest regression model is the most suitable to predict student performance for both Mathematics and Portuguese.

# **Part 3: Classification Model for Bank Marketing**

The Banking Marketing dataset has 4521 observations with 16 input variables and 1 output variable. The dataset is from Portuguese banking institution and data was collected through telemarketing campaigns. The input variables have mixture of categorical and numerical attributes. The output variable, Y, is a binary variable which denotes whether the client has subscribed to a term deposit. Therefore, the objective is to build a classification model with the highest accuracy in predicting the dependent variable Y. Some variables are being converted into numeric class to fit the models better. No missing data was found; this is a clean dataset.

## Chart, bar chart Description automatically generatedChart, bar chart Description automatically generatedExploratory Data Analysis

Figure 14

Figure 15

Bar chart and histogram are used for analysing the dataset to understand the frequency distribution of variables.

## Classification and Regression Trees (CART)

CART is known for explaining decision tree algorithms with classification and regression learning tasks. For Bank Marketing dataset, we will focus on using Classification trees because the target variable Y is categorical. The dataset will be spilt into classes and the CART algorithm will identify which “class” target variable Y would most likely be classified under. (Mehta, 2020)

1. Dataset is partitioned into 80% train set and 20% test set. Grow the tree to the maximum using trainset.
2. Prune the Tree to the minimum with optimal Complexity Parameter (CP) value to reduce overfitting. To calculate the optimal CP value, we need to choose the lowest cross-validation error (xerror = 0.88599) and following the 1 Standard Error Rule, add this value to its standard error: (0.88599 + 0.043444 = 0.929434). This forms the Cross Validation Error Cap where the optimal CP region will be located between the Error Cap as shown in *Figure 17*.
3. Calculate the geometric mean of the two identified CP values in the optimum region if the optimal tree has at least one split: and this represents the optimal CP value.
4. Chart, histogram

   Description automatically generatedAfter pruning the tree with the optimal CP value, we can identify the optimal tree by selecting the lowest cross validation error. According to the pruned tree, optimal tree *(Figure 16)* is the 3rd tree and there are 8 splits.

Figure 17

Figure 16

Table

Description automatically generatedA picture containing table

Description automatically generatedUsing CART, we are also able to identify variable importance in optimal tree for Bank Marketing. “Duration” is ranked the most important variable followed by “Previous campaign outcome”, hence the bank should pay more attention to these two variables during marketing. Next, it is also important to find out what is the misclassification rate for the train set and cross validation. The root node error is reported to be: 421/3616 = 0.116, using this value the trainset misclassification rate is calculated as follows: 0.810 x 0.116 x 100% = 9.40% and CV misclassification rate is: 0.910 x 0.116 x 100% = 10.56%. Lastly, CART will predict on test set and Confusion Matrix is used to assess the model performance. CART has an accuracy rate of 89.28% and it is predicted that 5.4% of the clients will subscribe to a term deposit as shown in Figure 18 and 19.

Figure 19

Figure 18

## K-Nearest Neighbour (KNN)

A picture containing text

Description automatically generatedTable

Description automatically generatedThe principle behind KNN is to predict the target output Y by finding the nearest neighbour class. The closest class is often identified by using measurements like Euclidean distance. KNN is able to use clients’ available data to predict future client’s decision to subscribe a term deposit based on the comparison of similar data. However, choosing the right value for parameter K (number of nearest neighbour) is the most crucial part in KNN as it determines the efficacy of the model. A small value of K might lead to overfitting, a large value K might develop biasness and ignore the useful patterns (Saxena, 2016). We will conduct KNN as follows: Train the data using KNN function and to obtain the best K value, use 10-folds Cross Validation. This technique allows the algorithm to test different values of K and produces the optimal value with the highest accuracy. Among ten K values generated, the best K value is 17 which represents the count of the nearest neighbours and it has accuracy up to 88.7%. The training set outcome is 3195 “no” and 421 “yes”. Then, we can predict data on the test set and assess the model performance by Confusion Matrix. Based on the table results below, KNN has achieved an accuracy of 90.06% and predicted 1.7% of the clients will subscribe to the term deposit.

Figure 21

Figure 20

## Random Forest Classification

Table

Description automatically generatedTable

Description automatically generatedRandom forest classification tree is able to use here because Y is a binary/categorical variable output. Trees are grown similarly to random forest regression as mentioned above. The only difference will be mtry which is the square root of the number of input variables for classification. After training 80% of dataset, a measure of variable importance revealed that “duration” has the highest importance followed by “age” and “balance”. Finally, predictions were made on the test set and the model performance is assessed by Confusion Matrix. Results shown in Figure 22 and 23 stated that random forest classification has an accuracy of 89.94% and it is predicted that 5.8% of the clients will subscribe to the term deposit.

Figure 23

Figure 22

## Conclusion

In terms of accuracy, KNN technique has the lowest misclassification error rate (9.94%) among the three classification techniques. KNN also has quick calculation time and high interpretability in comparison with other techniques, it is recommended to adopt this technique to predict future client’s subscription on term deposit.

# **Appendices**

**Page count: 10**

**excluding** cover page, table of contents and appendix.

## Reference

Bhalla, D., 2014. A complete guide to Random Forest in R. *ListenData*. Available at: https://www.listendata.com/2014/11/random-forest-with-r.html [Accessed March 29, 2021].

Mehta, A., 2020. A Beginner's Guide To Classification And Regression Trees. *Digital Vidya*. Available at: https://www.digitalvidya.com/blog/classification-and-regression-trees/ [Accessed March 29, 2021].

Kim, B., 2015. University of Virginia Library Research Data Services + Sciences. *Research Data Services + Sciences*. Available at: https://data.library.virginia.edu/diagnostic-plots/ [Accessed March 29, 2021].

Saxena , R., 2016. Knn Classifier, Introduction to K-Nearest Neighbor Algorithm. *Dataaspirant*. Available at: https://dataaspirant.com/k-nearest-neighbor-classifier-intro/ [Accessed March 29, 2021].

James, G. et al., 2017. In *An introduction to statistical learning: with applications in R*. New York, New York: Springer, pp. 384.