Coursework - EMFSS ST3189 – Machine Learning

Student Number - **200692630**

Part 1

In this section, the given EWCS data set is explored and visualized via unsupervised learning methods of Principal Components Analysis (PCA) and Cluster Analysis.

The data set contains, 7647 useable observations after cleaning up the out-of-scope values. The distribution of Gender, Age as well as responses for the questionnaire is shown in *Figure 1*. Also *Figure 2* highlights the distribution of age. Referring to them following conclusions can be made briefly.

• There were slightly more males who answered the questionnaire than females.
• Looking at the distribution of answers for the questionnaire, it's clear that it is skewed to left (Positively skewed). From a real-world interpretation this means, majority of people have chosen options all of the time, most of the time or half of the time as their option, hinting a positive output on his/her job satisfaction.
• The age is almost normally distributed with min age of 15 and max age of 87 with a mean age of 43.16.

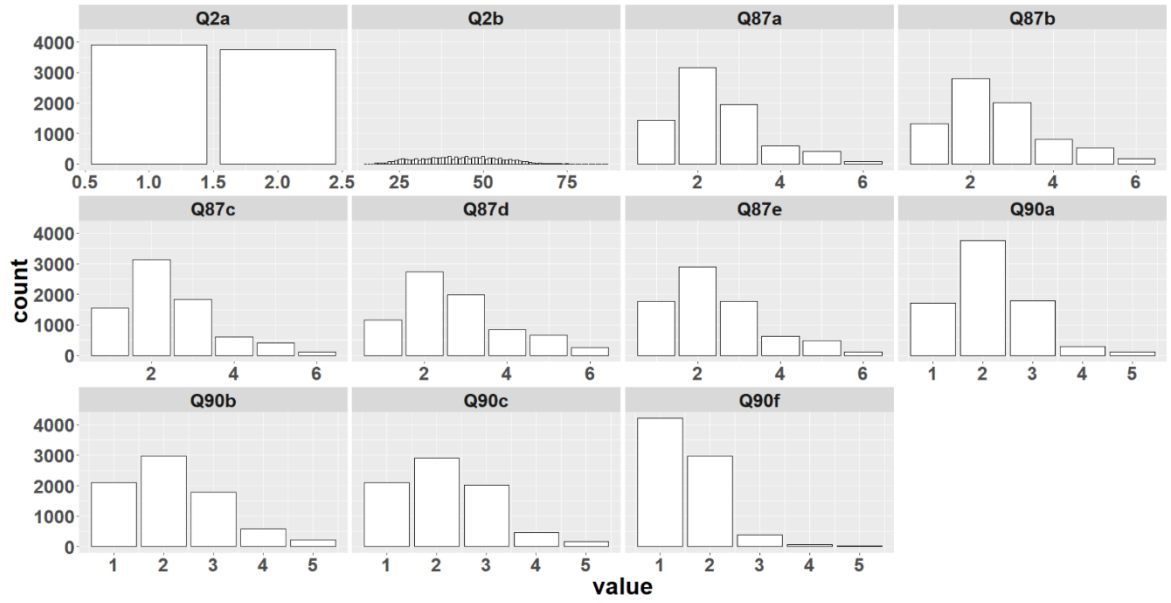


Figure 1. EWCS – distribution of data

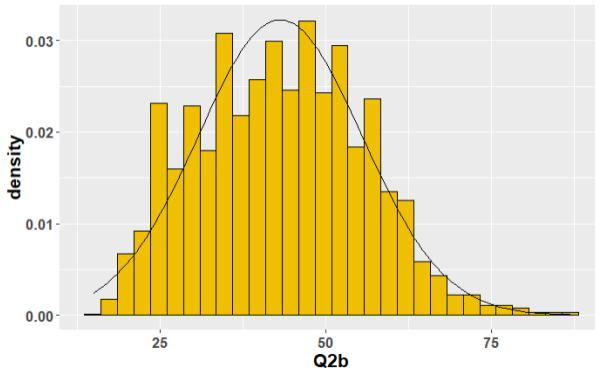


Figure 2. Distribution of age(Q2b)

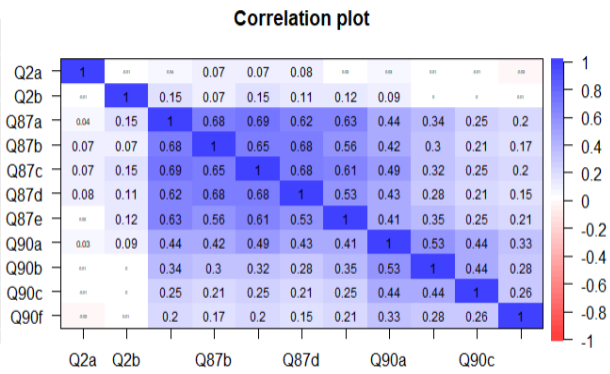


Figure 3. Correlation between variables

In order to see whether there are any inter or outer dependencies of personal wellbeing and job performance of individuals, we plot the correlation matrix (*figure 3*). From the correlation plot we can see that there's a significant inter dependency in Q87 questions category. The correlation between gender and personal wellbeing/ job performance is quite low.

Principal Component Analysis –

We use PCA method to reduce dimensionality and analyze the data by dividing them into components. First, data should be scaled in order standardize the range of variables. After conducting PCA on the scaled data, we get the following Scree plot (*Figure 4*) which would explain the proportion of variance explained by each component. More than 50% percent of variance is explained by the first 2 principal components. Hence first 2 PCs are further analyzed via a PCA Biplot (*Figure 5*).

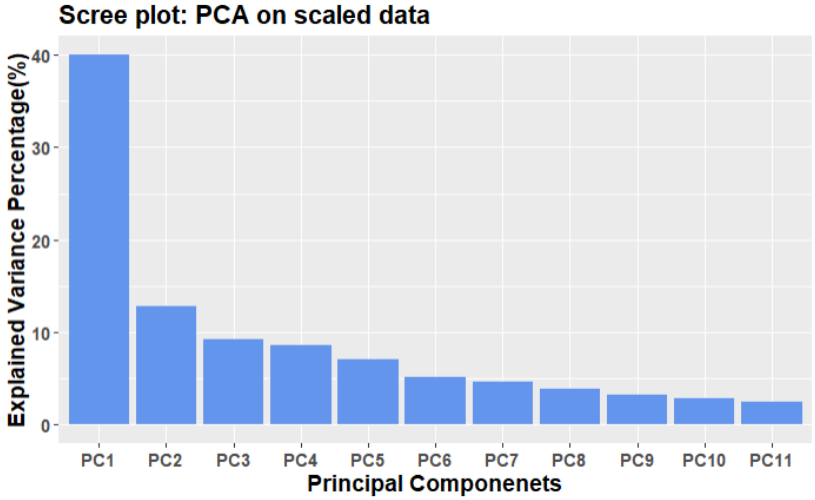


Figure 4. PCA Scree plot

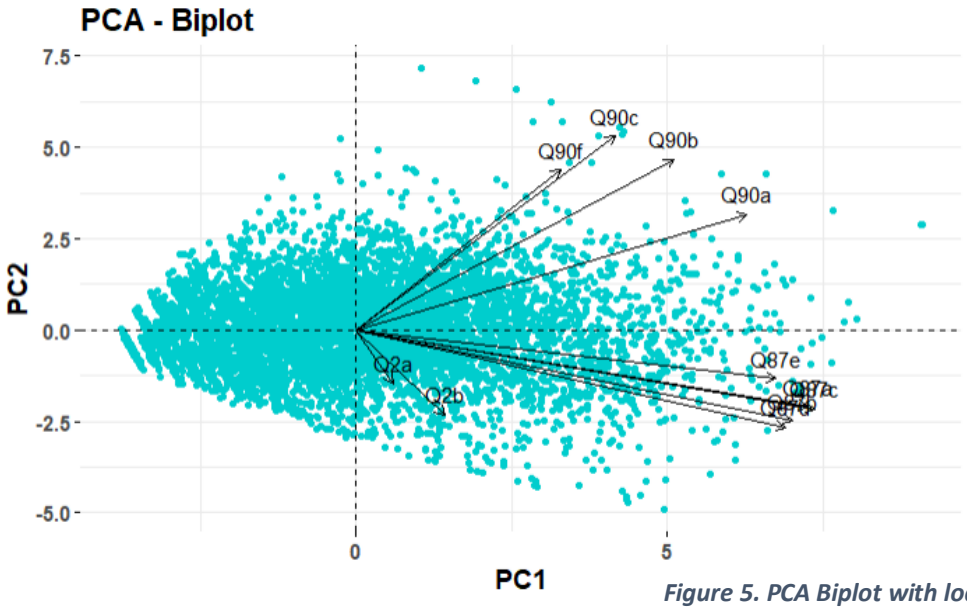


Figure 5. PCA Biplot with loading vectors

In the Biplot we can see that the angles between vectors corresponding to same category questions are lower compared to angles between vectors of different question categories. This means there's a high positive correlation between the responses to same category question. (E.g., A person who has responded negatively to Q87a is highly likely to respond negatively to any other question in Q87

category). Also, all the vectors are in the same direction hinting a positive correlation between all responses to the questions.

K-means Clustering –

Another possible way to summarize the data is by K -means clustering method. First the optimal number of clusters is found using elbow method (*Figure 6* **-** since total within sum of squares starts decreasing slowly after k=3, we pick 3 clusters). The resulting cluster plot is shown in *Figure 7.*

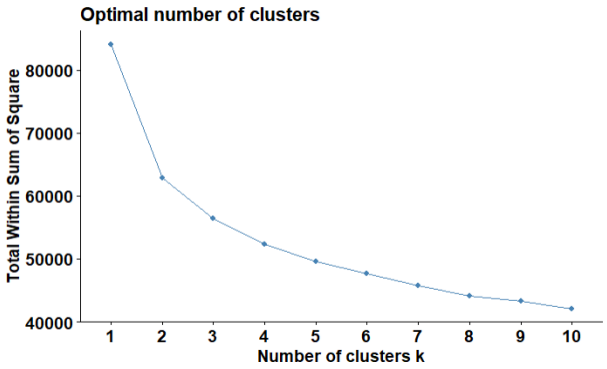


Figure 6. Elbow plot

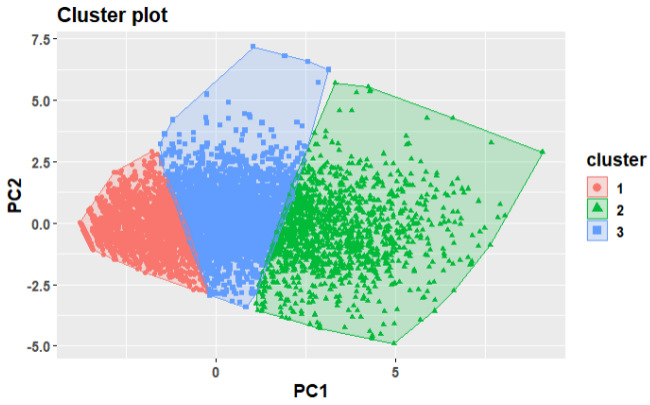


Figure 7. Cluster plot k=3

Further inferences can now be made by segmenting data with clusters. Referring to the distribution of Gender (*Figure 9)* It is evident that, 2nd cluster is dominated by females whereas other two clusters have more males. By looking at the mean values of answers to two question categories (Q87 and Q90) according to cluster (*Figure 8*) we can define the clusters as follows,

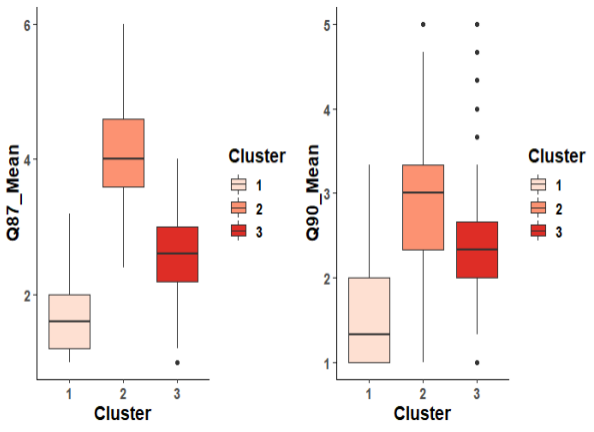


Figure 8. Q87 Mean Boxplot according to cluster

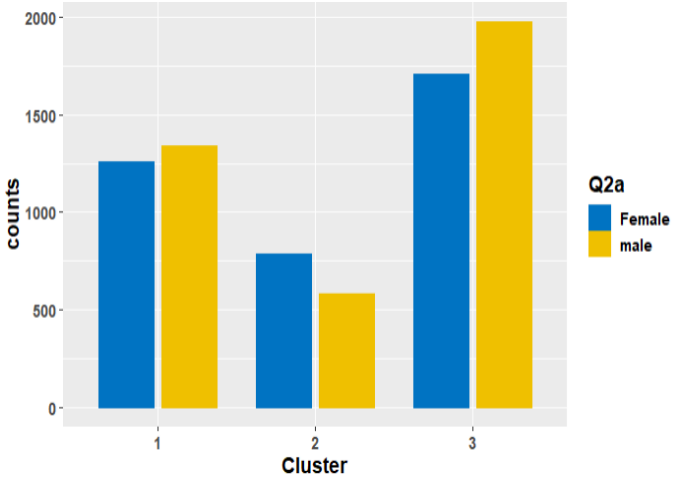


Figure 9. Distribution of Gender according to cluster

Cluster 1 – People who generally feel positive at work and home all the time or most of the time
Cluster 2- People who generally feel positive at work and home Rarely or Less than half of the time
Cluster 3- People who generally feel positive at work and home sometimes or more than half of the time.

Part 2

The objective of the second part is to use the given data of two Portuguese schools to build a model predicting the final grades that a student would get for subjects Math and Portuguese. The initial dataset contains two interim grades obtained by students for those subjects, however, those grades will not be used when predicting the final grade.

Looking at the descriptive statistics of the dataset, there were no visible missing values after data from two schools were merged. Overall, 382 observations were used in training the models.

Exploratory Data Analysis (EDA) –

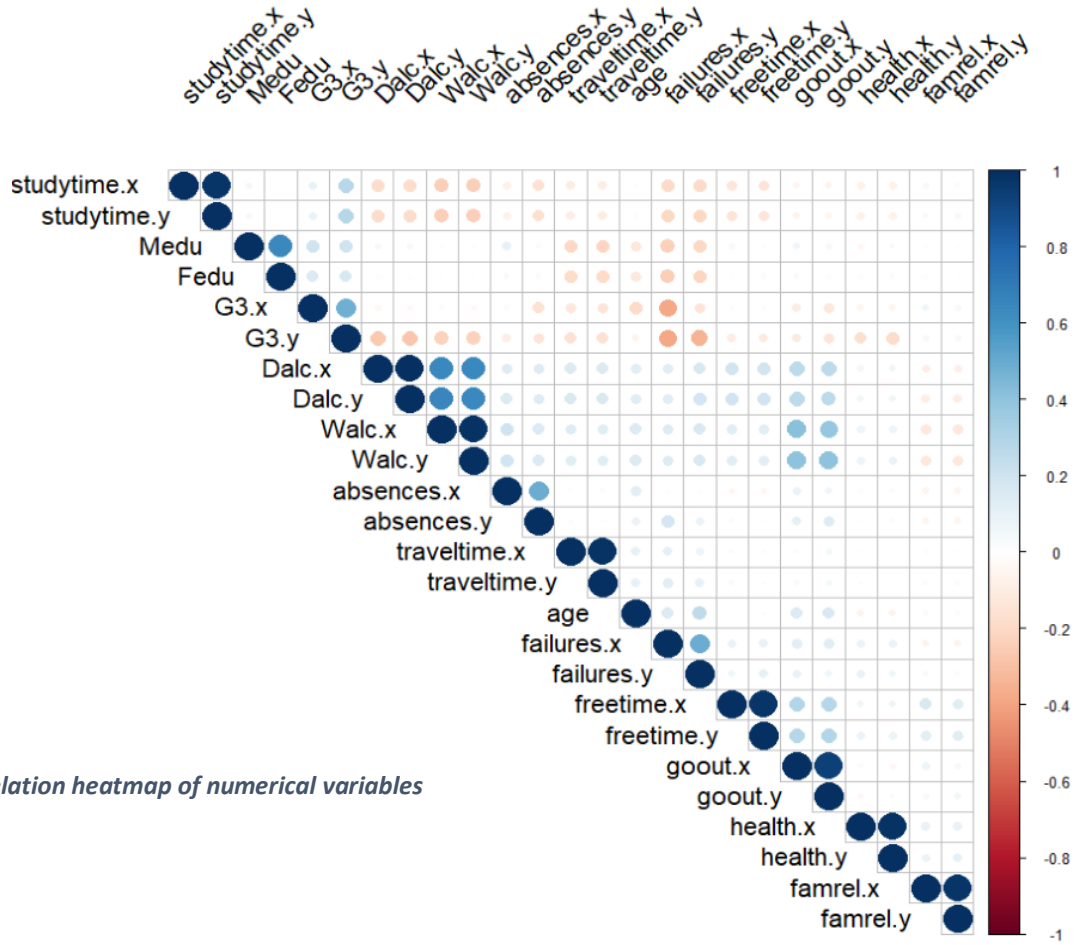


Figure 10. Correlation heatmap of numerical variables

First, in ***Figure 10*** we look at the correlation between numerical variables and final Math and Portuguese grades, it is evident that Father education, Mother education, study time are positively correlated with final grades. Also, there is a significant positive intercorrelation between Grade of Portuguese and Math.

Failures, travel time, workday alcohol consumption and weekend alcohol consumption are negatively correlated with final grades.

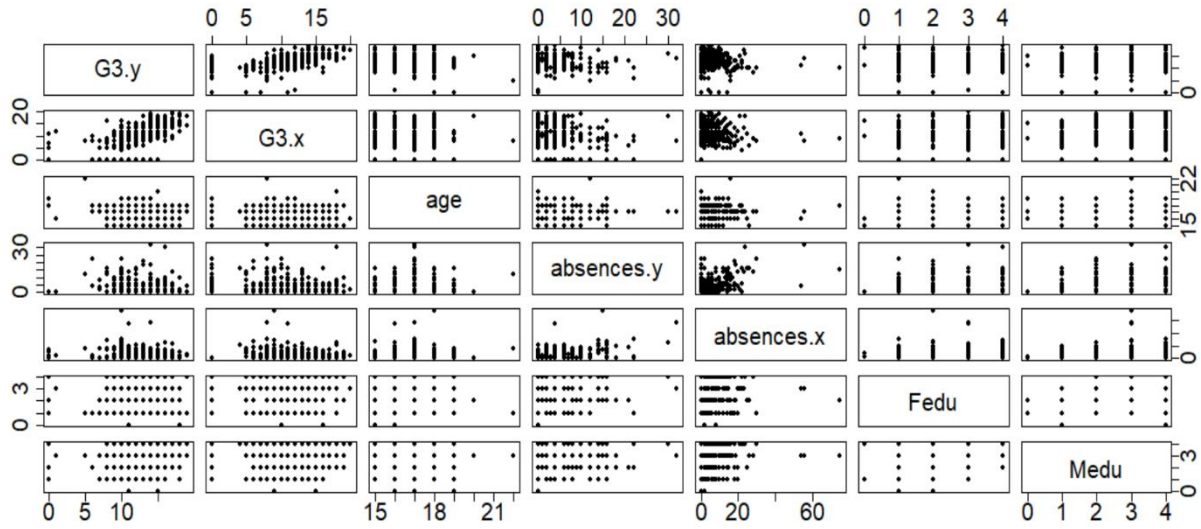


Figure 11. Pairwise scatter plots

Figure 11 shows pairwise scatter plots of the distribution of some significant features according to the final Math and Portuguese grades. Looking at the dataset as whole, we can see that almost half the variables are categorical. Therefore, in order to incorporate them into the model, those variables are given numerical values using one hot encoding method.

Models and Performance analysis –

The five techniques that were used to create predictive models were Linear Regression, Lasso Regression, Support vector machines(svm), Random Forest and Gradient Boosting regression. Here Linear Regression was taken as the baseline model. Since there a significant number of features to be considered, a regularization method like lasso regression was also tested as it will shrink the redundant coefficients in the model towards zero. Since there is a considerable number of observations than the features in the model support vector machines method was tested. Random Forest and Gradient boosting techniques were also used as they are more advanced and modern machine learning techniques presuming these models may give better results.

The predictive performances of the models are evaluated via two methods. First by K – fold cross validation method which uses 5 folds and metric R-squared. Secondly by hold-out train – test method with 80-20 split (train – 305 observations, test-77 observations).

Predict MATH GRADE K -FOLDS

Model	Fold1	Fold2	Fold3	Fold4	Fold5
1.Linear Regression	0.3572	0.1164	0.4140	0.3936	0.5017
2.Lasso Regression	0.0329	0.4965	0.2791	0.0001	0.3347
3.Support Vector Machines	0.3178	0.3384	0.1989	0.4802	0.4509
4.Random Forest	0.3876	0.4358	0.3941	0.5071	0.3346
5.GradientBoosting Regression	0.4517	0.3966	0.4055	0.5304	0.5210

Predict PORTUGUESE GRADE					K -FOLDS
Model	**Fold1**	**Fold2**	**Fold3**	**Fold4**	**Fold5**
1.Linear Regression	0.4892	0.2634	0.4253	0.4904	0.5152
2.Lasso Regression	0.4638	0.4742	0.0208	0.3916	0.3320
3.Support Vector Machines	0.4461	0.3153	0.3092	0.5675	0.5541
4.Random Forest	0.4851	0.3262	0.3701	0.5341	0.5047
5.GradientBoosting Regression	0.3893	0.3985	0.4547	0.5387	0.4774

Figure 12. K-fold cross validation results

Predict	Final Math Grade(G3.x)			Portuguese Grade(G3.y)		
Model	**RMSE**	**R^2**	**MAE**	**RMSE**	**R^2**	**MAE**
1.Linear Regression	4.1453	0.3489	3.1477	2.2369	0.4528	1.7736
2.Lasso Regression	4.3212	0.1265	3.4734	3.3212	0.0861	2.3433
3.Support Vector	4.2103	0.3151	2.9650	2.3116	0.4184	1.6261
4.Random Forest	3.7365	0.4747	2.9650	2.0447	0.5274	1.6261
5.GradientBoosting Regression	3.8294	0.4372	3.011	2.1254	0.5272	1.6805

Figure 13. Hold out method performance on test set (train – 305 observations, test- 77 observations)

In the first evaluation method, data is splatted into 5 folds and iteratively accuracy is calculated by taking one-fold to test the model and remaining folds to train the model. Looking at the R – Squared values of the results (Higher the better), it could be concluded that Gradient Boosting Regression outperforms other models when predicting both Math grade and Portuguese grades.

In the second type of evaluation, the models are trained on 80% of the data and tested on the remaining 20% of data. Looking at the results obtained by the accuracy tests, we can see both Random Forest model and Gradient Boosting regression have a lower RMSE and better R-squared values implying that those 2 models fit relatively well compared to Linear and Lasso regression verifying the conclusions made in the earlier evaluation method. However, there's a possibility of model over-fitting when this evaluation method is used.

Additionally further improvements to the models could be made via feature engineering such as removal of some highly correlated features with each other and taking linear combinations of independent variables when training the models.

Part 3

In the 3rd part, a classification models are built to predict whether clients of a Portuguese bank will subscribe to a term deposit or not.

Exploratory Data Analysis (EDA) –

Overall, there are 4521 observations with 17 features available in the dataset with no missing values. Looking at the correlation between numerical features of the dataset (***Figure 13***), a significant correlation is present between variables pdays and previous. Boxplot of numeric variables (***Figure 14***) shows the existence of possible outliers in the balance feature. Looking at the histogram of responses for a subscription according to marital status, there is no significant comparison as majority of people are declining to subscribe to a deposit despite marital status. However, married people had relatively subscribed more to term deposits compared to divorced and single people. Overall, only 521 people out 4521 have subscribed to a term deposit.

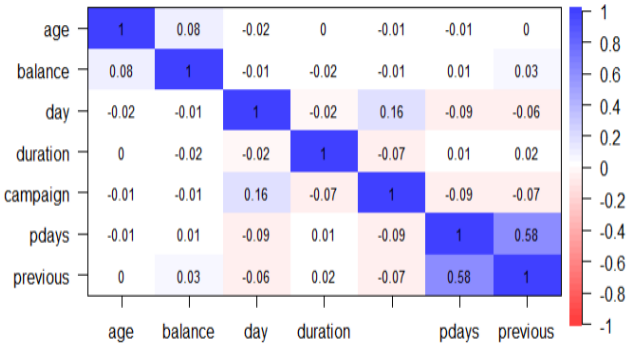


Figure 13. Correlation between Numerical variables

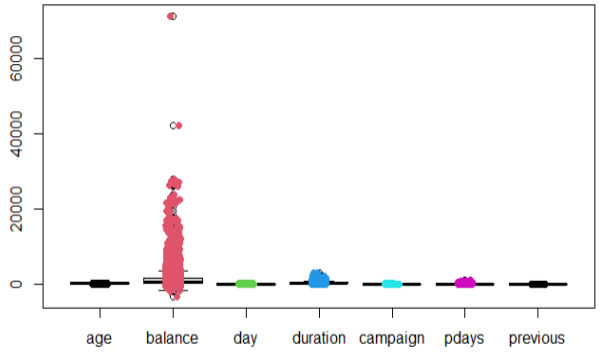


Figure 14. boxplots of numeric variables

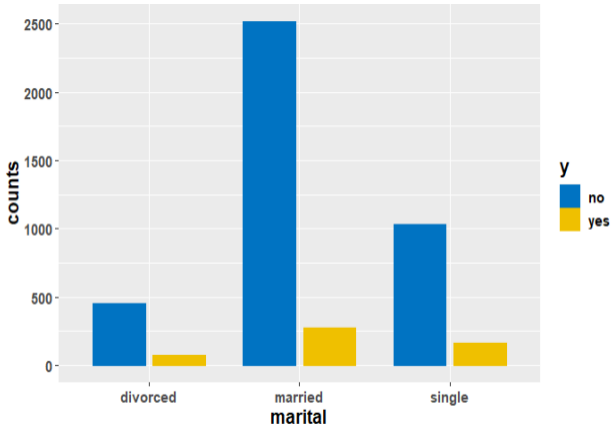


Figure 15. Subscription to a deposit according to marital status.

Models and Performance analysis –

The classification models that were used are Logistic Regression, K- Nearest Neighbors and Random Forests Classifier.

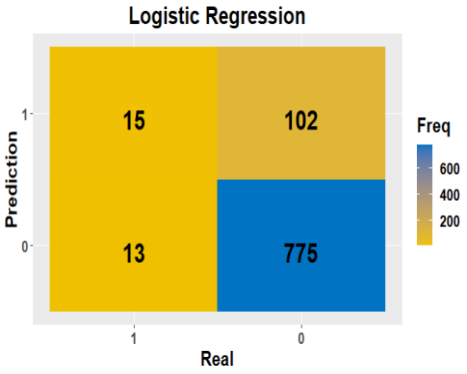
In order to fight multicollinearity, variable pdays is excluded from our model. It is a fact that data for the last contract duration cannot be collected before making a phone call. Hence in order to build a realistic predictive model the variable duration is not included in the models. All the categorical variable values in the data frame were encoded into numerical values using one hot encoding method, in order for those to be incorporated in the model. The performance of the models is evaluated with creating a confusion matrix and calculating metrics of Accuracy, F1 score, Precision and recall.

Models were trained on 80% of the data and tested on both test data (remaining 20%) and train data itself. The results are given below. The confusion matrices shown below are based on results on test data.

Model 1: Logistic Regression –

Simple binary logistic model. This model is taken as the baseline model. In general model performs well on both testing and training data as prediction accuracy on both sets are over 80%. However, as mentioned earlier in the EDA, class distribution of the target variable y is highly uneven. Therefore, F1 score should be a better evaluator of the models.

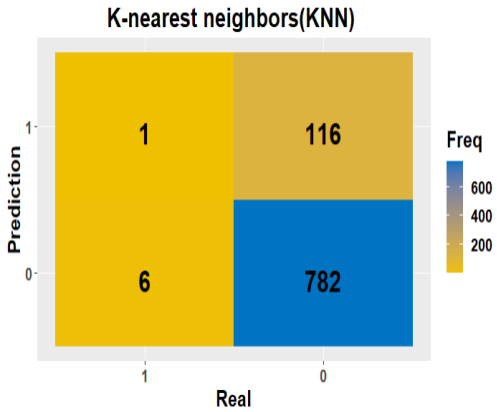
	Train Sample (3616-observations)	**Test Sample** (905-observations)
Accuracy	0.8985	0.8729
F1 Score	0.22500	0.19130
Precision	0.66176	0.73333
Recall	0.13554	0.11000



Model 2: K-Nearest neighbors' classification (*KNN)* –

KNN method was tested as supervised nonlinear classification model with the number of neighbors k equal to 10. Here the ideal value for k was calculated manually checking performance for a range of k values. Overall, the model seems to be underperforming compared to the baseline model with lower F1 scores.

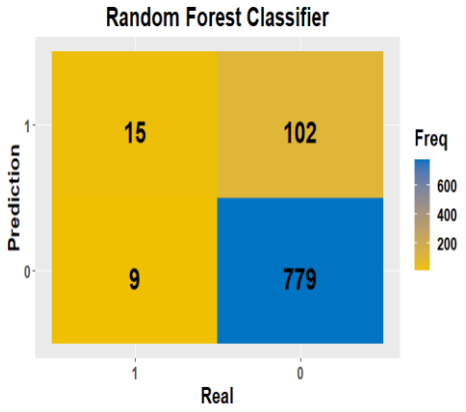
	Train Sample (3616-observations)	**Test Sample** (905-observations)
Accuracy	0.8946	0.8681
F1 Score	0.0986	0.0720
Precision	0.5454	0.3636
Recall	0.0542	0.0400



Model 3: Random Forest Classifier –

Nonlinear tree-based classification algorithms of random forests were tested to check whether it would make an improvement over the baseline model. Overall, Random Forest method predicts better on Train sample compared to baseline model with 97% accuracy and precision of 1. F1 score on train data predictions is also very high, however on testing data model accuracy and f1 scores are quite similar to that of baseline model.

	Train Sample (3616-observations)	**Test Sample** (905-observations)
Accuracy	0.9785	0.8796
F1 Score	0.8877	0.1607
Precision	1.0000	0.7500
Recall	0.7981	0.0900



In conclusion, out of the 3 models under consideration, Random Forest classifier works well on seen data whereas Logistic regression works better on unseen data. However, the overall objective of this model is to check whether a client would subscribe to term deposit. Therefore, number of true positive values are important. Hence model 3 with higher precision is the preferred classifier model. There is a possibility of further improvement of models with more feature engineering and hyperparameters tuning.

Reference:

- James, G. Witten, D. Hastie T. & Tibshirani, R. ISLR: Data for an Introduction to Statistical Learning with Applications in R 2017.
- Logistic Regression – [http://www.sthda.com/english/articles/36-classification-methods-essentials/151-logistic-regression-essentials-r/#:~:text=Logistic%20regression%20is%20used%20to,%2C%20diseased%20or%20non%2Ddiseased]
- F1 score – [https://www.statology.org/f1-score-vs-accuracy/]
- Precision, Recall, F1 – Score – [https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance measures]
- Data Visualization – [https://stackoverflow.com/questions/23891140/r-how-to-visualize-confusion-matrix-using-the-caret-package]
- Plots – [http://www.sthda.com/english/wiki/be-awesome-in-ggplot2-a-practical-guide-to-be-highly-effective-r-software-and-data-visualization]