COURSEWORK – EMFSS ST3189 – Machine Learning

Table of Content

Task 1

We have a dataset of 7647 perceptions from a study with 11 elements, the most important of which is to depict a person's state and temperament. It is worth mentioning that at first, none of the factors have lost values. This could be used for cluster analysis and interpretation.

EDA

Table 1.

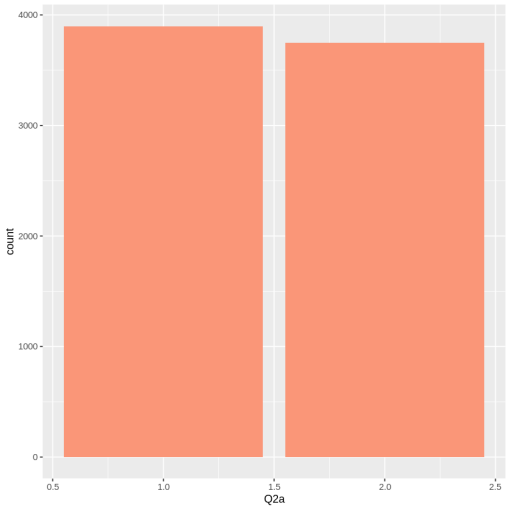
Q2a	Q2b	Q87a	Q87b	Q87c
Min. :1.00	Min. :15.00	Min. :1.000	Min. :1.000	Min. :1.000
1st Qu.:1.00	1st Qu.:34.00	1st Qu.:2.000	1st Qu.:2.000	1st Qu.:2.000
Median :1.00	Median :43.00	Median :2.000	Median :2.000	Median :2.000
Mean :1.49	Mean :43.16	Mean :2.426	Mean :2.606	Mean :2.415
3rd Qu.:2.00	3rd Qu.:52.00	3rd Qu.:3.000	3rd Qu.:3.000	3rd Qu.:3.000
Max. :2.00	Max. :87.00	Max. :6.000	Max. :6.000	Max. :6.000

Q87d	Q87e	Q90a	Q90b
Min. :1.000	Min. :1.000	Min. :1.000	Min. :1.000
1st Qu.:2.000	1st Qu.:2.000	1st Qu.:2.000	1st Qu.:1.000
Median :2.000	Median :2.000	Median :2.000	Median :2.000
Mean :2.717	Mean :2.408	Mean :2.126	Mean :2.194
3rd Qu.:3.000	3rd Qu.:3.000	3rd Qu.:3.000	3rd Qu.:3.000
Max. :6.000	Max. :6.000	Max. :5.000	Max. :5.000

Q90c	Q90f
Min. :1.000	Min. :1.000
1st Qu.:1.000	1st Qu.:1.000
Median :2.000	Median :1.000
Mean :2.175	Mean :1.531
3rd Qu.:3.000	3rd Qu.:2.000
Max. :5.000	Max. :5.000

1. Q2a – Gender:

Fig .1



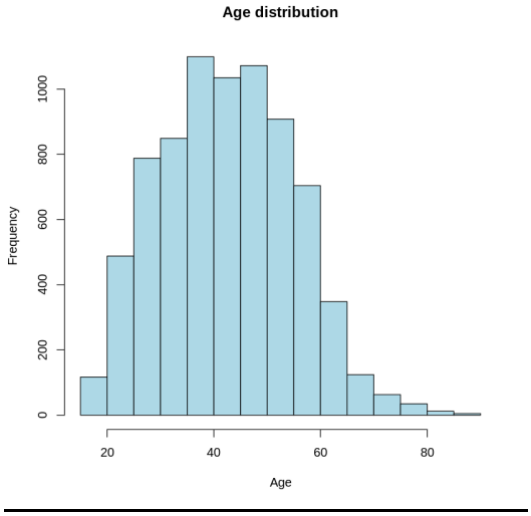
1 – Males

2- Females

According to the Fig.1 the number of women are slightly less than men. However, we can say that there are approximately equal number of women and men.

2. Q2b – Age

Fig.2

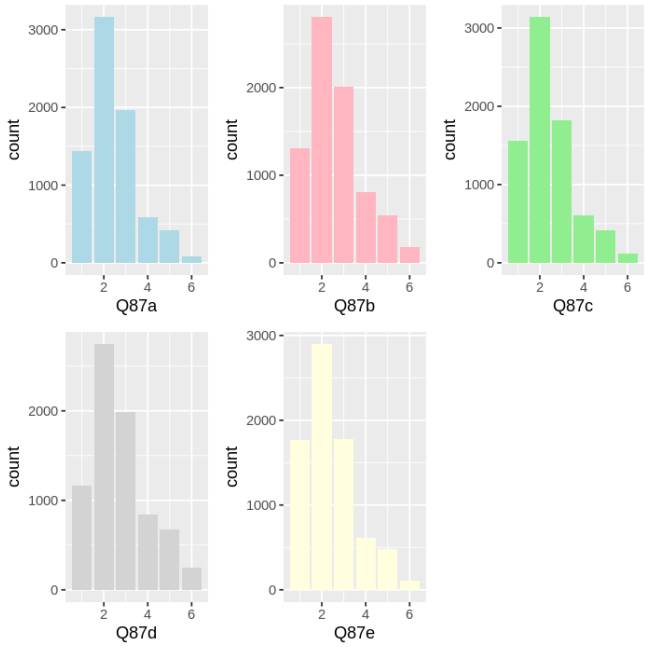


The mean age is 43 (table 1). Data is a slightly skewed. So, the major of people are in the group 34 – 52. And there are a few people older than 60.

3. Q87a - Q87e

   Estimation of persons mood and behavior.

Fig. 3



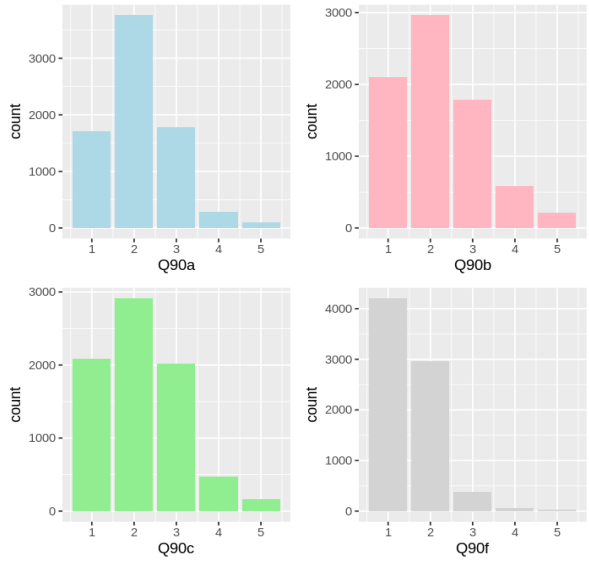
There are 6 values indicating mood estimation.

According to Fig.3 the frequent answer is 2 (Most of the time), and the rarest is 6 (At no time)

4. Q90a - Q90f

   Measure of persons attitude to life.

Fig.4

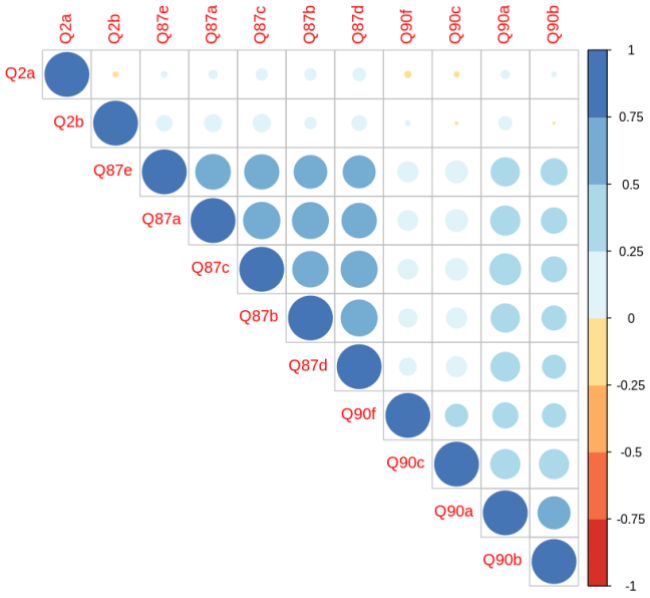


There are 5 variables. The most popular is 1 (Always) and 2 (Most of the Time).

According to this, the significant part of observations is people, who loves their job.

5. Correlation analysis

Fig. 5



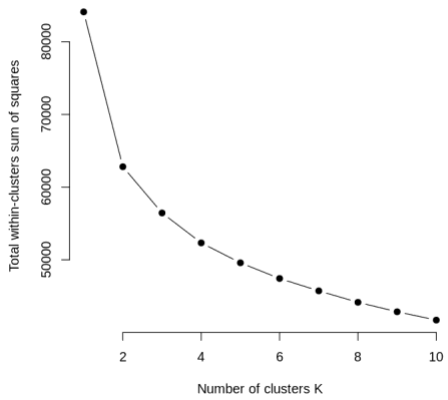
According to the correlation map(Fig. 5), the mood variables are highly correlated with each other. It is logically, because people in good or bad mood would respond equally to all questions.

Clustering

I decided to use cluster segmentations.

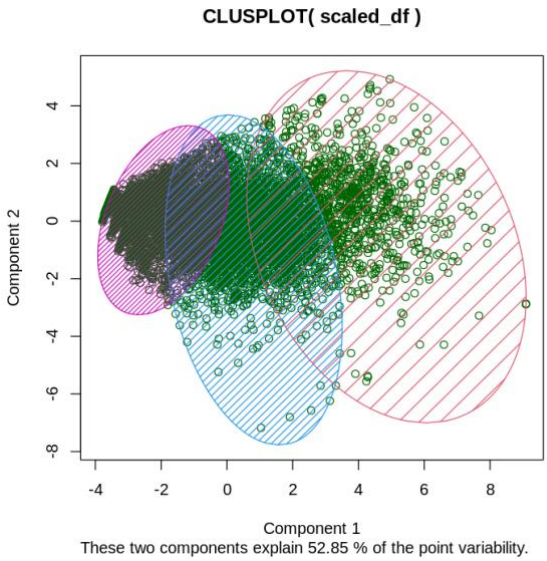
At first step, we need to estimate the best number of clusters. I used “Elbow Method” for this.

Fig.6



The best number of clusters for our data is 3 (fig.6), because our data is well divided into 3 clusters.

Fig. 7



A tibble: 3 × 12

Cluster	Q2a	Q2b	Q87a	Q87b	Q87c	Q87d	Q87e	Q90a	Q90b	Q90c	Q90f
<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	1.463209	43.03910	2.437687	2.638881	2.435515	2.734999	2.475156	2.270703	2.431442	2.387999	1.675265
2	1.574671	46.37848	4.002196	4.240117	4.053441	4.431918	3.859444	2.971449	2.912884	2.699122	1.797950
3	1.483834	41.63972	1.581216	1.700539	1.524634	1.790608	1.548499	1.477290	1.479600	1.598537	1.184758

Table 2

There are 3 segments (Fig.7) and Summary mean information about clusters (Table 2).

1 cluster – People with ordinary (most frequent) marks (men majority)

2 cluster –People with low marks (women majority)

3 cluster – People with extremely high marks (men majority)

Results

This analysis can be use in future, for example for HR management improvements, like creating good working environment and preventing attrition. So, according to our results women tend to be less satisfied with their positions at work than men.

Task 2

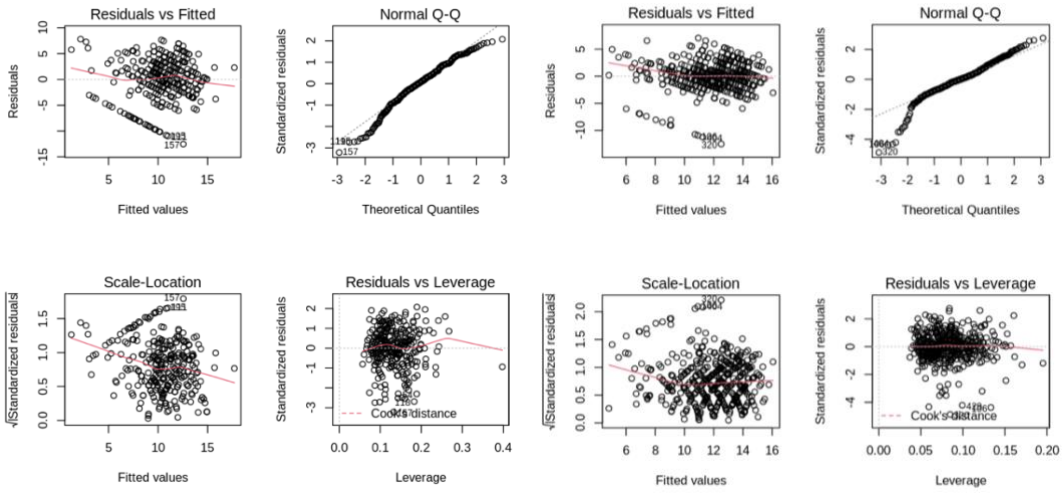
There are 2 datasets for math and language.

1. We don't have missing values in both datasets
2. All categorical features should be converted.
3. We should drop G1 and G2

Linear Regression

All categorical data was encoded by One Hot Encoder and split in to train and test sample.

Fig.8 (math) Fig.9 (language)



We can see that a person receives an average grade for math (intercept) of 11.40, whereas the average grade for language is 9.7

Results on test data:

RMSE: 887445.545292266 R2: 0.294427746291415

Table 3 (math)

RMSE: 1254711.9147298 R2: 0.384717688917388

Table 4 (language)

The residuals vs fitted graphs (Figs. 8, Figs. 9) illustrate that such models is fundamentally difficult to predict the data for both regressions. We can notice that the values of RMSE are really high and R^2 values are low (table 3, table 4).
Results

Since the model has low results on the test sample we can confirm that without variables G1 and G2 it is really hard to predict G3. It is vital to note that with so many variables, it is critical not to overlook the issue of multicollinearity.

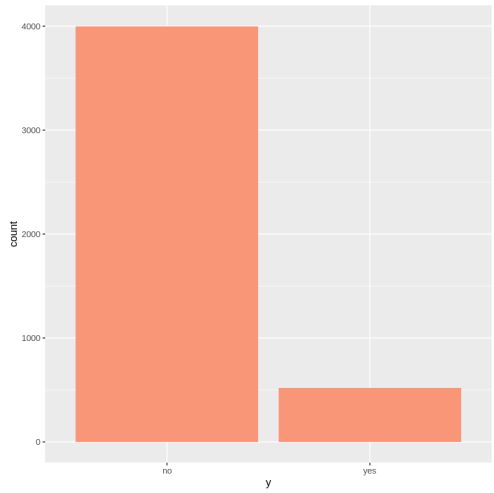
Task 3

Classification

In this task the goal is to predict the subscription on bank service.
The data does not contain NAN values.

The target variable is imbalanced 88% for “no” and 12% for “yes”. (Fig.10)

Fig.10



At first step, all categorical variables were transformed into numerical by Label encoder and Duration variable was deleted from the sample.
Then, The Logistic regression model was applied as baseline model for our dataset.
Since we are handled with imbalanced dataset, we can assess the quality of prediction only by Accuracy because it is not count the class difference. That is why the F1 score is was also used.

The data imbalance can be overcoming by some methods like “oversampling”,”undersampling”,”boosting” and setting weights for classes. I tried the “oversampling” method and setting weights, however the “oversampling” method did not improve the baseline score. Therefore it was decided to setting different weights for improving predictions.

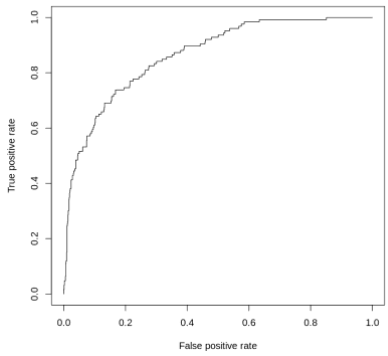
Models comparison

According to the Table 5 it can be observed that the baseline score is high, the data was good classified. The accuracy and f1 score is 88% and 94% respectively (table 5) and the value of ROC-AUC metric was 0.88 (Fig.11)

[1] "Logistic Regression"	[1] "Decision Tree"	[1] "Random Forest"
[1] "Accuracy - 0.8896"	[1] "Accuracy - 0.8942"	[1] "Accuracy - 0.8942"
[1] "F1 score - 0.941"	[1] "F1 score - 0.94281"	[1] "F1 score - 0.94293"

Table 5 Table 6 Table 7

Fig.11



After, applying baseline model the Random Forest model and Decision tree model was applied. Based on results both models performed better than the baseline, however prediction improvements are quite small (Table 6 ,Table 7).

The Decision tree model and Random Forest model has the same results, so the further analysis of the prediction will be carried out on the Decision tree model.
Based on graph on Fig 12 and Fig 13. the most importance feature for classification will customer subscribe or not are “poutcome”, “mounth” , “day” and “job”.

Fig. 12

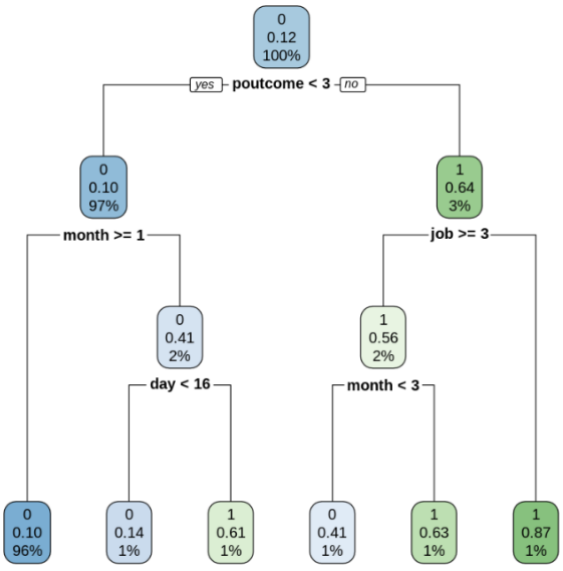
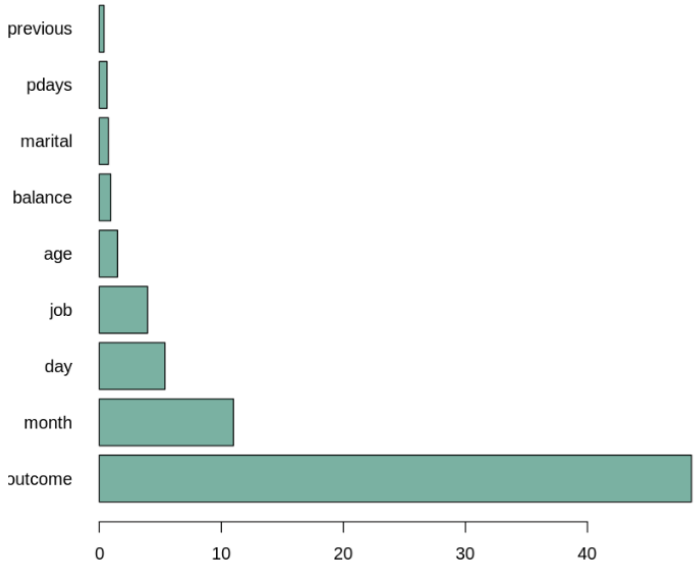


Fig.13



Results

To sum up, the decision tree and random forest models perform well for this particular problem in the financial sector.

This model will be useful to the bank's partners, as it shows which clients should be paid attention to attract them to subscribe a term deposit. Based on the results of the predictive general model. clients with a "successful" past marketing campaign are more willing to subscribe than others. Also, the people with following jobs : 'admin', 'blue-collar', 'entrepreneur' tend to deny subscription more than others and people who contacted the previous company less than 16 days are inclined to accept the offer.