University of London
London School of Economics and Political Science

Coursework – EMFSS ST3189 – Machine Learning

Student Number: 200692700
Candidate number: A28699

Part 1 – EWCS

In the first part we asked to explore EWCS dataset using unsupervised learning techniques such as Principal Component Analysis (PCA) to illustrate the information in the data with relevant tables and graphs.

Firstly, let's make Exploratory Data Analysis which is a very important step when we work with raw dataset.

Observing the structure, we see that there are 7813 observations and 11 variables in EWCS dataset, no null features. However, if we look at the statistics we can see that the minimal value in each variable is -999 which means we have incorrect data here, so let's just remove the rows containing such values. Here what we have after cleaning the data:

Data statistics #2

	count	mean	std	min	25%	50%	75%	max
Q2a	7647.0	1.490127	0.499935	1.0	1.0	1.0	2.0	2.0
Q2b	7647.0	43.160194	12.366371	15.0	34.0	43.0	52.0	87.0
Q87a	7647.0	2.426180	1.108552	1.0	2.0	2.0	3.0	6.0
Q87b	7647.0	2.606120	1.222427	1.0	2.0	2.0	3.0	6.0
Q87c	7647.0	2.415065	1.145142	1.0	2.0	2.0	3.0	6.0
Q87d	7647.0	2.717275	1.279363	1.0	2.0	2.0	3.0	6.0
Q87e	7647.0	2.407611	1.188074	1.0	2.0	2.0	3.0	6.0
Q90a	7647.0	2.126324	0.846588	1.0	2.0	2.0	3.0	5.0
Q90b	7647.0	2.194063	1.013382	1.0	1.0	2.0	3.0	5.0
Q90c	7647.0	2.175363	0.969037	1.0	1.0	2.0	3.0	5.0
Q90f	7647.0	1.530535	0.673537	1.0	1.0	1.0	2.0	5.0

Figure 1: Statisctical table

We can see that there are less observations and data looks correct now in terms of min, max according to the data description. Now, let's investigate how features are correlated and distributed in order to understand it better:

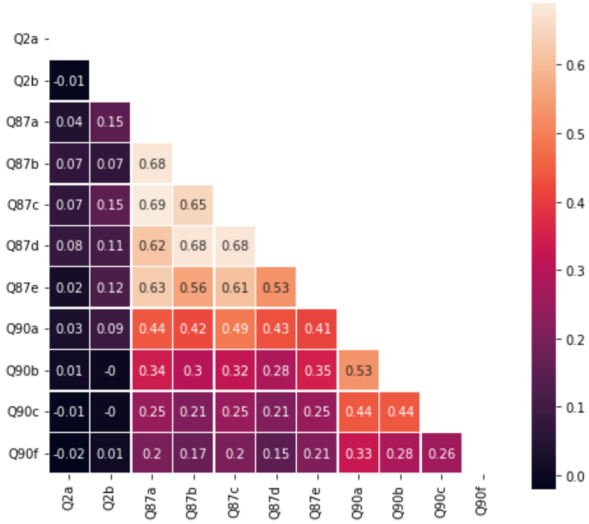


Figure 2: Heatmap

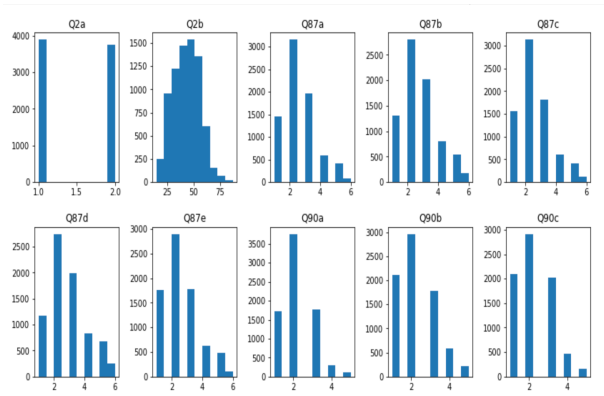
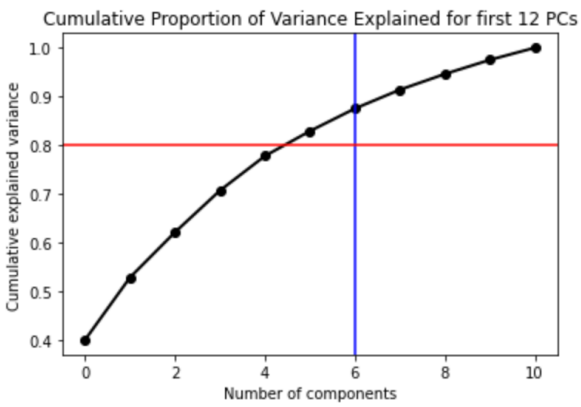
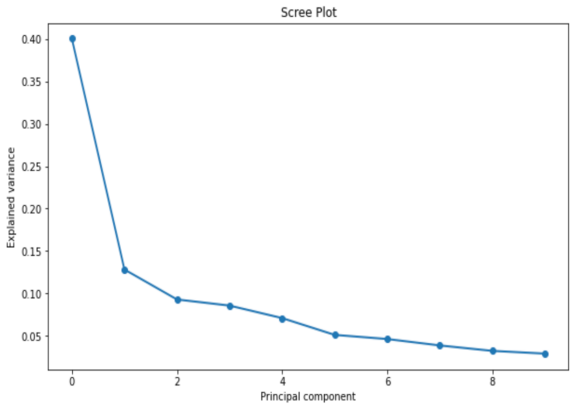


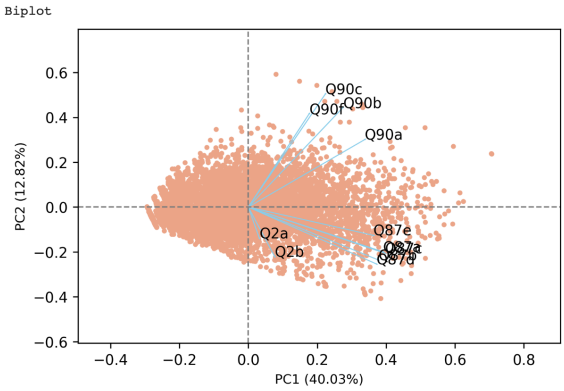
Figure 3: Histograms

Considering Figure 2 we can highlight few points:
1) There are positive moderate correlations between Q87a,b,c,d,e features with each other and the same level of relationships have Q90a with Q90b which is not something surprising considering the meaning of these features;
2) There is almost no correlation between Q2a&b features with all other features, which means the gender and age parameters have almost no or very little effect on other parameters.
By observing Histograms it is clear that almost all features are skewed right (positively).

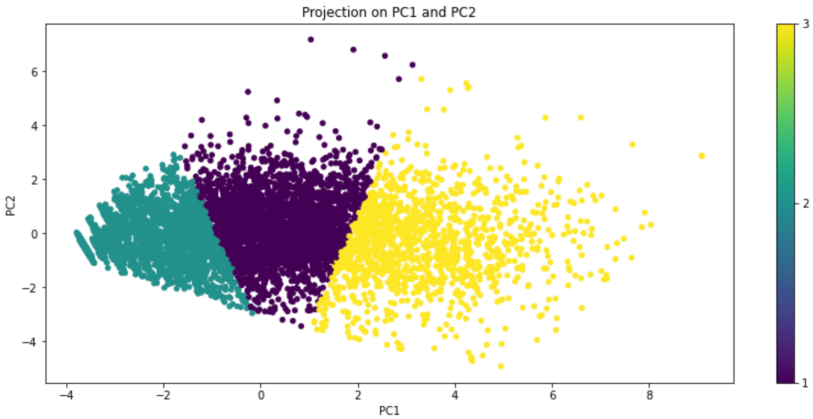
Let's move to the applied PCA and observe results.







We obtain 11 components, with 5 of them explaining slightly more than 80% of the variance in the dataset according to the first plot. From the both Biplot and Scree plot we can see that there are 3 main clusters which correspond to the Q2, Q87 and Q90 groups. Now, we can visualize these clusters on PC1and PC2:



The boundaries between clusters are very clear on the projection of 561's dimensional space on first two PCs.

Part 2 - Student Performance Dataset

In this section, we are asked to build a regression model for the variable G3 (final grade) and evaluate its predictive performance.

At the beginning we need to prepare our dataesets by removing G1, G2 and G3 from dataframes that will be used for testing and training. Then we need to deal with categorical variables assuming one-hot encoding on datasets. There is no missing data needed to be improved or deleted.

After little preparation, we can proceed to the splitting process. All ations for one dataset are duplicated for the second. We need to explore different models and decide which model performance is better for each of them. The Simple Linear Regression model will be treated as a baseline which we fit on all features and we will consider 4 more models: Lasso Regression, Ridge, SVR and Random Forest. Performance is being evaluated by the R^2 metrics, mean squared error (better interpret it by taking a square root) and mean absolute error. So, let's take a look at the tables, on the left part we have test and train sets for the Math and on the right for the Portuguese respectively:

	R^2	MSE	MAE
Linear Regression	0.231324	16.354659	3.186930
Lasso	0.035124	20.529078	3.398820
Ridge	0.236830	16.237519	3.176645
SVR	0.106034	19.020368	3.223672
Random Forest	0.364192	13.527699	2.693165

	R^2	MSE	MAE
Linear Regression	0.263691	15.353030	2.962797
Lasso	0.028086	20.265712	3.373258
Ridge	0.263600	15.354923	2.962868
SVR	0.082397	19.133248	3.197625
Random Forest	0.900166	2.081680	1.053987

	R^2	MSE	MAE
Linear Regression	0.300980	5.965909	1.836380
Lasso	-0.006491	8.590075	2.232949
Ridge	0.302797	5.950401	1.832644
SVR	0.198866	6.837417	1.881959
Random Forest	0.393393	5.177197	1.781385

	R^2	MSE	MAE
Linear Regression	0.359469	6.966823	1.904964
Lasso	0.005185	10.820238	2.433267
Ridge	0.359447	6.967064	1.904723
SVR	0.216083	8.526383	2.050132
Random Forest	0.891774	1.177133	0.778844

It is clear that Random Forest performs significantly better than any other model on both datasets; the MSE and MAE values for both train and test are far lower than those of others; and for training sets only, this model has a R squared value near to 1, while others are just over 0. Aside from that, both the simple linear model and ridge regression appear to produce reasonable results, although not as well as the tree-based approach.

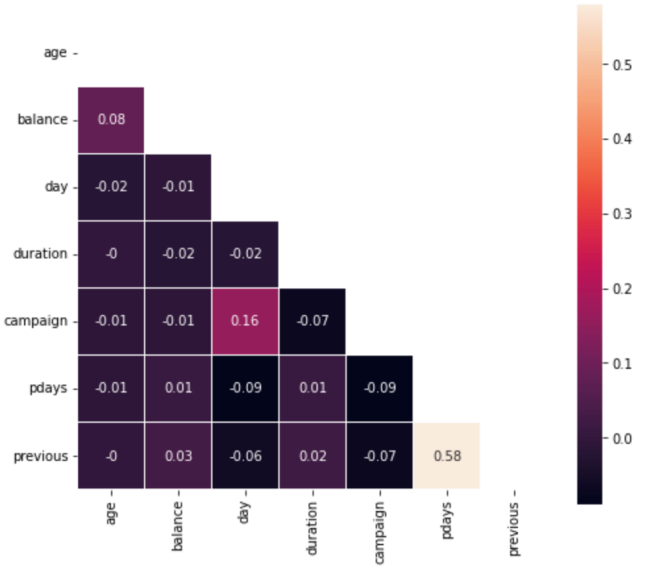
Based on the results, the top three models are Random Forest, Ridge Regression, and Linear Regression; these results may also be enhanced by performing hyperparameter tweaking.

Part 3

The purpose of this section of coursework is to create classification models using the available dataset to assess if a client would subscribe to a term deposit.

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
age	4521.0	NaN	NaN	NaN	41.170095	10.576211	19.0	33.0	39.0	49.0	87.0
job	4521	12	management	969	NaN	NaN	NaN	NaN	NaN	NaN	NaN
marital	4521	3	married	2797	NaN	NaN	NaN	NaN	NaN	NaN	NaN
education	4521	4	secondary	2306	NaN	NaN	NaN	NaN	NaN	NaN	NaN
default	4521	2	no	4445	NaN	NaN	NaN	NaN	NaN	NaN	NaN
balance	4521.0	NaN	NaN	NaN	1422.657819	3009.638142	-3313.0	69.0	444.0	1480.0	71188.0
housing	4521	2	yes	2559	NaN	NaN	NaN	NaN	NaN	NaN	NaN
loan	4521	2	no	3830	NaN	NaN	NaN	NaN	NaN	NaN	NaN
contact	4521	3	cellular	2896	NaN	NaN	NaN	NaN	NaN	NaN	NaN
day	4521.0	NaN	NaN	NaN	15.915284	8.247667	1.0	9.0	16.0	21.0	31.0
month	4521	12	may	1398	NaN	NaN	NaN	NaN	NaN	NaN	NaN
duration	4521.0	NaN	NaN	NaN	263.961292	259.856633	4.0	104.0	185.0	329.0	3025.0
campaign	4521.0	NaN	NaN	NaN	2.79363	3.109807	1.0	1.0	2.0	3.0	50.0
pdays	4521.0	NaN	NaN	NaN	39.766645	100.121124	-1.0	-1.0	-1.0	-1.0	871.0
previous	4521.0	NaN	NaN	NaN	0.542579	1.693562	0.0	0.0	0.0	0.0	25.0
poutcome	4521	4	unknown	3705	NaN	NaN	NaN	NaN	NaN	NaN	NaN
y	4521	2	no	4000	NaN	NaN	NaN	NaN	NaN	NaN	NaN

As we can see the dataset consists neither missing values, and when the statistics of the variables are examined, it is clear that the standard error of balance is relatively high (due to its range), the age, day, and campaign variables have means that are nearly identical to their medians, and previous, balance, and campaign have right skewed tails.



When seen as a heatmap, pdays and previous have the largest connection of 0.58, whereas previous has little to no correlation with others. Aside from that, all other factors show essentially little association with each other, with the exception of day and campaign, which have a 0.16 correlation. We should then use 4 KNN as a baseline and Logistic regression, LDA, QDA and Random Forest Classifier as other models.

References

- James, G., Witten, D., Hastie, T., & Tibshirani, R. ISLR: Data for an Introduction to Statistical Learning with Applications in R, 2017.
- https://plotly.com/python/pca-visualization/
- https://scikit-learn.org/stable/auto_examples/preprocessing/plot_scaling_importance.html