EMFSS ST3189
Machine Learning
Coursework

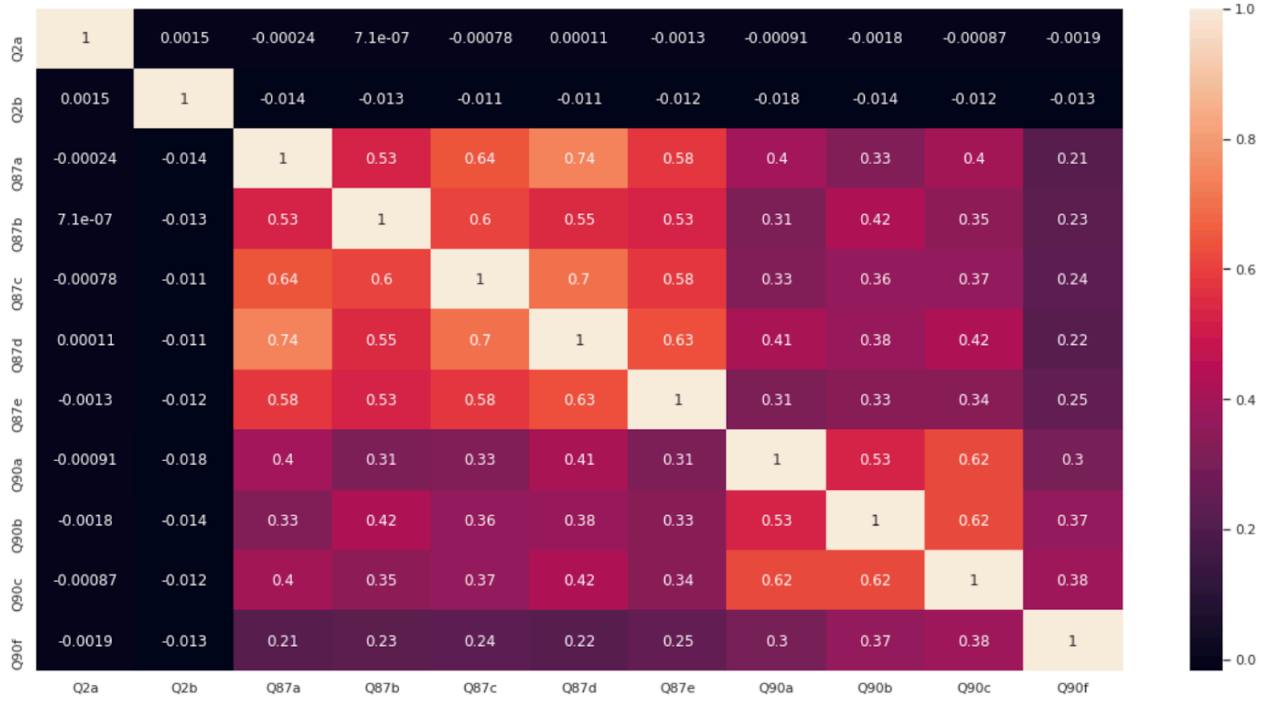
Student Number: 190534002

Table of contents:

Student Number: 190534002

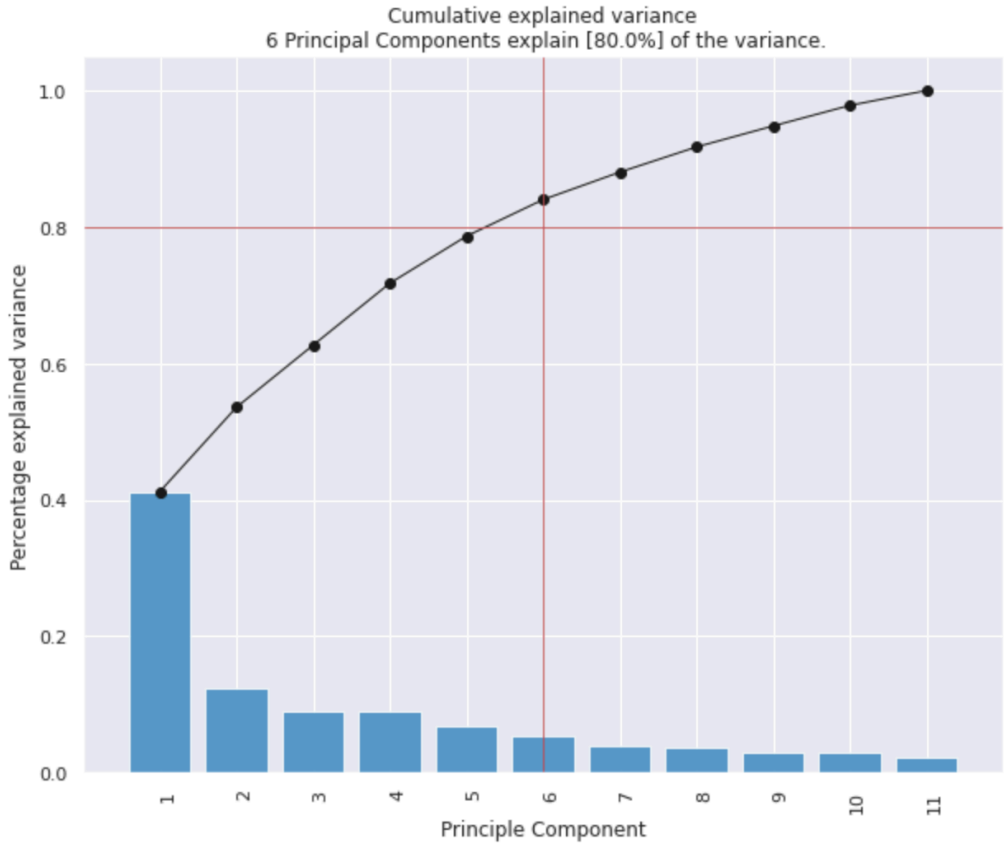
Part 1

The dataset given in the first part is constructed on the data, collected via the European Working Conditions Survey 2016. The main goal of the analysis in this part is to summarise the information, using various visualizations, and describe the data by applying the unsupervised learning techniques to the provided dataset. To explore the data the heatmap correlation matrix is constructed.



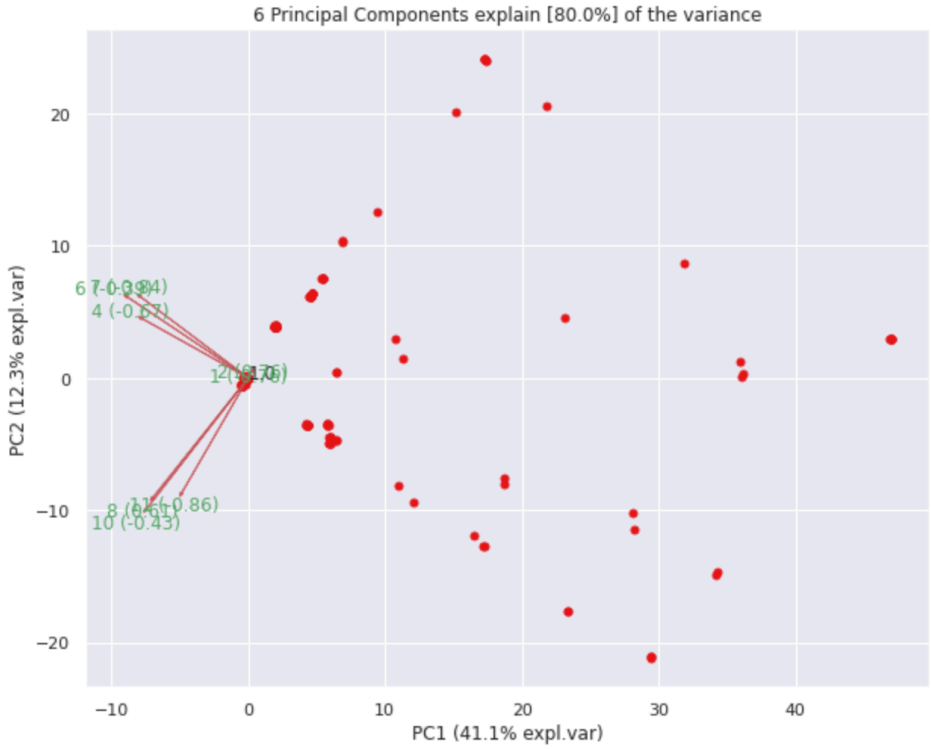
It can be seen on the matrix that there are a lot of correlated features in the dataset. For example, the first two columns (and rows), Q2a and Q2b, which are gender and age respectively, are almost not correlating with other attributes, whereas answers on all other questions are strongly correlating with each other. Such results of a survey with a large number of correlated answers are hard to analyze, because you cannot track the influence of each answer separately, and attempting to analyze such data will lead only to confusion.

The model that is used to cope with collinearity is the Principal Component Analysis(PCA). First of all we should scale the data. It is very important for Principal Component Analysis to have scaled data with a mean 0 and standard deviation of 1. This will improve understanding data by the model influence of features to each other and as result, it will improve the quality of the model.



After scaling the PCA is applied. The model is constructed in such a way that 80 percent of data variability is explained. We can see the results of building models with different amounts of principal components.

We can see that we have chosen 6 principal components to have 80% variance explained. Also it can be noticed that the first principal component takes the most variance. Other principal components are not so large and it seems like the rest of the variance is divided equally among them. Now let us inspect the results on biplot.



It can be seen here that some of the arrows are overlapped. It means that variables represented by this arrows are correlated. So we can see that PCA successfully caught this correlation.

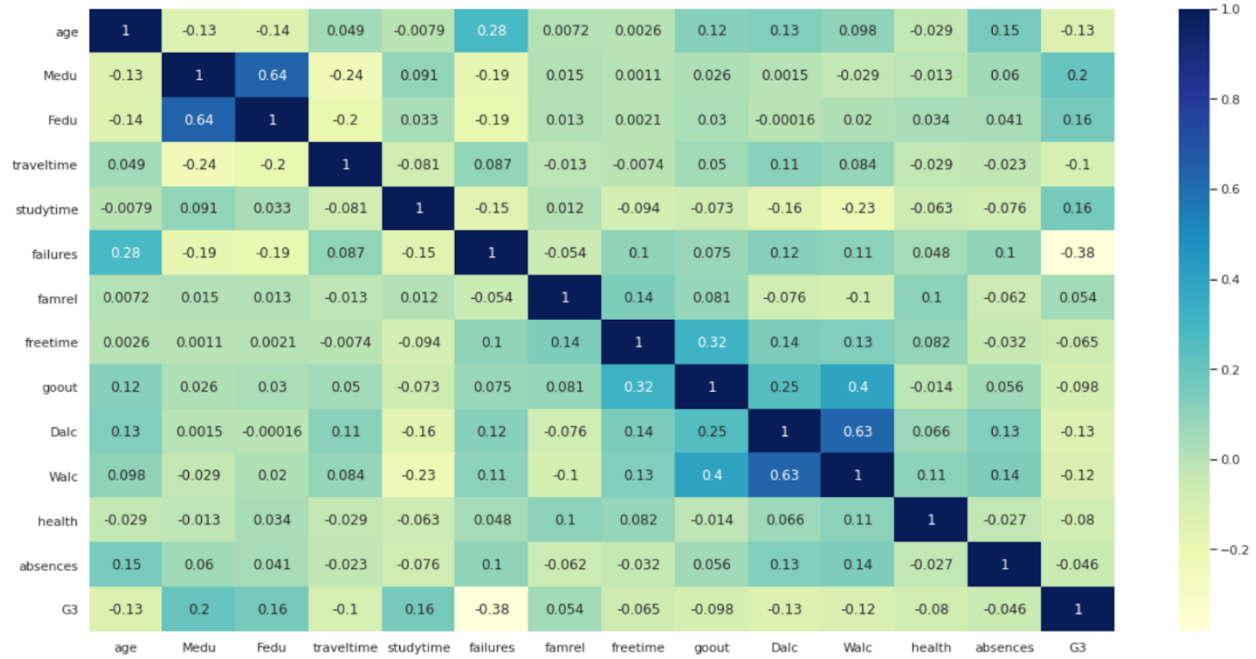
PC1	Q87d	-0.3870423751321271	best
PC2	Q90c	-0.4344536776177707	best
PC3	Q2a	-0.7602485977653248	best
PC4	Q2b	0.7600922794647411	best
PC5	Q90f	-0.8616607682652841	best
PC6	Q87b	-0.6711584262164054	best
PC7	Q87e	-0.8404484970014545	best
PC8	Q90a	0.6109788794575037	best
PC9	Q90c	-0.6351484826045629	best
PC10	Q87c	-0.6045323184965207	best
PC11	Q87d	-0.7955763452153696	best
PC10	Q87a	0.5242280068896257	weak
PC6	Q90b	-0.4766457622537542	weak

There is a table, provided above, in which it can be seen that various features have different impacts on the creation of the principal components. For instance, it can be noticed that not only age and gender are used in the forming of principal components, but also meaningful questions are contributing to it. For example, Q87d, which is the answer “I woke up feeling fresh and rested” has proven to be the best. Whereas, there are two weaf attributes, they are Q87a and Q90b, which correspond to “ I have felt cheerful and in good spirits” and “I am enthusiastic about my job” respectively.

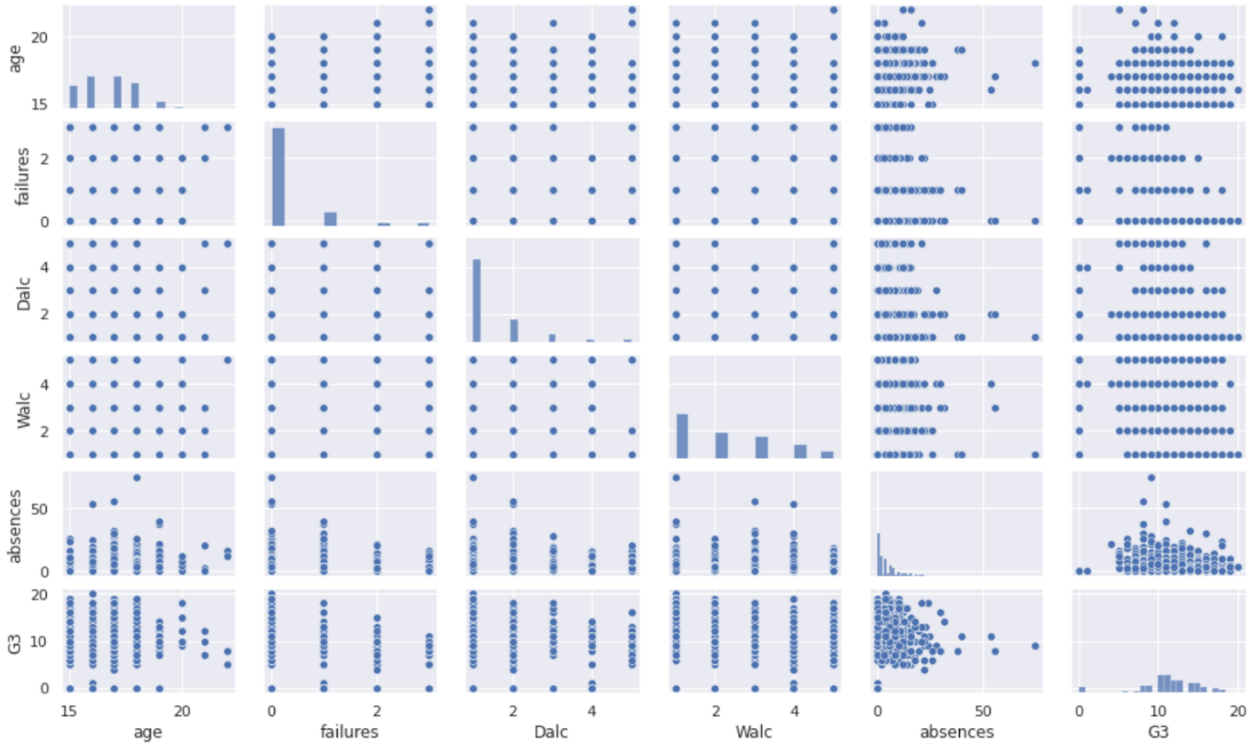
Part 2

In this part, two datasets are given. The first one is connected with the performance of students in mathematics and the second is connected with the Portuguese language. The data was obtained from two Portuguese schools and shows the achievements of students there. The information included in the dataset is: student grades, demographic, social and school-related features. Such elements of data were collected by analyzing school reports and questionnaires. The goal of the analysis of this dataset is to predict the final year grade, G3, using the attributes, except for the middle grads, G1 and G2.

As these two datasets are quite small they can be united into one more informative dataset to improve the predictions. To begin with, the variables “G1” and “G2” were excluded from the dataframe, as it was mentioned before, they should not be used in the prediction.



To explore the data the heatmap correlation matrix is constructed. It can be seen that “Medu” and “Fedu” attributes are strongly correlated, so it is better to drop them from the dataframe. Other highly correlating variables are “Walc” and “Dalc”.



It can be noticed that “Walc” has less impact on the final grade than “Dalc” and they are performing similar templates, so one of these attributes can be dropped. Then, “Walc” is dropped from the dataset. To prove observations and consider the data better the pairplot is constructed, where it can be seen that “romantic” is also an extra attribute, which can be excluded.

The next step is model evaluation. There are several models considered: Random Forest Regressor, LGBMRegressor, GradientBoostingRegressor and XGBRegressor. You can see the result in the following table:

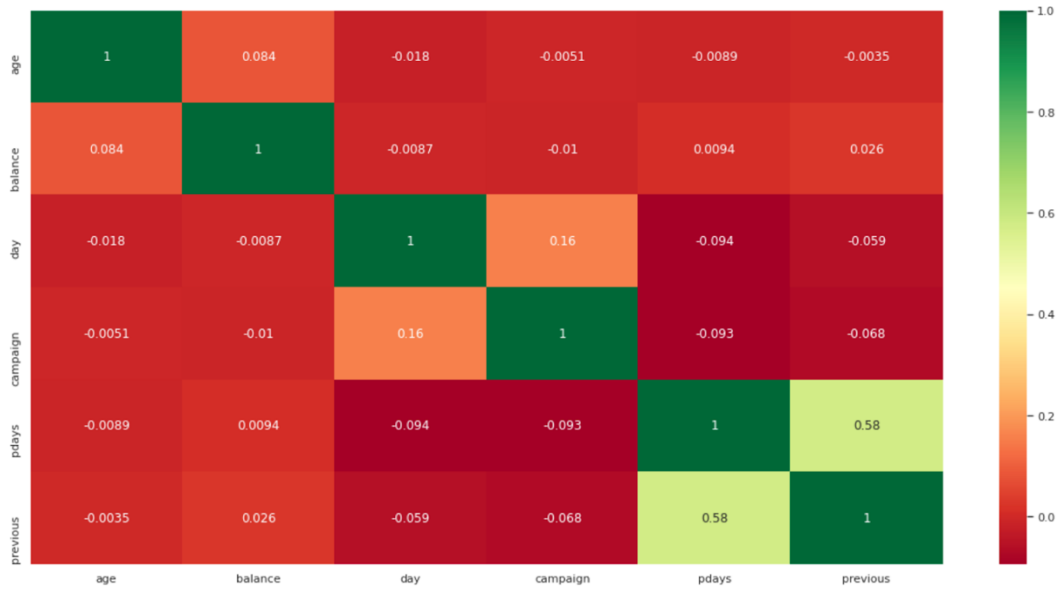
	Model object	MSE float64
0	Random Forest	9.330208770206504
1	Gradient boosting	9.884264263375744
2	Light GBM	10.187595181741656
3	XGBoost	10.520628002731415

These models were applied to calculate MSE and the results were compared. So, the best MSE is performed by the Random Forest Regressor model. The MSE for Random Forest Regressor is equal to 9.884264263375744. To define the value of the error calculating the RMSE or taking the square root out of MSE: $\sqrt{9.330208770206504}$= 3.055.

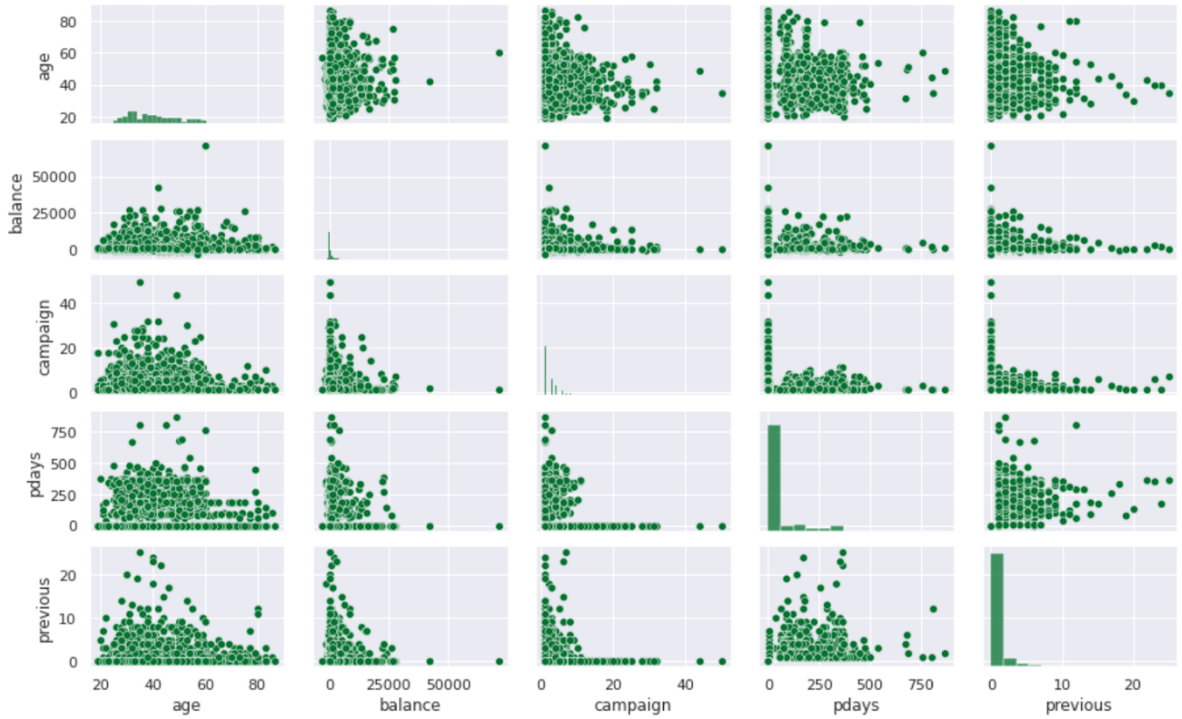
Overall, the average error is 3.055. So, in the twenty-point system of grading the average value of the error is 3.055, which is high enough, but the train set was not large. However model seems to be sustainable. We can improve this score by increasing the size of training set and adding new variables with more information about past students performance.

Part 3

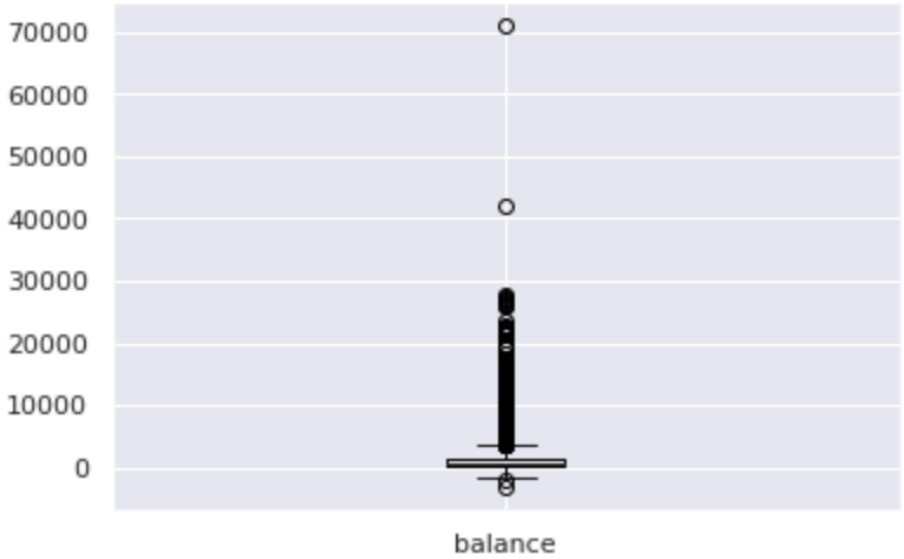
The dataset given is related to the marketing campaigns of a Portuguese banking institution. The marketing campaigns are based on the data, obtained from the telephone calls, for example, data about the person, like job or education and data about the call, like day and duration. The goal of the analysis of this dataset is to predict whether the client will subscribe to a term deposit.



To explore the data the heatmap correlation matrix is constructed. It can be seen that “previous” and “pdays” are strongly correlated. Other variables are correlated slightly.



One more graph constructed is pairplot. It can be noticed that “balance” variable has outliers. The box plot is constructed to take a better look at it. Other variables are not so informative.



It can be seen that there are really many outliers, so we want to catch them. A new variable “detector” can be introduced, which will indicate the presence of a balance outlier by “1” and have “0” value otherwise. Also we should encode text variables somehow. Extra variables can bring additional variance to the model which is not quite good, but absence of necessary ones will lead to high bias. If one column is chosen to leave, then all the different values in it must be encoded with numbers from 1 to n. This approach allows not to add many unnecessary variables, but

assumes the same distance between the two values. That is, for example, assumed that the difference between primary and secondary education is the same as between secondary and higher, which may not be the case. In the given case, the same distance is suitable for months, they can be encoded as they go in a year, and for the rest of the variables, new columns will be created. Also in the variable age, there are only patterns associated with a very large age, which will be caught by the variable 'retired', so the age can be dropped.

The next step is model estimation. An accuracy score cannot be used, as it tends to overestimate performance of the model. For example, the default LGBM model, that I decided to use, gave almost 90 percent score, which cannot be the truth. We can see that this happens because of small percentage of predicted 1s:

	Predicted Negative int64	Predicted Positive int64
Actual Negative	1154	28
Actual Positive	146	29

So, the weighted F1-score is chosen for the estimation. It works better for the prediction of “1”. According to the F1-score calculated the best model is selected. It is “LGBMClassifier” And also we should apply a re-sampling method to balance the dataset for better prediction of 1s. The undersampling algorithm is chosen: deleting samples from the majority class. It removes “0” to make the training set more equal. More precisely, the Instance Hardness Threshold (IHT) method is used. IHT is removing hard samples to eliminate the imbalance of classes. IHT is implemented here, using the “imbalanced-learn” library, which works like removing the classes with low probability. Finally, the IHT constructs the model, based on the undersampled data. We can see that now we can better predict 1s.

	Predicted Negative int64	Predicted Positive int64
Actual Negative	557	625
Actual Positive	40	135

Now we can see the final result, which is obtained through parameters tuning. We can see that now we have not got really great accuracy, but we are able to predict most of 1s correctly. The percentage of predicted 1s which are actually 1s is not very big, but natural for task.

	Predicted Negative int64	Predicted Positive int64
Actual Negative	635	49
Actual Positive	547	126

The result of our work is that now we can classify about a half of our clients as potential subscribers of term deposit and among them will be about 70% of all subscribers.