Machine Learning - Programming Assignment

Marketing campaign

Bill Mono
Institutional ID: 479379
Email:
bill.mono01@universitadipavia.it

I. Introduction

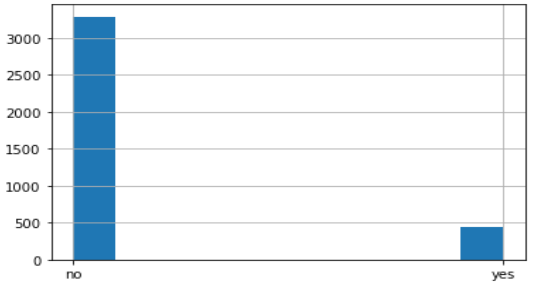
The goal of this work is to be able to develop a machine learning system capable of predicting the final decision of the customer to subscribe or not the deposit plan, based on the data collected by the operators through the different telephone calls that they have made. First of all, an exploratory analysis was carried out on the data in order to discover anomalies or relationships, using statistics tools and graphs. Then various predictive models were used and their results were compared to each other, so that the best performing model could be chosen.

II. Exploratory Analysis

The data were divided into 3 different datasets: 'bank_train.csv' with 3721 observations and 17 features, used for the model training phase, 'bank_validation.csv' and 'bank_test.csv' with 400 observations for both, used in the validation and test phase.

The data was loaded using the pandas data analysis library, after renaming the columns of the three datasets, an inspection was carried out on the data, to understand if there were any missing or duplicate values. One of the main steps in data preprocessing is handling missing data. Feeding missing data to your machine learning model could lead to wrong prediction or classification. Hence it is necessary to identify missing values and treat them. In our case there were no missing data or duplicates. Subsequently the columns with categories and the columns with continuous numerical values were identified, and the type of data present in each feature was checked, because it could be that the value of the features is different from what it should be.

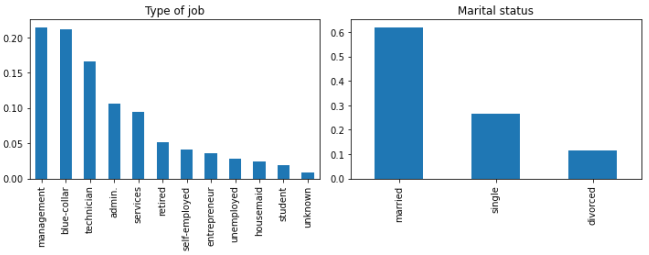
Another important part of data analysis is to see if there is an imbalance between classes. Since most machine learning algorithms assume that data is equally distributed, applying them on imbalanced data often results in bias towards majority classes and poor classification of minority classes. Hence we need to identify and deal with class imbalance.

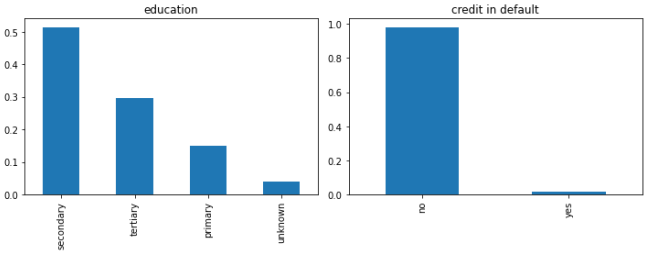


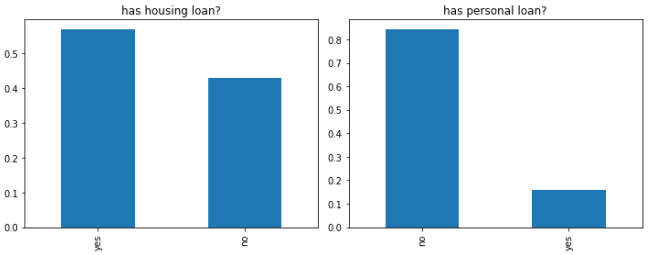
From the graph above it can be seen that there is a clear imbalance in the distribution of the target classes, with almost 89% of the target values belonging to the 'no' class and only 11% to the 'yes' class. This factor will need to be handled before applying the learning algorithms.

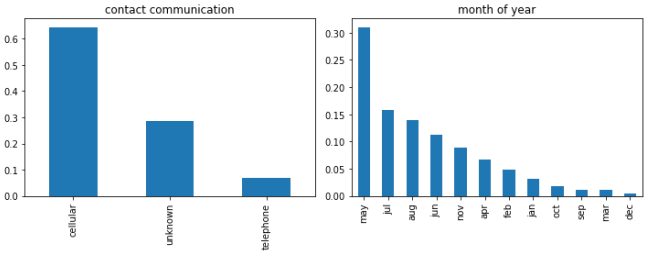
To better visualize the data, various graphs were made, barplots to better visualize the values of all categorical variables, and histogram for continuous variables.

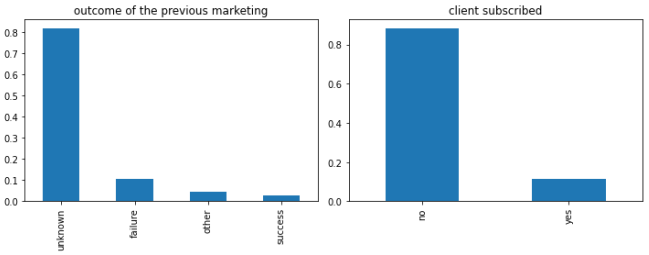
From the graphs below we can make some observations, for example the customers who were mostly called by phone, worked as blue-collar, in the administration and as technicians. Other considerations we can make, is that most of the customers are married people, many of the customers do not have credit in default, past customers have applied for a housing loan but very few have applied for personal loans, the mobile phones seem to be the preferred method of reaching customers, and the month in which the most customers were contacted it's May.



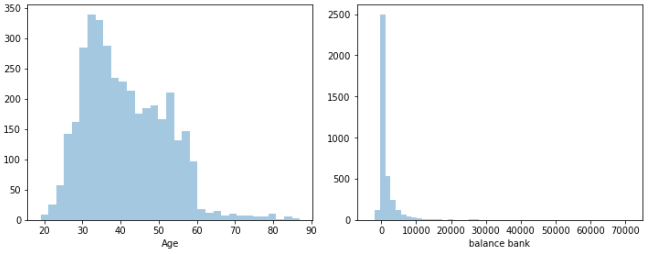


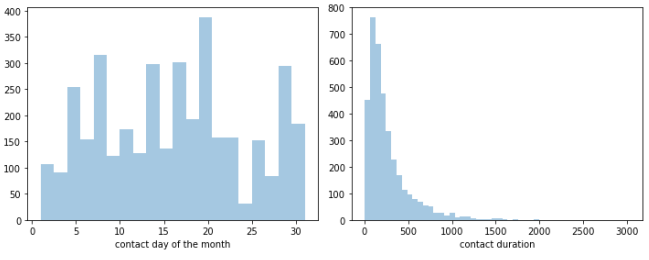


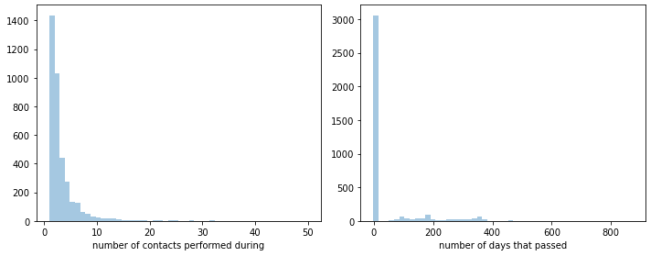




For columns that have continuous numerical values, we can make graphs to see, for each column, its distribution of values.

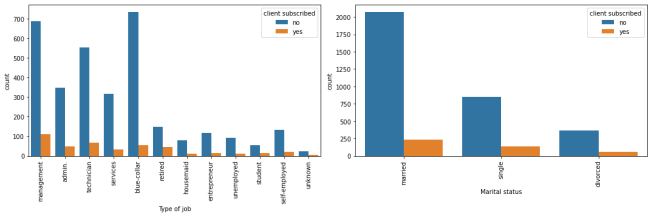


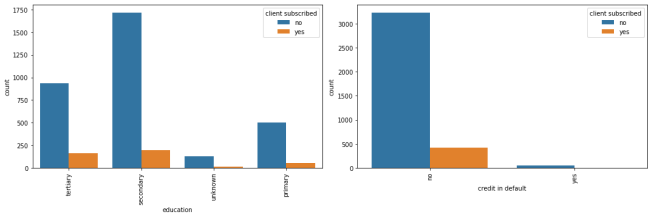


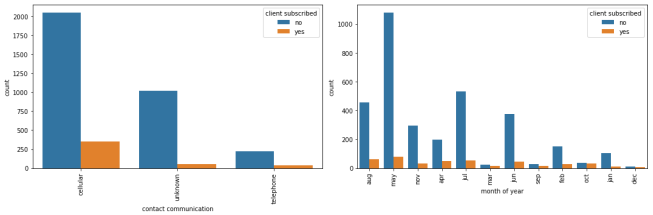


Here too we can make several considerations, for example one of these is that the age range of the most contacted customers goes from 30 to 40 years, and this is probably due to the fact that at that age it is more likely that the client has a job. For some variables, we can see the presence of many zeros, this could be due to the fact that many customers have been contacted for the first time.

In the graphs below, there are other representations useful for understanding the relationship between the target and the various categorical variables. In this way it is possible to see the common characteristics for the customers who have subscribed the term deposit.

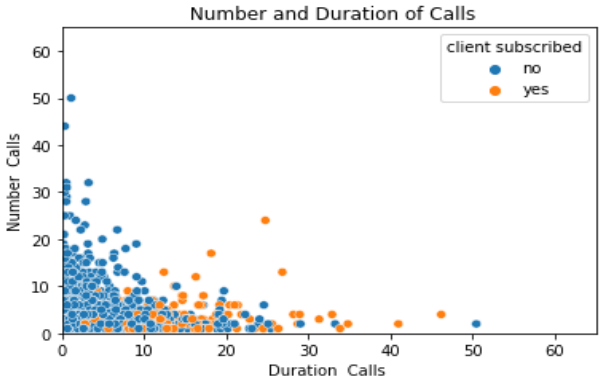


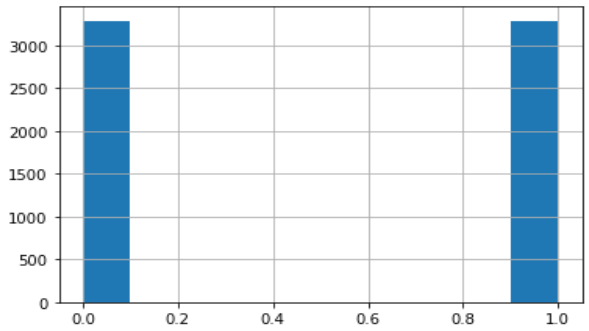




From the graphs that relate the target with the categorical variables, it is possible to say that the customers who have signed the term deposit are customers who have administrative duties,are married, have at least a degree, do not hold a credit in default, and mobile phones should be the preferred way of contacting them.

Another interesting relationship that was analyzed, was to see if there was a correlation between the duration of the call and the target variable, this is because the more the bank talks to a target customer, the greater the probability that the target customer opens a term deposit, since a longer duration means a greater interest from the prospect. Then a scatter plot was made, where the duration of the calls and the number of calls made were put as variables. In the graph below we can see how the longer the call duration, greater is the probability of subscribing to the deposit plan.





III. Data Preprocessing

After having done this analysis on the data, we begin to do a data preprocessing phase, so that the machine learning algorithms can learn better and have better performance.
First let's identify if there are any outliers, i.e. observations that lie far away from the majority of observations in the dataset and can be represented mathematically in different ways. We can see that there are mainly for the variables: '*balance bank*', '*number of contacts performed*', '*number of days that passed*'. To eliminate the outliers, a method called Winsorization was used. In this method we define a confidence interval of let's say 90% and then replace all the outliers below the 5th percentile with the value at 5th percentile and all the values above 95th percentile with the value at the 95th percentile. It is pretty useful when there are negative values and zeros in the features which cannot be treated with log transforms or square roots.

Another transformation that was done on the data was to scale it. Standardizing a data set involves scaling the distribution of values so that the mean of the observed values is 0 and the standard deviation is 1. This can be thought of as a subtraction of the mean value or a centering of the data. Like normalization, standardization can be useful and even required in some machine learning algorithms when the data has input values with different scales. To do this, a *scikit-learn* library was used.

Before applying our machine learning algorithm, we need to recollect that any algorithm can only read numerical values. It is therefore essential to encode categorical features into numerical values. Encoding of categorical variables can be performed in two ways, with *Label Encoding* or with *One-Hot Encoding*.
For the given dataset, we are going to label encode the categorical columns, using another scikit-learn library.

After encoding the categories, the problem of target balance in the dataset must be solved. We had seen before that there was a clear imbalance between classes, which could made the performance of our model worse, so to solve this problem we used the *RandomOverSampler* library from Imblearn, with the strategy of increasing the observations of the minority class to the same number as the majority.

IV. Feature Selection

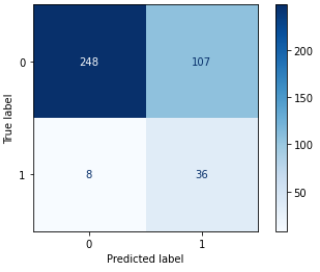
Feature selection is the process of choosing variables that are useful in predicting the response . It is considered a good practice to identify which features are important when building predictive models. The *backward elimination* was used as a method of selection of the features.

This method checks the performance of the model and then iteratively removes the worst performing features one by one till the overall performance of the model comes in acceptable range.
The performance metric used here to evaluate feature performance is p-value. If the p-value is above 0.05 then we remove the feature, else we keep it. This approach give the final set of variables which are: *['balance bank', 'has housing loan?', 'has personal loan?', 'contact communication', 'contact duration', 'number of days that passed', 'number of contacts performed', 'outcome of the previous marketing']*.

At this point, we can start using the various classifiers. The first that was used is Naïve Bayes, which are simple and not computationally expensive. Even if they were to provide non-optimal results, the outcomes would still serve as reference points for building better models through more advanced techniques. In this case BernoulliNB from the scikit-learn library was used. Since there are no parameters to be altered directly, the accuracy only depends on the features. Training is quick and the results are shown below:

Naïve Bayes results

Accuracy Train : 73.84%
Accuracy Test : 71.17%

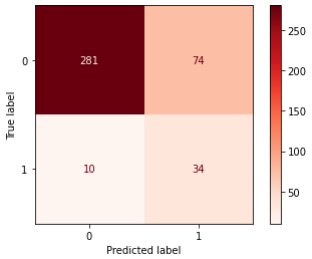


The results are rather low, both of the train set and of the validation set. The second classifier used is the Logistic

Regression, again using the scikit-learn library, first the standard configuration was tested, with the value C (Inverse of regularization strength) equal to 1.0, while the max number of iterations was 100, the results obtained for this classifier are:

Logistic Regression

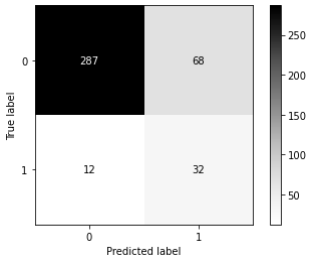
Accuracy Train : 79.44%
Accuracy Test : 78.94%



The results here as expected are better than those of Naïve Bayes. By modifying the various parameters of the Logistic Regression, no major improvements are obtained compared to the standard model.
The last classifier used is SVM (Support Vector Machine) , both in its linear and radial version, it was tested with various gamma and C configurations, these are the results obtained from the best combination of parameters:

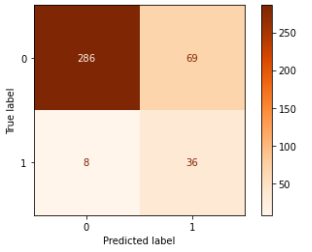
SVM (Linear Kernel)

Accuracy train is: 78.69% C : 0.1
Accuracy test is: 79.94 % C : 0.1



SVM (Radio Kernel)

Accuracy train is: 82.54% C : 10 gamma : 0.1
Accuracy test is: 80.70 % C : 10 gamma : 0.1



The radio kernel, as expected, is the one that shows better performance than other models, and was chosen as the best model, at this point using the test data an accuracy of *75.6%* was obtained.

V. CONCLUSIONS

The results obtained are not very satisfactory, the proposed model can be optimized, perhaps using more data or applying other more complex learning algorithms.

Run : ML_project.py

I AFFIRM THAT THIS REPORT IS THE RESULT OF MY OWN WORK AND THAT I DID NOT SHARE ANY PART OF IT WITH ANYONE ELSE EXCEPT THE TEACHER.