Marketing Campaigns Classification

Guanhua Zhu

Brown University

https://github.com/gzhu7/bank-marketing-campains-classification

Introduction

In this project, I want to predict whether the clients of a Portuguese bank will subscribe a term deposit at the bank. The data includes bank client data, social-economic context attribute, and the information of marketing campaigns.
The goal is to use classification methods to predict if a client will subscribe a term deposit based on the attribute information from 20 features. Supervised ML methods will be applied and different methods are to be compared.
For a commercial bank, one important job is to attract more deposits. Certain group of customers usually contribute much more than others to the total subscription. One major task for banks is to find such target clients. Also, the banks need to pay attention to the patterns shown in previous campaigns and the potential effects of macroeconomic factors, since many of these factors are directly related to the probability of success of a campaign, and may be valuable for future campaigns.
The dataset is from Kaggle and is initially posted on the paper *A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems (2014)* by Sergio Moro. The dataset includes 41188 rows of data and 21 columns. There are 10 numeric feature variables and 10 categorical feature variables. The variables are listed below(detailed descriptions see link above):

age, ***job, marital***, ***education***, ***default, housing***, ***loan***, ***contact***, ***month***, ***day_of_week***, ***duration***, ***campaign, pdays***, ***previous***, ***poutcome***, ***emp.var.rate***, ***cons.price.idx***, ***cons.conf.idx***, ***euribor3m*** , ***nr.employed***, ***y***

On Kaggle, some models like KNN, SVM are performed for the prediction. I want to use different set of features and test on different classification methods to see if a higher performance can be achieved.

EDA

Below is the bar graph(Figure1) of dependent variable. The balance is 0.88 so the dataset is highly imbalanced.

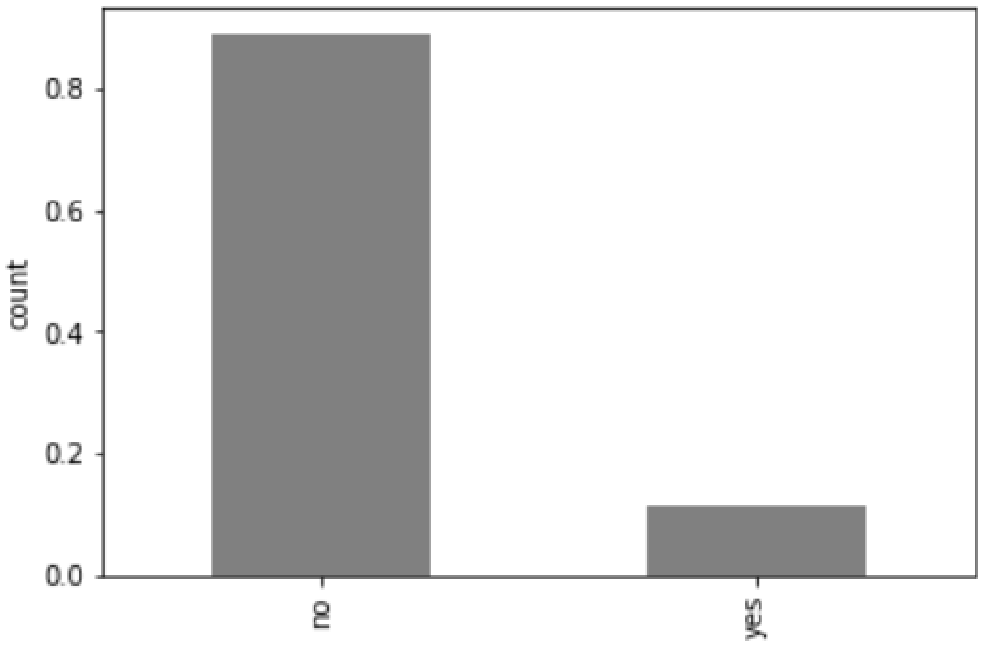


Figure 1: y

The following graph(Figure2) shows the amount of calls made each month, both successful contacts and failure contacts. The distribution shows a clear seasonality. There are more calls been made in spring and early summer, but less in winter. We can also see that in the period of more frequent calls,

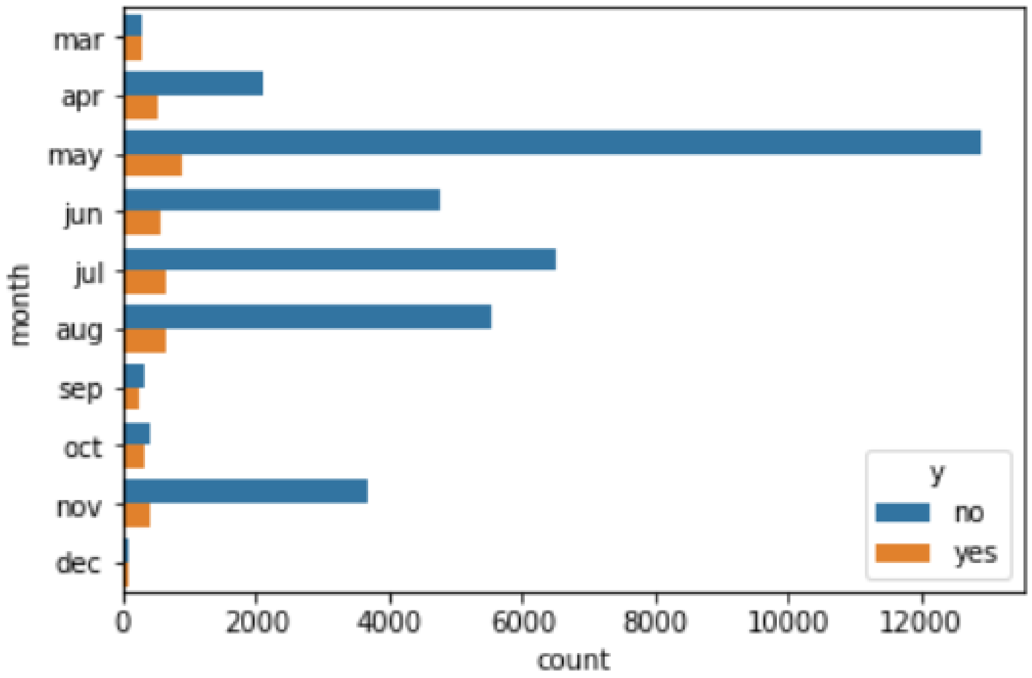


Figure 2: contacts counts for months

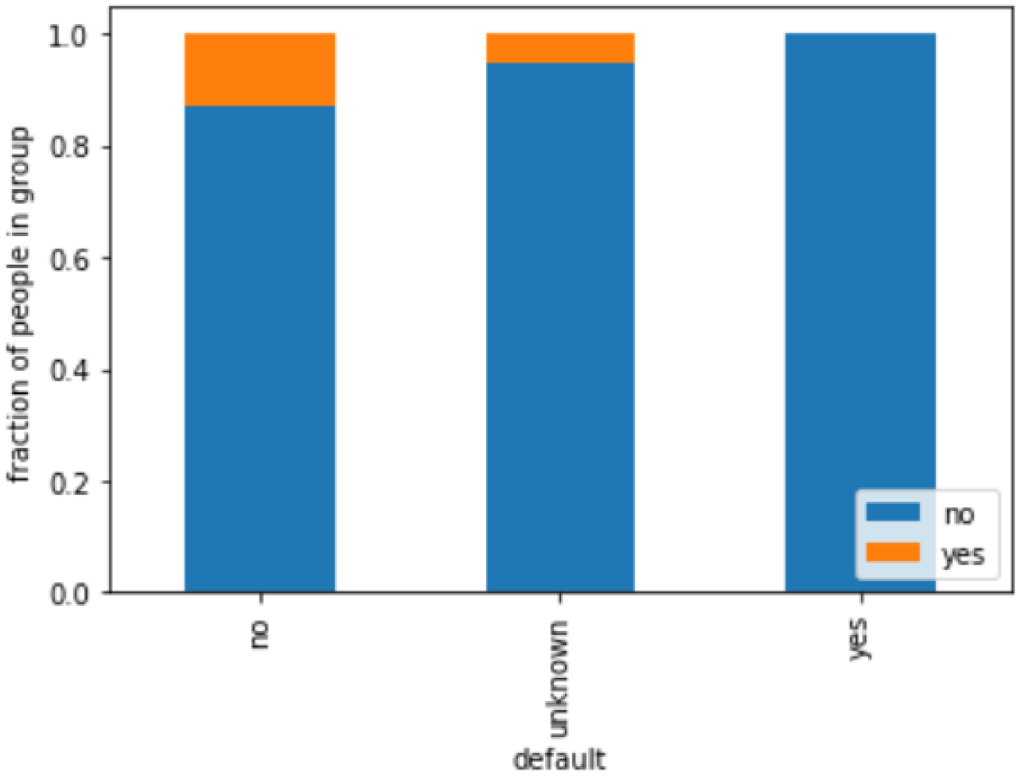


Figure 3:default counts

From the figures above(Figure3) we can see that the client with defaults are not likely to subscribe to the term deposits.
The figure below(Figure4) shows the distribution of contact duration, it follows a right skewed distribution. We can also infer from the interaction plots(Figure5) that this variable may contribute a lot in the predictions.

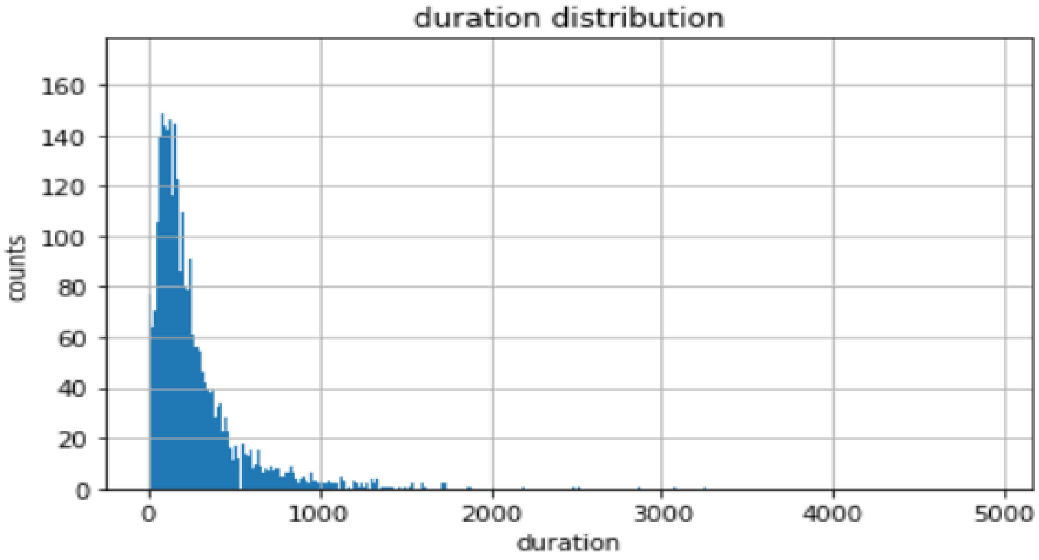


Figure 4:duration distribution

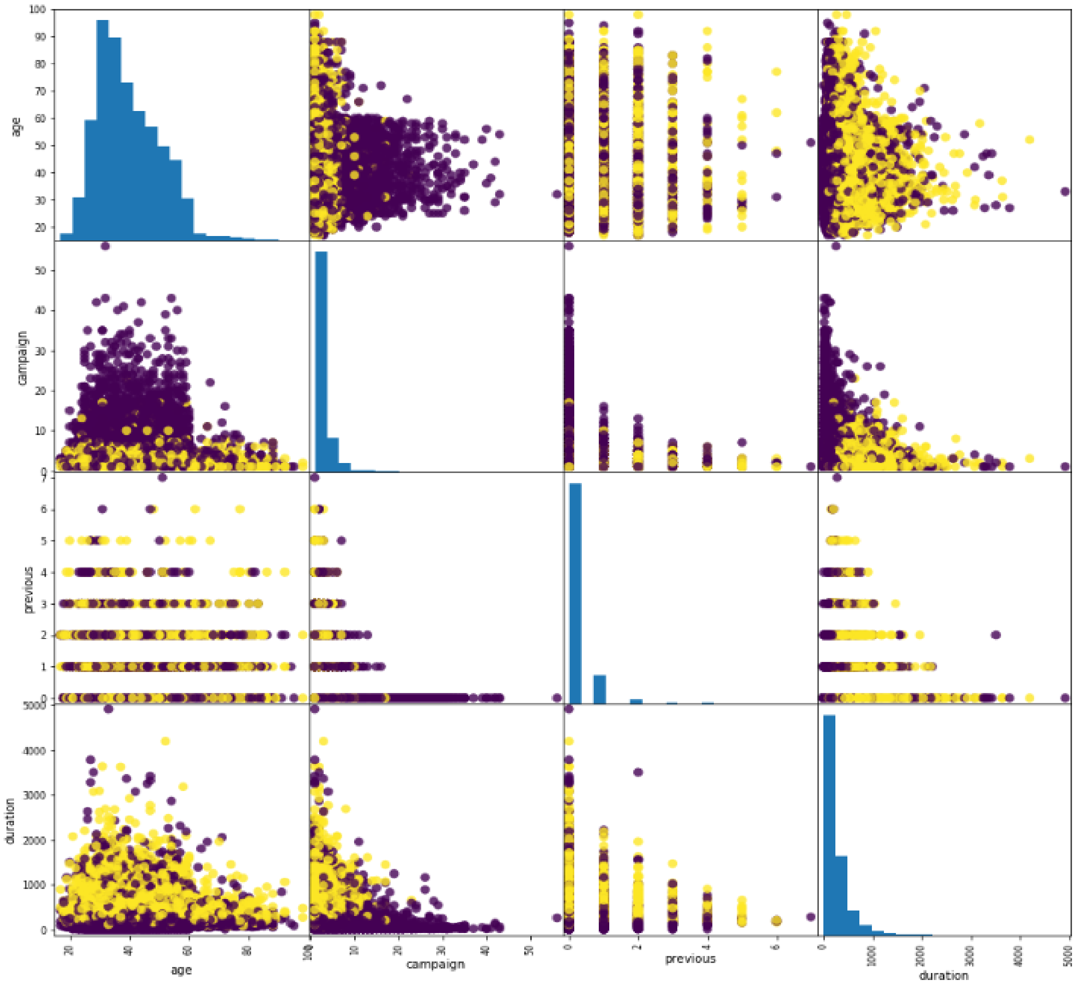


Figure 5:interactions plot

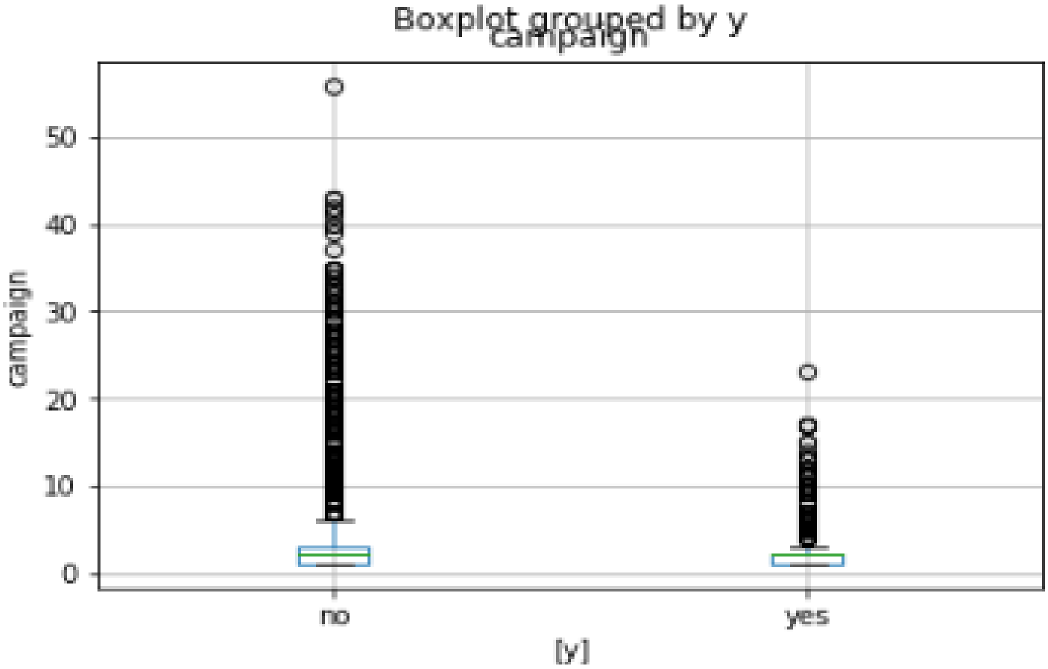


Figure 6: campaign vs y boxplot

The relationship between the number of contacts performed during this campaign is showed above(Figure6). We can see that if there’s already 18 contacts, additional contacts may not be useful to increase the successful rate of current campaign. The more is not always the better.

From the correlation matrix(Figure7) below, it’s clear that the variable euribor3m is highly correlated with two other variables: emp.var.rate and nr.employed and the correlation coefficients are greater than 0.9. So they contain a large amount of mutual information. So I remove the two variables: emp.var.rate and nr.employed, only leaves the euribor3m in our features data frame.

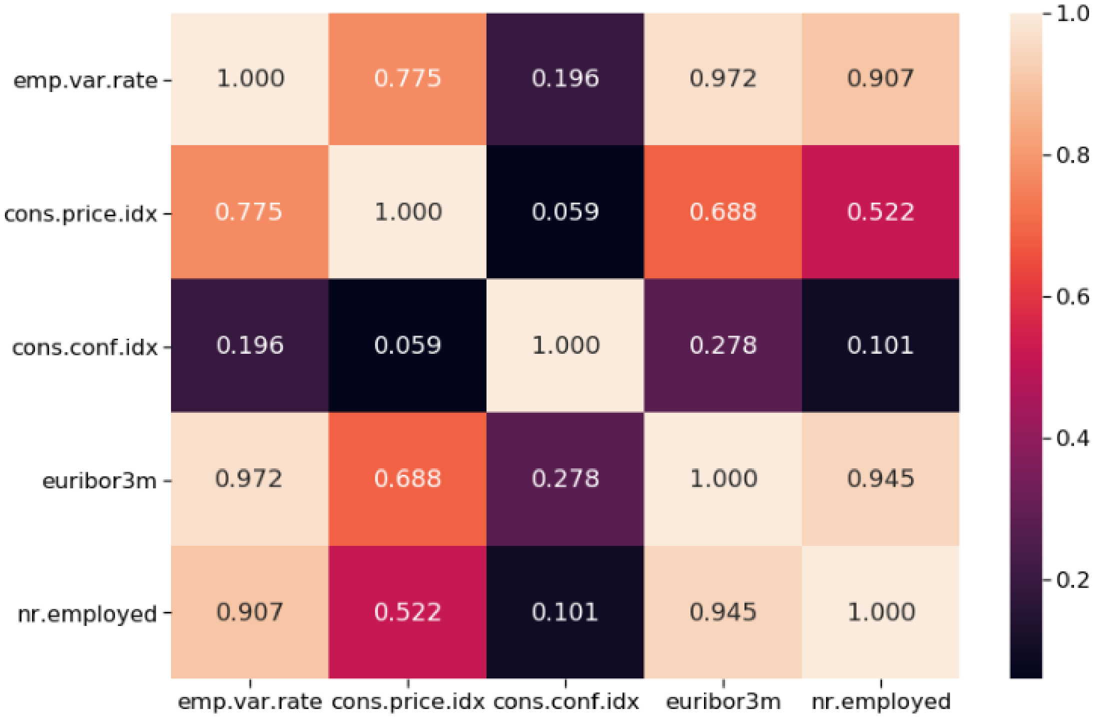


Figure 7:correlation matrix(social-economic var)

Methods

In the features preprocessing, the numeric features(age, campaign, previous, pdayscon, duration, cons.price.idx, consconf.idx, euribor3m) are scaled through standard scaler since they are in a tailed distribution. The categorical features(job, marital, education, default, housing, loan, contact, poutcome, month, day_of_week) are all transformed using one-hot encoding. There is no missing value in this dataset. For the variable pdays which indicating the number of days that passed by after the client was last contacted from a previous campaign, if the client was not previously contacted, pday = 999.

The binary variable indicating whether a client was previously contacted(pdayscon) was used instead.

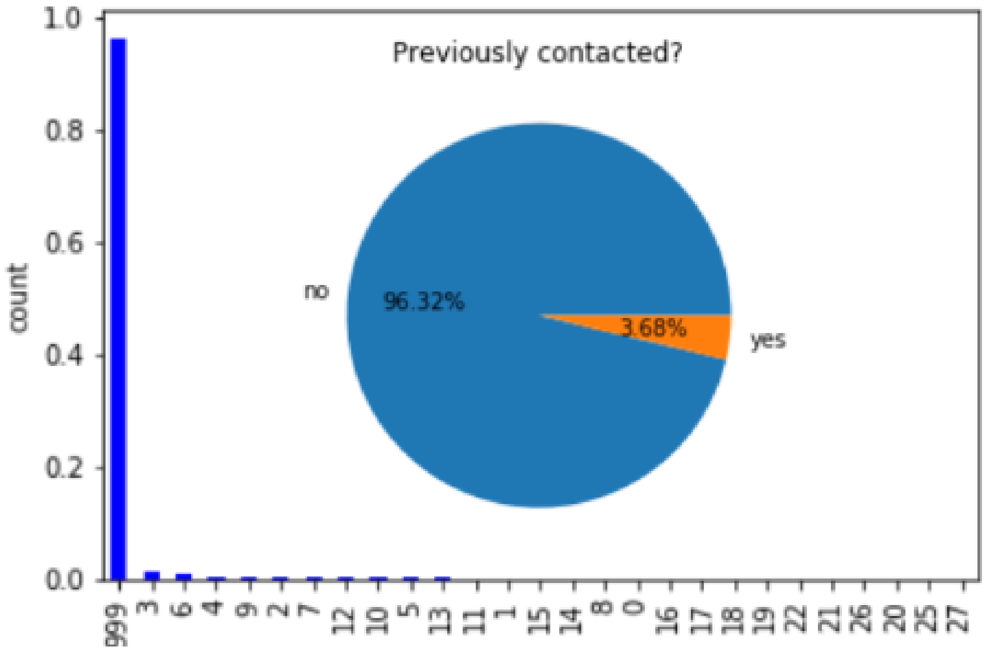


Figure 8:pdays->pdayscon

Also, the three variables that are highly correlated with euribor3m is removed. As a result, there are 61 features in the preprocessed feature dataset after the above manipulations and feature selections.

The dataset is imbalanced. Therefore, I applied stratified splits to the imbalanced data in the K fold validation. 4 classification models are tried and the parameters are tuned through Grid Search. All models are supervised models and are using the sklearn packages to build the model pipelines. And 5 random state are used.

Algorithm	Hyper-parameters	Range
Random Forest Classification	n_estimators, max_depth	[1,100], [5,11]
Logistic Regression	C, penalty	[0.1, 10], [l1&l2]
Gradient Boosting Classification	learning_rate, max_depth, n_estimators	[5,20], [2,4], [0.05,0.2]
Support Vector Classification	C, gamma	[1&10], [0.1&1]

Table 1:ML methods&parameters

First, the random forest classifier is deployed. In this model, two parameters are tuned in the grid: n_estimators and max_depth. The two parameters have big influences on the performance. There are four n_estimators from 1 to 100 and four max_depth(5,7,9,11) in the grid. Two metrics are used to evaluate the models' performance respectively, the accuracy and the recall score. The accuracy serves as a general evaluation which indicates the fraction of predictions got right in the model. The recall score is used to better compare the model with the baseline. It will be discussed in the results part. Since it's a non-deterministic ML methods, I performed the methods for 5 times and calculate the mean and the standard deviation of the best test score for the 5 rounds. And the global feature importance is also calculated and plotted.

The second model I deployed is the logistic regression. In this model, two parameters are tuned in the grid: the C value and the penalty term. The C values ranging from 0.1 to 10 are tried. The performance of two types of penalty L1 and L2 will be compared. The method is also run for 5 times and I calculate the mean and the standard deviation of the results.

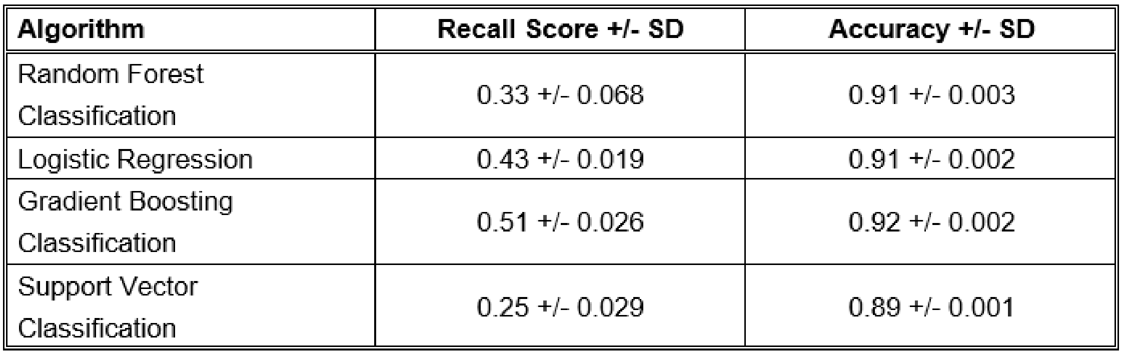
The third model is the Gradient-boosting method. Three parameters are tuned in the gird for this model: learning_rate, max_depth and n_estimators. The values of n_estimators range from 5 to 20, the values of max_depth range from 2 to 4, and the values of learning_rate range from 0.05 to 0.2. The accuracy and the recall scores are used to evaluate the model performance. The method is also run for 5 times and I calculate the mean and the standard deviation of the results. And the global feature importance is also calculated and plotted using the build-in function.

The last model is the support vector classification. Two parameters are tuned in the grid: the gamma and the C value. Since SVC takes a long time to train if there are many values in the grid, it will take long to go through a complete tuning, only 2 values of gamma(0.1,1) and 2 values of C(1,10) are tried in the grid for this model and the method is running for 3 times to calculate the mean and the standard deviation. The accuracy and the recall score both used to evaluate the model performance respectively.

In the pipeline, all models are using stratified K fold methods due to the imbalance. And random state of the random forest model is fixed in the pipeline function. For the random forest classification and the Gradient-boosting method, the one-hot features are also scaled by applying standard scaler in the preprocessing in order to prepare scaled coefficients to interpret feature importance.

Results

In the baseline model, we predict the most frequent(popular) class(0) to each data point. As a result, the TN is 36548, FN is 4640, and both FP and TP are 0. So the baseline recall score is 0. And the accuracy will be 0.8873. If the recall score of our model is greater than zero, we can say that the model performs better than the baseline.



Algorithm	Recall Score +/- SD	Accuracy +/- SD
Random Forest Classification	0.33 +/- 0.068	0.91 +/- 0.003
Logistic Regression	0.43 +/- 0.019	0.91 +/- 0.002
Gradient Boosting Classification	0.51 +/- 0.026	0.92 +/- 0.002
Support Vector Classification	0.25 +/- 0.029	0.89 +/- 0.001

Table 2:Results

The results show that the test accuracy for the random forest classification is 0.33 with a standard deviation of 0.0678. So it's better than the baseline model. The accuracy is 0.91 with a standard deviation of 0.0025 which also indicates that it performs better than baseline. The best parameters found for each training varies for many rounds of training: (max_depth: 9, n_estimators:1), (max_depth:11, n_estimators:4), (max_depth: 11, n_estimators:1), (max_depth: 11, n_estimators:1), (max_depth: 9, n_estimators:1). The most common parameters group is (max_depth:11&9, n_estimators:1).

For the logistic regression model, the recall score is 0.43 with a standard deviation of 0.0189. So it's better than the baseline model. The accuracy is 0.91 with a standard deviation of 0.002 which also indicates that it performs better than baseline. With L1 penalty, the model can shrink the less important feature's coefficient to zero, thus it can be used for feature selection if there is a huge number of features. While the model with L2 penalty performs better most of the time in the trainings. The best C value found in the grid is 1.

For the Gradient-Boosting classifier, the recall score is 0.51 with a standard deviation of 0.0257. So it's better than the baseline model. The accuracy is 0.92 with a standard deviation of 0.002 which also indicates that it performs better than baseline. The best parameter group is(learning_rate:0.2, max_depth:4, n_estimators:20).

For the SVC, the recall score is 0.25 with a standard deviation of 0.02. So it's better than the baseline model. The accuracy is 0.89 with a standard deviation of 0.001 which also indicates that it performs better than baseline. But the improvement is the worst among all models. It might be due to the fact that only a small portion of parameters are tuned. The best parameter group is(C:10, gamma:0.1).

Below is the graph(Figure9) of global feature importance for the random forest model:

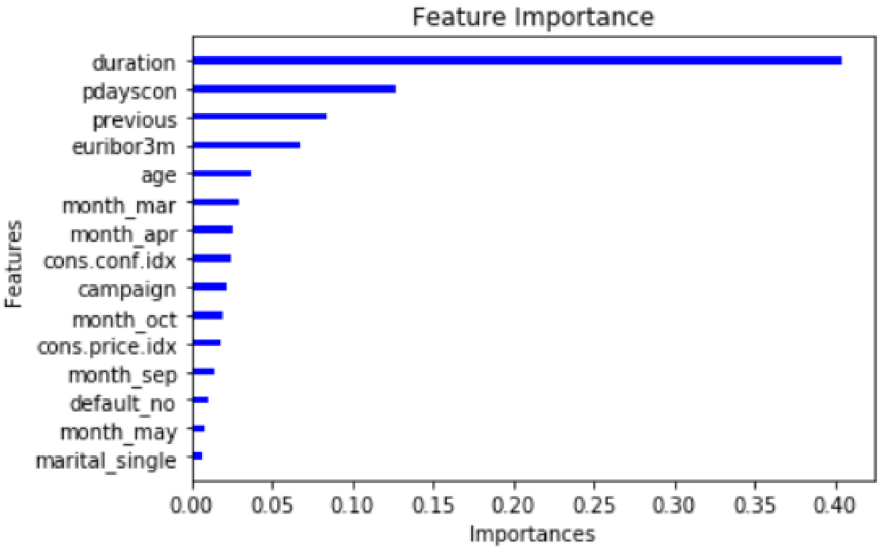


Figure 9:Feature importance for random forest

We can see that the *duration* is the most important factor in the classification, then comes *pdayscon*, *previous* and *eurbor3m*. It means that the duration of the last contacts is highly relevant to the results of the marketing campaign. It makes sense because if a client is attracted and is going to subscribe the term deposit, he may ask more questions so the call tends to be longer. If the client shows no interest, the call is likely to be hung up at the beginning. Similarly, if a client is not previous contacted(*pdayscon*) which means he may hear about the product for the first time, or he has not received enough calls in the previous campaign(*previous*),he is not likely to subscribe, which encourages the bank to have more campaigns to let target clients learn about their products.

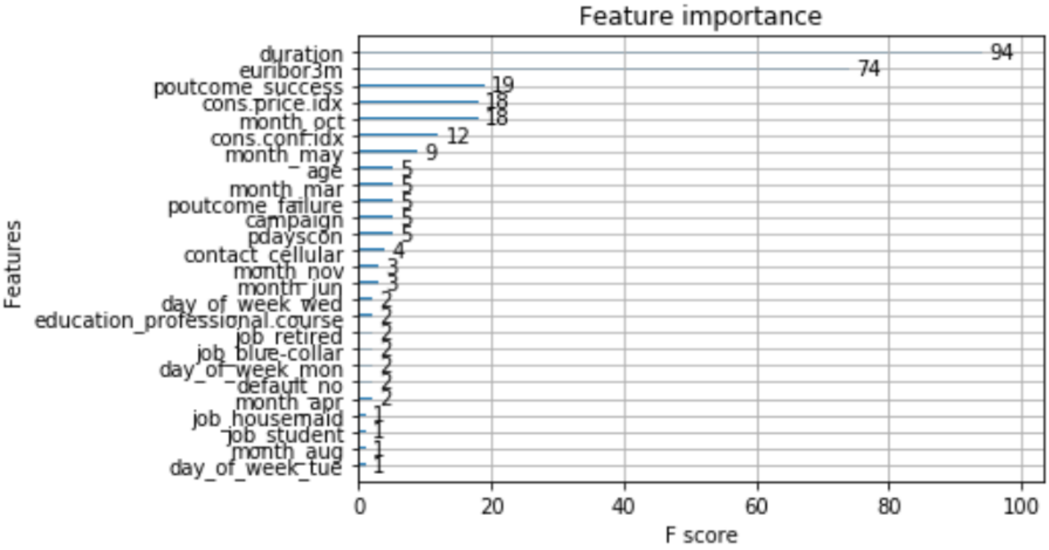


Figure 10:Feature importance for Gradient Boosting

The graph of feature importance of the Gradient Boosting model is a little different. But for the top lists of important factors, they have many variables in common. And the variable *duration* keeps being the most important factor. The variable poutcome_success is in the top 3, which is in agreement with our previous EDA findings. It means if a client made a subscription in one previous campaign, he is more likely to subscribe in the future.

From the two plots of feature importance, we can see some common macroeconomic factors that play important roles. Generally, when there's high Euro-interbank offered rate, it means this is a period of monetary tightening and there's not enough money deposit in banks, which means the willingness of subscription might be low. When consumer price index is high, it indicates that there's inflation and the purchasing power of money goes down. So the clients may find other investment substitutions in order to preserve values. These are all practical significance of the results.

Outlook

To improve the models further, we can tune more hyperparameters and try more values in the grid search, especially for SVC which I only tested on small amount of values but still takes a whole day to train.

If more data or features can be collected and provided, the models are expected to have better performance. Such features can relate to information like: clients' home location, the average income of households of the clients' living areas, and funds, stocks and other investment substitutions in hand, etc. Also, since there 61 features to feed in, we can try other methods like neural networks to get better performance if the computation power permits.

References

Moro, Sérgio & Cortez, Paulo & Rita, Paulo. (2014). A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems. 62. 10.1016/j.dss.2014.03.001.

https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

https://www.kaggle.com/mpalaghian/bank-marketing-imbalanced-classes-resampling

https://www.kaggle.com/fengdanye/machine-learning-4-support-vector-machine