Marketing Campaign in a Portuguese Bank

Facundo Sigal

6/10/2020

Introduction

In this work, I analyse the data from a marketing campaign of a Portuguese Bank. The marketing campaign was based on phone calls and sometimes, it required more than one call in order to get the potential client subscribed. The database includes 45211 observation of 17 variables. The output is a binary variable representing if the candidate had a subscription ("yes") or not ("no"). In order to predict the result of a single campaign, I'll train a Machine Learning algorithm based on some variables in this database. Some of them are data about the client itself, as age, marital status, job, educational level or credit history. Other variables represents the moment of the final contact between the bank and the client, as month, day o contacts. The main objective of this work is to find an algorithm that predicts if a potential client will become an account subscriber or not, using the information contained in the mentioned database. In order to achieve this goal, I'm training different machine learning techniques: Logistic Regression, Linear and Quadratic Discriminant Analysis, K Nearest Neighbors and Random Forest. An ensemble model that combines these three techniques will be adjusted also. The database will be divided into two separate sets: the train set which includes the 80% of the database, and the test set, that contains the remaining proportion. I've chosen this proportions **80/20** because in this case, I have a large sample, but not as large as the one I used in the Movielens project. This is a large set (almost 50k observations), so I sample a minor set to test the trained models.

Methods and Analysis

The first step is getting the data I need. The database was downloaded directly from UCI Machine Learning Repository with this code:

```
fs <- httr::GET("https://archive.ics.uci.edu/ml/machine-learning-databases/00222/bank.zip")
temp <- tempfile()
download.file(fs$url,temp)
```

After that, I must import the downloaded file with this `read.csv` code:

```
bancos <- read.csv(unz(temp, "bank-full.csv"), sep = ";")
```

Then, some exploratory analysis have to be performed. The global rate of subscription is

```
0.1169848
```

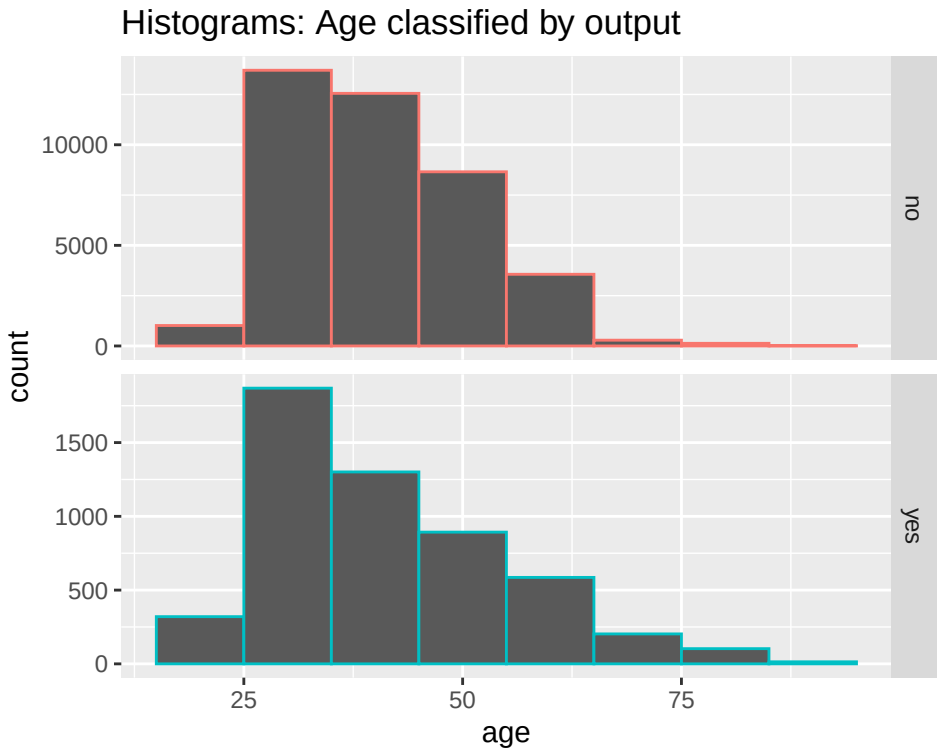
I should be careful with the measures used to evaluate the different models because the subscription rate is low (about 12%). This small prevalence produces "Accuracy" values that don't represent the real effectiveness of each model.

Now let's take a look to some of the explanatory variables I've chosen and their relation with the output variable.

Age Next table contains summary measures of the age of the candidates:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.00   33.00   39.00   40.94   48.00   95.00
```

The youngest candidate is 18 years old and the oldest is 95. The average age is almost 41 and the median is 39. The fact that the median is smaller than the mean is reasonable because the distribution of the variable *age* is positively asymmetric.



As seen in this graph, the distribution of age is similar in both subscribers and non-subscribers group. The average age between subscribers is

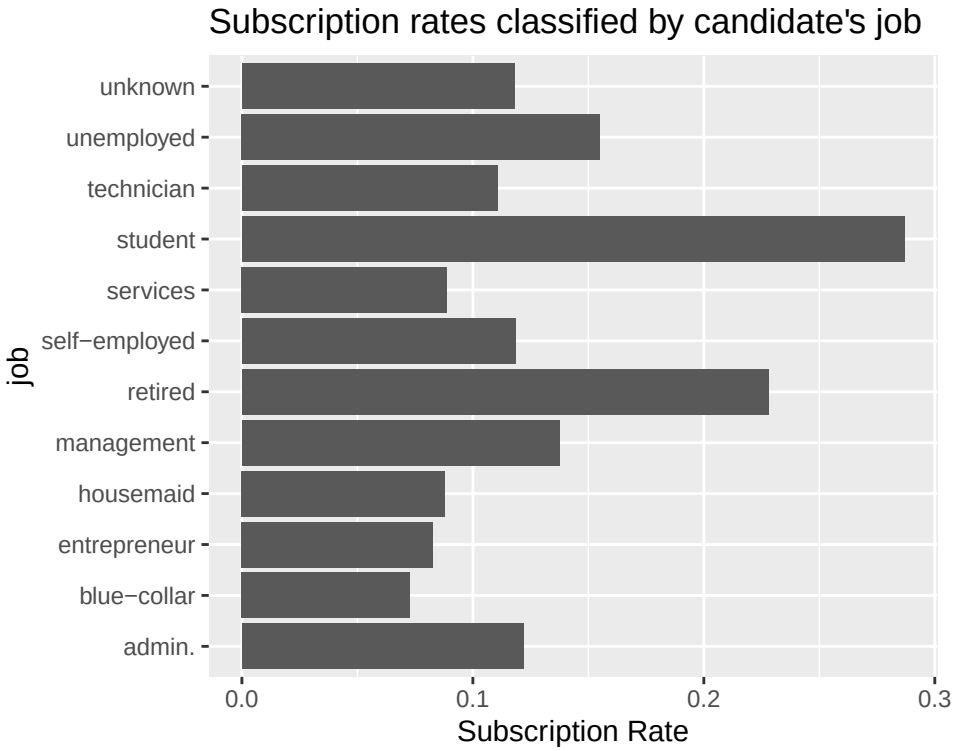
```
41.67
```

The average age between non-subscribers is

```
40.84
```

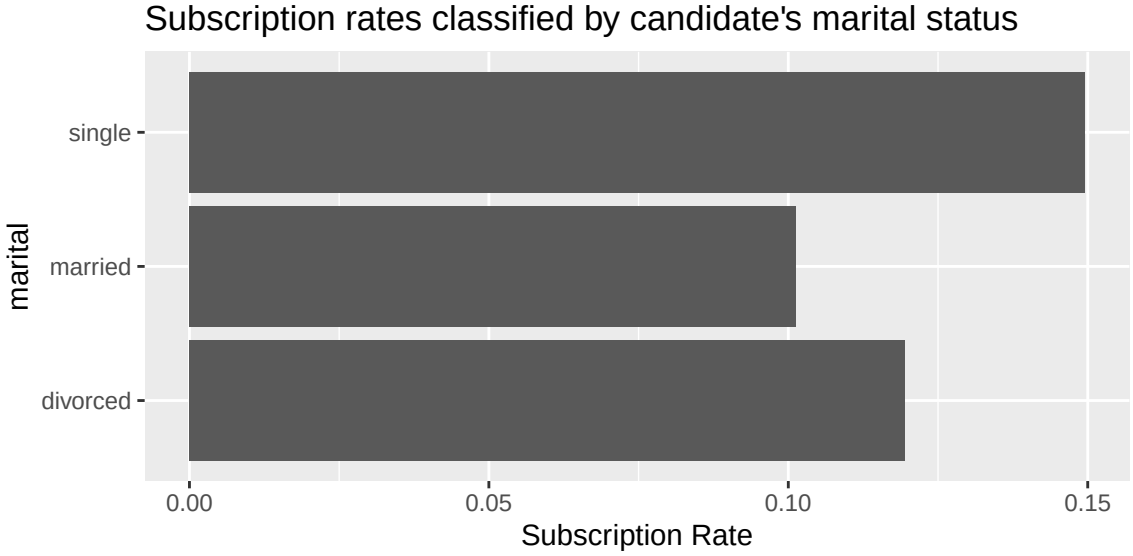
The difference between the average age of subscribers and non-subscribers is less than a year.

Job In the following graph, the subscription rate is classified by the candidate's job.



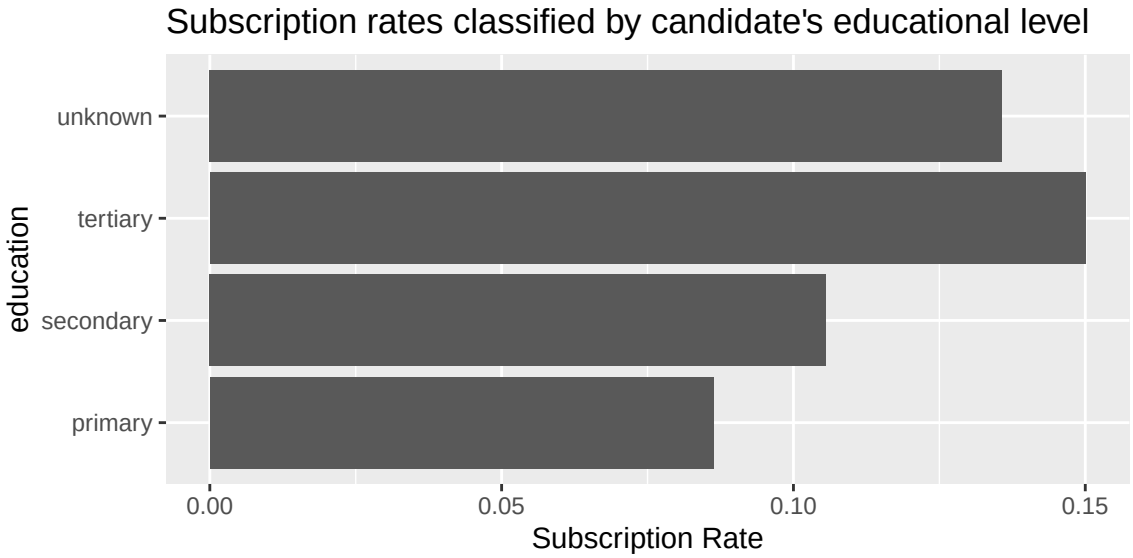
Students have the largest subscription rate. More than 25% of the contacted students got subscribed. Also between retired people, the marketing campaign was successful, with more than 20% of acceptance. The lower subscriptions rate correspond to people who work in Services, Housemaids, Entrepreneurs and Blue-Collar jobs, with less than 10% of acceptance.

Marital Status In the next graph, there is a comparison of subscription rates between single, married and divorced people. This last category includes divorced and widowed candidates.



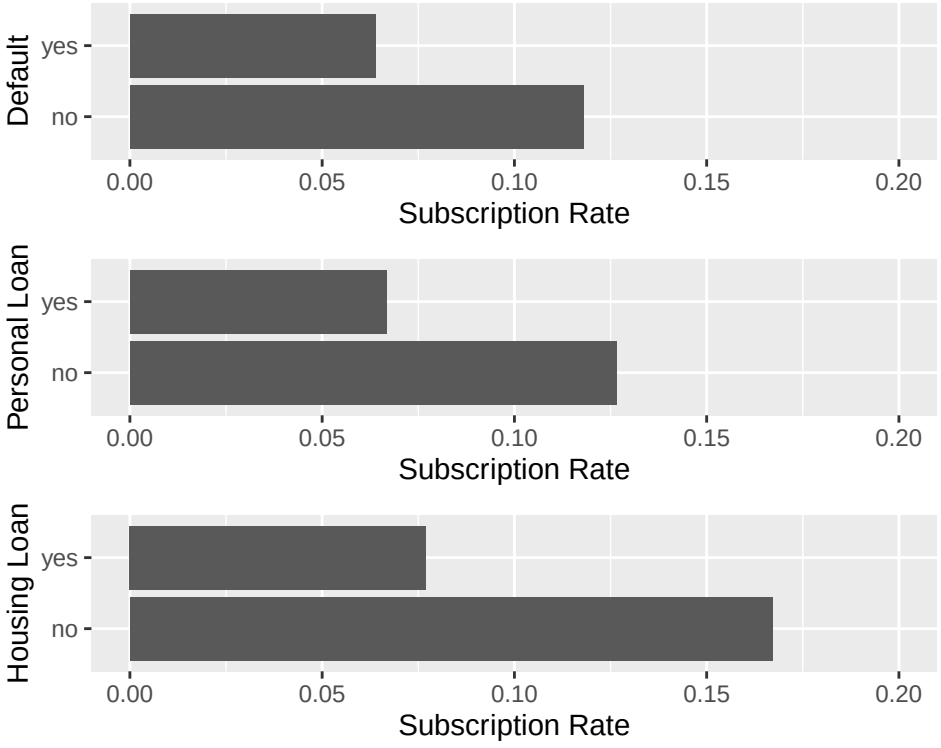
Almost 15% of single people got their subscription and about 10% of married people did. The acceptance between divorced/widowed people is similar to global acceptance rate.

Education In the next figure, the subscription rate is classified by the candidate's maximum educational level reached.



As I expected, the rate of subscription increases for higher levels of education. The rate between people with at least tertiary education is 15%. For secondary and primary is about 10% a 8% respectively.

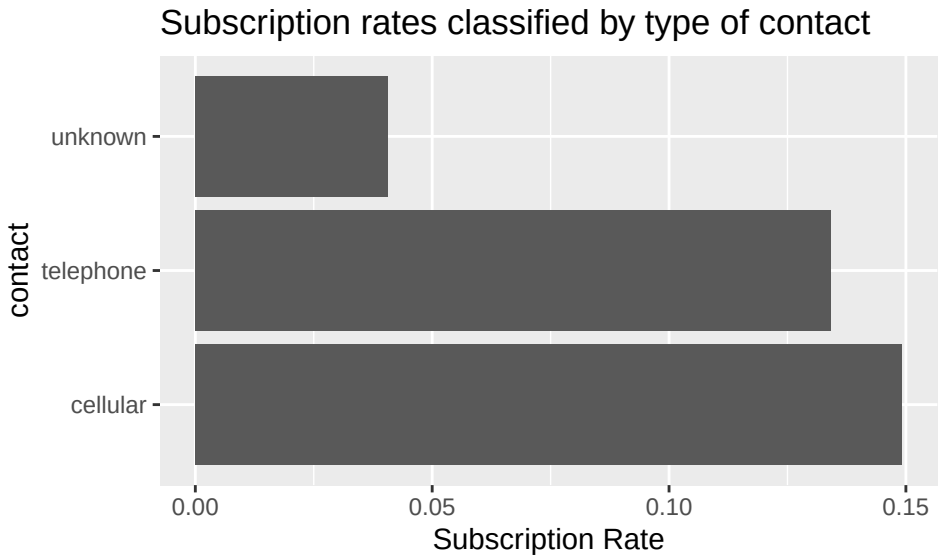
Credit History Now it's time to analyze the credit records of each candidate. Here we can see the subscription rate according if they had defaults before, and if they had previous personal or housing loans.



Only 6% of the candidates that had previous defaults got an account subscription. People with previous loans rarely got their subscription. These rates were about 7% and 8% for previous personal and housing

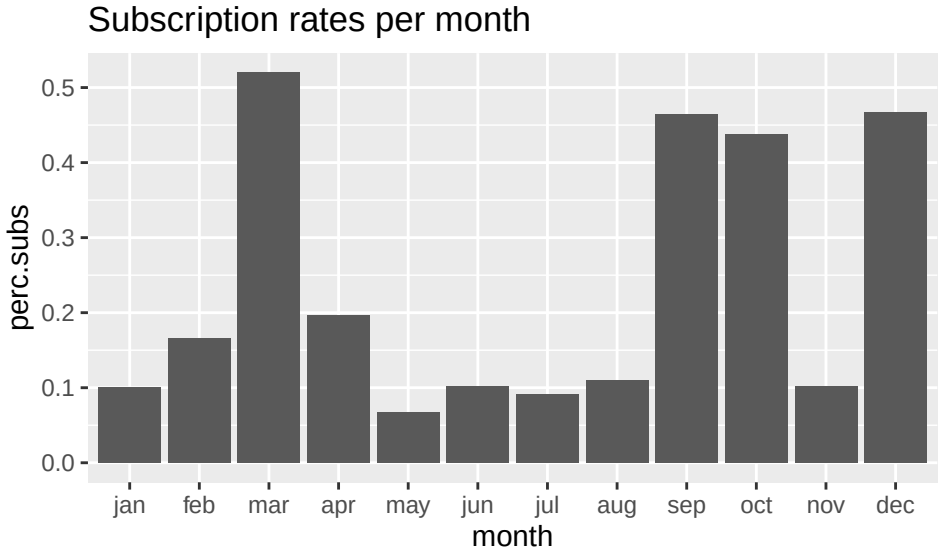
loans respectively.

Type of contact In the following figure, there is a comparison of subscription rates between the type of last contact. Telephone means fixed-line phone.



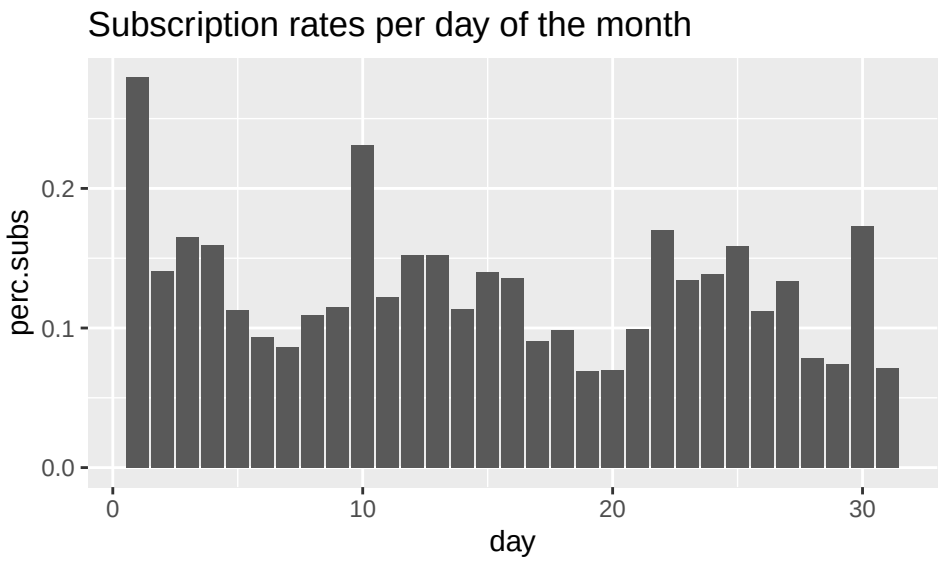
We can't see major differences in subscription rates between fixed-line and mobile phone. A very low subscription is observed in clients with no registration of the type of contact they had.

Month of last contact In this graph, the subscription rate is classified by the month when the subscription was confirmed.



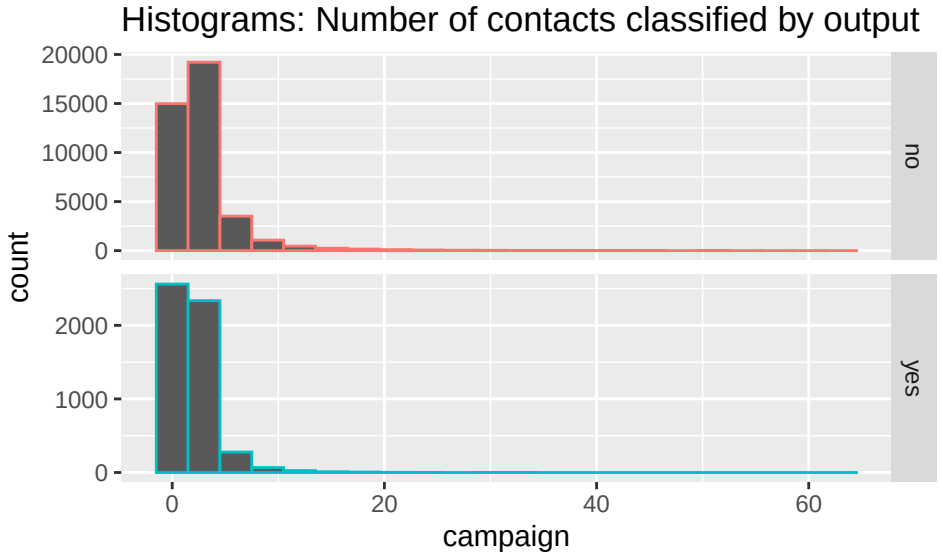
There are four months with very high effectiveness. In march, more than a half of the people reached had their account subscription. In September, October and December the subscription rates were larger than 40%. On the other hand, in May the rate was about 6%.

Day of the Month of last contact Now it's time to check if there is any time of the month when the subscription rates are larger. Here we can see the subscription rates classified by the day of the month.



It doesn't seem to be a certain trend. The first and the tenth day of the month have larger subscription rates, over 20%. Most of the days between 22nd and 30th days, the daily rates are larger than the global rate of 12%.

Number of contacts The last explanatory variable describe is the number of contacts performed during the campaign with each client. Here we see two histograms for the number of contacts. The first one corresponds to people who didn't subscribe to the bank and the second belongs to subscribers. Each bar has a width of three contacts.



Subscribers has a bigger proportion of people that received less than four contacts, while non-subscribers has larger proportion for the interval from 4 to 6 contacts.

Machine learning models/algorithms

Once the exploratory analysis is done, it's time to begin with the machine learning techniques. The database was divided into two sets by random sampling. The 80% of the original database is used to train the models - called **train set** - and the remaining set is used to evaluate each model trained - called **test set**.

The models/algorithms to be evaluated are (a) Logistic Regression; (b) Linear Discriminant Analysis; (c) Quadratic Discriminant Analysis; (d) K nearest neighbor and (e) Random Forest. There is a sixth technique that ensembles every model used before. The ensemble model assigns for each candidate the output that is chosen in three or more models from (a) to (e).

The evaluation uses some measures that indicate how accurate are those models. As the output has a low prevalence of "yes", just looking to "Accuracy" isn't enough. We also must evaluate other measure as Sensitivity and Specificity.

Sensitivity measure the ability of the algorithm to detect a "no" and Specificity indicates the proficiency to detect a "yes". As there is more than 88% of "no's", we expect to have much higher Sensitivity than Specificity.

Results

Before we start training models, let's see what happens if we assign "no" to every candidate in the **test set**.

"No" for everybody

```
Sensitivity: 1

Specificity: 0

Accuracy: 0.883
```

The first model to train is logistic regression. This model estimate the probability of each candidate to have a "yes" using regression technique.

In a first estimation, there were two variables that weren't statistically significant: *Default* and *Day of the Month*. Hence, the final logistic regression doesn't include those two.

Logistic Regression

```
Sensitivity: 0.98986

Specificity: 0.08696

Accuracy: 0.88422
```

The next techniques to train are Discriminant Analyses. Firstly with a Linear Model and then with a Quadratic Model.

Linear Discriminant Analysis

```
Sensitivity: 0.97658

Specificity: 0.18336

Accuracy: 0.88378
```

Quadratic Discriminant Analysis

`Sensitivity: 0.93388`

`Specificity: 0.3242`

`Accuracy: 0.86255`

Finally we have the most complex algorithms: K nearest neighbor and Random Forest. These techniques include an optimization of their parameters using the `tuneGrid` option.

K nearest neighbor

`Sensitivity: 0.99198`

`Specificity: 0.07089`

`Accuracy: 0.88422`

Random Forest

`Sensitivity: 0.99123`

`Specificity: 0.09452`

`Accuracy: 0.88632`

After all these training, we finally use the ensemble to combined every technique used to estimate the output for the test set.

Ensemble

`Sensitivity: 0.9866`

`Specificity: 0.12949`

`Accuracy: 0.88632`

Conclusion

This work has the main objective of finding an algorithm capable of predicting if the potential clients finally get their account subscription in the bank. In order to achieve that, several models and algorithms were applied for training the database.

These methods produced different results. There is no model or algorithm that perform well according to every measure used to evaluate them. Thus, the conclusion are presented separately for each measure.

When we consider the capability of the procedures to identify the "no's", all methods have excellent performance. The highest sensibilities are produced by K Nearest Neighbor, Random Forest and Logistic Regression. Their sensibilities are about 99%.

However, these procedures weren't able to identify the "yes" accurately, because this is an event with very small prevalence (under 12%). The best performance was achieved by Quadratic Discriminant Analysis, by far. It produced a specificity of 32%. QDA's specificity almost doubles the specificity of the next model in descending order - Linear Discriminant Analysis has a specificity of 18%.

Accuracy is highly influenced by prevalence. As the prevalence is low, larger values of Accuracy will be closely related to higher sensitivities. The best accuracy is achieved by Ensemble, Random Forest K nearest neighbor and Logistic Regression. Quadratic Discriminant Analysis, which is the model that produced the highest specificity, has the lowest accuracy. QDA was the only model that produced lower accuracy than assigning "no" to every candidate.

The Ensemble of the five methods produced a specificity of almost 13% (nearly a median value) without losing sensibility. The accuracy of the Ensemble method is the highest one, similar to the accuracy achieved by K Nearest Neighbor.

This work presents a strong background for estimating the result of a marketing campaign, just considering some information about the potential client. There is a limitation about the results found in this work. The algorithms trained here aren't capable to identify the successful campaigns with much accuracy. The only model that identifies almost one out of three subscribers is the Quadratic Discriminant Analysis. There is a lot of work to do in the future. Basically, there is a deficiency that must be improved in the detection of the candidates who will end up being subscribers.