Bank Marketing Campaign

Virtual Internship

15-July-2022

Team member’s details

- **Group Name:** The Greek Scientist
- **Name:** Nikolaos Sintoris
- **Email:** nikolaossintoris@gmail.com
- **Country:** Greece
- **College/Company:** Data Glacier
- **Specialization:** Data Science
- **GitHub Repo Link:** https://github.com/NikolaosSintoris/Data-Glacier-Virtual-Internship/tree/main/Week%2013

Problem Description

ABC Bank wants to sell its term deposit product to customers and before launching the product they want to develop a model which help them in understanding whether a particular customer will buy their product or not (based on customer's past interaction with bank or other Financial Institution).

Data Understanding

- **Data Information:** The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be (or not) subscribed.
- **Summary of the data:** The bank-additional-ful.csv file, consists of 41188 observations (rows) and 21 attributes (columns). The size of the file is 6.6 MB.
- **Attributes:** Age, job, marital, education, default, housing, loan, contact, month, day of week, duration, campaign, pdays, previous, poutcome, emp.var.state, cons.price.idx, cons.conf.idx, euribor3m, nr.employed as well as the output variable y (whether the client subscribed a term deposit).

Missing Values

In our data, missing values is the value "unknown" in the categorical attributes, so job has 330 missing values, marital has 1731, default has 8597, housing has 990 and loan has 990.as they are shown in the screenshot below:

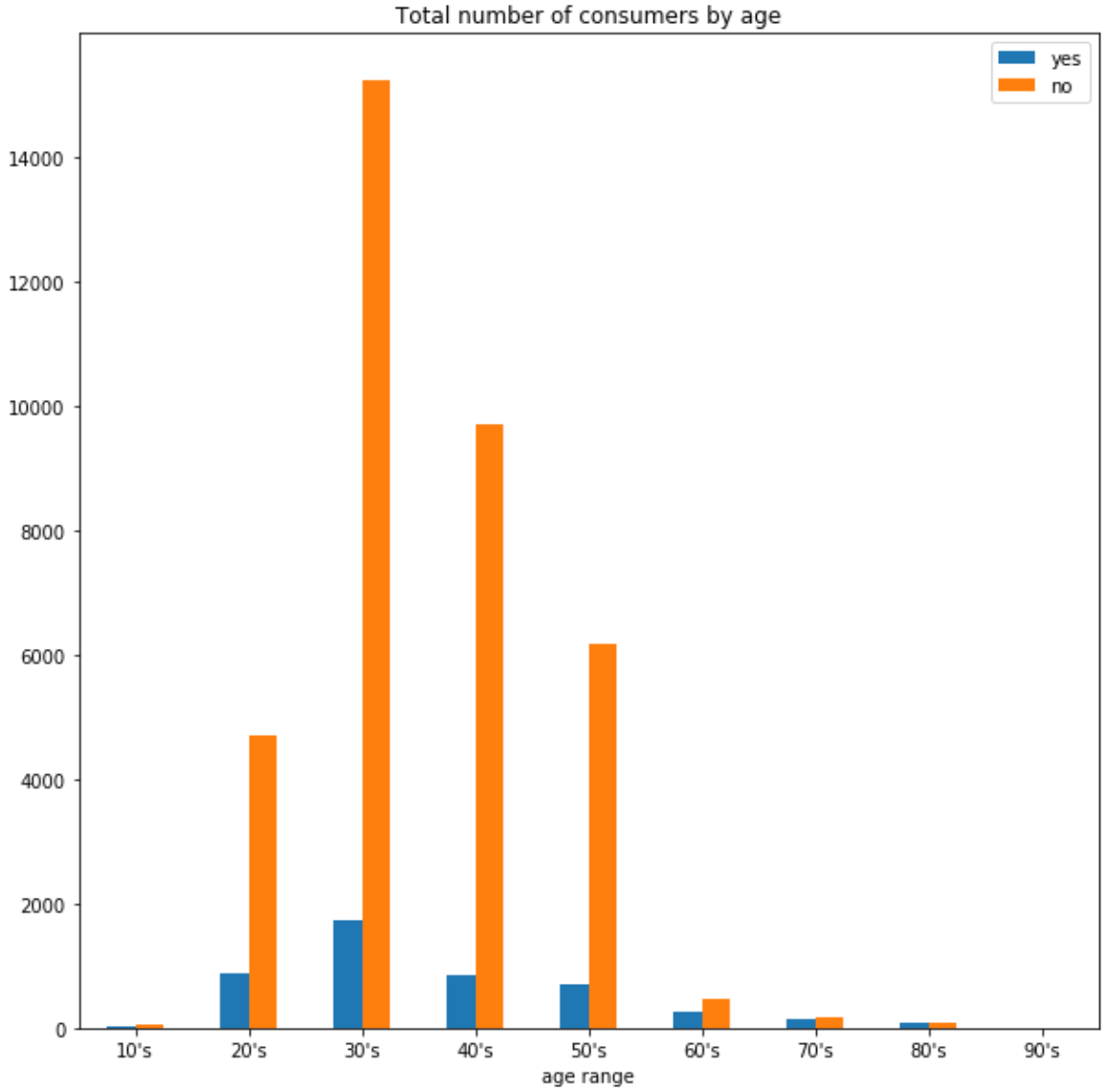
	Column name	# of missing values
0	age	0
1	job	330
2	marital	80
3	education	1731
4	default	8597
5	housing	990
6	loan	990
7	contact	0
8	month	0
9	day_of_week	0
10	duration	0
11	campaign	0
12	pdays	0
13	previous	0
14	poutcome	0
15	emp.var.rate	0
16	cons.price.idx	0
17	cons.conf.idx	0
18	euribor3m	0
19	nr.employed	0
20	y	0

For the missing values we are going to use 2 techniques. When the most frequent word has by far more data points we, are going to replace the missing values with the most frequent one. Otherwise we are going to replace the missing values with a random value. So, for the job, education and housing attributes we are going to use the second technique and the first technique for the other attributes.

Exploratory Data Analysis

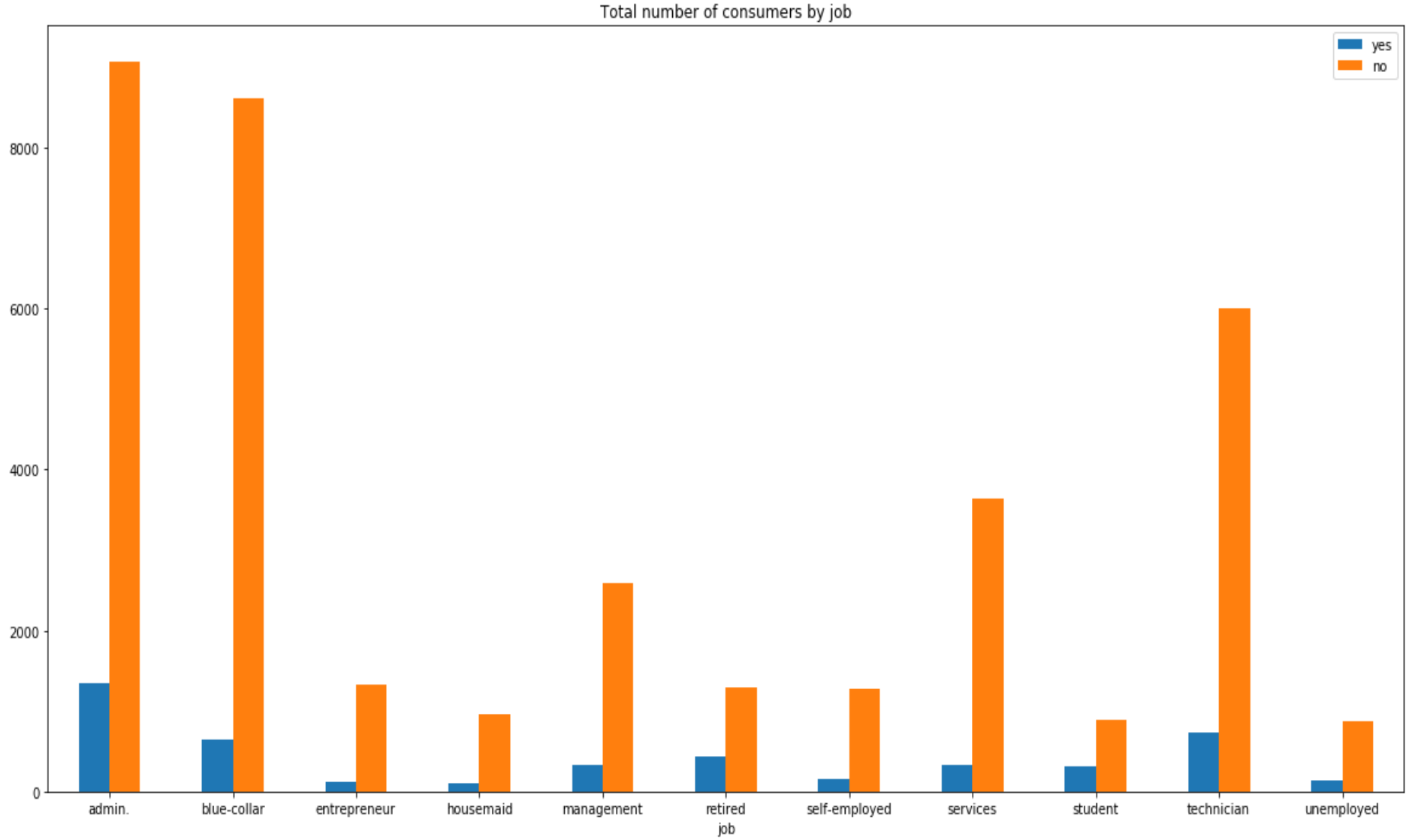
I performed Exploratory Data Analysis on the data. First I examined the pearson correlation coefficient between the attributes and the result was that there was no correlation between them. So I examined the relationship between the output variable (“y”) and some of the rest attributes.

Total Number of Consumers by Age



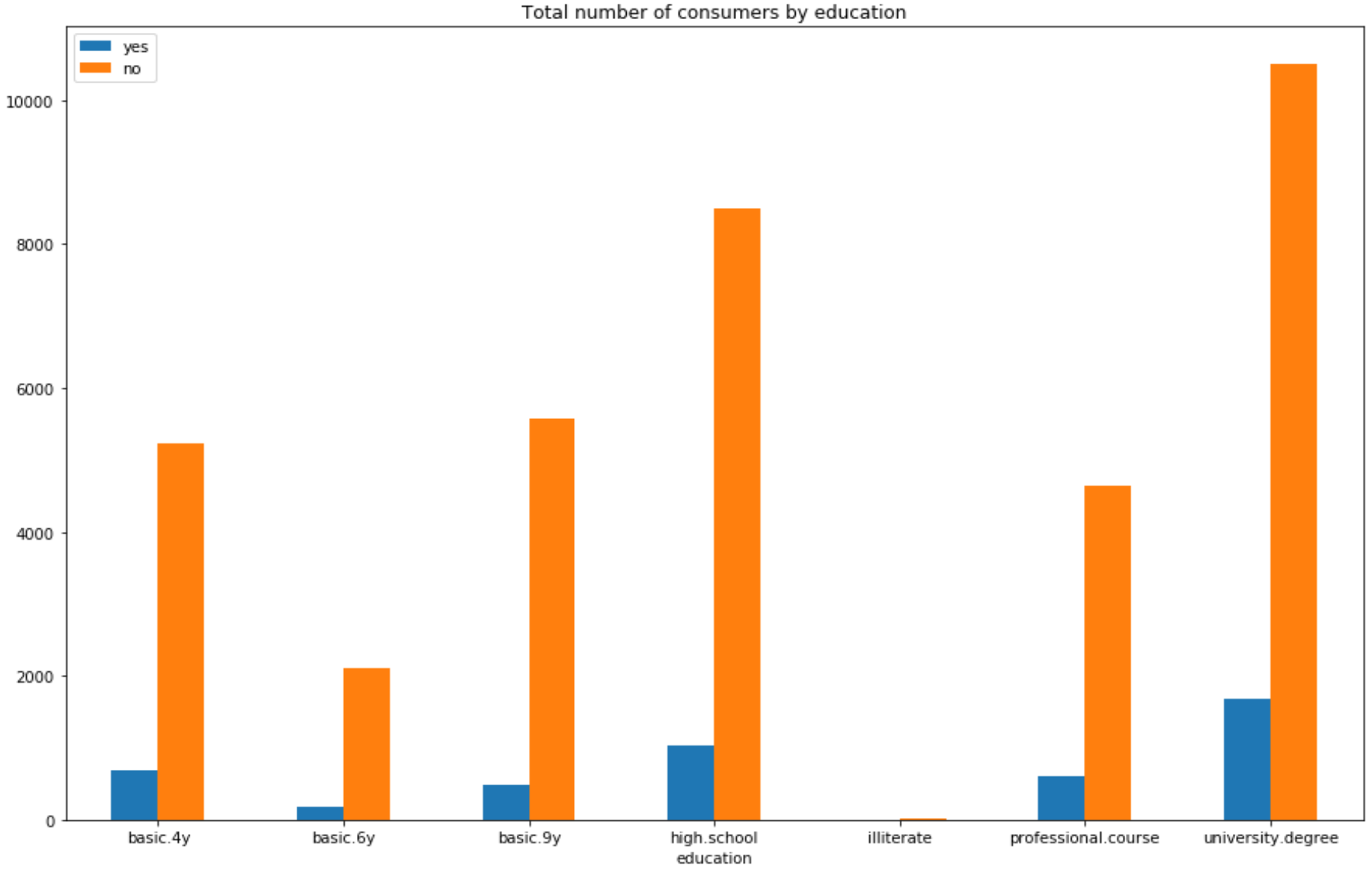
- In the graph we can see that for the ages 20 to 60, there is a significant difference between those who applied and those who did not.
- We also observe that at every age, those who did not apply far outnumbered those who did.

Total Number of Consumers by Job



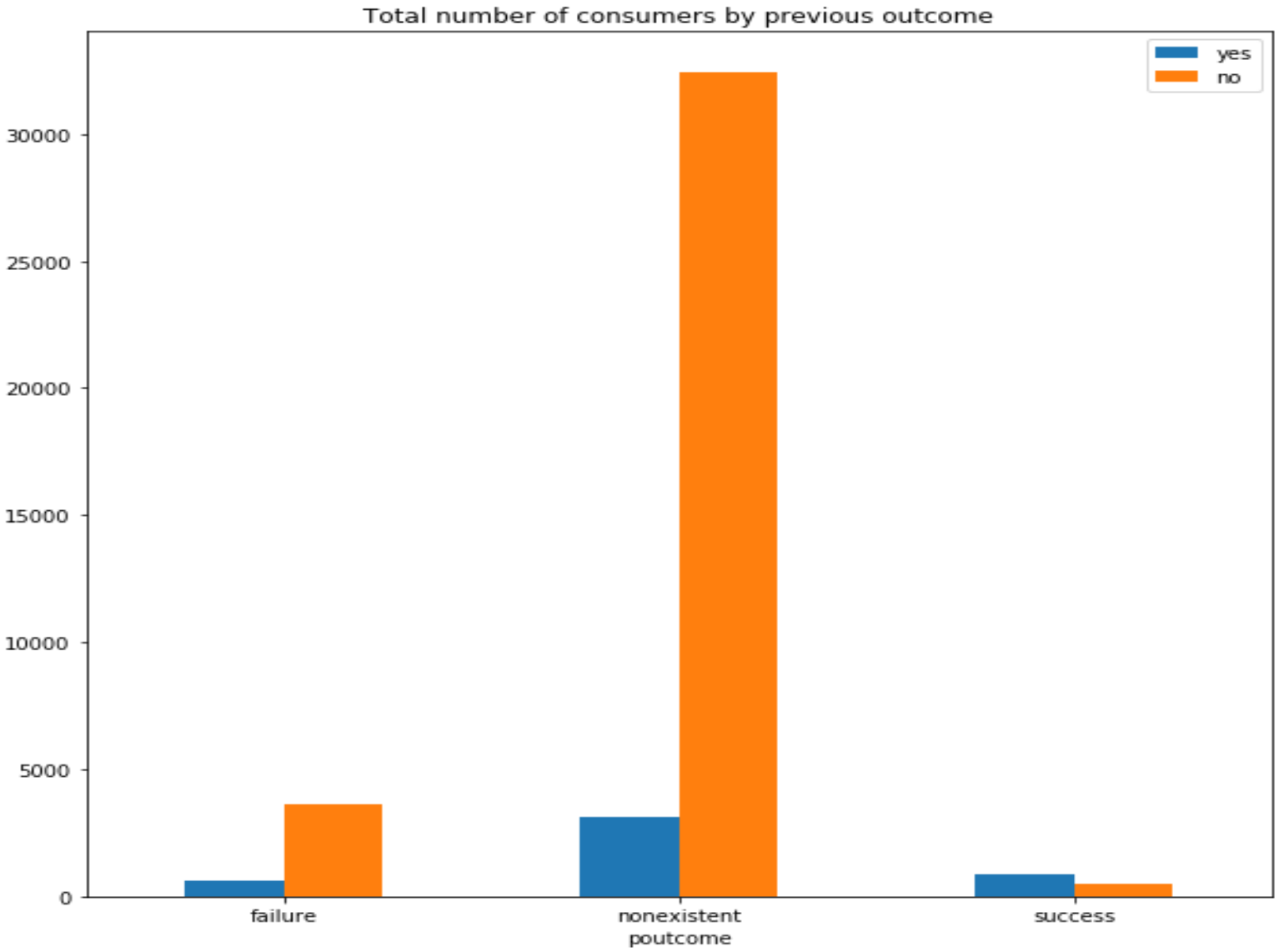
- In the graph we can see that the jobs admin, blu-collar and technician have the highest number of consumers.
- We also observe that at every job, those who did not apply far outnumbered those who did.

Total Number of Consumers by Education



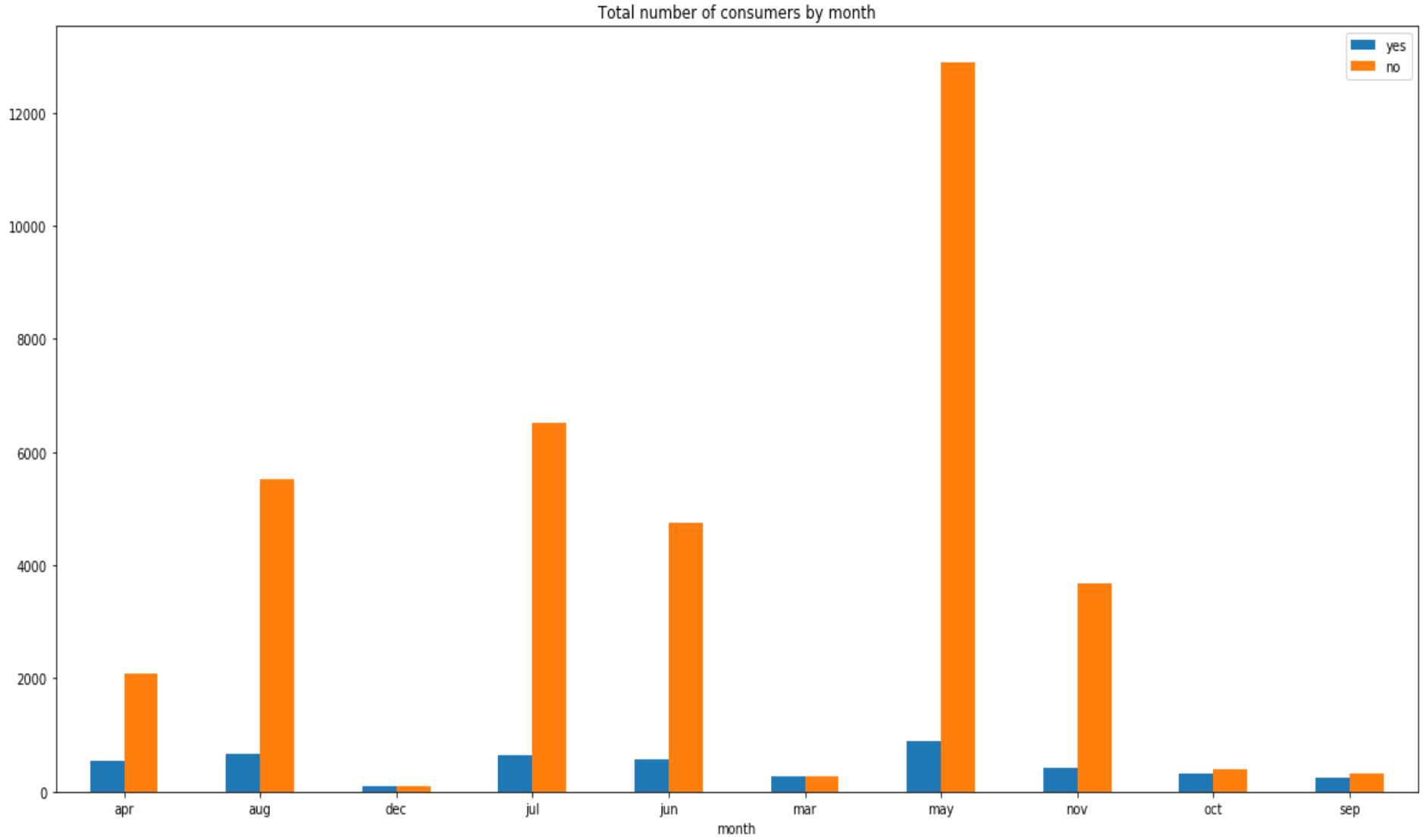
- In the graph we can see that university degree has the highest number of consumers in both categories.
- We also observe that at every education, those who did not apply far outnumbered those who did.

Total Number of Consumers by Previous Outcome



- In the graph we can see that nonexistent poutcome has the highest number of consumers in both categories.
- We also observe that those who have previous campaign and it was successful they subscribe in a term deposit.

Total Number of Consumers by Month



- In the graph we can see in May we have the most contacts in both cases..
- Generally, months in summer have the most contacts in both cases.
- We also observe that December, March, October and September we have small differences between those 2 cases.

Recommendations

Based on the above the EDA showed that age, duration, month and poutcome may have important role for the bank to sell his term deposit to the customers.

The recommended models for this data set are:

- Logistic Regression
- Random Forest Classifier
- Gradient Boosting Classifier

Metrics to evaluate models

Accuracy

- Accuracy is the ratio of the number of correct predictions to the total number of predictions made for a dataset
- Works well, only if there are equal number of samples belonging to each class (balanced data)

Precision

- Which proportion of positive identificators was actually correct (eas actually positive)
- It is a good measure to determine, when the cost of false positive is high

Recall

- Which proportion of actual positives, was identified correctly

F1-Score

- Tells you how precise(how many instances it classifies correctly) and robust(it does not miss a significant number of instances) your model is
- Works well on uneven class distribution (unbalanced data)
- The higher, the better

Our data are unbalanced. So, maybe f1-score is a more valid metric for this dataset.

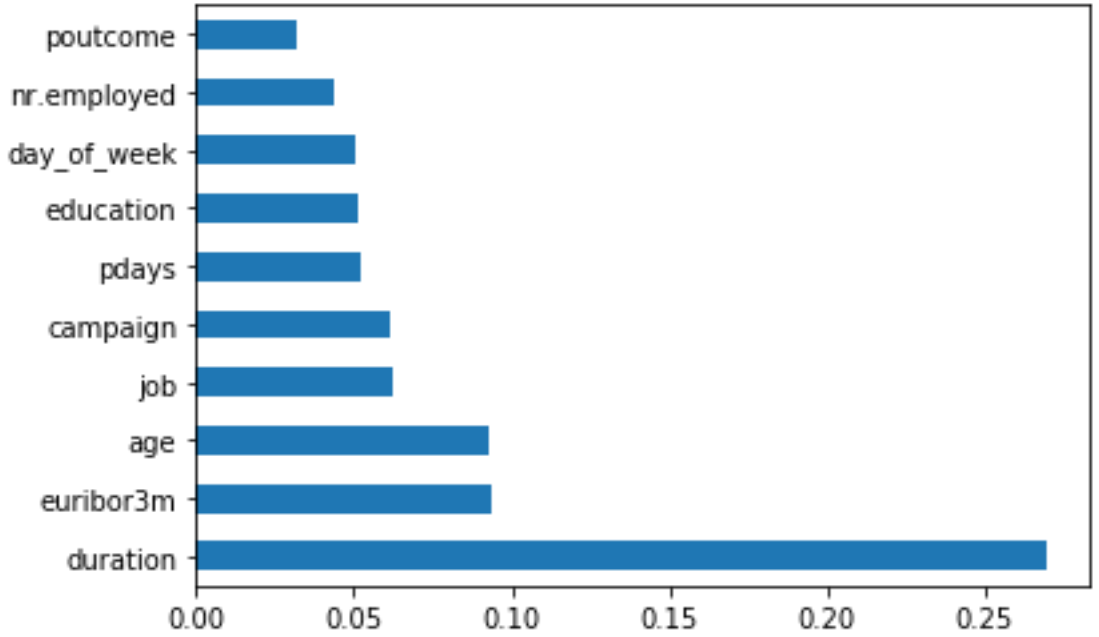
Results

	Logistic Regression	Random Forest Classifier	XGB Classifier
Accuracy	0.903194	0.960530	0.926725
F1-Score	0.646856	0.925535	0.794130
Recall	0.617951	0.924063	0.759743
Precision	0.751228	0.937497	0.854115

Based on the above we can point out that the Random Forest Classifier has the best performance on our dataset.

Feature Importance

- Before we proceed into the final step of our project, we are going to find out which features are most important in our dataset.
- From our EDA, we ended up that age, duration, month and poutcome may have important role in our model. Now, we are going to use feature importance.
- Feature importance gives you a score for each feature of your data, the higher the score more important or relevant is the feature towards your output variable.



As we can see, the most important features are duration, euribor3m, age, job, campaign, pdays, education, day_of_week, nr.employed and poutcome. So we are going to train and test our Random Forest Classifier only using these attributes.

Feature Importance

	Random Forest Classifier
Accuracy	0.964781
F1-Score	0.934930
Recall	0.932654
Precision	0.939855

There is a slightly increase in our results after the feature importance process.

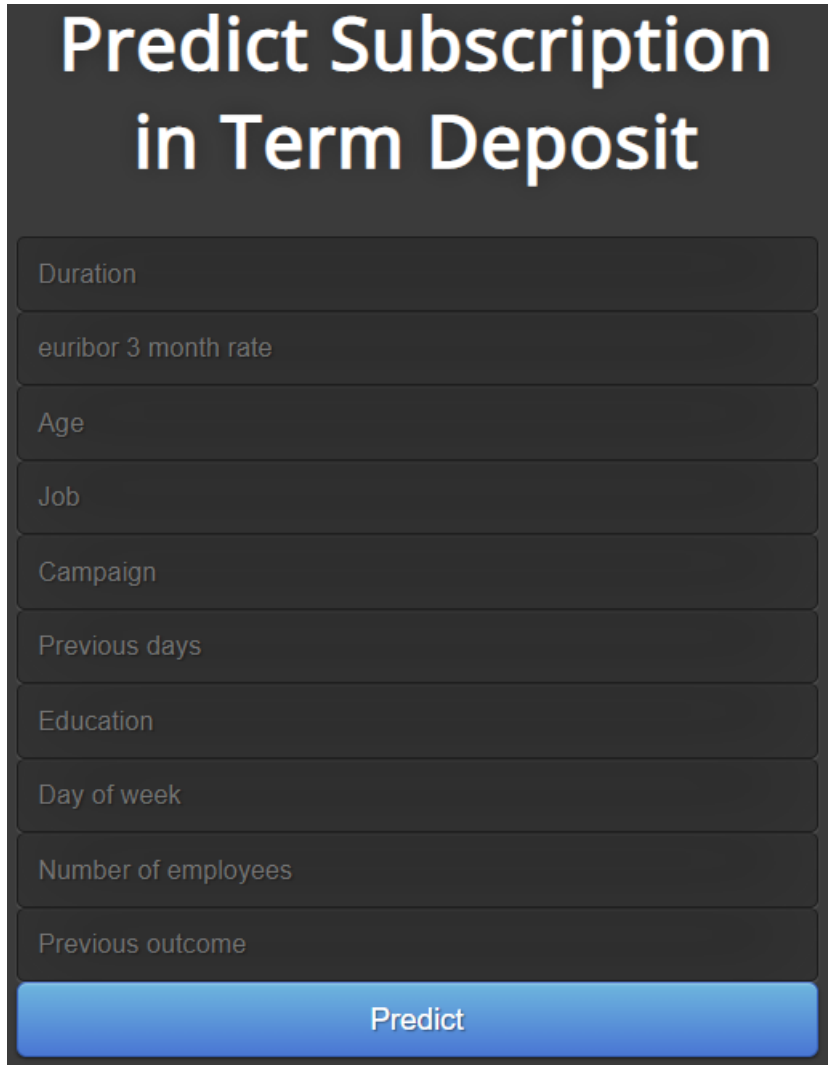
Model Selection

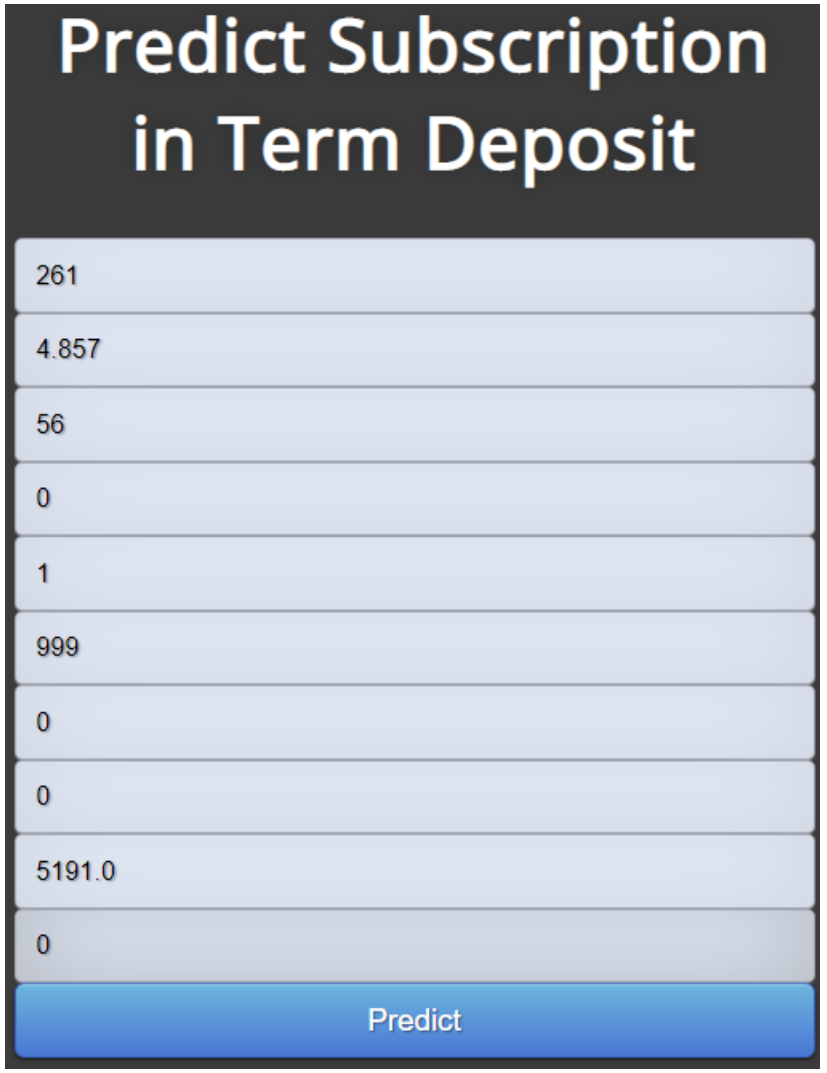
- Based on the result, we can suggest that the **Random Forest Classifier** will fit the needs of the company.
- It has provided us satisfactory results.
- It is a good model for the bank to try and reach out to new customers.

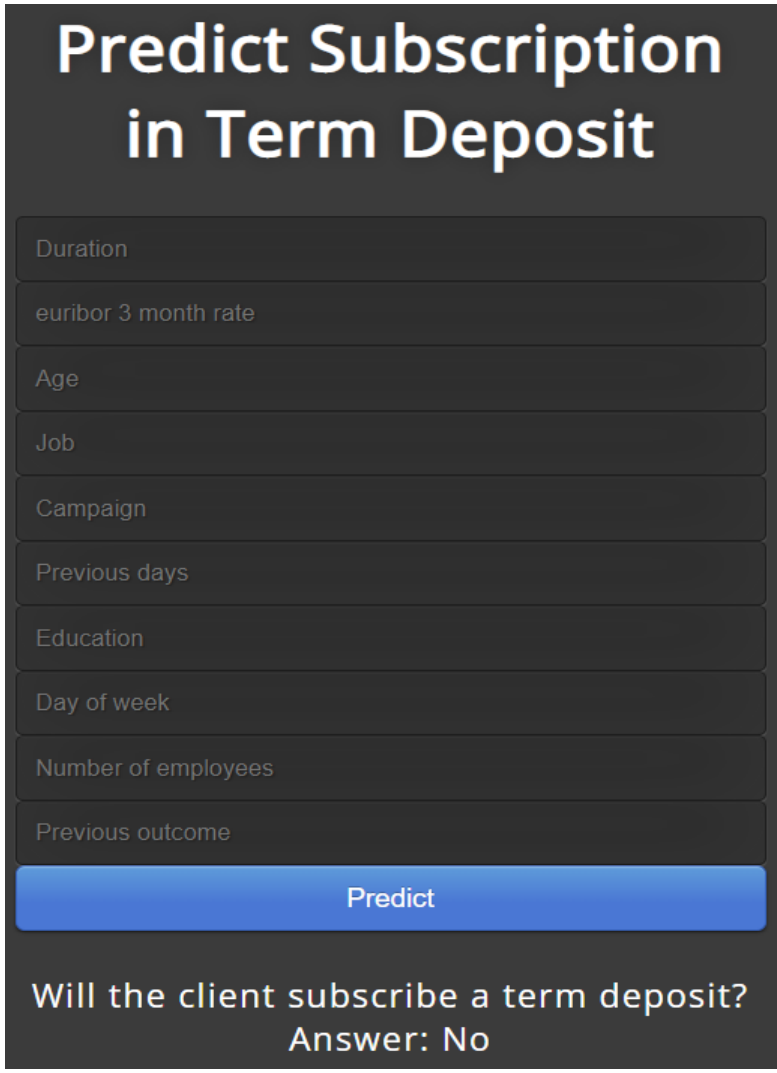
Deploy Model

- We develop a web application that consist of a simple web page with 10 fields that let the user fill the values.
- After submitting the values, the user press the predict button and a message shows and informs him if the client will subscribe a term deposit or not.

Deploy Model







Thank You