

# PREDICTIVE ANALYTICS FOR RETAIL BANKING

## 1. Objectives

Analytics is helping the banking industry become smarter in managing the myriad challenges.

Challenges:

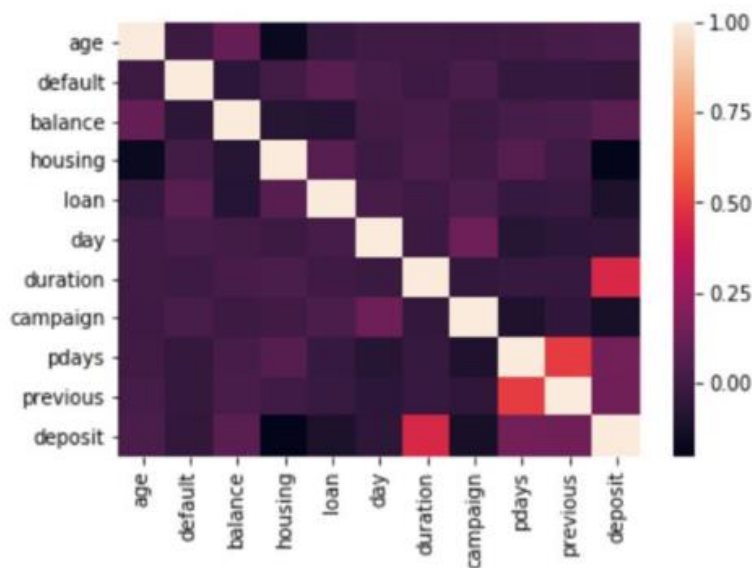
- What is a suitable product to recommend to a customer?
- What is the best time to market the product?
- Which is the most effective channel to contact a customer?

The data is related with direct marketing campaigns of a 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 ('yes') or not ('no') subscribed.

The goal is to predict if the client will subscribe a term deposit.

## 2. Problem Statement

The data-set is related with direct marketing campaigns (were based on phone calls) of a banking institution. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed. The goal is to predict if the client will subscribe a term deposit.



CORRELATION  
USING  
HEATMAP

### 3. Description of the Dataset

This is the classic marketing bank dataset uploaded originally in the UCI Machine Learning Repository. The dataset gives you information about a marketing campaign of a financial institution in which you will have to analyse in order to find ways to look for future strategies in order to improve future marketing campaigns for the bank.

age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	outcome	deposit
59	admin.	married	secondary	no	2343	yes	no	unknown	5	may	1042	1	-1	0	unknown	yes
56	admin.	married	secondary	no	45	no	no	unknown	5	may	1467	1	-1	0	unknown	yes
41	technician	married	secondary	no	1270	yes	no	unknown	5	may	1389	1	-1	0	unknown	yes
55	services	married	secondary	no	2476	yes	no	unknown	5	may	579	1	-1	0	unknown	yes
54	admin.	married	tertiary	no	184	no	no	unknown	5	may	673	2	-1	0	unknown	yes
42	managemt	single	tertiary	no	0	yes	yes	unknown	5	may	562	2	-1	0	unknown	yes
56	managemt	married	tertiary	no	830	yes	yes	unknown	6	may	1201	1	-1	0	unknown	yes
60	retired	divorced	secondary	no	545	yes	no	unknown	6	may	1030	1	-1	0	unknown	yes
37	technician	married	secondary	no	1	yes	no	unknown	6	may	608	1	-1	0	unknown	yes
28	services	single	secondary	no	5090	yes	no	unknown	6	may	1297	3	-1	0	unknown	yes
38	admin.	single	secondary	no	100	yes	no	unknown	7	may	786	1	-1	0	unknown	yes
30	blue-collar	married	secondary	no	309	yes	no	unknown	7	may	1574	2	-1	0	unknown	yes
29	managemt	married	tertiary	no	199	yes	yes	unknown	7	may	1689	4	-1	0	unknown	yes
46	blue-collar	single	tertiary	no	460	yes	no	unknown	7	may	1102	2	-1	0	unknown	yes
31	technician	single	tertiary	no	703	yes	no	unknown	8	may	943	2	-1	0	unknown	yes
35	managemt	divorced	tertiary	no	3837	yes	no	unknown	8	may	1084	1	-1	0	unknown	yes
32	blue-collar	single	primary	no	611	yes	no	unknown	8	may	541	3	-1	0	unknown	yes
49	services	married	secondary	no	-8	yes	no	unknown	8	may	1119	1	-1	0	unknown	yes
41	admin.	married	secondary	no	55	yes	no	unknown	8	may	1120	2	-1	0	unknown	yes
49	admin.	divorced	secondary	no	168	yes	yes	unknown	8	may	513	1	-1	0	unknown	yes
28	admin.	divorced	secondary	no	785	yes	no	unknown	8	may	442	2	-1	0	unknown	yes
43	managemt	single	tertiary	no	2067	yes	no	unknown	8	may	756	1	-1	0	unknown	yes
43	managemt	divorced	tertiary	no	388	yes	no	unknown	8	may	2087	2	-1	0	unknown	yes

#### 3.1 Columns in the Dataset

Age : Age of the client- (numeric)

Job : Client's occupation - (categorical) (admin, blue-collar, entrepreneur, housemaid, management, retired, self employed, services, student, technician, unemployed, unknown)

Marital : Client's marital status - (categorical) (divorced, married, single, unknown, note: divorced means divorced or widowed)

Education : Client's education level - (categorical)

Default : Indicates if the client has credit in default - (categorical) (no, yes)

Balance : average yearly balance, in euros (numeric).

Housing : Does the client as a housing loan? - (categorical) (no, yes)

Loan : Does the client as a personal loan? - (categorical) (no, yes)

Contact : Type of communication contact - (categorical) (unknown, cellular, telephone)

Day : Day of last contact with client.

Month : Month of last contact with client - (categorical) (Jan - Dec)

Duration : Duration of last contact with client, in seconds - (numeric)  
For benchmark purposes only, and not reliable for predictive modelling.

Campaign : number of contacts performed during this campaign and for this client (numeric, includes last contact) - (numeric)  
(includes last contact)

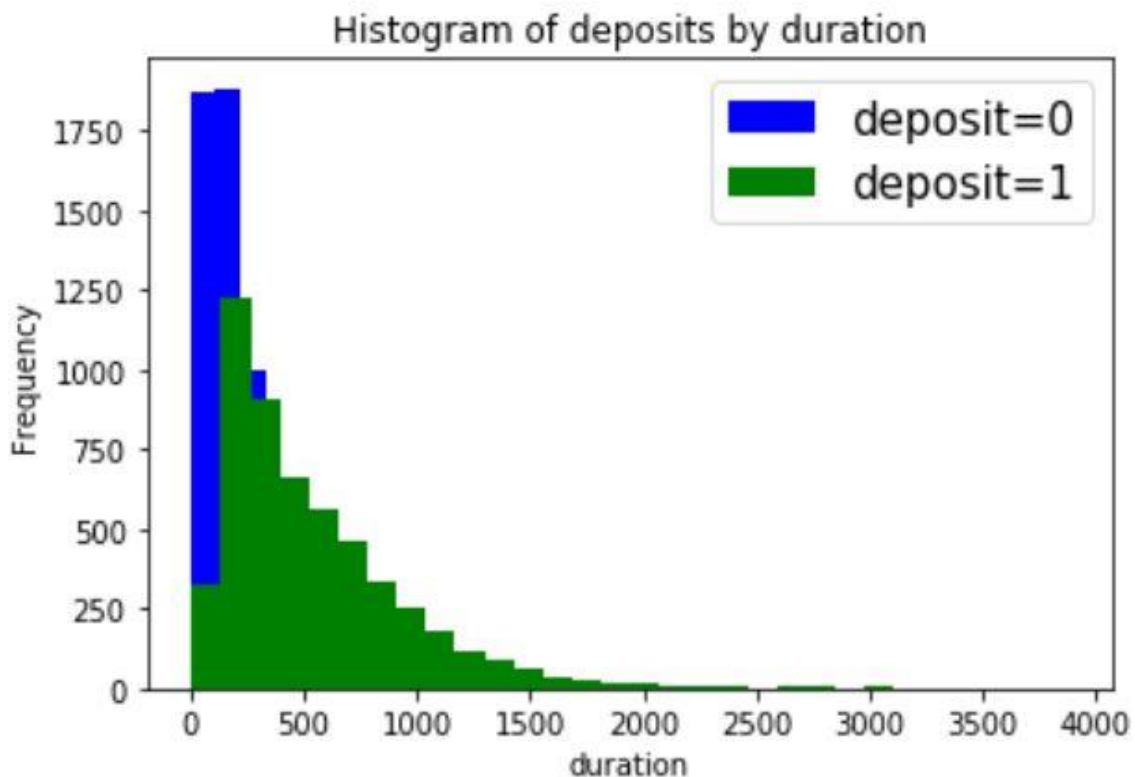
Pdays : Number of days passed client was last contacted - (numeric)  
(-1 means client was not previously contacted)

Previous : Number of client contacts performed before this campaign - (numeric)

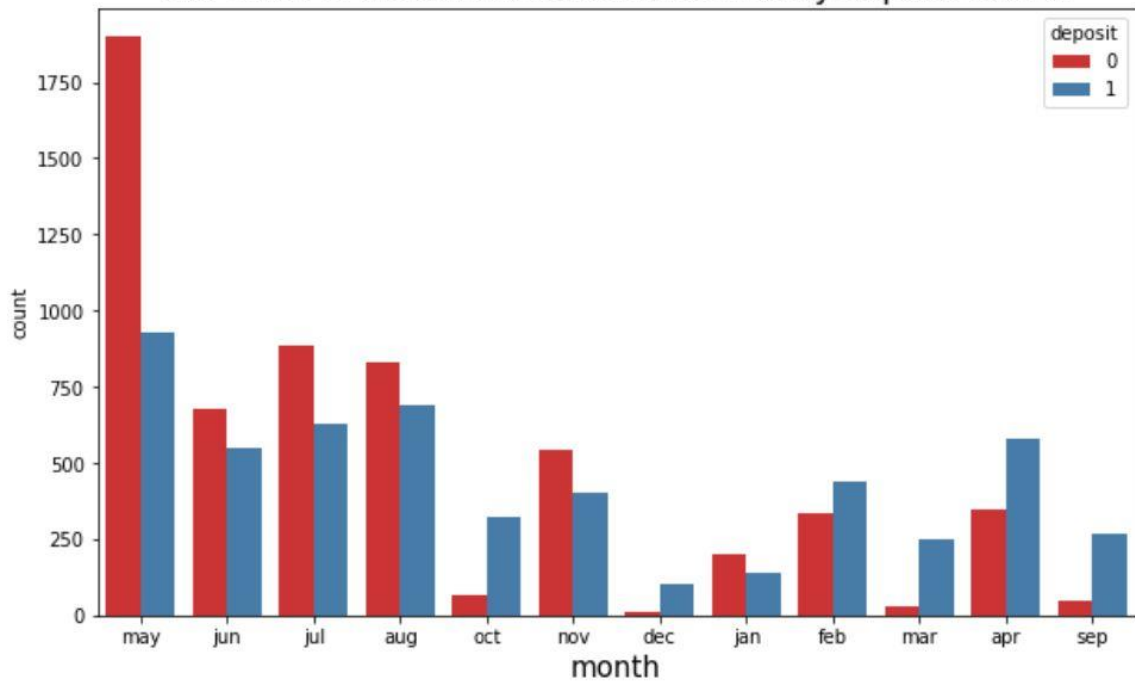
Poutcome : Previous marketing campaign outcome - (categorical)

Deposit : subscription verified. (output)

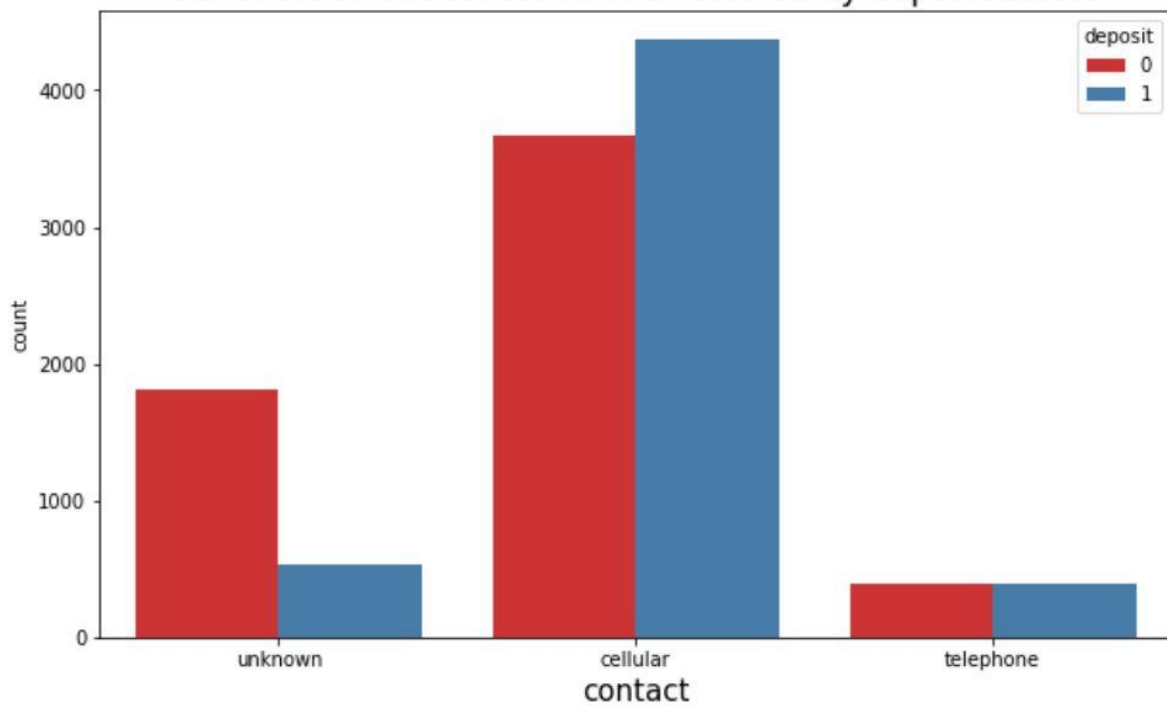
### 3.2 EXPLORATORY DATA ANALYSIS (EDA)



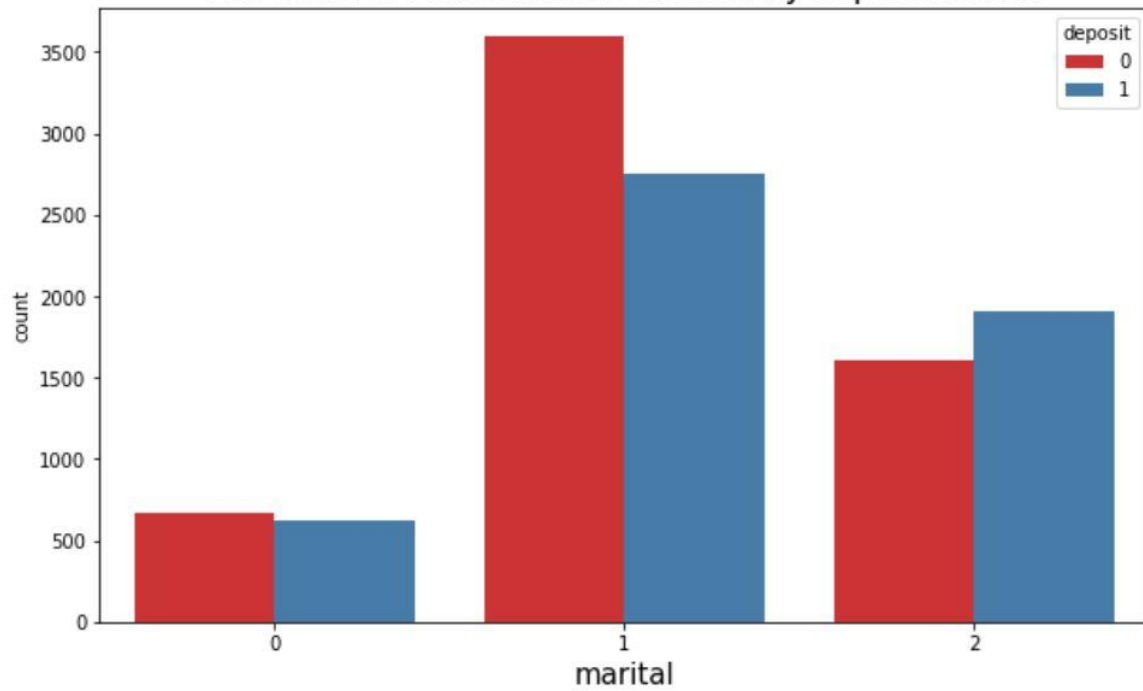
Bar chart of last contact month colored by deposit status



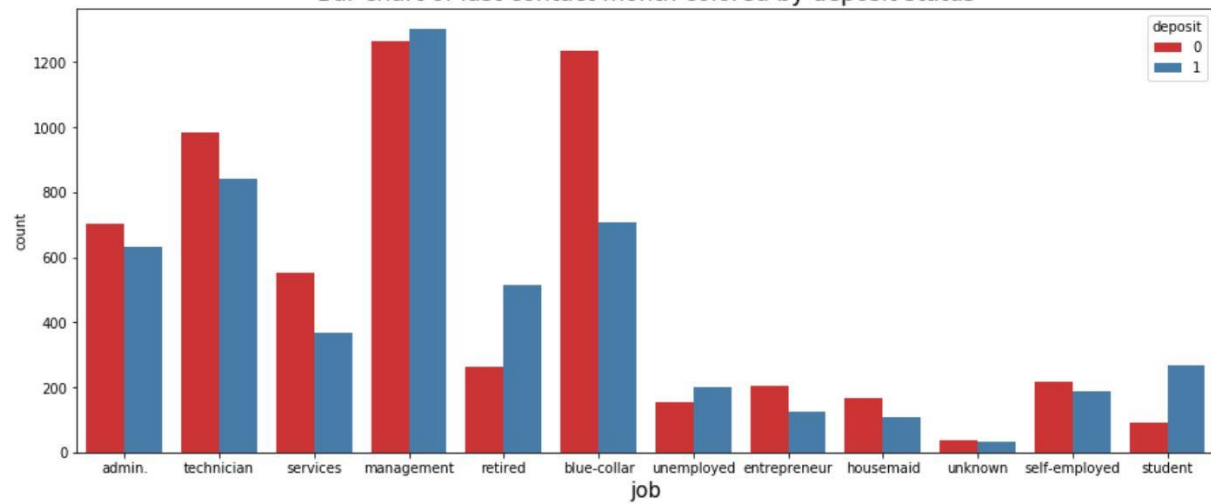
Bar chart of last contact month colored by deposit status

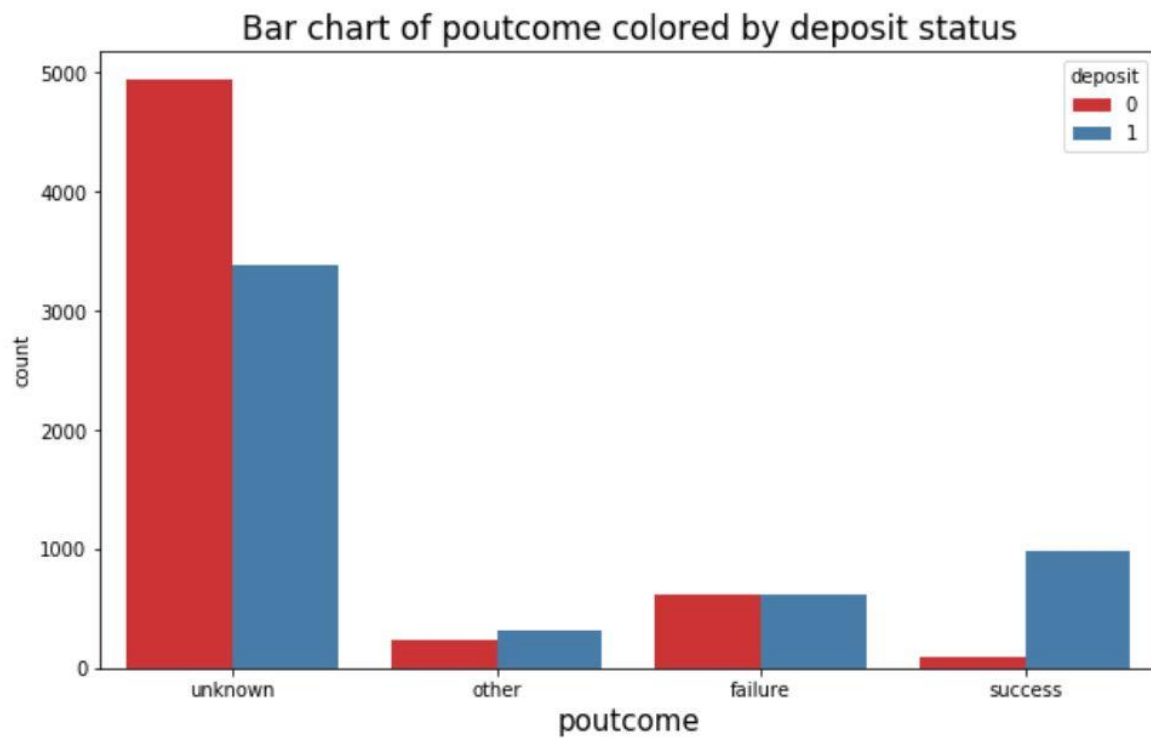
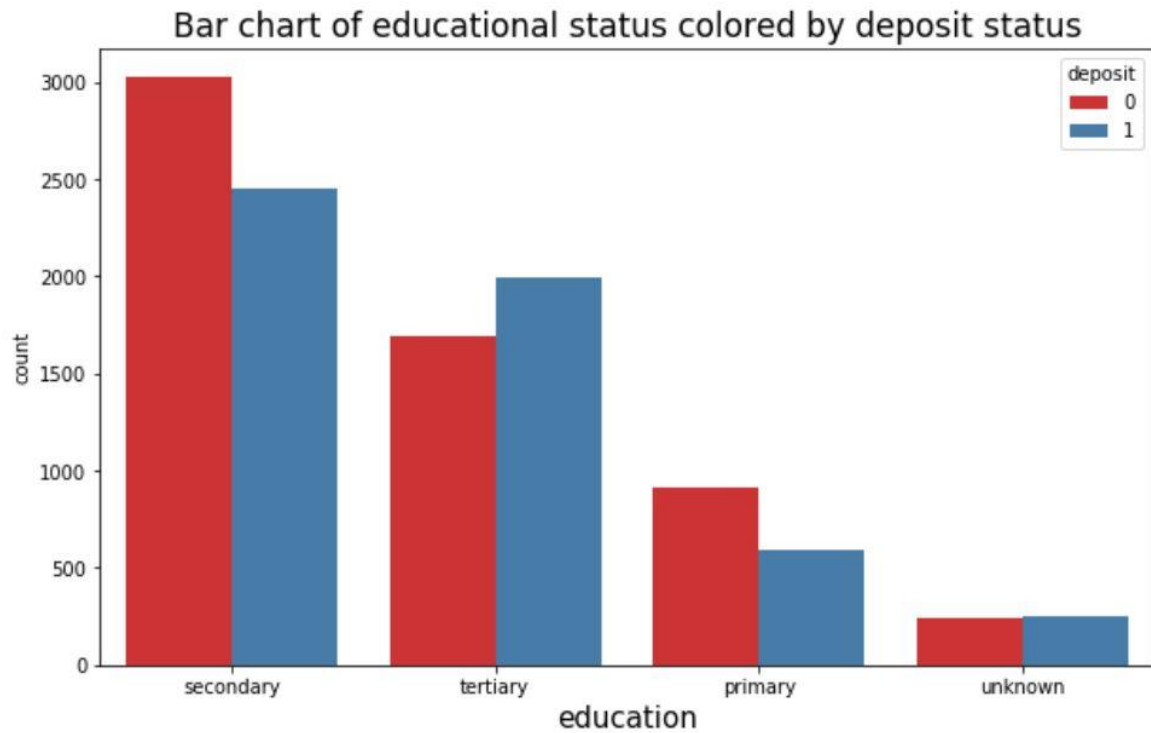


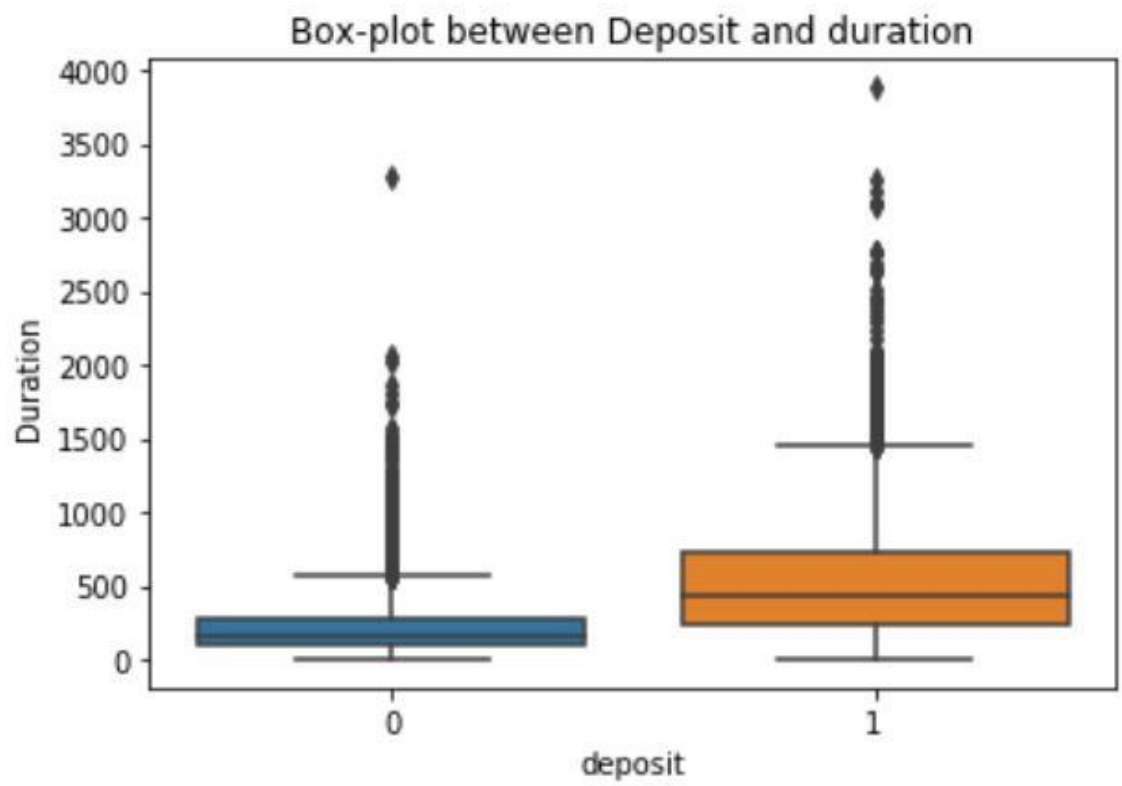
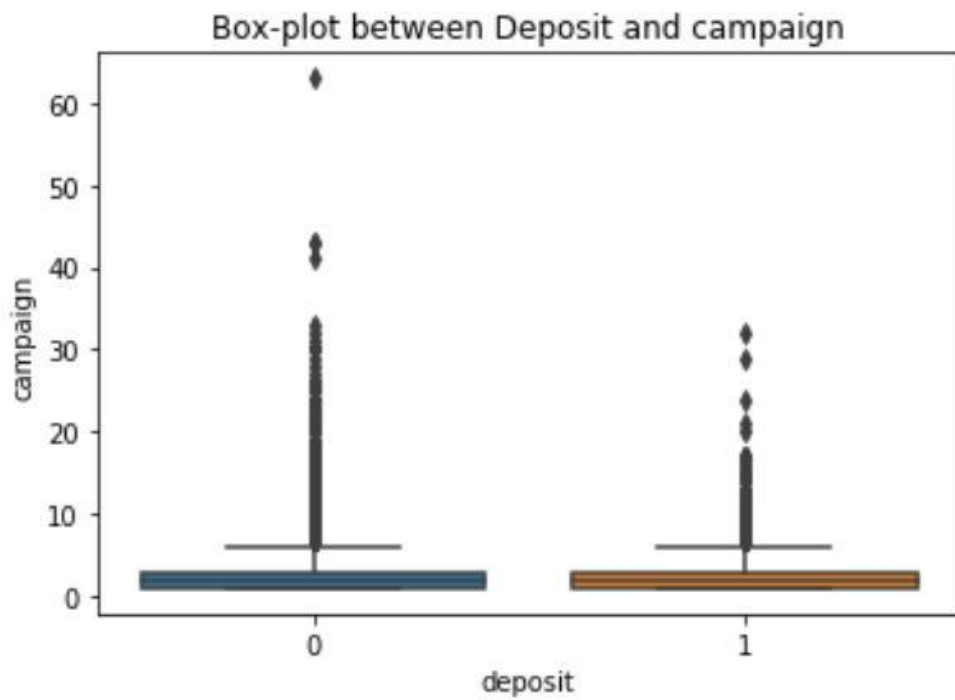
Bar chart of Marital status colored by deposit status



Bar chart of last contact month colored by deposit status







## 4. Model Selection

Since the output variable is in 0's and 1's form, We need to scale down our feature variables to the range of 0 and 1. So we are using Min – Max scaler. Test size = 80-20.

### 4.1 Accuracies compared

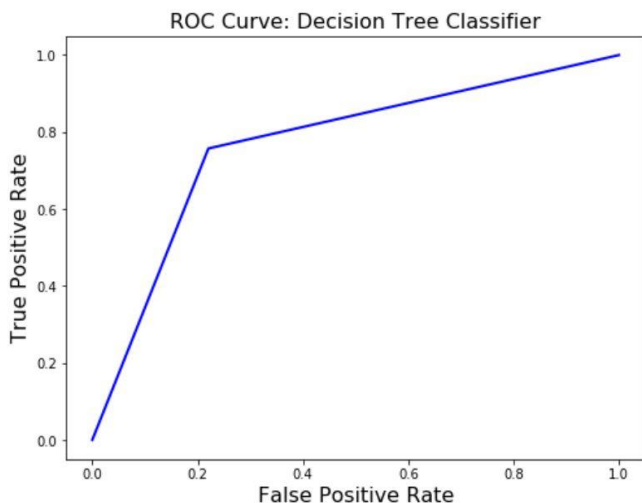
- K-nearest Neighbour: 75.3%
- Logistic Regression: 80.9%
- Decision Tree: 78.2%
- Random Forest Classifier: 78%
- Support vector Machine: 53%

### 4.2 Confusion Matrices

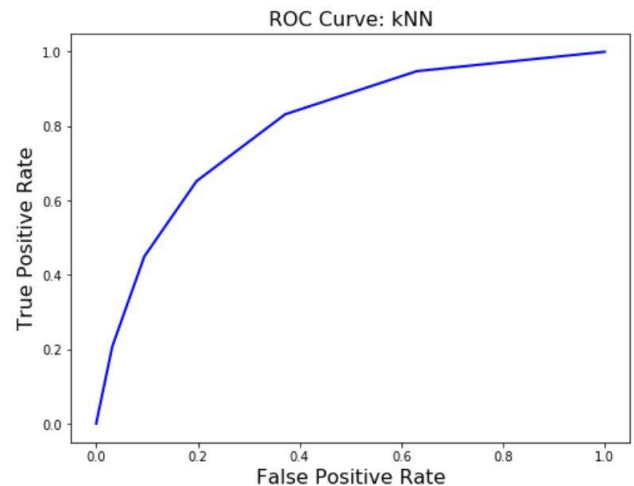
$\begin{bmatrix} 903 & 284 \\ 297 & 749 \end{bmatrix}$	$\begin{bmatrix} 972 & 215 \\ 255 & 791 \end{bmatrix}$	$\begin{bmatrix} 928 & 259 \\ 248 & 798 \end{bmatrix}$	$\begin{bmatrix} 904 & 283 \\ 253 & 793 \end{bmatrix}$	$\begin{bmatrix} 1183 & 4 \\ 1042 & 4 \end{bmatrix}$
KNN	Logistic Regression	Decision Tree	Random Forest	SVM

### 4.3 Graphs

AUC Score (Decision Tree Classifier): 0.77

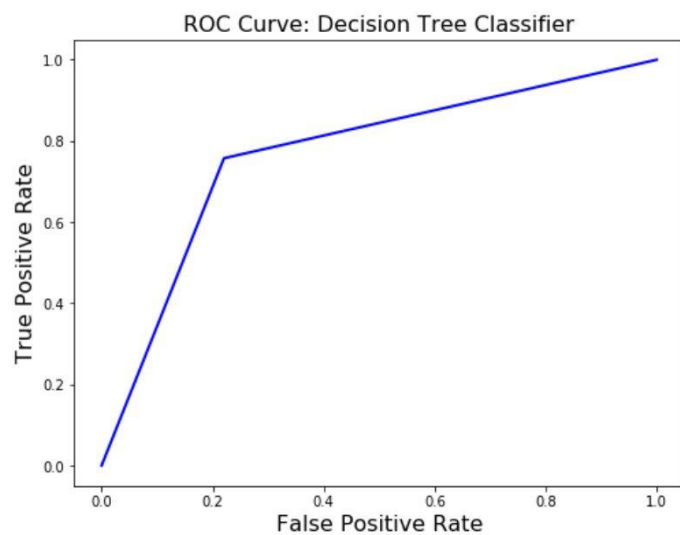


AUC Score (kNN): 0.80

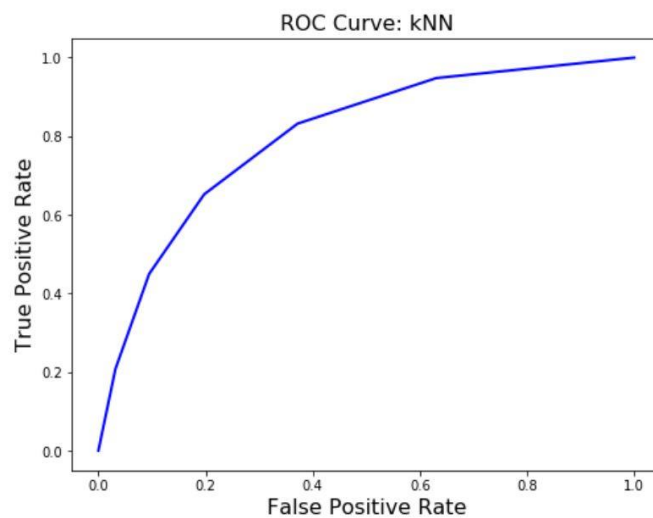




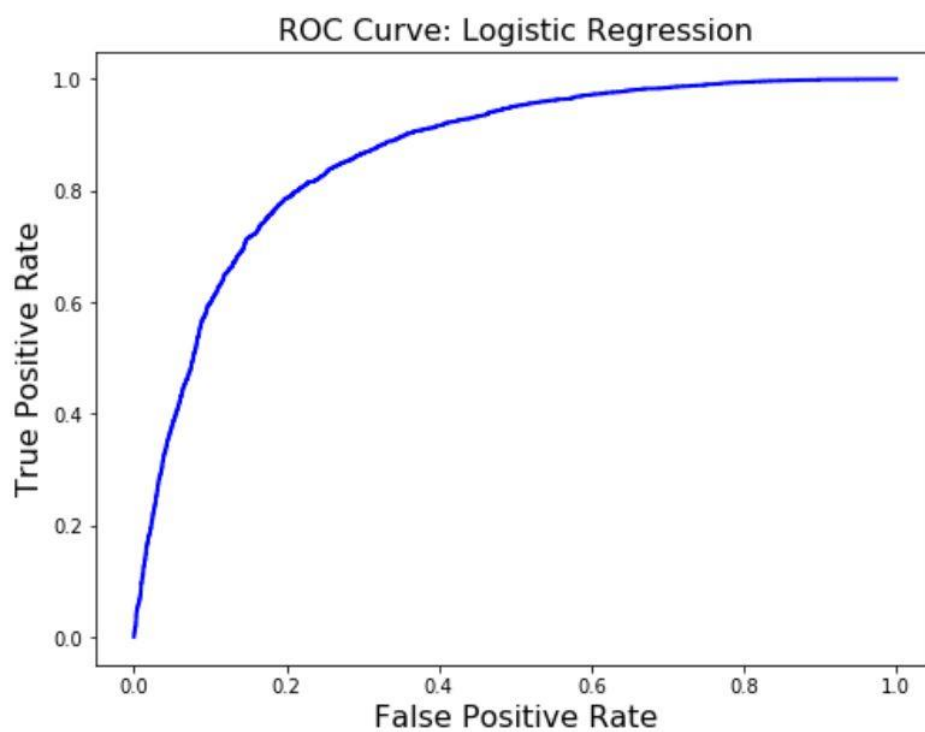
AUC Score (Decision Tree Classifier): 0.77



AUC Score (kNN): 0.80



AUC Score (Logistic Regression): 0.86



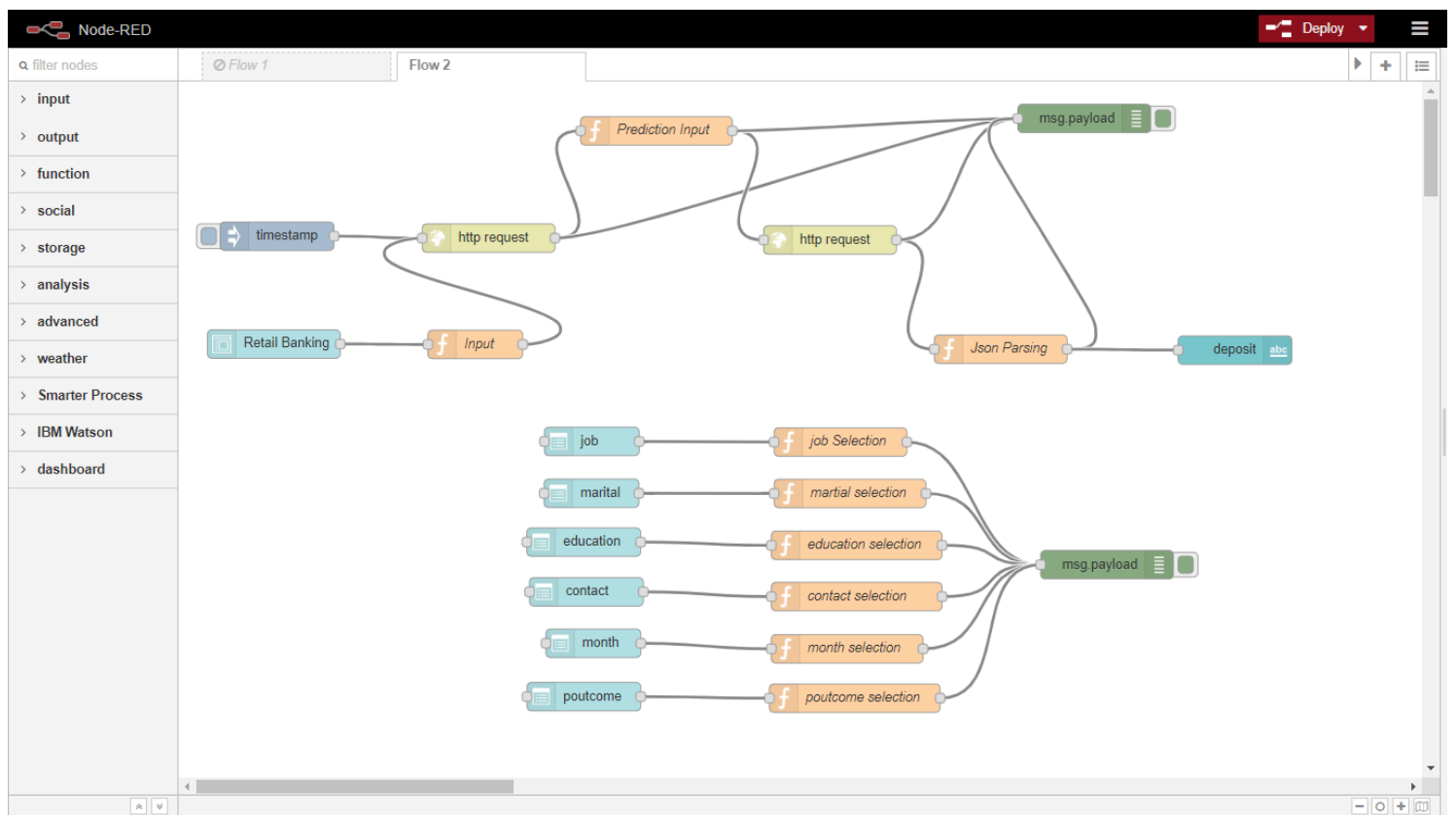
So we choose:

# LOGISTIC REGRESSION

Accuracy = 80.9%

## 5. ScreenShots

### 5.1 Node Red Flow



# 5.2 ML Model UI

Predictive Analytics For Retail Banking

ML Model

jobretired

contactcelluar

monthapr

poutcomeunknown

maritalmarried

educationsecondary

age63

defaulter0

balance\*2030

housing\*0

loan\*0

duration\*61

campaign\*6

pdays\*0

previous\*0

SUBMIT

CANCEL

deposit

## 6. Conclusion

- Most classification problems in the real world are imbalanced. Also, almost always data sets have missing values. In this post, we covered strategies to deal with both missing values and imbalanced data sets. We also explored different ways of building ensembles in sklearn. Below are some takeaway points:
- Sometimes we may be willing to give up some improvement to the model if that would increase the complexity much more than the percentage change in the improvement to the evaluation metrics.
- When building ensemble models, try to use good models that are as different as possible to reduce correlation between the base learners. We could've enhanced our stacked ensemble model by adding *Dense Neural Network* and some other kind of base learners as well as adding more layers to the stacked model.
- Easy Ensemble usually performs better than any other resampling methods.