Using Machine Learning Algorithms to make "Bank Marketing Campaign" more effective

Submitted by:

Md Javed,

MBA Business Analytics Sem 4

Reg. no. 20MBMB13

University of Hyderabad

Under the guidance of:

Dr. Varsha Mamidi,

Assistant Professor,

School of Management Studies

University of Hyderabad

Introduction

The number of participants in the banking industry is high in numbers, and they're equipped with the latest technology to serve their customers better.

Machine learning, a branch of Artificial Intelligence, has started a new age of knowledge discovery & hence intensifying the business competition.

The main objective of this project is to make the telemarketing campaign of a Portuguese bank, which is struggling with revenue, more effective. We tried addressing the problem by creating a business plan that is inspired by the "80-20 rule" which is identifying the 20% of customers that can help generate 80% of the revenue. We have used supervised machine learning-based classification methods to understand the existing customer's behaviour so that the outcome of any marketing lead can be forecasted & hence the business resources are better utilized to maximize the revenue.

Business Problem

The revenue of Portuguese bank is falling & it was observed that the customers were not making frequent deposits. Hence, the bank planned to run a marketing campaign that will try to increase the subscription of "term deposits" (TD) as it will lock the money of savers for a certain period.

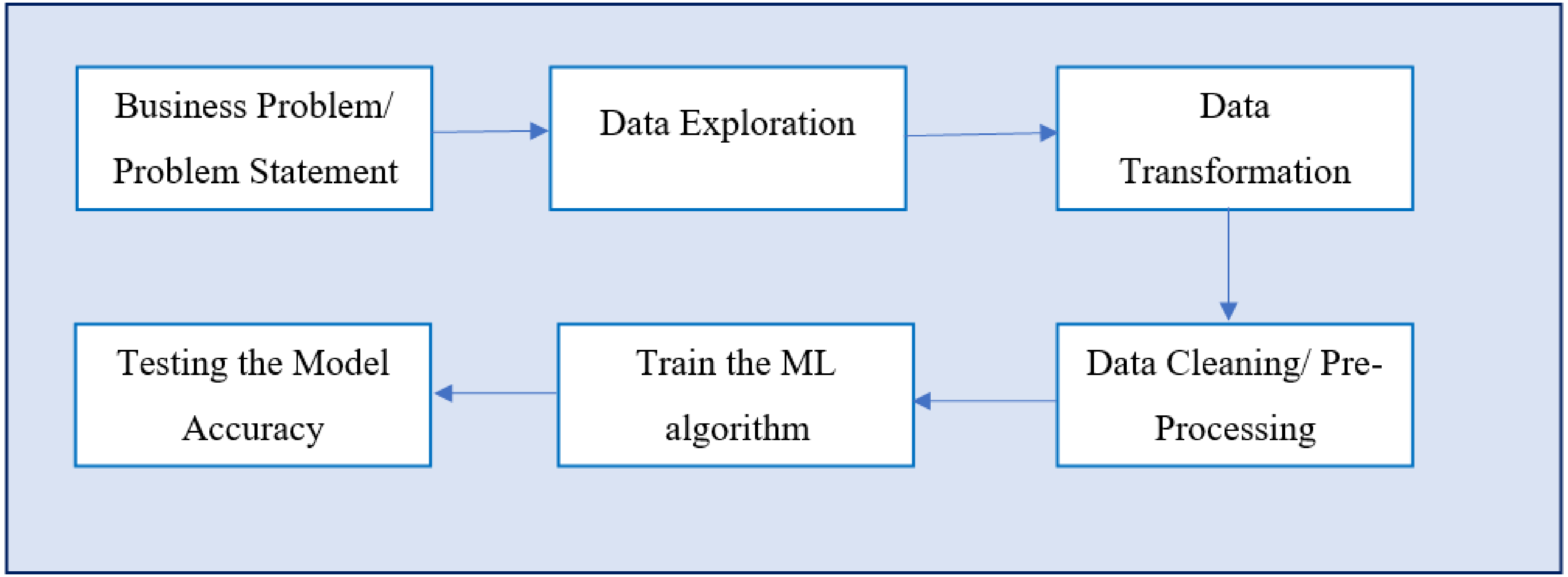
Additionally, the bank has better scope to run parallel marketing for insurance products or mutual funds that can further improve the revenue. Hence, the bank intends to identify the target customer from the existing client base with a higher chance of subscribing to a term deposit and focusing market efforts on such clusters of customers.

Proposed Solution:

- ✓ To create a system that automatically places the bank's perspective leads on having a term deposit.
- ✓ We will be creating a classification algorithm and also suggest to them the insights we derive from this dataset and also help them to narrow down their leads into the marketing funnel and in the end, make a term deposit.

Project Architecture

The project architecture is inspired by the CRISP-DM methodology.



Problem Statement - To analyze the Bank lead's dataset and create a classification model.

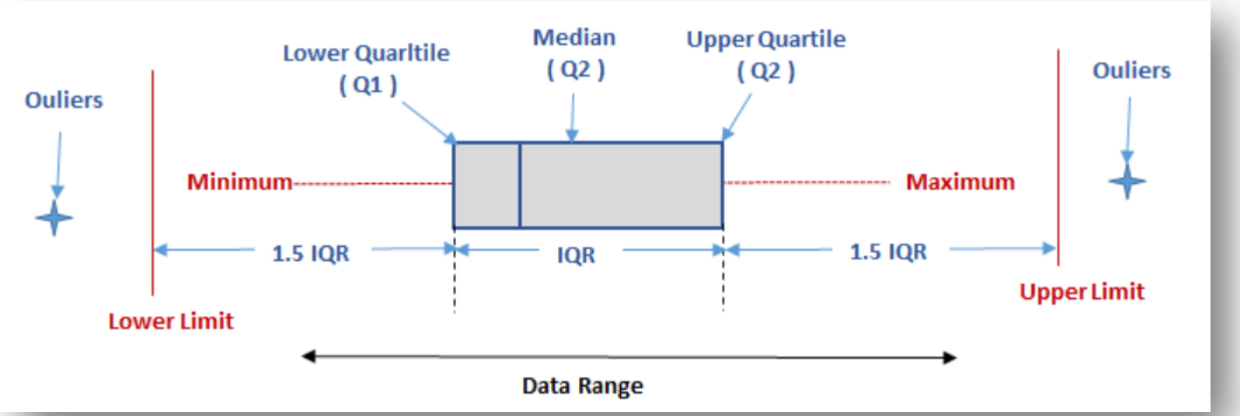
Dataset

```
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";"eu
56;"housemaid";"married";"basic.4y";"no";"no";"no";"telephone";"may";"mon";261;1;999;0;"nonexistent";1.1;93.994;-36.4;4.857;5191;"no"
57;"services";"married";"high.school";"unknown";"no";"no";"telephone";"may";"mon";149;1;999;0;"nonexistent";1.1;93.994;-36.4;4.857;5191;"no"
37;"services";"married";"high.school";"no";"yes";"no";"telephone";"may";"mon";226;1;999;0;"nonexistent";1.1;93.994;-36.4;4.857;5191;"no"
40;"admin.";"married";"basic.6y";"no";"no";"no";"telephone";"may";"mon";151;1;999;0;"nonexistent";1.1;93.994;-36.4;4.857;5191;"no"
56;"services";"married";"high.school";"no";"no";"yes";"telephone";"may";"mon";307;1;999;0;"nonexistent";1.1;93.994;-36.4;4.857;5191;"no"
45;"services";"married";"basic.9y";"unknown";"no";"no";"telephone";"may";"mon";198;1;999;0;"nonexistent";1.1;93.994;-36.4;4.857;5191;"no"
59;"admin.";"married";"professional.course";"no";"no";"no";"telephone";"may";"mon";139;1;999;0;"nonexistent";1.1;93.994;-36.4;4.857;5191;"no"
41;"blue-collar";"married";"unknown";"unknown";"no";"no";"telephone";"may";"mon";217;1;999;0;"nonexistent";1.1;93.994;-36.4;4.857;5191;"no"
24;"technician";"single";"professional.course";"no";"yes";"no";"telephone";"may";"mon";380;1;999;0;"nonexistent";1.1;93.994;-36.4;4.857;5191;"no"
25;"services";"single";"high.school";"no";"yes";"no";"telephone";"may";"mon";50;1;999;0;"nonexistent";1.1;93.994;-36.4;4.857;5191;"no"
```

- Source: "Machine Learning Repository" of "The University of California Irvine"
- Dataset in CSV format with 41,188 rows & 21 columns
- Has Portugal Bank [tele]marketing campaign results
- If client agreed to place a deposit, target variable – 'y' is marked as 'yes' otherwise 'no'.

Data Pre-processing

- Checked for the null values
- Dealing with outliers



$IQR = Q_3 – Q_1$

$\text{Lower limit} = Q_1 – 1.5 * IQR$

$\text{Upper limit} = Q_3 + 1.5 * IQR$

- **Feature engineering steps**:
  1. In feature ‘education’, the clubbed ‘basic.9y’, ‘basic.4y’ & ‘basic.6y’ into one category ‘middle.school’
  2. Coding the months & days of the week as per their respective numbers

3. Coding ‘999’ in pdays as ‘0’

4. For features – ‘housing’, ‘default’ & ‘loan’, we have coded as [(yes, no, unknown) = (1,0,-1)]

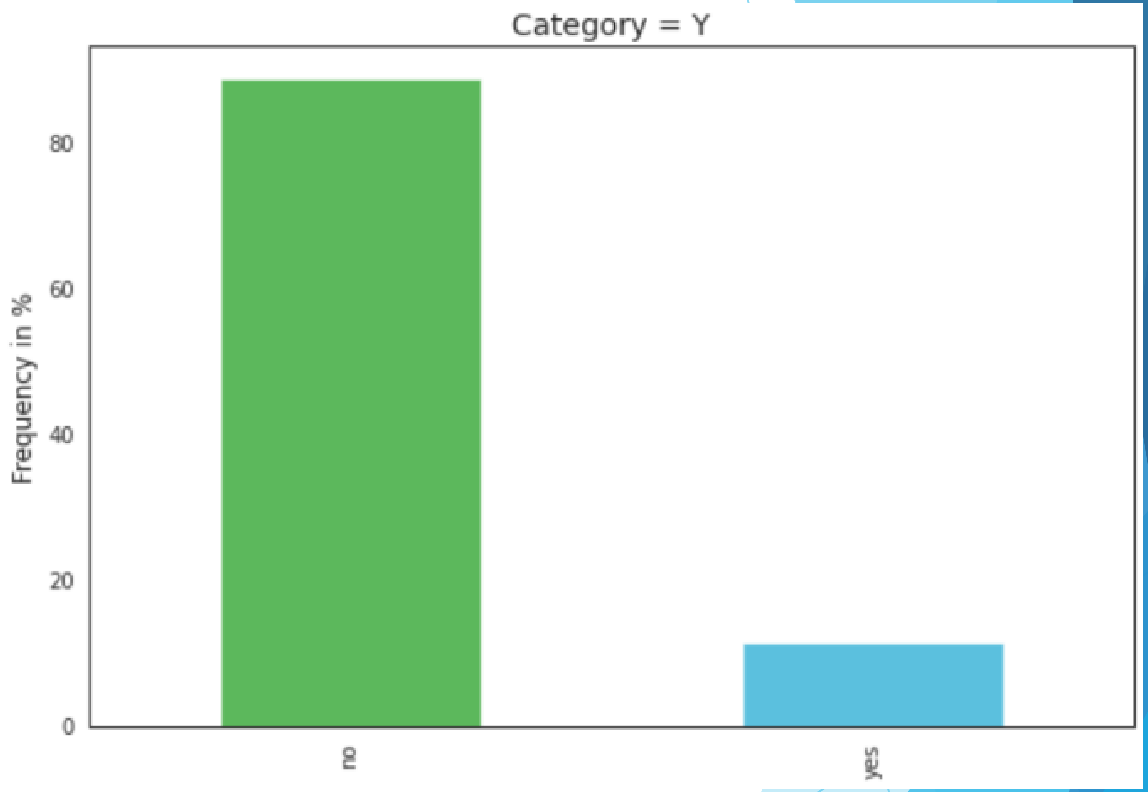
5. Label encoding of [‘contact’, ‘poutcome’] features & then drop them from the dataset

6. Encoding [‘job’, ‘education’] as per their frequencies in each of their respective categories

7. Label encoding of ‘marital’ features

- Standardisation using the standard scalar method
- Picked the important features and rest were dropped
- Creating ‘training: testing’ data split based on ‘80:20’ ratio

- Applying SMOTE (Synthetic Minority Oversampling Technique) to reduce the data imbalance:
- 89% of customers have not subscribed to the term deposit (TD) product
- we have used SMOTE method to oversample the minority class
- The original sample count of training features is 28,448 & up-sampled is 52,276.

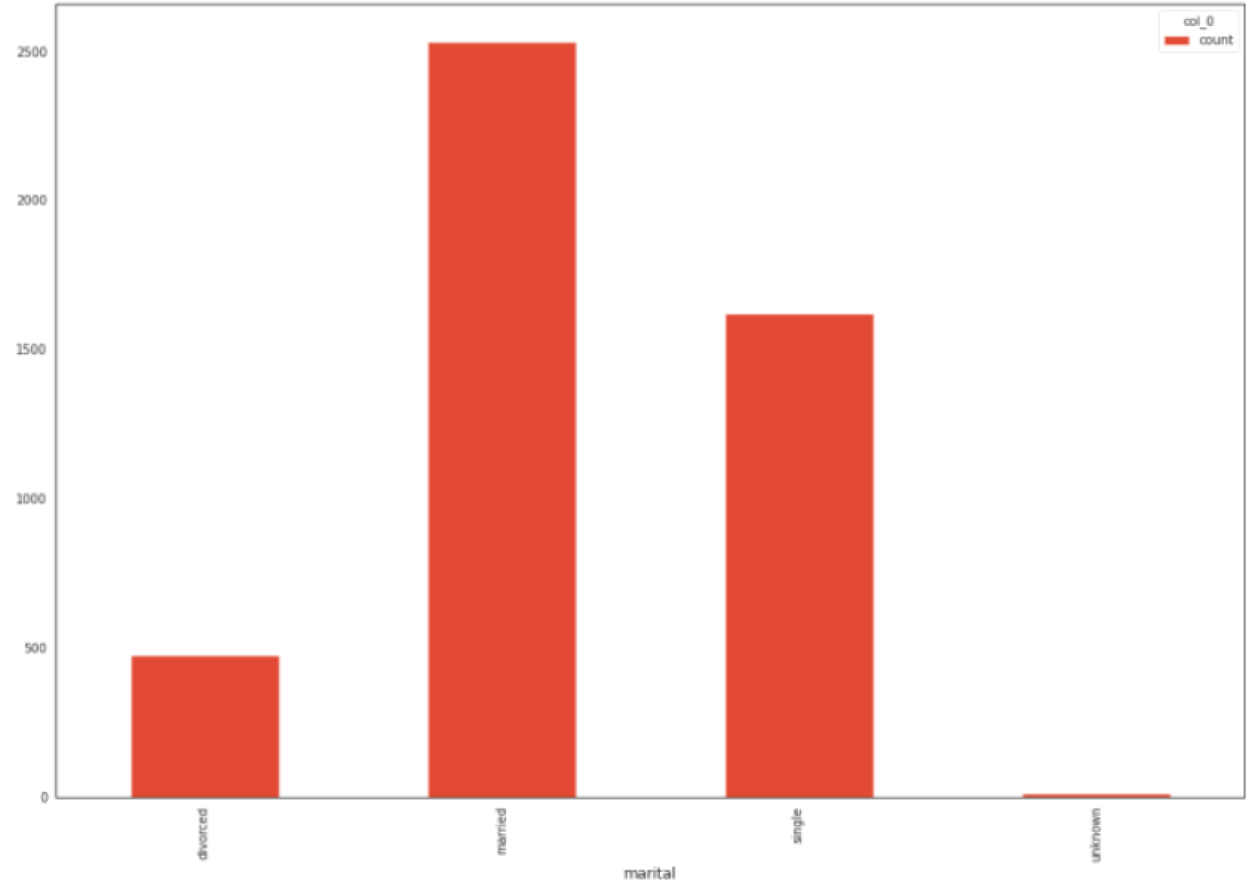


Exploratory Data Analysis

Duration of calls vs Job roles

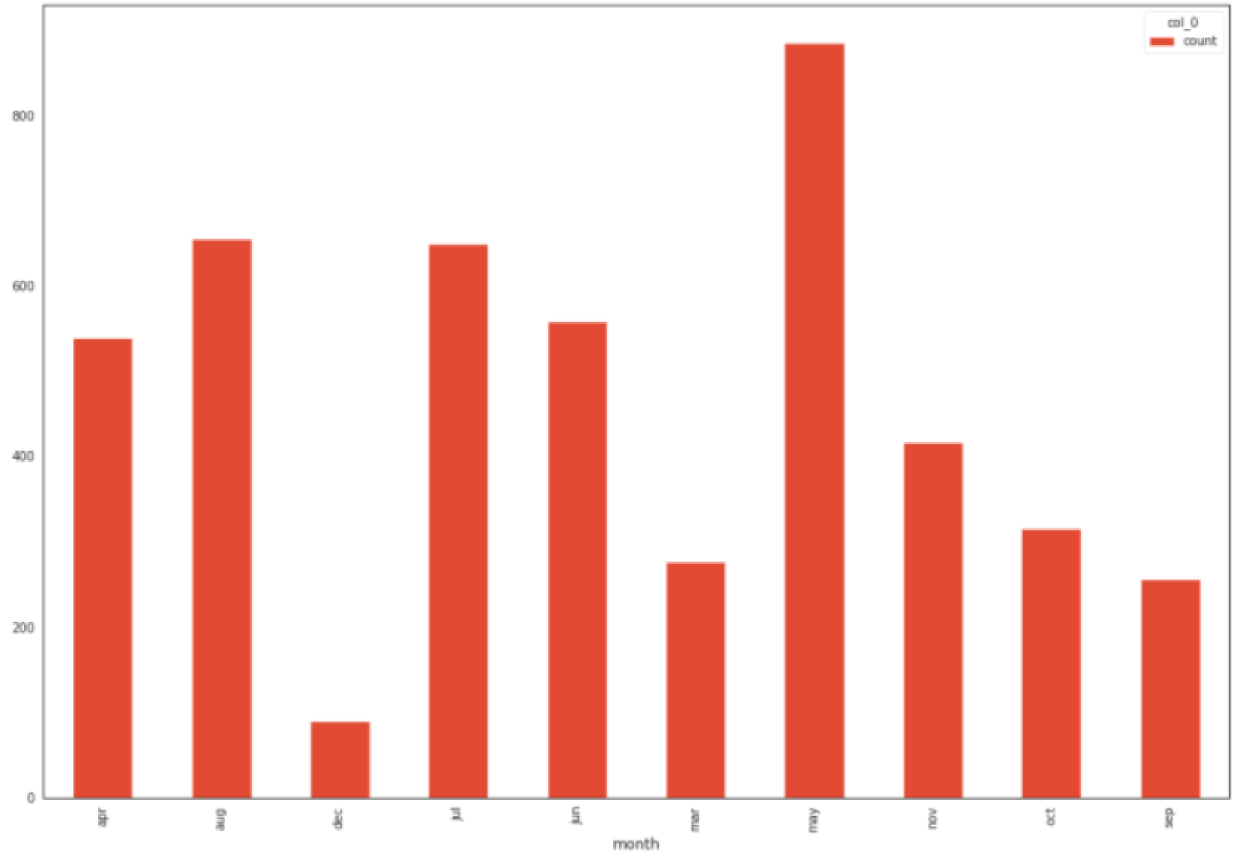
- Leads with lesser call durations yields negative outcome
- More positive outcome are being supplied from the occupations - blue-collar & entrepreneurs

Positive Deposits vs Marital Status



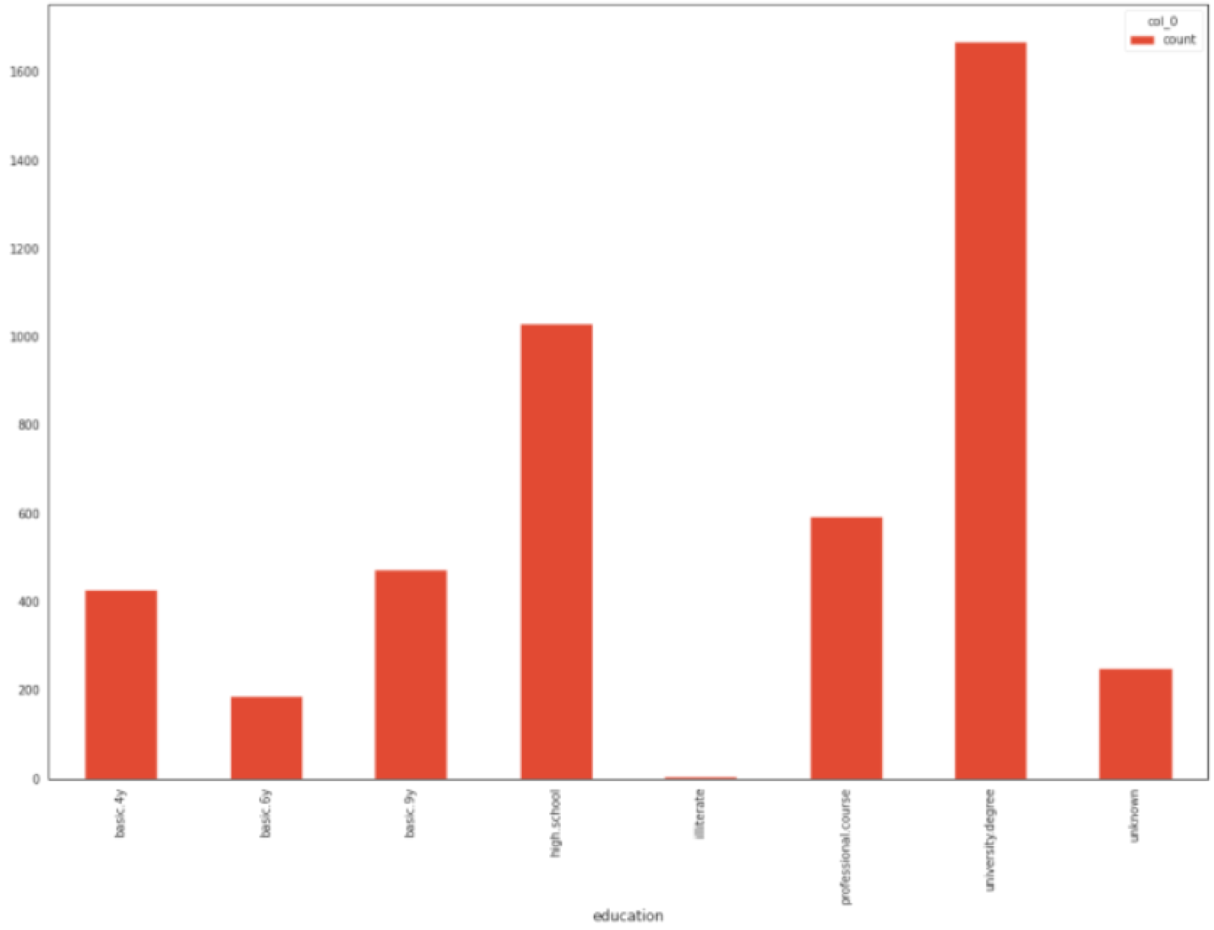
- Married customers are more likely to open the term deposit account

Positive Deposits vs month



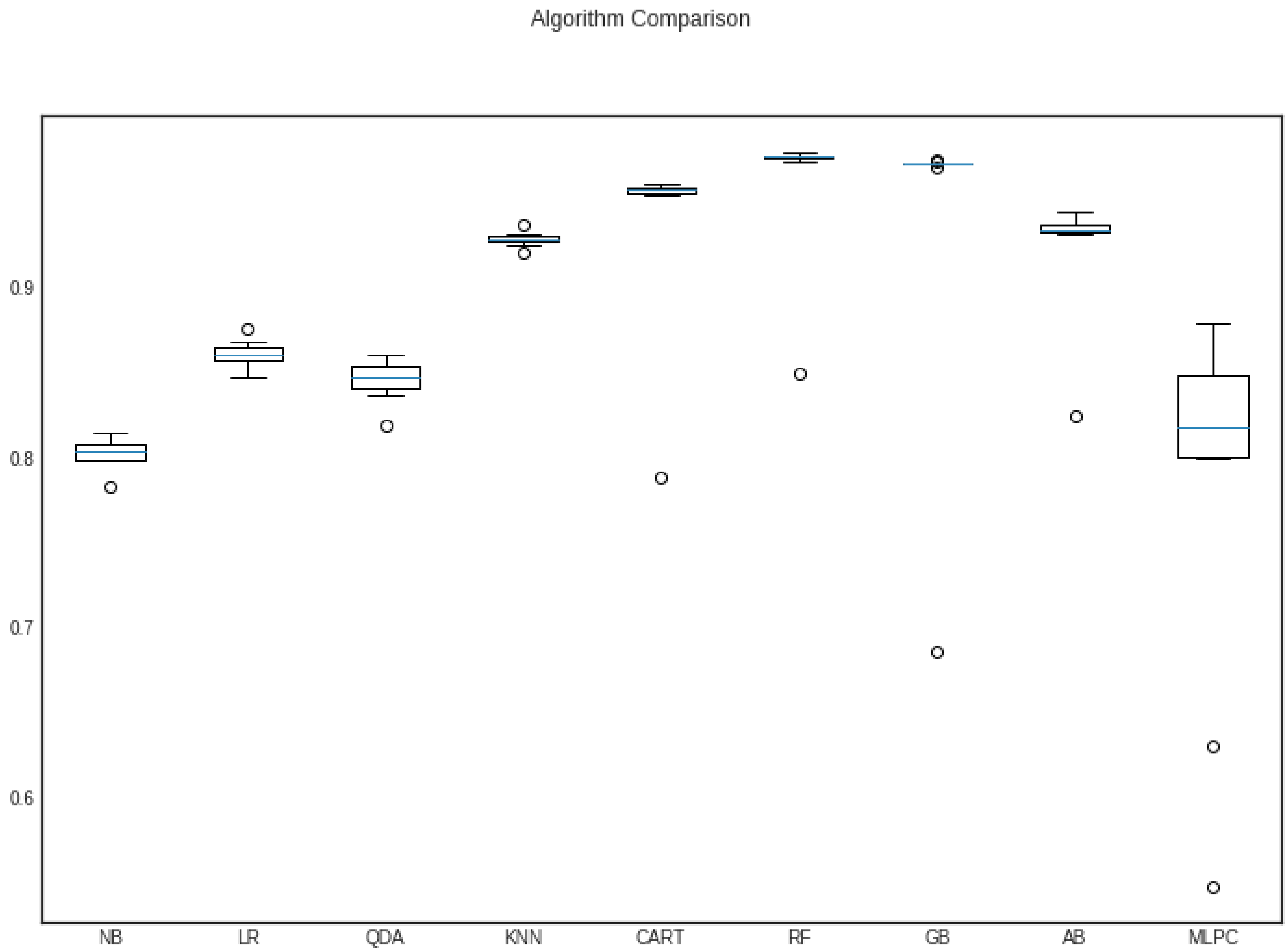
- In the month of May, maximum customers have subscribed

Positive Deposits vs education level



- Leads who had at least a university degree had made the deposits followed by high school

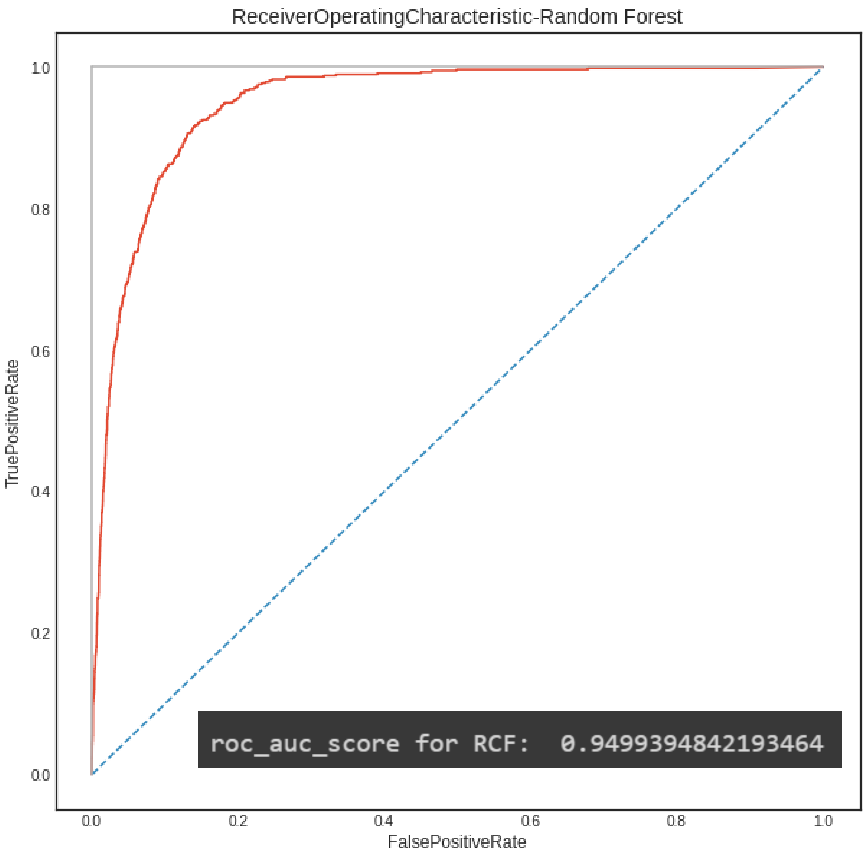
Modelling – with upscaled data



Abbreviation key:
- NB – Bernoulli Naïve Bayes
- LR – Logistic Regression
- QDA – Quadratic Discriminant Analysis
- Decision tree classifier
- Random Forest Classifier
- Gradient Boosting Classifier
- Ada Boosting Classifier
- MLPC – Multi Layer Perceptron Classifier

- RF is giving best accuracy score of 96% (approx.)

Building model with Random Forest Classification

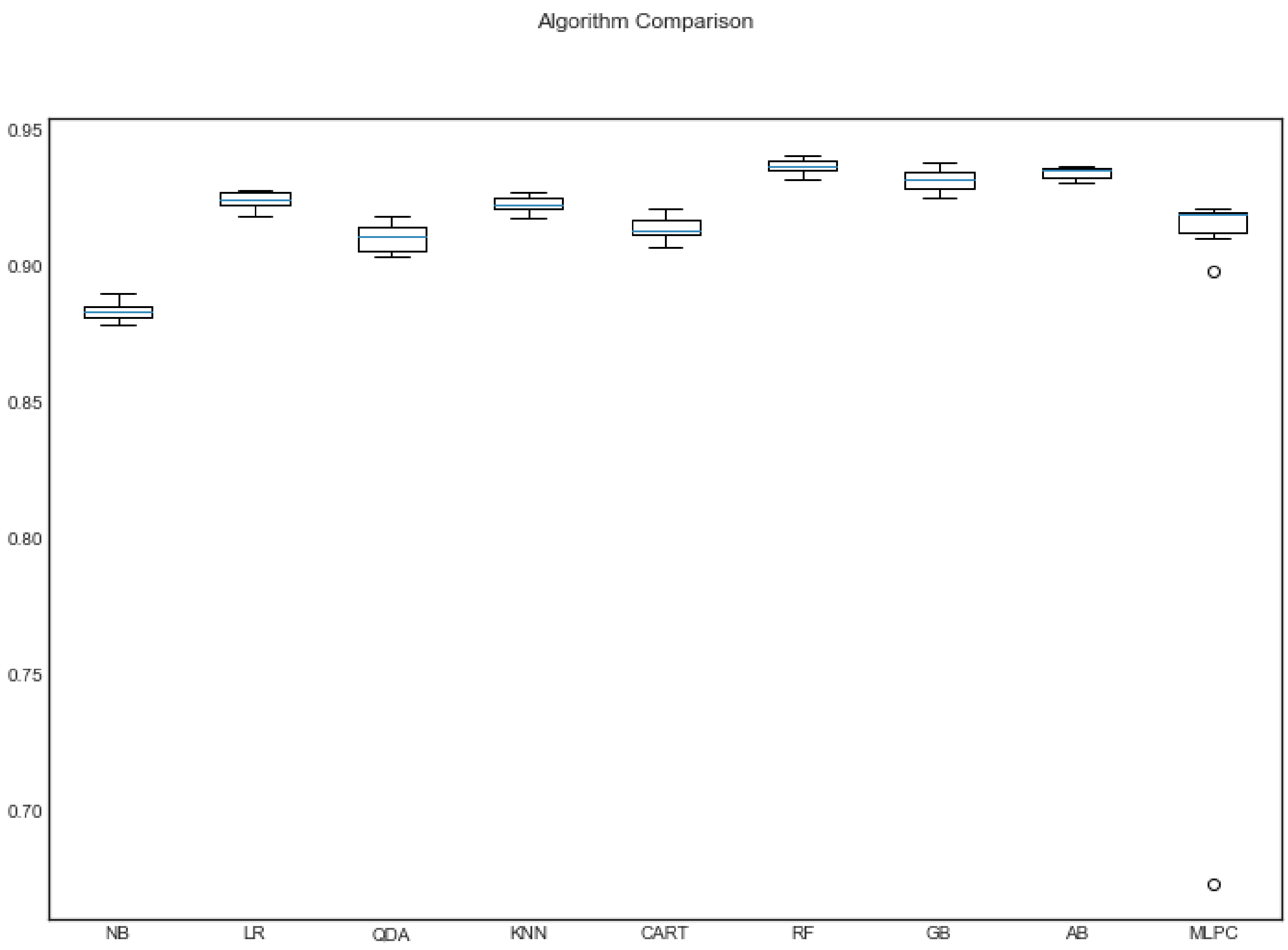



```
Confusion Matrix:
 [[6225  313]
 [ 175  400]]
Accuracy: 0.9313932236749614
Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.95      0.96      6538
           1       0.56      0.70      0.62       575

    accuracy                           0.93      7113
   macro avg       0.77      0.82      0.79      7113
weighted avg       0.94      0.93      0.93      7113
```

Tuned Parameters:
‘n_estimators=128’ & ‘max_depth=16’

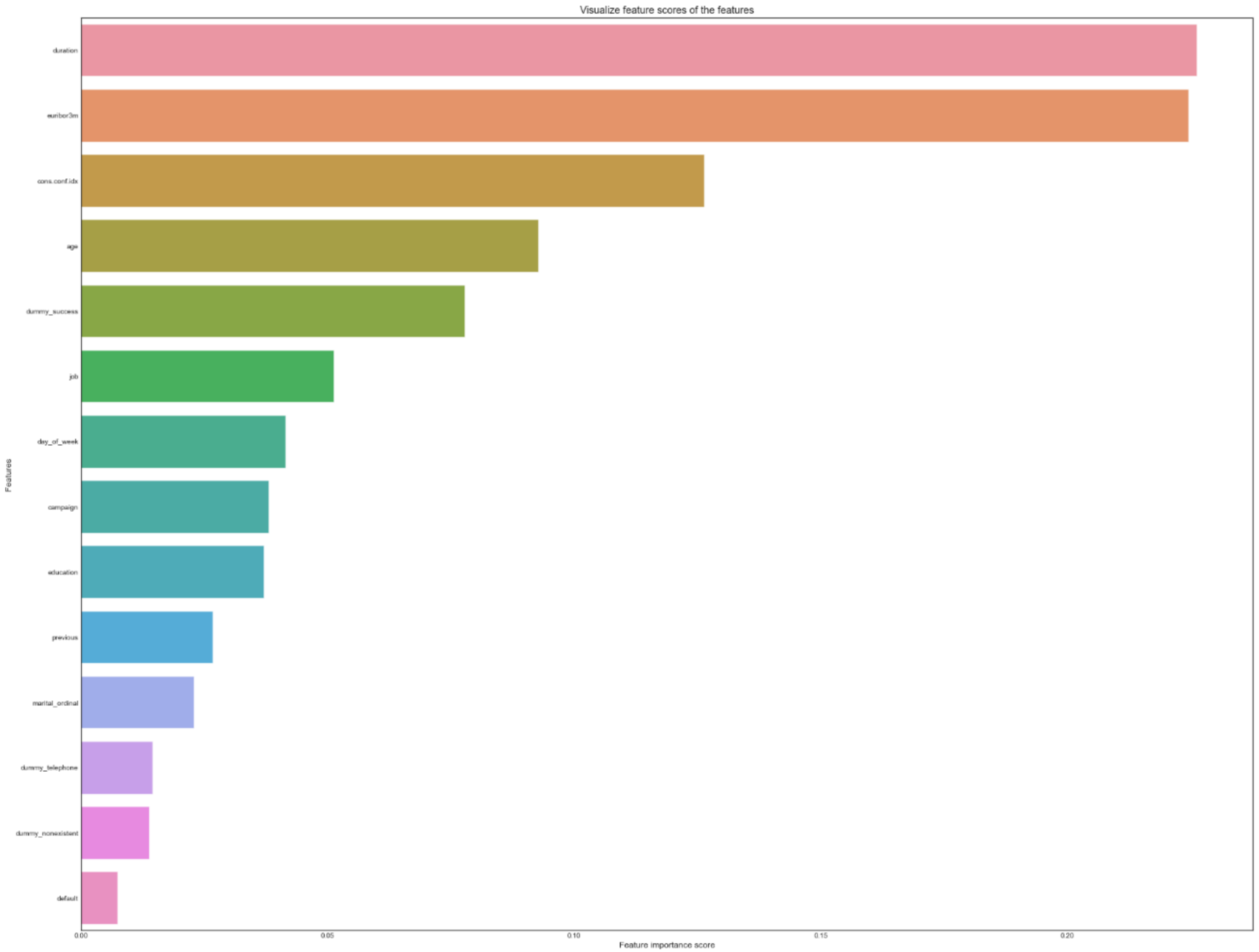
- The AUC score of ‘0.9499’ indicates that our model can distinguish between the positive & negative classes with a 95% chance.

Modelling – with original data



- Random forest classifier has performed the best with 94% accuracy

Top features as per Random Forest Classifier



- Top features are – Duration, Euribor & con.confi.index

Conclusion

Our recommendations:

- Campaign should target the potential customers who are married, having the university degree at least & belongs to the blue-collar occupation
- The campaign will give best results in the month of May & the target duration of the call should be 3 mins.
- Bank should avoid any campaign efforts when the economy is not doing good as it affect the consumer confidence index.

Further Improvements:

- Bank should collect more data on the number of successful subscriptions
- The model can be deployed on cloud to provide the scalable solution.