VIETNAM NATIONAL UNVERSITY, HANOI

INTERNATIONAL SCHOOL

INS3063 – Enterprise Analytics for Data Science

Project report

A project report on Portuguese's bank marketing campaign

Prepare by:

Đỗ Anh Luyện – 19071575

Hanoi, 2021

Contents

Introduction

Data analysis is the process of discovering, interpreting, and communicating meaningful patterns in data. Particularly valuable in areas where a lot of information is recorded, analysis

relies on the simultaneous application of statistics, computer programming, and operations research to quantify performance.

Nowadays, big data analytics is widely used in service industries in general and in the banking industry in particular. Banks hold a huge amount of information about customers such as customers' spending habits and behaviours. Exploiting that huge data will help the bank a lot in its campaigns in both quantity and quality. In this report, we will have access to the huge data source related with direct marketing campaigns of a Portuguese banking institution

The main business problem:

Improve marketing campaign of Portuguese bank by analysing their past making campaign data.

Recommendation which customer to target.

Dataset in used:

To handle the business problem, we will use the bank marketing dataset of a Portuguese banking institution in UCI published in 2014 with 41188 samples, 20 inputs and 1 output. Input variables provide information about the customers, the previous campaign of the bank and others social and economic information. The output is to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

Input variables (Independent variables):

Attribute informations:

Bank client data:

Age (numeric)

Job: type of job (categorical: 'admin.', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown')

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

Education (categorical: 'basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown')

Default: has credit in default? (Categorical: 'no', 'yes', 'unknown')

Related with the last contact of the current campaign:

Contact: contact communication type (categorical: 'cellular', 'telephone')

Month: last contact month of year (categorical: 'jan', 'feb', 'mar',…, 'nov', 'dec')

Dayofweek: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')

Duration: last contact duration, in seconds (numeric).

Housing: has housing loan? (Categorical: 'no', 'yes', 'unknown')

Loan: has personal loan? (Categorical: 'no', 'yes', 'unknown')

Other attributes:

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

Pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)

Previous: number of contacts performed before this campaign and for this client (numeric)

Poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')

Social and economic context attributes:
Emp.var.rate: employment variation rate - quarterly indicator (numeric)

Cons.price.idx: consumer price index - monthly indicator (numeric)

Cons.conf.idx: consumer confidence index - monthly indicator (numeric)

Euribor3m: Euribor 3-month rate - daily indicator (numeric)

Nr.employed: number of employees - quarterly indicator (numeric)

Output variable (dependent variable) (desired target):

y - has the client subscribed a term deposit? (binary: 'yes', 'no')

Some interesting questions:

1. Do older people tend to deposit the most?
2. Which jobs will be having more term deposit? Which job has the highest marketing success rate?
3. Can be concluded that people with higher education level will have more term deposits?
4. What are their marital and economic status (has credit in default, housing loan, personal loan)?
5. What type of contact communication is most effective and they should contact at what time? Does the duration of the last call say anything?
6. Campaign styles play an important role in converting a sale or not?
7. What can we learn from the previous campaign?
8. If the bank has been in contact with that customer before this campaign started, so are they more likely to have a term deposit in this campaign?
9. Do other social and economic context attributes have any effect?

Method used in the report:

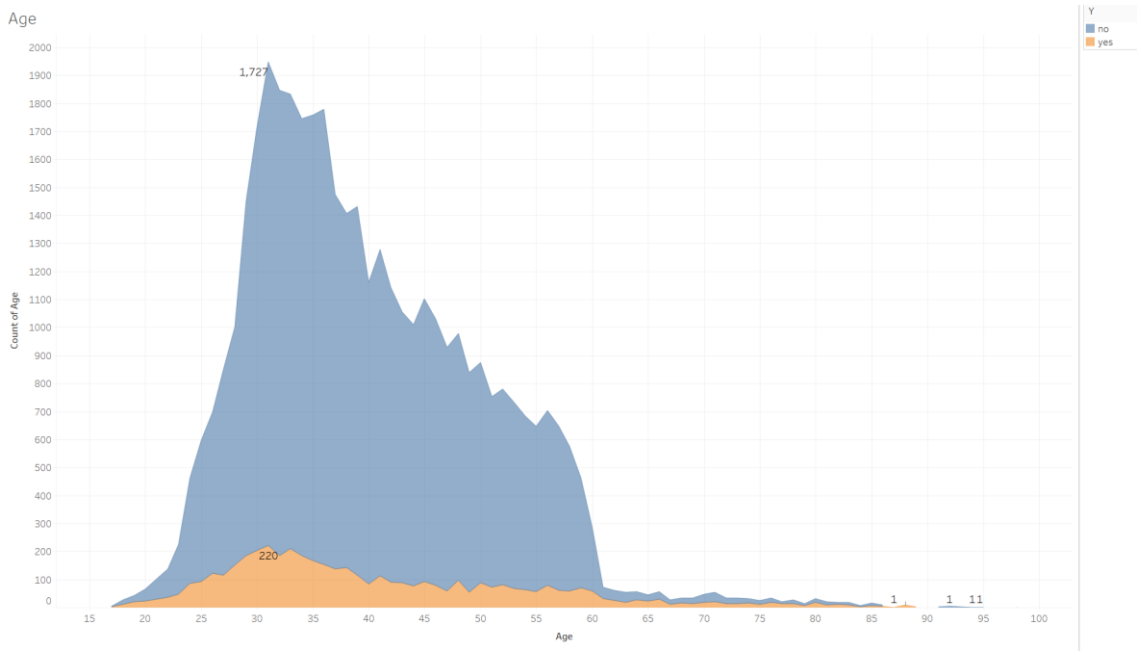
For data analysis, we use both descriptive and inferential analysis. In data pre-processing, data cleaning, handling imbalance data, data encoding and data scaling were in used. Build and train model, applying logistic and random forest for classification model, evaluation based on the confusion matrix, finally selecting the best parameters and evaluation again. Comparing model before and after handling imbalance data, comparison between each model after evaluation, selecting the best model. For visualization, using several charts such as line chart, bar chart, pie chart, area chart, treemaps, mixed chart…

Tools:

Tableau, Knime, R

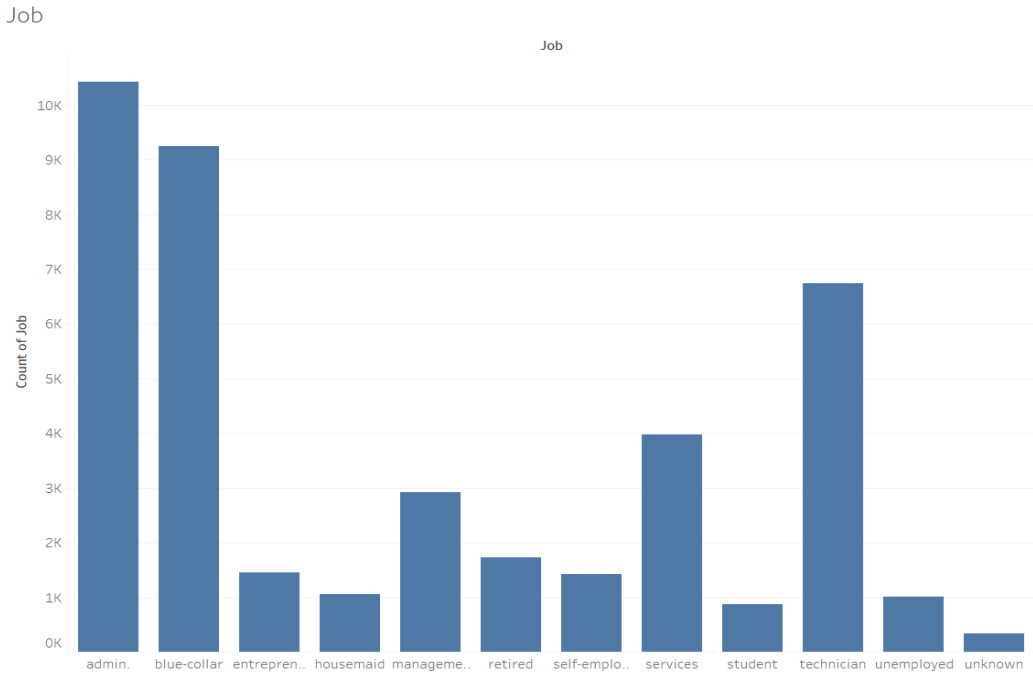
Data Analysis

In the first question: Do older people tend to deposit the most?

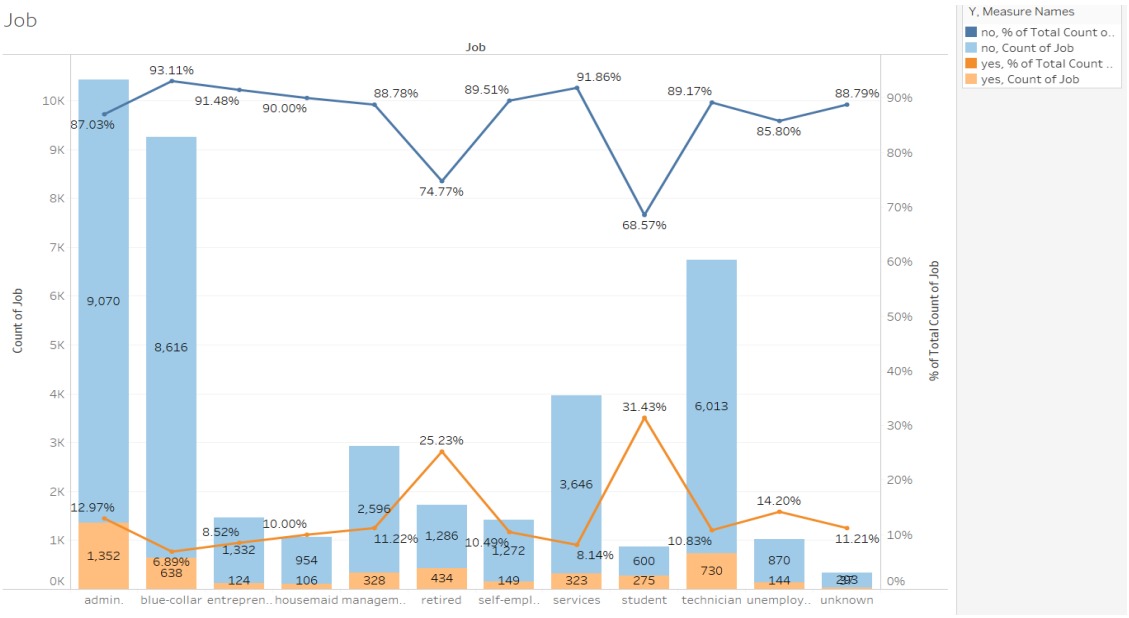


Based on this graph, the dispersion is high across the 17-98 age range. The age of consenting to deposit at the bank is highly concentrated in the age group of 20 – 55, especially from 30 to 40. As the older ages, the number decreases. This shows that older people are not more inclined to deposit but are people of working age.

In the second question: Which jobs will be having more term deposit? Which job has the highest marketing success rate?

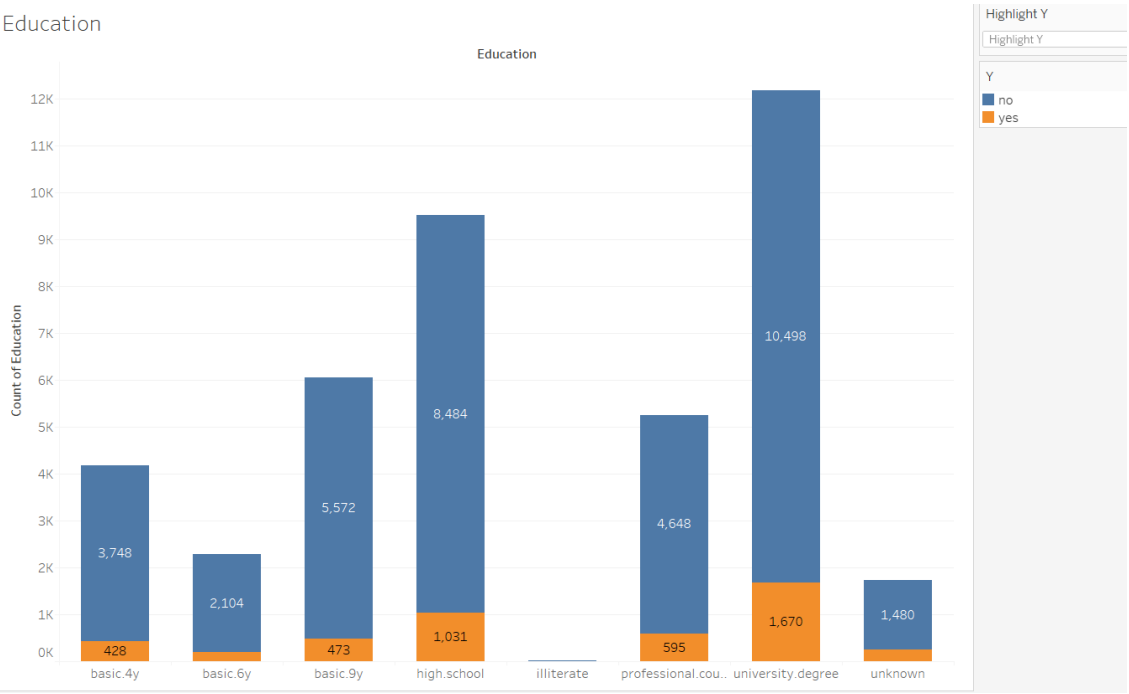


The bank's customers span many different sectors. The plural of customers is admin, blue-collar, services and technician. The minority is housemaid, student and unemployed.



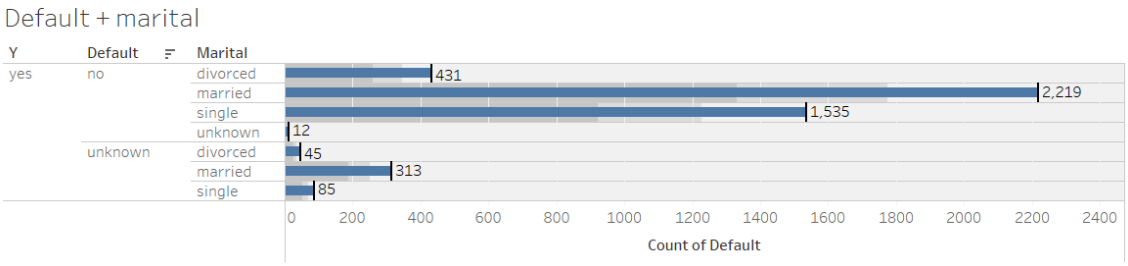
The number of clients spans many industries. However, we can see that the success rate of marketing with retired and student is higher than in other groups, but this is the group with the least number of people.

In the third question: Can be concluded that people with higher education level will have more term deposits?



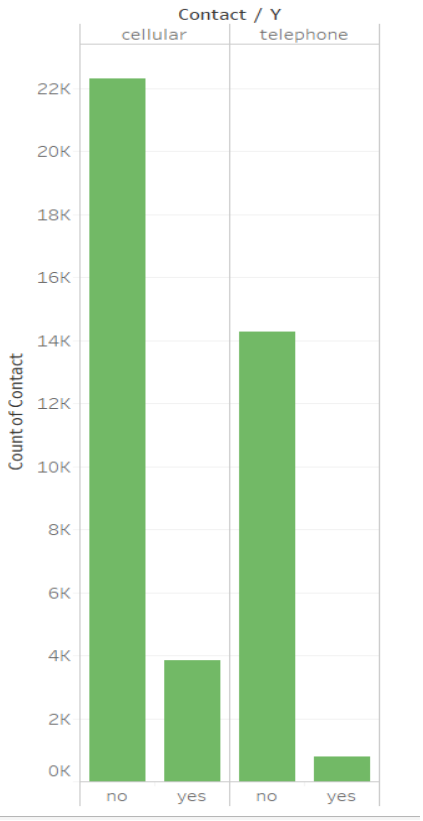
The majority of the bank's customers are high school and college degrees. Therefore, their potential customers are also these two sets of customers (high.school: 1.031 and university.degree: 1.670), so we can also conclude that learners with advanced level are more likely to have a term deposit.

In the fourth question: What are their marital and economic status (has credit in default, housing loan, personal loan)?



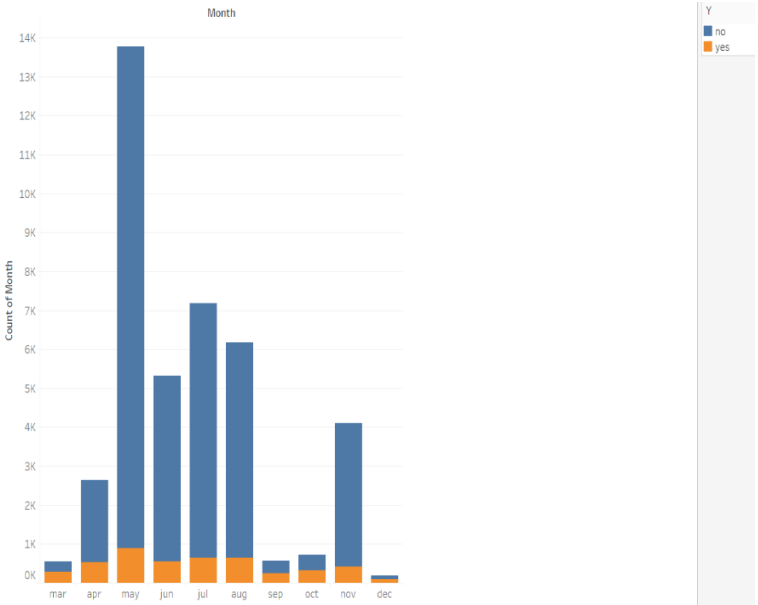
Based on this, we see that customers with no default who are single and married are more likely to deposit than other groups and no one who defaults have a term deposit.

In the fifth question: What type of contact communication is most effective and they should contact at what time? Does the duration of the last call say anything?



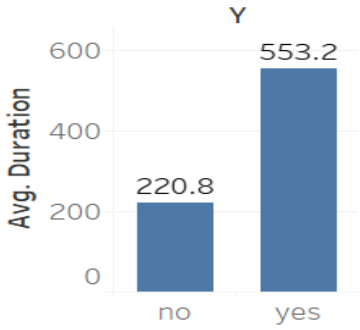
Day Of Week

Y	mon	tue	wed	thu	fri
no	7,667	7,137	7,185	7,578	6,981
yes	847	953	949	1,045	846



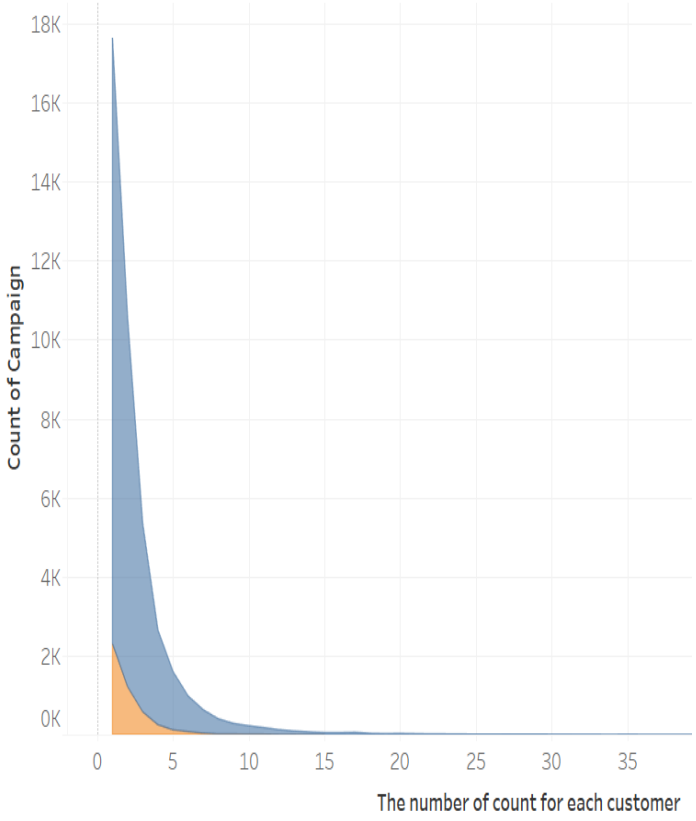
The probability of accepting an offer from a bank by cell phone is higher than telephone and others. Moreover, the number of calls during the week seem to be have not much information with us.

March, September, December & October are great months for the campaign with a 40% or more probability of customers accepting the marketting offer. It generally seems like the last month of each quarter is a fruitful time for campaigning. But the number of calls in those months is very low, we will find out the reason in the last question.



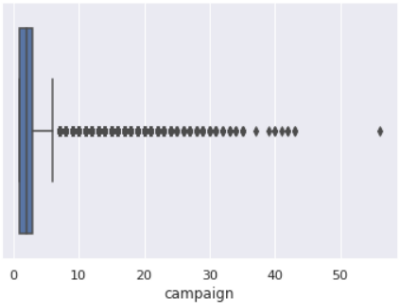
By calculate the average of the last call duration, we can see that the duration of the last call if successful is much higher than defeat, more than 2 times. That means, the duration of the last call is very important to decide whether successful or not.

In the sixth question: Campaign styles play an important role in converting a sale or not?



Blue colour is 'no' class, orange colour is 'yes' class. As we can see, most of success or unsuccess call is known in the first or second call.

Increasing the number of calls to a customer only reduces the chances of this customer accepting the product or even make them feel uncomfortable and some campaigns went for too long should be change.



In the seventh question: What can we learn from the previous campaign?

We have 39.678 values '999' means 39,678 customers did not return from the last call in the previous campaign, so we will not count it in this table below:

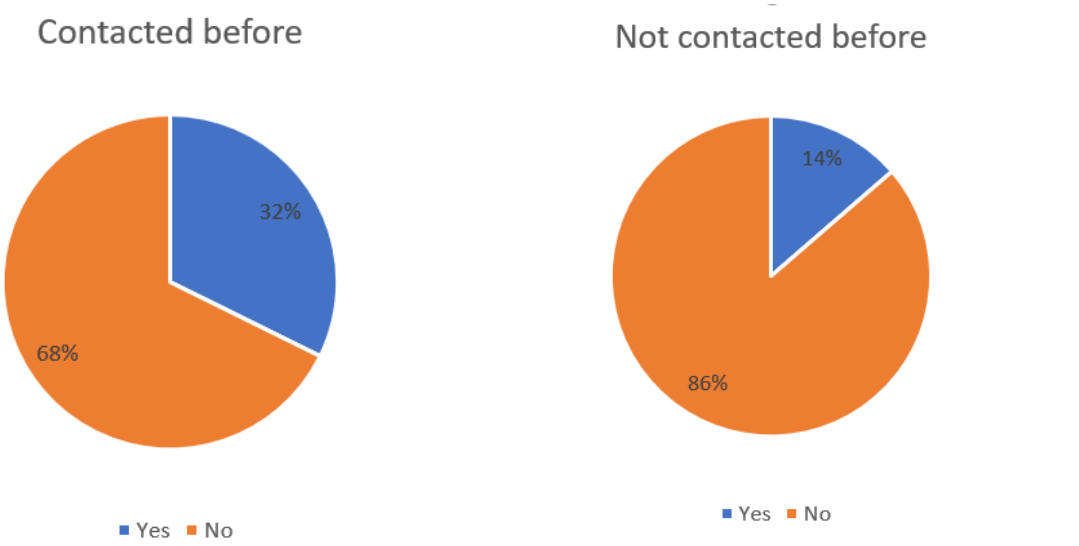
Pdays																									
3	6	4	9	2	7	12	10	5	13	11	1	15	14	8	0	16	17	18	19	22	21	20	25	26	27
439	412	118	64	61	60	58	52	46	36	28	26	24	20	18	15	11	8	7	3	3	2	1	1	1	1

As we can see from the previous campaign, mostly after 3 and 6 days from the last contacted, the bank will know if this customer has approved or not.

In the eighth question: We will find out if the bank has been in contact with that customer before this campaign started, so are they more likely to have a term deposit in this campaign?

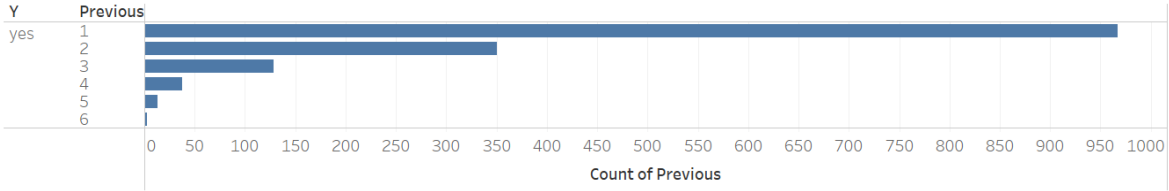
- At first, we find out if the customer has been contacted by the bank before this campaign started, will it affect their decision in this campaign of the bank?

The percentage of the 'yes' class of the total number of people who have been in contact with the bank since before the campaign started:



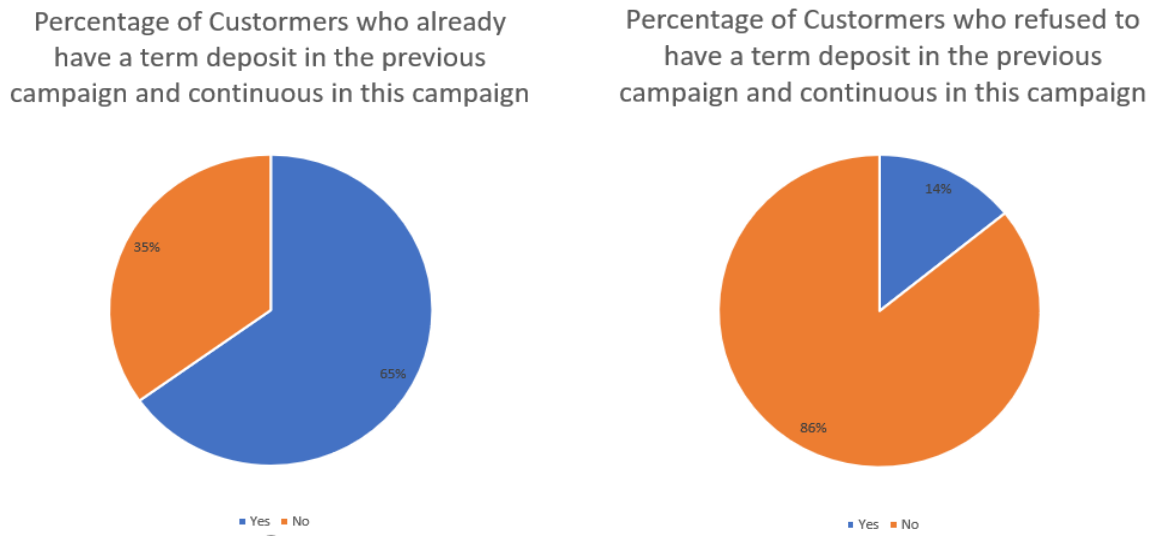
As we can see, with the customer have contacted before this campaign, up to 32% that they will have a term deposit in this campaign, compared to 14% who have not contacted before.

- Next, how many calls should the bank make for that customer before campaign started?



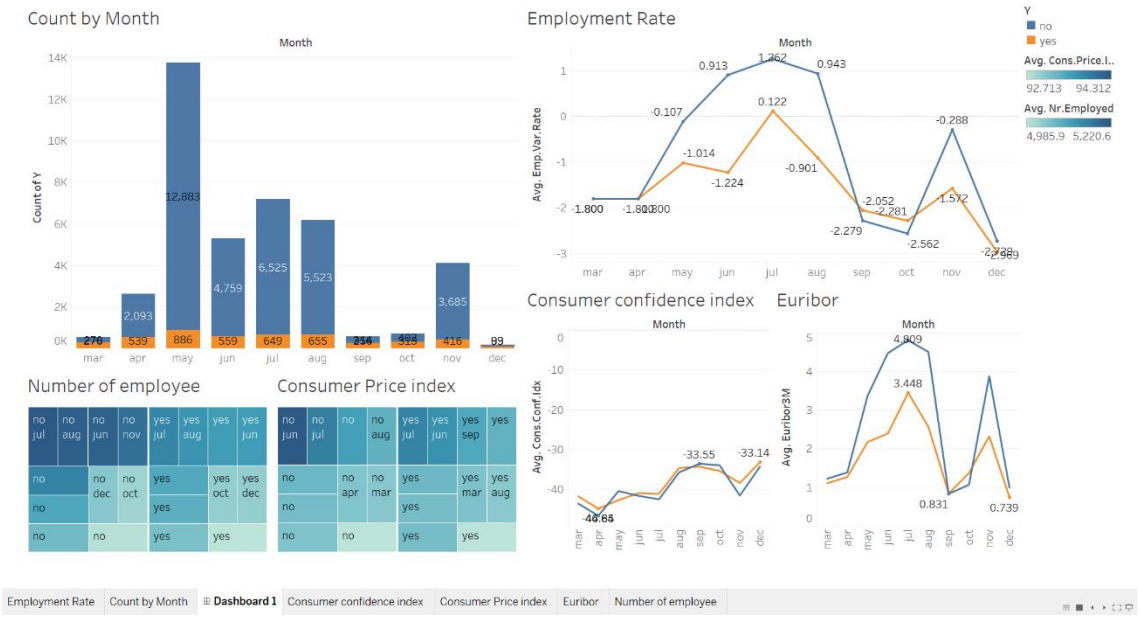
As we can see, from 1 to 3 call should be the best.

- Thirdly, we'll see whether or not the customer from the previous campaign has an impact on the outcome of this campaign?



We can see that people who were had a term deposit in the previous campaign that will have 65% comeback in this campaign. 14% of people who was refused to make a term deposit in the previous campaign that will make in this campaign, much higher than 8.8322% with new people.

In the nineth question: Do other social and economic context attributes have any effect?



In this task, we will create a dashboard to know whether or not the number of customers varies according to the socioeconomic context. The dashboard contents 5 charts, the first one on top left hand side is the count of the number of customers who have a term deposit or not in each month, at the bottom of is the average of number of employee and consumer Price index by month, on the right is 3 line charts which are the average of employment rate, consumer confidence index and Euribor, all chart are compared between the number of customers who have a term deposit or not. 6 months have the highest number of successful are April, May, June, July, August and November. The bank was very acumen in business, all 6 months have highest employment rate and Euribor rate in the whole year. The number of employees, CPI and consumer confidence index have not much change over the year so it will not affect to our business campaign.

Data-Preprocessing

1. Cleaning data:

Luckily, we don't have any missing value. In ‘pdays’ column, number 999 means client was not previously contacted, this number is quite large compared to the rest, we originally planned to replace it with a 0 however the 0 means the customer has agreed to a term deposit since the last call and no need to call back, so we still keep this number.

2. Categorize age values:

- First group from 0 to 22: is the age group still going to school.

- Second group from 23 to 40: is the age group start working and have family.

- Third group 40-60: is the age group that have children and stable job.

- The last over 60: is the retired age group.

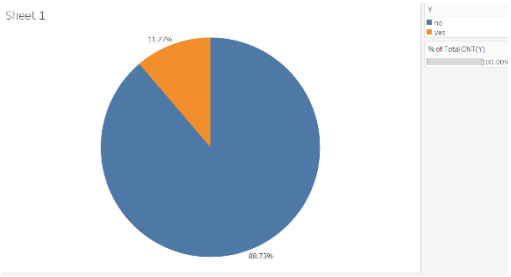
3. Encoding character columns:

Output variable: 1 means ‘yes’ and 0 means ‘no’

Input variables:

"job", "marital", "education", "default", "housing", "loan", "contact", "month", "day_of_week", "poutcome" will be encoded into numbers.

4. Handling imbalanced data:



As we can see, ‘No’ class is almost 90%, so we will use Synthetic Minority Over-Sampling Technique (SMOTE) Sampling in training data set. SMOTE is simple adding exact replicas of minority instances to the main dataset.

5. Scaling data:

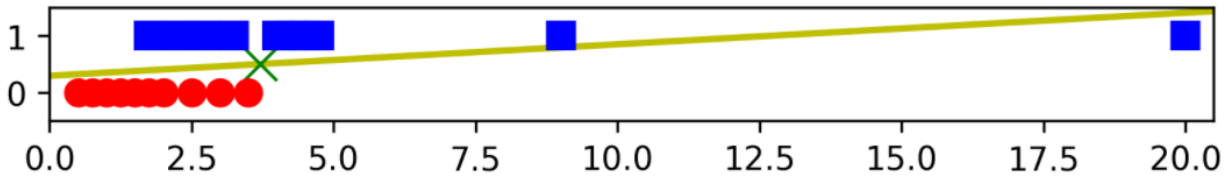
Using min-max scaler(normalization) basically shrink the data interval so that the range is fixed between 0 and 1 (or -1 to 1 if there is a negative value). We use min-max scaler in this case because in our data the distribution is not Gaussian and the standard deviation is very small.

Build and Train model

I. Logistic Regression:

1. Build model:

We use logistic regression because the output probability is blocked for about [0,1] while the linear regression produces the output within [-∞, +∞]

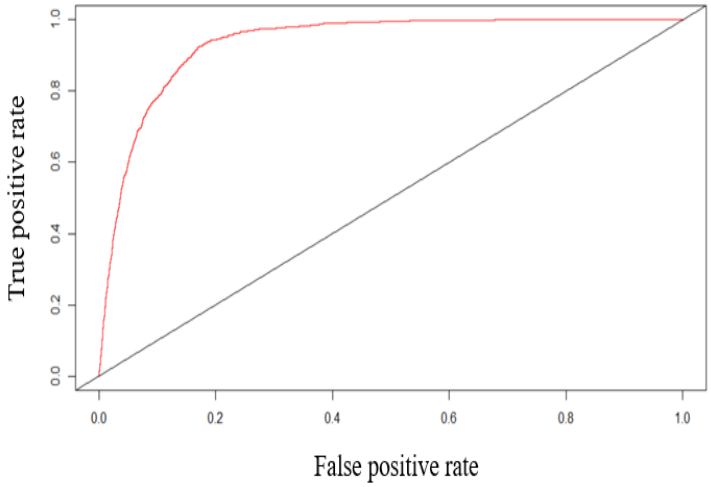


Here is the example if we use linear regression, some blue point will be predicted to be 0 class.

2. Model evaluation:

```
          Reference
Prediction    0    1
         0 7923  150
         1 1262  962
```

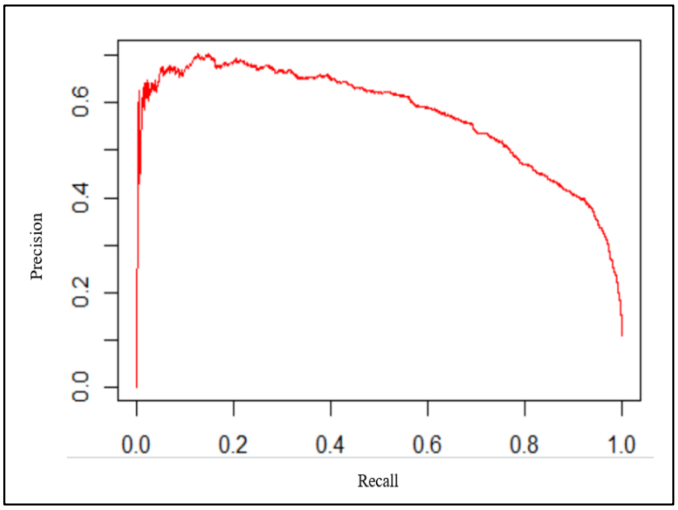
```
"AUC:  0.934646240547025"
```



As TP, TN, FP, FN are 962, 7923, 1262,150; the ROC graph above will show what is the Sensitivity and False positive rate for each cut point corresponding to it. AUC is 0.93 means the model's classification ability is good

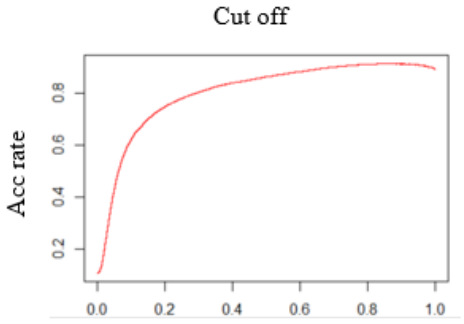
We need calculate precision, recall and f1_score for deeply understanding the result.

```
"precision:  0.432553956834532"
"Recall:  0.865107913669065"
"F1_Score:  0.57673860911271"
```



Precision is 0.43 means in all positive point that model has predicted, there are 43% of true positive while recall is 0.86 that the percentage of true positive among those that are actually positive. F1_Score tells that how good is the classification model (0 is extremely bad, 1 is perfectly). We have f1 equals to 0.57, this is not so good so we will try to improve this later.

Next, we will select the maximum accuracy and the cut off corresponding to that:



With cut off is 0.84, we have highest accuracy (0.91) that means most of output prediction are belong to negative class.

In conclusion, we have a table below:

	Accurancy	f1-score	AUC
columns	Accurancy	f1-score	AUC
values	0.9144411	0.5767386	0.9346462

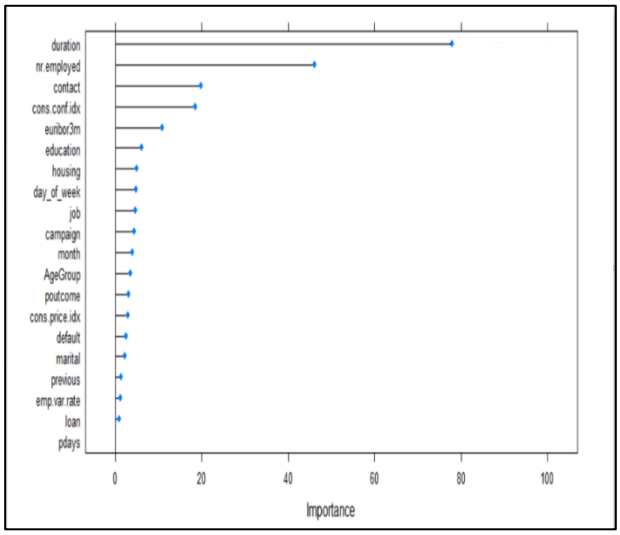
3. Select the best parameter:

The only thing we can do for logistic regression that is finding the best learning rate. However, R has already done it for us. Moreover, glm (logistic regression) package in R doesn't have any parameter for fine-tuning and searching, so we use cross-validation (10-fold) and we see that AUC and Accuracy decrease many so we will assume that we have already built the best logistic regression model.

II. Random forest:

1.Build model

Random forest classifiers are one of the ensembles learning methods for classification with tree-based algorithm. It constructs multiple decision trees (a "forest") at the training time and the output prediction is the class which is the mode of the predictions made by the individual decision trees in the ensemble.

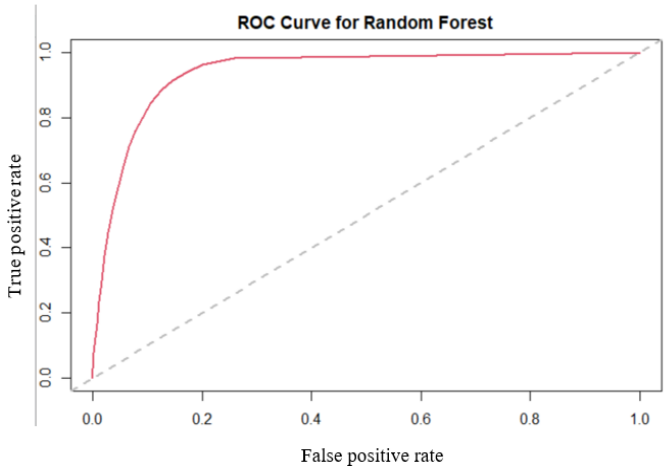


We will choose the number of trees to be 20. And we have the feature importance besides.

Highest importance is duration, followed by nr.employed, contact and consumer price index.

2. Model evaluation:

We will draw ROC curve and calculate AUC to see how good is the classification model




```
          Reference
Prediction    0    1
         0 8736  449
         1  439  673
```

As TP, TN, FP, FN are 673, 8736, 439, 449 and AUC is 0.94

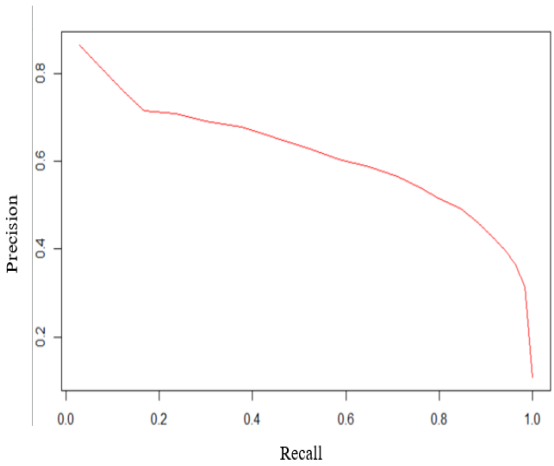
```
"AUC:  0.939132069412516"
```

With AUC up to 0.94, we can assume that random forest did a pretty good job of classifying.

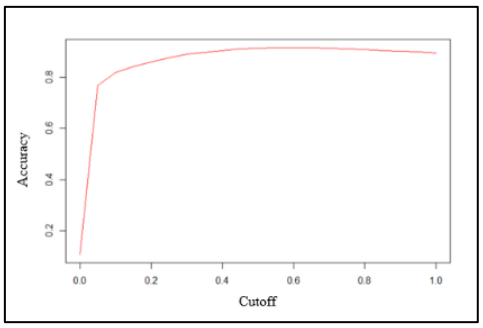
We still need calculate precision, recall and f1_score for deeply understanding the result.

```
"precision:  0.597147950089127"
"Recall:  0.602517985611511"
"F1_Score:  0.599820948970457"
```

Precision is 0.60 means in all positive point that model has predicted, there are 60% of true positive while recall is 0.60 that the percentage of true positive among those that are actually positive. F1_Score tells that how good is the classification model (0 is extremely bad, 1 is perfectly). In this case f1_score is almost 60%, its quite good.



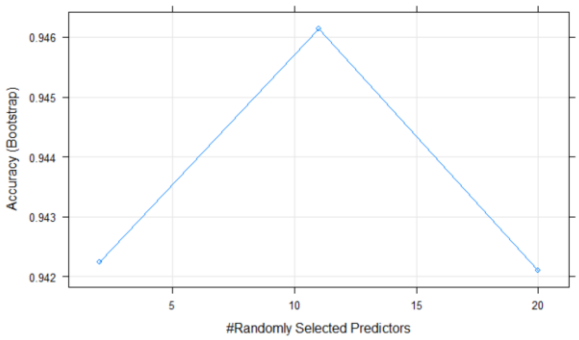
Next, we will select the maximum accuracy and the cutoff corresponding to that:



With cutoff is 0.60, we have highest accuracy (0.915) that means most of output prediction are belong to negative class. In conclusion, we have a table below:

Accurancy	f1-score	AUC
0.9150238	0.5998209	0.9391321

3. Select the best parameter (fine tuning-hyperparameter):




```
Random Forest

52059 samples
   20 predictor
    2 classes: 'first_class', 'second_class'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 46853, 46853, 46854, 46852, 46853, 46853, ...
Resampling results:

  ROC        Sens       Spec
  0.9912295  0.9544638  0.9403956

Tuning parameter 'mtry' was held constant at a value of 2
```

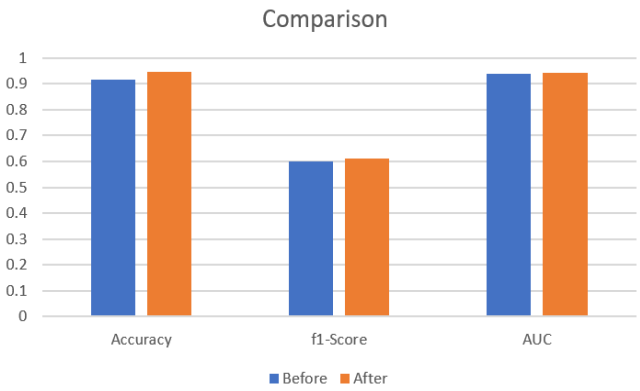
As we can see, the number of trees is mtry is 2 and using cross-validation (10 folds) that give us highest accuracy score.

After fine-tunning, we have table below:

Accurancy	f1-score	AUC
0.9462134	0.6098377	0.9423434

In conclusion, there are many parameters for searching to improve our score, but in this subject, we are not focus on fine-tuning too much, so we make this step as simple as possible.

4. Comparing after fine-tuning:

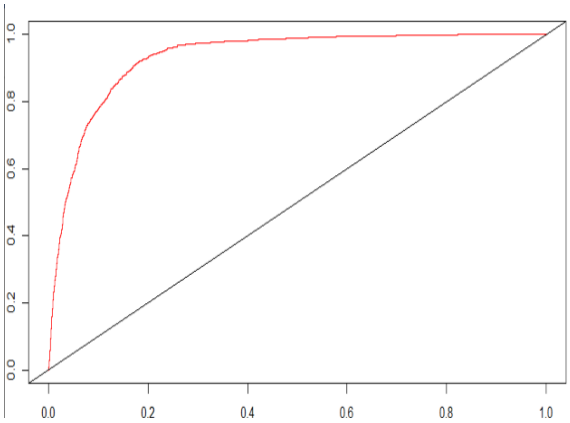


As we can see after fine-tuning, our model become better with all of 3 ways to evaluation.

III. Evaluation before handling imbalanced data:

1. Logistic Regression

ROC Curve for Logistic Regression



```
            Reference
Prediction    0    1
         0 8941  648
         1  244  464
```

```
"precision:  0.655367231638418"
"Recall:  0.41726618705036"
"F1_Score:  0.50989010989011"
"AUC:  0.931390228046214"
```

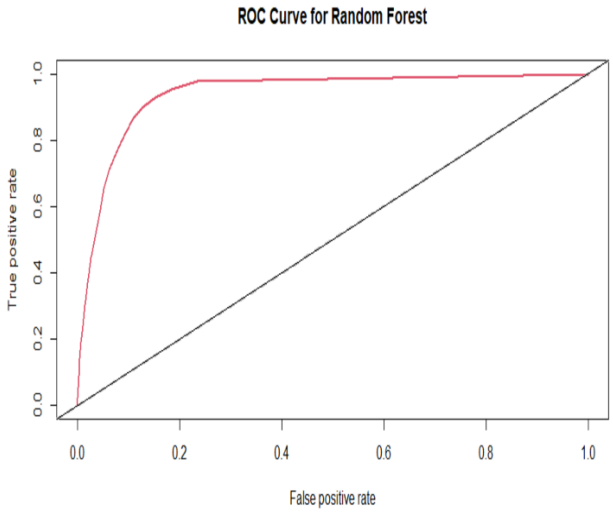
And accuracy and cut off:

```
  accuracy cutoff.9687
 0.9158978   0.4155200
```

In conclusion, we have table below:

Accurancy	f1-score	AUC
0.9158978	0.5098901	0.9313902

2. Random Forest




```
            Reference
Prediction    0    1
         0 8838  516
         1  347  596
```

```
"precision:  0.632025450689289"
"Recall:  0.535971223021583"
"F1_Score:  0.580048661800487"
"AUC:  0.938804079218933"
```

And accuracy and cut off:

```
 accuracy    cutoff
0.9159949 0.6000000
```

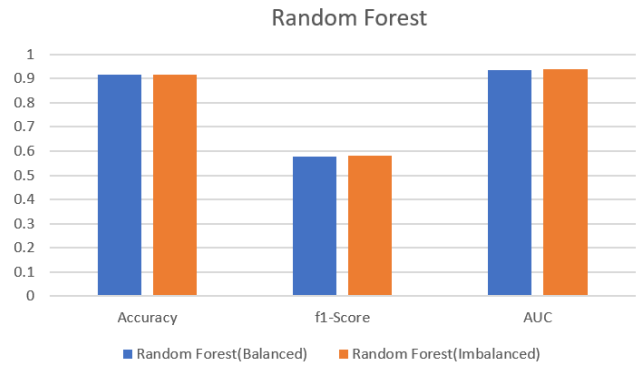
In conclusion, we have table below:

Accurancy	f1-score	AUC
0.9159949	0.5800487	0.9388041

3. Comparing after and before handling imbalanced data:

- Logistic Regression:

After handling imbalanced data, accurancy and AUC seems to be unchanged, however f1-score increased by 7% that means we reduced the imbalance between precision and recall, with a higher f1 score, the better the classification model.

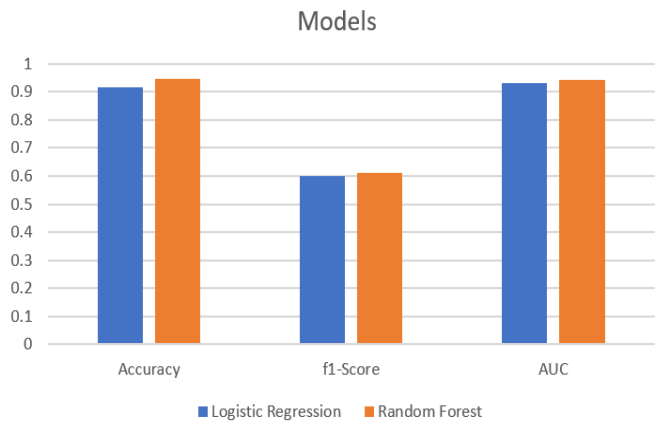


- Random forest:

After handling imbalanced data, accuracy and AUC seems to be unchanged, f1-score increased a little bit by 2%, with a higher f1 score, the better the classification model.

In conclusion, accuracy and AUC unchanged even after handling imbalance data and very little increase in f1-Score.

IV. Comparing between models:



In conclusion, Both Random forest and Logistic Regression give pretty good results. However, Random forest has a little higher accuracy, f1-score and AUC than logistic regression, so Random forest is the best choice.

Summary

Insight after analysis:

- The age group with the most term deposits at banks is from 20 to 55, special focus on age from 30 to 40.
- People whose jobs are admin, blue-collar, management, services, technical are the majority of the bank's customers. However, retired and student have the highest marketing success rate, but the number of this group of people is very small.
- Those who have education level from high school or higher that will have higher the number of term deposit than others.
- Most of the bank's customers are married people, followed by single people and divorced. The bank has no customer who goes default.
- The most success calls are called by cellular; the highest percentage of success calls are in March, September, December & October but the campaign of bank was not focus in those months because those 4 months have lowest employment rate and Euribor so it will have a lot of risk if the bank focus on those months.
- The duration of the last call is very important; as the longer the duration, the higher the success rate.
- From the previous campaign, the bank will know the product (bank term deposit) would be ('yes') or not ('no') subscribed in from the first 1 to 3 calls (mainly in first or second call). The more calls, the lower the success rate.
- From the previous campaign, after 3 or 6 days from the last call, the bank will know if this customer has approved or not.
- If the bank has been contacted with the customer before the campaign started, 32% they will say ‘yes’ with a term deposit. Customers have never contacted before this campaign, just 14% they will say ‘yes’, much lower. And they should call for that customer from 1 to 3 calls before the campaign started.
- With those who have been contacted by the bank since the previous campaign, 65% is the percentage of people who said ‘yes’ from the previous campaign and keep saying ‘yes’ in this campaign; people who said ‘no’ from the previous campaign have 14% saying ‘yes’ in this campaign comparison with 8.83222% is the new customer saying ‘yes’.

Answer the business problem

(recommendation)

From the previous campaign, we can improve some of the following for future campaign of the bank:

- The most successful marketing method is the cellular.
- The most appropriate time to promote the campaign also depends on the social and economic context, should start when employment rate and Euribor high to minimize risk as much as possible.
- In each new campaign, bank must call to invite old customers from the previous campaign whether or not they have agreed to term deposits in the previous campaign.
- Every time before the new campaign starts, the bank should call and get to know the customers who will be the target when the campaign starts.
- If it has been more than 6 days since the last call and no response from that customer, the bank should switch to another customer.
- Banks should focus on convincing customers on the first calls (from 1-3 first calls) because those are the calls that determine success or failure. After those calls, the bank can almost know if they have that customers or not. If it has been more than three calls and the customer still has no intention, the bank should move to others customer.
- It is impossible to know which will be the last call, but in each call, the bank must try to keep the duration as long as possible.

From customer analysis, we can recommend which customer to target for the campaign of the bank:

- The bank should focus on age group from 22 to 55 and specially from 30 to 40. Moreover, student and old people(retied) have highest percentage of success rate but these people are very small in number, it is difficult for the bank to contact. Bank should find as much as possible people in group student and retired.
- People whose jobs are admin, blue-collar, management, services, technical and people who have education level from high school or higher should still be the main object of the campaign.
- Banks don't need to call bankrupt people because they will definitely say ‘no’, instead they should call people who have family first and after that is single people and finally is divorced.

Combined with the use of predictive models, the bank can predict whether the customer they have selected for this campaign will agree to deposit or not, repeatedly selecting potential customers and predicting whether that customer will subscribe or not will bring the best results for the bank's campaign.

For stakeholders such as government, they can rely on the success rate of the bank's campaign to gauge social and economic factors like employment rate or Euribor…

Conclusion

After using several analysis techniques, we have indicated to the bank potential customer groups for their campaign. Moreover, we analysed the previous campaign and pointed out things to keep in mind for the next campaign. Besides, we were able to predict whether the customer would agree to subscribe or not. Finally, we've made suggestions or recommendations to improve the bank's upcoming campaigns. To sum up, we've done all the work a data scientist or analyst can contribute to an organization.