Predicting Customer Responses to Bank's Telemarketing Campaign (2008-2010)

By GILANG SATRIYA UTAMA

Outlines

01 Project Background

✓Define business problems

02 Data Features and Analysis

✓ Bank’s Client Data
✓ Socio-Economic Attributes
✓ Other attributes

03 Insights Discovery

✓Relation between some features
✓Relation between features and target

04 Machine Learning

✓Best algorithm models selection
✓Confusion matrix and classification report analysis
✓Feature Importances

05 Conclusion and Recommendation

✓ Conclusion and Recommendation discovery

Background

At the moment, the spending of marketing in the banking industry is tremendous. In other words this is crucial for banks to optimize marketing strategies and enhance its effectiveness. If we can understand the customer's need, it will lead to more effective marketing plans, smarter product designs and improved customer satisfaction.

Main Objectives :

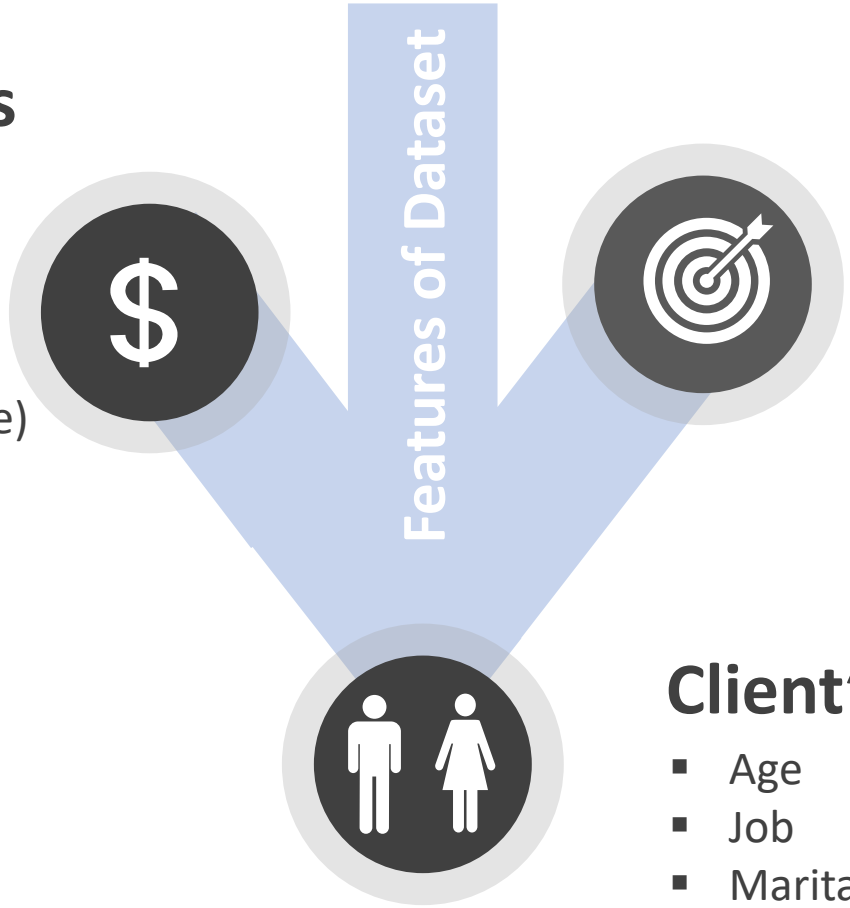
- Enhance the effectiveness of bank's telemarketing campaign

In this project, we will focus on to how we can enable the bank to develop a better understanding of its customer base, predict customer's response to bank's telemarketing campaign and set up a target customer profile for future marketing plans.

Data Features

Socio-Economic Features

- Employment Variation Rate
- Number of Employees
- Consumer Price Index (CPI)
- Consumer Confidence Index (CCI)
- Euribor 3 Month Rate (Interest Rate)



Other Features

- Campaign
- pdays (days passed after last contact)
- Previous (number of contacts performed before current campaign)
- Poutcome (outcome of previous campaign)

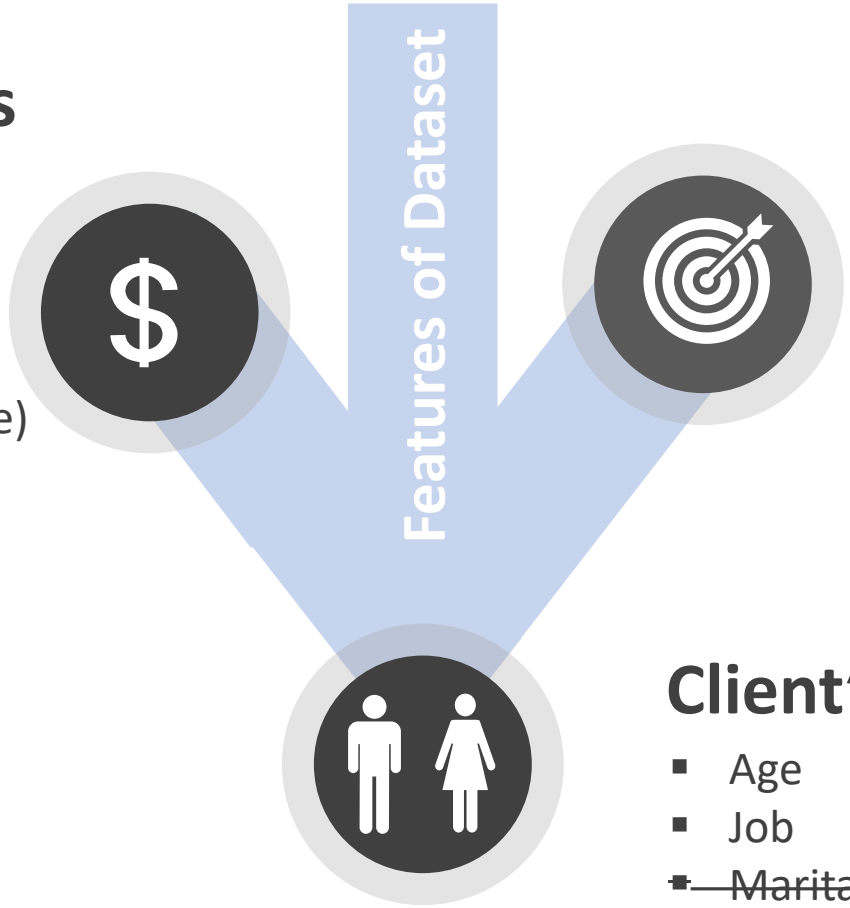
Client's Data

- Age
- Job
- Marital
- Education
- Default
- Housing
- Loan

Data Features After Feature Selection

Socio-Economic Features

- ~~Employment Variation Rate~~
- ~~Number of Employees~~
- Consumer Price Index (CPI)
- Consumer Confidence Index (CCI)
- Euribor 3 Month Rate (Interest Rate)



Other Features

- ~~Campaign~~
- ~~pdays~~ (days passed after last contact)
- ~~Previous~~ (number of contacts performed before current campaign)
- ~~Poutcome~~ (outcome of previous campaign) Missing Values > 50%

Client's Data

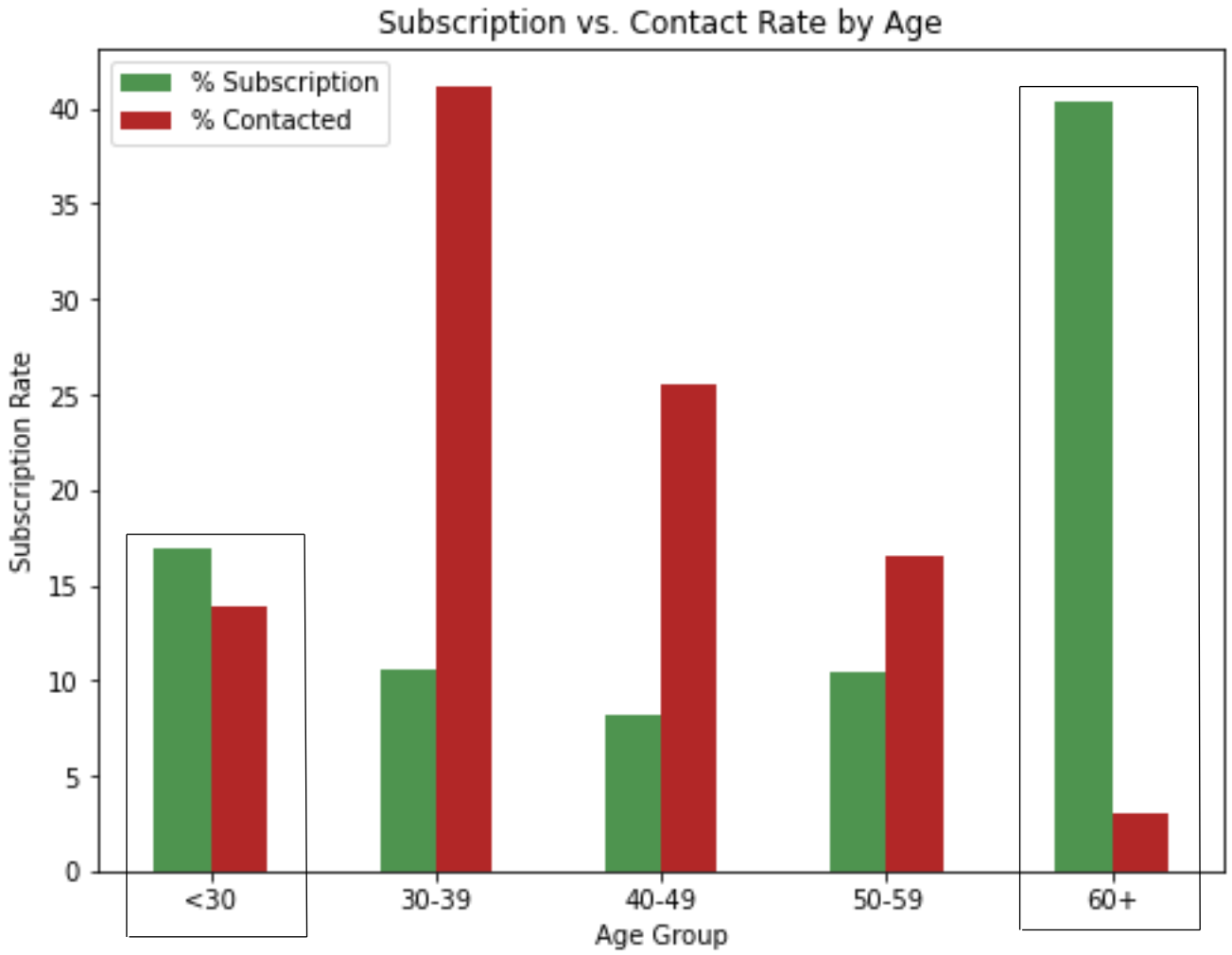
- Age
- Job
- ~~Marital~~
- Education
- Default
- Housing
- Loan

Insights Discovery

Insights: Target young clients and oldest clients instead of middle-aged clients.

- Green vertical bars are indicating that clients with an age of more than 60 or 60+ have the highest subscription rate. About 16.9 or 17% of the campaign subscriptions came from the client between 17 to 29. More than 50% of the campaign subscriptions are contributed by the youngest and the eldest clients.

- Nonetheless, red vertical bars in the chart show that the bank mainly focused its telemarketing campaign efforts on the middle-aged group which is between 30-49 years old which returned lower subscription rates than the younger and older groups. Thus, to make the telemarketing campaign more effective, the bank should target the younger and older clients in the next campaign.

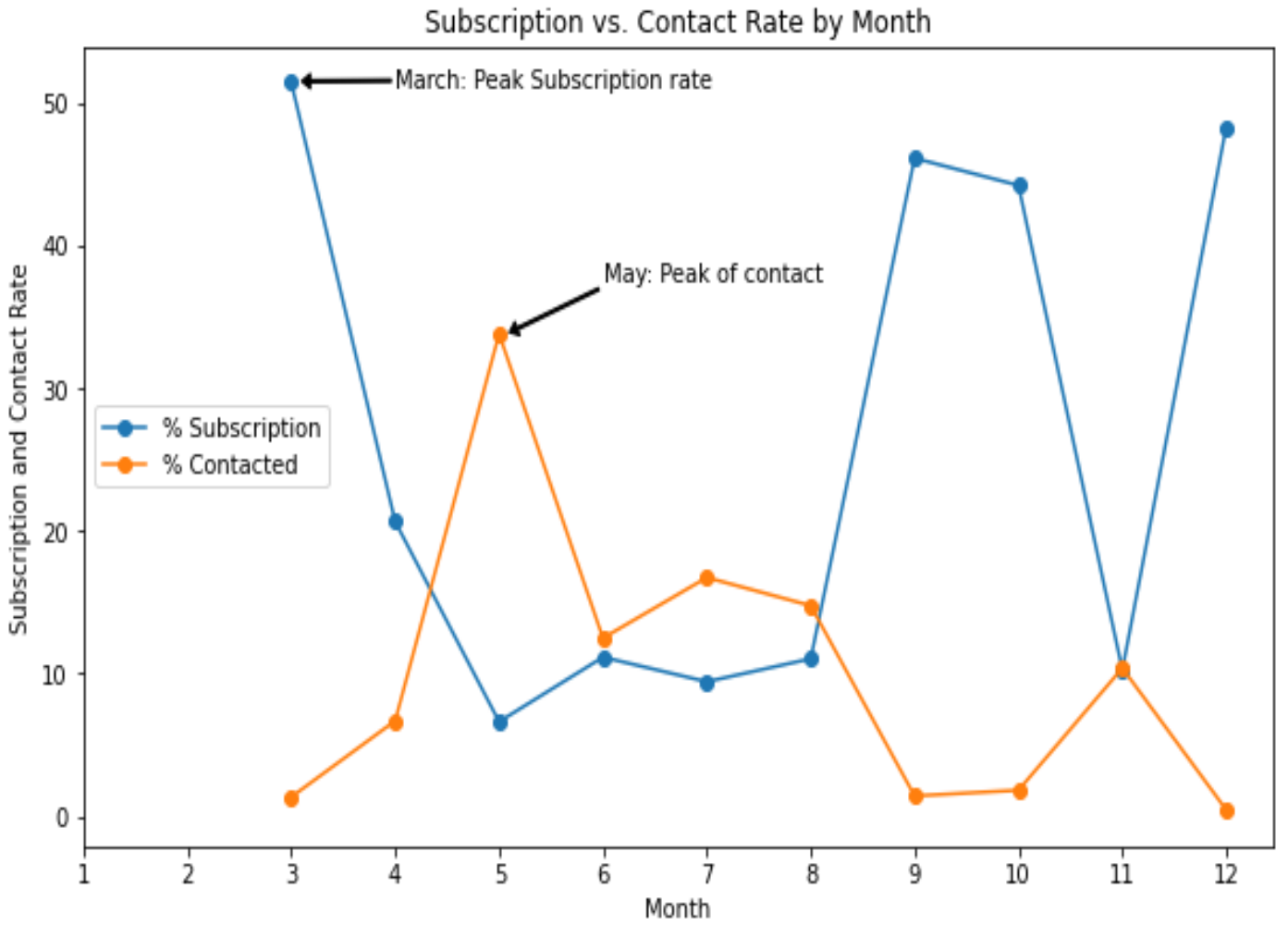


Insights Discovery

Insights: Initiate the telemarketing in fall or spring.

- The bank contacted the most clients between May and August. The highest contact rate is around 33%, which happened in May, while contact rate is closer to 0 in March, September, October and December.

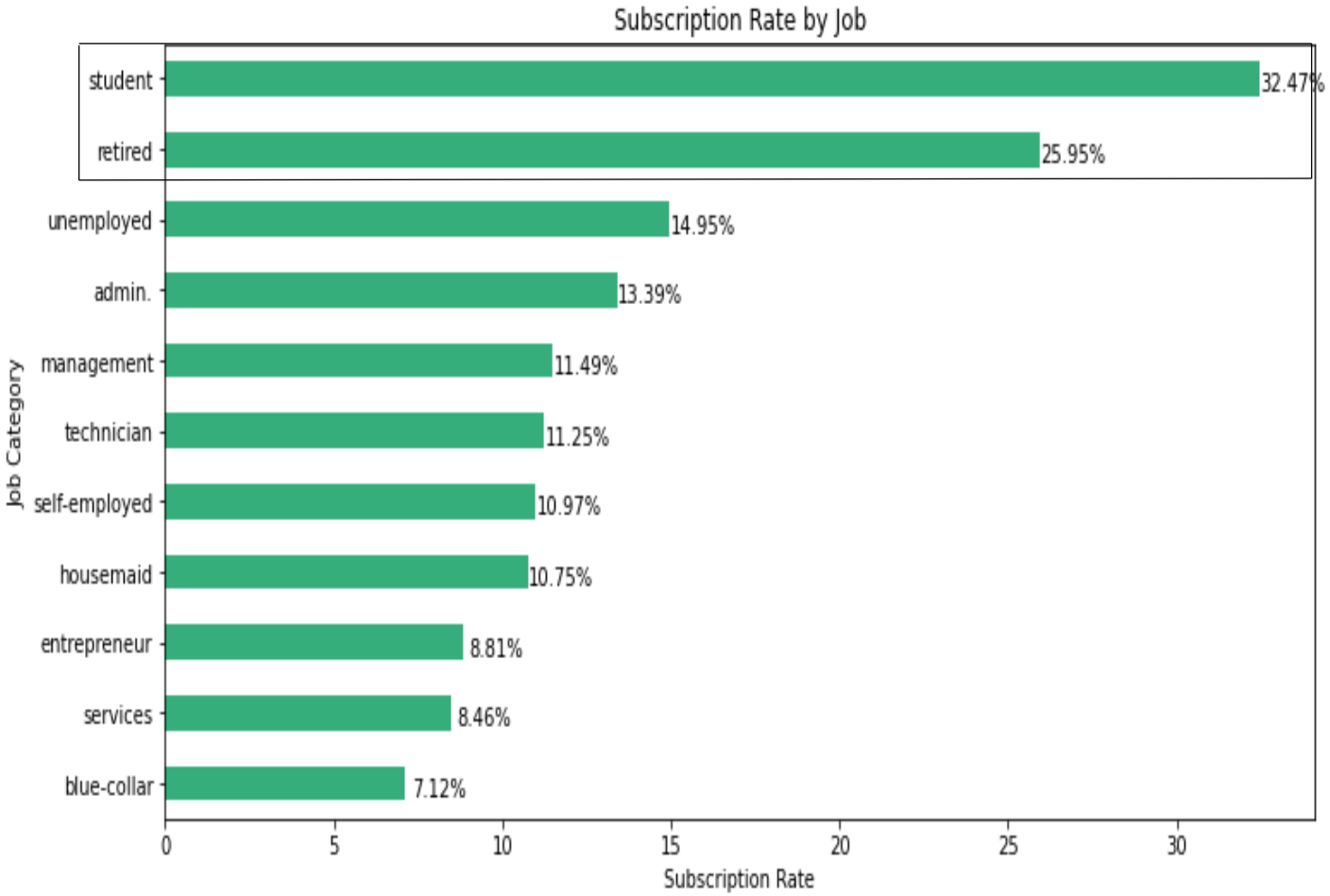
- Nonetheless, the subscription rate showed a different trend. The highest subscription rate occured in March, which is over 50% and all subscription rates in September, October and December are over 40%.



Insights Discovery

Insights: target students and retired clients

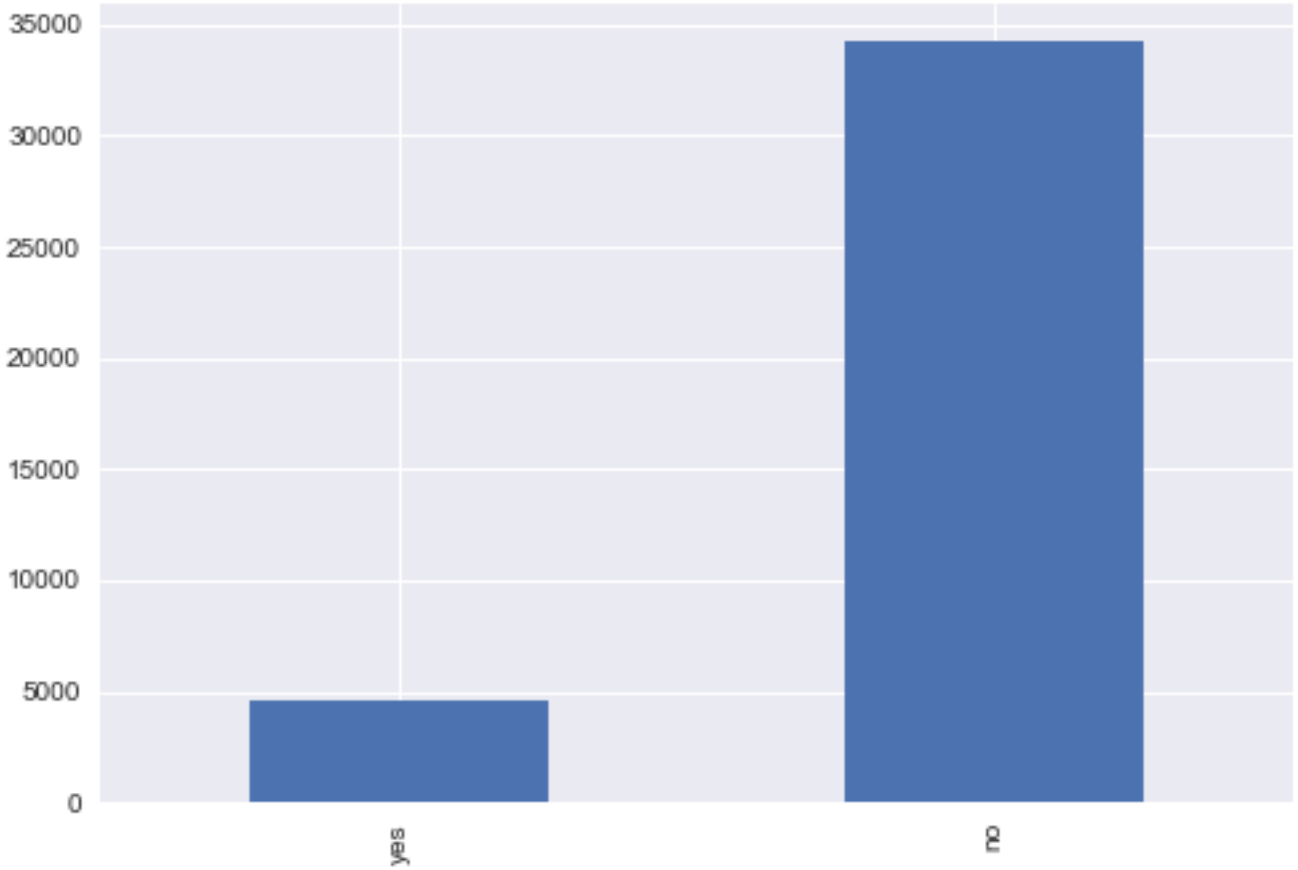
As noted from the horizontal bar chart, students and retired clients account for more than 50% of subscription, which is consistent with the previous finding of higher subscription rates among the younger and older.



Machine Learning and Interpretation

THE DISTRIBUTION OF THE TARGET IS HIGHLY UNBALANCED! BECAUSE THE RATIO OF THE RESPONSE OF CLIENTS THAT DECLINE THE CAMPAIGN OFFER AND CLIENT THAT ACCEPT CAMPAIGN ARE NEAR 9:1 (88.9% NO AND 11.1% YES)

The Distribution of Target



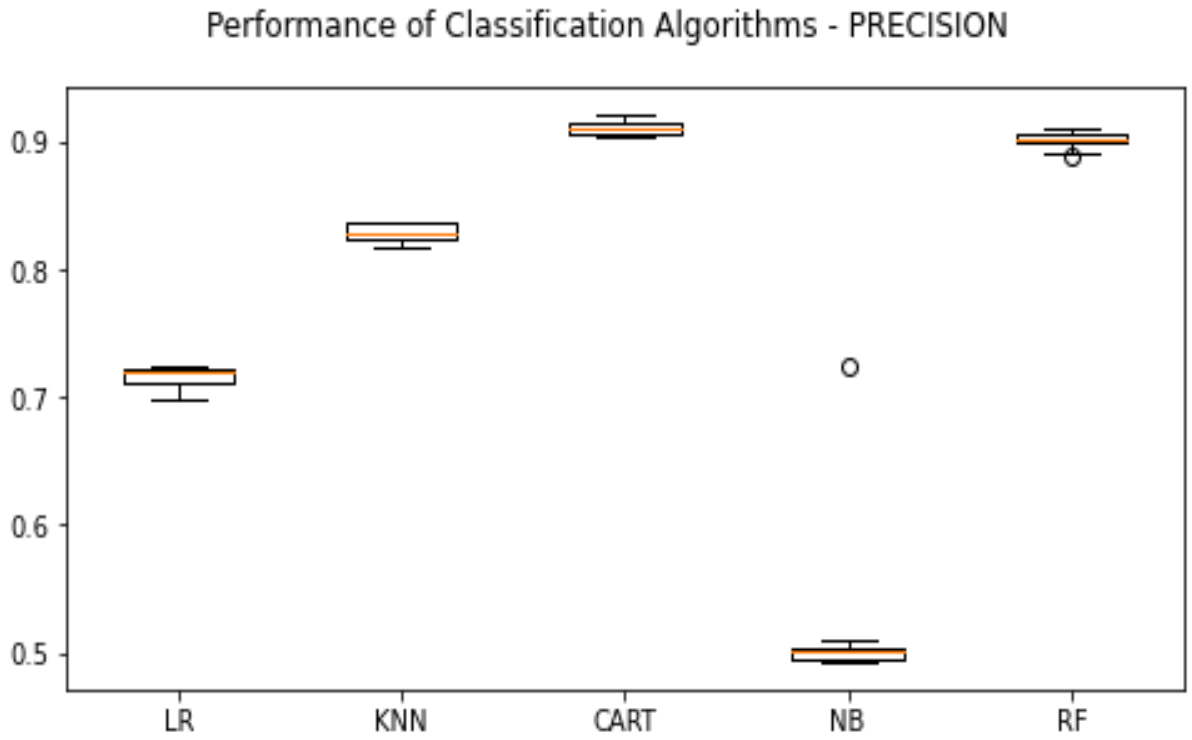
Number of Clients that:

- Decline the offer = 32.205
- Accept the offer = 4.529

To overcome the unbalance problem we done the Oversampling with Smote

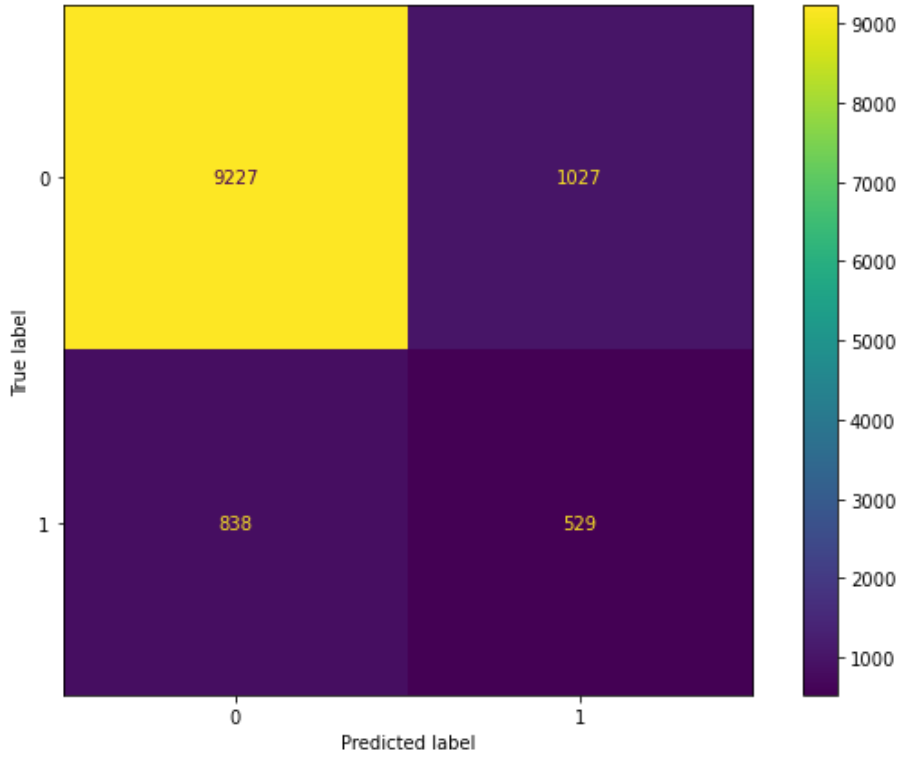
WE COMPARE CLASSIFICATION ALGORITHMS OF BASE MODELS WITH LOGISTIC REGRESSION(LR), K-NEAREST NEIGHBORS(KNN), DECISION TREE(CART), NAIVE BAYES(NB) AND RANDOM FOREST(RF) BY COMPARING ITS PRECISION

MODEL	PRECISION
LR	71%
KNN	82%
RF	90%
NB	52%
CART	91%



WE DECIDE TO CHOOSE 3 MODELS WITH HIGHEST PRECISION TO TUNING AND CHECK IF THE PRECISION STILL SAME WITH TEST DATA.

PRECISION IS SIMPLY THE RATIO OF CORRECT POSITIVE PREDICTIONS OUT OF ALL POSITIVE PREDICTIONS MADE, OR THE ACCURACY OF MINORITY CLASS PREDICTIONS. THE KNN ALGORITHM ACHIEVES AN PRECISION OF 34%, SUGGESTING LOW LEVEL OF STRENGTH OF THIS MODEL TO CLASSIFY THE CUSTOMER RESPONSE GIVEN ALL THE DEFINED CUSTOMER FEATURES.

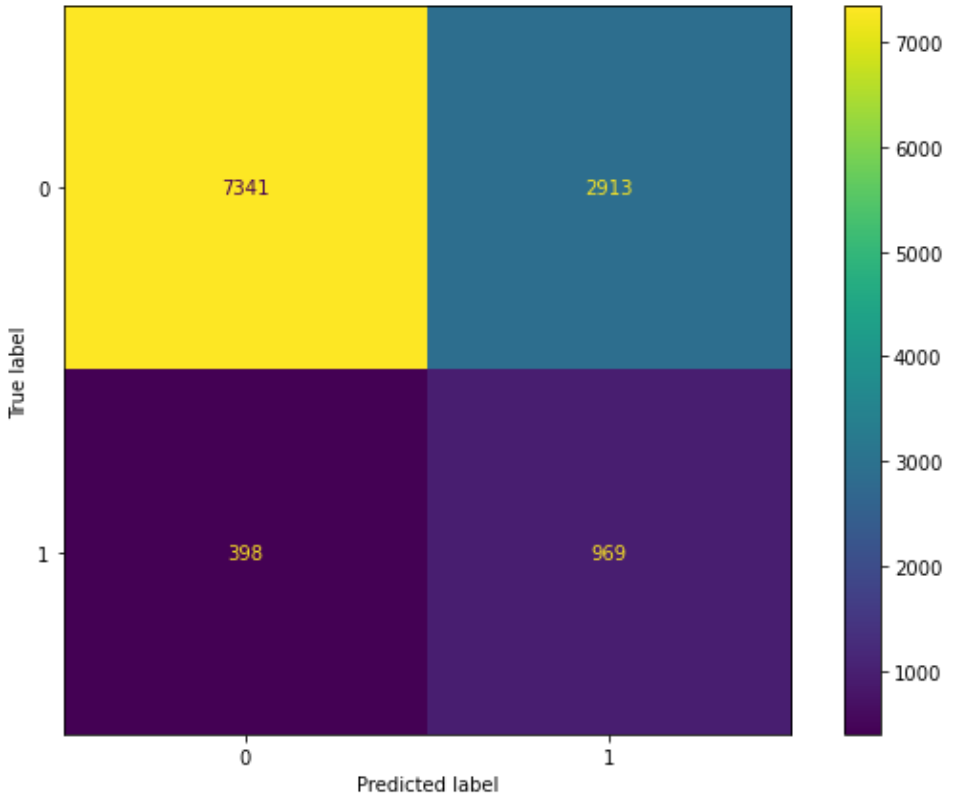


KNN CLASSIFICATION REPORT

	precision	recall	f1-score	support
0	0.92	0.90	0.91	10254
1	0.34	0.39	0.36	1367
accuracy			0.84	11621
macro avg	0.63	0.64	0.64	11621
weighted avg	0.85	0.84	0.84	11621

NONETHELESS, THE RESULT OF ACCURACY SCORE CAN POSSIBLY YIELD MISLEADING RESULT IF THE DATA SET IS UNBALANCED, BECAUSE THE NUMBER OF OBSERVATIONS IN DIFFERENT CLASSES LARGELY VARY. SO WE DON'T USE ACCURACY IN THIS PROJECT, INSTEAD WE USE PRECISION.

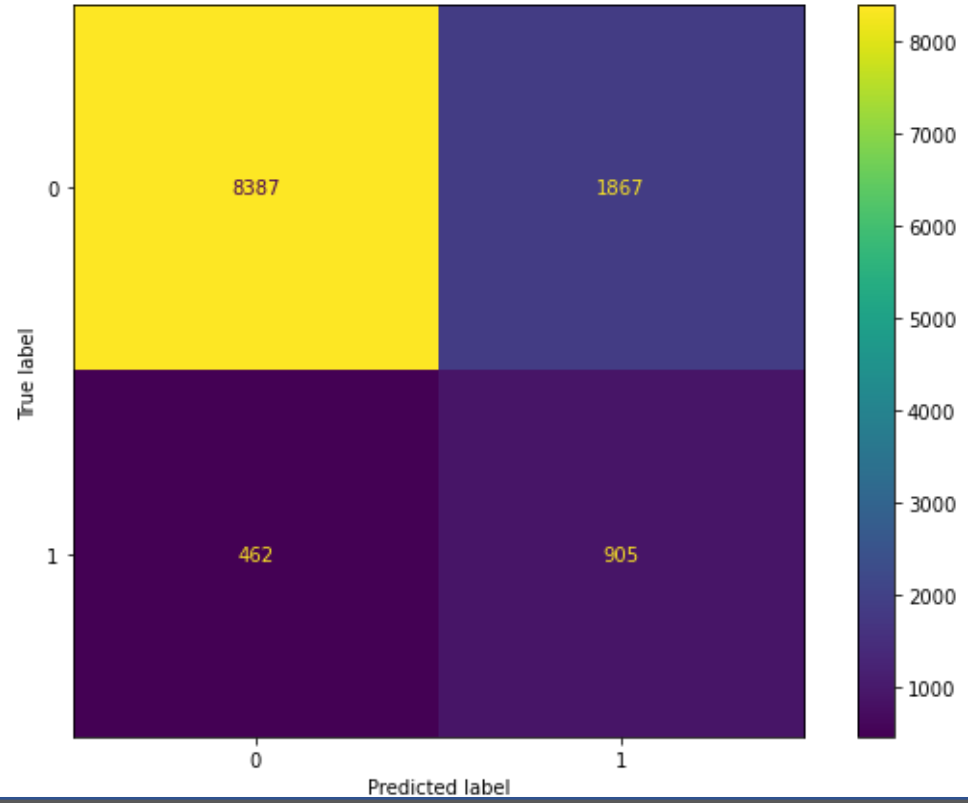
THE RECALL FROM TUNED-LOGISTIC REGRESSION IS BETTER WHICH IS 71% THAN KNN(39%) BUT THE PRECISION IS LOWER THAN KNN, BECAUSE WE PRIORITIZE THE PRECISION OVER RECALL IN THIS PROJECT (WE DON'T WANT THE MODEL PREDICT THE CLIENT AS CUSTOMER THAT ACCEPTED CAMPAIGN BUT ACTUALLY NOT) OR HIGH FALSE POSITIVE. SO WE BETTER CHECK IF THE PRECISION OF TUNED-RANDOM FOREST BETTER THAN THIS MODEL(LR).



LOGISTIC REGRESSION CLASSIFICATION REPORT

	precision	recall	f1-score	support
0	0.95	0.72	0.82	10254
1	0.25	0.71	0.37	1367
accuracy			0.72	11621
macro avg	0.60	0.71	0.59	11621
weighted avg	0.87	0.72	0.76	11621

IN GENERAL, THE REPORT SHOWS THAT RF MODEL HAS GREAT PREDICTIVE POWER TO IDENTIFY THE CUSTOMERS WHO WOULD NOT SUBSCRIBE TO THE TERM DEPOSIT. BECAUSE OF THE LIMITED NUMBER OF CLIENTS ACCEPTING THE TERM DEPOSIT. IN THIS CASE WE AIM FOR HIGH PRECISION AND THE HIGHEST PRECISION IS 33%, WE MAY NEED TO FIND OTHER MODELS OR USING NEW DATASET WITH MORE BALANCED RATIO.



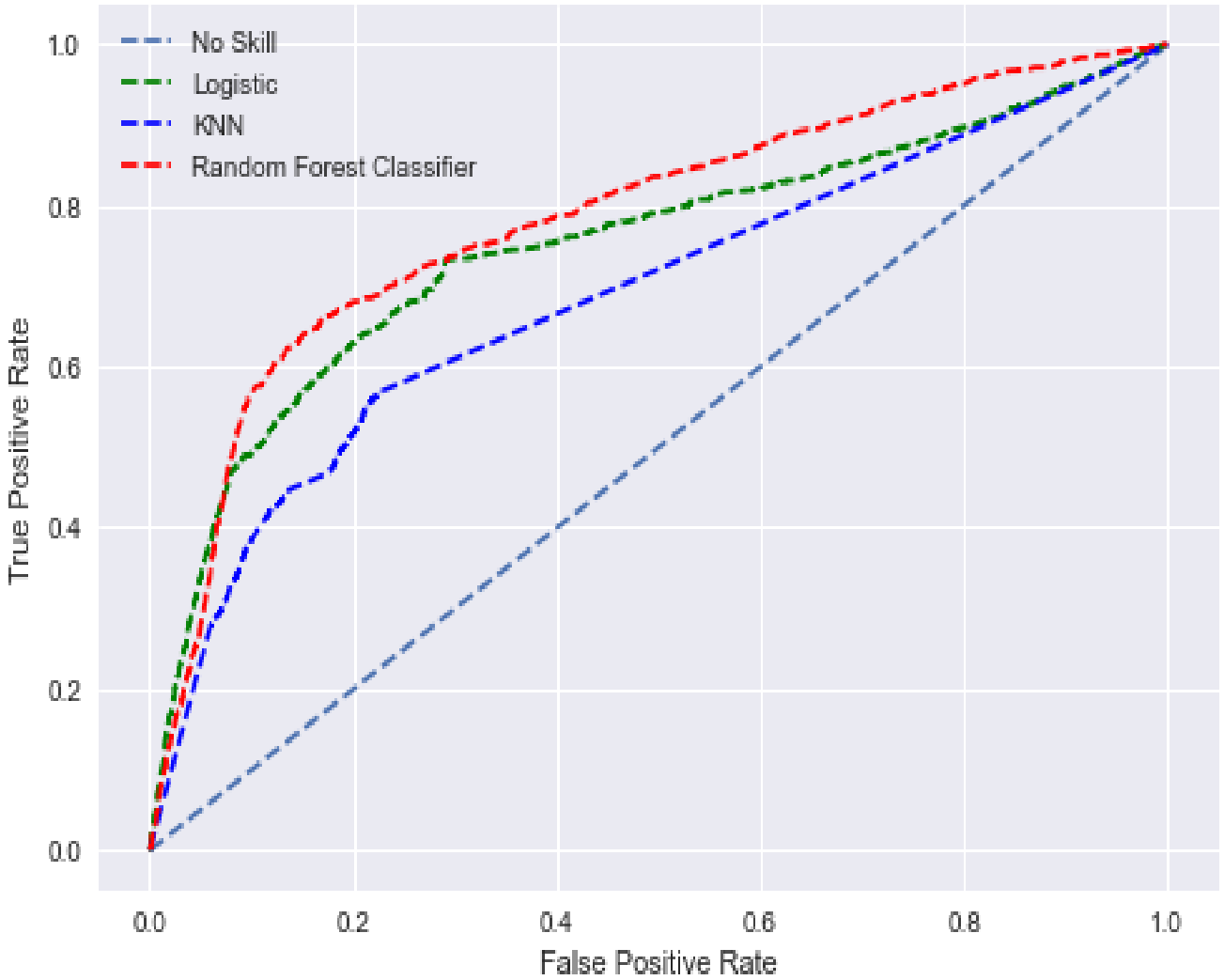
RANDOM FOREST CLASSIFICATION REPORT

	precision	recall	f1-score	support
0	0.95	0.82	0.88	10254
1	0.33	0.66	0.44	1367
accuracy			0.80	11621
macro avg	0.64	0.74	0.66	11621
weighted avg	0.87	0.80	0.83	11621

EVEN THOUGH WE ALREADY DONE OVERSAMPLING WITH SMOTE FOR X_TRAIN AND Y_TRAIN, THE MODEL STILL PREDICTED POORLY ON TEST DATASET. F-1 SCORE CAN BE CONSIDERED AS GOOD METRICS BECAUSE WE NEED HIGH PRECISION BUT THE DATA IS IMBALANCED, SO WE TAKE THE RANDOM FOREST MODEL WITH F-1 SCORE OF 44%.

IT IS EVIDENT FROM THE PLOT THAT THE AUC FOR THE RANDOM FOREST ROC CURVE IS HIGHER THAN OVER THE KNN ROC CURVE OR LOGISTIC REGRESSION ROC CURVE.

ROC-AUC CURVE



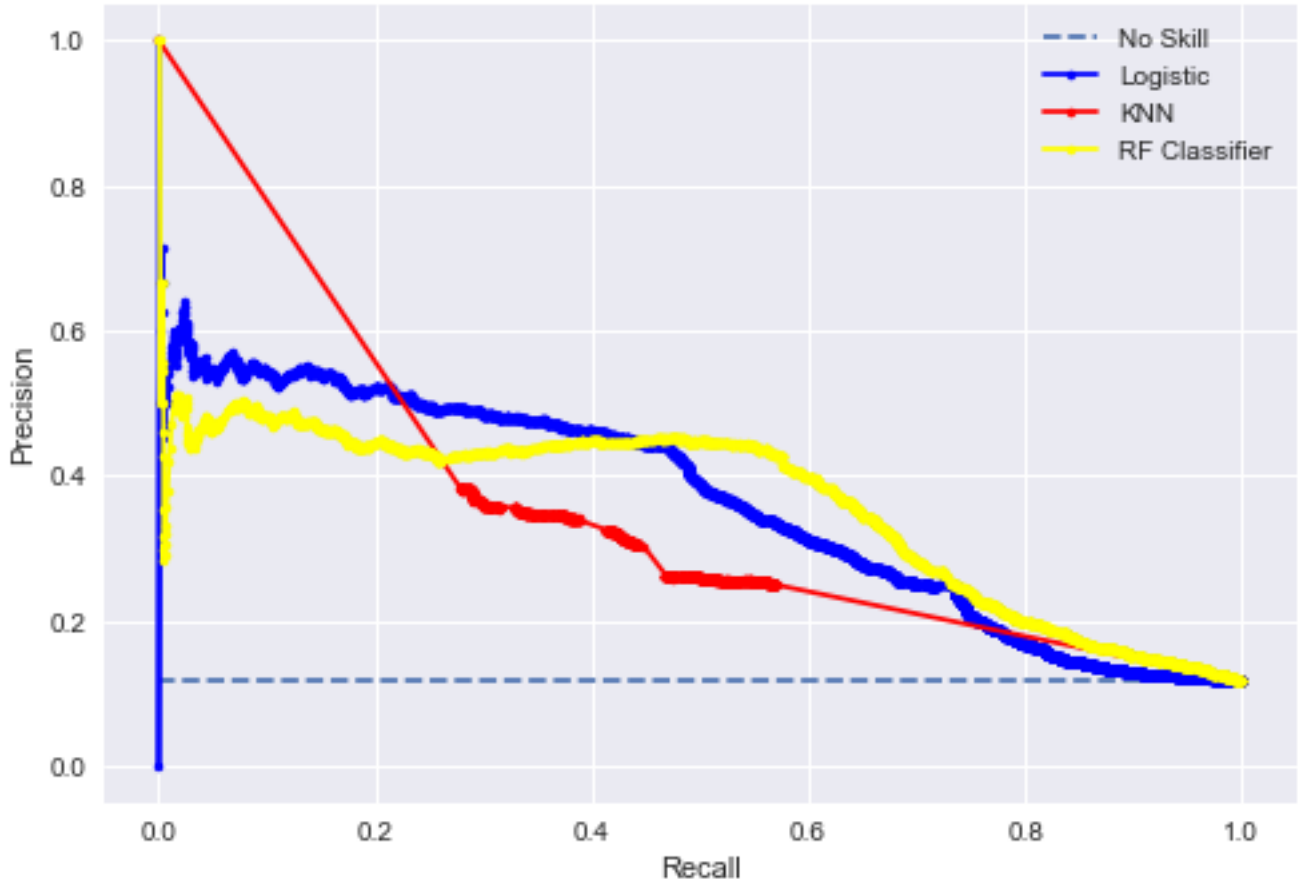
MODELS	ROC-AUC
NO SKILL	0.50
LOGISTIC REGRESSION	0.74
KNN	0.68
RANDOM FOREST	0.78

Therefore, we can say that Random Forest did a better job of classifying the positive class in the dataset. Also because the AUC score of random forest is 0.78 we can consider that the random forest model quite good in ROC-AUC **(IF THE DATASET IS BALANCED**).

We just want to check the AUC-ROC, but because we have imbalanced dataset we will use Precision-Recall Curves instead AUC-ROC.

THE AUC VALUE BETWEEN THESE MODELS ARE NEAR WHICH ARE IN RANGE 0.36. ON TOP OF THAT THE PRECISION OF THE MODELS ARE QUITE LOW TOO. SO IN THE END, WE DECIDE TO USE THE BEST MODEL WITH HIGHEST F-1 SCORE WHICH IS RANDOM FOREST WITH VALUE 44%.

PRECISION RECALL-AUC CURVE



MODELS	PR-AUC
LOGISTIC REGRESSION	0.361
KNN	0.362
RANDOM FOREST	0.356

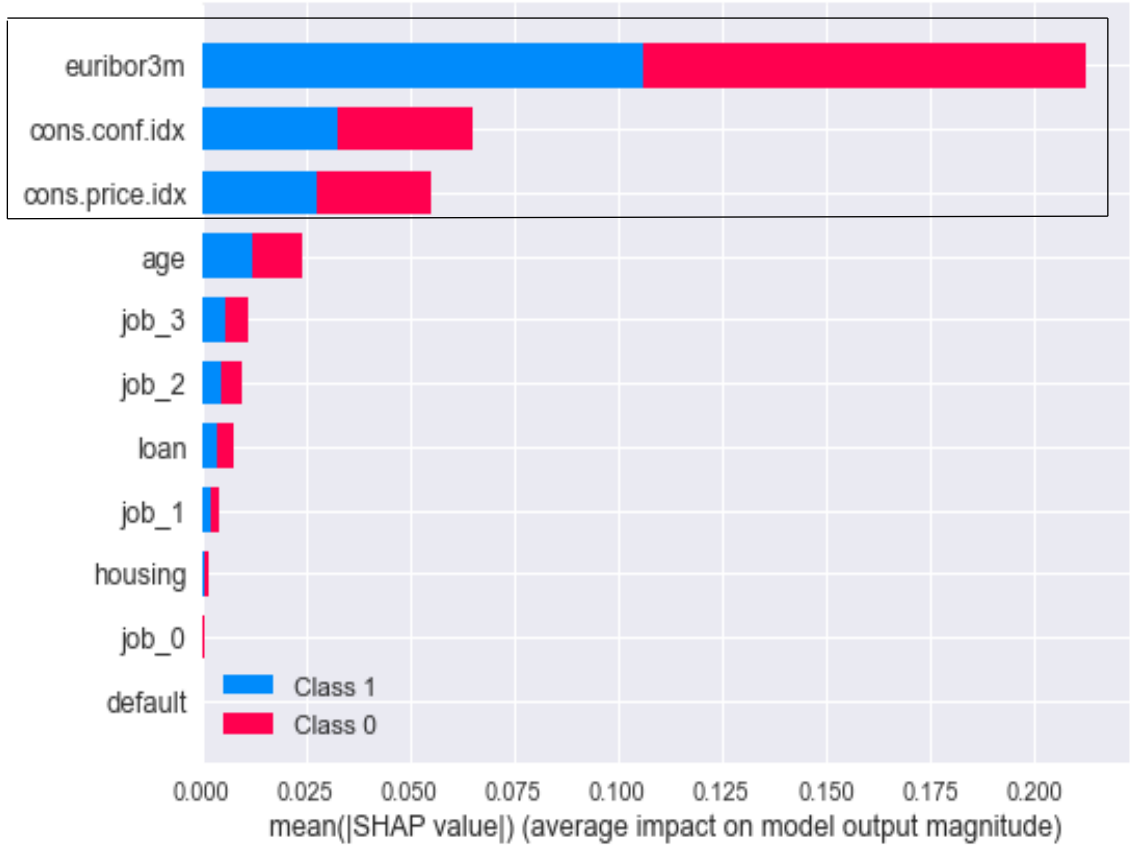
Precision-Recall curve is a curve that combines precision (PPV) and Recall (TPR) in a single visualization. For every threshold, you calculate PPV and TPR and plot it. The higher on y-axis your curve is the better your model performance.

You can use this plot to make an educated decision when it comes to the classic precision/recall dilemma. Obviously, the higher the recall the lower the precision. Knowing at which recall your precision starts to fall fast can help you choose the threshold and deliver a better model.

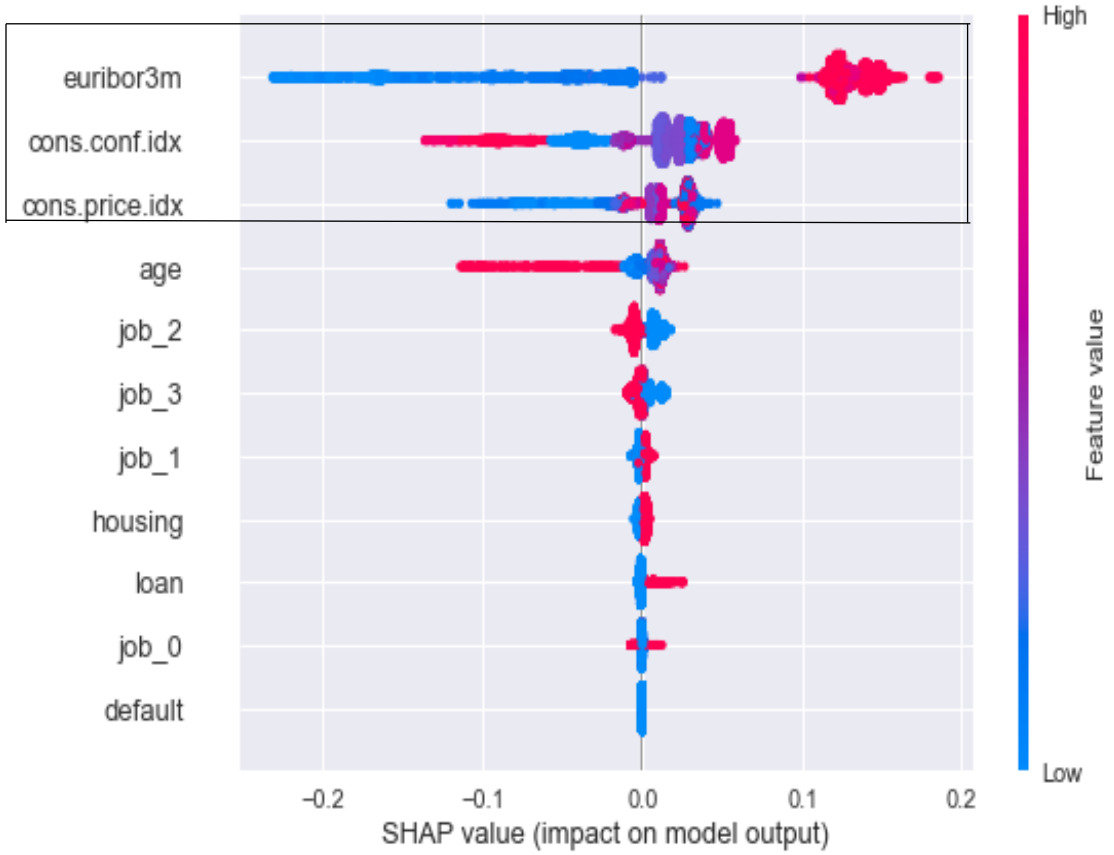
IN THE CHART BELOW, WE CAN CONCLUDE THE FOLLOWING INSIGHTS:

- **HIGHER VALUE OF "EURIBOR" (HIGH INTEREST RATE)** LEADS **TO HIGHER CHANCE OF CLIENTS TO ACCEPT THE CAMPAIGN**. LOWER VALUE OF "EURIBOR" (LOW INTEREST RATE) LEADS TO HIGHER CHANCE OF CLIENTS TO DO NOT ACCEPT CAMPAIGN(DECLINE THE OFFER).
- **HIGHER VALUE OF "CONS.CONF.IDX"** LEADS TO **HIGHER CHANCE TO OF CLIENTS TO ACCEPT THE CAMPAIGN AS WELL**.
- **LOWER VALUE OF "CONS.PRICE.IDX"** LEADS TO **HIGHER THE CHANCE OF CLIENTS TO DECLINE THE CAMPAIGN(DECLINE THE OFFER).**

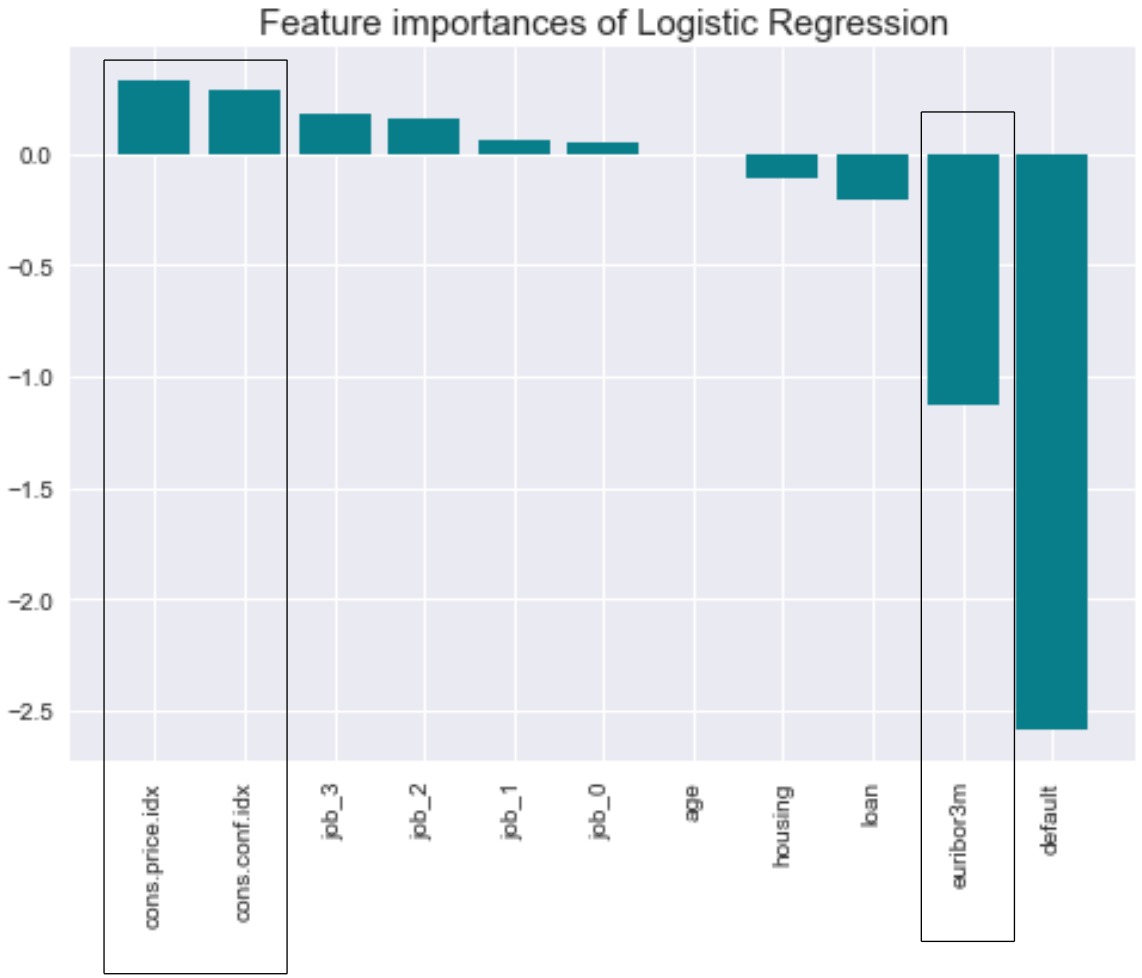
FEATURE IMPORTANCES OF RANDOM FOREST – SHAPE VALUES BAR



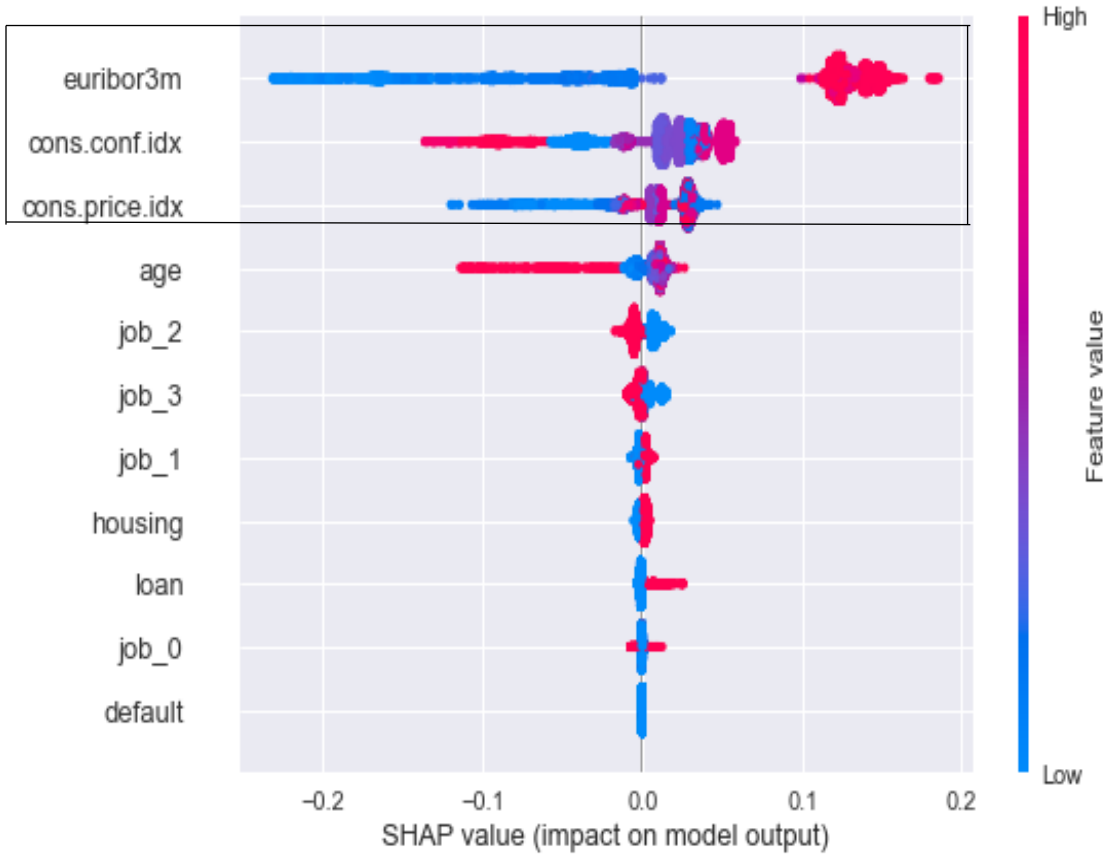
FEATURE IMPORTANCES OF RANDOM FOREST – SHAPE VALUES BEE SWARM



BY DOING FEATURES IMPORTANCE WITH LOGISTIC REGRESSION, WE CAN COMPARE WITH RANDOM FOREST FEATURE IMPORTANCES. FROM THESE 2 MODELS, THERE ARE 2 FEATURES THAT HAVE SAME PATTERN WHICH ARE `CONS.PRICE.IDX` AND `CONS.CONF.IDX`, BOTH OF THEM AFFECT THE TARGET RESPONSES WHEN THE RESPONSE IS "YES".



FEATURE IMPORTANCES OF RANDOM FOREST – SHAPE VALUES BEE SWARM



CONCLUSION

According to previous analysis, a target customer profile can be established. **the most responsive customers possess these features:**

- feature 1 : age <30 or age >60
- feature 2 : students or retired people

By analyzing the socio-economic factors**. the most best environtment for deposit to clients are :**

- feature 3 : high interest rate or high euribor3m
- feature 4 : clients prefer to deposit when cpi is low(interest rate<inflation) - low cpi
- feature 5 : client's confidence to then/current economy is high (high cci)

Also Because this dataset also taken during 2008-2009 financial crisis, we should consider other situation which is why clients even though the interest rate is low they accept the campaign and keep doing term deposit.

RECOMMENDATION

1. **MORE APPROPRIATE TIMING :**
   - When implementing a marketing strategy, external factors, such as the time of calling, should also be carefully considered. The previous analysis points out that March, September, October and December had the highest success rates.

2. **SMARTER MARKETING DESIGN:**
   - By targeting the right customers, the bank will have more and more positive responses. Hence, more accurate information will be presented to the bank for improving the subscriptions. Meanwhile,, the bank should re-evaluate the content and design of its current campaign, making it more appealing to its target customers.

3. **BETTER SERVICES PROVISION**
   - With a more granular understanding of its customer base, the bank has the ability to provide better banking services.


THANK YOU

FOR MORE INFORMATION ABOUT THIS PROJECT, YOU CAN ACCESS THIS NOTEBOOK :

https://colab.research.google.com/drive/1N90U9oXMgSXv8yaOp8YF37pGmcj9bXqR?authuser=2