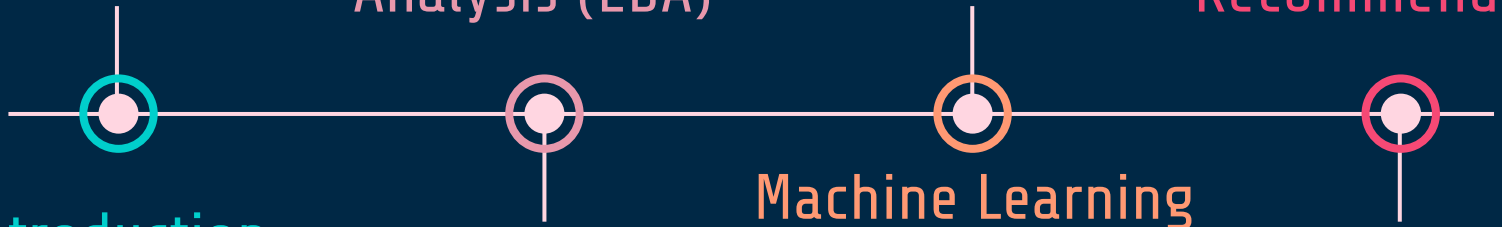
DATA SCIENCE FINAL PROJECT

Machine Learning Classification:
Marketing Target Analysis

By Rafi Wirawan

End-To-End Project



Introduction

Background & Objective
Problem Statement

Analysis (EDA)

Data Cleaning
Data Explorations
Data Visualization
Data Preprocessing

Machine Learning Modelling

Model Train
Model Validation

Recommendation

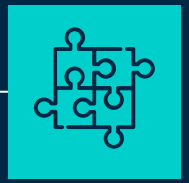
Business case implementation / suggestion (conclusion)

INTRODUCTION

01

Background & Objective
Problem Statement

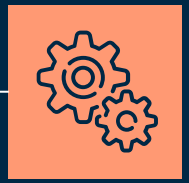
Background & Objective



01

BACKGROUND

The company will have telephonic campaign. The goal is to have customers subscribe for term deposits.



02

PROBLEM

By reaching up to 31000 customers, the company will spend 2000 hours of working and cost $140.000 (with the assumption $ 1 / minutes cost)

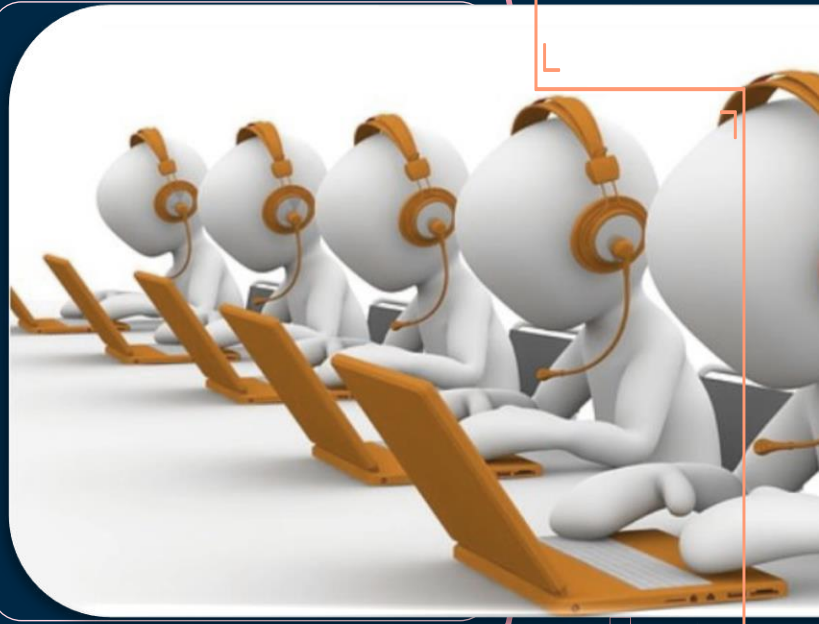
03

TARGET

Creating a model to help the telemarketing team works effectively and reduces cost.

Problem Statement

1. Which features contributes high subscribed rate?
2. Which analytics model has the highest score prediction for Banking Dataset?
3. What strategy the marketing team could use to succeed in their campaign?



Data Acquisition

Marketing Target Analysis dataset was obtained from kaggle.com with 31647 rows and 17 columns

The data set includes information about:

1. Customer's age – the column is called age
2. Customer type of job– the column is called job
3. Customer's education level – the column is called education
4. Outcome of the previous marketing campaign – the column is called poutcome

Data Dictionary

NO	COLUMNS	DESCRIPTION
1	age	Numeric of customer age
2	job	Type of job (categorical: "admin.","blue-collar","entrepreneur","housemaid",etc)
3	marital	Marital status (categorical: "divorced","married","single"; note: "divorced" means divorced or widowed)
4	education	Categorical: "primary","secondary","tertiary,"unknown"
5	default	Having credit in default or not (categorical: "no","yes")
6	balance	How much money in bank account (numeric)
7	housing	Currently having a housing loan or not? (categorical: "no","yes")
8	loan	Has personal loan or not (categorical: "no","yes")
9	contact	Contact communication type (categorical: "cellular", "telephone")
10	month	Last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")
11	day	Last contact day of the week
12	duration	Last contact duration, in seconds (numeric). (e.g., if duration=0 then y="no")
13	campaign	Number of contacts performed during this campaign and for this client (numeric, includes last contact)
14	pdays	Number of days that passed by after the client was last contacted from a previous campaign.
15	previous	Number of contacts performed before this campaign and for this client (numeric)
16	poutcome	Outcome of the previous marketing campaign (categorical: "failure","unknown","other","success")
17	subscribed	Has the client subscribed a term deposit? (binary: "yes","no")

Analysis (EDA)

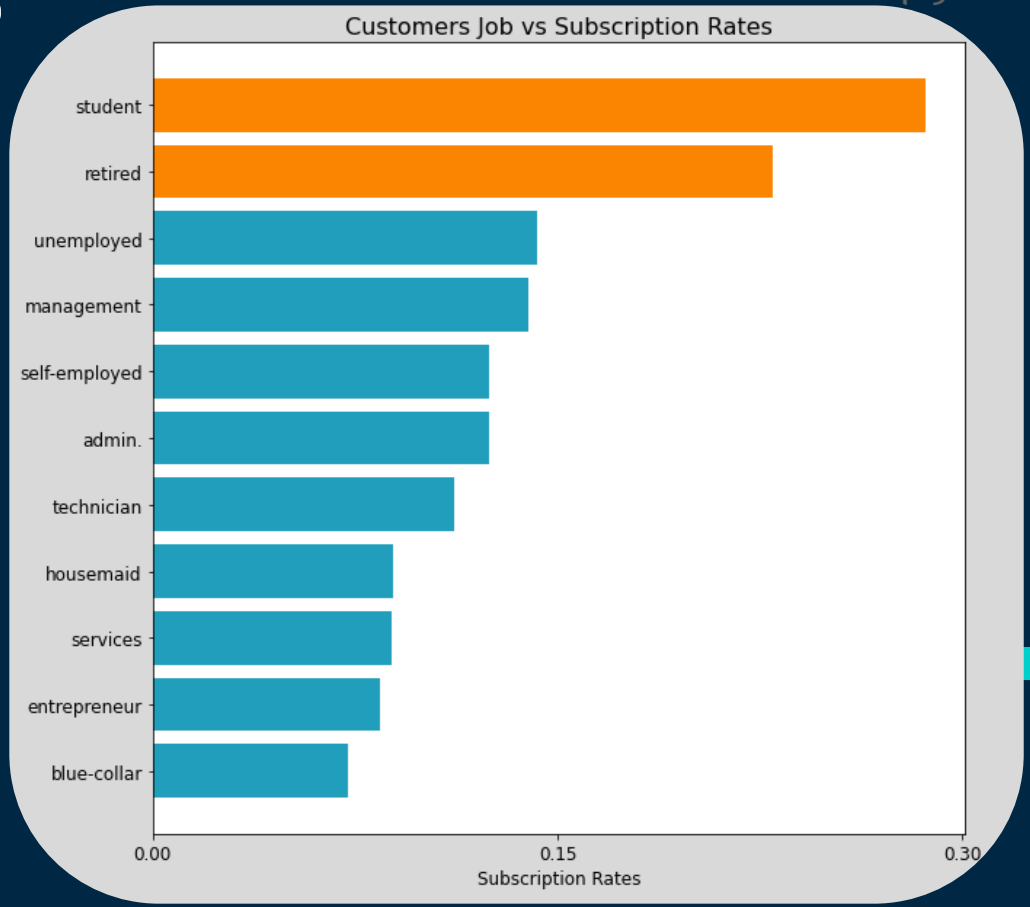
Data Cleaning | Data Explorations |
Data Visualization | Data Preprocessing

02

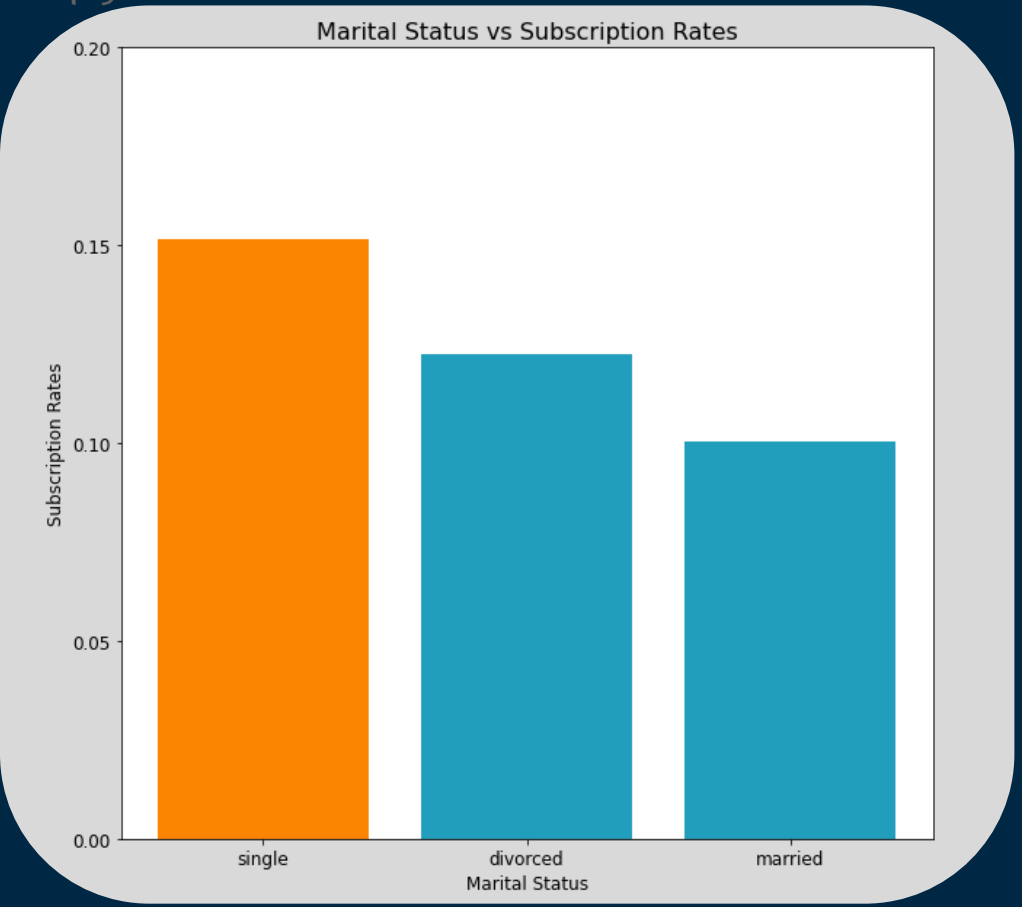
Job vs Subscription Rates

Based on customers type of job, student and retired customers have the highest subscription rates which are 0.28 for student and 0.22 for retired.

	job	subscribed
8	student	0.286614
5	retired	0.229987
10	unemployed	0.142541
4	management	0.138997
6	self-employed	0.124666
0	admin.	0.124484
9	technician	0.111928
3	housemaid	0.088913
7	services	0.088415
2	entrepreneur	0.084325
1	blue-collar	0.072489



Marital Status vs Subscription Rates



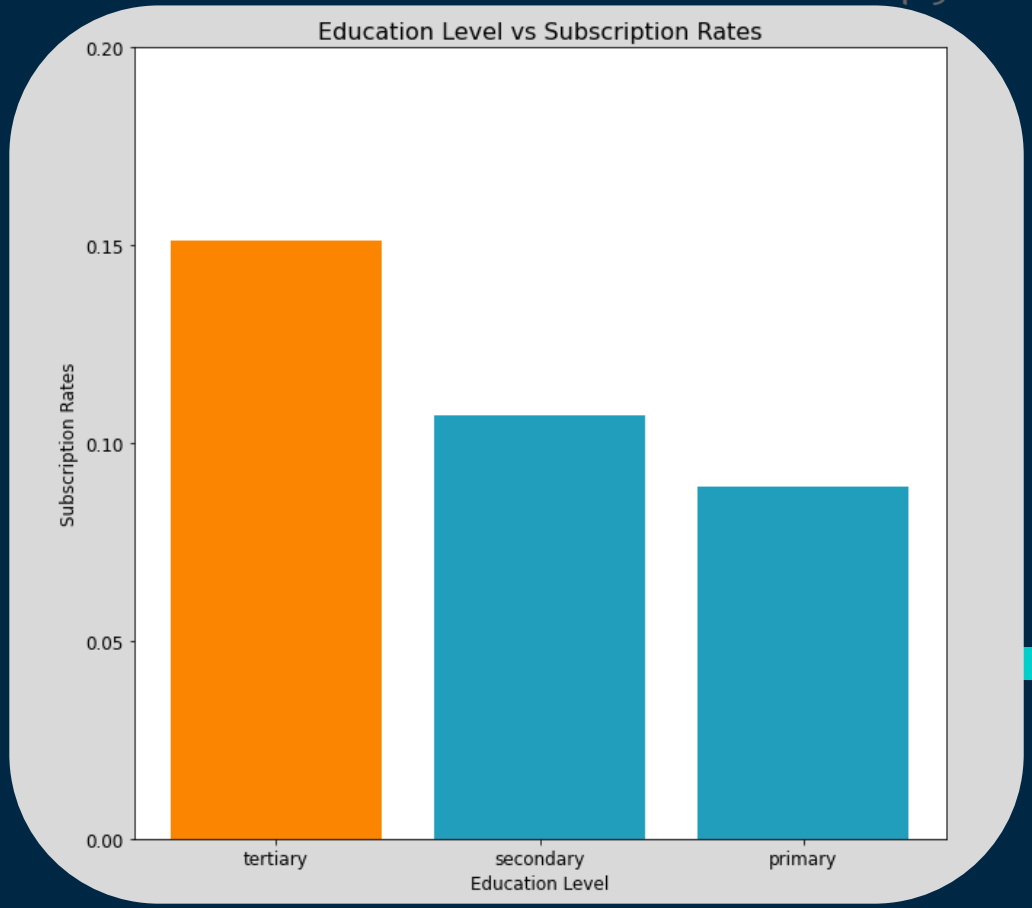
Among 31647 customers, those who are still single has higher subscription rates which is 0.15. This is 20% higher rates than those who already divorced and 33% higher than married customers.

	marital	subscribed
2	single	0.151423
0	divorced	0.122590
1	married	0.100498

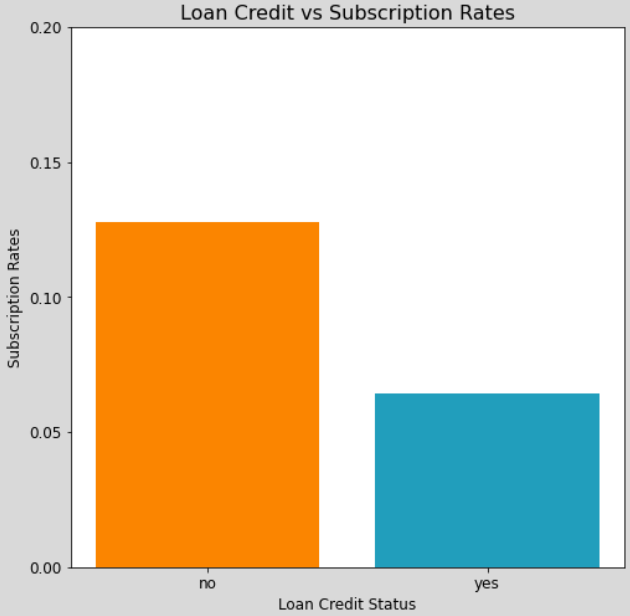
Education vs Subscription Rates

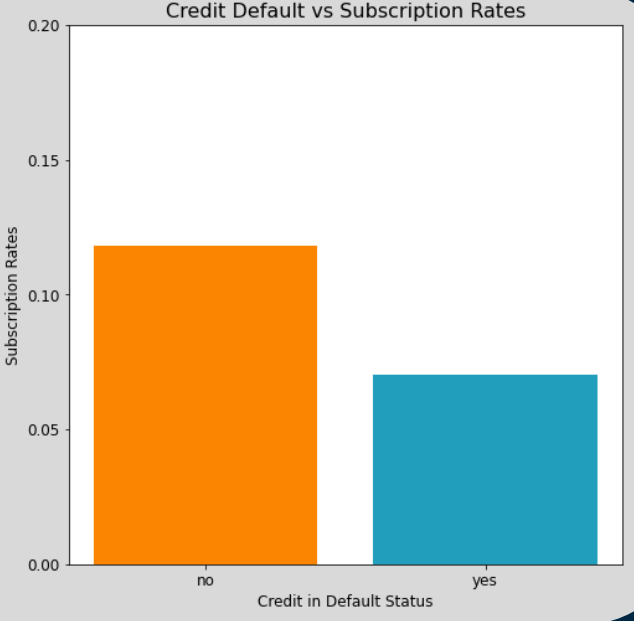
If we classified the customers by their education level, customers with education level of tertiary has the highest subscription rates which is 0.15.

	education	subscribed
2	tertiary	0.151017
1	secondary	0.106916
0	primary	0.088810



Loan, Housing, Credit Default vs Subscription Rates



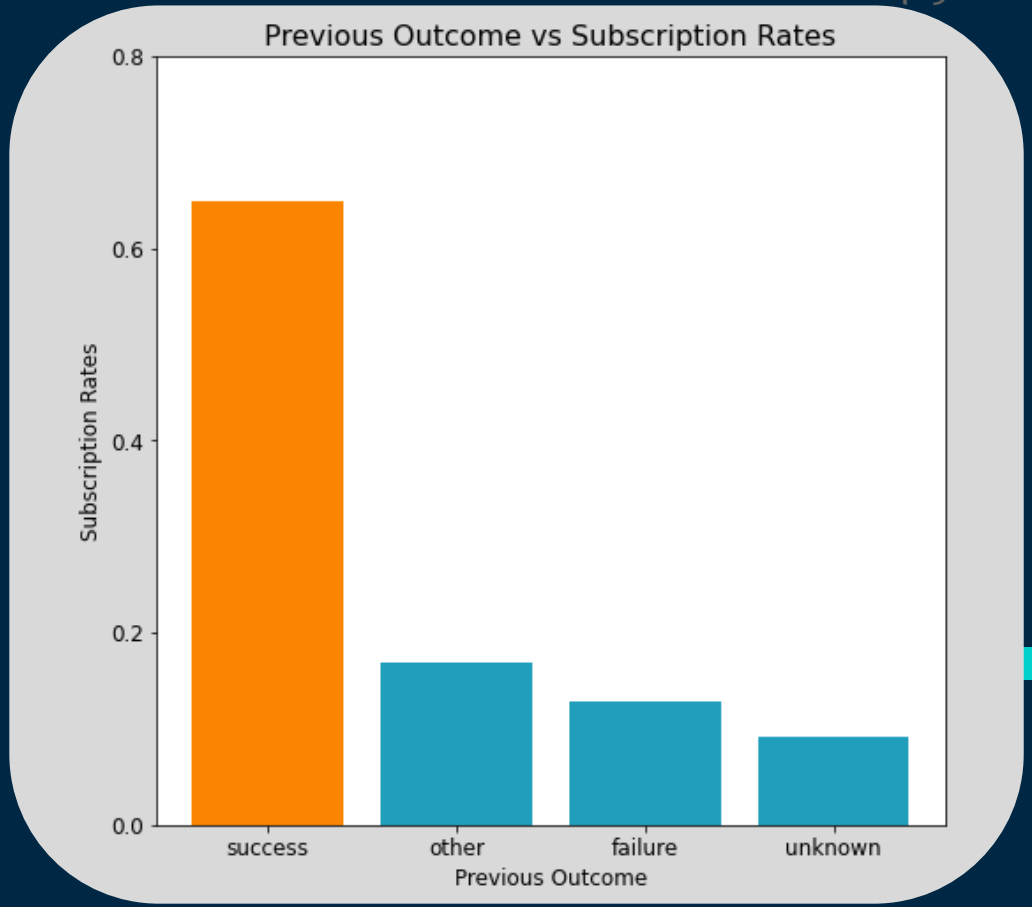


Loan, Housing, and Credit Default have similarities if we compare it to subscription rates. Customer who currently having debt in the bank tends to reject the offering of marketing campaign

Previous Outcome vs Subscription Rates

We could see there is a correlation with the customers who succeed in previous campaign have highest subscription rates which is 0.64. There is a significant differences of 48% than the others.

	poutcome	subscribed
2	success	0.649813
1	other	0.168478
0	failure	0.128198
3	unknown	0.091519



Previous Outcome vs Subscription Rates by Age

By classifying the customers into 4 group of age, we could see the differences of subscription rate for customers who are older than 60 years old. This corresponds to duration spend by the telemarketing calling the old customers, which is 206.770 minutes or 3446 Hours in total.

	age_category	subscribed
3	> 60	0.130611
0	0-20	0.002478
1	21-40	0.002022
2	41-60	0.000692

	age_category	durasi_telfon
3	> 60	206770
2	41-60	1446
1	21-40	989
0	0-20	807

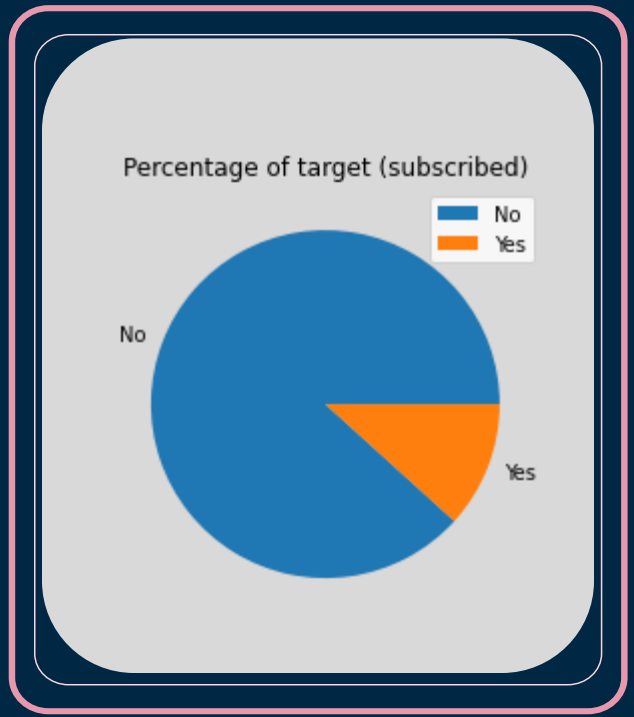
Data Encoding

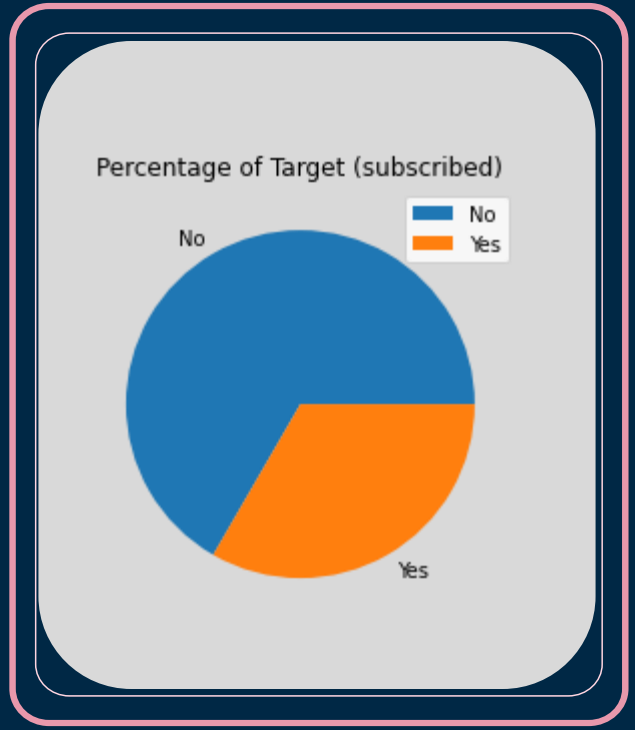
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 31647 entries, 0 to 31646
Data columns (total 17 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   age         31647 non-null  int64
 1   job         31647 non-null  object
 2   marital     31647 non-null  object
 3   education   31647 non-null  object
 4   default     31647 non-null  object
 5   balance     31647 non-null  int64
 6   housing     31647 non-null  object
 7   loan        31647 non-null  object
 8   contact     31647 non-null  object
 9   day         31647 non-null  int64
 10  month       31647 non-null  object
 11  duration    31647 non-null  int64
 12  campaign    31647 non-null  int64
 13  pdays       31647 non-null  int64
 14  previous    31647 non-null  int64
 15  poutcome    31647 non-null  object
 16  subscribed  31647 non-null  object
dtypes: int64(7), object(10)
memory usage: 4.1+ MB
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11145 entries, 0 to 11144
Data columns (total 20 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   age                  11145 non-null  int64
 1   default              11145 non-null  int64
 2   balance              11145 non-null  int64
 3   housing              11145 non-null  int64
 4   loan                 11145 non-null  int64
 5   duration             11145 non-null  int64
 6   campaign             11145 non-null  int64
 7   pdays                11145 non-null  float64
 8   previous             11145 non-null  int64
 9   marital_divorced     11145 non-null  uint8
 10  marital_married      11145 non-null  uint8
 11  marital_single       11145 non-null  uint8
 12  education_primary    11145 non-null  uint8
 13  education_secondary  11145 non-null  uint8
 14  education_tertiary   11145 non-null  uint8
 15  poutcome_failure     11145 non-null  uint8
 16  poutcome_other       11145 non-null  uint8
 17  poutcome_success     11145 non-null  uint8
 18  poutcome_unknown     11145 non-null  uint8
 19  subscribed           11145 non-null  int64
dtypes: float64(1), int64(9), uint8(10)
memory usage: 979.7 KB
```

There are some features in the dataset that need to be encoded I used label encoding to encode value consists of 'yes' and 'no', and I used one-hot encoding to process features with more than one unique value.

Handling Data Imbalance





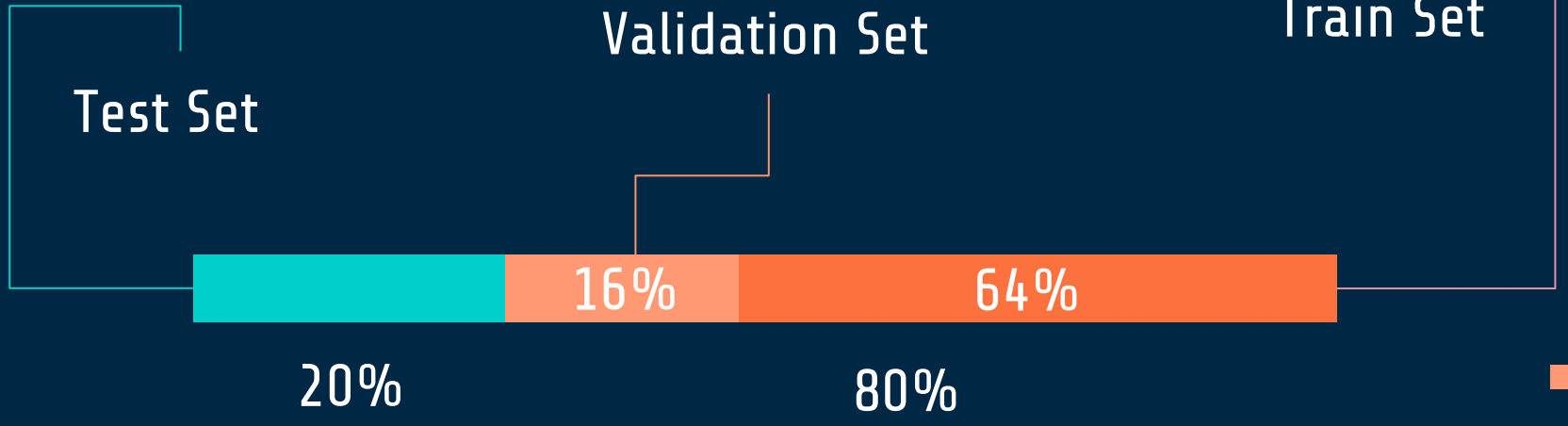
In order to improve the performance of machine learning, the imbalance target must be handled first. I used undersampling with the ratio of 0.5.

Modelling

Model Train | Model Validation | Model Evaluation

03

SPLIT THE DATASET



In this project I split the dataset using the ratio of 80 : 20. I split the data into 80% of training data and 20% of test data. Then I split the training data into 16% of validation data and 64% of training data. This project used 5 different classification model, they are : Decision Tree, Logistic Regression, Random Forest, Support Vector Machine, and Naive Bayes.

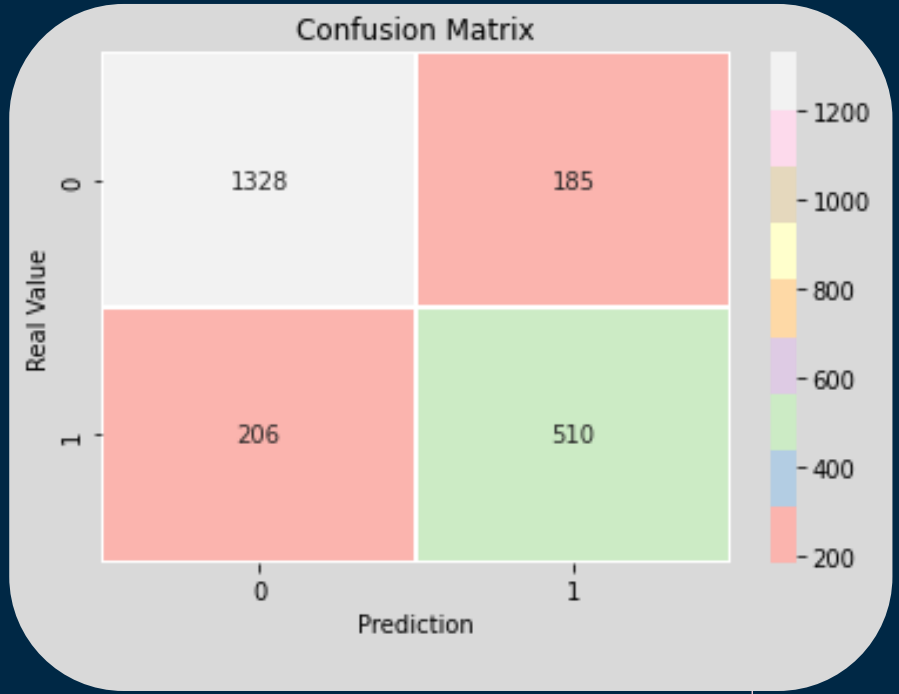
Model Evaluation

	Precision		Recall		F1 Score	
	Train	Test	Train	Test	Train	Test
Decision Tree(Base Model)	-	0.36	-	0.41	-	0.38
Decision Tree	0.61	0.62	0.61	0.63	0.61	0.63
Logistic Regression	0.76	0.78	0.59	0.60	0.67	0.67
Random Forest	0.74	0.73	0.68	0.71	0.71	0.72
Support Vector Machine	0.69	0.72	0.38	0.38	0.48	0.50
Naive Bayes	0.70	0.69	0.51	0.53	0.59	0.60

Random Forest

Due to the concern of creating the model is customers recommendation, therefore the precision of the machine learning model is important. However, the model needs to be sensitive enough to classify the data.

Random Forest has the precision score of 0.73 and recall score of 0.71. This leads to F1 score of 0.72. To simplify, our model predicted 510 customers within the actual value of 716 customers.



Precision		Recall		F1 Score	
Train	Test	Train	Test	Train	Test
0.74	0.73	0.68	0.71	0.71	0.72

Hyperparameter Tuning

Define Range

STEP 1

Define the parameters and range for hyperparameter tuning

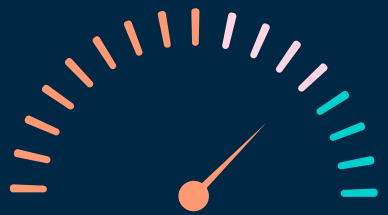
Optimization Process

STEP 2

Run the process using GridSearchCV

Extract Best Parameter

STEP 3



After the process completed we will have the best parameter for our model

Hyperparameter Tuning

BEFORE

Precision		Recall		F1 Score	
Train	Test	Train	Test	Train	Test
0.74	0.73	0.68	0.71	0.71	0.72

TUNING

```
parameter
n_estimators          [10, 17, 25, 33, 41, 48, 56, 64, 72, 80]
max_features                                        [auto, sqrt]
max_depth                                                [2, 4]
min_samples_split                                        [2, 5]
min_samples_leaf                                         [1, 2]
bootstrap                                          [True, False]
Name: param_grid, dtype: object
```

AFTER

Precision		Recall		F1 Score	
Train	Test	Train	Test	Train	Test
0.74	0.75	0.61	0.61	0.67	0.68

After the hyperparameter tuning process, the score of random forest model has affected. The precision score increases to 0.75 for testing and 0.74 for training. Meanwhile The recall score and F1 Score moves down to 0.61 and 0.68 for testing and 0.61 and 0.67 for training. To simplify, the model doesn't undergo overfitting.

Recommendation

Business case implementation | suggestion (conclusion)

04

Recommendation

Special cashback for student

In some investment platform, they usually give some cashback for those who subscribes the term deposits. Due to 20 % of our customers are student, it could lead to positive impact to our profit.

Focus on the loan-free

The telemarketing should reach more customers who currently having no debt in the bank.

Special promo for new customers

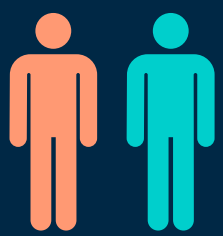
The company needs to attract more new customers by giving special promo like cashback or discount.

Target the duration call

Duration of the call has major correlation with the succeed of telemarketing approach.

Business Implementation

BEFORE



100
Telemarketing

Salary : $ 15 / Hours
Total : $ 2.970.000 / Month

100% Cost : $ 140.000

100% Duration : 2000 Hours or 84 days

Subscription Rate : 11,73 %

AFTER

75
Telemarketing

Salary : $ 2.227.500 / Month

75% Cost : $ 105.000

75% Duration : 1500 Hours or 63 Days

Subscription Rate : 75,59 %

$ 777.500.000

The cost that the company will save
if using our machine learning model

Do you have any questions?

Email : rafiwirawan@gmail.com
Instagram : rafiwirawan
LinkedIn : https://www.linkedin.com/in/muhammadrafiwirawanputra/

THANKS