Kampus Merdeka x MyEdusolve

The Term Deposit Subscription Prediction

OUR TEAMS

Alex Mario Simanjuntak

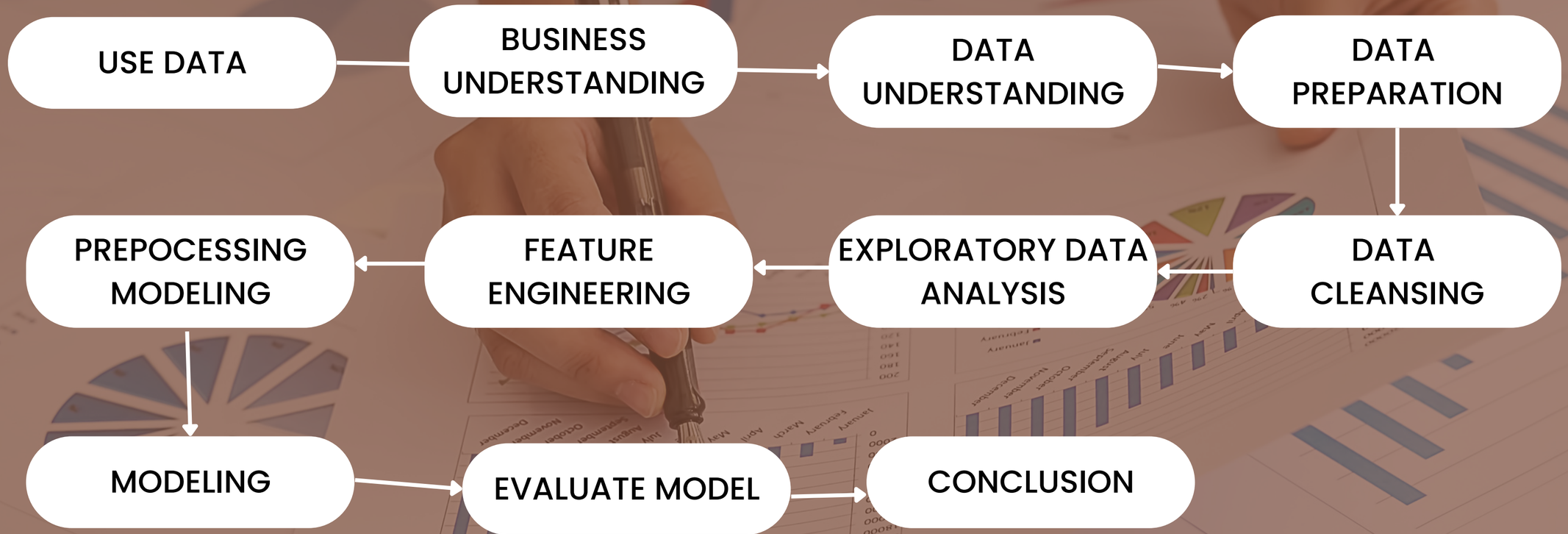
Audric Lysander

Muhammad Reza Abdillah

Latifah Sinta

Yemima Sipayung

WORKFLOW


USE DATA
BUSINESS UNDERSTANDING
DATA UNDERSTANDING
DATA PREPARATION
DATA CLEANSING
EXPLORATORY DATA ANALYSIS
FEATURE ENGINEERING
PREPOCESSING MODELING
MODELING
EVALUATE MODEL
CONCLUSION

USE CASE

Objectives

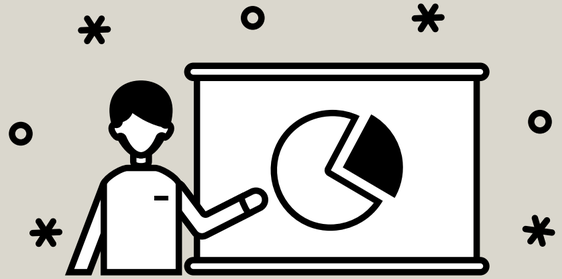
- To find out how many customers are subscribe an not subscribe.
- To gain insight on subscribed and non subscribed based on customer demographic info such as age, balance, job, education, etc.
- To gain insight into whether in each feature there is a relationship.
- To build the right model to predict customers and find out whether the product (bank term deposit) is subscribed by the client or not.
- Optimize model using Hyperparameter Tuning

USE CASE

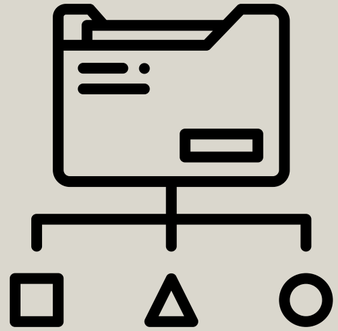
Methodology

Exploratory Analysis

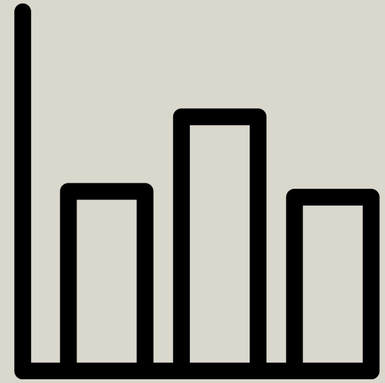
Descriptive Statistic



Classification report



Statistical Analysis



Business Understanding

BACKGROUND

Marketing to potential clients has always been a crucial challenge in attaining success for banking institutions. It’s not a surprise that banks usually deploy mediums such as social media, customer service, digital media and strategic partnerships to reach out to customers. But how can banks market to a specific location, demographic, and society with increased accuracy? With the inception of machine learning - reaching out to specific groups of people have been revolutionized by using data and analytics to provide detailed strategies to inform banks which customers are more likely to subscribe to a financial product. In this project on bank marketing with machine learning, we will explain how a particular Portuguese bank can use predictive analytics from data science to help prioritize customers which would subscribe to a bank deposit.

Data Understanding

Dataset
Input variables:

- 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', note: 'divorced' means divorced or widowed)
- education (categorical: "unknown", "secondary", "primary", "tertiary")
- default: has credit in default? (categorical: 'no', 'yes', 'unknown')
- housing: has housing loan? (categorical: 'no', 'yes', 'unknown')
- loan: has personal loan? (categorical: 'no', 'yes', 'unknown')
- y : has the client subscribed a term deposit? (binary: 'yes','no')

Data Understanding

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')
- day: last contact day of the month (numeric)
- duration: last contact duration, in seconds (numeric)

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')

Data Preparation

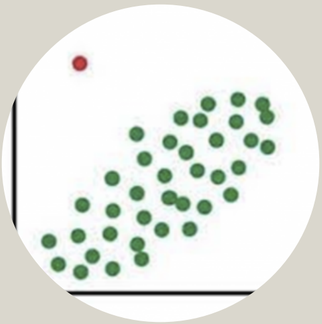
Code Used

Languange: Python
Libraries : Pandas, Numpy, Matplotlib, Seaborn, Sklearn, time, warning

Data has no Null Values

Total of 35 Columns

No Columns Has Missing Values



Columns Has Outliers

Data Cleansing

All data is clean from Null Values and is in accordance with the data type,

Handling Outliers

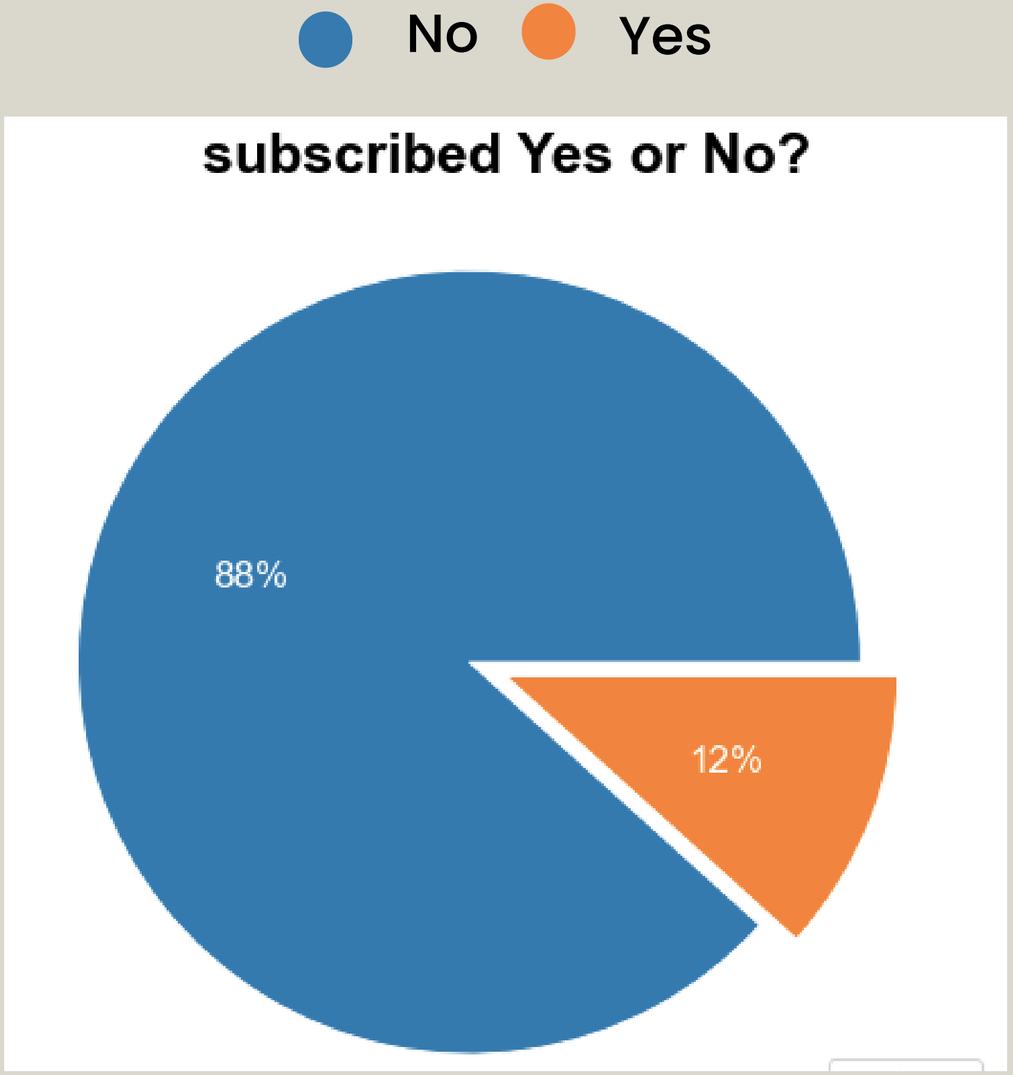
- We only take a few columns for modeling purposes, and there are 3 columns that have outliers, the percentage of each outlier is balance has 10%outliers. campaigns has 6% outliers, and age has 1%outliers
- We detect outliers by using a boxplot, by looking for the lower and upper of each column
- We clean outliers using winsorize

EXPLORATORY DATA ANALYSIS

EDA

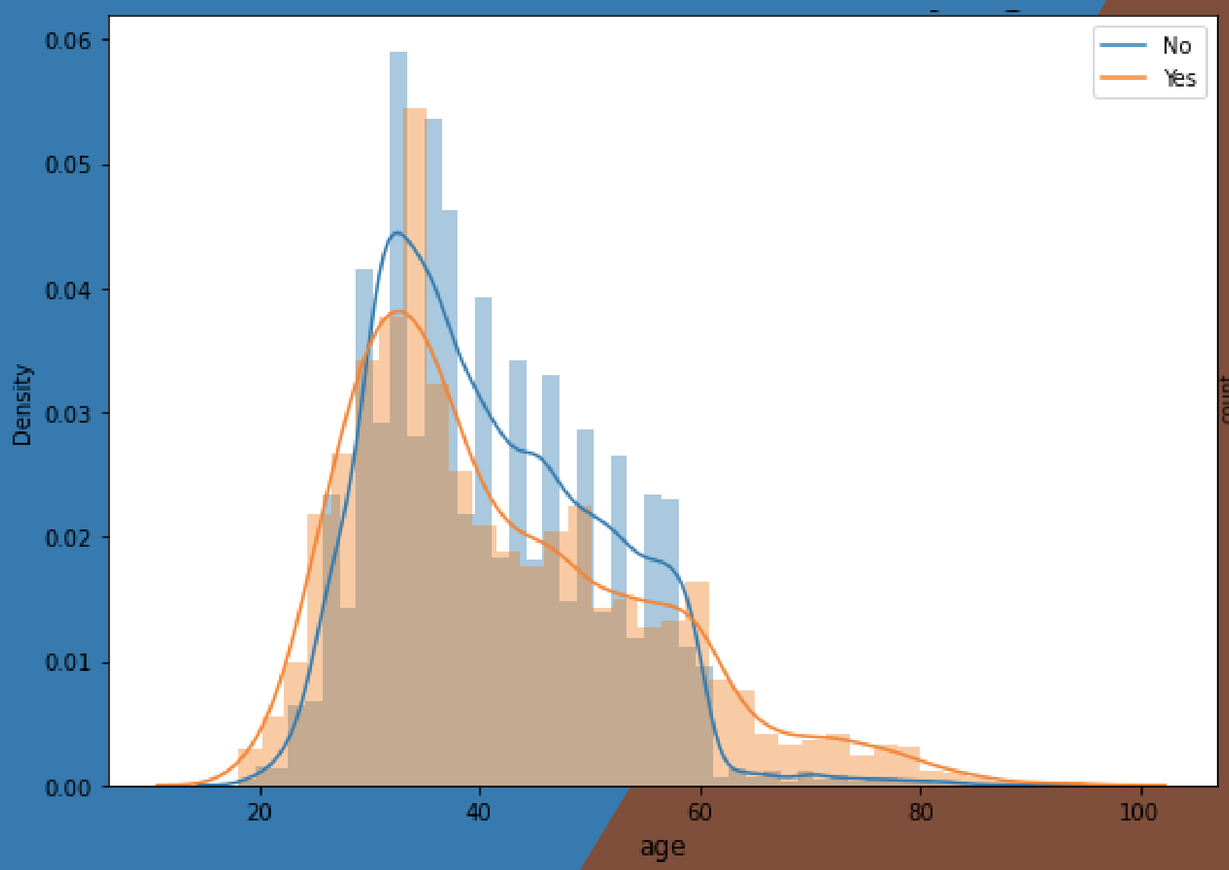
SUBSCRIBED COMPARASION

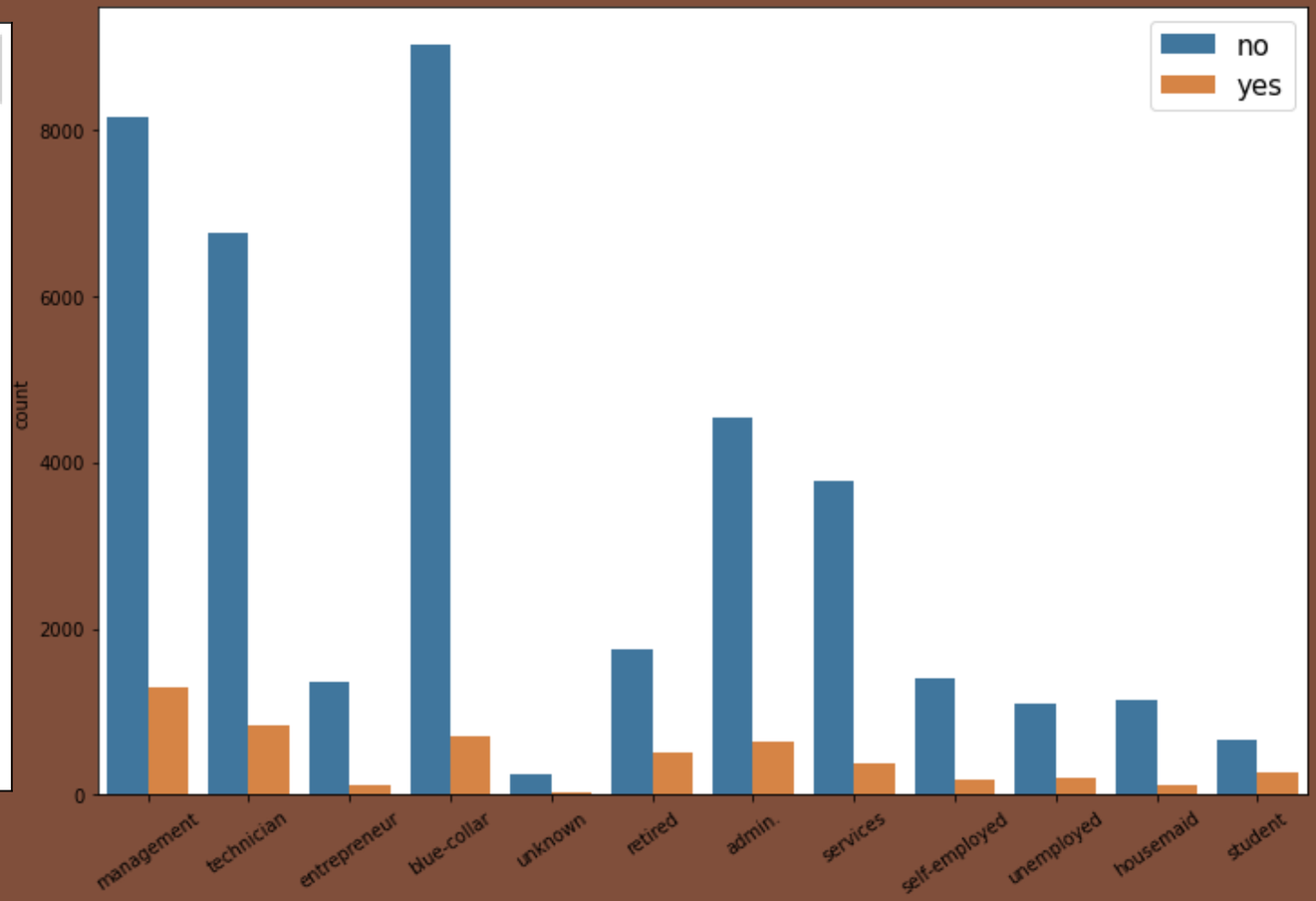
As we can see based on the pie chart above, we can see that there are about 12% (5289) of clients who have deposited, while 88% (39922) have not.



EDA

SUBSCRIPTION BASED ON AGE DISTRIBUTION AND CUSTOMER OCCUPATION



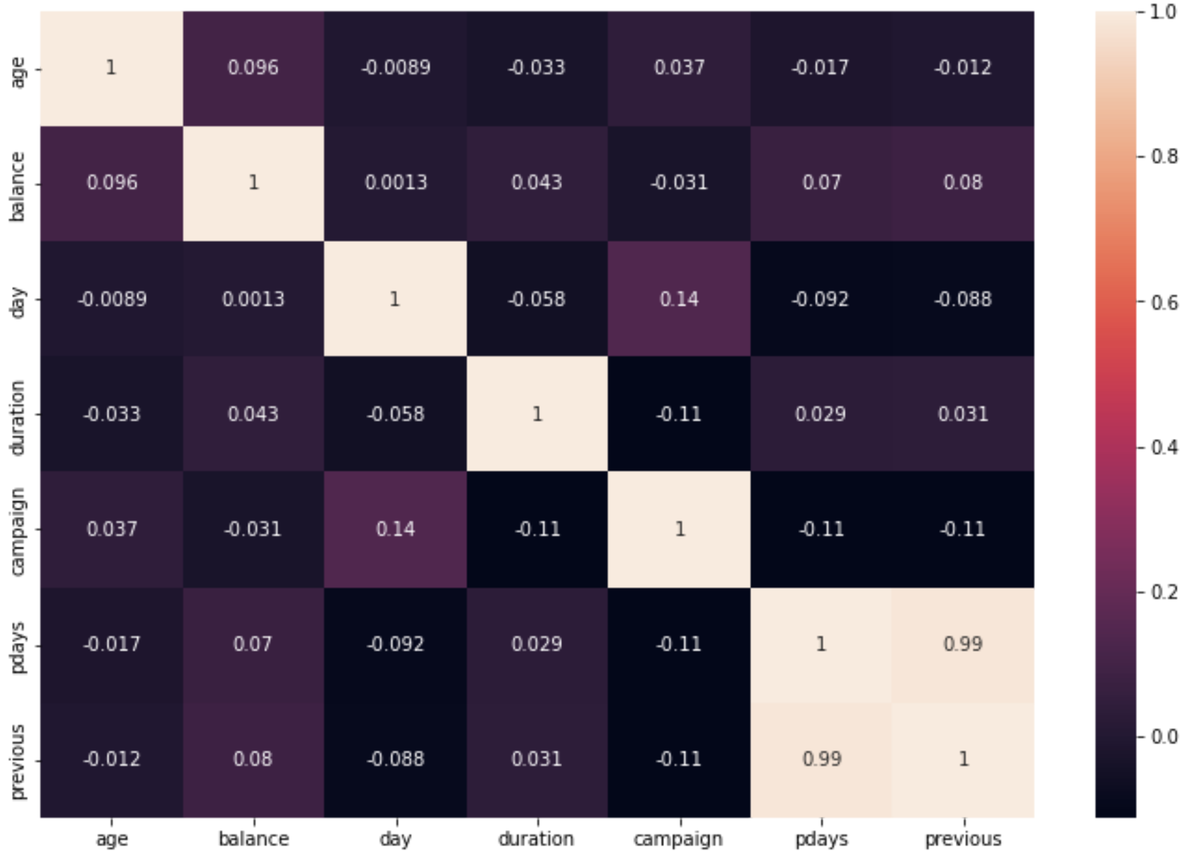


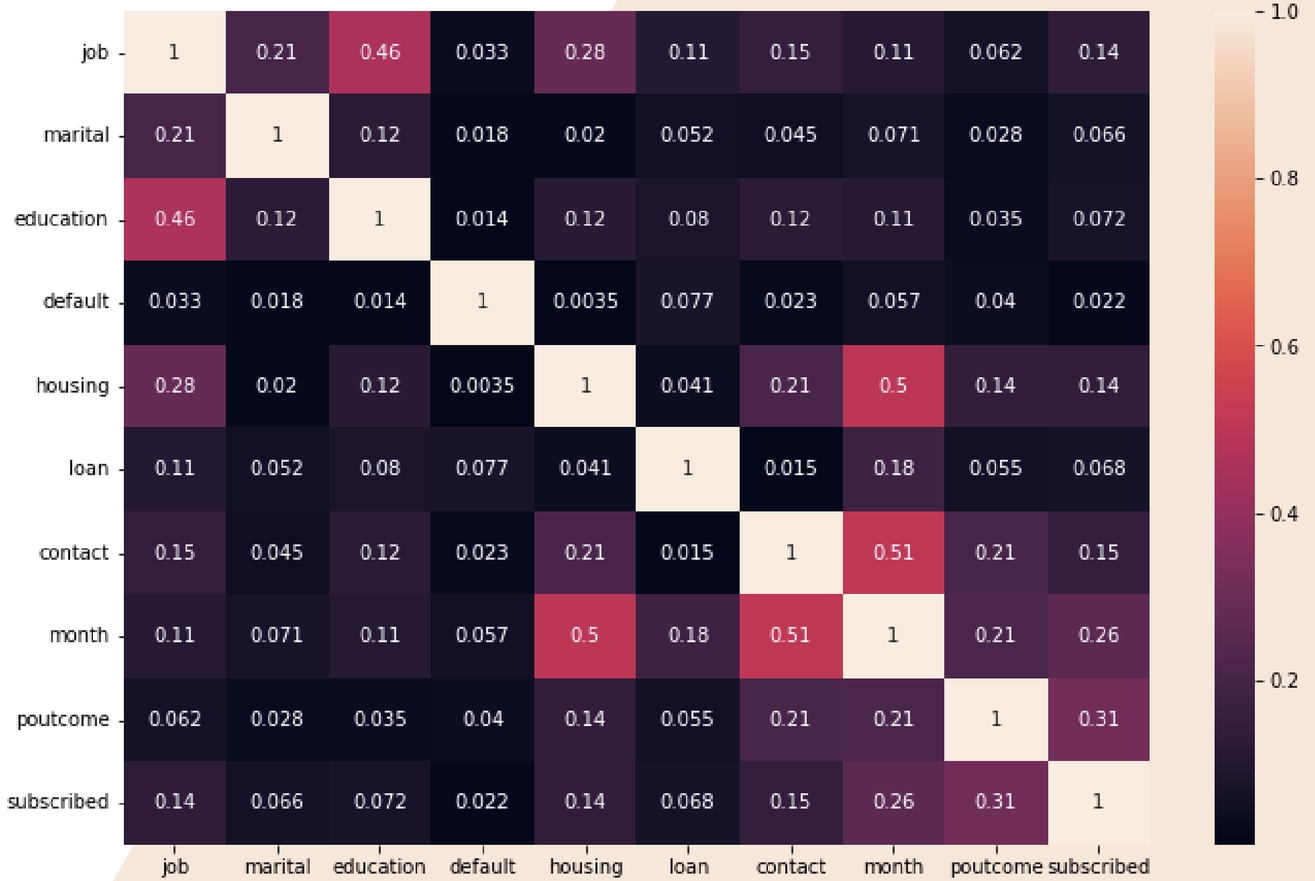
In the age column, it can be seen that the age of the clients in the 30-35 age range, those who have and those who have not made a deposit, there is no significant difference from the data.

From the plot above, we can conclude that most clients who do not make deposits are clients who have jobs as blue-collars, management, and technicians.

EDA

CORELATION ON NUMECRICAL AND CATEGORICAL FEATURES





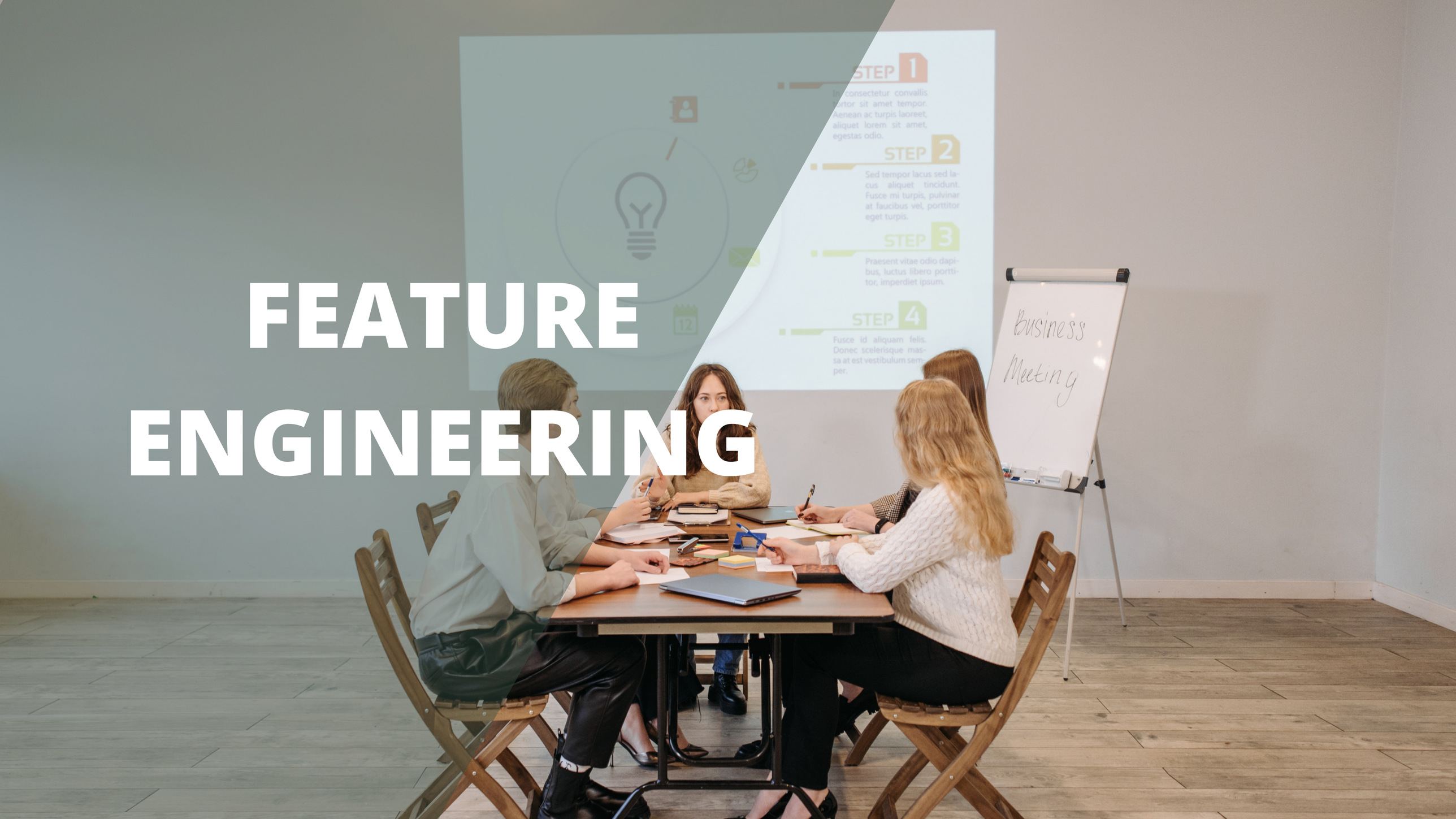
Numerical

in the heatmap above, it can be seen that there is still no significant correlation, while the correlation is

- pdays with previous
- campaign with day

Categorical

In the heatmap above, we can see the moderate correlation, namely job with education, housing with month, and contact with month. there is also a fairly low correlation, namely job with housing, job with martial and housing with contact



FEATURE ENGINEERING

Feature Engineering

FEATURE ENCODING

- ONE-HOT ENCODING

Using one-hot encoding on features: (job, marital, contact, poutcome), (default, housing, loan, subscribed)

- ORDINAL ENCODING

Using ordinal encoding in features: education

- LABEL ENCODING

The encoding label is only used to encode the target variable

FEATURE SCALING

- We will feature scaling on the numerical features using MinMax Scaler


PREPROCESSING
MODELING

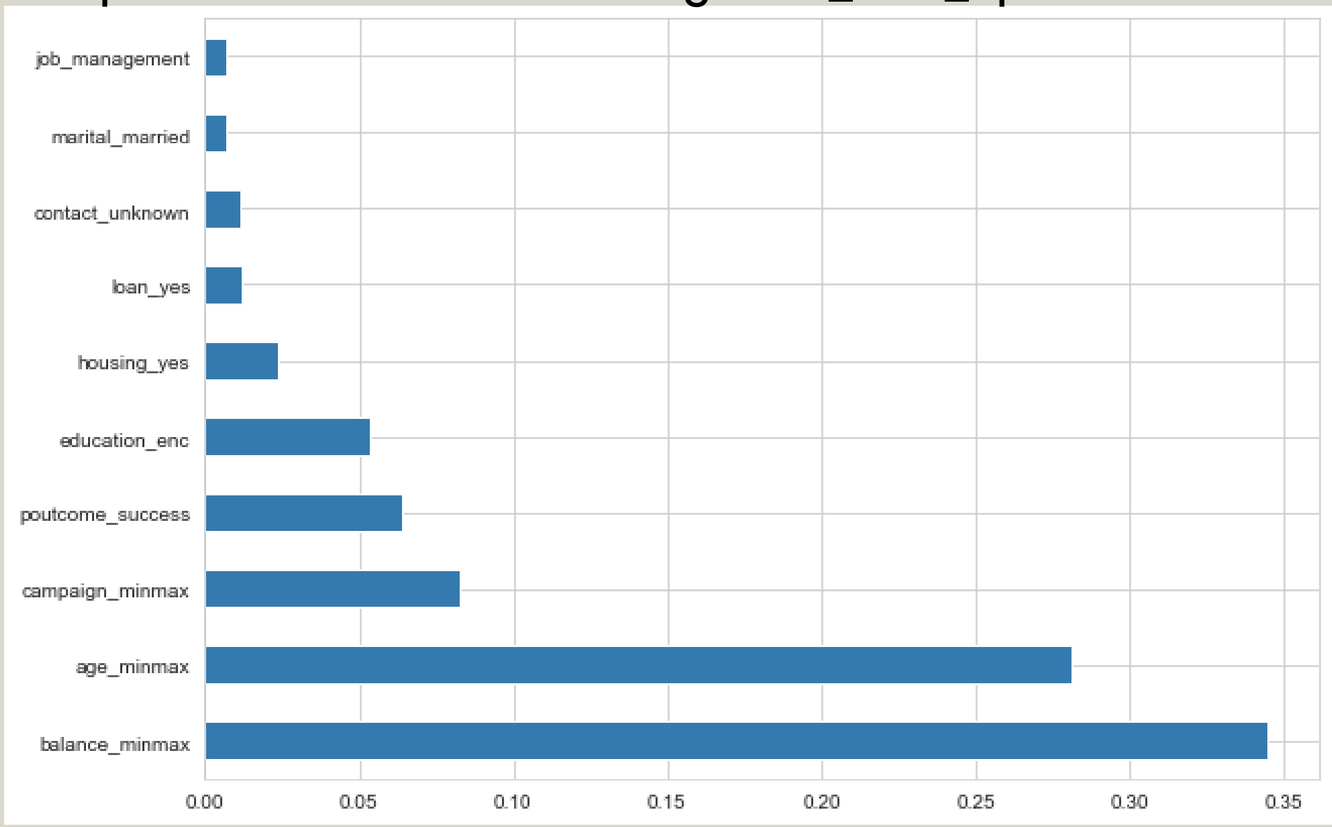
Preprocessing Modeling

Before doing modeling we need to do model preprocessing, so that our model can learn based on important features.

FEATURE SELECTION

- The feature_importances have multicollinearity (poutcome_succes with poutcome_unknown, contact_cellular and contact_unknown, etc.), therefore we will drop the labels.
- Feature Selection using EMBEDDED METHOD (FEATURE IMPORTANCES)

Split train-test the data using train_test_split function.



Modeling

Modeling

Machine Learning Logistic Regression | Ensemble Methods

- Logistic regression is a fundamental classification technique. Logistic regression is fast and relatively uncomplicated, and it’s convenient for you to interpret the results. in this model we train and fitting data using LogisticRegression()

- Ensemble modeling is a process where multiple diverse models are created to predict an outcome, either by using many different modeling algorithms or using different training data sets. The ensemble model then aggregates the prediction of each base model and results in once final prediction for the unseen data. in this model we train and fitting data using GradientBoostingClassifier()

the purpose of using these two models is to compare which model is the best fit for our data

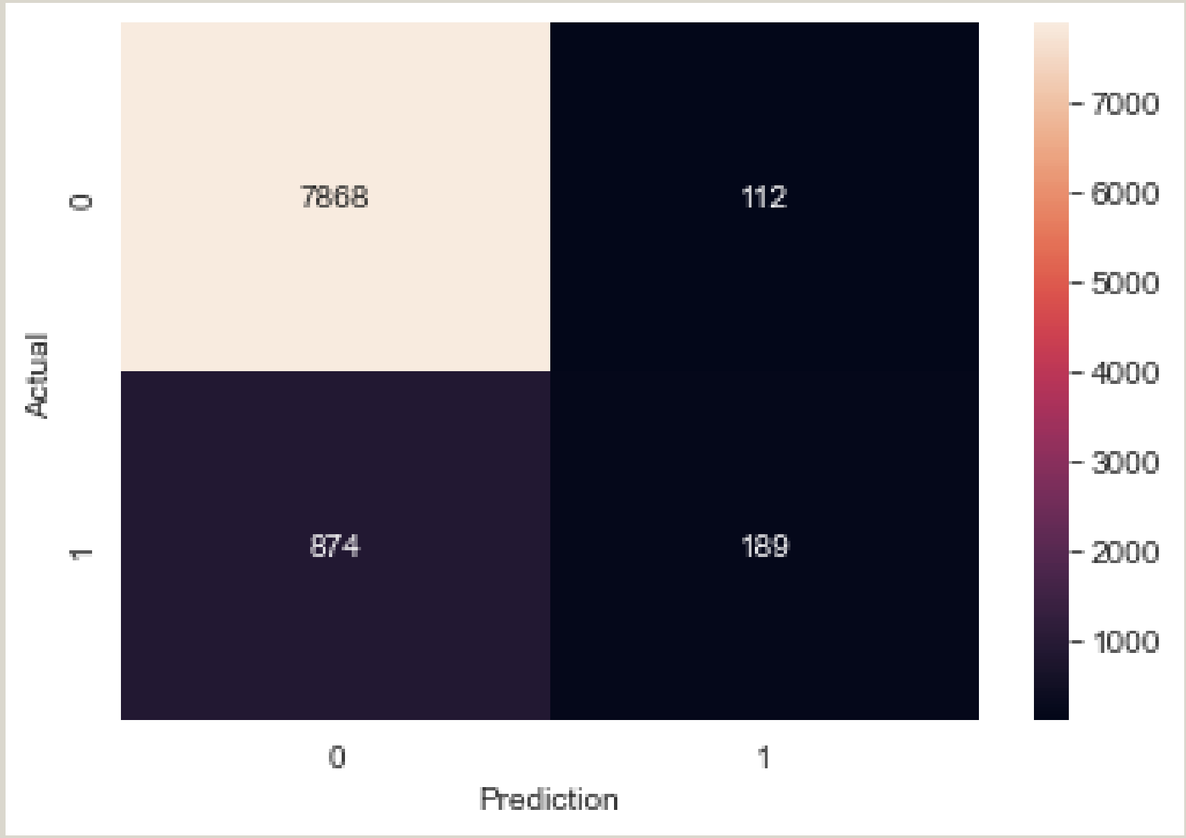
Modeling

Logistic Regression

In this case, in our opinion, False Positive (Prediction:1, Actual:0) is more important because it can be detrimental to the company.
So we will prioritize precision, therefore we don't need to handle imbalance data

we can see that the accuracy of the model is already bestfit, but the precision and recall sections are still overfit

	precision	recall	f1-score	support
0	0.90	0.99	0.94	7980
1	0.63	0.18	0.28	1063
accuracy			0.89	9043
macro avg	0.76	0.58	0.61	9043
weighted avg	0.87	0.89	0.86	9043




```
CV: (0.893635137940446, 0.8935799428792274)
CV: (0.6803496687627075, 0.8935799428792274)
CV: (0.1696318907954204, 0.8935799428792274)
```

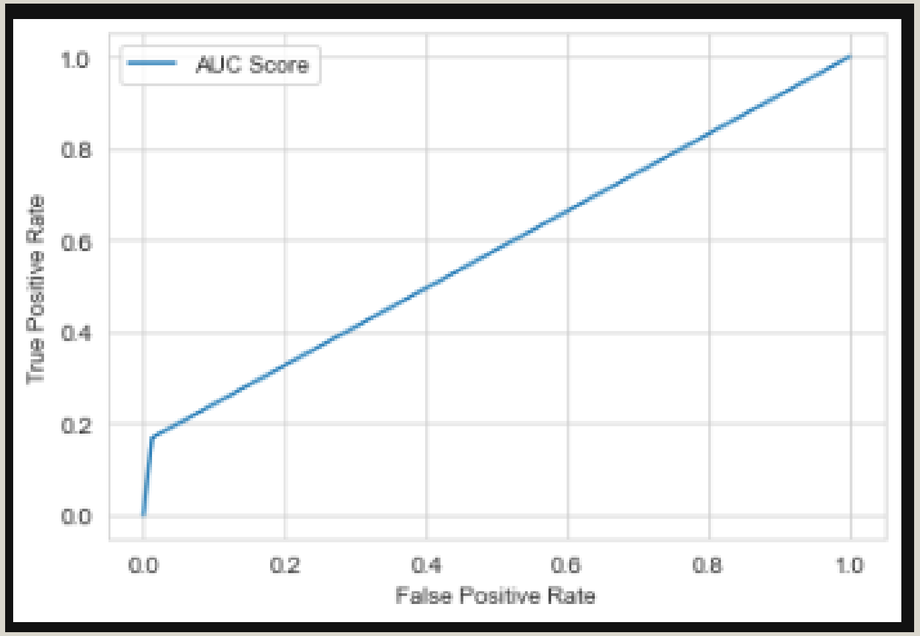
Modeling

Ensemble Model

we can see in the confusion matrix in the precision section that false positives have increased by 3% while the recal true positives have also decreased by 1%,but the accuracy is still the same as the logistic regression model

	precision	recall	f1-score	support
0	0.90	0.99	0.94	7980
1	0.66	0.17	0.27	1063
accuracy			0.89	9043
macro avg	0.78	0.58	0.60	9043
weighted avg	0.87	0.89	0.86	9043

```
CV: (0.893054519533299, 0.892805782454782l)
CV: (0.6684657024683889, 0.8927781314109552)
CV: (0.16844195552425587, 0.8927781314109552)
```




Evaluate
Model

Model Evaluation

HYPERPARAMETER TUNING

Logistic Regression

	precision	recall	f1-score	support
0	0.89	0.99	0.94	7980
1	0.66	0.11	0.19	1063
accuracy			0.89	9043
macro avg	0.78	0.55	0.57	9043
weighted avg	0.87	0.89	0.85	9043

```
model AUC Base : 0.5818817976267102

model AUC Random : 0.5529670248056642
```

The results of a random search show that there is a decrease in model performance when compared to the base model without optimization / tuning parameters.but in the precision false positive section it increased by 3%, and in the true negative recall section it decreased by 7%

HYPERPARAMETER TUNING

Ensemble Model

the results are not much different but there is a slight difference after we tuning to the model, in the precision true negative section it decreased by 1% but in the false positive there was an increase of 1%

	precision	recall	f1-score	support
0	0.89	1.00	0.94	7980
1	0.67	0.04	0.07	1063
accuracy			0.88	9043
macro avg	0.78	0.52	0.50	9043
weighted avg	0.86	0.88	0.84	9043

```
model AUC Base : 0.5783059483138703

model AUC Random : 0.5166834654840299
```

CONCLUSION

> The clients who have not subscribed to a term deposit is 88% (39922).
- clients in the 30-35 age range, those who have and those who have not made a deposit, there is no significant difference from the data.
- most clients who do not make deposits are clients who have jobs as blue-collars, management, and technicians.

> In Evalutation Counfussion Matrix we prioritize precision with the assumption that the campaign is by telephone, targeting those who have not subscribed to deposits.
- if FN = they are considered unsubscribed, but actually they have, I assume it's impossible to call again.
- else FP = they are considered to have subscribed, but actually they haven't, this is dangerous, because their target is those who have not subscribed, so the campaign is useless.

> The accuracy rate of predicting customer using machine learning Logistics Regression and after evaluating the model using Hyperparameter Tuning is 89%.(0.89).

>The accuracy rate of predicting customer using machine learning Ensamble Model and after evaluating the model using Hyperparameter Tuning is decrease from 89% to 88%.

>The hyperparameter tuning improved the models, but not by much. This is most likely due to the fact that we have a high variance situation.

THANK YOU