PRODUS

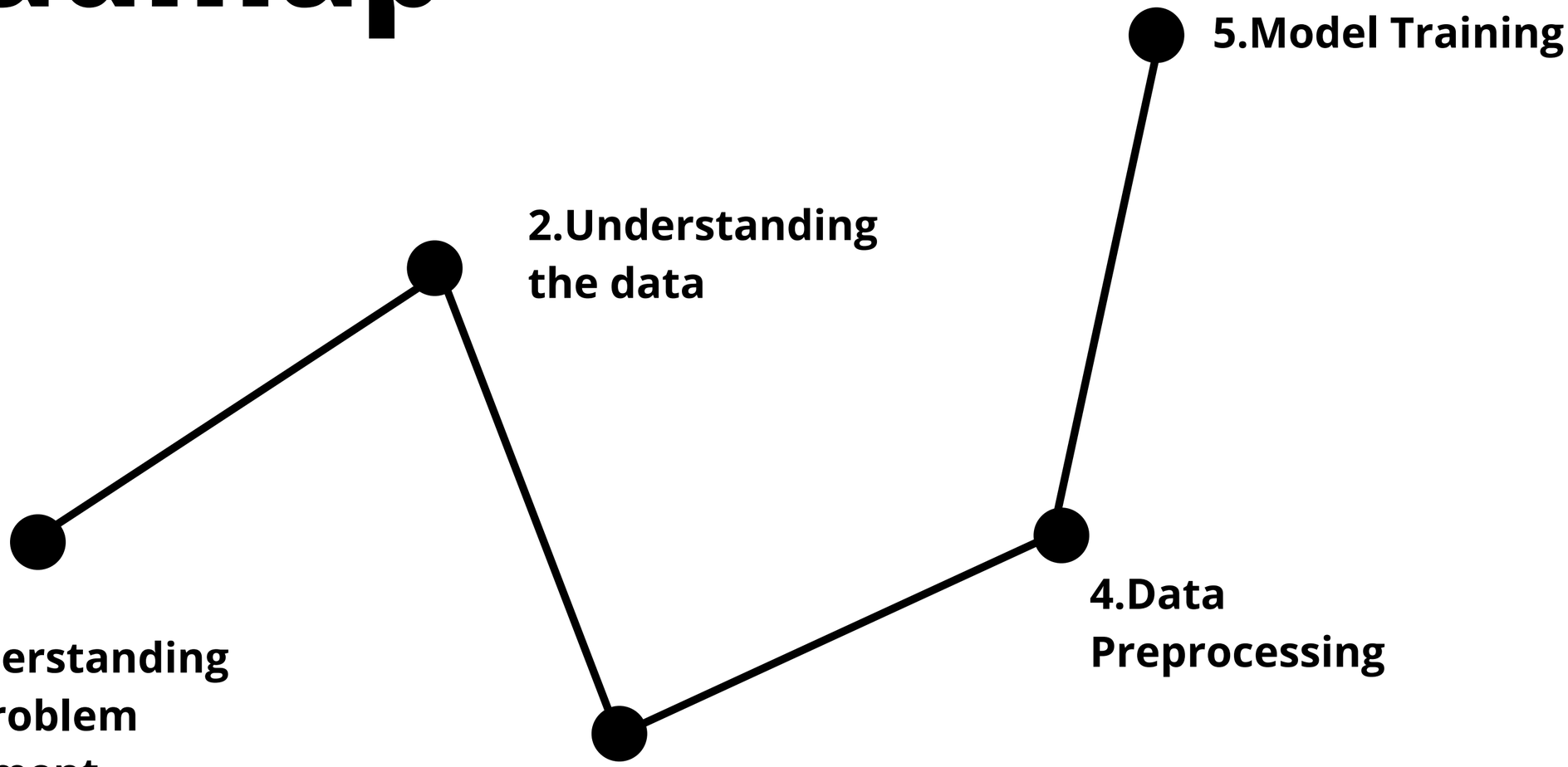
Hack-the-Data

Solution Presented by

PaperBoat

Dhrubojyoti Das | Nilanjan Sengupta

Roadmap


5.Model Training
2.Understanding the data
1.Understanding the problem statement
3.Exploratory Data analysis
4.Data Preprocessing

UNDERSTANDING THE PROBLEM STATEMENT

- XYZ is an Indian bank, whose one of the most profitable service is **Term Deposit.**
- XYZ has launched a market campaign to **contact customers on phone and advertise their product.**
- They want to **analyze the data** that they collected during campaigns **to get the insights about customers.**
- Lastly, they want to build a **model which automates the process of classification.**

UNDERSTANDING THE DATA

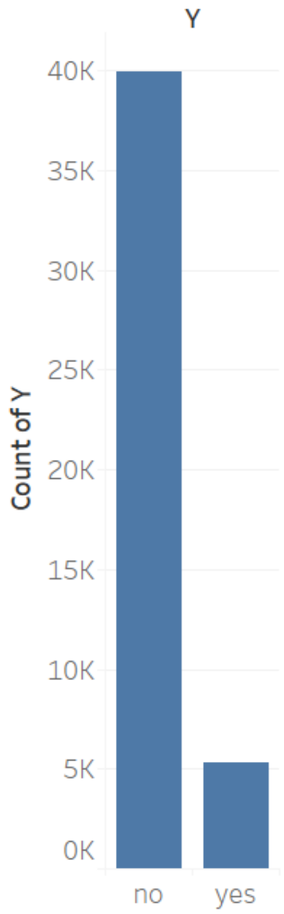
- The given data has **15 columns** : **7 Numeric** Columns & **8 Categorical** Columns
- There are no **missing values.**
- There are **no duplicates**.
- **64.02%, 4.76%, and 0.64%,10.02%** of 'p_outcome', 'education' ,'job' and 'communication ' columns has **'unknown'** entries.
- **11.70%** of all entries are '**yes**' under 'Y'.
- **71.53%** of the customers have '**balance**' between **0-2k.**

.

ANOMALIES IN DATA (THAT REQUIRES PREPROCESSING)

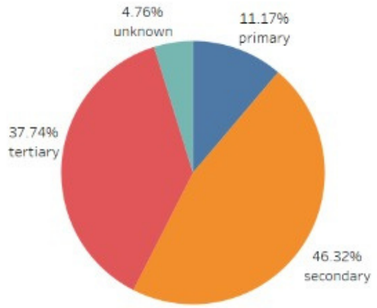
- The Dataset has categorical variables which requires Label Encoding.
- There are outliers in the dataset which requires scaling.
- The Dataset is highly imbalanced.

.

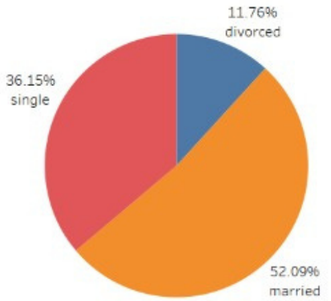


In the given dataset, **88.3% (39,922)** of the customers said NO in our previous campaign and only **11.70% (5,289)** customers AGREED to subscribe.

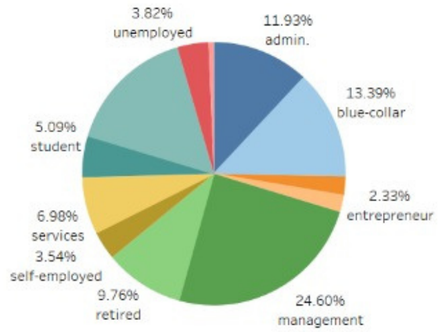
Exploring the previously converted clients



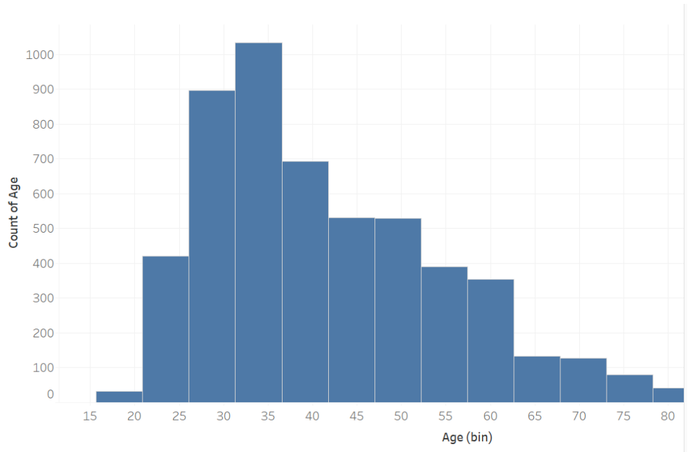
Education



Marital Status



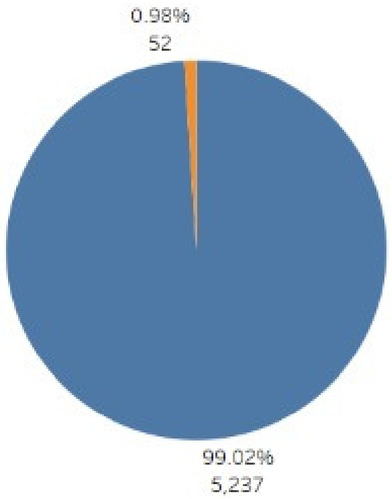
Jobs



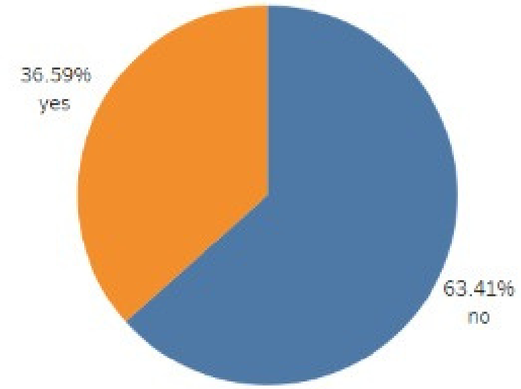
Age

- **84.06%** of the audience who subscribed for the term deposit had **at least secondary education**.
- **Married people (52.09%)** tend to subscribe to this scheme at a much higher rate than single and divorced people.
- **80%** of customers who have subscribed the term deposit are between the age of **25-55.**

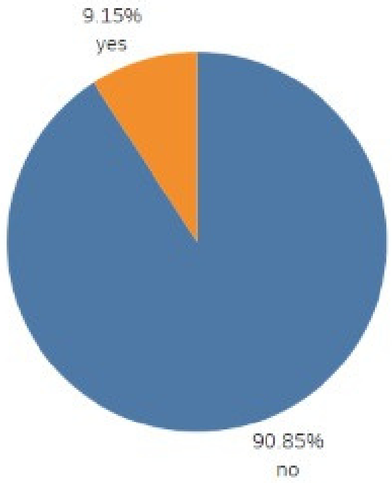
Exploring the previously converted clients



Default



Housing Loan



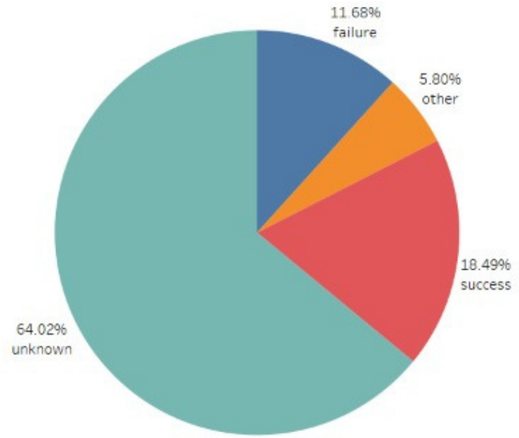
Personal Loan

Among the people subscribed to the previous campaign,

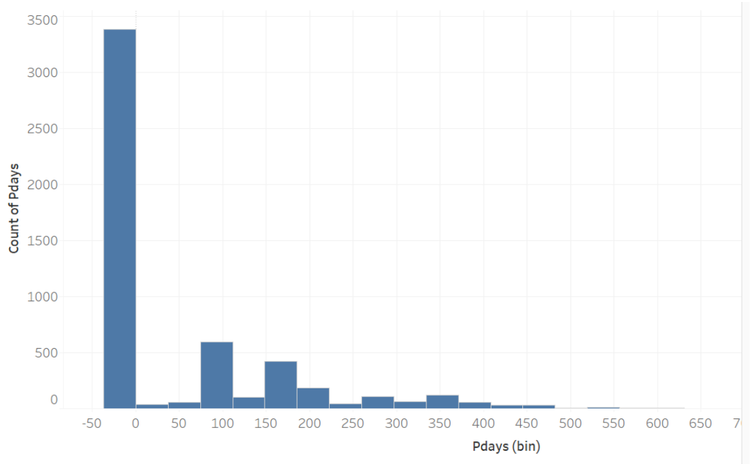
- **99.02%** of people didn't had any credit in default.
- **63.41%** of people didn't had any HOUSING LOAN.
- **90.85%** of people didn't had any personal loans upon them.

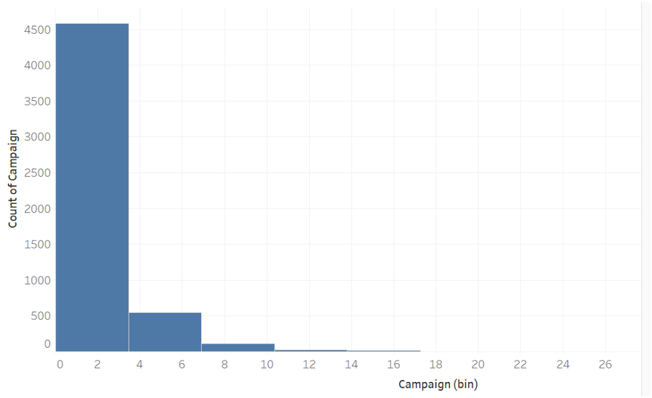
Exploring the previously converted clients

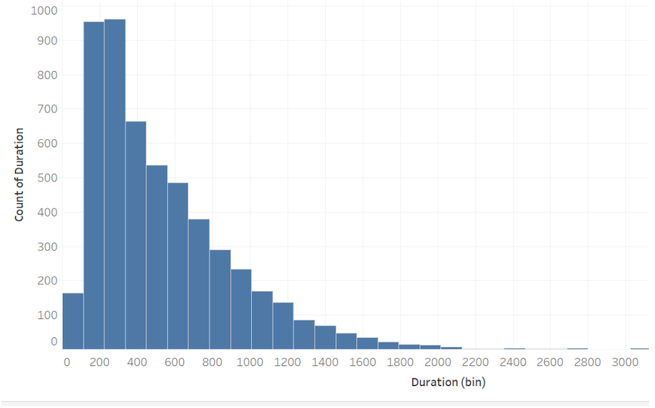
Communication Channels		
Cellular	4369	82.61%
Telephone	390	7.37%
Unknown	530	10.02%



P_outcome







- **63.98%** of the people who subscribed to the term deposit were last contacted **under one day**.
- **92.59%** of the customers who have subscribed were contacted **less than 5 times** during this campaign.
- The last contact duration for **89.42%** converted clients was **under 400 seconds**.
- For **64.02%** of cases, the outcome of the previous campaign was **unknown**.

THE MODEL

Data Description

	age	balance	day	duration	campaign	pdays	previous
count	45211.000000	45211.000000	45211.000000	45211.000000	45211.000000	45211.000000	45211.000000
mean	40.936210	1362.272058	15.806419	258.163080	2.763841	40.197828	0.580323
std	10.618762	3044.765829	8.322476	257.527812	3.098021	100.128746	2.303441
min	18.000000	-8019.000000	1.000000	0.000000	1.000000	-1.000000	0.000000
25%	33.000000	72.000000	8.000000	103.000000	1.000000	-1.000000	0.000000
50%	39.000000	448.000000	16.000000	180.000000	2.000000	-1.000000	0.000000
75%	48.000000	1428.000000	21.000000	319.000000	3.000000	-1.000000	0.000000
max	95.000000	102127.000000	31.000000	4918.000000	63.000000	871.000000	275.000000

- **We prepared a classification model based on the given dataset.**
- **Label Encoding was used to treat the categorical variables.**
- **Used Logistic Regression, Naive Bayes, XGB Classifier, Decision Tree to prepare the model.**
- **Used Confusion Matrix to evaluate the model accuracy.**

EVALUATION AND RESULTS

In [35]:
```
#Details about various models
rd
```

Out[35]:
```
{'Logistic Regression': {'Precision': 0.8544990349259355,
  'Recall': 0.8839986730067455,
  'Accuracy': 0.8839986730067455,
  'F1 Score': 0.8550026170862134},
 'Naive Bayes': {'Precision': 0.8572547638693171,
  'Recall': 0.8444100409156253,
  'Accuracy': 0.8444100409156253,
  'F1 Score': 0.8503006089728973},
 'XGB Classifier': {'Precision': 0.8936273255070852,
  'Recall': 0.9035718235098972,
  'Accuracy': 0.9035718235098972,
  'F1 Score': 0.8965604952197974},
 'Decision Tree': {'Precision': 0.8760644142312257,
  'Recall': 0.8789118655313503,
  'Accuracy': 0.8789118655313503,
  'F1 Score': 0.8774318419129685},
 'Random Forest Classifier': {'Precision': 0.8904500856406075,
  'Recall': 0.9036824062811014,
  'Accuracy': 0.9036824062811014,
  'F1 Score': 0.89078180795337}}
```

- We used Confusion Matrix for evaluating our Models.
- XGB Classifier and Random Forests were the best performing models.
- We chose Random Forests over XGB Classifier because there were outliers in the data and the dataset was highly imbalanced. Random Forest deals with these issues on its own in a better manner, and it performs better in case of yes-no classification problem.

Summary

We solved a classification problem for XYZ bank for their 'Term Deposit' service. The dataset did not have any missing values or duplicate values. We did thorough Exploratory Data Analysis to find important insights and anomalies in the data. It contained categorical values which required label encoding. The dataset was highly imbalanced and contained outliers in many columns. We used Confusion Matrix for Model Evaluation, where we found out that XGB Classifier and Random Forest were the best performing models. We chose Random Forest over other models because it handles outliers and is better for a yes no classification problem.

Thank You