Bank Marketing Effectiveness Prediction

Classification Project by
Jayesh Dahiwale

1. Table of Content

Introduction of Project

- The Data is related to direct marketing campaigns(phone calls) of a Portuguese banking institution.The marketing campaigns were based on phone calls. Often, more than one call to the same client, in order to access if the product(bank term deposit) would be **'YES'** or **not 'NO' subscribed**. The classification goal is to predict if the client will subscribe to the term deposit (variable y).

Term Deposit

- A term deposit is a type of deposit account held at a financial institution where money is locked up for some set period of time.

- Term deposits are usually short-term deposits with maturities ranging from one month to a few years.

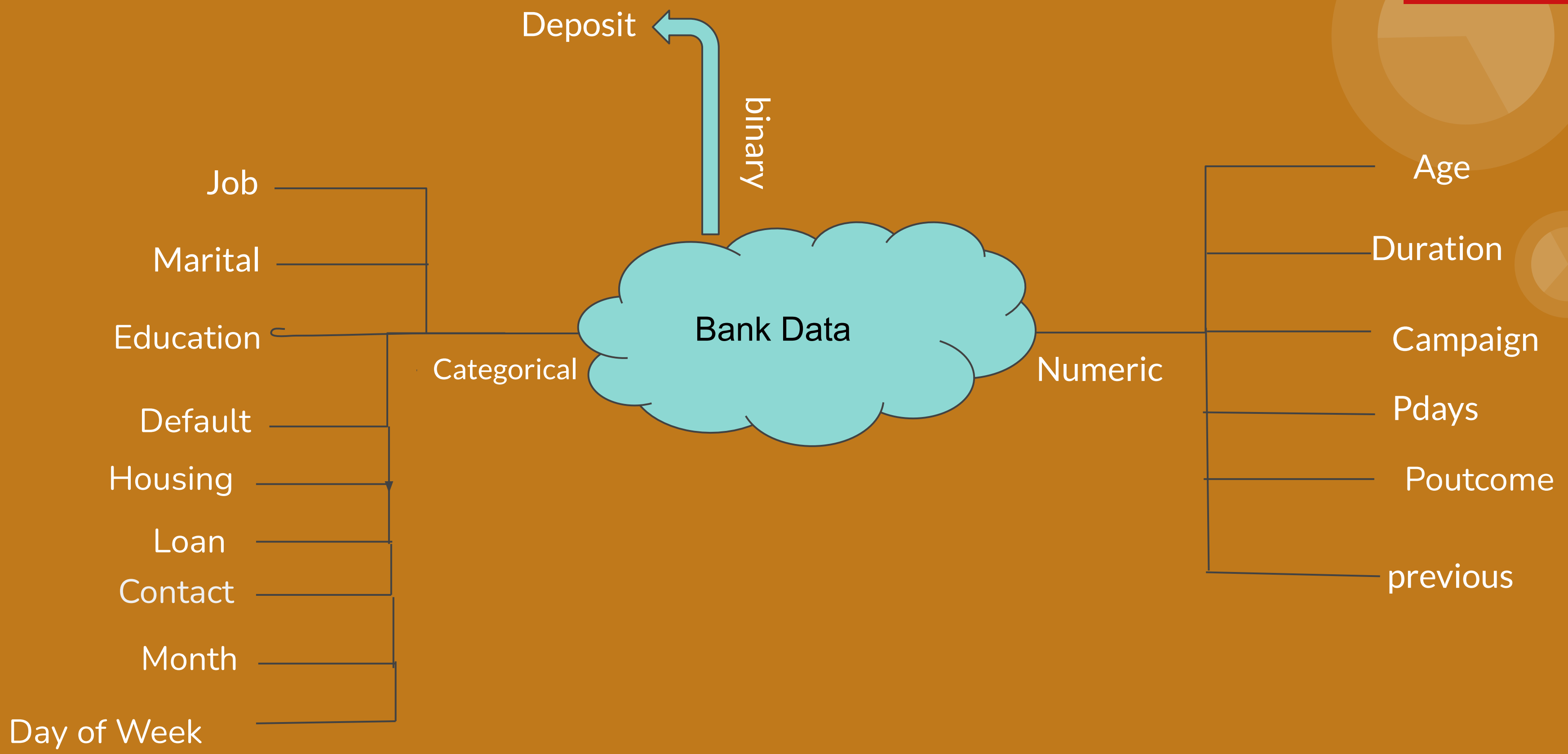
Problem Statement

AIM

- *Using the given dataset and developing a ML model out of it with **TARGET:Deposit (YES/NO)** For Classifying a new customer based on given features and also **Determining the most relevant features** of classification*

- *We need to classify the customers with very good accuracy so that organization can only contact those customers which having high chances of subscribing to the term deposit.*

Data Summary Bank Dataset



Variables in Brief

1. Age - **(numeric)** Age of the customer
2. Job - **(Categorical)** Represents the diferent types of Jobs. eg. {'Admin','BlueCollar','Entrepreneur','Housemaid','Management','Retired'} etc.
3. Marital - **(Categorical)** {Divorced, Married, Single,Unknown.} Note: Divorced meaning Divorced or Widowed.
4. Education - **(Categorical)** '' {'Basic4y','Basic6y','Basic9y','highschool','illiterate'} etc.
5. Dafault - **(Cateogircal)** Has credit in default {'YES','NO', 'Unknown'}
6. Housing - **(Categorical)** has housing loan ? {'YES' , 'NO' , 'Unknown'}

7. loan - **(Categorical)** has personal loan ? {'YES' , 'NO' , 'Unknown'}

8. Contact - **(Categorical)** contact communication types {'Cellular', 'Telephone'}
9. Month - **Categorical** Last contact month of the year {'Jan' , 'Feb' , 'Mar'..........,'Dec'}

Variables in Brief

10. day_of_week - (**Categorical)** Last contact day of week. {'Mon','Tue','Wed','Thu','Fri','Sat'} etc.
11. Duration - **(numeric)** Last contact duration in seconds. **Important note:** This feature highly attributed to the target value Y. i.e. if the call duration is zero the output is "NO". Also after the end of the call y is obviously known. Thus this feaute should only be included for benchmark purposes and should be discarded if the intention is to build realstic predictive model.

12. Campaign - **(numeric)** No of contacts performed during this campaign and for this client

13. pDays **- (numeric)** No of days that passed by after the client was last contacted. **999** means cliendt was not contacted.
14. previous - **(numeric)** number of contacts performed before this campaign for this client
15. pOutcome - **(categorical)** Outcome of the previous campaingn. {'Success','Failure','Non-existence'}
16. y-target Variable - **(Binary)** has the client subscriber to a term deposit. {'Yes' , 'NO'}

Graphs used for EDA :

- Count Plot
- Bar Plot
- Dist Plot
- Box Plot
- HeatMap
- Pie Chart

Python Libraries used for EDA :

- Matplotlib
- Numpy
- Pandas
- Seaborn
- ScikitLearn

Bank Dataset :

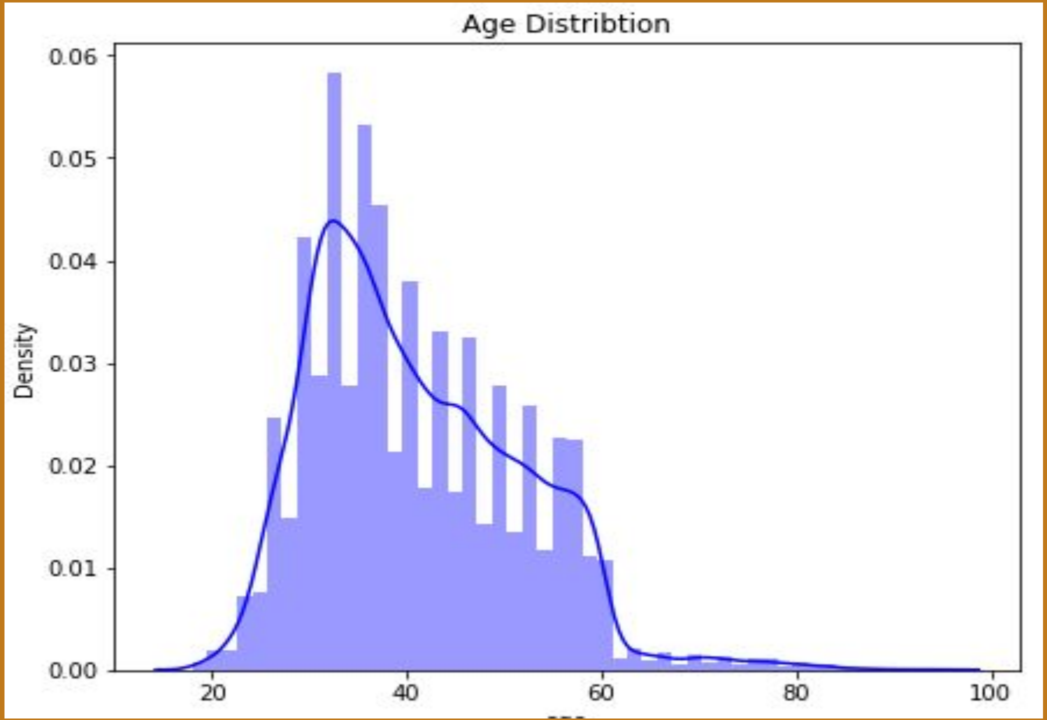
- Lets check the null values in the dataset:

#	Column	Non-Null Count	Dtype
0	age	45211 non-null	int64
1	job	45211 non-null	object
2	marital	45211 non-null	object
3	education	45211 non-null	object
4	default	45211 non-null	object
5	balance	45211 non-null	int64
6	housing	45211 non-null	object
7	loan	45211 non-null	object
8	contact	45211 non-null	object
9	day	45211 non-null	int64
10	month	45211 non-null	object
11	duration	45211 non-null	int64
12	campaign	45211 non-null	int64
13	pdays	45211 non-null	int64
14	previous	45211 non-null	int64
15	poutcome	45211 non-null	object
16	y	45211 non-null	object

Features	count
age	0
job	0
marital	0
education	0
default	0
balance	0
housing	0
loan	0
contact	0
day	0
month	0
duration	0
campaign	0
pdays	0
previous	0
poutcome	0
y	0

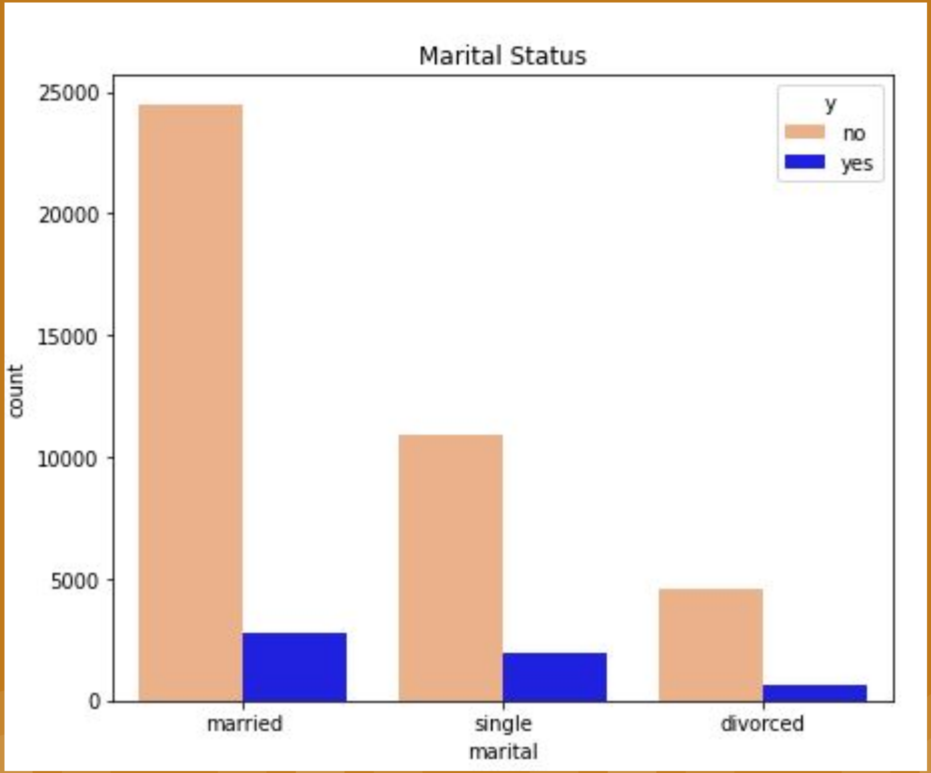
- We can see that in the given dataset there is no null value
- There are total 45,200 rows and 17 columns in the dataset

1. Visualizing the distribution of "Ages"



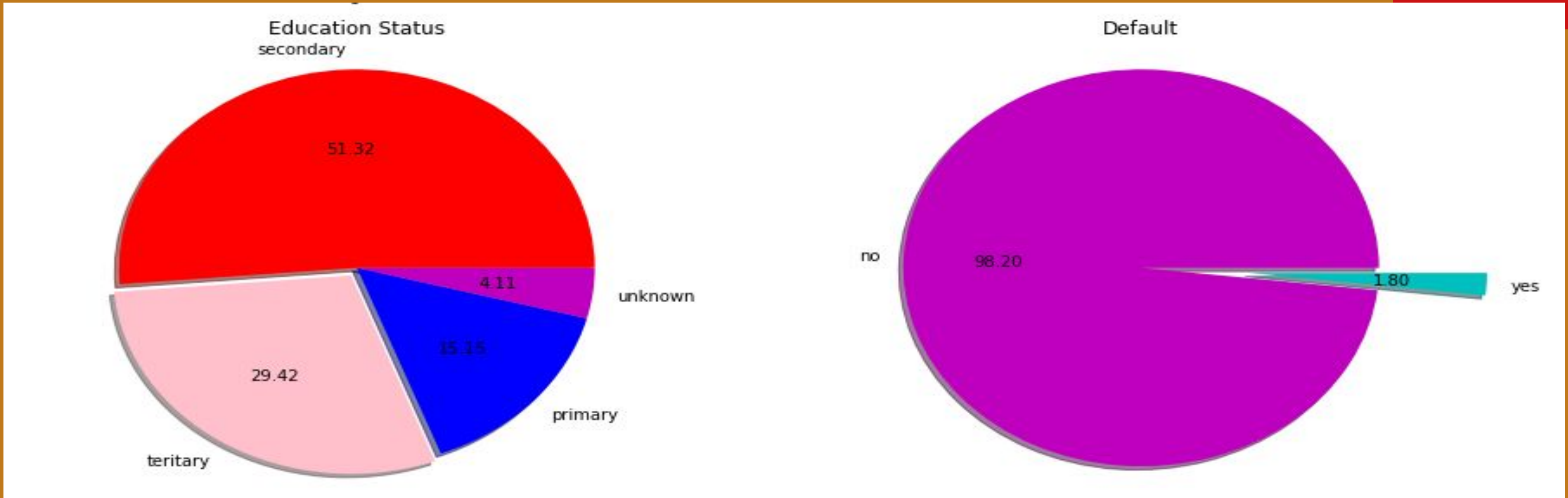
- Here in the graph we can see that the distribution is **positively skewed.**
- It shows that there may exist some **outliers.**
- We can see the **mean** is around **40.93** and **median** is **39**.

2. Visualizing the countplot of “Marital Status”



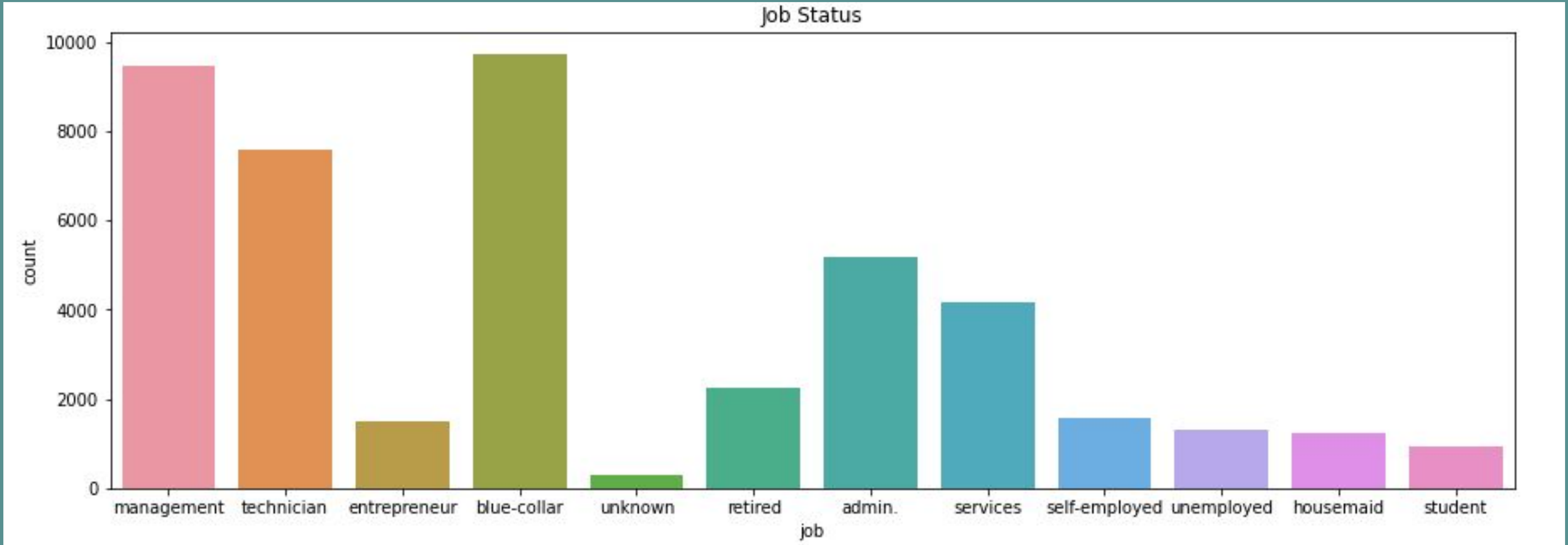
- From the marital status count bar plot , we can conclude that **very less percentage** of **married couple** opt for Term Deposit. But those who are **single** has higher proportion of opting for term deposit.

3. Visualizing "Education" and "Default" status



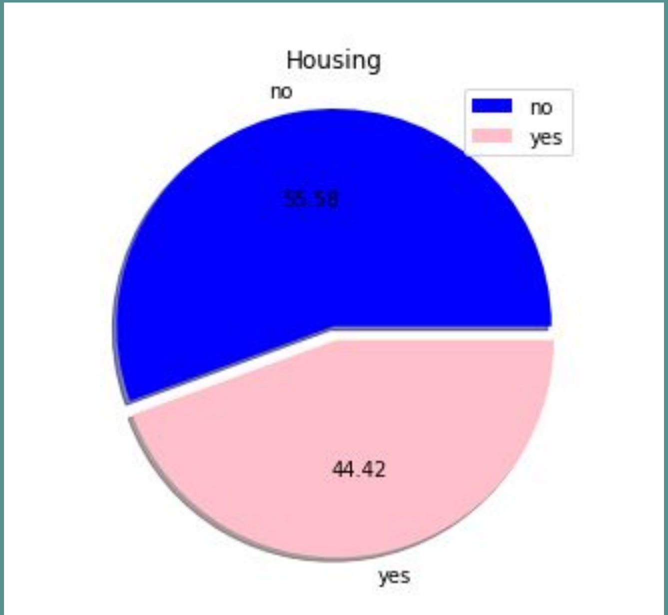
- For education status , one can infer that most of the customers have completed with their **Secondary education.**
- Very less no of customers have credit in default.

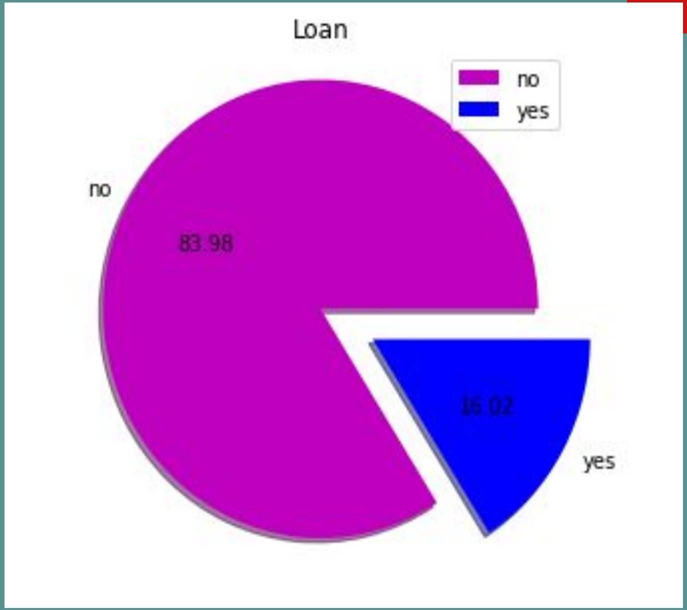
4. Visualizing Job Status



- Here , I can observe that most customers have blue collar Jobs. And least number of customer are students.
- Customer Having Blue Collar Jobs : **9732 ..** Student Customer : **938**

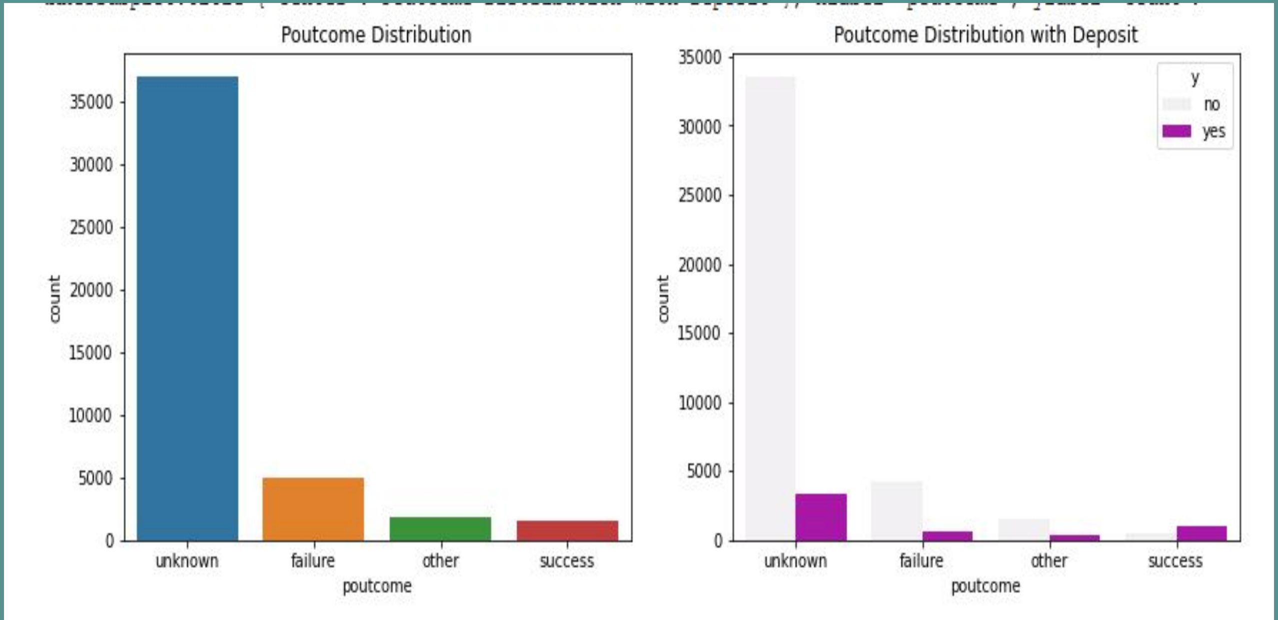
4. Visualizing Housing And Loan Status





- Here we can see that **44.42%** customer has opter for **housing loan.**
- In case of Personal Loan, **16.02%** of the people opted for it.

5. Previous Outcome distribution with Deposit



		y
poutcome	y	
failure	no	4283
	yes	618
other	no	1533
	yes	307
success	no	533
	yes	978
unknown	no	33573
	yes	3386

- We can see higher no. of customers has unknown previous outcome. But we can observe one thing even if the **Previously** those people who successfully subscribed to the term, those are in high numbers of renewing their subscription.
- From poutcome, those who has successfully subscribed to deposit has higher proportion that they subscribe to deposit.

Removal Of Outliers Job Columns

- In **'Job'** columns , we found that the total percentage of **outliers are 1.08%.**
- So decided to replace the **lower outliers with 10.5**
- Decided to replace the **upper outliers with 70.5**

Removal Of Outliers Campaign Columns

- It contains, we found out the total percentage of **outliers are 6.78%**
- So decided to replace the **lower outlier with -2.0**
- Decided to replace the **upper outlier with 6.0**

Model Selection

AI

- The metric for selection of Classification Model is **Cross Validation Scores.**

Random Forest Classifier :

- The **Random Forest Classifier** score for 5 cross-fold-validation is [0.90643599, 0.90268549, 0.90199336, 0.90531561, 0.90420819]

- The mean score is 0.90412772879

XGBoost Classifier :

- The **XGBoost Classifier** score for 5 cross-fold validation is [0.90560554 0.90033223 0.90268549 0.90739203 0.90753045]

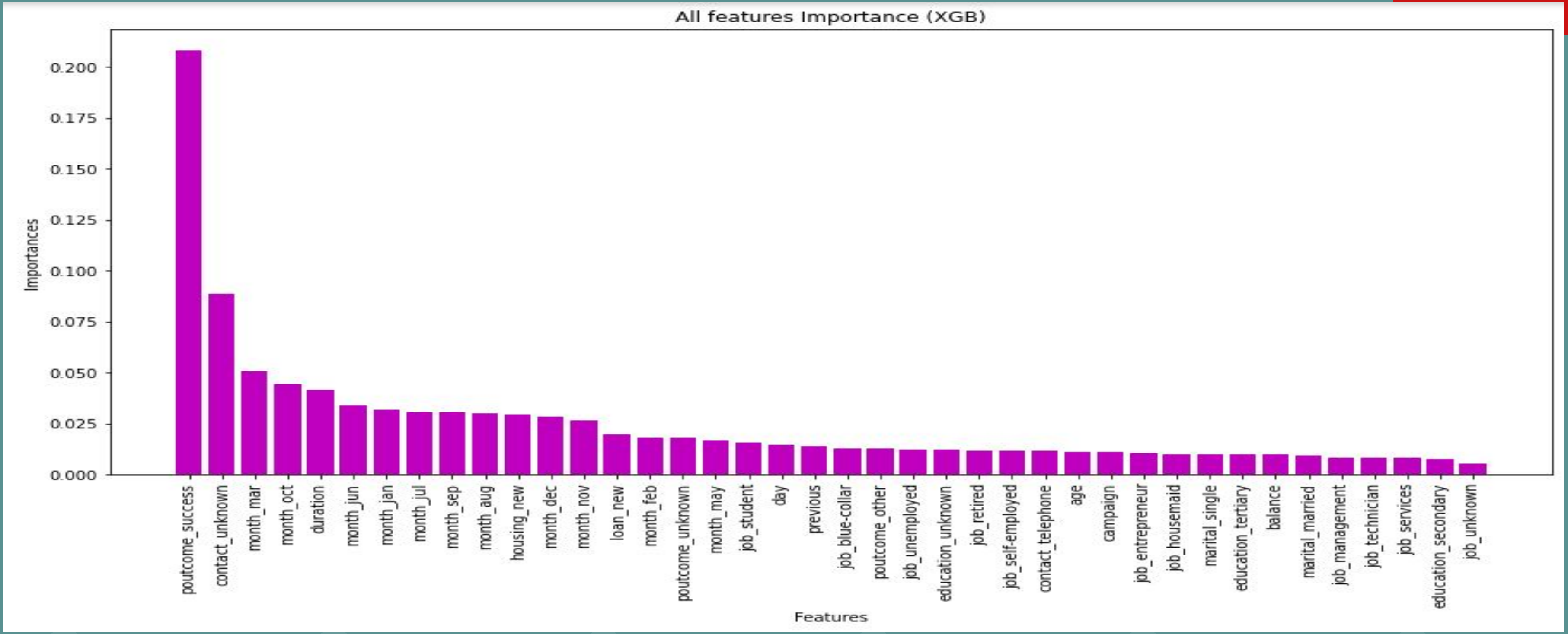
- The mean score is 0.90470914

- Finally I chose, XGBoost algorithm for our current classification model.
- Important feature used for classifications are :

	feature_name	score
0	poutcome_success	0.208168
1	contact_unknown	0.088568
2	month_mar	0.050951
3	month_oct	0.044368
4	duration	0.041616
5	month_jun	0.034124
6	month_jan	0.031845
7	month_jul	0.030888
8	month_sep	0.030875
9	month_aug	0.029980
10	housing_new	0.029795
11	month_dec	0.028590
12	month_nov	0.026859
13	loan_new	0.019642
14	month_feb	0.018186
15	poutcome_unknown	0.018026
16	month_may	0.016863
17	job_student	0.015735
18	day	0.014676

18	day	0.014676
19	previous	0.014205
20	job_blue-collar	0.013127
21	poutcome_other	0.012901
22	job_unemployed	0.012354
23	education_unknown	0.012073
24	job_retired	0.011831
25	job_self-employed	0.011699
26	contact_telephone	0.011598
27	age	0.011388
28	campaign	0.011139
29	job_entrepreneur	0.010416
30	job_housemaid	0.010127
31	marital_single	0.010116
32	education_tertiary	0.009916
33	balance	0.009814
34	marital_married	0.009665
35	job_management	0.008360
36	job_technician	0.008228
37	job_services	0.008074
38	education_secondary	0.007867
39	job_unknown	0.005348

Feature Importance Plot



- Here , poutcome_success is the most important feature in classification.

Classification Matrix

	precision	recall	f1-score	support
0	0.93	0.97	0.95	7942
1	0.67	0.49	0.57	1089
accuracy			0.91	9031
macro avg	0.80	0.73	0.76	9031
weighted avg	0.90	0.91	0.90	9031

- Receiver Operating Characteristic score is : 0.728
- This score is good and acceptable,
- Accuracy of the model is 0.91.

Conclusion

1) **Very less percentage** of **married couple** opt for Term Deposit. But those who are **single** has higher proportion of opting for term deposit.

2) For education status , one can infer that most of the customers have completed with their **Secondary education.**

3) Very less no of customers have credit in default.

4) Here , I can observe that most customers have blue collar Jobs. And least number of customer are students.

5)Customer Having Blue Collar Jobs : **9732 ..** Student Customer : **938**

6) Here we can see that **44.42%** customer has opter for **housing loan.**

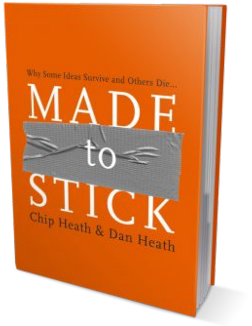
7) In case of Personal Loan, **16.02%** of the people opted for it.

Continued…….

8.We can see higher no. of customers has unknown previous outcome. But we can observe one thing even if the **Previously** those people who successfully subscribed to the term, those are in high numbers of renewing their subscription.

9. From poutcome, those who has successfully subscribed to deposit has higher proportion that they subscribe to deposit

Thank You!


Why Some Ideas Survive and Others Die...
MADE
to
STICK
Chip Heath & Dan Heath