Term Deposit Marketing

Background:

We are a small startup focusing mainly on providing machine learning solutions in the European banking market. We work on a variety of problems including fraud detection, sentiment classification and customer intention prediction and classification.
We are interested in developing a robust machine learning system that leverages information coming from call centre data.
Ultimately, we are looking for ways to improve the success rate for calls made to customers for any product that our clients offer. Towards this goal we are working on designing an ever-evolving machine learning product that offers high success outcomes while offering interpretability for our clients to make informed decisions.

Data Description:

The data comes from direct marketing efforts of a European banking institution. The marketing campaign involves making a phone call to a customer, often multiple times to ensure a product subscription, in this case a term deposit. Term deposits are usually short-term deposits with maturities ranging from one month to a few years. The customer must understand when buying a term deposit that they can withdraw their funds only after the term ends. All customer information that might reveal personal information is removed due to privacy concerns.

Attributes:
age : age of customer (numeric)
job : type of job (categorical)
marital : marital status (categorical)
education (categorical)
default: has credit in default? (binary)
balance: average yearly balance, in euros (numeric)
housing: has a housing loan? (binary)
loan: has personal loan? (binary)
contact: contact communication type (categorical)
day: last contact day of the month (numeric)
month: last contact month of year (categorical)
duration: last contact duration, in seconds (numeric)
campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)

- Output (desired target):

y - has the client subscribed to a term deposit? (binary)

Expletory Data Analysis

- The data consist of 40000 customer attributes and the target label that correspond to subscription status.

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	y
0	58	management	married	tertiary	no	2143	yes	no	unknown	5	may	261	1	no
1	44	technician	single	secondary	no	29	yes	no	unknown	5	may	151	1	no
2	33	entrepreneur	married	secondary	no	2	yes	yes	unknown	5	may	76	1	no
3	47	blue-collar	married	unknown	no	1506	yes	no	unknown	5	may	92	1	no
4	33	unknown	single	unknown	no	1	no	no	unknown	5	may	198	1	no

- Checking the data for missing values
  - age 0
  - job 235
  - marital 0
  - education 1531
  - default 0
  - balance 0
  - housing 0
  - loan 0
  - contact 12765
  - day 0
  - month 0
  - duration 0
  - campaign 0
  - y 0

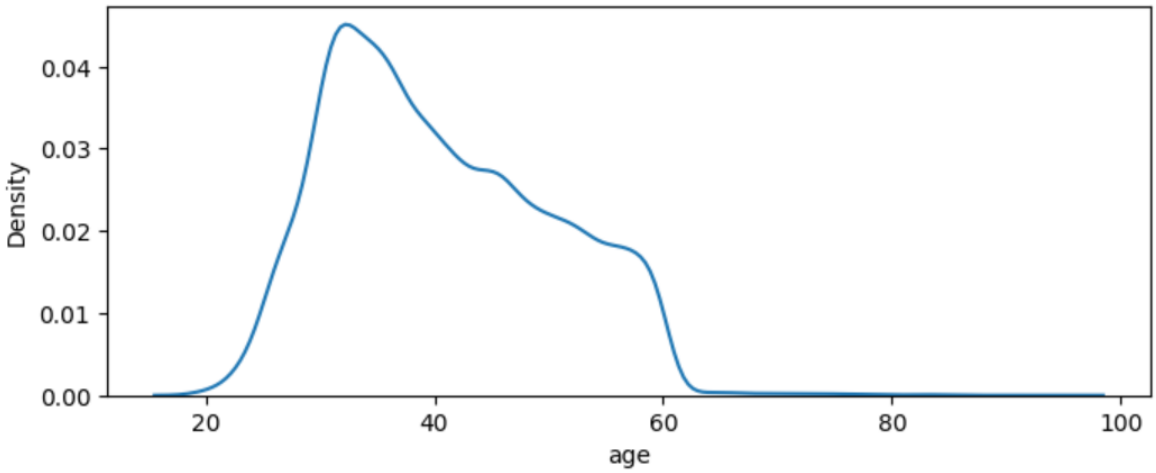
The data has total of 14531 missing value, all categorical type.

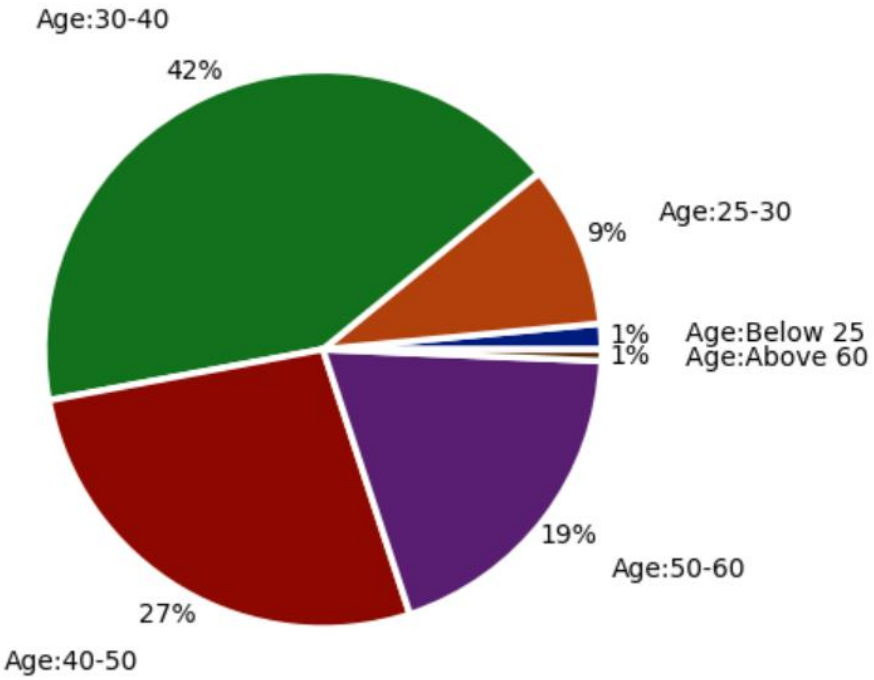
To treat this, mode imputation is applied to the attributes with missing values replacing missing values with the most frequent in each attribute.

- Numerical data statistics

	age	balance	day	duration	campaign
count	40000.000000	40000.000000	40000.000000	40000.000000	40000.000000
mean	40.544600	1274.277550	16.017225	254.824300	2.882175
std	9.641776	2903.769716	8.278127	259.366498	3.239051
min	19.000000	-8019.000000	1.000000	0.000000	1.000000
25%	33.000000	54.000000	8.000000	100.000000	1.000000
50%	39.000000	407.000000	17.000000	175.000000	2.000000
75%	48.000000	1319.000000	21.000000	313.000000	3.000000
max	95.000000	102127.000000	31.000000	4918.000000	63.000000

Age Attribute



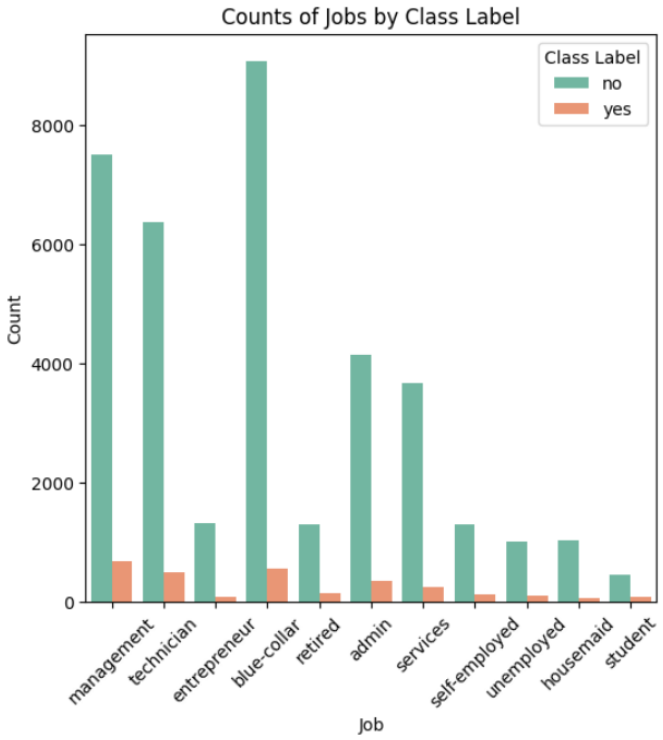


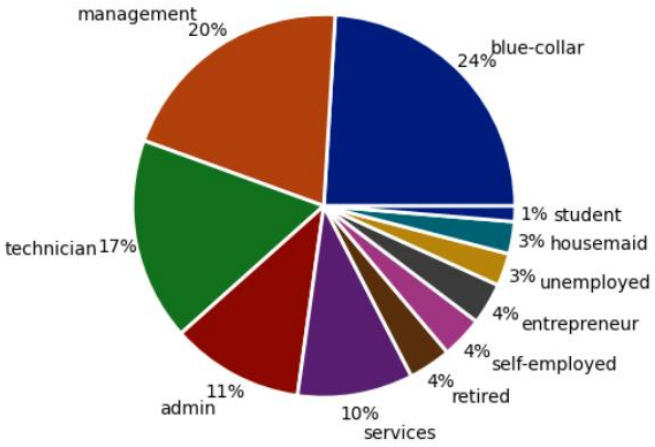
- The figures above shows that majority of the customers are from age 30 to 50, while there is a small minority of customers with age below 25 and above 60.

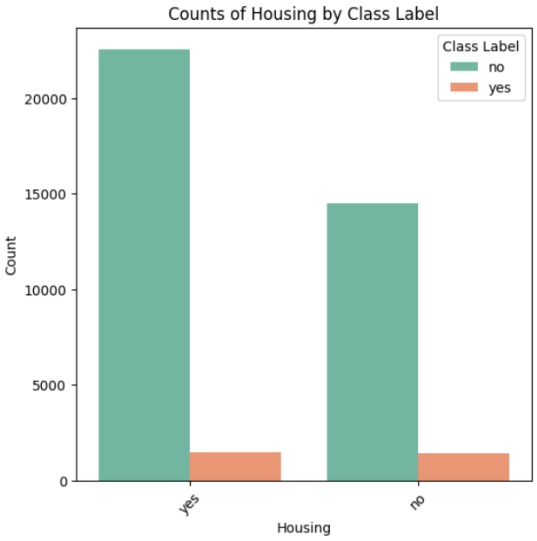
Job Attribute

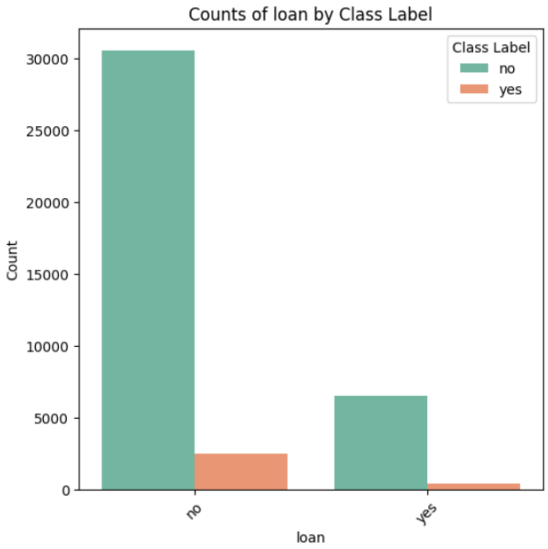
The count plot shows that majority of the customers are from job class " blue-collar.

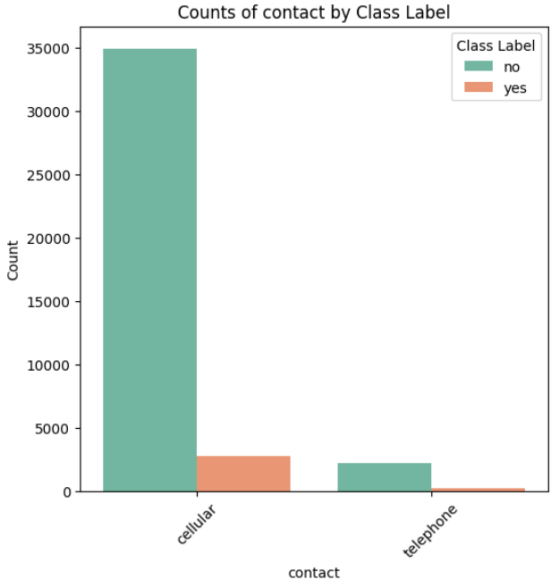
While subscribed customers majority are from class "management " and class " blue-collar"

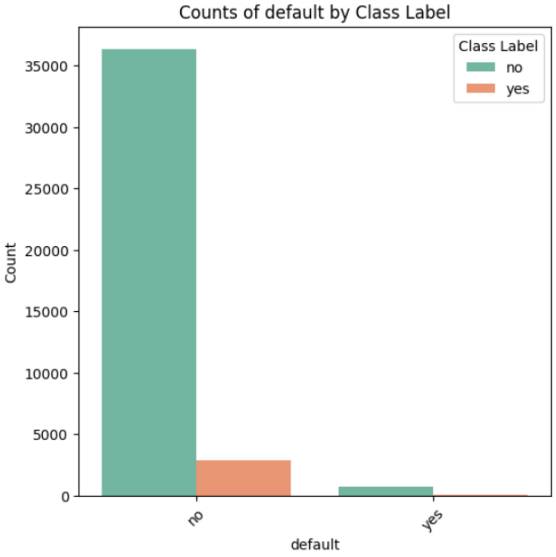


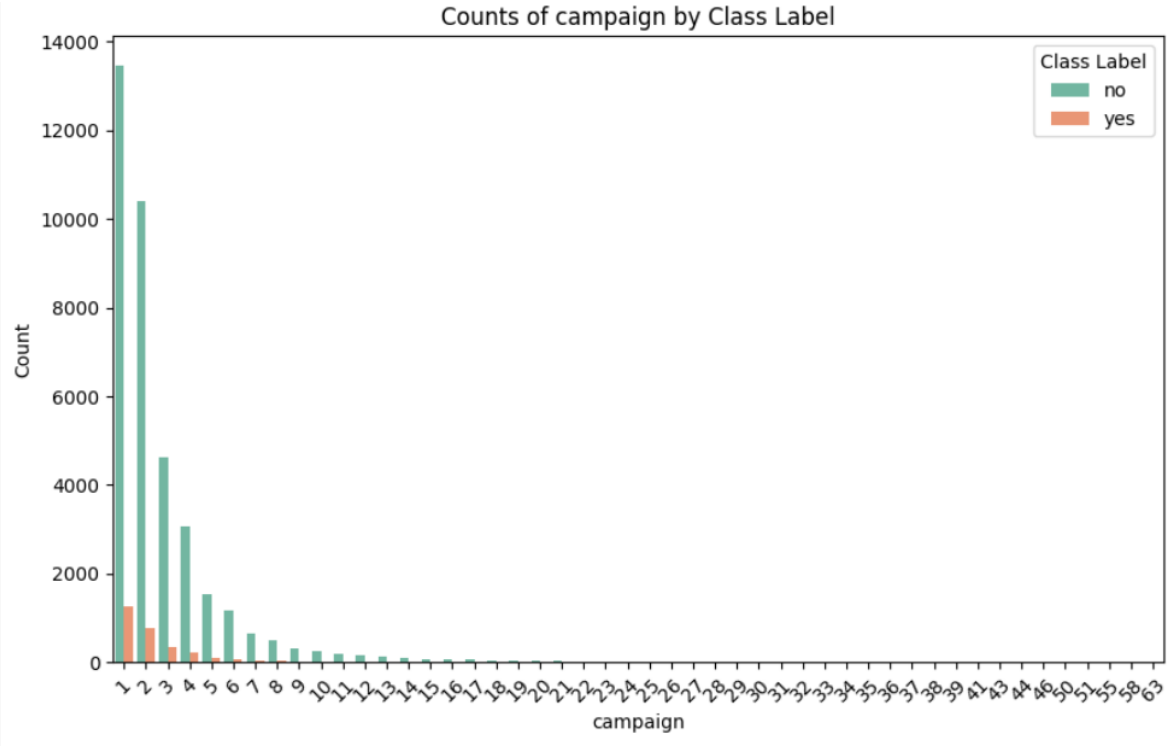



Counts of Housing by Class Label
Class Label
no
yes
Count
0
5000
10000
15000
20000
yes
no
Housing


Counts of loan by Class Label
Class Label
no
yes
Count
0
5000
10000
15000
20000
25000
30000
no
yes
loan


Counts of contact by Class Label
Class Label
no
yes
Count
0
5000
10000
15000
20000
25000
30000
35000
cellular
telephone
contact

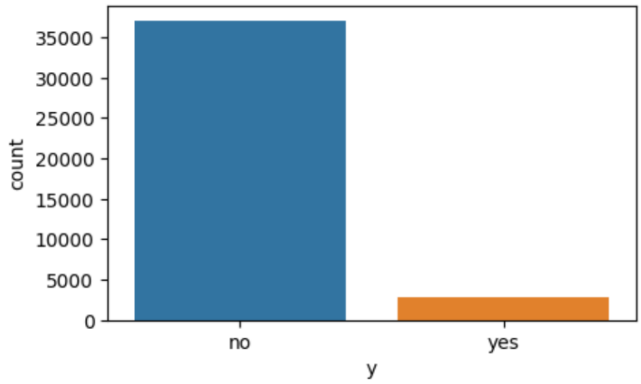

Counts of default by Class Label
Class Label
no
yes
Count
0
5000
10000
15000
20000
25000
30000
35000
no
yes
default



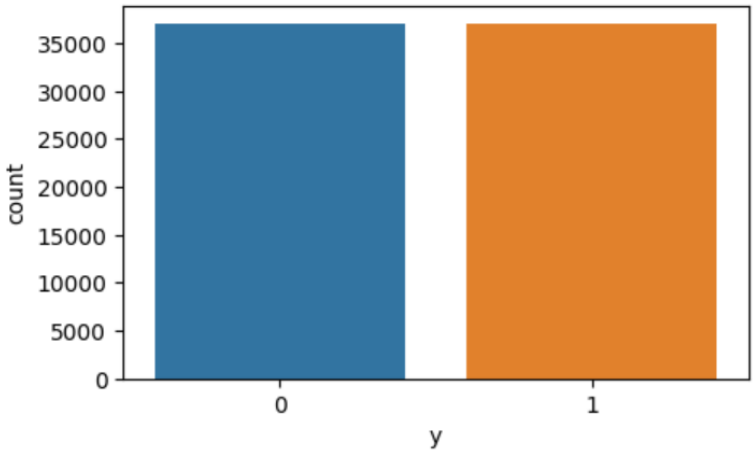
Majority of customers gets contacted 1-3 time, while subscription occur after 1-2 call on average.

Target label Y (Subscription Status)

The data target variable is imbalance and has 93% of the customer with no subscription, this will affect the model performance and show bias toward the majority class.



SMOTE (Over Sampling) is applied to artificially generate data points for the minority class, result is a balanced data set



Model and conclusion

After analysing the data, treating missing values, and fixing data imbalance issue, 6 machine learning algorithms were used to train and test the

- SVM: F1 score = 0.878
- Logistic Regression: F1 score = 0.837
- Decision Tree: F1 score = 0.929
- Random Forest: F1 score = 0.940
- K Nearest Neighbour: F1 score = 0.931
- Xgboost: F1 score = 0.959

The initial result shows that xgboost achieved the best f1 score, hyperparameter tuning is applied using grid search to find the optimal parameters.

The optimized xgboos with parameters

- gamma: 0.1
- learning rate: 0.3
- max depth: 6
- n estimators: 150

And an f1 score of 0.97