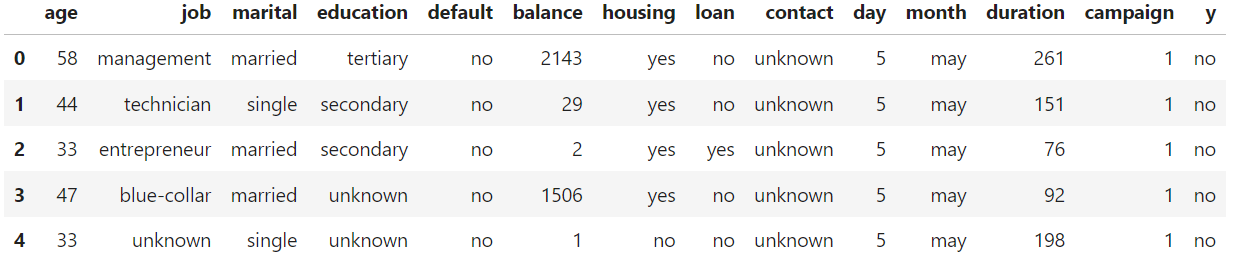
**Term Deposit Marketing**

**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.



* Checking the data for missing values
* age 0
* job 235

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

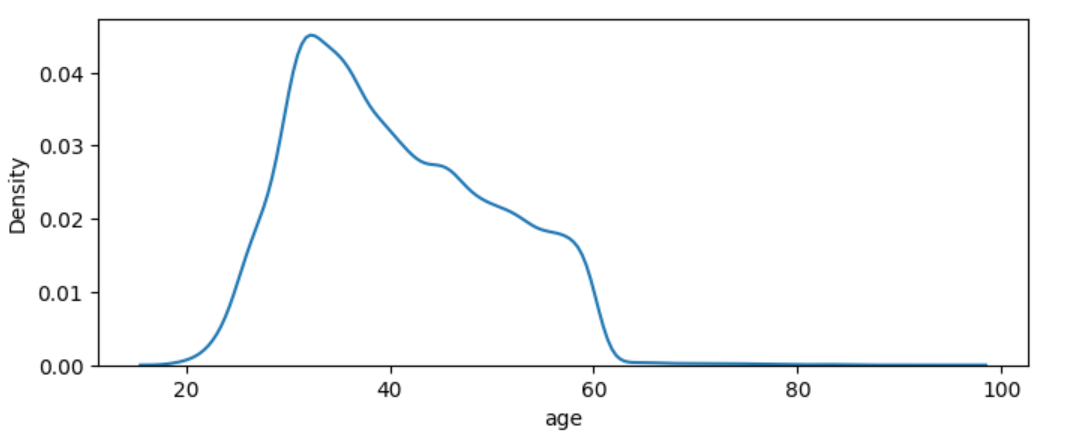
To treat this, mode imputation is applied to the attributes with missing values replacing missing values with the most frequent in each attribute.

* marital 0
* education 1531
* default 0
* balance 0
* housing 0
* loan 0
* contact 12765
* day 0
* month 0
* duration 0
* campaign 0
* y 0
* Numerical data statistics

A screenshot of a table with numbers

Description automatically generated

**Age Attribute**

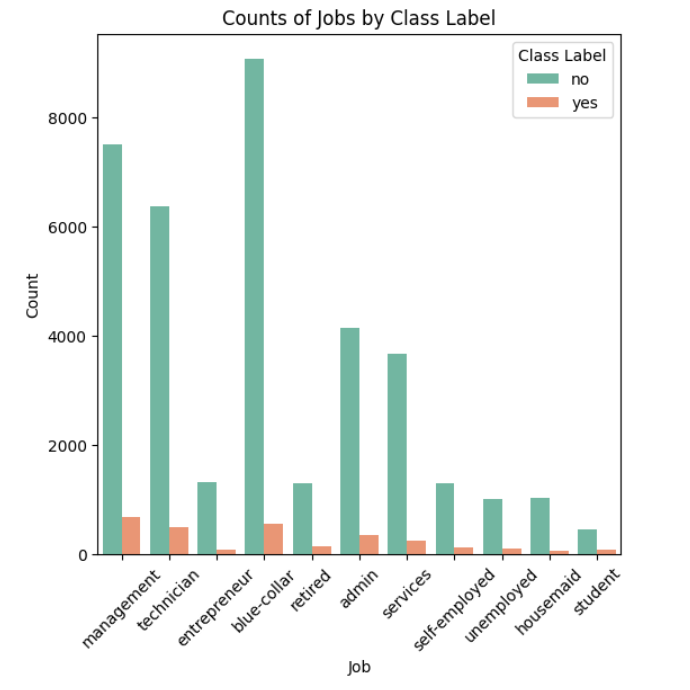


A pie chart with different colored sections

Description automatically generated

* 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.

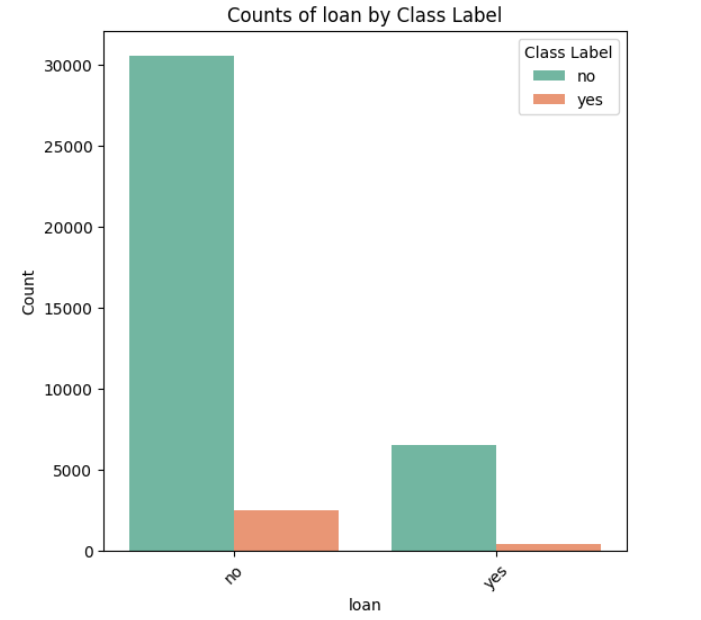
While subscribed customers majority are from class “management “ and class “ blue-collar”

A pie chart with numbers and text

Description automatically generated

A graph of a number of bars

Description automatically generated with medium confidence



A graph of a number of numbers

Description automatically generated with medium confidenceA graph with green and orange bars

Description automatically generated

A graph of a person with a number of numbers

Description automatically generated with medium confidence

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.

A blue and orange rectangular graph

Description automatically generated

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

A blue and orange rectangular bars

Description automatically generated

**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