Customer Conversion Prediction

Problem Statement:

You are working for a new-age insurance company and employ multiple outreach plans to sell term insurance to your customers. Telephonic marketing campaigns still remain one of the most effective ways to reach out to people however they incur a lot of cost. Hence, it is important to identify the customers that are most likely to convert beforehand so that they can be specifically targeted via call. We are given the historical marketing data of the insurance company and are required to build a ML model that will predict if a client will subscribe to the insurance.

DATASET:

The historical sales data is available as a compressed file here.

Features:

- age (numeric)
- job : type of job
- marital : marital status
- educational_qual : education status
- call_type : contact communication type
- day: last contact day of the month (numeric)
- mon: last contact month of year
- dur: last contact duration, in seconds (numeric)
- num_calls: number of contacts performed during this campaign and for this client
- prev_outcome: outcome of the previous marketing campaign (categorical: "unknown","other","failure","success")

Output variable (desired target):

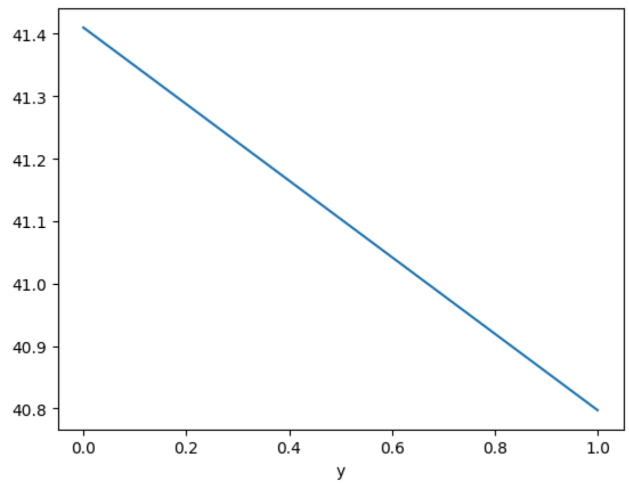
- y - has the client subscribed to the insurance?

APPROACH:

1. Import the required packages.
2. Load the dataset.
3. Clean the dataset.
   a. Remove duplicates(total data 45211 after removing duplicates total data 45205).
   b. Missing values(no Nan values in the dataset).
   c. Data Type conversion(no incorrect format).
   d. Structured dataset.
   e. Remove outliers(column ‘age’,’dur’,’num_calls’ consists of outliers).

4. Target variable y has maximum ‘no’ so mapped ‘no’ as 1 and ‘yes’ as 0 to perform EDA.
5. EDA(Exploratory Data Analysis) and Encode.
    a. Group the column ‘y’ and mean of column ‘age’ and plot.

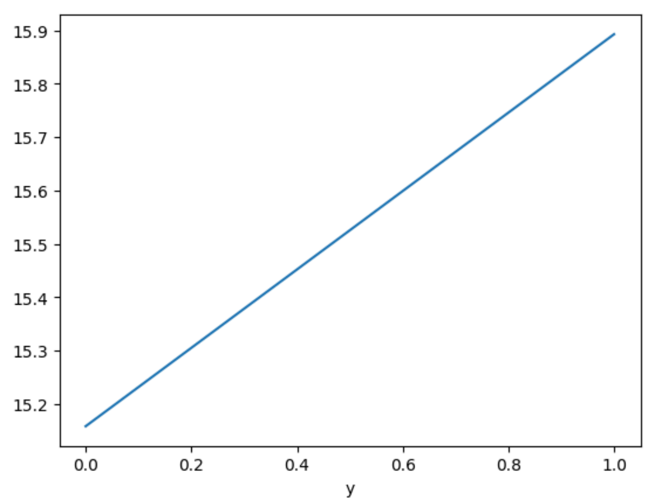
y
0   41.410096
1   40.797362



From this can conclude that people above age 41 will subscribe insurance.And people age below 41 will not subscribe insurance.

    b. Group the column ‘y’ and mean of column ‘day’ and plot.

y
0   15.158253
1   15.892825



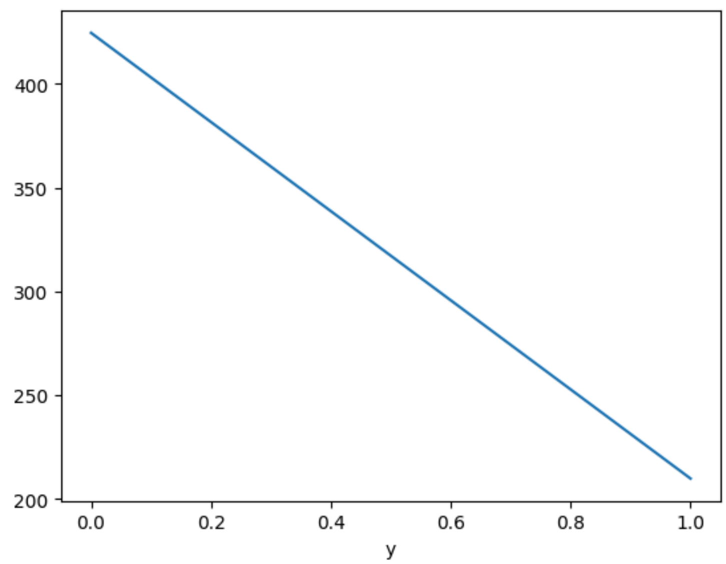
From this can conclude that days above 15 will not subscribe insurance and days below 15 will subscribe insurance.

c. Group the column ‘y’ and mean of column ‘dur’ and plot.

y

0 424.640953

1 209.822352



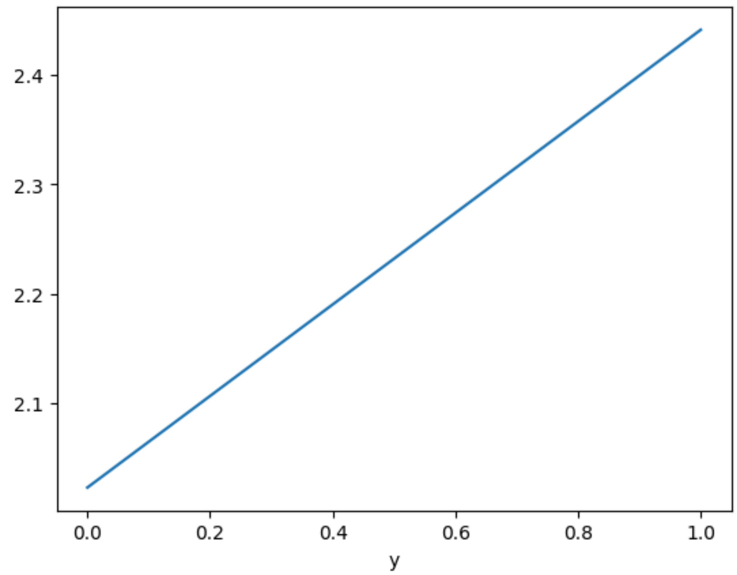
From this can conclude that dur above 400 will subscribe insurance and dur below 250 will not subscribe insurance.

d. Group the column ‘y’ and mean of column ‘num_calls’ and plot.

y

0 2.022689

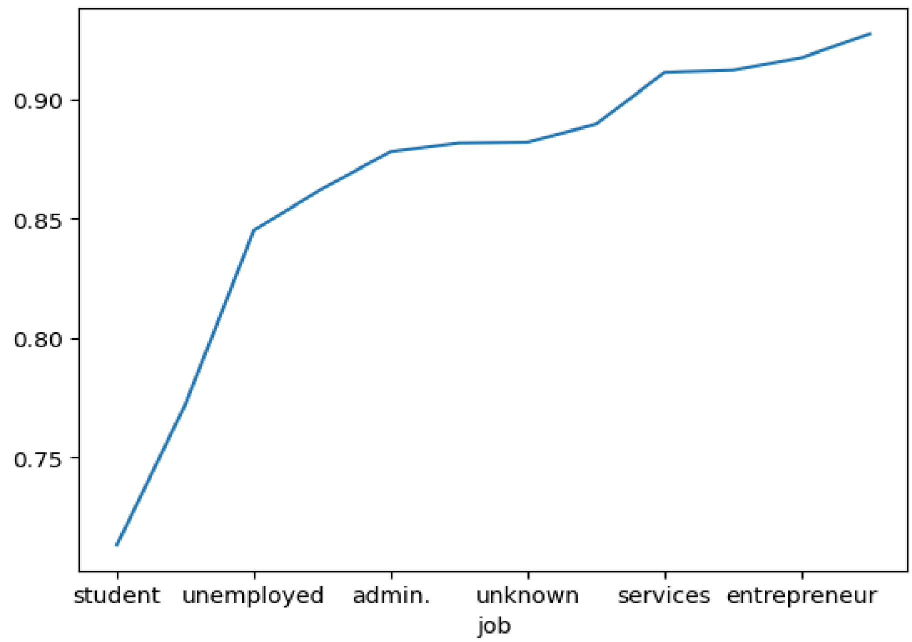
1 2.441202



From this can conclude that num_calls above 2 will not subscribe insurance and num_calls below 2 will subscribe insurance.

e. Group the column 'job' and mean of column 'y' and plot.

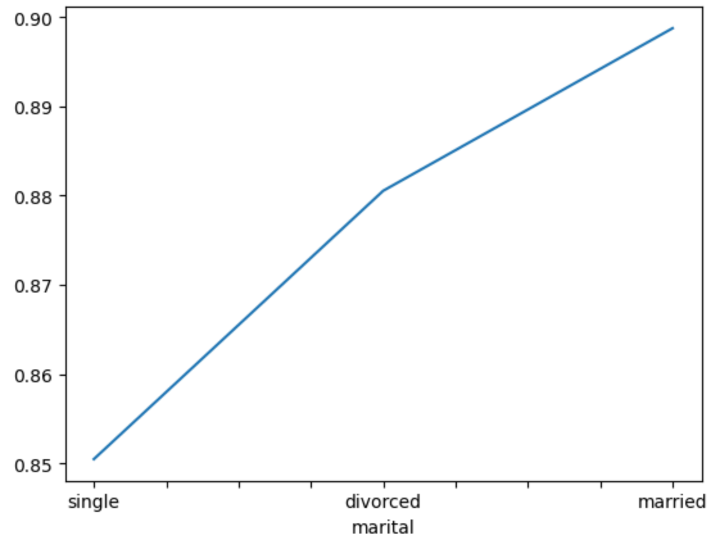
job	
admin.	**0.877950**
blue-collar	**0.927235**
entrepreneur	**0.917283**
housemaid	**0.912097**
management	**0.862430**
retired	**0.772085**
self-employed	**0.881571**
services	**0.911149**
student	**0.713220**
technician	**0.889415**
unemployed	**0.844973**
unknown	**0.881944**



From this can conclude that ‘student’,’retired’ people are having a high chance to subscribe to insurance policy and ‘blue-collar’, ’entrepreneur’ people will not subscribe insurance.

Mapped the column job based on mean values in this ‘blue- collar’ mean is 0.92 so mapped with 12 then ‘entrepreneur’ mapped with 11 likewise mapped according to mean values.

f. Group the column ‘marital’ and mean of column ‘y’ and plot.
marital
divorced 0.880545
married 0.898750
single 0.850485

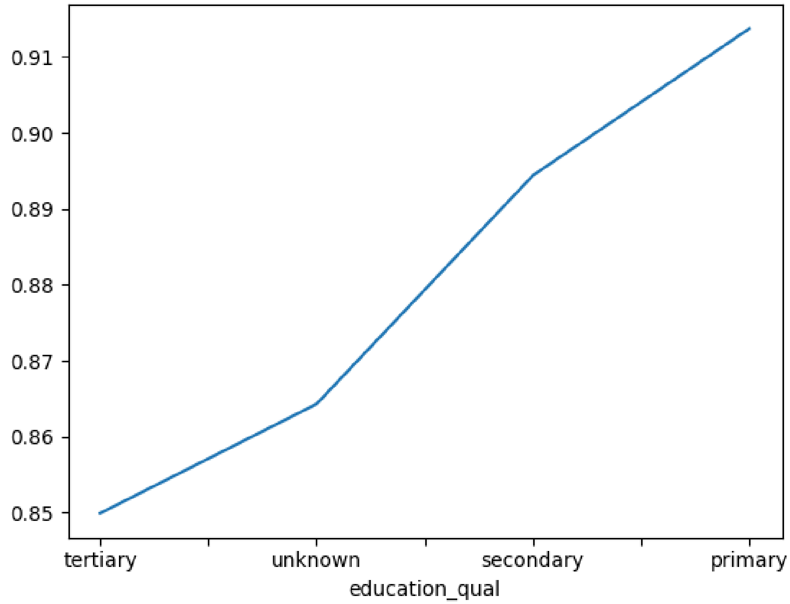


From this can conclude that ‘single’, ‘divorced’ people are having high chance to subscribe insurance and ‘married’ people will not subscribe insurance.

Mapped the column ‘marital’ according to mean values ‘single’ as 1, ‘divorced’ as 2 and ‘married’ as 3.

g. Group the column ‘education_qual’ and mean of column ‘y’ and plot.

education_qual
primary 0.913723
secondary 0.894392
tertiary 0.849914
unknown 0.864297

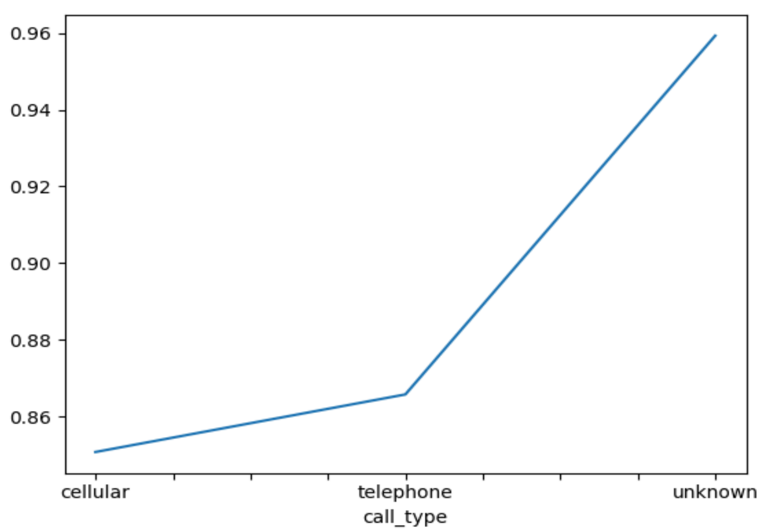


From this can conclude that 'tertiary' and 'unknown' qualified people have a high chance to subscribe insurance.And 'secondary' and 'primary' qualified people will not subscribe insurance.

Mapped the column 'education_qual' according to mean values.

h. Group the column 'call_type' and mean of column 'y' and plot.

call_type
cellular 0.850796
telephone 0.865795
unknown 0.959284

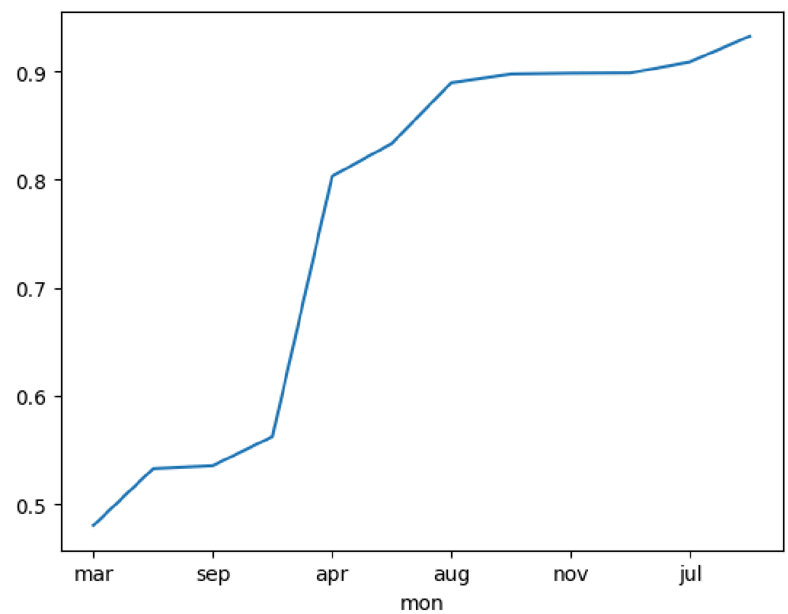


From this can conclude that call type through ‘cellular’ subscribe insurance and call type through ‘unknown’ will not subscribe insurance.

Map the column ‘call_type’ according to the mean values.

i. Group the column ‘mon’ and mean of column ‘y’ and plot.

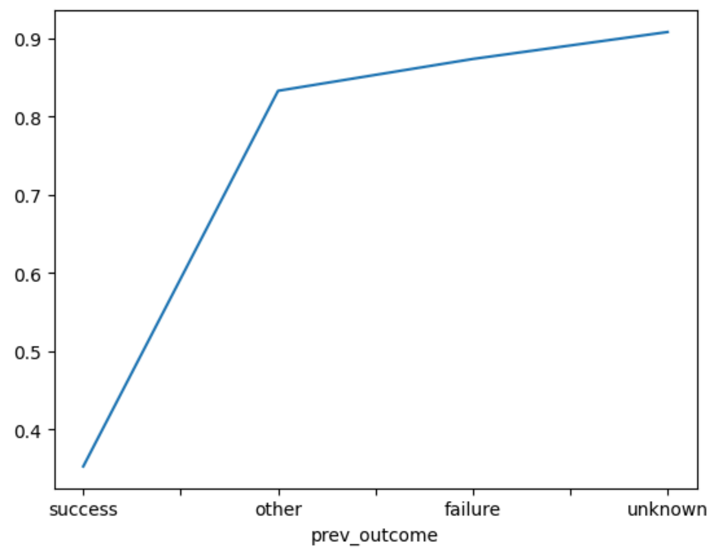
mon
mar 0.480084
dec 0.532710
sep 0.535406
oct 0.562331
apr 0.803206
feb 0.833522
aug 0.889832
jun 0.897734
nov 0.898489
jan 0.898788
jul 0.909051
may 0.932801



From this can conclude that calls during month ‘mar’ , ’dec’, ’oct’,’sep’ will subscribe insurance and calls during month ’nov’,’jan’, ’jul’, ‘may’ will not subscribe insurance.

Map the column ‘mon’ according to the highest mean map with higher value.

j. Group the column ‘prev_outcome’ and mean of column ‘y’ and plot.
prev_outcome
failure 0.873903
other 0.833152
success 0.352747
unknown 0.908370



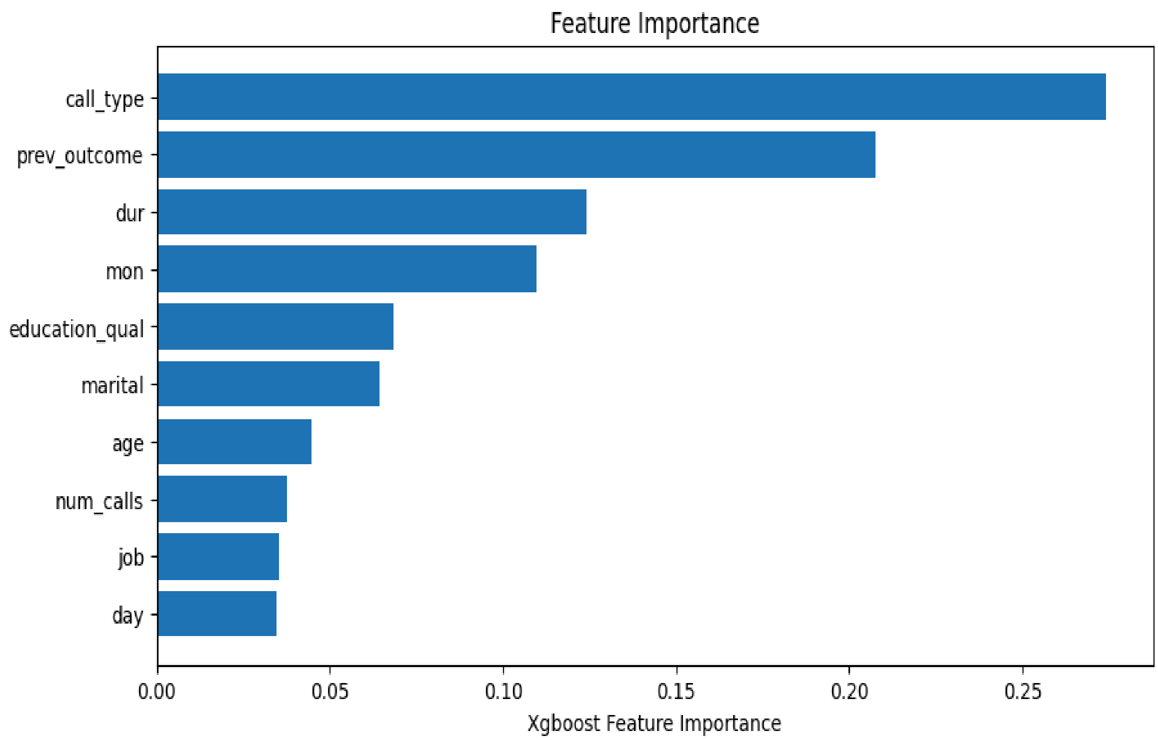
From this can conclude that last call ‘success’ subscribes insurance and last call ‘other’,’failure’,’unknown’ will not subscribe insurance.

Map the column ‘prev_outcome’ according to the mean values.

6. Save the mapped column using pickle.
7. Given dataset is an imbalanced dataset 88% ‘no’ and only 11% ‘yes’.
8. Split the dataset into train and test data.
9. Balance the dataset using SMOTE(Synthetic Minority Oversampling Technique). Can also use Cluster- centroid and Smoteenn. But in this dataset using smote got the best f1_score.
10. Model fit and evaluation.
    a. LogisticRegression F1 score: 0.9007163703900238
    b. DecisionTreeClassifier F1 score: 0.9191093861709734
       After cross validation at max_depth 18
       DecisionTreeClassifier F1 score: 0.9211865713845047
    c. RandomForestClassifier F1 score: 0.9329923273657289
       After cross validation at n_estimators 500 max_depth 25 and max_features ‘log2’
       RandomForestClassifier F1 score: 0.9339473011000256

d. XGBClassifier F1 score: 0.9378427787934186
After cross validation at learning_rate 0.6
XGBClassifier F1 score: 0.9383451059535822
e. So the best model is the XGBClassifier, save fine tuned XGBClassifier model using pickle.

11. Feature importance of fine tuned XGBClassifier



12. Then load the saved encoder and model in streamlit with the select box and text_input to predict whether the client will subscribe or will not subscribe to the insurance policy.