This is a bank data.The requirement is to develop a statistical model which predicts whether a customer will subscribe to the term deposit or not ?.The model should also be efficient in targeting and less bothering to uninterested customers as the bank has faced complains regarding irrelevant product calls.We also need to devise a framework to choose which customer to call and which one to leave alone

Since the data is binomial (yes/no).We will be making a classification.We will start with Logistic Regression .Since any model in sklearn needs numerical value ,we will use replace function for variables with 2 levels and getDummies for levels more than two.

The model has been deployed and parameters to check how good the model is explained in the code.Coming back to the original requirements where the model should be efficient in targeting and less bothering to uninterested customers

We can have an ideal situation where there are absolutely no misclassification in a model where the accuracy is 1.00 but the accuracy of this model is 0.90,so there will be misclassification which we could not avoid

Comparison of different ML Classification Algorithms

Logistic Regression -

KNN

SVC

Decision Tree

Data Exploration and Visualization

Statistical Properties of the numerical variables (Observations)

Mean Age is 40.The Age group of 30-42 has 21977 Data points i.e. 48.6 % of the total sample.The distribution plot of this variable shows that the data points are right skewed and do not possess normality.

Credit in Default Ratio is 1.8 %.Euro 1362(Portugal) is the mean bank balance.Euro (72 - 1428) is the IQR for balance

55.58% of the sample population have a housing loan while only 16.02% of the sample population have a personal loan

Avg. Last contact duration was 4 mins 18 secs.Avg. person was contacted at least two times in this campaign.Mean days passed since last contact is 40 and max is 871 days (2.3 years)

The Blue Collar and Management and technician groups are the prominent job categories of the total sample.60% of the sample population is married.More than half the population's educational level is secondary ,followed up by tertiary level at second place.

Now we need to see the behaviour of each variable in comparison with the target variable y.Starting with age which is a continuous variable needs to be grouped in age brackets.Cross tabbing these age groups with y variable,you get number of people in those groups who said yes or no.

Crosstab Age Group results are as follows.The first age group 18-20(Young Adults) has around 50 people out of which 18 people i.e. 38.3 % of the people subscribed to the term deposit.The next age group 20-25 has 762 people of which 24.8 % of the sample subscribed .The next age group 25-30 has 4464 people of which 16.2 % of the people subscribed .The next age group 30-35 has 9740 people of which 10.8 % of the sample subscribed.The next age group 35-42 has 10995 people of which 10 % of the people subscribed.The next age group 42-50 has 9009 people of which 9.2 % of the sample subscribed.The next age group 50-60 has 8410 people of which 9.3 % of the sample subscribed.The next age group 60-70 has 1230 people of which 30 % of the sample subscribed.The next age group 70-80 has 424 people of which 42.5 % of the sample(180 people) subscribed.The next age group 80-90 has 121 people of which 40% of the sample subscribed.The final age group 90-100 has 9 people of which 7 people (77.77 %) of the sample subscribed.

Crosstab Job Group results are as follows.The admin group has 5171 people in total out of which people who subscribed are 12.2 % of the sample. The blue-collar group has 9732 people in total out of which people who subscribed are 7.3 % of the sample.The entrepreneur group has 1487 people in total out of which people who subscribed are 8.3 % of the sample.The housemaid group has 1240 people in total out of which people who subscribed are 8.8 % of the sample.The management job group has 9458 people in total out of which people who subscribed are 13.8 % of the sample.The retired group has 2264 people in total out of which people who subscribed are 22.8 % of the sample.The self-employed group has 1579 people in total out of which people who subscribed are 11.9 % of the sample.The services group has 4154 people in total out of which people who subscribed are 8.9 % of the sample.The student group has 938 people in total out of which people who subscribed are 28.7 % of the sample.The technician group has 7597 people in

total out of which people who subscribed are 11.1 % of the sample.The unemployed group has 1303 people in total out of which people who subscribed are 15.5 % of the sample.The unknown group has 288 people in total out of which people who subscribed are 11.8 % of the sample.

Insights

Age plays a huge factor in figuring out the behaviour of your customers and the mean age sits right skewed of the maximum frequency area of the age variable of this dataset which is between 30 to 42 which accounts for 48.6% of the total sample.