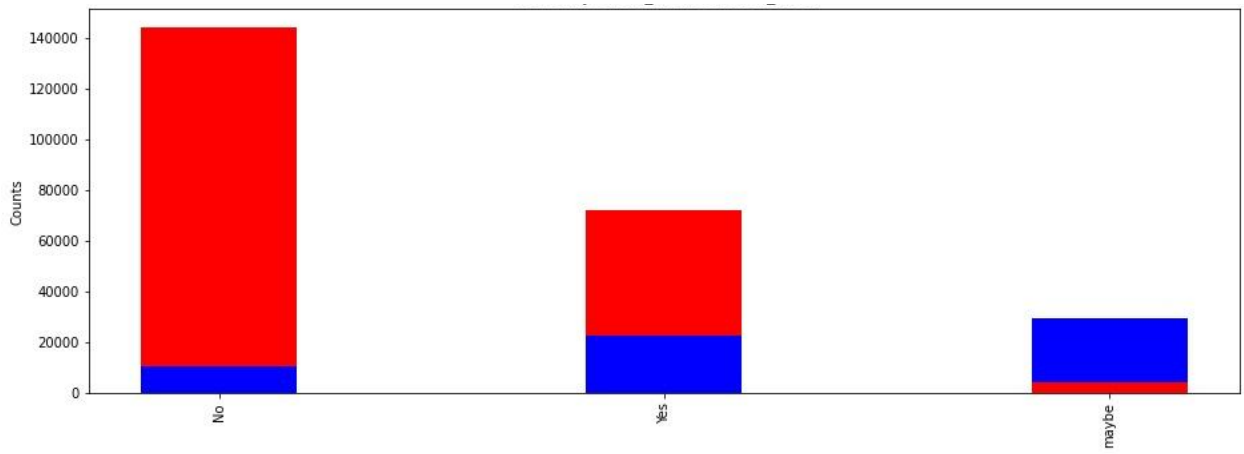
Approach

Data Modeling

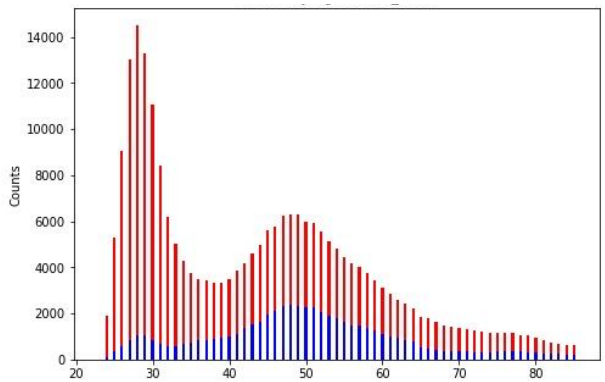
Columns ‘Gender’, 'Region_Code', 'Occupation', 'Channel_Code', 'Credit_Product' and 'Is_Active' are categorical. Each of them consists of only a few distinct values. So they are used without any modifications.

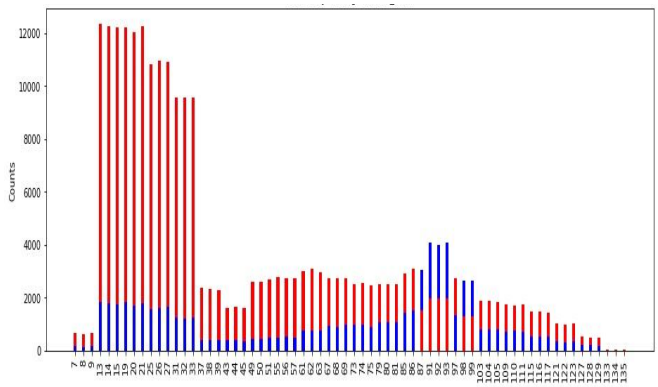
Column 'Credit_Product' contained missing values. From the frequency distribution plot of this column taking the target classes separately, we can see that the samples with missing values show an opposite trend compared to the rest.

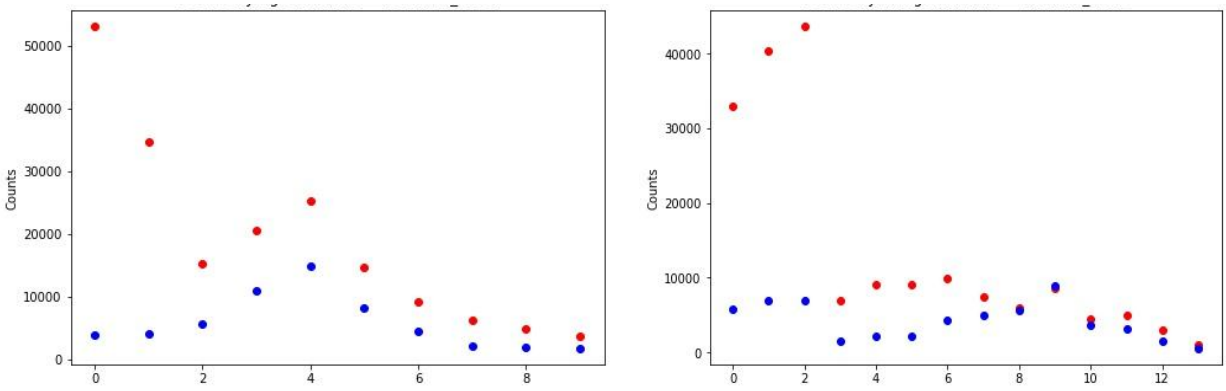


(* Red : ‘Is_Lead’ = 0, Blue : ‘Is_Lead’ = 1. ‘Maybe’ denotes the samples with missing values.)

‘Age’ and ‘Vintage’ have numeric data. To avoid any potential overfits they both are discretized into bins. The number of bins 10 and 14 for ‘Age’ and ‘Vintage’ respectively, were chosen as they preserve the general trend shown in their frequency plots.
(Below original frequency distribution of ‘Age’ (left) and ‘Vintage’ (right))







(Above frequency distribution of ‘Age’ and ‘Vintage’ after discretizing.)

Column 'Avg_Account_Balance' contains numeric values with a very large range. Discretizing into bins and log transforms were considered to reduce the range. As log transform performed better, it was used in the final model.

Accounting for Imbalance

The number of negative samples (‘Is_Lead’=0) is roughly 3 times that of the positive samples. This was accounted for by setting a weight of 0.25 for negative and 0.75 for positive samples while fitting the classifier.

Classifier

Extra Trees Classifier and GBM were considered. GBM performed better even with no hyper parameter tuning.
Some tuning was done in GBM to improve the score, but no significant improvements were made.