**JOB-A-THON**

Graphical user interface

Description automatically generated

**Approach:**

**EDA:**

Exploratory Data Analysis is always the step of any machine learning lifecycle. It involves different steps like Univariate Analysis, Bi-Variate analysis, checking for missing values, outliers, missing value imputation etc. Steps followed and some salient observations were as follows,

1. **Univariate Analysis:**
2. Only Credit\_Product variable had missing values.
3. It was found that the Response variable was not balanced. Around 76% of the respondents did not show higher intent towards a recommended credit card and only around 24% showed interest towards the credit card offered.
4. It was also found that Region\_Code was a high cardinality variable with 35 unique values.
5. It was found that most of the customers who were reached out to were Self Employed and a very few were Entrepreneurs.
6. X1 seemed to be the most popular Channel code and X4 was the least popular.
7. Majority of the customers did not have any existing credit product with the bank.
8. It was also found that majority of the customers were not active in the past 3 months.
9. It was also observed that the Avg\_Account\_Balance variable was positively skewed.
10. It was also observed that most the customers were below 35 and there was a sizeable population of customers between 40 and 55.
11. Most of the customers were with the bank for about 15 to 35 months. Very few customers had long relationship with the bank.
12. **Bi-Variate Analysis:**
13. During the Bi-Variate analysis, it was found that if the customer was an entrepreneur, he or she had a higher chance of showing higher intent towards the credit card. This is contrary to all other variables.
14. It was found that the response rate was higher if the customer was with the bank for less than 15 months and greater than 130 months.
15. It was also found that higher the Avg\_Account\_Balance, higher is the chance that the customer will subscribe to the offer.
16. It was observed that the Average Age of the customers who subscribed to the offer was higher than customers who did not subscribe to the offer.
17. It was also found that customers who subscribed to the offer had longer relationship with the bank.
18. **Correlation:**

It was found most of the numerical dependent variables were not correlated with the independent variable, the correlation was not strong enough in either direction. Only the Vintage variable had a better correlation compared to the other variables. It was also found that Vintage and Age were correlated but the correlation was not strong enough. Below is the diagram representing the correlation,

Chart, treemap chart

Description automatically generated

1. **Missing Value Treatment & Feature Generation:**
2. Missing values in the Credit\_Product were treated in 3 ways, first with the mode of the variable, second with “Missing” value, and the third method was to let the model handle the missing value. Out of all these methods, the third method proved to be useful.
3. Since Region\_Code had high cardinality, instead of Label Encoding the variable, it was replaced with its frequency counts. This method was more useful than Label Encoding the variable.
4. Following feature Generation techniques were tried, but were not found to be useful, Binning of Numerical columns like Age, Vintage etc., creating interaction features between categorical variables. Grouped features were generated using Avg\_Account\_Balance and categorical variables.

**Modelling:**

Modelling always comes second to feature engineering and exploratory data analysis, having said that, it plays an important role in the entire predictive modelling purpose.

1. LightGBM, CatBoost and XGBoost were tried individually but none of them performed better either in the local validation or the leaderboard.
2. For validation purpose, a 5-fold StratifiedKFold CV model was used. Stratified Cross Validation technique was used as the data is slightly skewed and with normal KFold CV, the data distribution in the folds is not the same.
3. Up Sampling, Down Sampling and Synthetic Samples were generated to offset the class imbalance, but none of these methods were helpful in the end.
4. Instead of Up sampling, down sampling, I had to use the in-built feature of the algorithms to balance the weight of positive and negative samples (**is\_unbalance=” True”** in LightGBM and **class\_weight=” balanced”** in XGBoost).
5. Hyperparameter tuning was extensively performed tuning the values of learning rate, max depth, subsample, column samples, etc.
6. Regularization was also performed so that the model does not overfit on the training data.
7. Instead of individual models, an ensemble of LightGBM and XGBoost performed better in the local validation as well as in the leaderboard. LightGBM was again used on the predictions of the individual models.

**Room for Improvement:**

There is always room for improvement in the field of machine learning.

1. Generating more features based on domain knowledge always helps the cause.