**AmExpert CodeLab**

Approach By: **Vishwanath Kulkarni**

**Approach:**

**EDA:**

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

1. **Univariate Analysis:**
2. There were missing values in multiple columns, but the percentage of missing values was less.
3. The dataset was imbalanced with only 8.1% positive values.
4. Dataset contains mix of numerical and categorical variables.
5. There are 45528 unique customers in training data and 11383 customers in testing dataset.
6. Age of customers range from 29 to 55 in both training and testing datasets.
7. Majority of customers are Female followed b Male customers, one of the customers genders is incorrectly mentioned as XNA.
8. 65% of the customers do not own a car. There are missing values in this variable.
9. Around 70% of the customers own a house.
10. Majority of customers do not have children, followed by 1 child. There are missing values here too.
11. Yearly income variable as an extremely high value which is an outlier, it has been dealt with appropriately. Log transformation of this variable looks normally distributed.
12. Number of days employed variable is interesting as a significant number of values are more than 35000 which will be 1000 years, this is not practically possible. These values are replaced with median value. Even after imputation, some values were greater than the Age of the customer, these values were again imputed appropriately. There were also some negative values in the Test dataset which ere imputed.
13. Occupation Type variable has a cardinality of 19. Interestingly, “Unknown” value of occupation type corresponds to records where the number of days employed is more than 35000.
14. Most of the customers have a family of 2, followed by 1. There are missing values in this variable.
15. Around 81% of the customers are not migrant workers.
16. Yearly Debt Payments variable seems positively skewed.
17. Credit Limit had one extreme value which was probably a data entry error. This was replaced with the median value.
18. Credit Limit Used ranged from 0 to 99% of the credit limit available.
19. Credit Score ranged from 500 to 949. Majority of the customer credit scores were above 700.
20. Majority of the customers didn’t default previously. And very few defaulted in last 6 months.
21. **Bi-Variate Analysis:**
22. It was observed that customers who had used between 70 to 100% of their credit limit had higher chance of defaulting than other customers.

A picture containing shape

Description automatically generated

1. It as also observed that customers with credit score of less than 700 had higher chance of defaulting tan other customers.

A picture containing logo

Description automatically generated

1. It was also observed tat if a customer defaulted in last 6 months, it as highly likely that he would default on the credit card payment.
2. It as also observed that if the customers had previous defaults, there was high chance that he would default on the credit card bill payment.
3. We could also see that the probability of defaulting as almost the same across all age groups.

Shape

Description automatically generated with medium confidence

1. **Correlation:**

From the correlation plot below, we can observe that credit limit used, previous defaults, and defaults in last 6 months were positively correlated to the dependent variable. Also, the credit score had negative correlation with the dependent variable (As the credit score decreases, the probability of defaulting on payment increases).

A picture containing timeline

Description automatically generated

We can also observe that some of the independent variables are correlated with each other (previous defaults wit defaults in last 6 months etc.).

1. **Missing Value, Outlier Treatment & Feature Generation:**
2. Since there were outliers in some of the variables, they were detected only after careful examination of the variable. These were imputed wit appropriate values.
3. Some of the values in variables such as number of children, total family members were combined because of the low count of the values.
4. Missing values in numerical variables were replaced with median and in the categorical variables, it was replaced with a ne value.
5. Continuous variables were also binned based on their values.
6. Label Encoding was used to encode the categorical variables. One hot encoding resulted in a lower leaderboard score.
7. Features were also generated based on grouping by categorical variable and takin the mean, median, std dev of continuous variables. Rule based feature engineering techniques were also employed.
8. Continuous variables were also log transformed to offset the effect of their skewness. It was also observed that Log transforming these features improved the LB score to a good extent.
9. Since the Dependent variable is highly imbalanced, up sampling and down sampling techniques (SMOTE) were also employed to balance the positive and negative classes, but these techniques didn’t improve the score.

**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 and XGBoost were tried individually but XGBoost performed poorly as compared to LightGBM.
2. Since the Dependent variable is highly imbalanced, Stratified K Fold sampling technique was used as it splits the data based on the dependent variable.
3. Hyperparameter tuning was extensively performed tuning the values of learning rate, max depth, subsample, column samples, etc.
4. Regularization was also performed so that the model does not overfit on the training data.

**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.
2. It was also observed that extensive tuning of Hyperparameters did not yield great results, feature engineering was the key in doing well in this competition.