Churn Prediction Model Report

Problem Statement:

The goal of this project is to develop a base predictive model that accurately identifies customers at risk of churning from our commercial platform. By analyzing historical customer data, including purchase history, location and other demographic correlations, we aim to create a base algorithm that can serve as a blueprint model to forecast the likelihood of a customer discontinuing their visit to the platform. As a result, it will enable us to execute retention strategies such as discounts and other offers focused on the right customers to retain them before churn."

Featured received directly from the customer:

Demography: Customer ID, DOB, City, Country, Gender

Channel Valid: Email Valid, SMS, WhatsApp, Push Notification

Transaction: Transaction Date, Purchase, Product Category

Synthetic data generation:

All the columns mentioned above are generated using data generation python libraries (mimesis & faker).

- Adding random none
- Creating anomalies in the numeric data columns ("Purchase")

Data Profiling:

- Finding for null value columns
- Knowing the datatypes and value count (frequency)
- Finding Anomalies in data

Data Preprocessing:

- Removing data anomalies
- Missing value imputation: Removed the rows where significant features such as Purchase, Customer ID are missing. Then missing values are imputed using Univariate and multivariate imputer based on the feature context, feature correlation and imputed values effect on prediction.

- Feature encoding: Features with multiple categorical data are one hot encoded (each category as a new feature) and simply binary encoding for 2 category features.
- **EDA:** Finding correlation among the features and perform univariate as well as multivariate analysis.

- **Feature Engineering:** Created new features (CLTV , RFM) which are the ultimate significant features to determine the customer churn based on defined conditions.

Model cross validation:

- Train test split before model training

As the target column (churn) will have categorical data, therefore we used classification prediction algorithms such as logistic regression and decision tree.

As decision tree was overfeeding hence logistic regression proved better algorithm to go with in this case.

Model was validated by the test data after training with the train data.

Model deployment:

Suggestions:

The target feature (churn) is a categorical feature which straight forward classifies the customer as churn (1) or not churn (0) which has drawbacks to decide the severity of churn when deciding to target only a part of "churning" customers to run campaign. Hence having the churn values on a scale of 0 to 100 which will indicate the probability of churn in percentage. Thus, it will provide broader scope to target the customer on a priority basis.