Mvend Ltd Proposal Description by Fonyuy Boris Lami

After importing the sm_subscriber_orders.csv and sm_transaction_history.csv. The first step was to identify the number of observations were respectively (178077, 10) and (135300, 21). Prior to loading sm_transaction_history.csv in a python notebook, the data was cleaned using a bash script titled clean_trans_hist.sh. This resulted in the retrieval of over 99.997% of the records correctly loaded to the notebook. For easy data manipulation, active_return_msg and pay_ref_no were dropped. It should be noted that these features are relevant as they can be used alongside machine translation followed sentiment analysis based on keyword extraction that can give an idea if a subscriber is likely to make a subscription renewal. This is a perspective as it was not done in the context of this work. Sentiment analysis will give an idea of the view of how a client feels about the service. A positive feedback being proportional to a high chance of subscription renewal.

Commenced by checking for missing values in both datasets and observed active_return_msg to have over 99.43% missing values (dropped due to variety of languages - English, French and Kinyarwanda, more in depth study using unstructured data analytics), slip_no with over 51.4% missing values and drawers_bank with over 40.46% missing values. These three features were the top three that had missing values. The strategy for categorical data was the use of mode while for numeric data, the mean was used to fill missing values. A grouping strategy was applied to sm_subscriber_orders.csv based on the ‘transaction_id’ accompanied by mode and sum aggregation. An inner join of the two datasets was then done using 'transaction_id', 'transaction_date', 'provider_id' as these features are common to both individual data sets. It was assumed that these three features are similar to foreign keys in the database and therefore unique in the database. The ‘handled_at’ was normalized by setting the seconds to zero so as to match ‘transaction_date’ in a bit to have an estimate of the duration of a service prior to performing the inner join.

The joined data had dimensions (134472, 26). A labelEncoder was used to convert categorical features. In this merged data, a client is uniquely identified by an account_number. In order to make “A forecast into what the chances are that a given client will make a payment through the same teller as the most recent payment.” This was reformulated to predicting the teller_id to which a client will likely do a payment. This problem is considered as a classification problem

with target feature teller_id. A variety of models were used to fit the data and accuracy_score metric evaluated with training data being 80% and test data being 20%. The model that gave the best performance was BaggingClassifier with a score of 0.824428 followed by DecisionTreeClassifier with a score of 0.805317. As a perspective, CatBoostClassifier, LGBMClassifier and XGBClassifier can be experimented and tuned.

Due to time constraints, doing hyper-parameter tuning was possible using GridSearchCV accompanied by cross validation in order to obtain a better model. The approach used for predicting the teller_id could be extended to predictions of when certain clients would most likely make their subscription renewal but using regression instead of classification. Due to limited information, it was difficult to tell the different packages that exist. Notwithstanding, the prediction of a package upgrade by a client can be done with a machine learning regression approach. It was unclear what is meant by ‘sales’, the reason why this was not elaborated. Once the definition of ‘sales per year’ is clarified, the data can be grouped per year in a bit to generate a visualization showing the trend in sales per year and hence a realistic forecast. It would have been great to explore the effects of data normalization such as min_max, absolute, standard and more normalizations on model accuracy. Also, data balancing using undersampling, oversampling and other data balancing techniques can be explored given that the data is unbalanced.