**APPROACH DOCUMENT**

**Problem Statement:** Can you predict whether a customer will churn or not?

1. **Understanding Problem Statement**

Decreasing the Customer Churn is a key goal for any business. Predicting Customer Churn (also known as Customer Attrition) represents an additional potential revenue source for any business. Customer Churn impacts the cost to the business. Higher Customer Churn leads to loss in revenue and the additional marketing costs involved with replacing those customers with new ones.

In this challenge, as a data scientist of a bank, you are asked to analyze the past data and predict whether the customer will churn or not in the next 6 months. This would help the bank to have the right engagement with customers at the right time.

1. **There are two main aspects of the business understanding part:**

* What is the business problem?
* What are its implications?

In the beginning, you should understand what the business problem is and what its implications are. These implications might be cost-based, revenue-based, time-based and so on.

**3. Solution Approach:**

Once you have the business understanding and an understanding of the available data, we can start building the solution approach. The approach should be in tune with the hypotheses formulated and should give an idea of your approach of the overall solution i.e. the data that you’ll be using, the EDA that you’ll perform, the ML algorithm that you’ll be using, the Evaluation metrics you’ll be tracking, mapping those evaluation metrics to KPIs and making business decisions.

**4. Exploratory Data Analysis**

Once the solution approach is finalised, it's time for the EDA steps to begin. Here’s a brief summary of EDA steps that can be used:

● Data staging and clean up​: This is the basic data cleanup and preparation stage. You collect the data from various sources, clean it and prepare the master dataset. In the data of this problem, no null values were present.

● Sanity checks​: The next step is doing a quick sanity check of the entire dataset to observe any unusual data points that should not exist.

● Class Imbalance: check if any columns have class imbalance. We found that target variable has class imbalance of 23%

● Univariate Analysis​: Finally, we begin with the univariate analysis part. This is where visualisation tools like histograms and boxplots come in handy as they help in analysing numerical features.

● Bivariate Analysis​: Then, you go ahead and evaluate the relationship between the target variable and the rest of the features. Here plots like scatter plots, pair plots, correlation matrices come in very handy to do the analysis. Some of the insights are:

1. Gender: Female customers will churn more than male customers.

2. Transaction\_status: Customers who have not made any transaction in past 3 months can churn.

3. Product\_holdings: Customers with less holdings can churn more as compared to customers with more product holdings with the bank.

● Feature Engineering​: Finally, if you want, you can do feature engineering to extract useful features from the given dataset.

**4. Model Building**

Here are the steps that you need to perform during the model building phase.

● First, build a simple enough model with good interpretability so that you can demonstrate the results to your client. This will act as a proof of concept that the suggested solution approach is feasible. We started with building a logistic regression model.

● If possible, check the statistical significance as well. For example, we checked the statistical of variables by p-values in logistic regression model.

● After that, increase the complexity of the model and try to optimise the parameters involved to get the best results. Then we tried to make more complex problems like Decision Trees and Random Forest.

● Avoid overfitting the results or else your model can't be generalised for unseen data. For dealing with this we did hyperparameter tuning in decision tree and random forest model.

● We used SMOTE sampling technique to deal with class imbalance.

**5. Model Evaluation**

Now that you've prepared the model, it's time to evaluate the results and present it to your client. Here's a summary of the model evaluation steps that need to be performed here:

● Evaluate performance on unseen data

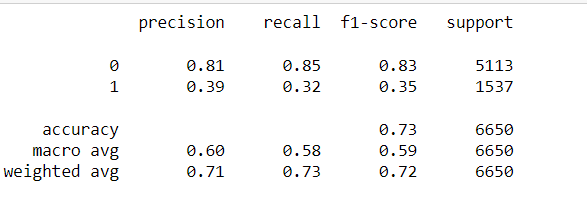
● Identify the right evaluation metrics - Depending on the problem statement at hand, you should be tracking the correct evaluation metrics. In this problem statement, it is important to predict the churn (1) records than no churn (0) records. Therefore sensitivity/recall is the important measure.

● Evaluation metrics should align with business outcomes

**6. Final Insights according to Business Impact**

* Accuracy cannot be counted on when dealing with unbalanced datasets since it cannot detect all the churn records.
* Recall says out of all the churning customers, what percentage were correctly identified.
* Precision says out of all churning customers predicted to be churn; how many have actually churned.

*Decision tree with smote sampling gives better precision and recall score along with better F1-score and good AUC scores.  So, we have used Decision Tree (with SMOTE sampling) model to predict whether the customer will churn or not in the next six months. Below is the performance metrics of the model:*

**