# **Proposal**:

The bank transaction categorization model aims to classify transactions description into predefined categories based on descriptions feature data. The model leverages machine learning Multiclass classification techniques, particularly text classification algorithms, to automate the categorization process.

## Data Understanding and Preprocessing:

- Explore and understand the provided datasets (bank_transaction.csv and user_profile.csv).
- Preprocess the descriptions feature (e.g., lowercase, remove punctuation) and merge user profiles with transaction data based on client_id.

## Exploratory Data Analysis (EDA):

- Analyze the distribution of transaction categories and Description
- And there is uncategorized data and nan data in category feature. So I separated the labeled and unlabeled data into two dataset

## Feature Engineering:

- In this I used natural language processing (NLP) involves selecting the most relevant and informative features (words or phrases) from the description data to improve model performance and reduce overfitting
- And I used CountVectorizer technique commonly used in natural language processing (NLP) tasks for converting a collection of text documents into a numerical representation
- And finally I used Frequency based Word Embedding technique Tfidftransformer which is used to convert text into numeric form

## Balance the Data

- I used RandomOverSampling from Imblearn to balance the dataset

## Model Selection and Development:

- Train test split is used to split the dataset in training and testing data

- And created a pipeline with countvectorizer , tfidftransformer and a classification model

- Experiment with various classification algorithms suitable for text data (e.g., SVM, Random Forest, naive baye, xgboost, etc ). Random Forest Classifier gave good result of 92% accuracy in prediction.

- Train the model using labeled transaction data.

- And predict the class data of unlabeled dataset

## Model  Evaluation

- I evalvated the model's performance using accuracy score, confusion matrix, classification report

- And finally I used cross validation

## Save the Model

- I use joblib to save the model pkl format

## Short-term Plan:

- Develop a base-line model using a simple classification algorithm.

- Implement basic feature engineering techniques and assess model performance.

- Establish a workflow for data preprocessing and model training.

## Medium-term Plan:

- Explore advanced NLP techniques for feature extraction from transaction descriptions.

- Incorporate user feedback and domain expertise to refine model predictions.

- Implement cross-validation to evaluate model generalization and stability.

## Long-term Plan:

- Develop a scalable and robust system for continuous model training and deployment.

- Monitor model performance over time and retrain periodically to adapt to evolving transaction patterns.

- Explore model interpretability techniques to provide insights into model predictions.

## Viability:

- **Accuracy:** The model achieves satisfactory accuracy in categorizing transactions type, as demonstrated by evaluation metrics. In random Forest Classifier an accuracy of 92%

- **Interpretability:** Model predictions are interpretable, allowing users to understand the rational behind each categorization.

- **Scalability:** The model can handle a large volume of transactions and users efficiently, facilitating scalability.

- **User Experience:** The model provides valuable financial insights to users, aiding in financial management and decision-making using the description provied.

- **Data Privacy:** The model ensures data privacy by protecting users' sensitive information during the categorization process.

## Considerations and Constraints:

- **Model Maintenance:** Regular monitoring and retraining are necessary to maintain model performance over time.

- **Data Quality:** The model's performance heavily relies on the quality and consistency of transaction data.

- **Regulatory Compliance:** Ensure compliance with data privacy regulations and ethical considerations in handling user data.

- **Resource Allocation:** Allocate sufficient resources for model development, deployment, and maintenance to ensure its long-term viability.