Project Proposal: Predicting Future Events and Offer Acceptance Probability

I - Project Overview

Our proposal is to develop a machine learning engineering system that predicts future events and proability of the events using AWS Machine Learning Services (SageMaker) following standards. The system will make forecasting of the a particular offer given the current data of the user enabling customer success for the organization and follow up engagement strategies.

II – Project Statement

The project challenge is to predict the next event and its probability of being accepted, but also we will output all the probabilities of the possible events to gather more information about the customer to take further actions even if our main goal is to see if an offer will be accepted.

We will start analyzing customer offers history and the final model will tell us how is probable to the user to accepts the offer. Our focus will be minimizing the cost of False Negatives, because missing out a successful offer completion is deemed more expensive than a False Positive.

III - Objectives

- **First objective**: Prediction of the future events and probabilities of these events with focuse on offer acceptance.
- **Second objective**: We know at first we will not get the best model due to limitation of the dataset but we will focus later to improve precision and recall metrics.

IV. Data and Approach

We will use fake datasets provided to do the job.

- **Profile.json**: This dataset has customer demographic data (e.g., age, gender, income).
- **Transcript.json**: This is the event data (e.g., offer received, viewed, completed).
- **Offer Data**: This is the information on the offers, such as type, duration, reward, and difficulty.

Our approach includes:

- **Shift the target variable ‘event’**: We will group customers and shift the history of the events (future event will become the target variable) to predict the probability of the next occurrence of an event (that could be the offer completion and other types of events).

- **Exploratory Data Analysis (EDA)**: We will investigate the trends of the customer age, gender, income and other variables to understand the behavior of the customer under this patterns, i.e. first analyzing a cliente and later the groups.

V - Solution Architecture

A - Data Preprocessing:

- We will merge the datasets to get one big data that follows our proposal.
- Handle basic missing data (not to much because our model will provide a transformation pipeline)
- Extract an transform from the data using OHE (One-Hot-Encoding) some categorical values.

B - Model Selection:

- We will start with a basic **XGBoostClassifier** for traing. After evaluating varios metrics, we will shift to get the precision and recall metrics to be maximized, focusing at first on precision and later on recal.
- We know that this is the proposal but we got the model results sooner
    1. **Base model metrics**: Precision of 27%, recall of 89%.
    2. **Final model metrics**: Precision of 28%, recall of 71%
- We will test the pipeline with any of the models and finally put in production the best model. If the base model is better the put the base model, if not then the final model.

C - Hyperparameter Tuning:

- We will start to conduct an hyperparameter tuning job using **Amazon SageMaker** at first with 12 different models (same XGBoost but different hyperparameters).
- Our constraints will be limited to these first models to let us know if is a good idea to improve further and spending more money.
- There will be a door open to improvements but need approval.

D - Deployment:

- Because the base model and the final model are similar in resource input, we will deploy the trained model using **AWS SageMaker**, and make predictions available using **AWS Lambda** and **API Gateway**.
- We will test the predictios through REST API endpoints using curl and also the API gateway interface.

VI - Results and Insights (After the proposal, what we got)

- **Model Performance**:
    - The base XGBoost achieved good recall results, the recall metric (89% in the base model, 71% after optimization was not good enough because we need to go further looking over the hyperparameters).
    - Precision remained extremely low, recalling we need to improve the quality of the data of the non lowest occurrence results of the future event acceptance.
- **Trendings of the dataset**:
    - **Demographics**:
        - Customer between 45-65 are more likely to complete offers, with higher incomes showing similar tendencies.
    - **Types of Offer**:
        - Discount and BOGO offers were more popular, with higher difficulty offers still seeing high completion rates when rewards were significant.
- **Recommendations**:
    - We tested the pipeline with our hyperparametrized model, but the base model was better, at first, put in production the base model, keep improving by hyperparameter and finally update the endpoint.

VII - Challenges and Recommendations

- **Data Constraints**: The training dataset for our approach is not sufficient to get a good model for edge cases (e.g., extremely high or low values of percentiles).
- **Future Improvements**:
    - **Generative Adversarial Networks (GANs)**:
        - Introduce GANs to generate synthetic data for underrepresented groups.
        - Make statistical data (bottom 5% and top 95% of the data) to improve the model's ability to handle outliers.
    - **Additional Hyperparameter Tuning**:
        - Further exploration of different models and tuning parameters using a larger budget would enhance the model's overall performance.

VIII - Future Work

- Implement a GAN Implement advanced techniques like GANs for synthetic data generation.
- Test more XGBoost Models or change the model like using a Random Fores with wider variety of hyperparameters to enhance predictive performance.

IX. Conclusion

The project demonstrates how to do a complete pipeline and approach for the next event of offer acceptance using SageMaker.

The model shows a strong performance in recall but needs improvement on the precision metric. We will need further enhacements of this part.

We leverage AWS services such as SageMaker, Lambda, and API Gateway, to put the model in service and being able to integrate into real-time systems, offering businesses a scalable, data-driven approach to optimizing their offers.

We need to refine the dataset with synthetic (time-series) approach.

The solution could become an essential tool for companies looking to maximize offer success and customer engagement.