# Project Report

This project focused on developing predictive models for customer purchase behavior by addressing two main tasks. **Task 1** involved predicting the probability of purchasing any product in the next month, while **Task 2** aimed to estimate the probability of purchasing each individual product. These tasks allowed us to approach the problem from both a general and a granular perspective.

Data Exploration and Analysis

For the EDA analysis, the **MDB_mini** dataset — comprising 10% of the full dataset — was used due to resource constraints. In the analysis, I examined the seasonal and temporal dependencies in transactions and dialogues, explored the distribution of categorical features, and identified outliers. PCA and KMeans clustering were applied to dialogue embeddings, which later assisted in feature engineering. At this stage and onward, only transaction and dialogue data were considered, while geo data were not used due to time limitations for this task.
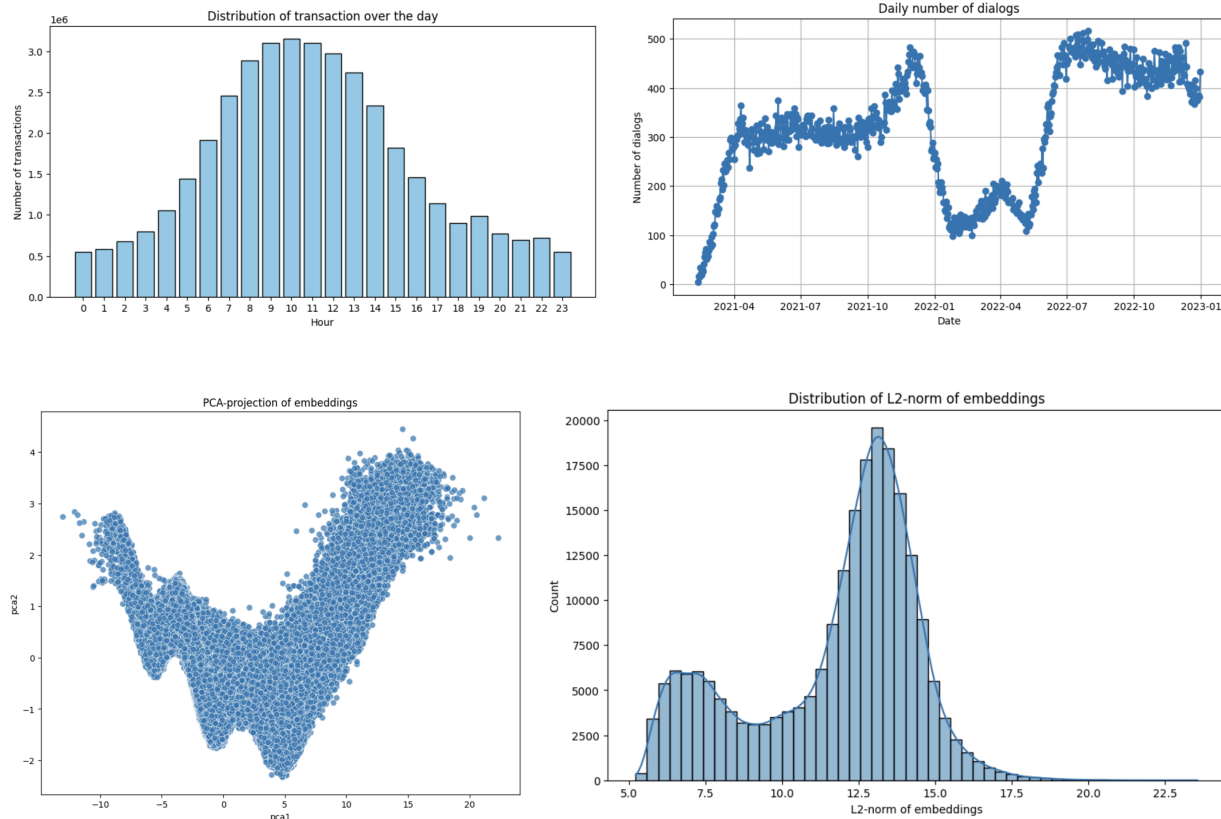
Feature Engineering

A significant part of the project was devoted to feature engineering. In total, **2266 features** were computed, encompassing transactional metrics, behavioral patterns, and aggregated statistics.

Modeling Approach

The data show a severe **class imbalance**: for the single-target task, only about 1% of samples are positive, and for the four-target task, each positive class makes up less than 0.5% of the data. I addressed this by applying both upsampling and downsampling techniques.

For the predictive modeling, **LightBoost was utilized with L1 and L2 regularization and 5-fold cross-validation**. The model was trained and evaluated on the two tasks, and optimal thresholds were determined to maximize the F1 score and overall accuracy.

Results

A model for Task 1 achieved an AUC of 0.7756 and an F1 of 0.6230, using an optimal threshold of 0.46. These results indicate promising performance with room for further tuning and improvement.

For the four-target task (Task 2), Target 4 showed the strongest performance with an AUC of 0.8367 and an F1 of 0.6796, while the remaining targets yielded moderate results. Optimal thresholds varied across targets, suggesting additional tuning and balancing could further enhance performance.