Data collection outline for ML techniques to predict B2B and B2C churn

Data Collection Tool & Process

I'll be working with Python alongside PostgreSQL to pull data directly from Cartrack's database. This will cover all the key churn indicators, variables like subscriptions, usage patterns, device health, repair history, and support tickets. I'll write SQL queries to extract what's needed, then process and organise it in Python into monthly snapshots so it's clean and ready for modelling. Before moving ahead with the modelling phase, I'll check in with the Finance team to make sure the churn labels are correct and to clear up any anomalies

Planned Analysis

After the data is cleaned and verified, I'll use it to train churn prediction models. My starting point will be gradient-boosted trees (LightGBM/XGBoost) since they're proven performers for this type of data. I'll evaluate the models with metrics like PR-AUC, precision@K, and probability calibration, and I'll use forward-chaining validation to avoid data leakage. SHAP values will be used to understand which factors are having the biggest influence on the predictions.

Answering the Research Question

This setup answers my research question because it builds an accurate dataset and then uses it to predict churn for both B2B and B2C customers. Pulling from our DB and confirming with Finance means my models will be trained on correct churn data, so the predictions should actually be reliable.