Capstone Project Phase 5 Proposal

Net Guard: Predicting Customer Churn Risk in Internet Service Subscribers Using Machine Learning *By Group 9*

Business Understanding

As a group of six team members, we discussed several ideas and finally agreed on a problem that is common but highly important, **customer churn**. Customer churn is a critical challenge for internet service providers, directly impacting revenue and customer lifetime value. Acquiring new customers is more costly than retaining existing ones. Our objective is to develop a supervised machine learning model capable of predicting customer churn risk using behavioral, transactional, and feedback data.

Our Capstone project aims to support data-driven retention strategies by identifying customers likely to leave the service and allowing pre-emptive interventions. The target audience includes Internet Service Providers marketing teams, customer retention units, and data science practitioners within the telecom sector.

Data Understanding

We are using a structured dataset containing 36,992 customer records with 23 features, including customer demographics such as age, gender, and region. Subscription attributes like membership category and internet option. Behavioral indicators include average time spent and days since last login. Customer interactions such as past complaints, feedback, and churn risk score. The dataset is in CSV format and contains both structured and unstructured components, enabling both classical modeling and natural language processing.

Below is the data set link

https://huggingface.co/datasets/d0r1h/customer_churn

Data Preparation

Our initial preprocessing step involves handling missing values in essential features such as region category, points_in_wallet, and preferred-offer-types. Next, we will encode categorical variables, including membership category, feedback,

and internet option, using appropriate encoding techniques such as label encoding or one-hot encoding.

We will also apply feature engineering techniques to derive customer tenure from the joining date to quantify the length of customer engagement and perform sentiment analysis on textual feedback to extract useful insights that may correlate with churn behavior. Finally, we will standardize numerical features where needed to ensure that the scales of different variables do not negatively impact the learning algorithms that are important for models that are sensitive to feature magnitudes, such as logistic regression or neural networks.

Modeling

We frame this as a supervised classification task to predict the churn_risk_score. Depending on the score distribution, we may treat this as a binary or ordinal classification problem. Our modeling roadmap includes a baseline model using Logistic Regression, followed by more advanced tree-based models such as Random Forest, XGBoost, and LightGBM. For stretch goals, we will explore neural networks to capture complex, non-linear relationships in the data. Model interpretability will be examined using SHAP values or permutation importance techniques.

Evaluation

Model performance will be evaluated using standard classification metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. Emphasis will be placed on recall due to the business goal of minimizing churn by correctly identifying high-risk customers.

Our minimum viable product (MVP) will consist of a fully cleaned dataset, comprehensive exploratory data analysis, and at least one baseline model such as Logistic Regression or Random Forest achieving a minimum accuracy of 70%. Additionally, we will present insights on feature importance to support model interpretation and stakeholder understanding.

As stretch goals, we plan to enhance model performance through hyperparameter tuning using GridSearchCV. We will also incorporate sentiment analysis on the customer feedback field to improve predictive power and explore neural network architectures to compare performance with traditional machine learning models.

Deployment

Final results will be reported through a written report and visualizations. If time permits the model will be deployed via a Streamlit. Key features will include a file upload interface for CSVs, customer-level churn prediction output, model explain ability using top predictive features, and recommended retention strategies based on prediction level.

Tools and Methodologies

- **Programming Language:** Python
- **Libraries for Data Handling & Visualization:** pandas, NumPy, matplotlib, seaborn
- **Modeling & Evaluation:** scikit-learn, XGboost, lightgbm, tensor flow (optional)
- **Deployment:** Streamlit (If time allows)

Development Platform:

- Local Jupyter Notebooks with backup on Google Drive
- Local data storage during development phase