Phase 1 – Project Proposal

Customer Churn Prediction System

Team 2

Liam Knapps - Gleb Ignatov - Gautam Singh - Minh Le Nguyen

Table of Contents

I. Project Idea

- Customer Churn Prediction

The purpose is to predict whether someone is going to cancel a subscription to a certain service or not, identify common features that affect the churning rate between different datasets to determine what affects the decision to cancel the subscription, and provide businesses with preventive measures to decrease the churning rate and keep customers subscribed. We also plan to train and compare models using several different datasets that have churning rate information for 2 different services, such as shopping and music services. Two models will be trained individually using assigned single service-related datasets, then we will train an additional model that is using combined datasets for training.

II. Dataset Used

a. Sources (Small Dataset): https://www.kaggle.com/datasets/muhammadshahidazeem/customer-churn-dataset

b. Size: 3.28 MB
c. Rows x Columns: 12 columns, 64,374 records
d. Description:

   This dataset encompasses various features related to customer shopping preferences, gathering essential information for businesses seeking to enhance their understanding of their customer base. The features include customer age, gender, purchase amount, preferred payment methods, frequency of purchases, and feedback ratings. Additionally, data on the type of items purchased, shopping frequency, preferred shopping seasons, and interactions with promotional offers is included. With a collection of 3900 records, this dataset serves as a foundation for businesses looking to apply data-driven insights for better decision-making and customer-centric strategies.

e. Sources (Large Dataset): https://www.kaggle.com/datasets/joy11117/customer-churn-dataset
f. Size: 12.5 GB

g. Rows x Columns: 18 columns, 26,259,199 records
h. Description

The Customer log dataset is a 12.5 GB JSON file and it contains 18 columns and 26,259,199 records. There are 12 string columns and 6 numeric columns, which may also contain null or NaN values. The columns include userId, artist, auth, firstName, gender, itemInSession, lastName, length, level, location, method, page, registration, sessionId, song,status, ts and userAgent. As evident from the column names, the dataset contains various user-related information, such as user identifiers, demographic details (firstName, lastName, gender), interaction details (artist, song, length, itemInSession, sessionId, registration, lastinteraction) and technical details (userAgent, method, page, location, status, level, auth).

III. Relevance and Impact on the Real-World Problem

Customer churn is a major challenge for subscription-based businesses, directly impacting revenue, profitability, and long-term sustainability. The ability to predict which customers are likely to cancel their subscriptions allows businesses to take proactive measures, improving customer retention and satisfaction. By leveraging predictive analytics, we aim to develop a machine learning model that accurately forecasts churn, identifies key contributing factors, and suggests effective intervention strategies. This will enable businesses to implement targeted retention campaigns, optimize customer support, and enhance the overall user experience. Our solution has broad applicability across industries, including streaming services, SaaS platforms, telecom providers, and financial institutions. Comparing performance evaluation metrics for models that are trained on datasets for different services might provide insights into how the relevance of the same features used in all datasets may differ depending on the service, such as a shopping or food delivery services.

IV. Project Scope

The companies that have customer subscription services allow them to register and gain profit from it. This project focuses on businesses operating under a subscription model, where customers pay a recurring fee for services. The scope includes analyzing customer data from various industries such as e-commerce to build a robust predictive model with key features

- Collecting and preprocessing customer data, including demographics, subscription details, engagement levels, and customer support interactions.
- Identifying the most relevant features influencing customer churn.

- Developing and optimizing a machine learning model for churn prediction.
- Providing actionable insights to businesses on intervention strategies for retaining at-risk customers.

V. Project Objectives

Customer retention is critical, we will use customer data to predict churn and identify contributing factors for a customer's leaving and identify preventative actions.

- **Predict customer churn**: Develop a reliable predictive model that classifies customers as either likely to churn or retain.
- **Identify key factors influencing churn**: Use feature importance analysis to determine the reasons behind customer attrition, such as pricing issues, low engagement, or customer dissatisfaction.
- **Develop personalized retention strategies**: Provide businesses with actionable insights, such as offering discounts, improving customer support, or adjusting product features to reduce churn.
- **Optimize intervention timing**: Predict when a customer is most at risk of churning, allowing businesses to intervene at the right moment with retention strategies.
- **Enhance customer lifetime value (CLV)**: Improve overall retention rates, thereby increasing long-term customer revenue and loyalty.

VI. Project Approach

Use a classification model such as random forests to predict churn. If a customer is predicted to unsubscribe the system will then identify the cause aspect such as the product does not meet their needs, high subscription fees, or any others aspect that we consider inside our dataset to give the business methods for preventative unsubscribed actions like reduced fees, 1 month free (threshold: 1 times only/3 months for renewable prevention policy).

VII. Project Timeline (Weekly)

Time Range	Task
Week 9	- Investigate Dataset and prepare Phase 1 Document - Exploratory Data Analysis (EDA) to find out the data distribution, missing values, duplicated
Week 10	- Clean and preprocessing data to handle the missing values, normalize/standardize numerical features, and encode categorical features

	- Add more features from the raw dataset - Split data set to training validation and test or plan cross validation - Calculate the correlation score between features and label
Week 11	- Training the baseline model: Logistic Regression, Decision Tree, Random Forest Tree, KNN - Train deep learning model (NN) with basic hyperparameters - Evaluate performance using key metrics (accuracy, precision, recall, F1-score, AUC-ROC) - Document insights and performance benchmarks
Week 12	- Experiment with hyperparameter tuning using Grid Search - Implement ensemble approach combining the different models and NN - Fine-tune model architecture - Calculate the error analysis again
Week 13	- Finalize model performance evaluation and compare the results with baseline - Prepare the model deployment strategy (save to file…) - Prepare the documentation for Phase 2 Document (key findings, challenges, enhancement plan)
Week 14	- Submit Phase 2 Document

VIII. Preliminary Plan for Deployment of the Model:

Model deployment will be done using AWS Lambda and can receive features via an HTTP request and respond with the predicted result.

IX. Tools Used to Manage the Project Tasks:

- Discord
- GitHub
- Email