Project Goals: Renewal Prediction & Timing (Group 52)

1. Background

This work aims to improve renewal forecasting and operational planning by predicting for Elara Health (fictitious company):

- **Whether an account/member renews** (binary outcome)
- **When the renewal occurs** in months from first unlock (timing outcome)

Accurate renewal is crucial for forecasting (billings/revenue), targeting enablement touchpoints, and aligning staffing with demand. The analysis will also identify which member demographics, customer characteristics, and touchpoints are most predictive of renewals and especially those associated with earlier renewals.

2. Objectives

Primary

- Predict renewal (Yes/No)
- Predict timing of renewal (months to second unlock)

Secondary

- Quantify which features drive renewal and early renewal
- Explore PCA/clusters associated with renewal propensity and earlier timing

3. Research Questions

Key Questions Project Seeks to Answer

- Which member demographics, customer characteristics, and plan design features most strongly predict renewal versus non-renewal?
- Among accounts that do renew, which factors are most associated with earlier renewals (shorter time between first and second unlock)?

- How do specific touchpoints (i.e., app sessions, chats, credit card requests, provider matching) influence both the likelihood and timing of renewal?
- Are there identifiable clusters of customers (via PCA or segmentation analysis) with distinct renewal behaviors that can inform targeted engagement strategies?
- What is the expected accuracy and reliability of renewal and timing predictions across different customer / member configurations that can be incorporated into a predictive mobile application?

4. Study Design & Scope

Cohort & Observation Window

- Index date: First Unlock Date (first_unlock_cp_created_date)
- Window: First unlocks from 12/1/22 through 11/30/24
- Follow-up: Up to 18 months after first unlock

Outcomes & Definitions

- **Renewal (binary):** 1 if second_unlock_date is present AND second_unlock_journey ≠ missing; else 0
- **Timing (months):** Integer difference between first and second unlock dates (only for renewals)

Inclusion Criteria

- First unlock between 2022-12-01 and 2024-11-30
- Age > 18 and non-missing sex
- Non–health plan customers

Exclusion Criteria

- Termed cases disabled within 18 months of unlock
- Health plan customers
- Age ≤ 18 or missing sex

5. Data Sources, Variables, & Technical Tools

Data Sources

- Extracted from Snowflake SQL (12/1/22–11/30/24) using first unlock create date

- Includes renewal data, demographics, customer characteristics, plan design, coverage features, and touchpoints

Technical Tools

- **Python packages for cleaning:** pandas, numpy, pyjanitor, dateutil, scikit-learn preprocessing utilities
- **Python packages for modeling:**
    - scikit-learn (logistic regression, decision trees, random forests, gradient boosting)
    - XGBoost, LightGBM (boosted trees)
    - statsmodels (Poisson, quasi-Poisson, negative binomial)
    - lifelines (survival models)
    - SHAP (interpretability/feature importance)

Variables

- **Geography & Segmentation:** Country, market segment
- **Contract/Pricing:** Case-rate contract, program type
- **Plan design & cost sharing:** Deductible status, co-pay/coinsurance, annual/lifetime max
- **Coverage features:** Medically necessary preservation, elective ART, adoption, gender-affirming care, menopause, low-T, pregnancy & postpartum, gestational carrier, doula, childbirth classes, milk shipping
- **Communication:** Registration emails allowed, email on file
- **Demographics:** Sex, employee age
- **Touchpoints:** App sessions, chats (benefits, plan, expert, IVF), credit card requests, expenses created, group sessions, logins, support messages, phone support, referrals, operational emails

6. Data Preparation & Feature Engineering

Here are the steps for the exploratory data analysis that will prepare the data for machine learning modeling.

- **Cleaning:** Standardize variable names, parse dates, combine unlock fields, remove ineligible records

- **Outcomes:** Create binary renewal indicator; calculate timing outcome

- **One-Hot Encodings:** Convert booleans to 0/1; one-hot encode key categoricals

- **Outliers/Transforms:** Winsorize at 1st/99th percentile; add outlier flags; log-transform skewed values

- **Missing Values:** Replace with zero when "none"; otherwise impute or keep as NA

7. Methods

For this analysis, the objectives are to compare a suite of machine-learning models to measure two related problems: (1) predicting whether a renewal occurs and (2) predicting when it occurs (in months from first unlock). For robustness, machine learning models will start with interpretable baseline models (regression and decision trees) and then evaluate more complex models (ensemble tree models and neural networks), using the same feature variables. Model outputs will be interpreted (e.g., via coefficients or feature importance using SHAP plot) to highlight the characteristics and touchpoints most associated with renewals and earlier timing.

Binary outcome (Renewal vs. not):

For predicting renewals (yes or no), this analysis will use logistic regression, decision trees, random forests, gradient-boosted trees, XGBoost, and a shallow neural network.

Timing outcome (Months to renewal):

For the month count between first and second unlock, predictive modeling will use Poisson and quasi-Poisson regression models, test for over-dispersion, and use Negative Binomial regression (if warranted due to high propensity of zeros). In parallel, this analysis will try tree/boosting regressors (gradient boosting/XGBoost/random forest) to capture non-linearities and interactions. Due to members being able to renew at different times within the observation window (early renewal months 9-12 and after original engagement has expired), the analysis may also do a sensitivity analysis using survival methods: discrete-time hazard models (complementary log-log link) and/or Cox proportional hazards to provide time-to-event insights.

Evaluation Metrics

The following is how this analysis will evaluate model performance for both the classification (binary) and timing (month to renewal):

Binary (renewal vs. not):

- **Accuracy** is the share of records correctly classified at the chosen probability threshold (e.g., 0.5).

- **Receiver Operator Curve (ROC) / Area Under Curve (AUC)** summarizes ranking quality across all thresholds; it's the probability the model scores a random renewal higher than a random non-renewal (1.0 = perfect, 0.5 = no better than chance). The closer the AUC is to 1 the better.

Timing (months to renewal):

- **Root Mean Squared Error (RMSE)** is the square root of the mean squared error. This metric will be in months and will measure the size of a prediction error (lower is better, large misses are penalized more).
- **Adjusted R^2** estimates the proportion of variance in months explained by the model, adjusted for model complexity (closer to 1 is better).
- **Prediction thresholds (% within ±1 and ±2 months)** reports how often predictions land within one or two months of the actual renewal month ad is an easy-to-read accuracy style metric for timing.

8. Deliverables

The following deliverables will enable leadership to more accurately forecast revenue, optimize resource allocation, and design member engagement strategies that improve retention.

- ***Final Report***

  The final report will include the background, objectives, data sources and collection approach, inclusion and exclusion criteria, data preparation, the modeling methods for both renewals and timing, the validation approach, results, key drivers of renewal, touchpoint insights tied to earlier renewals, limitations, recommendations, and a clear implementation plan.

- ***Executive Summary (PDF)***

  A two-page summary in plain language that highlights the problem, the approach, headline results, expected business impact, and the next steps.

- ***Summary Presentation***

  A concise slide deck for leadership that covers the "why," "what," and "how," shows results and what they mean for the business, and lays out the action plan and timeline.

- ***Mobile Application***

A simple phone-friendly tool where a user enters a few fields and receives a renewal likelihood and an estimated number of months to renewal, along with the top factors that drove the score and suggested next actions.

- ***Interactive Dashboard***

  A web dashboard to explore results and monitor performance over time. It will include filters, side-by-side model comparisons, key metrics, the top factors that influence predictions, segment and plan type breakouts.

- ***Final Presentation***

  The recorded presentation will walk through the project objectives, methods, and key research questions, highlighting each team member's specific contributions to data preparation, modeling, and visualization. It will also present results, showcase a preview of the mobile app and interactive dashboard, and conclude with final insights and recommendations for business impact.

- ***Final Recommendations***

  - The project will deliver concrete recommendations that link predictive insights to business actions, including:
  - How to target enablement and customer success resources to accounts with high renewal likelihood or earlier renewal timing.
  - Which customer characteristics, demographics, and plan features drive renewal outcomes and should inform product design and pricing strategies.
  - Where to invest in touchpoints (e.g., chats, application views, provider matching) that are most predictive of renewal.
  - Forecasting and staffing guidance tied to predicted renewal volumes and timing, enabling proactive resource alignment.

9. Data Privacy

To protect customer and member privacy, all direct identifiers will be removed from the analytic dataset. Customer identifiers (including customer names) will be stripped, and no member-level identifying information (i.e. name, date of birth, or any other unique identifier) will be included. To ensure there can be no linkage back to customer or employee, additional safeguards will be done to replace the native account_id and employee_id with randomly generated tokens so the data cannot be linked back to customers or individual members. Only de-identified fields needed for analysis (e.g., age in years, not date of birth) will be retained. Access to the dataset will be limited to approved classmates, and results will be shared only in aggregate to prevent re-identification.