Predicting Rare Medical Conditions from Retail and Pharmacy Data

CVS Health Pre-Internship Project

May 2025

Oreoluwa Alade
Data Science Intern, Summer 2025
North Dakota State University
oreoluwa.alade@gmail.com
(701) 491–3828

This project was completed independently in under four days to explore predictive insights from behavioral data in preparation for my internship with the CVS Media Exchange Team (CMX).

Overview

This project simulates a real-world pipeline to predict rare medical conditions using only retail and pharmacy purchase data. The goal is to flag people who might need follow-up or support, without using private health records.

I generated synthetic data with over 5 million purchase records across 100,000 customers. I then cleaned the data, explored patterns, engineered features, and trained machine learning models to detect people with rare conditions based on behavioral signals like how often they shop, what they buy, and when.

I used models like Random Forest and XGBoost, and evaluated them using recall, precision, and AUC score. I also applied SHAP to explain the model's decisions and built a simple Streamlit dashboard to simulate how it could be used in practice.

This work was done independently in under four days, and reflects the kind of fast, focused exploration I plan to bring to my internship with the CVS Media Exchange (CMX) team.

Key Insights and Model Performance

Behavioral Patterns from EDA

- About 25% of customers had no insurance information, simulating real-world out-of-pocket purchases or data gaps.
- People with rare conditions were more likely to buy a wider variety of product categories (e.g., pain relief, immune boosters).
- Behavioral features such as total purchases and number of unique shopping days were strong signals.
- Gender, age, and state had minimal influence on condition prediction, which supports a behavior-first targeting strategy.

Modeling and Evaluation

- I trained two models: Random Forest and XGBoost.
- XGBoost performed best, achieving:
  - ROC AUC: **0.853**
  - Recall (label = 1): **0.75**
  - Precision: **0.30**
  - Accuracy: **0.80**
- SHAP analysis showed that the most important features were:
  - `Product category`

  - `Purchase date`
  - `Purchase frequency`
- I also built a Streamlit dashboard to input customer profiles and get predictions instantly.

Business Relevance

This project shows that it is possible to detect rare medical conditions using only retail and pharmacy behavior, without needing personal health records.

Why This Matters for CVS Health

- **Privacy-first**: The model uses everyday purchase data and avoids sensitive medical details.
- **Early targeting**: Customers with more diverse and frequent purchases were more likely to have rare conditions. This can help CVS flag cases earlier.
- **Scalable**: The approach can work across millions of retail records, just like in a real CMX pipeline.

Fairness and Transparency

- Demographic factors had little influence on the model's output.
- SHAP helped me explain why a prediction was made, both for the group and for each person.
- These tools can support fair, data-driven decisions without bias or guesswork.

Potential Use Cases

- Pharmacy or clinical follow-up
- Adherence or refill support
- Health screening campaigns
- Personalized product targeting or education

This project shows how simple behavior data can be turned into something useful, respectful, and actionable for CVS Health and its customers.