Reducing Customer Churn in Retail with Personalized Incentives

Vaibhav Rouduri
vr2470

Abhishek Agrawal
aa9360

Shubham Naik
svn9724

Big Data
Section C
Fall Semester, 2024

Project Abstract:

Customer churn is very important for every retail business, as existing customers are more valuable than new ones. Our project analyzes the spending patterns of many households. We aim to use a system to detect customer churn by analyzing transactional data, household data, and product data. We will update the churn score everyday for each household, even if they haven't purchased anything that day, and identify if a household's churn score has crossed a certain threshold. If this happens, we will send those households to a pipeline that will analyze their purchasing trends, and subsequently provide tailor made discounts on certain products, to retain customers and boost loyalty.

Project Statement:

High customer churn is an issue for any retail business, as it directly affects revenue. Using customer behavior to identify churn risk is extremely important to get ahead of and prevent churn as much as possible. This project seeks to identify churn risk in real time on a daily basis, and provide personalized discounts on certain products to prevent churn. Predicting churn and taking preventive measures is a challenging process, requiring analysis of varying customer habits of diverse types of customers, which is what we aim to achieve in this project.

Objectives:

1. Data Integration: Identify and integrate data from multiple sources, including transaction history, individual household information, and individual product data.

2. Feature Engineering: Create features that capture behavior we want, like days since last purchase, purchase frequency, and total amount spent in a given timeframe
3. Labeling and Modelling: Analyze the data and define churn based on days since last purchase, and label the data as at risk of churn or not as it crosses a certain threshold (for example, 30 days since last purchase). Now, train a model that finds the relationship between the features and the churn label.

4. Real Time Detection: Implement a system that updates the churn risk of each household every day, and flags the household if it is at risk of churn.

5. Product Analysis Pipeline: Once a household is flagged, send it to a pipeline that analyzes the flagged household's transaction histories to identify favorite products, brands, stores and other factors.

6. Personalized Incentives: Generate personalized incentives, like discounts on favorite products, or discounts at their favorite store, to incentivize at risk of churn customers to return to the store and remain loyal customers.

7. Scalability: Deploy the system on a cloud platform that can deal with large volumes of data with daily updates and is capable of real time processing, like Azure.

8. Compliance and Regulation: Make sure that the system follows all necessary guidelines and ethical considerations, and prevent customer churn without misusing data related to customers.

Datasource:

- **Name:** Dunnhumby - The Complete Journey
- **Links to datasets:**
    - https://www.kaggle.com/datasets/frtgnn/dunnhumby-the-complete-journey?select=transaction_data.csv
    - https://www.kaggle.com/datasets/frtgnn/dunnhumby-the-complete-journey?select=hh_demographic.csv
    - https://www.kaggle.com/datasets/frtgnn/dunnhumby-the-complete-journey?select=product.csv

Data Size: 200 mb

Approximate Number of Records: ~2.6 million

Proposed technologies and Programming Languages:

S3, Lambda Function, Kafka, Airflow, Python, Pyspark, EMR Serverless, Scikit-learn, and more.