Big Data (CSGY-6513-C) Fall 2024

Project Report

Reducing Customer Churn in Retail with Personalized Incentives

Professor: Amit Patel

Authors –

Abhishek Agrawal - aa9360
Shubham Naik - svn9724
Vaibhav Rouduri - vr2470

NYU

1. Code Execution Instructions

- The code execution instructions are in the README of our GitHub repository, in the 'Technical Overview' section under 'Pipeline Execution Steps'.
- Click here: Technical Overview.

2. Technological Challenges

- Converting raw transaction data into meaningful features was a significant challenge. This involved aggregating transaction data at the household level, extracting relevant metrics such as Total Monetary spend, Frequency of purchases, Average Basket Size, Discount Utilization, and Discount Count. These features were crucial for training the model to predict churn accurately
- Labeling households to predict churn was a critical step. This involved analyzing historical transaction data and identifying patterns that could predict future churn. Each household was assigned a binary churn indicator: '1' for households likely to churn and '0' for those likely to stay.
- Deciding the best model was a challenge. We experimented with various machine learning models to predict churn, including Logistic Regression, MLPClassifier, and Random Forest Classifier. After extensive testing and validation, we found that the Random Forest Classifier outperformed the other models with an accuracy of 96%.
- Identity Access Management was a challenge since we had to create and configure IAM roles for cross-resource access management, i.e., SageMaker required necessary permissions to access the data in S3 and also to share the same AWS ecosystem with the team.

3. Changes in Technology

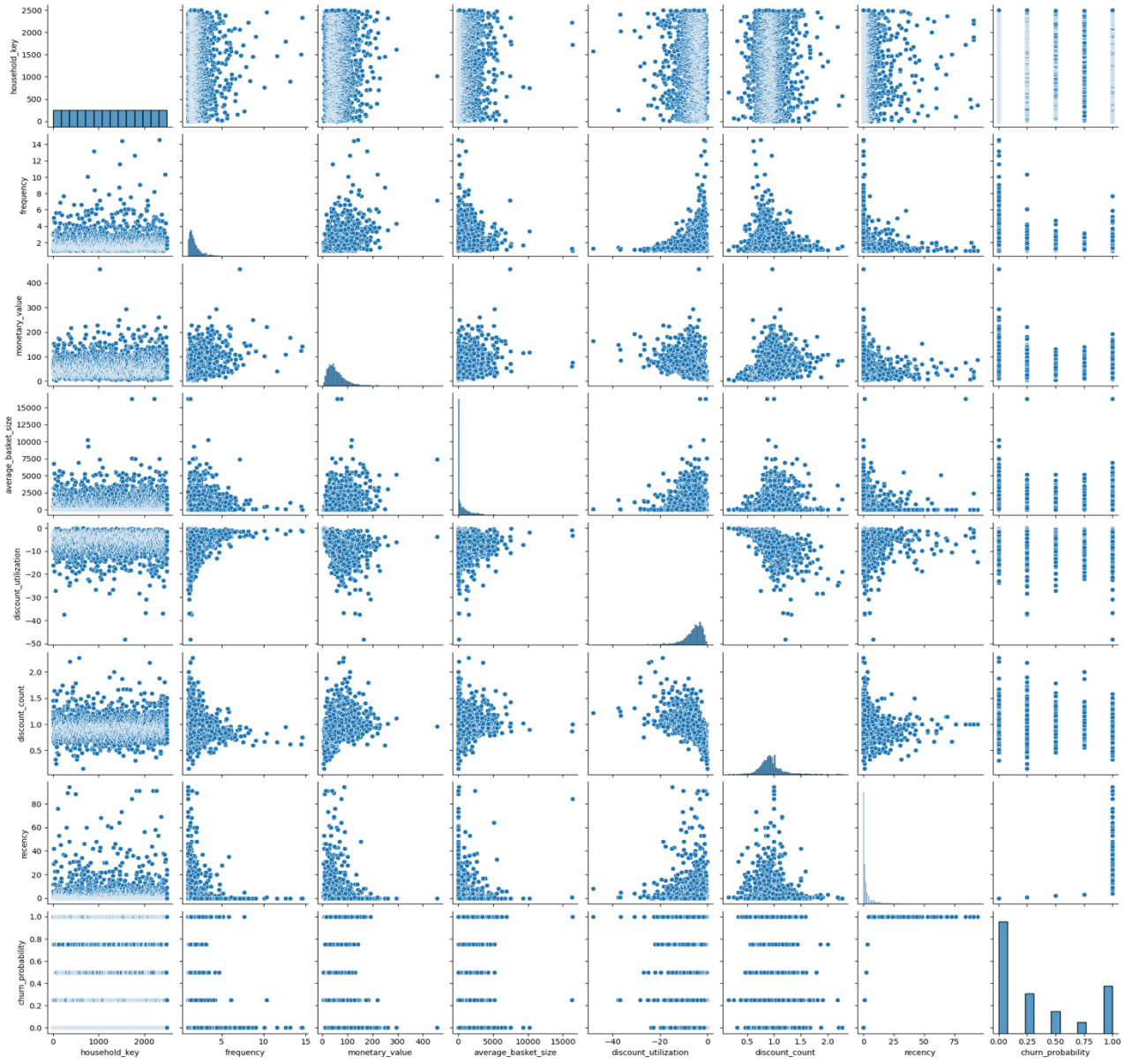
- Initially, we planned to use Airflow, Lambda Function, and EMR Serverless for orchestration, cluster management, and ML model management. However, we decided to replace these with AWS SageMaker. This change was driven by cost-effectiveness and the convenience of having all necessary resources within a single platform. AWS SageMaker provided a more integrated and streamlined environment for our machine learning workflows
- We initially considered using PySpark for data processing. However, since our entire codebase was written in Python, we found it more efficient to use Pandas for data-

frame management. Pandas is well-supported within AWS SageMaker, making it a more suitable choice for our project.

– At the beginning of the project, we were uncertain whether to base our implementation on historical or incremental data. This led us to consider using Kafka for real-time data streaming. However, as the project evolved, we decided to focus on historical data analysis. Consequently, we no longer needed Kafka or any live data streaming solutions.

4. Uncovered Aspects of Presentation

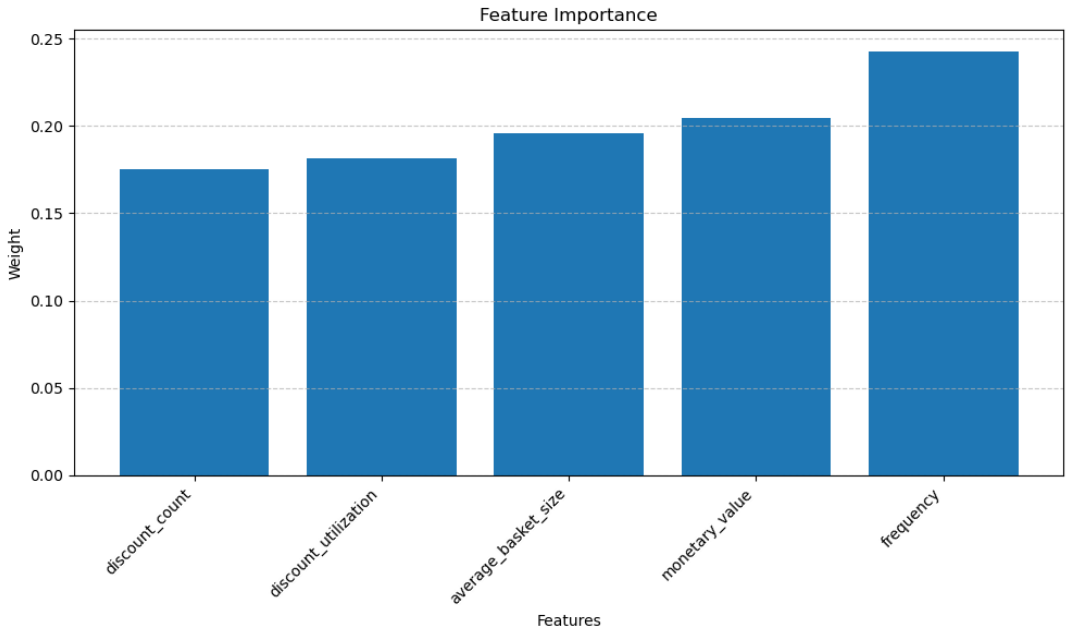
– Feature Importance: The feature importance weights were determined as follows:



- Some features showed a linear relationship, indicating that as one feature increases, the other tends to increase as well. This was particularly useful for identifying potential predictors of churn.
- Other features displayed more complex, non-linear relationships, suggesting the need for advanced modeling techniques to capture these interactions.
- The distribution of data points in certain plots highlighted outliers and anomalies, which we addressed through data cleaning and preprocessing.

5. Lessons Learnt

- The main challenge was to get the data into the right format for training the models. The Exploratory Data Analysis (shown in the GitHub Repository) using scatter plot matrix helped us get a better understanding of the data and the relationships between different features.
    - Frequency: 0.24
    - Monetary Value: 0.20
    - Average Basket Size: 0.20
    - Discount Count: 0.18
    - Discount Utilization: 0.18



- Out of the 2500 households in our dataset, only 800 have registered demographic information. This lack of demographic data for the majority of households posed a challenge in creating personalized offers and understanding customer behavior.

Enhancing the dataset with more comprehensive demographic details could significantly improve the accuracy of our churn predictions and the effectiveness of our personalized offers.

- Over a 52-week period, our model predicted that more than 25% of the total households would churn. This was a prediction to identify and manage churn by establishing stronger bonds with customers and retaining them, with strategies like personalized incentives and tailored offers. This is a very important aspect of any business, since it is more expensive to acquire new customers than to retain existing ones. The average annual churn rate should not be more than 5-7% for any business.
- The current offers and coupons are linked to a limited range of products, which restricts their appeal to a broader customer base. Expanding the range of products associated with these offers could increase customer engagement and satisfaction. By diversifying the product offerings, we can cater to a wider array of customer preferences and needs, thereby enhancing the overall effectiveness of our promotional strategies.

6. Future Improvements

- Enhance the dataset with more demographic information for households. This could include details such as age, income, household size, and geographic location. Having a richer set of demographic data would allow for more precise segmentation and targeting of customers, leading to more effective personalized offers and marketing strategies.
- Expand the range of products associated with offers to increase customer engagement. By including a wider variety of products in the promotional offers, we can cater to diverse customer preferences and needs. This approach can help in attracting different segments of customers and encouraging them to take advantage of the offers, thereby boosting overall engagement and sales.
- Distribute offers through various channels to optimize customer engagement. Utilizing multiple distribution channels such as email, SMS, mobile apps, and social media can ensure that offers reach customers through their preferred communication mediums. This multi-channel approach can enhance the visibility and accessibility of the offers, leading to higher redemption rates and improved customer satisfaction.
- Giving incentives based on previous purchases alone is not sufficient to retain customers. It is crucial to understand and identify the customer's potential preferences for new or unused products or services. Often, due to budget constraints or other limitations, customers may not purchase these items. However, a well-crafted

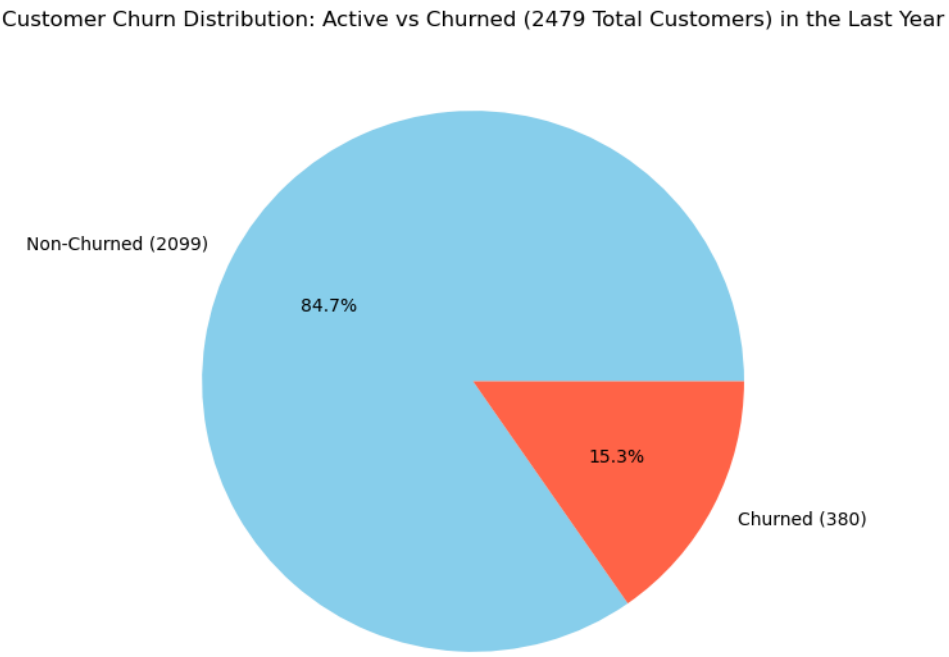
personalized offer could influence their behavior positively. By analyzing customer data and predicting their interests, we can create targeted offers that encourage customers to try new products or services. This approach not only boosts sales but also enhances customer satisfaction by providing them with relevant and appealing options.

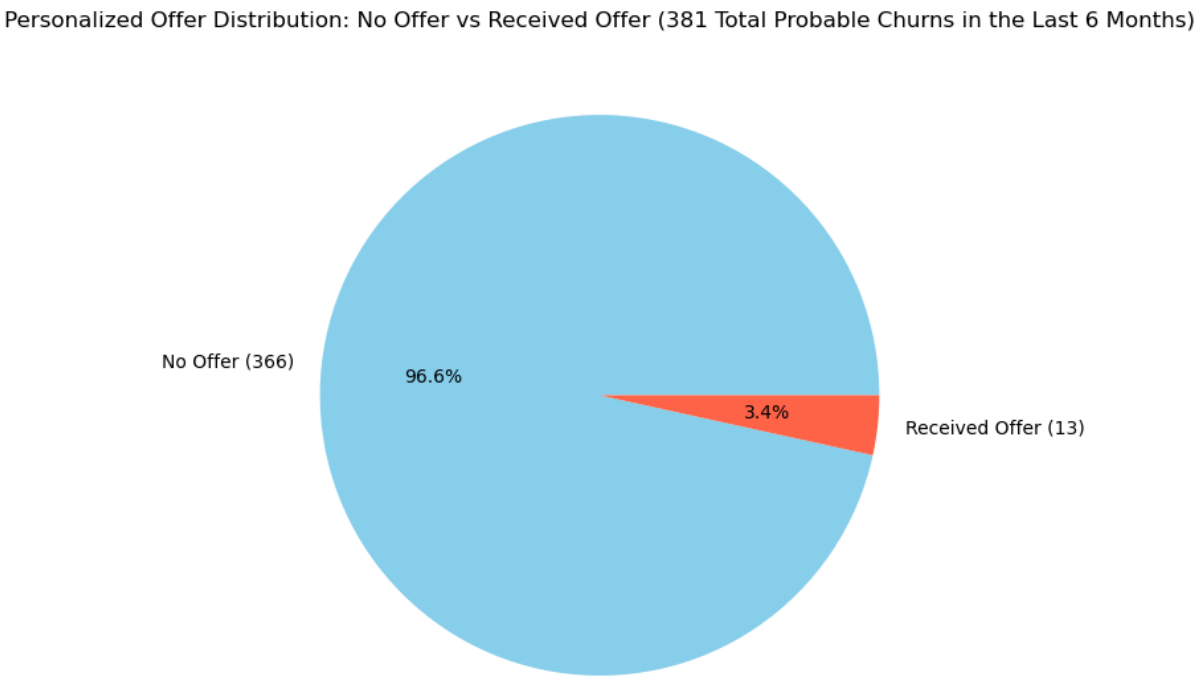
- Customer feedback and engagement metrics could be used to refine and improve personalized offers over time. Collecting and analyzing feedback from customers about their experiences with the offers can provide valuable insights into what works and what doesn't. Additionally, tracking engagement metrics such as offer redemption rates, click-through rates, and customer responses can help in continuously optimizing the personalization strategies to better meet customer expectations and drive higher engagement.

7. Data Sources and Results

- Sources:
  - transaction_data.csv
  - products.csv
  - hh_demographic.csv
  - coupon.csv

- Results:





- Based on the analysis, approximately 3.5% of the households identified as probable or already churned have received offers. The following recommendations are proposed:
- Enhance the dataset with additional demographic details for households.
- Expand the range of products associated with offers to increase customer engagement.
- Distribute offers through various channels to optimize customer engagement.