Repeat Buyers Prediction

CECS 550 : Pattern Recognition

Spring 2023

Group :- 7

Anthony Martinez
Diksha Patil
Pratik Jadhav
Pavan More
Sudarshan Powar

Instructor: Prof Mahshid Fardadi
Teaching Assistant: Rahul Deo Vishwakarma

CALIFORNIA STATE UNIVERSITY
LONG BEACH

Agenda

- Background and Goal
- Datasets Interpretation
- Feature Engineering
- Dataset statistics and feature ranking
- Prediction model
- Model evaluation
- Results
- Conclusion

Background and Goals

1. A shop runs big promotions on “Double 11” - the biggest online shopping event, in order to attract a large number of new buyers.
2. Unfortunately, many of the attracted buyers are one-time deal hunters, and these promotions may barely a have long-lasting impact on sales.
3. To reduce the promotion cost and enhance the return on investment (ROI), they want to identify who can be converted into repeated buyers.

Goal

1. Predict the probability of the given user becoming a repeat buyer of the given merchant in the future
2. To find the most important factor to predict repeat buyers

Data Interpretation

The data set contains anonymized user’s shopping logs in the past 6 months before and on the “Double 11” day.

The dataset has -

User profile :-

age_range , Gender , User_id

User Behavior Logs :-

user_id, item_id, cat_id, merchant_id, brand_id, time_stamp, action_type

Data Visualization

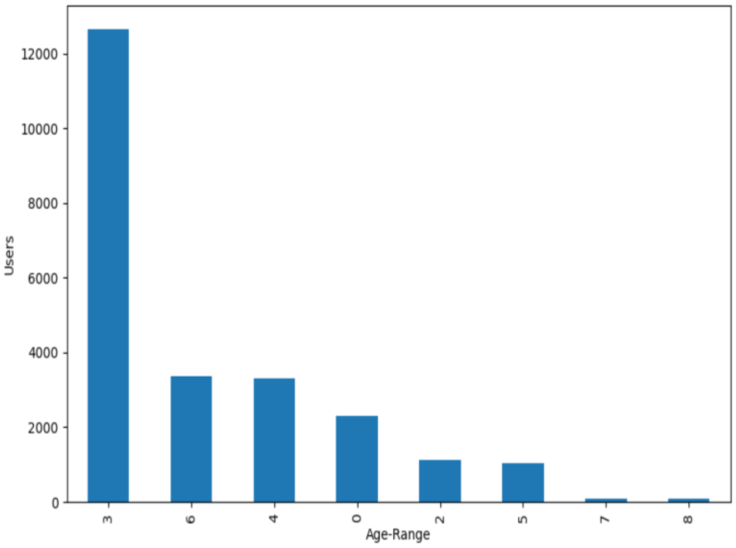


Figure 1: Users from different age-range

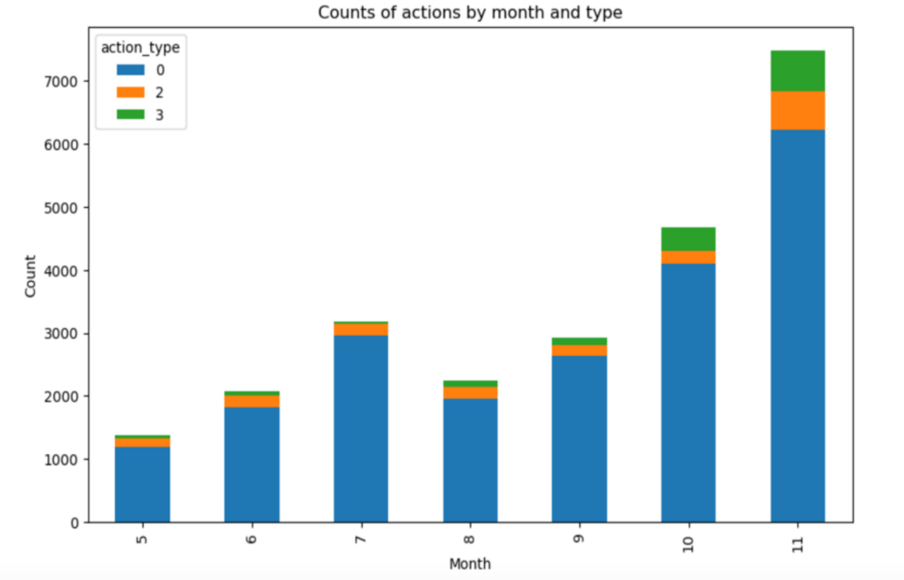


Figure 2: Counts of actions by month and type

Feature Engineering

1. **Average User Age for each Category**
2. **Purchase Average Time**
3. **Purchase Ratio**
4. **Purchase Frequency**
5. **Average User Age for each Merchant**
6. **Ratio of add-to-cart actions to clicks for each Merchant**
7. **Number of distinct brands a user has interacted with for each Merchant**
8. **Number of distinct categories a user has interacted with for each Merchant**

Feature Ranking

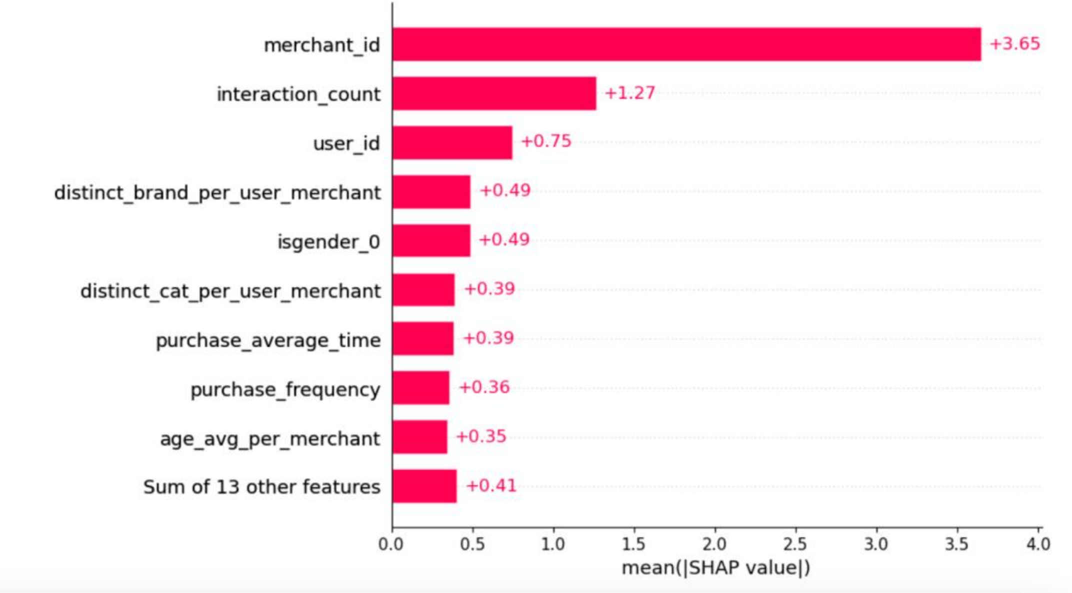


Figure 3: Important features based on SHAP

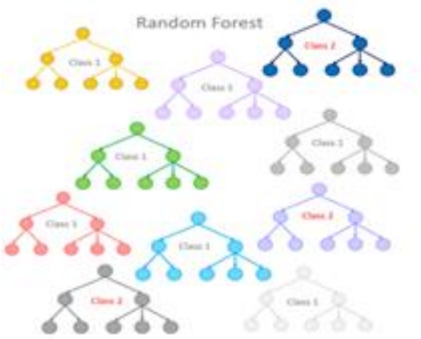
Model Predictivity

Split Data

- Randomly split data into
  - Training set : 80%, for model training.
  - Testing set : 20%, for model testing.
- First, we use the original data to train baseline models with 3 different algorithms
  - Random Forest
  - XGBoost
  - Neural Networks
  - Bayes classifier
  - KNN

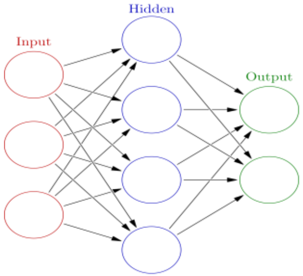
Random forest:

Different classifiers overfit the data in a different way, and through voting those differences are **averaged out.**



Neural Network

They consist of interconnected nodes or neurons that process and transmit information to make a prediction or decision. A neural network typically Consists of multiple layers, including an input layer, one or more hidden layers, and an output layer.

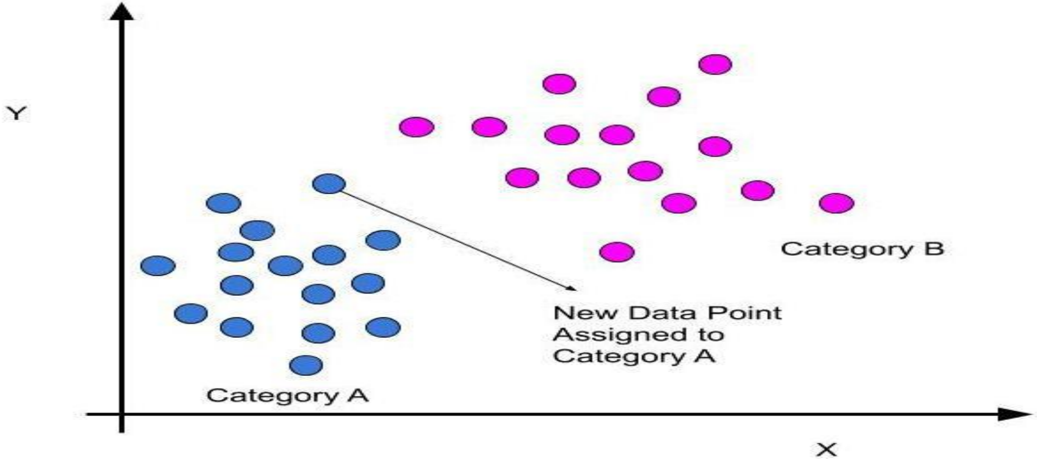


Bayes classifier:

Bayes' theorem is a fundamental principle in probability theory that describes the probability of an event based on prior knowledge or information.

KNN:

It is a non-parametric algorithm that makes predictions based on the k nearest neighbors of a new data point in the training data.



Performance on unseen tasks