Instacart Market Basket Analysis

Contents

You could add something here

Introduction

Instacar market basket analysis is a classic customer behavior prediction case. Instacart's data team has open sourced about 300000 order data. We will analyze users' shopping behavior through these data;

The project provides about 3000000 order data generated by about 200000 users, and each user provides 4 to 100 order data with product series. These data contain the time information of order generation and other important information (detailed research will be carried out in the third session). These data are divided into several CSV files and can be downloaded with one click.

Methodology

Data Analysis

Understanding data by exploring data sets is an important part of the project, as it can find valuable perspective, so as to obtain clean data and facilitate processing. The more we know about the data, the more we have the ability to accurately select and manipulate features to obtain the best results.

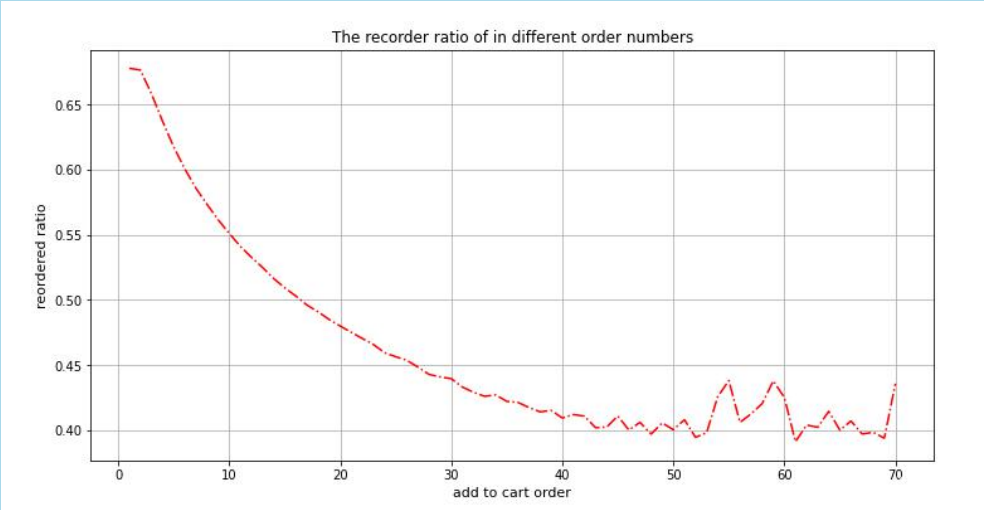
Data observation

1. Fruit and vegetable products have the highest consumption frequency, and the three products that sell the most are banana, bag of organic banana and orange strawberries. 2. The higher the average position of products in the shopping cart, the greater the possibility of re purchase; During precision marketing, products that are added to the shopping cart first should be recommended to users first.

Research Models

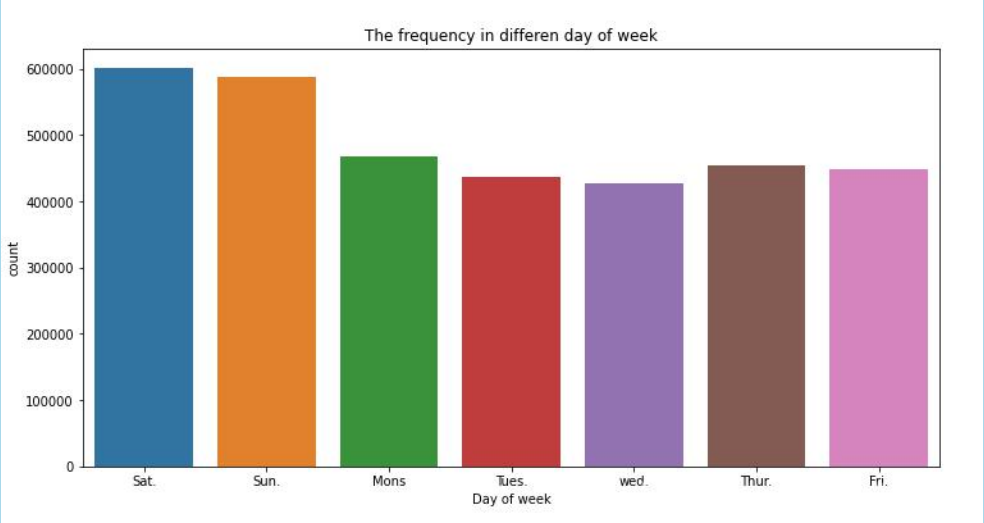
XGBoost & Lightgbm

Data observation



Reorder ratio

Through the dotted line diagram, we can clearly see that the higher the order in which products are added to the shopping cart, the greater the possibility of re purchase.



Shopping habits

We can look at the relationship between users' shopping habits and reordered. Obviously, Saturday and Sunday (0, 1) are the most frequent weekend shopping, and Wednesday is the lowest.

Research Models

XGBoost

Xgboost is a machine learning function library focusing on gradient lifting algorithm, which was born in February 2014. This function library has attracted extensive attention because of its excellent learning effect and efficient training speed. In 2015 alone, 17 of the 29 algorithms that won in the kaggle competition used the xgboost library, while for comparison, the data of the popular deep neural network method in recent years was 11. In the KDDCUP 2015 competition, all the top ten teams used the xgboost library. Xgboost not only has good learning effect, but also has fast speed. Compared with the implementation of gradient lifting algorithm in another common machine learning library, scikit learn, the performance of xgboost is often improved by more than ten times.

Light GBM

Gbdt (gradient boosting decision tree) is an enduring model in machine learning. Its main idea is to use weak classifier (decision tree) iterative training to obtain the optimal model. The model has the advantages of good training effect and difficult over fitting. Gbdt is not only widely used in industry, but also commonly used in multi classification, click through rate prediction, search sorting and other tasks; It is also a deadly weapon in various data mining competitions. According to statistics, more than half of the champion schemes in the kaggle competition are based on gbdt. Lightgbm (light gradient boosting machine) is a framework for implementing gbdt algorithm. It supports efficient parallel training, and has the advantages of faster training speed, lower memory consumption, better accuracy, distributed support, and rapid processing of massive data.

Implementation

Data Pre-processing

Data processing technologies are used before data mining, which greatly improves the quality of data mining patterns and reduces the time required for actual mining.

Training and Testing

The purpose of this competition is to predict what kind of goods instacart consumers will buy again according to customers' historical purchase records, so that the supply of goods can be sufficient when customers need this product.

Evaluation

This project will use the F1 evaluation algorithm recommended by kaggle.

Evaluation

This project will use the F1 evaluation algorithm recommended by kaggle.

$$F_1 = 2 \cdot \frac{1}{\frac{1}{\text{recall}} + \frac{1}{\text{precision}}} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Results and Findings

1. Fruit and vegetable products have the highest consumption frequency, and the three products that sell the most are banana, bag of organic banana and orange strawberries.
2. The higher the average position of products in the shopping cart, the greater the possibility of re purchase; During precision marketing, products that are added to the shopping cart first should be recommended to users first.
3. The interval for users to increase the number of products in each order is [5,8]. If this is the most acceptable shopping volume for most users, you can recommend [6,10] products to users each time and show users as many products as possible without affecting the user experience.

4. Users usually shop on Saturday and Sunday. Users are the least active on Tuesday and Wednesday. The daily passenger flow is mainly concentrated in 8:00-18:00 during the day. Generally speaking, users reach the peak from 6:00-9:00 on Sunday. These time periods can be preferentially selected for new product promotion advertising display.
5. Users of the same product re purchased two small peaks one week and one month after purchase. In addition, there was a small peak two weeks later. At these three time points, the products purchased by users should be recommended to users first.
6. The repurchase rate of personal care products is the lowest, so it is necessary to analyze the reasons with more characteristics.

Thank You For Watching !

Zhengjie He A1808830