**The second semester of the 2023-2024 academic year**

**South China Normal University**

华南师范大学阿伯丁数据科学与人工智能学院

**《Data Analysis and Application with Pyton》 final assignment report**

|  |  |
| --- | --- |
|  |  |
| **Major：** | IMIS |
| **StudentID：** | 20233803078 |
| **name：** | Fengrui Gan |
| **Teacher：** | Hai Liu |

1.Project objects

The research object of this project is an Amazon transaction report file named Table task, which includes many columns, mainly product\_id, discounted\_price, Actual\_price, discount\_percentage, orderTime, rating, rate\_count, and user\_id

2.Data set description

Discounted price reflects the discounted price, actual price reflects the original price, and discount percentage indicates the strength of the discount. Rating is the rating given by each user for the current transaction, rate\_count is the total score of the product, and user\_id reflects the user's ID information.

3.Project pipeline

Firstly, we imported a series of modules in the first stage.

In the second stage, we mainly carried out data preparation and cleaning. Firstly, we read the CSV file using pd.read\_csv() and named it df. Then, we reviewed the basic information of the file.

 In the process of data cleaning, we first check whether the product\_id is of the same size and whether they are duplicated. Subsequently, we cleared and removed the NaN values in actual\_price and discounted\_price, and also cleared the discount\_percentage in the row where NaN values existed. Afterwards, we cleaned up the rating column and rating count column, and processed the NaN values, which is the general clear process. In the third stage, we conducted exploratory analysis and visualized the results. Firstly, we calculated the sum of actual price and discounted price, and analyzed the price trend from January to December, visualizing it using a line chart. Secondly, we calculated the average value of ratings and visualized the average distribution of ratings using histograms and boxplots. Subsequently, we analyzed orderTime to determine which day of the week had the highest trading volume and the distribution of trading volume at each time period of each day. We visualized the results using a histogram. Afterwards, we analyzed the customer's repurchase rate based on user\_id and which month had the highest repurchase rate.

In the fourth stage, we conducted diagnostic analysis. The first question I raised is, what is the relationship between discounted\_price and actual\_price? Does discount\_percentage accurately reflect each discount level? To answer this question, I calculated the actual discount rate and compared it with the discount rate in the statistics. I printed out the inaccurate data and displayed the distribution of the inaccurate data in a scatter plot.

The second question I raised is whether discount\_percentage has a significant impact on ratings? To answer this question, I calculated the correlation coefficient between the discount rate and the average score, and I hope to use a scatter plot to fit the straight line.

I raised the third question, is there a correlation between reviewcontent and ratings? Regarding this, I first checked high rated comments and low rated comments, and conducted sentiment analysis on user comments to evaluate the consistency between comment content and ratings. I visualized the correlation coefficient between ratings and sentiment scores using a boxplot. Finally, I checked the rating distribution to see if there were any abnormally high or low ratings.

The fourth question, can the user\_id and orderTime data provide assistance in analyzing user behavior and habits? To answer this question, I counted the purchase frequency of each user and identified occasional and loyal customers. Then we classify users based on the purchase price and amount, such as high-value customers and potential customers. And a user's user type can be determined by entering user\_id.

In the fifth stage, we conducted predictive analysis and visualization. After importing the necessary libraries, we defined the target variable rating and the feature variables used for prediction, including discounted\_price, actual\_price, discount\_percentage, and rating\_count. Subsequently, we divided the dataset, defined the size of the validation and testing sets, the number of trees used in the random forest, and the number of parallel jobs. So we initialized a random forest regression model for model prediction. Calculate and output the mean square error and R ² score of the model to evaluate its performance.

4.Achievement in assignment

In this assignment, we preliminarily analyzed that there is a higher distribution of general trading on weekends, and the peak of trading starts from 3pm to 12pm every day, with few trading occurring in the early morning. Secondly, there is no clear pattern in the repurchase rate of users, which fluctuates in a straight line. There are a few discount\_percentages that are not very accurate, and we analyzed the relationship between discount intensity and ratings, and their correlation coefficient is -0.4701, indicating a negative correlation between them, that is, discounts will have a negative impact on ratings. Then we checked the consistency between comments and ratings, and also found that their correlation coefficient was 0.274. This means that there is some correlation between the ratings given by users and their emotional scores (emotional polarity) in their comments, but this correlation is not particularly strong. Subsequently, we analyzed the user types and found that no customers were high-value or loyal, which also reflects that the repurchase rate of customers is not high enough. Finally, we used a random forest model to predict the score and found that the mean square error was very small but the R ² score was also not high, indicating that the model's ability to interpret data was weak and the prediction effect was not ideal.

**References:**