Project description

The aim of this project was to try to predict the rating of users by analysing the historical data of Brazilian e-commerce companies.

At the beginning of the project, in order to identify the features that would help the model to make predictions, the team used word-cloud technology to search for keywords in the historical reviews of the company and found that among the many reviews there were keywords related to: **product, delivery, time** and **cost**.

So based on the above keywords, the team chose four feature values as follows.

1. shipping cost ratio -> (payment value-item price)/order price

2. difference between estimated time to real delivered -> Order estimated delivery date - Order delivered carrier date

3、 Product description length -> Assume that the length of the product description has an impact on the product information

4、 Product photo qty -> Assume that the number of product images has an impact on the product information

The team divided the Y values into two cases.

The first: a group of 1-3 customer ratings, a group of 4, a group of 5, a total of three classes

The second: a group of customers with ratings below 4, a group of customers with ratings above or equal to 4, a total of two classes

Predictive models

A total of five algorithms were used in this study: support vector machines, decision trees, logistic regression, random forests and deep learning algorithms.

The best result is Random Forest ,f1-score = 0.4349136314110886 for three classes.

The best result is Random Forest ,f1-score = 0.6251761807865373 for two classes.

(Results may vary depending on each optimization! )

In the exercise of the model, the team optimized the data collected from the company, eliminating all the null values and balancing the distribution of the data in the algorithm optimization