Store Management for online shopping

1. INTRODUCTION

With the online shopping rising, retailing is severely impacted. Layoffs and reasonable purchase have become the top priority options for river closure. In this program, we aim to find the best schema of layoff and purchase by using machine learning to create the optimal model for predicting future trendy with related history dataset.

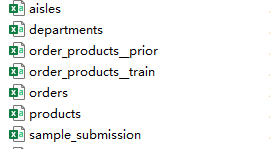
1. GOALS

As the trendy of benefit maximization, store management should be optimized with machine learning. We choose history data information about Orders, Products, Hour of a Day, The Day of a Week and so forth as the dataset, to analyze and predict two things to help reducing the cost of managing a store.

Specifically, we set two goals to do this.

1. Predict the product that the user would buy again in the next purchase.
2. According to dataset we can predict in one week which day is the most busy and which day order number is the least. Accordingly, store supervisor would like to arrange schedule more reasonable according to this prediction.
3. DATA CLEANING
4. Input Data

The data are given as CSV format, and the data amount is up to 200M.



After investigation, all the order records within a week are included in a CSV file, and they are divided into three different sets, Prior, Train and Test. Prior means all the history orders which the user made. Train data are the ones that need to be trained to build the model, while Test are the one we are going to use to predict the final results.

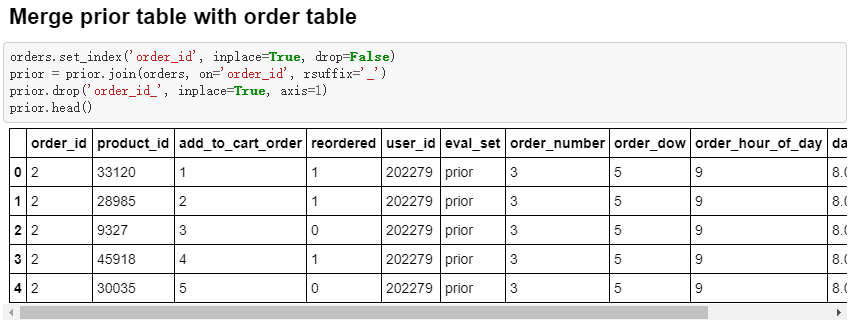
1. Remove Null/Empty elements

There is no NAN value in any table except the column orders['days\_since\_prior\_order'].

And the NAN value in orders['days\_since\_prior\_order'] means the order is the first order for that user so it doesn't have prior order. Therefore, there is no need for us to remove any row with NAN value.

1. Merge Tables

For the first goal, we need to get different features from users’ order history. Therefore, our first step is to combine the “order\_products\_\_prior.csv” with “orders.csv”. In that way, we get all the necessary information that we need.



1. ALGORITHMS

* Predict the product that the user would buy again in the next purchase.

1. **LightGBM**

LightGBM is a gradient boosting algorithm framework based on decision tree developed by Microsoft. It is designed to be distributed and efficient with the following advantages:

1. Faster training speed and higher efficiency.
2. Lower memory usage.
3. Better accuracy.
4. Support of parallel and GPU learning.
5. Capable of handling large-scale data.

The base idea of LightGBM is still decision tree while it improves the accuracy and efficiency. Therefore,