Group Project Description

IE4211, Modeling and Analytics

Li Xiaobo

February 28, 2023

1 Introduction

In this project, you are asked to form a group of size 3 to 5 and solve a real-world analytics problem. The detailed description is provided in the following sections. The due day for submitting the project is 23:59, Sunday, 16 April, 2022.

2 Project Description

We manage to obtain the sales data of around 100 products for around 60 days in an e-commerce company. Some attributes of products, customers and the day. We have done the preprocessing of the data and split the data into the training and test data set. We provide complete information on the training table but hide the column 'sales' on the test table. Note that these products belong to the same category.

You need to predict the sales on the test table. Moreover, you would need to make the inventory decision. The performance of your project depends on the prediction accuracy and the profit induced by your decision.

2.1 Data

In the "Data-Train" table, you are provided with daily sales of different products of the same type together with different types of attributes. The meaning for each column is as follows.

- productID: a unque code for each product.
- brandID: a unique id for each brand.
- attribute1, attribute2, attribute3, attribute4: different attributes for the product.
- clickVolume: the number people who clicks the product on that day.

- avgOriginalUnitPrice: among the purchase of that product on that day, the average listed price of that product. Note that the listed price is the price you saw when the customer click the product, which could be different for different time.
- avgFinalUnitPrice: among the purchase of that product on that day, the average final price of that product. Note that the final price is the price that customer actually pay, which could be different from the listed price since customers might have coupons or enjoy discount for bulk purchase.
- ma14SalesVolume: average of the sales over the last 14 days.
- weekday: day of the week, 1 for Monday and 7 for Sunday.
- meanAge: average of the estimated age among the customers that purchase the product.
- gender: percentage of male among the customers that purchase the product.
- maritalStatus: percentage of customers purchasing the product that are married.
- meanEducation: average education level of customers that purchase the product. The higher education level means that the customer has a higher degree.
- plus: percentage of customers that are the member of the company.
- meanPurchasePower: average purchase power of the customers that purchase the product.
- meanUserLevel: average user level of the customers that purchase the product.
- meanCityLevel: average city level of the customers that purchase the product. The higher city level means the city is larger.
- sales: total number of sales of the product on that day.

In the “Data-Test” table, you can find all of the above columns except for the sales.

2.2 Tasks

In this project, you need to find patterns from Data-train.csv and predict the sales for Data-test.csv. The performance of the prediction would be judged by the mean square error. The smaller the error is, the better your prediction is.

Moreover, we assume that the company need to make the inventory decision every day. The product costs 12 dollars to purchase and can be sold at the price of 20 dollars. If the product could not be sold by the end of the day, the firm can salvage the product (for instance quick sell to a vendor) at the price 8 dollars. Given this situation, you need to make the inventory decision for each day. We would then calculate the profit induced by your inventory decisions.

Note that in reality, the firm can stock the inventory to the next period. To solve this more realistic setting, one has to use the dynamic programming method, which is out of the scope of this course. Our setting is a reasonable simplification of the reality that can capture some insights into the problem.

2.3 Grouping

You can form a group of size 3 to 5 to do this project. You need to sign up your group members through the google docs: `https://docs.google.com/document/d/1qnRP1nwlCDFBiMVmvx0oc1ckw8` `edit?usp=sharing`

Please sign up the names of your group members before the deadline. After that, I will randomly assign the rest of you into groups. I will also assign a unique group number for each group. All group members have to contribute to the project. If some group members are taking the free ride and/or is not responsive, please write to me and I would investigate that.

2.4 Submission

To complete this project, each group need to submit the followings.

1. The csv file that contains your predictions of the sales and the inventory decision. The format of the file (including the name) should follow strictly group##.csv where ## is your group number. You can run the "TestYourSubmission.ipynb" to check whether the format of your csv file is correct.

2. Jupyter notebook (both .ipynb and .html files) that contains your data exploration and the production of your prediction. Make sure that you make necessary commenting on your codes, in particular those parts not taught in lab. Make sure that your notebook is runnable and your csv file can be reproduced through running your notebook.

3. Report. In the report, you must include the followings:

    - explain how to arrive at your prediction and the inventory decision;
    - briefly summarize the data exploration you have done in this project;
    - summarize your key findings.

    Please make sure that your report is concise and precise. It should be a summary and highlight of your jupyter notebook.

Be careful about the followings

1. Each group can submit only one file into each folder.

2. You need to make sure that your group number is the same as the number indicated on google docs, which will be finalized after the sign up deadline.

3. The submission of the CSV has to follow strictly the format.
4. The deadline for submission is strict. If you submit the project one day late (even if only one minute late), your score would be deducted by 10%. The deduction would be 20% if you submit two days late. Any submission after two days passing the deadline is not acceptable.
5. Make sure that your notebook is runnable and your csv file is reproducible. Grades would be deducted if this is not the case.
6. In your notebook, we recommend you not to load other files than the files we provide on LumiNUS. If there is a need to load other files, please include them in your submission.
7. If you refer to some research papers, make sure that you cite them in your notebook or in the report.

2.5 Rubrics

The final grade (100 pts) is composed of the followings.

- Correctness(20 pts). Can codes be run without any bugs? Is the format of your submission correct? Is your prediction based on an analytics model, rather than a random guess or an "eyeball" prediction?
- Report and notebook(20 pts). Do you write your report and notebook clearly? Do you make necessary explanation on your codes? Note that the report point has nothing to do with the length of the report. Also, using the methods not taught in this course will not grant you extra points, unless you have evidence that it improves the prediction accuracy or the generated profit significantly.
- Out-of-sample error for your prediction of sales (20 pts). We will calculate the mean square error between your prediction and the ground truth. Your grade is higher if your prediction is more accurate.
- Out-of-sample profit generated by your inventory decision (20 pts). The higher the profit you achieve, the better scores you would get for this part.
- In-sample analysis (20 pt). You can validate your methods using in-sample analysis such as cross-validation. Because the test data might be noisy, the best method that works in-sample might not be the best method that works for the test data. If your methods perform well based in-sample analysis, you can earn some points. Make sure that you do cross-validation correctly to get these points. For ease of comparison, please perform 10 fold cross validation for your methods. Please include the results for the 10-fold cross validation result for your best models into the report. Note that you can also do in-sample analysis for the profit, if you would like to.

3 Permissions and Cautions

Permissions:

1. You are allowed to use any existing packages/libraries/functions in python.
2. You can refer to textbooks and/or research articles etc from Internet, library, newspapers etc.. But you need to cite the sources properly if you use the resources other than our textbooks.

Cautions:

1. Make sure that your prediction file is reproducable. You np.random.seed() or other methods to ensure that.
2. For all the knowledge that is not discussed in the lecture/lab/textbook, you need to indicate the reference from which you learn about them. If you know this from other courses/projects, you can indicate that as well.