Project 2: Walmart Store Sales Forcasting

Fall 2024

Contents

Overview

Given historical sales data from 45 Walmart stores spread across different regions, your task is to predict the future weekly sales for every department in each store.

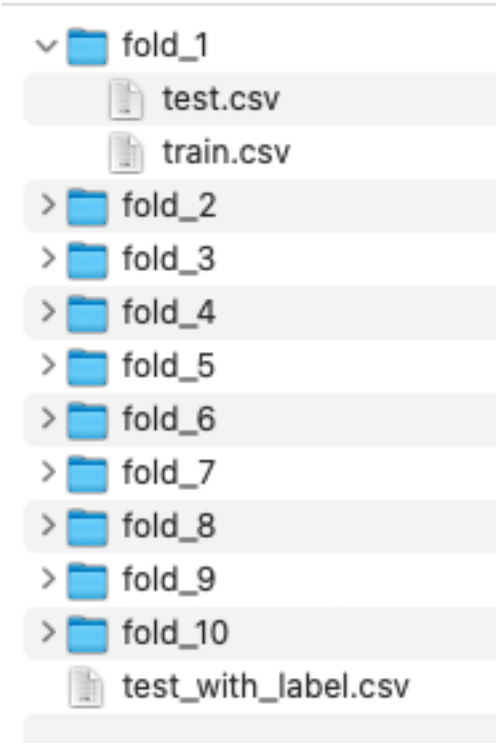
The dataset is from https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting.

Please Note::

1. We will **only** be using the **training data** in this project.
2. Kaggle's competition includes an additional CSV file with features such as temperature, fuel price, CPI, etc. This feature set is **not** utilized in our project.

Datasets

Download the dataset from the link: [proj2.zip]. Extract the zip to find 10 folders and a file named `test_with_label.csv`. In each folder, there are two files: `train.csv` and `test.csv`.



- **train.csv** contains 5 columns ("Store", "Dept", "Date", "Weekly_Sales", "IsHoliday") and ranges
  - from 2010-02 (February 2010) to 2011-02 (February 2011) in fold_1,
  - from 2010-02 to 2011-04 in fold_2,
  - from 2010-02 to 2011-06 in fold_3,
  - .....
  - from 2010-02 to 2012-08 in fold_10.
- **test.csv** contains 4 columns ("Store", "Dept", "Date", "IsHoliday") and ranges
  - from 2011-03 to 2011-04 in fold_1,
  - from 2011-05 to 2011-06 in fold_2,
  - from 2011-07 to 2011-08 in fold_3,
  - .....
  - from 2012-09 to 2012-10 in fold_10.
- **test_with_label.csv** is formatted similarly to **train.csv** and ranges from 2011-03 to 2012-10.

Objective

Predict the weekly sales for the subsequent two months for every combination of Store, Dept, and Date in **test.csv** using the historical data from **train.csv**.

Code Evaluation

Name your script as **mymain.R** (for R) or **mymain.py** (for Python).

Execution:

- For R: We'll run `source(mymain.R)` in RStudio from a clean environment (meaning, no pre-loaded libraries).
- For Python: We'll execute `python mymain.py` from the command line.

We'll execute your code **inside** each of the 10 folders.

- Please avoid using any commands that access or reset the current directory.
- Please ensure that your script only accepts 'train.csv' and 'test.csv' as inputs, without any additional path specifications. For example, your script should NOT include paths like 'XX/Proj2_Data/fold_1/train.csv'.
- No evaluation is required in your script. Your submitted script should not attempt to access 'test_with_label.csv'
- Please do not write your code to automatically iterate through all 10 folders. We will run your code inside each folder individually.

The 10-folder setting here differs from Project 1, as in Project 2, you can check the `Date` column to detect which folder your code is being tested on. Therefore, students are **allowed** to use different methods for different folders.

After successful execution, we anticipate finding a new CSV file named **mypred.csv** in the respective directory. The file **mypred.csv** should look as follows:

```
Store,Dept,Date,IsHoliday,Weekly_Pred
1,1,2011-03-04,FALSE,21827.9
1,1,2011-03-11,FALSE,21043.39
1,1,2011-03-18,FALSE,22136.64
1,1,2011-03-25,FALSE,26229.21
1,1,2011-04-01,FALSE,57258.43
1,1,2011-04-08,FALSE,42960.91
......
```

Evaluation Metric. We use the same evaluation metric as the [one described on Kaggle], which uses higher weights on the following **four** holiday weeks:

- Super Bowl
- Labor Day
- Thanksgiving
- Christmas

Performance Target. See Campuswire

Submission Guidelines

Submit the following **two** items on Coursera/Canvas:

- **Code**: Your R/Python script should be in a single file named either **mymain.R** or **mymain.py**. This script should:
  - Accept `train.csv` and `test.csv` as inputs.
  - Generate one file named `mypred.csv` based on the specified format (described before).
  - Important: Do not submit ZIP files or markdown/notebook files.
- **Report**: Submit a concise report (maximum of 2 pages, in PDF format) which contains two sections:
  - **Section 1: Technical Details**: Discuss key details such as data pre-processing and implementation aspects of your models. Do NOT paste your code in the report; instead, explain the technical steps in plain English. Your description should be detailed enough to allow your fellow PSL classmates to accurately replicate your results.
  - **Section 2: Performance Metrics**: Report the accuracy of your prediction on each of the 10 test datasets (refer to the evaluation metric described above), the execution time of your code, and details of the computer system you used (e.g., Macbook Pro, 2.53 GHz, 4GB memory or AWS t2.large) for each of the 10 folders.

- Students are permitted to use the code we provided on Campuswire, so you can disregard the Plagiarism Report on **Coursera**.
- On **Canvas**, when you submit your file multiple times, it may rename your file from 'mymain' to 'mymain-1' or 'mymain-2'. That's fine, no need to worry about it.