CSE 519 -- Data Science (Fall 2021)
Prof. Steven Skiena
Homework 3: Data Integration and Modeling
Due: Tuesday, October 19, 2021

This homework will investigate data integration and model building in IPython. It will be based on the Rossmann Stores Sales prediction task on Kaggle, where you will be forecasting the daily store sales up to six weeks in advance. More than just data exploration, you must also join the challenge and submit your model, to get a score from Kaggle. You are to explore the data and uncover interesting observations about store sales. You will need to submit your code files in three different formats (.ipynb, .pdf and .py). Make sure to have your code documented with proper comments and the exact sequence of operations you need to produce the resulting tables and figures. The submission steps are discussed below.

Data downloading

First of all, you need to join the challenge and download the data here. The description of the data can also be found at this page.

Python Installation

Instead of installing python and other tools manually, we suggest installing **Anaconda**, which is a Python distribution with a package and environment manager. It simplifies a lot of common problems when installing tools for data science. More introduction can be found here. Installation instructions can be found here.

Another option can be using **Google Colaboratory.** This is an option for those who want to run their Jupyter notebook remotely instead of installing the required packages locally. Colab allows you to write and execute Python in your browser, with

- Zero configuration required
- Free access to GPUs
- Easy sharing

If you are an expert of Python and data science, what you need to do is install some packages relevant to data science. Packages that I believe you will definitely use for this homework include:

- pandas
- scikit-learn
- numpy
- Matplotlib
- seaborn

The google colab notebook contains boilerplate code to download the data to your google drive and a dictionary containing the features along with its data type. The colab notebook has been posted to the google classroom.

All comments by competition host on this contest: FlorianKnauer | Discussion Novice

Tasks (100 pts)

1. Take a look at the training data. Combine the tables in store.csv and train.csv into a single dataframe. (5 points)
2. Do people shop more during the holidays or before the holidays? Analyze how different types of holidays affect the sales. (5 points)
3. Amongst the stores with at least 6 months of sales data, list the IDs of:
    - The five stores with the highest cumulative sales
    - The five stores with the least cumulative sales

    a. Plot the sales per week over time for these two sets of stores. (5 points)
    b. How similar are the patterns of sales each week amongst these two sets of stores? Make plots to reveal them. (5 points)
4. Plot the sales per week against
    a. **Distance of the closest competitor.** Do the stores farther to competitors have a better sale per week than the closer ones? (5 points)
    b. **DELETED**
5. Select a set of the five most interesting features (includings *sales*). Compute the Pearson correlation between all pairs of these variables. Show the result using a heat map, and list the feature-pairs with the strongest correlations. Which feature correlates the best with *sales*? How does this change with Spearman correlation?
   You can use the seaborn library to plot the heatmap, with instructions found here. (10 points)
6. For each of three different variables (one likely good, one presumably meaningless, one at random), build single-variable regression models, and do a permutation test to determine a *p*-value on whether the predictions of the sales are better than chance. Use root-mean-squared error of the log(*sale*) as the statistic to score your model. In other words, compare how your model ranks by this metric on the real data compared to 100 (or more) random permutations of the sales assigned to the real data records. (15 points)
7. Produce five informative plots revealing aspects of the combined data. For each plot, describe interesting properties your visualization reveals. These must include:
    - at least one line chart
    - at least one scatter plot
    - at least one histogram or bar chart

   (15 Points)
8. Create a training set and a validation set using the data given in train.csv. The validation

set must contain all the data from May, June, and July of 2015. The training set will consist of the rest of the data. Build **two different** prediction models to solve the task. Evaluate your model on the validation set using Root Mean Square Percentage Error ($RMSPE$). You are free to do any kind of preprocessing, and use any algorithm for training. Explain the hyperparameters of your model. Report how the performance of the model and the time taken for training changes for different hyperparameter settings. You should try at least three different hyperparameter settings for each model. (15 points)

9. For the two models from above, perform a t-test to evaluate whether their predictions are significantly different. (10 points)
10. Predict the sales for all the test instances in "test.csv". Write the result into a csv file following the format of the file "sample_submission.csv" and submit it to the website. Do this for both models you develop. Report the private score, the public score for your highest scoring model and the total number of submissions you have made on Kaggle. Include a snapshot of your best score from my submission page as confirmation. Be sure to provide a link to your Kaggle profile. (5 points)

Rules of the Game

This assignment must be done **individually by each student**. It is not a group activity.

1. If you do not have much experience with Python and the associated tools, this homework will be a substantial amount of work. Get started on it as early as possible!
2. All of your written responses should be put in the appropriate place in your notebook template. Get the template notebook form from Google Classroom.
   You are allowed to add more cells, but definitely fill out the cells we give.
3. We will discuss topics like logistic regression in detail only after the HW is due. Muddle along for now, and we will understand the issues better when we discuss them in the course.
4. To ensure that you are who you are when submitting your models, have your Kaggle profile show your face as well as a Stony Brook affiliation.
5. There are some public discussions and demos relevant to this problem on Kaggle. It is okay for students to read these discussions, but they must write the code and analyze the data by themselves.
6. You will submit your code so we can run it through MOSS to detect copying and plagiarism. Do your own work!!
7. Our class Piazza account is an excellent place to discuss the assignment. Check it out at piazza.com/stonybrook/fall2021/cse519.

Submission

Submit everything through Google classroom. As mentioned above, you will need to upload:

1. The Jupyter notebook all your work is in (.ipynb file), derived from the provided template
2. Python file (export the notebook as .py)

3. PDF (export the notebook as a pdf file)

These files should be named with the following format, where the italicized parts should be replaced with the corresponding values:

1. cse519_hw3_*lastname_firstname_sbuid*.ipynb
2. cse519_hw3_*lastname_firstname_sbuid*.py
3. Cse519_hw3_*lastname_firstname_sbuid*.pdf