Assignment 4: Data Wrangling and Data Analysis

Due November 26st 2023

In this task, you'll use time-series forecasting to forecast store sales on data from Corporaci ́on Favorita, a large Ecuadorian-based grocery retailer Data on Kaggle. Specifically, you'll build a model that accurately predicts the unit sales for thousands of items sold at different Favorita stores. You'll practice your machine learning and time series analysis skills with an approachable training dataset of dates, store, and item information, promotions, and unit sales.

Question 1. [2.5 points] Load oil.csv. This file contains years worth of data of the daily oil price. However, the data is missing for a few days. Make sure that every day contains a value using any data imputation technique that you learned during the data preparation week or during the missing values imputation week.

Question 2. [2.5 points] Augment the data in test.csv and train.csv with the oil price

Question 3. [2.5 points] Note that the training set contains a 'sales' column while the test set does not. Use the training set to train a model of your choice and use that model to predict which values for sales should be in the test set. You should try training at least 2 models and compare their accuracy later.

Question 4. [2.5 points] Compare your prediction with the prediction found in submission.csv with 3 different methods:

(a) Root Mean Square Error (RMSE)

(b) Mean absolute deviation

(c) Another metric of your choice Compare the three errors. Are they in agreement? Do you think any of the methods is objectively better than the others in this case?

Submission Guidelines:

Create a PYTHON notebook (On Google Colab of jupyter notebook) and write the question, your analysis and the code for solving each task. Submit the PYTHON notebook as well ad a pdf copy of the notebook on Moodle.