Machine Learning – Assignment 2

Due Date: 14.05.2022

Data

You have been provided with sales data (*sales.csv*) for 111 products whose sales may be affected by the weather (such as milk, bread, umbrellas, etc.). These 111 products are sold in stores at 45 different Walmart locations. Some of the products may be a similar item (such as milk) but have a different id in different stores/regions/suppliers.

The 45 locations are covered by 20 weather stations (i.e., some of the stores are nearby and share a weather station). The key data (*key.csv*) indicates for each store to which weather station it belongs.

In addition, you have been provided with the weather data (*weather.csv*) of each weather station.

The data can be found here.

The sales and weather data are daily observations, from January 2012 to October 2014.

You should use the observations from the years 2012-2013 as your **training** data, and the observations from the year 2014 as your **test** data.

Section A (Data Exploration and Visualization) *10 pts*

Explore the data using tables, visualizations, and other relevant methods.

- Plots should have an informative main title, axis labels and a legend.
- For each plot or table, provide a short description of **key observations**. Make sure to only include content which would be **meaningful** for a Walmart store district manager.
- The goal of this section is to get insights on the data which may or may not be relevant for the following sections.

<u>Section B (Data Pre-processing)</u> *30 pts*

Apply different methods of pre-processing to the data in order to prepare it for the models you wish to apply in the next sections.

- Apply at least one type of imputation, one of transformations, and one of exclusion (i.e., feature selection).
- Provide an explanation to each method you apply. Your choice should reflect an understanding of the method and why it's needed.

<u>Section C (Unit Sales Prediction)</u> *25 pts*

Sum up the unit sales of product 5, 6, 9, 16, 45 to a new value called key_sum. Use at least **two** different machine learning models to predict the daily sales figures of key_sum per store on a given day using the weather and sales data.

- You can use all the data **except** for the features used to create key_sum (*this rule only applies for this section*).
- Feel free to create additional features based on analysis that you have produced.
- The implementation should include parameter tuning.
- Report a suitable measure to evaluate the performance of each model and compare the results.
- Present the models' results in a plot.

<u>Section D (Rainy Day Prediction)</u> *25 pts*

Based solely on the sales data table, use **two** different machine learning models to predict if on a given day it rained or not for store number 11.

- A rainy day is defined as a day in which the precipitation (preciptotal column), is greater than 0. Trace (T) is defined to be greater than 0.
- The implementation should include parameter tuning.
- Report a suitable measure to evaluate the performance of each model and compare the results.

Presentation *10 pts*

Create a short presentation (no more than 6 slides) that includes interesting findings of your choice. 3-4 presentations will be chosen to be presented in front of the class. The goal is to learn from other students' work.

Section E (Elevation Estimation - Bonus) *10 pts*

Estimate the elevation of each weather station based on the weather data.

- Note that there is no elevation data for the weather stations so you must come up with a way to estimate it using the given data (This is not a guess, there is a way to measure this using the given data).

Section F (Clustering - Bonus) *10 pts*

Apply a clustering algorithm on the weather data to cluster the weather stations. Can you identify similarities inside the clusters or differences between them? Discuss your findings and find a way to demonstrate visually what similarities the clusters may have.

Section G (Performance - Bonus) *5 pts*

Machine learning models that outperformed other students' models for either the unit sales predictions or the rainy-day classification may get additional points as long as the non-standard methodology to obtain superior results is also explained.

- In order to get the bonus points you may want to apply multiple performance measures to ensure that we can compare your performance on an equal basis to other projects, and that you did not sacrifice performance in a specific measure to outperform in another.

Submission

- The assignment should be submitted in pairs (only one submission).
- You are required to submit two files including sections A-F. One in **.ipynb** format and one in **.html**. Both files should also include the program's outputs. In addition, you are required to upload a **pdf** file of the presentation you prepared.
- The files' names should be of the form: **ML_HW2_#ID1_#ID2**.
- Assignments submitted late will receive a penalty of **3 points** for each day, up to one week. Later submissions will not be accepted.