MACHINE LEARNING I WORKGROUP PROJECT

Table of Contents

1) Introduction

Welcome to the workgroup project of *MBD-EN2024A-2_38R192_442434* course. In this project you will be working with a real-world dataset from the Kaggle platform corresponding to the “AMS 2013-2014 Solar Energy Prediction Contest”.

2) Objectives Summary

Your goals in this project will be two:

1. Perform some Exploratory Data Analysis, EDA, clean data and data pre-processing steps on this dataset (see Section 6.1) – **5 points**.
2. Train a Machine Learning model to try to predict solar energy production of one solar station using the given dataset (see Section 6.2) – **5 points**.

You do not need to achieve the two goals as they will be evaluated independently.

3) Submission

You must submit the following items

1. The Dataiku DSS project used to achieve both goals.
2. A .ppt or .pdf file with the visualizations carried out for the EDA section.
3. A brief document detailing the steps carried out. You do not need to write a formal or detailed technical document, just a list of the steps implemented, and the decisions made, so I do not “get lost” when looking at your project.
4. A .csv file of the predictions in the format indicated in the Evaluation section of the Kaggle competition, with estimated values for the station ACME.

The submission deadline for this project will be **Friday 18th, October 2024 (included).**

4) About Kaggle

Kaggle, https://www.kaggle.com/, is an online community of data scientists and machine learners, owned by Google, Inc., which allows users to find and publish data sets, explore and build models in a web-based data-science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges.

In particular, we will be working on the “AMS 2013-2014 Solar Energy Prediction Contest”. You can check the description, evaluation metric used, leaderboard, how to make a submission and the discussion forum corresponding this competition accessing this url: https://www.kaggle.com/c/ams-2014-solar-energy-prediction-contest/overview/evaluation. It is highly recommended to carefully read this information.

Bear in mind than in this project it will only be required to make predictions for the station ACME, while the Kaggle contest requires to submit predictions for the 98 solar stations available.

5) Dataset

In this project you will work with a partly preprocessed dataset created from the original data given in the Kaggle dataset. You can find this dataset in the file *solar_dataset.csv*, which contains data with the following properties:

- A total dimension of 6909 rows and 456 columns.

- Each row corresponds to information of a particular day, ranging from 1994-01-01 to 2012-11-30. The first column, ‘Date’, informs you of which day corresponds to each row.

- The next 98 columns (from 2nd to 99th position) gives the real values of solar production recorded in 98 different weather stations. These columns are only informed until 2007-12-31 (row 5113); after this date these 98 columns contain NA or missing values. These missing values are the ones that you must predict for the station ACME to achieve the second goal of the project (see Section 6.2).

- The remaining columns are variables created from different weather predictors given in the Kaggle competition. They are the result of performing Principal Component Analysis, PCA, over the original data.

You have also available two other files:

1. **station_info.csv:** File with name, latitude, longitude, and elevation of each of the 98 stations.

2. **additional_variables.csv:** 100 new variables to optionally add to the ones in solar_dataset.csv. All these variables correspond to real Numerical Weather Prediction, NWP, values. As in solar_dataset.csv, each row corresponds to a particular day.

It is not mandatory to use these files, use them only if you think they can be helpful in your analysis.

6) Objectives Description

6.1) EDA, cleaning, and pre-processing

The goal here is to perform Exploratory Data Analysis, clean data, and pre-processing on the given dataset to extract information, prepare the data for the machine learning model, and/or visualize the dataset.

Some **ideas** of what can be done here are:

- Compute statistics of each column.
- Compute correlations.
- Outlier detection/removal/correction.
- Data scaling e.g. subtract mean and divide by standard deviation.
- Dimensionality reduction.
- Visualization of column values and distributions.
- Visualization of correlations.
- Visualization on a map (*Maps* section inside *Charts*).
- Anything that comes to your mind and makes sense. Creativity will be rewarded.

Evaluation criteria:

1. Quality of the Dataiku DSS project.
2. Correctness in a statistical/mathematical sense.
3. Variety in the analysis.
4. Usefulness of preprocessed steps regarding the other objectives of the project.
5. Use of new types of visualizations/statistics no covered during the course.
6. Creativity.

6.2) Machine Learning Model

Your second and final goal is to train a machine learning model to predict the solar production in the station ACME from dates ranging from 2008-01-01 to 2012-11-30 (both included). These predictions would be uploaded to Kaggle after the submission to check your score and compare it to the other class groups results.

Some **ideas** of what can be done are:

- Use the final data after the pre-processing steps of 6.1.
- Split the dataset in train (model training), validation/cross-validation (model hyperparameters tuning) and test (prediction).
- Use the same the evaluation metric employed in Kaggle for evaluation/validation purposes.
- Build first some basic model using linear regression (or similar) to get an initial benchmark.
- Then try a comprehensive grid search for linear regression (or similar) models.
- You can also try mode advanced models: random forests, neural networks, SVMs, XGBoost, deep learning, etc. and check if your score is improving.
- For the final submission, select the model and predictions that gave you the best score in your validation or after submitting to Kaggle with predictions for the other 95 stations fixed to a constant value (so they do not affect the evaluation of the three stations you want to predict here).

Evaluation criteria: 2.5 points will be given considering only your score on Kaggle after uploading your predictions (with the other 95 stations fixed to a constant value). This will be rewarded following this criterion:

- **Position 1:** 2.5 points.
- **Position 2:** 2 points.
- **Position 3:** 1.5 points.
- **Position 4:** 1.1 points.
- **Position 5:** 0.75 point.
- **Position 6:** 0.5 points.

Only groups that submit to campus online a valid prediction file and that properly prove how they got these predictions submitting the corresponding Dataiku DSS project used to obtain them will be rewarded with points. Models that “cheat”, i.e. use any information outside the given files, or predictions created “manually” (not the output of a machine learning model) will get a zero score.

The other 2.5 points will be awarded considering the following criteria:

1. Quality of the Dataiku DSS project.
2. Project optimization: No redundant steps, use of automatic computations by Dataiku when possible, etc.
3. Impact of pre-processing steps in the final score.
4. Correct splitting of data in train/validation(cross-validation)/test and proper use of these datasets.
5. Good hyperparameter tuning via train and validation.
6. Test at least one advanced model (you have some examples in the **ideas** section), even if the final prediction does not come from this model because other gave better results.