INF161 project - bike traffic prediction

The goal of this project is to apply all material learned in INF161 and successfully complete a data science project. Please read the description carefully!

This project is a compulsory part of the course. This project contributes 50% to the final grade. The grade will be based on good choice of methods, correctness of answers, clarity of code and thoroughness and clarity of reporting.

The data for this project comes from Statens vegvesen and Geofysisk institutt. The trafficdata contains among othersthe columns *Dato*, *Fra tidspunkt*, *Felt*, and *Volum*. Our goal is to predict the *Volum* for those rows, where *Felt* is "Totalt". The 2022 *Volum* is missing and should be predicted. The weather data is split into one file per year and contains columns *Dato*, *Tid*, as well as weather-related columns. For prediction, you may use any weather-data that you think is relevant and was recorded no later than the *Fra tidspunkt*. The weather data uses the code 9999.99 for missing data.

Requirements

You will build a machine learning model to predict how many people cycle over Nygårdsbroen at a given time.

The system takes as its input the current date and time and current or past weather, and will output the predicted number of cycles. The work will consist of four parts:

- Data preparation and exploratory data analysis (40 pts):
  - Input: raw data
  - Output: data description and clean data ready for analysis
  - Features: This systems takes the provided data and generates a dataframe that can be used in the machine learning model. Data splitting, description, visualisation, and feature engineering are important parts your report should include reasons for the choices made within this system.
- Modelling and prediction (40 pts):
  - Input: prepared data
  - Output: machine learning model, expected generalisation RMSE
  - Features: This system takes the prepared dataframe and builds a machine learning model for predicting number of cycles. Model selection, feature selection and handling missing data are important parts of this system. You should evaluate at least 3 fundamentally different modelling approaches before selecting the final model. We evaluate the performance of the system by comparing the predicted scores with the known scores on a validation/test data set. Specifically, the system should be evaluated with the root mean squared error (RMSE) of predictions, i.e.
    $$\sqrt{\frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{N}},$$
    where N is the number of predictions, $\hat{y}_i$ is the i-th prediction and y_i is the corresponding true number of cycles. The system should report the expected generalisation RMSE.
- Prediction (10 pts):
  - Input: machine learning model and 2022 data
  - Output: predicted number of cycles
  - Features: Given new data, this system should return predictions of the scores.
- Website (10 pts):

  - Features: The website should allow users to enter a date, time and possibly weather information and return a predicted number of cycles. Note that this is a HTML document that exists on your personal computer that you open with your browser and not a website hosted on the internet.

Deadlines

The project has three distinct deadlines. For the first deadline, you will perform exploratory data analysis and prepare the data for analysis. For the second deadline, you will update your data exploration and preparation and design a machine learning model and predict numbers of cycles. For the final deadline, you will deliver a project that fulfills all the above requirements.

- Deadlines: Note that late submissions for any of the deadlines and peer reviews leads to lost points.
  - Deadline 1: Sunday, 18.09, 23.59
  - Peer review 1: Sunday, 25.09, 23.59
  - Deadline 2: Sunday, 16.10, 23.59
  - Peer review 2: Sunday, 23.10, 23.59
  - Final deadline: Sunday, 6.11, 23.59
- Deliver at MittUIB.no/assignments

Deliverables

All dealdines are mandatory. The first two parts will not be graded, but you will give and receive peer feedback that will improve your final project.

For the final submission, please provide the following:

- A jupyter notebook `preparation.ipynb` for data exploration and preperation. The notebook acts as both report and submitted code. It should contain all the code to reproduce your work and a report of all your methodological choices and results. Please "restart and run all" before submission, so that you submit a clean version.
- A jupyter notebook `model.ipynb` for modelling, fitting and selecting a machine learning model. The notebook acts as both report and submitted code. It should contain all the code to reproduce your work and a report of all your methodological choices and results. Please "restart and run all" before submission, so that you submit a clean version.
- A file `predictions.csv` with predictions for each date and time in 2022.
- A zip file that contains the file `app.py` for running a local website and all neccesary files such that `app.py` runs with the command `python3 app.py`. The website should then run at `localhost:8080/`.

Note that each notebook and the app need to run independently.

In addition to packages from the standard library, you may use the following python packages: `xlrd`, `numpy`, `pandas`, `scipy`, `sklearn`, `matplotlib`, `seaborn`, `requests`, `plotly`, `flask`, `django`, `waitress`. If you use any other packages we will not be able to run your app and you will fail the project.

Code should be documented and tricks (e.g. to avoid division by zero, to make sure it takes finite time to run, etc.) should be reported. The rational behind all steps in the code should be clear from the report.

NOTE: This project is a learning experience. If we see that you have copied your answers from online resources, you will get 0 points.

Model selection is an important part of the task and will be graded accordingly. All data exploration steps should be commented and conclusions drawn explained. Before applying machine learning algorithms, you should always consider (and report) what results you expect. When you have successfully applied machine learning algorithms, you should always comment on how well the results match your expectations.