Module 7: Data Engineering
Project Documentation

Marie-Christine Meyer

January 25, 2023

1 Goal

- ENFSolar.com is a comprehensive online directory of solar energy products, and particularly, solar panels
- The website only lists individual products, however. There is no tool to get an overview of what prices to expect for a given type of solar panel
- The goal of this project is to provide a simple tool for buyers to get a quick overview of the price they can be expected to pay for a given solar panel model

2 Data Processing

- The data was obtained through webscraping from the website ENFSolar.com
- Following webscraping, extensive data cleaning was necessary:
  - to obtain the adequate data type for numeric features like efficiency (in %), weight (in kg), and price (in EUR), the original string types had to be transformed into a float type and special characters (e.g., %) removed; range values were transformed into point values by taking the mean of the upper and lower range (this applied to efficiency)
  - For categorical features, preliminary EDA revealed that certain values occurred very rarely (e.g., certain brands or regions were represented by fewer than 5 datapoints, and often only 1 datapoint). These rare values were pooled into a 'Other' category for the features brand, region, and panel type
  - Finally, in preparation of the next step (modeling), the categorical features were dummy-coded using SKLearn's one-hot Encoder

3 Modeling

- As a baseline, a Linear Regression model was fitted and found to have relatively poor performance at ca. 0.65 accuracy
- Subsequently, a Random Forest Regressor model was fitted with the features Efficiency, Brand, Panel Type, Weight, and Region, and Price as target. The accuracy score obtained for the test set was 0.9

4 App Deployment

- After data preprocessing and modeling was complete, the next step was to build a price predictor app using Streamlit
- The cleaned-up dataset as well as the trained model were saved as pickle files in order to speed up running time of the app
- The app offers the option of inspecting the whole dataset as well as a quick visualization of manufacturing regions; a quick at-a-glance price summary displays the average as well as the min and max prices
- The core functionality of the app is the predictor: via drop-down menus, users can input their desired specifications and the expected price is displayed; behind the scenes this is achieved by using the model's *.predict* function with the user input as parameter

5 Tools

The following tools were used to complete this project:

- BeautifulSoup for webscraping
- Python libraries for pre-processing and statistical modeling such as pandas, numpy, and sklearn
- Python libraries seaborn and matplotlib for visualization
- Streamlit and Streamlit Cloud to create and deploy the app
- Github in conjunction with Streamlit Cloud and to make available the code and dataset