Apparent temperature Prediction using Weather variables

According to wikipedia - Apparent temperature, also known as feels like, is the temperature equivalent perceived by humans, caused by the combined effects of air temperature, relative humidity and wind speed. The measure is most commonly applied to the perceived outdoor temperature.

Is there a relationship between humidity and temperature? What about between humidity and apparent temperature? Can you predict the apparent temperature given the humidity?

Statistics and Machine Learning can help us answer these questions and identify various relations and predict variables given the appropriate data.

Apparent temperature is the temperature that it feels like to our body because of other weather variables other than temperature. This is generally higher in case of high humidity.

In this project we try to predict the apparent temperature given other weather variables in an hourly manner.

Prerequisites:

We would highly recommend that before the hack night you have some kind of toolchain and development environment already installed and ready. If you have no idea where to start with this, try a combination like:

- Python
- scikit-learn / sklearn
- Pandas
- NumPy
- matplotlib
- An environment to work in - something like Jupyter or Spyder

For Linux people, your package manager should be able to handle all of this. If it somehow can't, see if you can at least install Python and pip and then use pip to install the above packages.

Objectives in this project:

- Clean the data and drop useless columns.
- Make an EDA report .
- Visualize the distributions of various features and correlations between them.
- Feature engineering to extract the correct features for the model.
- Train a regression model to predict the apparent temperature

Dataset:

The dataset is in the form of a csv file and the link to download is given below:

Link:

https://drive.google.com/file/d/15JFnZhmpuBYSaK4JcoLL709g1LAfB0b6/view?usp=sharing

Dataset description:

The data set contains 6000 entries with 12 columns listing various environment variables and text descriptions

The csv file includes a hourly/daily summary for Szeged, Hungary area, between 2006 and 2008

Data available in the hourly response:

- time
- summary
- precipType
- temperature
- apparentTemperature
- humidity
- windSpeed
- windBearing
- visibility

- loudCover
- pressure

WorkFlow:

The workflow for the project is described in steps given below:

- Perform data cleaning using the pandas library. Which includes replacing the miscoded information and handling missing data.
- Make an Exploratory Data Analysis on the data.
- Visualize distributions and correlation of features
- Build a regression model taking the selected features and don't forget to tune the hyperparameters to prevent overfitting
- Predict the apparent temp for the split test data (Use 30% of the data for test)
- Experiment with all the algorithms and tune the hyperparameters to get the best results