Capstone project #2 project ideas

Mike (Xiangnan) Shi

1. **BUILDING ENERGY PREDICTOR (**[Kaggle link](https://www.kaggle.com/c/ashrae-energy-prediction/overview)**)**

**The Objective**

This is a Kaggle competition to develop models to predict metered hourly building energy usage in the following areas: chilled water, electric, hot water, and steam meters.

**Data**

* The training data comes from the historical energy usage of over 1,000 buildings over a three-year timeframe.
* Data features include datetime, location, building type, square footage, year built, local weather (temperature, humidity, wind velocity and others), etc.

**Limitations/Challenges**

This is a competition that has been closed. We can still do late submission and get scores, but we don’t seem to have access to the exact solution. We can consider splitting the 3-year data into train and test groups though to validate our model.

1. **SOLAR ENERGY PREDICTION CONTEST (**[**KAGGLE LINK**](https://www.kaggle.com/c/ams-2014-solar-energy-prediction-contest)**)**

**The Objective**

This is a Kaggle competition to predict the total daily incoming solar energy at 98 Oklahoma Mesonet solar sites.

**Data**

* The training data comes from the historical solar energy production from 98 sites from 1994 to 2007.
* Data features include datetime, site location, precipitation, radiative flux, temperature, humidity, etc.

**Limitations/Challenges**

Similar to topic 1, this is a competition that has been closed. We can still do late submission and get scores, but we don’t seem to have access to the exact solution. We can consider splitting the 13-year historical data into train and test groups though to validate our model.

1. **2015 FLIGHT DELAYS AND CANCELLATIONS (**[**KAGGLE LINK**](https://www.kaggle.com/usdot/flight-delays?select=flights.csv)**)**

**The Objective**

This is a dataset on the number of on-time, delayed, canceled, and diverted flights in 2015 published by U.S. Department of Transportation's (DOT). I propose to achieve two goals based on the dataset.

1. Conduct EDA to see if there is any pattern in delayed or canceled flights.
2. Build a model to predict the likelihood of a flight getting delayed or canceled.

**Data**

The data set has almost 6 million rows of historical flights. Each flight has 31 features including datetime, airline, route, scheduled time, actual time, canceled status, delayed status, delayed reasons, etc.

**Limitations/Challenges**

As flight delays and cancelations are not very common, most of the flights in the datasets are on schedule. Thus, there might be a data bias issue we have to be cautious about.