Capstone project #2 Proposal

Mike (Xiangnan) Shi

1. **Problem Statement and Context**

Building energy consumption forecast is often of great interest to building owners for the purpose of tracking energy usage, revealing energy usage anomalies, forecasting operating cost and budget planning. When combined with utility rate structure, this can further guide building owners and operators to operate the business to minimize expense and maximize profit. However, there isn’t always a tool available for every building owner to forecast energy consumption. The scope of this project is to develop such a tool and evaluate its performance when predicting energy consumption in both short term and long term.

1. **Solution and Approach**

The project will follow a Data Science Method that consists of 6 steps as below:

1. Problem identification
2. Data wrangling
   1. Examine and clean data to fix missing values, outliers and bad quality data
3. Exploratory data analysis (EDA)
   1. Explore the pattern of target data
   2. Explore the potential relationship between target data and relevant features
4. Pre-processing and Training Data Development
   1. Join the weather data and electricity data
   2. Define a holiday schedule and add it to the data as one feature
   3. Generate train and test data set
5. Modeling
   1. Build different models with both time series and random forest
   2. Tune the hyperparameters and find the best model
   3. Use the best model built to predict on test data
   4. Evaluate the model performance
6. Documentation
7. **Project Deliverables**

* A Jupyter notebook demonstrating the workflow of developing a model that’s accepted to meet the goal
* A project report documenting the workflow, assumptions and final results
* A slide deck summarizing the key steps and findings in the project

1. **Potential Constraints**

Most of the constraints come from the data. The dataset only covers one year of energy consumption for a group of buildings. It will be split into a train set (Jan – Sep) and test set (Oct – Dec). As a result, train set is not able to capture certain hidden features in test set because they’re different time periods. In addition, the only data features available are datetime related and weather. If the building consumption is impacted by other factors not captured in our data, our model will not be able to include them and the prediction quality will be compromised accordingly.

1. **References**

Kaggle data source link: https://www.kaggle.com/c/ashrae-energy-prediction/data