95845 Project Proposal

Lizhi Zhao lizhiz@andrew.cmu.edu
Meghan Clark meghanc@andrew.cmu.edu
Xuliang Sun xuliangs@andrew.cmu.edu

March 2018

1 Proposal Details

1.1 What is your proposed analysis? What are the likely outcomes?

Our proposed analysis is to estimate the global energy consumption of a building using data on whether or not it is a holiday, building location, historical weather data for geographic area close to the building, type of building, and historical building consumption data. With the combination of these data sources, our team will work towards creating a model that predicts building level energy use.

1.2 Why is your proposed analysis important?

The proposed analysis is important because the majority of energy inefficiencies occur during the transportation of energy. This is why there has been a recent movement for local energy production and energy grids. However for energy companies to confidently move from a national network to localized distribution, they need the tools to better match supply with demand. The only way to accomplish this match is by improving building level energy use forecasts.

1.3 How will your analysis contribute to existing work? Provide references.

The main goal of this project is select a machine learning algorithm to come up with a model that is more robust and precise to forecast building energy consumption with little data.

1.4 Describe the data. Please also define Y outcome(s), U treatment, V covariates, W population as applicable.

Four data sets are available for this project.

(1)Historical Consumption: A selected time series of consumption data for over 200 buildings.

(2)Building Metadata:Additional information about the included buildings.

(3)Historical Weather Data: This data set contains temperature data from several stations near each site. For each site several temperature measurements were retrieved from stations in a radius of 30 km if available.

(4)Public Holidays: Public holidays at the sites included in the data set, which may be helpful for identifying days where consumption may be lower than expected.

Submission frequency	ForecastId ForecastPeriodNS	6974 rows
Submission format	obsId, SiteId, Timestamp, ForecastId, Value	1,309,176 rows
Metadata	SiteID, Surface, Sampling, BaseTemperature, MondayIsDayOff, TuesdayIsDayOff, WednesdayIsDayOff, ThursdayIsDayOff, FridayIsDayOff, SaturdayIsDayOff, SundayIsDayOff	267 rows
Holiday	X, Date, Holiday, SiteId	8387 rows
Train	obsId, SiteId, Timestamp, ForecastId, Value	6,559,830 rows
Weather	SiteId, Timestamp, Temperature, Distance	20,017,278 rows

Y outcomes are the energy consumption values(variable name: Value)

U no treatment in this project

V covariates are The surface area of the building(variable name: Surface), The temperature as measured at the weather station(variable names: BaseTemperature, Temperature), The date information:(variable names: MondayIsDayOff, TuesdayIsDayOff, WednesdayIsDayOff, ThursdayIsDayOff, FridayIsDayOff, SaturdayIsDayOff, SundayIsDayOff, Date, Holiday)

W population are building sites being considered(variable name: SiteId)

1.5 What evaluation measures are appropriate for the analysis? Which measures will you use?

To identify if our algorithm is successful,for each building and test period, we will evaluate the forecasts of our neural network using Weighted Root Mean Squared Error (WRMSE) measure. According to the webpage of this competition, average NWRMSE for some categories of buildings will be computed to appreciate the actual performance on these categories. Since we're not given more data on these categories of buildings, what we're planning is to set a empirical benchmark for our accuracy during our project and try to reach this accuracy.

1.6 What study design, pre-processing, and machine learning methods do you intend to use? Justify that the analysis is of appropriate size for a course project.

We will have to pre-process the Timestamp data using basis functions to put dates on a 24 hr scale. We intend to use neural network for this project because we think the dataset may not be linear separable.In our dataset, more than 200 building sites are considered and three time horizons and time steps are distinguished, so it should be a reasonable size for a course project.

1.7 What are possible limitations of the study?

The possible limitation may be omitting some other important features which could have big effect on the energy consumption. In addition, having diverse building data may make our predictions for anyone building type less accurate. Lastly, our predictions are focused on energy consumption as electricity and may not apply to other utility usage such as natural gas and water.

1.8 References

https://www.drivendata.org/competitions/51/electricity-prediction-machine-learning/page/101/