INFO 7390 Midterm

Commercial Building Energy Consumption modeling

Instructions:

- You can use either Python or R to work on this case.
- No sharing of work. You can work in your teams only
- You are expected to submit a report that summarizes the key steps in your implementation as a flow chart and also submit fully functional code clearly packaged and fully documented with instructions to run your code.
- Deadline: 11/18/2016 11.59 PM. Late submissions not accepted.

Summary:

You have been offered an internship at Tarja and Pasi Inc. in Finland and you are excited to work on your first "international" project. Tarja and Pasi Inc. is an energy modeling consultancy and they have been approached by a Vokia Inc. to help monitor and reduce energy consumption in 33 of their buildings. Vokia Inc. owns these buildings and wants to understand and reduce energy usage and wants to make the buildings more energy efficient.

Data:

You have been given two input files.

RawData.csv

BuildingID	Building	Meter number	type	date	hour	Consumption (in KwH)

- Note that the data covers 33 buildings and has hourly power consumption info for close to 1 year
- Use only the **elect and dist_heat data** which covers electricity usage and heating energy consumption data

building	address	area_floor _m.sqr(in square meter)

- This data has the building address (of course in Finland) and the area of the building

Goals:

Your goal is to use the data science skills learnt in the class to understand the given data, get the weather data to do feature engineering and build models for prediction, classification and clustering as detailed below:

Part1: Data Ingestion and Wrangling (40 points)

- Use Python/R to ingest the data. Use missing value filling techniques to fill any data that isn't complete (Note: You need to ensure to ensure that the time series is complete)
- Extract features: Week of day, Month of year, Weekday/Weekend, Holiday (http://www.timeanddate.com/calendar/?year=2013&country=24) , Base_hour_Flag (Hours 0,1,2,3,4,22, 23 => True; Else False)

INFO 7390 Midterm

- Normalize Consumption with respect to area_floor _m.sqr to ensure you are working with Kwh/square meter
- Now you need to get the historical hourly weather information for each of the building based on the address. Here is the pseudocode:
    - Get the geo latitude and longitude from address for each building.
    - For example, you can use the Google API to do this.
        - https://developers.google.com/maps/documentation/geocoding/start
    - Once you determine the Latitude and Longitude for each building, retrieve the nearest airport location.
        - You can do this in many ways:
            - http://api.wunderground.com/auto/wui/geo/GeoLookupXML/index.xml?query=lat,long
            - https://pypi.python.org/pypi/get-weather-data etc. are examples.
    - Once you have the airport location, retrieve the hourly temperature for each building for the duration of the dataset.
    - Here is a sample from retrieved dataset.
        - *TimeEET,TemperatureF,Dew PointF,Humidity,Sea Level PressureIn,VisibilityMPH,Wind Direction,Wind SpeedMPH,Gust SpeedMPH,PrecipitationIn,Events,Conditions,WindDirDegrees,DateUTC 12:20 AM,44.6,42.8,93,29.53,5.0,SW,2.3,-,N/A,,Mostly Cloudy,220,2013-11-02 22:20:00*
- You can choose to retrieve this data from various other means. But do ensure you are getting hourly information and other information as illustrated above.
- Cleanup the weather data. Ensure:
    - You have exactly 1 data point for each hour for the time span covered in the consumption data set.
- Join the processed Raw Data, Address Data and Weather Data into one input file

Part 2: Modeling tasks. (50 points)

- You will need to create a daily base_hr_usage by {building, type, weekday, month, holiday}.
- Use the average hours {0,1,2,23} to compute this value.
- Create a column base_hr_usage and input the calculated value in that column. Note that you will have one value for each hour for the whole day.
- Create another column called "Base_Hour_Class". If the actual consumption is greater than the base_hr_usage, use High Else Low

Prediction:

Build a model to predict the consumption (KWH/sq m) for this dataset. Criteria:

1. Build Regression, KNN, Random Forest and Neural Network models for each **building dataset (33 different models). Optimize your model through feature engineering and model selection**
2. Build Regression, KNN, Random Forest and Neural Network models for all buildings **as one dataset. Optimize your model through feature engineering and model selection**

3. Compute MAPE, MAE, RMSE for all models.
4. Which model would you choose and why? Discuss in the report.
5. **Outlier Detection**
    - a. For the best model you want to use for each case (all models and one model (determined in 1 and 2), conduct residual analysis as follows:
        - i. Compute the predictions
        - ii. Compute Residuals
        - iii. Compute Standard Deviation on the Residuals
        - iv. Tag points as outliers if the residual is >= twice the Standard deviation.
        - v. (Note: When you are doing this for 1(33 models), you need to do residual analysis for each building.

Classification

Build a model to predict the **Base_Hour_Class** flag for this dataset. Criteria:

1. Build Logistic Regression, KNN, Random Forest and Neural Network models for each **building dataset (33 different models). Optimize your model through feature engineering and model selection**
2. Build Logistic Regression, KNN, Random Forest and Neural Network models for all buildings **as one dataset. Optimize your model through feature engineering and model selection**
3. Compute Confusion matrices and ROC charts for all models.
4. Which model would you choose and why? Discuss in the report.
5. **Outlier Detection**
    - a. For the best model you want to use for each case (all models and one model (determined in 1 and 2), conduct residual analysis as follows:
        - i. Compute the prediction flags
        - ii. Compute a new column called Outlier_day(If there is a mismatch in flags for 6 or more hours, tag all rows for that day as True else False)
        - iii. You need to do mismatch analysis for each building using both models (1 and 2)

Clustering

- Using the building features, cluster the buildings using K-means and Hierarchical clustering.
- Using Bend graphs, choose the optimal number of clusters. Discuss your cluster features in a report

Part 3: Visualization and Dashboard (10 points)

1. Bring the data from Part 1 and conduct exploratory data analysis in Tableau.
2. Build dashboards to discuss insights about each building and the usage for the year for both electricity and heating
3. Using R-serve, integrate your best prediction, classification and clustering models so that you can invoke it in Tableau.