Case: Machine learning with Energy datasets

Instructions:

- You should use Python 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 through github.
- Deadline: 11/02/2018 11.59 PM. Late submissions lose 10% points per day.
- 3 teams (picked at random) will present their work on 11/03/2018. Each team gets 10 minutes to present

Summary:

As interest in IOT and sensors pick up steam, companies are trying to build algorithms and systems to understand consumer behavior to help them make better decisions. One such application is energy modeling. Though, most consumers are aware of their aggregate consumption of energy, few are aware of how and where energy is consumed. With increasing sensors in equipment, it is becoming easier to find out which equipment/instruments consume the most power. AdaptiveAlgo Systems Inc. works on solutions to build algorithms and platforms to address energy modeling challenges. The company is putting together a solution for energy modeling and is interested in understanding consumer energy usage and the attributes that contribute to appliance energy usage. The data scientists there came across a recent paper and dataset and are interested in building various machine learning models that could contribute to understanding energy usage by appliances and the attributes that contribute to aggregate energy usage. With the knowledge of energy consumed by various equipment, seasonality and attributes like temperature and humidity, a machine learning model could be used to predict aggregate energy use.

Your team has been hired to conduct an in-depth analysis and provide insights on feature engineering and machine learning. AdaptiveAlgo only uses Python so your solution should be in Python. Each part should be a jupyter notebook

Part 1: Research

Review the following papers and provide a jupyter notebook for each paper.

A. https://www.sciencedirect.com/science/article/pii/S0378778816308970?via%3Dihub
B. https://www.sciencedirect.com/science/article/pii/S1364032116307420
C. https://www.sciencedirect.com/science/article/pii/S0360544212002903

Part 2: Exploratory Data Analysis

Preparation: Review the EDA tutorial here:

1. Video: https://www.youtube.com/watch?v=W5WE9Db2RLU
2. Code and Details: https://www.kdnuggets.com/2017/07/exploratory-data-analysis-python.html

Data for assignment: https://github.com/LuisM78/Appliances-energy-prediction-data

- Conduct an exploratory data analysis using Python packages (**plotly, seaborn, matplotlib etc.**) to understand the dataset.
- Put together a PowerPoint report with graphs and key insights garnered from this analysis.

Part 3: Feature engineering

Preparation:

https://jakevdp.github.io/PythonDataScienceHandbook/05.04-feature-engineering.html (1)

There are many features in the dataset. Which features are important? Do we need to do feature transformations? Is the data clean or are there problems that need to be addressed? Conduct a thorough feature analysis and use pre-processing techniques (1) that needs to be done to make the data usable. Use Python for this.

Part 4: Prediction algorithms

Try out Linear regression, Random forest, Neural networks to build prediction models in Python using sklearn. Compute RMS, MAPE, R2 and MAE for Training and Testing datasets. Which model would you recommend? Refer to Paper A for guidance on building models. (Note: The github has sample R code. You can refer to and implement "similar" models

Part 5: Feature Selection

Preparation:

http://scikit-learn.org/stable/modules/feature_selection.html#feature-selection (2)

https://machinelearningmastery.com/feature-selection-in-python-with-scikit-learn/

AdaptiveAlgo wants to understand the importance of the various features and how the features influence the output. Use different techniques and discuss which features are important. How do you quantify feature importance? How do you rank important features? Which features would you eliminate and why? Also try Exhaustive search, Forward search and Backward search for regression models

AdaptiveAlgo is curious about **tpot, featuretools(https://www.featuretools.com/demos), Boruta, tsfresh.** Use these packages and compare and contrast feature engineering in each approach.

Part 6: Model Validation and Selection

Preparation:

https://jakevdp.github.io/PythonDataScienceHandbook/05.03-hyperparameters-and-model-validation.html (3)

AdaptiveAlgo is interested in understanding hyperparameter tuning and model validation prior to model selection for production. Review (3) and discuss various approaches and discuss which approach you would recommend.

In addition, AdaptiveAlgo is intrigued by

- cross validation techniques
- Bias -variance tradeoff
- regularization (L1, L2, Elastic net)
- grid search options.

Illustrate all these in the context of model validation and selection.

Part 7: Final pipeline

Which model would you finally recommend and why?

Create a pipeline See (Feature selection as part of a pipeline in http://scikit-learn.org/stable/modules/feature_selection.html) on how to automate the entire model from data ingestion to final model prediction

Part 8: Report and Model development methodology*

- Put together a comprehensive report discussing your analysis in pdf.
- AdaptiveAlgo wants to see the whole project implemented as jupyter notebooks. Create jupyter notebooks for each part.
- AdaptiveAlgo also wants to see how you develop code and use github. Use and document the project using Github and use this for versioning of the code :http://nbdime.readthedocs.io/en/stable/ for version checking