UDACITY MLE ND – Capstone Project Proposal

Student: Bilgehan Kilinc

Date: 17.12.2019

1- Background

I've been working data science and machine learning as a self-learner enthusiast to improve my skills and toolbox as a business development professional and team leader.

Self-learning with online education has great merits but also has some drawbacks. One of the main ones is showing your skills on real world applications and data sets. Therefore, for my capstone project I've decided to enter an ongoing Kaggle competition to contest with experts and see my score against real people instead of only metrics. Also, before nanodegree my first attempt to enter a Kaggle competition was interrupted my health and family issues and this time I want to earn a respectful leaderboard place.

So, I've decided to enter ASHRAE - Great Energy Predictor III Kaggle Competition. Competition is a complex regression problem on energy consumption. Energy consumption predictions has a rich context since it has been studied all around the world.

In this huge context I tried to find researches similar to my project here and used key words such as:

- "*measuring efficiency of building retrofits with predicted consumption*"
- "*using regression to predict energy consumption*"
- "*measuring efficiency of building retrofits with machine learning*"
- "*measuring energy consumption of building retrofits with machine learning*"

I have found 2 of related research paper I'd like to mention here:

- https://viejournal.springeropen.com/track/pdf/10.1186/s40327-018-0064-7: This paper has given me idea of using Energy Performance Indicator which is energy consumption of building during a definite period normalized by floor area. With this new feature I will have one more feature to enrich my data set. Also, it uses the same mythology of applying various ML models.
- https://www.ashrae.org/File%20Library/Conferences/Specialty%20Conferences/2018%20Building%20Performance%20Analysis%20Conference%20and%20SimBuild/Papers/C013.pdf : Even though paper aims to solve energy classification of buildings and recommend retrofits, it has also supported yearly EPI as an important feature.

Url: https://www.kaggle.com/c/ashrae-energy-prediction

2- Problem Statement

ASHRAE - Great Energy Predictor III Kaggle Competition is all about creating models to predict hourly consumption data on 4 types of energy consumption meters (electricity, chilled-water, steam, hot-water).

In the competition overview competition host describes they have 3 years of data over 1000 buildings and their 4 types of consumption meters. Host has given 1 year of data as training data with labels and 2-year data for predictions as test. Final outcome will be predicted hourly meter readings which makes our problem as a complex regression problem.

As a problem statement host states that: building owners, financial intuitions and all related parties such as tenants has invested energy saving investments/retrofits. Therefore, buildings actual consumption data for last 2 years (scope of test data) has been changed. On the other hand, for comparison buildings owners needs estimation what would have been their consumption data if they hadn't invested efficiency retrofits. This is where competition predictions become important. Predictions will be used to measure efficiency of improvements on buildings. Which will lead to measure accuracy of investment decisions.

3- Data Sets and Inputs

Like all Kaggle Competitions problem data sets are available at Kaggle ASHRAE - Great Energy Predictor III web page for Kaggle contestants. There are 5 csv of folders:

- train: Host's given training data. Shape: (20216100, 4)
- test: Host's given training data. Shape: (41697600, 4)
- building_metadata: building definitions data. Shape: (1449, 6)
- weather_train: weather data for training data points. Shape: (139773, 9)
- weather_test: weather data for test data points. Shape: (277243, 9)

Our data set has different csv files with different features. For EDA first thing to do is merging additional information to train and test data. After merging, train and test data will have 16 original features. At first glance there are 1-datetime, 3-categorical, 11-numerical features and 1-numerical target variable. On the other hand, after EDA, some new features can be created (i.e. dayofweek or month) or some numerical features can be converted to categorical features due to high null entry percentage.

Finally, since it is an energy consumption prediction problem, I think weather data and time points will be important features.

Url: https://www.kaggle.com/c/ashrae-energy-prediction/data

4- Solution Statement

Problem is a regression problem for all consumption points(buildings) and meter types (electricity, chilled-water, steam, hot-water) with target value of meter readings.

My solution plan is first making exploratory data analysis and create cleaned data to feed my models. Then I am planning to train and compare 3 or 4 regression models such as classical Tree Based Regressor, SVR or Neural Network Regressors or new and effective algorithms such as XGBosst Regression or LightGBM Regression.

After model selection I will tune hyper parameters, record final local training metrics. Also, I will make predictions with completions unlabeled data make my final submission to it.

5- Benchmark Model

As explained at section 4- Solution Statement, I am planning to use 3 or 4 models to compare at performance and then will choose best model to tune make predictions. On the other hand, as suggested I am planning to use a simple Linear Regression model for benchmarking. My benchmark linear model will use same data input and will be evaluated same metrics. By this way before my final solution and submission to Kaggle competition I'll have local comparison between benchmark model and final tuned solution.

6- Evaluation Metrics

Competition host uses root-mean-squared-logarithmic-error as its final evaluation metric. Also some regression models does not support natively RMSLE, therefore I will consider root-mean-squared-error on implementation but final evaluation metric will be RMSLE.

Url: https://www.kaggle.com/c/ashrae-energy-prediction/overview/evaluation

7- Project Design

My proposed project outline:

- EDA: It will be performed on Jupyter Notebook. Results will be saved as csv.
- Model Building: 3 or 4 Regression models will be trained and RMSE and RMSLE values will be calculated. It will be performed on Pycharm or Anaconda Spyder Ide's.
- Model Selection: Lowest error valued-model will be selected to further investigation.
- Parameter Tuning: After selection model specific parameters will be investigated and tuning operation will be conducted on Pycharm or Anaconda Spyder Ide's.
- Predictions: After recording final local metrics, predictions will be made on host's unlabeled test data and it will be submitted to competition's leaderboard.

8- Desired Outcome

As a self-learner newbie on DS and ML with this project I want to have a well written and executed project code to finish my nanodegree. Also, I want to earn a respectful place competition's leaderboard against fellow DS and ML professionals and enthusiast.