# BUILDING ENERGY FORECAST

SPRINGBOARD DATA SCIENCE CAPSTONE PROJECT #2

PREPARED BY MIKE (XIANGNAN) SHI

MENTOR: BLAKE ARENSDORF

# OUTLINE

- Project overview
- Problem statement
- Data wrangling and exploratory data analysis
- Modeling
- Summary
- Future research

# PROJECT OVERVIEW

- Goal: Develop a model to forecast energy consumption in a building.
- Tool: Python, Jupyter Notebook
- Workflow:
  - Define the problem
  - Collect, clean and explore the data
  - Create and train models based on training set
  - Validate the models using test set and evaluate the performance

# PROBLEM STATEMENT

# PROBLEM STATEMENT

- Context:
  - Given a building's daily energy consumption in the first three quarters, what is the daily energy usage forecast in the fourth quarter?
- Stakeholders:
  - Building owners, building managers and operators, financial and accounting department
- Scope of solution
  - Build models that can take historical energy consumption as input and forecast future consumption
  - Try different models including time series and random forest
  - Compare the performance of different models and determine a model that best fits in the goal.
- Data:
  - Source: https://www.kaggle.com/c/ashrae-energy-prediction/data

# DATA WRANGLING AND EXPLORATORY DATA ANALYSIS (EDA)

# DATA OVERVIEW

- Data source: https://www.kaggle.com/c/ashrae-energy-prediction/data
- Data time span: 2016/1/1 – 2016/12/31
- Target: electricity usage (meter_reading)
- Features used in the study: Timestamp, air temperature, dew point temperature
- Data relationships:

**Energy_data**

- Buliding_id
- Meter
- Timestamp
- Meter_reading

**Building_metadata**

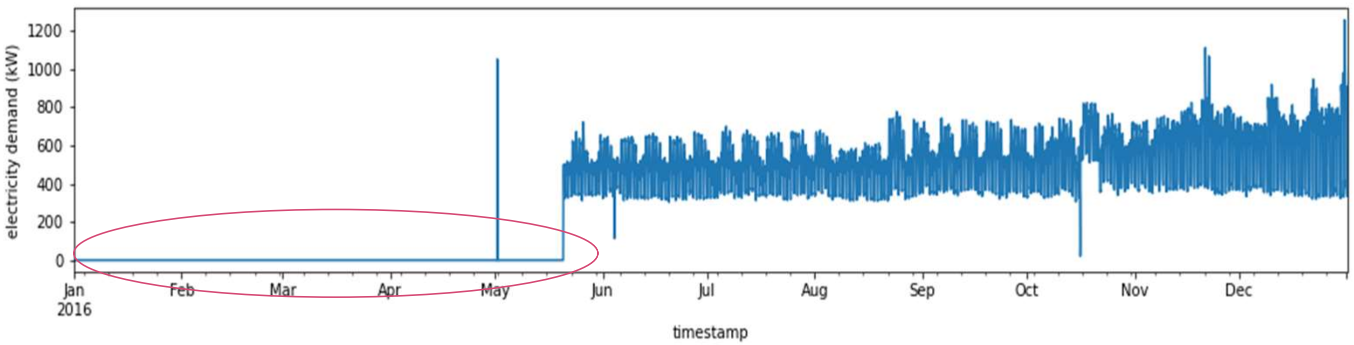
- Site_id
- building_id
- Primary_use
- Square_feet
- …

**Weather**

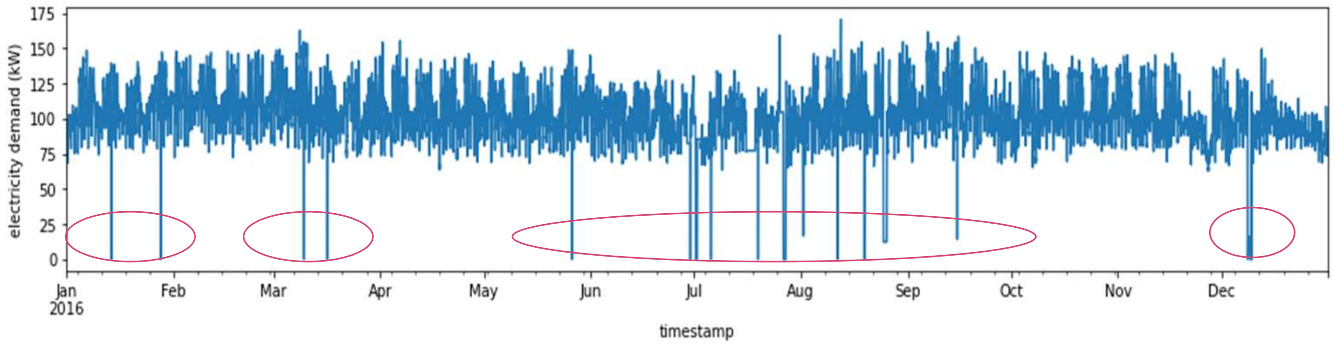
- Site_id
- Timestamp
- Air_temperature
- Dew_point_temperature
- …

# DATA ERRORS AND CLEANING

- Types of data errors
  - Zero readings
    - Extended periods of zero readings are considered as errors as building electricity rarely drops to zero. Those values are dropped out.

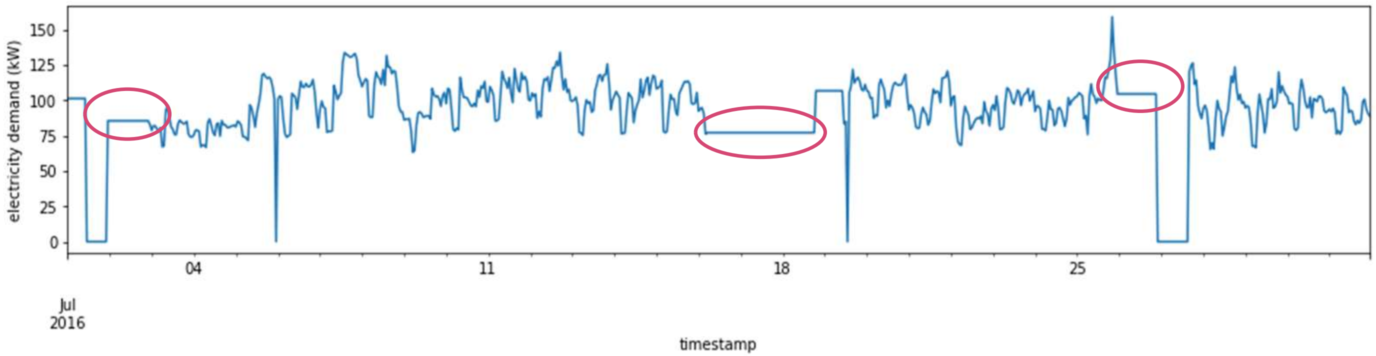



  - Anomalies
    - Abnormal electricity swings are considered sensor issues and dropped out as well.



# DATA ERRORS AND CLEANING

- Types of data errors
  - Frozen readings
    - Frozen readings are considered sensor failures and are dropped out as well.

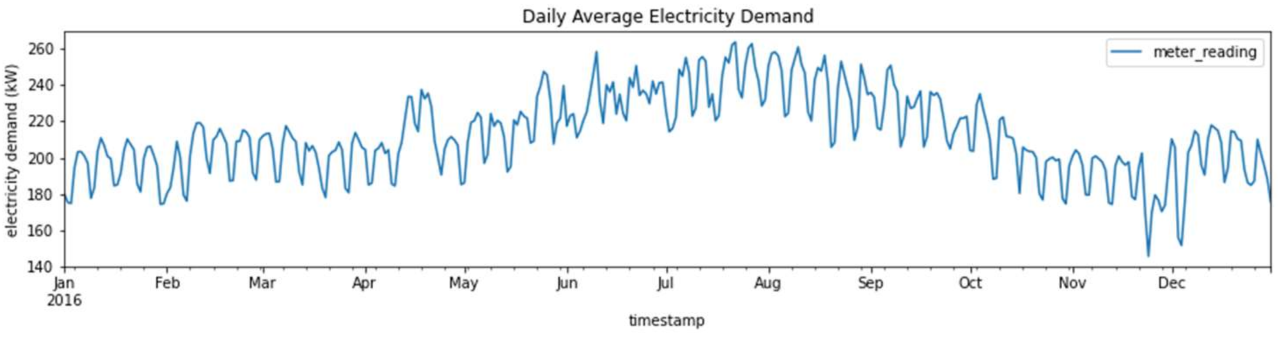



```
elec[elec.building_id==1287]['2016-7-17']
```

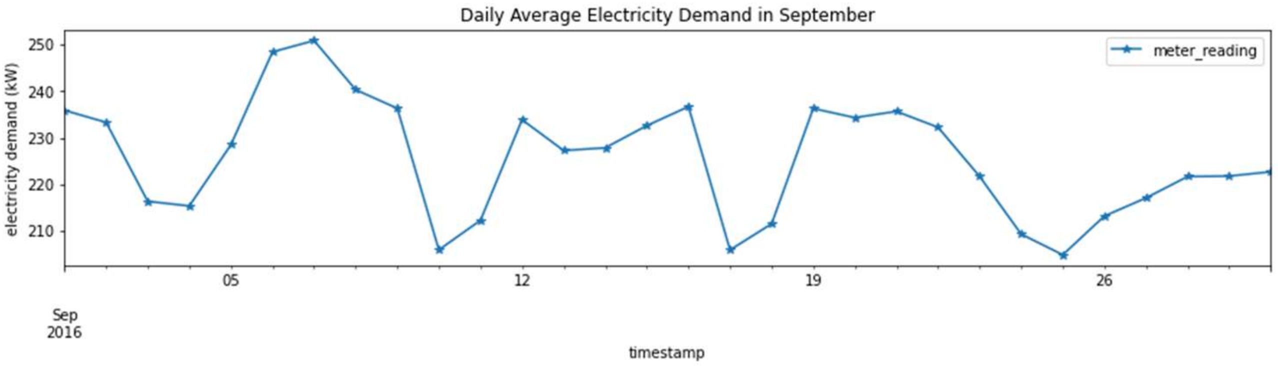
| | building_id | meter_reading |
|---|---|---|
| **timestamp** | | |
| **2016-07-17 00:00:00** | 1287 | 77.1729 |
| **2016-07-17 01:00:00** | 1287 | 77.1729 |
| **2016-07-17 02:00:00** | 1287 | 77.1729 |
| **2016-07-17 03:00:00** | 1287 | 77.1729 |
| **2016-07-17 04:00:00** | 1287 | 77.1729 |
| **2016-07-17 05:00:00** | 1287 | 77.1729 |
| **2016-07-17 06:00:00** | 1287 | 77.1729 |
| **2016-07-17 07:00:00** | 1287 | 77.1729 |
| **2016-07-17 08:00:00** | 1287 | 77.1729 |
| **2016-07-17 09:00:00** | 1287 | 77.1729 |
| **2016-07-17 10:00:00** | 1287 | 77.1729 |

# EDA

- The electricity usage appears higher in summer. This is likely due to more AC usage during that time.

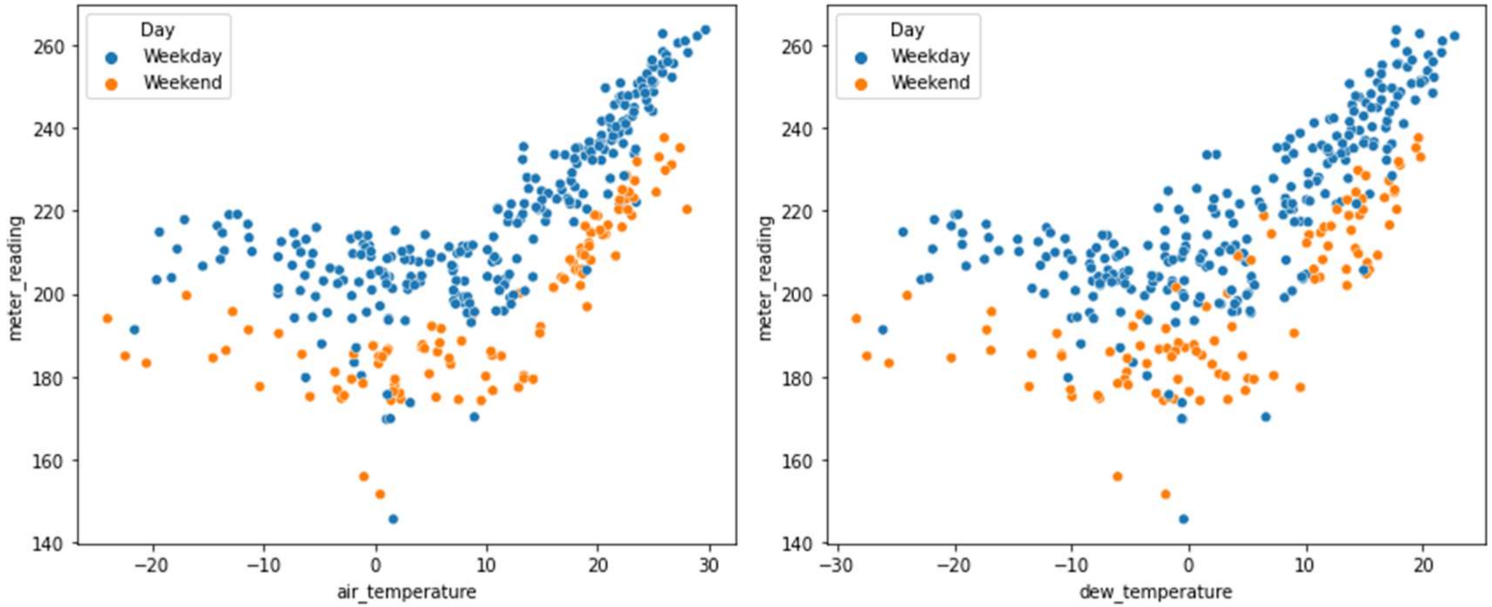



- There appears to be a regular fluctuation on a weekly basis. The electricity usage tends drop over the weekend and come back on weekdays.



# EDA

- When the air temperature and dew point temperature are above a threshold, the electricity usage starts to increase in correlation with those temperatures.



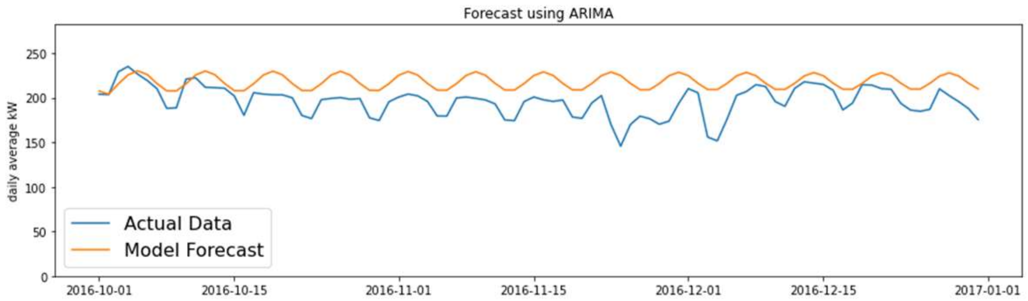
# MODELING

# MODELING

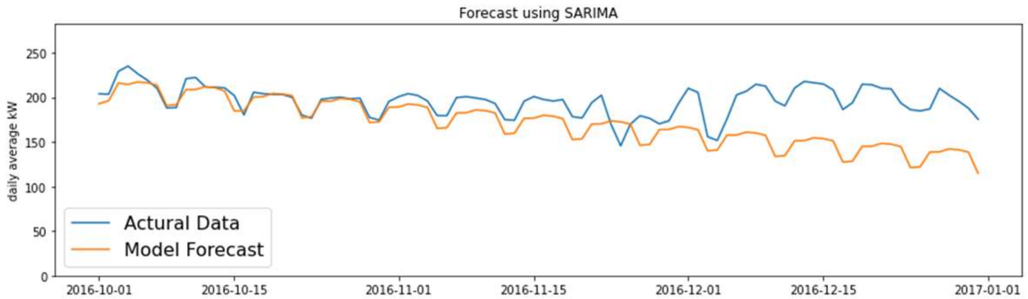
- Training / Test data split
  - Training: Jan – Sep
  - Test: Oct – Dec
- Models
  - Time Series
    - ARIMA
    - SARIMA
    - SARIMA with rolling forecast
    - SARIMAX with rolling forecast
  - Random forest
- Metrics
  - R squared
  - Mean Absolute Error (MAE)

# ARIMA AND SARIMA

- ARIMA
  - Optimal order is (2,1,5) based on auto arima search.
  - $R^2$: -10.95; MAE: 22.73
  - It's able to simulate the fluctuation patten, but the shape isn't quite right.

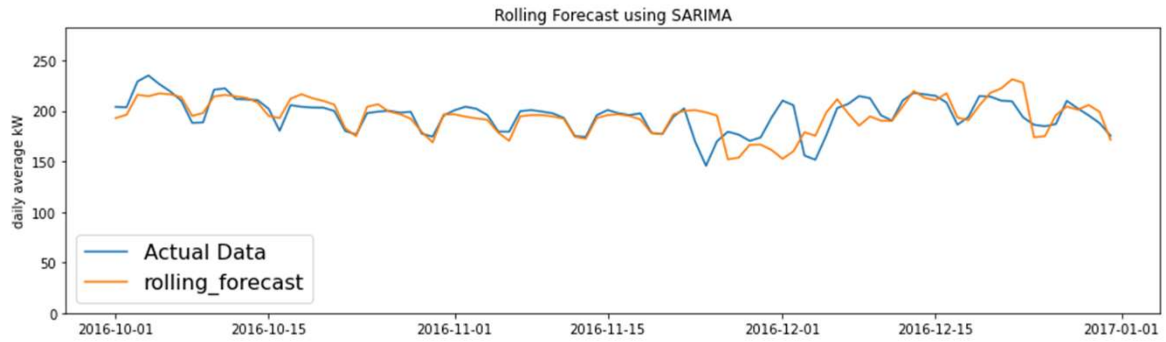



- SARIMA
  - Added seasonality into the model
  - Optimal order is (5,1,0)(2,1,0)[7] based on auto arima search.
  - $R^2$: -0.81; MAE: 25.5
  - The shape of forecast is much closer to actual data, but it mistakenly picks up a downward trend.

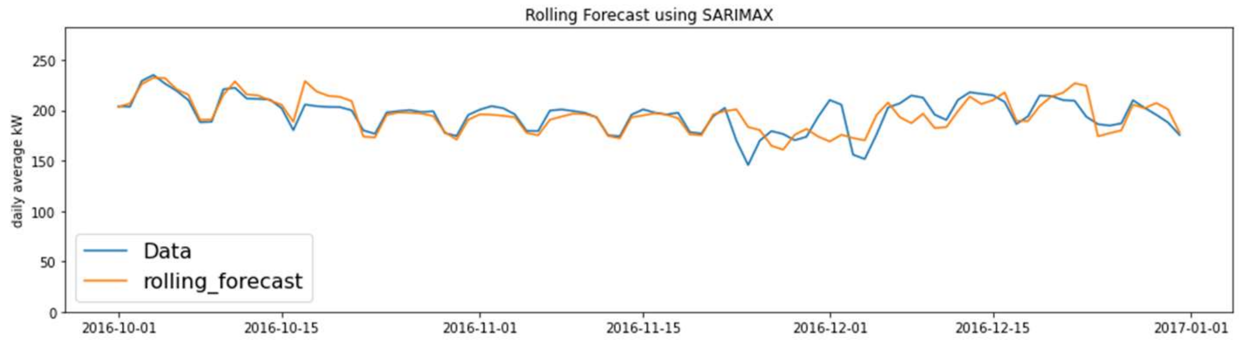


# ROLLING FORECAST USING SARIMA AND SARIMAX

- SARIMA with rolling forecast
  - Instead of a three month forecast at one time, weekly forecasts are done in a rolling manner.
  - $R^2$: 0.25; MAE: 9.84
  - It fixed the downward trend from last model.




- SARIMAX with rolling forecast
  - Added holiday schedule and temperatures into the model as exogenous inputs
  - $R^2$: 0.52; MAE: 8.23
  - The accuracy improved slightly



# RANDOM FOREST

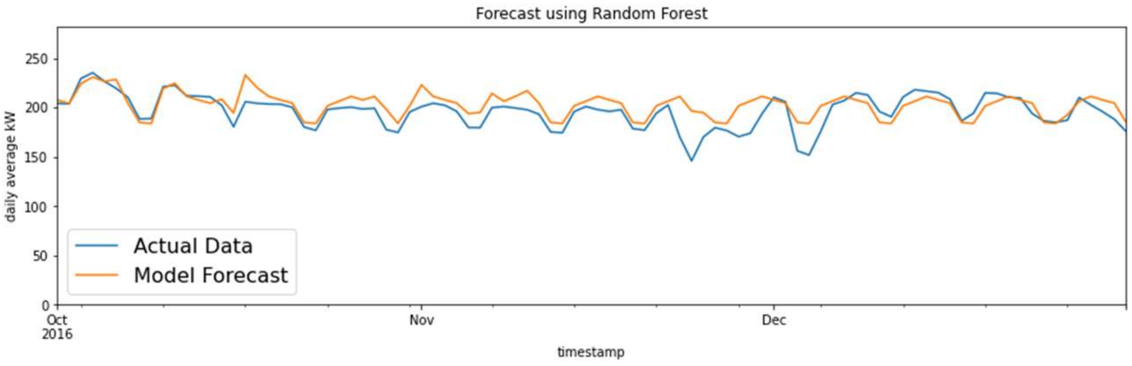
- Features:
  - Holiday
  - Day of week
  - Air temperature and dew point temperature

Training data set

| timestamp | holiday | temp | dewpoint | weekday_2 | weekday_3 | weekday_4 | weekday_5 | weekday_6 | weekday_7 |
|---|---|---|---|---|---|---|---|---|---|
| 2016-01-01 | 1 | 12.0 | 8.0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 2016-01-02 | 0 | 12.0 | 8.0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 2016-01-03 | 0 | 12.0 | 8.0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 2016-01-04 | 0 | 12.0 | 8.0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2016-01-05 | 0 | 12.0 | 8.0 | 1 | 0 | 0 | 0 | 0 | 0 |

- Grid search
  - A grid search is performed to find an optimal number of trees to be used, which turns out to be 10.

- Result
  - R2: 0.39; MAE: 9.93



# SUMMARY

- In this project, ARIMA and SARIMA don't seem to perform well enough for a one-time long term forecast. The models tend to pick up a trend from prior steps and assume the trend will continue in the future. As a result, the forecast is very sensitive to the last few steps in the train data set.
- Time series models perform much better forecast if done in a rolling manner, which is more focused on short term forecast.
- Adding exogenous features, such as holiday schedule, weather data in this case can improve the accuracy of time series models.
- Random forecast can be an alternative solution to consider if a one-time long term type forecast is needed.

| Model | R Squared | MAE |
|---|---|---|
| Basic ARIMA | -10.95 | 22.73 |
| SARIMA | -0.81 | 25.5 |
| SARIMA with rolling forecast | 0.25 | 9.84 |
| SARIMAX with rolling forecast and exogenous inputs | 0.52 | 8.23 |
| Random Forest | 0.39 | 9.93 |

# FUTURE RESEARCH

Additional features outside the data set may be explored and integrated into the models, such as building occupancy.

Additional models such as neural networks can be explored as well.

If the training data set can expand to at least a whole year, it may make the model more robust.

# CONTACT

**LinkedIn**
https://www.linkedin.com/in/mike-shi-pe-cem-14029a26/

**GitHub**
https://github.com/stonewatertx

**Email**
stonewatertx@email.com