1. Who are you (mini-bio) and what do you do professionally?

I am an IT consultant with around 18 years of industry experience. For the last 10 years, I have been consulting with utility (electricity, gas and water) companies around the globe in USA, Canada and Australia focusing on energy trading, customer care and billing and home services. I am a self-confessed MOOC addict ☺ enrolling myself in more MOOCs than I can ever complete.

1. High level summary of your approach: what did you do and why?

The solution is divided into 2 major parts

* Imputation of the data
* Model Building

**Imputation of the data**

Before running the model, imputation of the data was done for each of the site and penguin type combinations. The imputations were done in the following order

* Stine in case of R models and Linear in case of Python models
* Last Observation Carried Forward in case of R models only
* Next Observation Carried Backward in case of R and Python models
* Replace by Zero in case of R models and Python models

**Model Building**

Built 5 models for each of the site and penguin type combination. Therefore for each of the 648 combinations of site and penguin type, the following models were built

* XGBoost in Python
* RandomForest in Python
* ARIMA in R
* ETS in R
* Prophet in R

An average for all of these models was created for every 648 combinations.

1. What are some other things you tried that didn’t necessarily make it into the final workflow (quick overview)?

I used the Neural Network model but it did not improve the accuracy

1. Did you use any tools for data preparation or exploratory data analysis that aren’t listed in your code submission?

All the code is in the submission

1. How did you evaluate performance of the model other than the provided metric, if at all?

The provided metric was used to evaluate the performance of the model

1. Were there other fields or features you felt would have been very helpful to have? That is, what might the organization want to collect in the future that was not in the given data?

No suggestions in this regard.

1. Anything we should watch out for or be aware of in using your model (e.g. code quirks, memory requirements, numerical stability issues, etc.)?

There should not be quirky issues. I have provided the detailed instructions on how to install the software and the packages required for running the code.

The program uses R and Python. The Python code uses

* XGBoost. Install instructions of XGBoost with Anaconda on Windows is provided.
* randomForest

The R code uses

* Arima
* ETS
* Prophet

1. Do you have any useful charts, graphs, or visualizations from the process?

No useful charts, graphs or visualizations. Most of the data visualizations was throwaway. However a presentation ModelPresentation.pdf is being attached which explains the model in slides.

1. If you were to continue working on this problem for the next year, what methods or techniques might you try in order to build on your work so far?

It would be interesting to see if we could apply **LSTM Neural Network for Time Series Prediction** for this problem.

1. Are you willing to be interviewed for a blog post?

I am willing to be interviewed for a blog post.