Kevala Data Science Team

Take Home Challenge

Introduction

This challenge is intended to accomplish a dual purpose. Kevala's hiring committee gets a sense for how well-suited each candidate is for the prospective role, and the candidate gets a taste of the type of problems that Kevala's Data Science Team solves. Both out of respect for your valuable time, and so we can compare all candidates fairly, please limit the time you spend to 3 hours or less.

Challenge

Your team is setting out to develop a model to forecast the daily peak electrical demand of a home or business. This exercise represents the first steps in the development of that model:

- Writing code to parse the training data
- Conducting exploratory analysis
- Thinking about next steps for research and development

You should first answer the "All Candidates" questions, and then either the Software Engineer questions or the Data Scientist questions. You don't need to complete the other set, but feel free to read the other set as well and imagine that you have a teammate working on those!

What you will be evaluated on

- Your code, written in python 3, works well and can be read easily
- You can demonstrate your results and explain your thinking clearly
- Your approach to problems is analytical and creative

Dataset

The data for the challenge originates from the UCI Machine Learning Repository. We have cleaned and reduced the dataset provided by UCI into a CSV. The resulting dataset contains about 3 years of hourly demand measurements for 150 electrical meters. These meters are represented by columns "MT_001" to "MT_150". The first column in the .csv file contains an index of timestamps shared by all meters. There are no missing values. Each value in the time series represents the electricity demand in units of kW.

The data can be found here: cleaned_hourly_2012_2014_150_sps.csv.zip.

Questions

All Candidates

1. Load the data and do some exploration. Come up with a few (two or more) simple questions about the data, write code to investigate and answer the questions, and briefly discuss what you observe in the results. Please show that you have explored the dataset but don't spend too much time on this question!
2. Implement a function that calculates the daily maximum demand [kW] given a time series of demand values. The input to the function should be an (hourly) time series (timestamps and values) and the output should be another time series with the maximum value of the input time series on each day in the original time series.

Software Engineer Candidates

When developing a new predictive model, it is good practice to start by building a very simple version along with some code to evaluate its performance on a test dataset. Having a baseline model and performance metric enables an apples-to-apples comparison when evaluating more sophisticated candidate models. The following questions are an exercise in such baseline model development.

3. Implement a function for a simple model to forecast maximum daily demand. This model function should take as input a time series of known demand and a future date within one week of the end of the time series. It should output a predicted maximum demand for the specified date. The prediction algorithm is up to you to determine, but remember that the point here is to quickly establish a better-than-random baseline, not to develop a model for production. An algorithm as simple as "just use the last known value" would be perfectly fine!
4. Implement a method to compute the root mean squared error (RMSE) of a series of predictions, given the known true values as well.
5. Use the provided data to establish a test dataset that enables comparing predictions against known values.
6. Finally, calculate the RMSE of your baseline model using your test dataset, for:
    a. Predictions made 1 day in the future
    b. Predictions made 7 days in the future

    Compare the RMSE values - discuss how the results are similar or different from the two time horizons and why.

Data Scientist Candidates

In the exercises below you will complete some tasks that would occur before attempting to implement and evaluate a forecasting model. In question 3 you will work on feature engineering

to attempt to find a relationship in the data that could be of use for our forecasting problem. In question 4 you will brainstorm more details of a potential solution to our forecasting problem.

3. Using the provided data, create a feature (model input) that could be useful in predicting maximum daily load in the future. **Without implementing** a model, create a data visualization that illustrates how your feature might be of use for the forecasting problem described in the introduction. This is an intentionally open ended question - we would like to see how you would begin to evaluate solutions to the forecasting problem. Your visualization should show some relationship in the data that would be of interest when developing a model. Include the visualization in your written report.
4. Describe (don't implement!) a model to predict daily maximum demand values 7 days in the future. The model does not need to be complicated or high performing, just a possible solution to evaluate.
    a. Are there any additional datasets you would want to acquire?
    b. What features (model inputs) would you use?
    c. What is one prediction algorithm you would try?
    d. What sources of error do you anticipate in your model?

What to submit

If you worked in a notebook, please include your written answers and your code in the same document. Please export your jupyter notebook as a .html file and submit the .html file.

If you did not work in a notebook, please submit the .py file (or files) used to complete your analysis. Please submit a .pdf file containing your written analysis and any tables or graphs.

Please include your name in the filenames, i.e. ada_lovelace_kevala_takehome.py