sparkcognition™

DATA SCIENCE CHALLENGE

Thank you for your interest in SparkCognition. The following DS Challenge allows our team to better understand your technical abilities and gives you some insight into the type of work SparkCognition's data scientists encounter daily.

You will have three (3) days to complete the exercise which is estimated to take approximately four-six (4-6) hours. Please aggregate your responses to a single notebook or file with associated comments and include your full name in the file name. Acceptable submission formats include .ipynb, .html, and .pdf.

Each section will be weighted equally in the evaluation. In addition, please note that you will be evaluated on your quality of coding, thought process, and clarity in communication.

Exercise 1: Weather Forecast

Please complete the following exercise, using comments in your code to explain your reasoning.

Your customer is a power trading organization. They want to build a weather forecast model to better understand their asset's power generation to better bid in day ahead market. As a Data Scientist, they want you to build hourly windspeed forecast with horizon of 48 hours for their asset located **(lat: 37.7457611, lon: -121.6670387).**

Datasets:

- spark_renewable_ds_forecast_challenge.csv

Instructions:

- There are multiple ways of solving this problem,
    - 1. Using external sources (like met data, historical weather station data) to predict windspeed at given location.
    - 2. Using Machine Learning and traditional timeseries approach.
    - Use any one of these methods to solve the challenge.
- It's very important to use error metric appropriate to the end user.
- Benchmark the results against baseline models like persistence.

Exercise 2: Anomaly Detection

Please complete the following exercise in the same file as the other exercises, using comments in your code to explain your reasoning.

An oil and gas company with several offshore platforms is experimenting with anomaly detection on one of its platforms. An analyst has provided you with sample data for the pilot platform (anomaly_detection.csv).

The data is a time series dataset, consisting of average daily readings from 5 sensors between 01/01/2016 and 12/30/2016. You can assume that data preprocessing (filling missing values, scaling, etc.) has been handled by the analyst. You can also assume that the daily readings are independent and identically distributed.

The analyst has reviewed operation notes for the first 9 months (01/01/2016 to 09/30/2016) and identified that there were issues on the platform between 02/14/2016 and 02/21/2016. The analyst did not have time to review the final 3 months of data.

Use the first 9 months of data (01/01/2016 to 09/30/2016) to develop an anomaly detection model and test it on the final 3 months of data (10/01/2016 to 12/30/2016). How many anomalous periods were identified in the test period between 10/01/2016 and 12/30/2016? Only report anomalies that last longer than 2 days. Additionally, if an anomaly lasts for longer than 14 days, we consider that behavior to be the new normal, and we do not report it.

Exercise 3: Model Deployment

Please complete the following exercise in the same file as the other exercises, using comments in your code to explain your reasoning.

Customer wants to deploy the model built during exercise 2 in production. As a data scientist your job is to package the model in an API which takes given data (anomaly_detection.csv) as an input and gives prediction of alert prediction time range.

Bonus challenge: Packaging the model and API inside the docker container.