RMIT Vietnam University
School of Science and Technology

EEET2574 | Big Data for Engineering

Assignment 2: MongoDB and Spark

This assignment is worth 25% of your overall mark.

Introduction

In this assignment, you work with MongoDB, Databricks or Spark and MLflow to build a full data pipeline. This assignment is intended to give you practical experience with a simple big data stack and data pipeline.

The "Big Data for Engineering" Canvas contains further announcements and a discussion board for this assignment. Please be sure to check these on a regular basis - it is your responsibility to stay informed with regards to any announcements or changes.

Plagiarism

RMIT University takes plagiarism very seriously. All assignments will be checked with plagiarism-detection software; any student found to have plagiarised will be subject to disciplinary action as described in the course guide. Plagiarism includes submitting code that is not your own or submitting text that is not your own. Allowing others to copy your work is also plagiarism. All plagiarism will be penalised; there are no exceptions and no excuses. More information on Academic Integrity is available at https://www.rmit.edu.vn/students/my-studies/assessment-and-exams/academic-integrity

General Requirements

This section contains information about the general requirements that your assignment must meet. Please read all requirements carefully before you start.

Data

The dataset you will be working with in this assignment is a smaller version of this Kaggle dataset: https://www.kaggle.com/lucabasa/dutch-energy. Particularly, we only use the files of 3 different companies with data from 2018-2020 only. Please download the dataset from Canvas, and read the description of the dataset from Kaggle.

Important: The label column will be the "annual_consume". This is a regression problem.

Every file is from a network administrator from a specific year. The columns in each file are:

- net_manager: code of the regional network manager
- purchase_area: code of the area where the energy is purchased
- street: Name of the street
- zipcode_from and zipcode_to: 2 columns for the range of zipcodes covered, 4 numbers and 2 letters

- city: Name of the city
- num_connections: Number of connections in the range of zipcodes
- delivery_perc: percentage of the net consumption of electricity or gas. The lower, the more energy was given back to the grid (for example if you have solar panels)
- perc_of_active_connections: Percentage of active connections in the zipcode range
- type_of_connection: principal type of connection in the zipcode range. For electricity is # fuses X # ampère. For gas is G4, G6, G10, G16, G25
- type_conn_perc: percentage of presence of the principal type of connection in the zipcode range
- annual_consume: Annual consume. Kwh for electricity, m3 for gas
- annual_consume_lowtarif_perc: Percentage of consume during the low tarif hours. From 10 p.m. to 7 a.m. and during weekends.
- smartmeter_perc: percentage of smartmeters in the zipcode ranges

Task 1: MongoDB (18 pts)

Create a new database on MongoDB Atlas Cluster and load these datasets in it as collection(s) using PyMongo. You are free to decide on the number of collections and the schema.

Answer these questions:
Q1: How many collections do you have? Why?

Please include screenshot(s) of your collection.

Task 2: Data ingestion and data cleaning/transformation (18 pts)

Load data from MongoDB to Databricks using either MongoDB Spark Connector or PyMongo. You must provide the working uri (with user and password so I can run it and get the data). You can change the password after the course.

You should perform suitable data cleaning/transformation steps (missing values, impossible values, scaling, encoding, etc). Please put this in the same data pipeline with Task 3.

Answer these questions:
Q2-A: What are the chosen data cleaning steps? Why?
Q2-B: What are the chosen data transformation steps? Why?

Task 3: Model training and tracking with data pipeline and MLflow (45 pts)

You must use at least 2 different algorithms and 3 different parameter settings for each algorithm. All steps must be in the pipeline.

You must use the 2018 and 2019 as train data, and 2020 as test data. Please perform the correct

concatenation steps if needed. You also must use 3 evaluation metrics: MAE, R2, and RMSE.

All experiment info must be logged in MLflow, including algorithm name, parameter setting and 3 evaluation metrics.

Answer these questions:
Q3-A: What is/are your final model(s) based on the evaluation metrics?
Q3-B: Did you build one model for both electricity and gas or separate models? Why?
Q3-C: Should we build a separate model for each company or not? Why?

Note: Please include screenshots of MLflow UI showing all the experiment tracking.

Task 4: Visualisation (9 pts)

You can either use aggregation pipeline in MongoDB or PyMongo, or Spark on Databricks to transform data and put it in a new collection on MongoDB for visualisation on MongoDB Charts. If you think the current data is good enough, you can skip the data transformation step.

Create a visualisation dashboard on MongoDB Charts with at least 2 plots of your choice. Please include the public link to your dashboard on your submission (make sure it's public).

Code Readability and Documentation (10 pts)

You should organize your code structure and modularity for readability, with code documentation and explanatory comments.

What to Submit, When, and How

Assignments submitted after the due time will be subject to standard late submission penalties.

Please export the whole **project folder** for submission as zip file, name it after your student id ("s1234567.zip"). You must provide me with clear instructions on how/the order to run your notebook(s) (if you have multiple notebooks). If you use any custom packages, you must include instructions in the notebook for me to install it.

All the answers to questions (use the correct question code when answering), images, screenshots and the link to MongoDB Charts must be included in the notebook in the markdown cells. I won't grade a separate report word or pdf file.