Final Project
Data Science pipeline on-premises and on the cloud
Due on Sunday, 27 October (23:59)

Objective

In this assignment, you will be provided with a real-world dataset, and you must implement the whole pipeline of building a data science pipeline on-premises and on the cloud. This includes understanding the business problem, preparing data, exploring the data, performing feature engineering, and building and deploying models.

Introducing the business scenario

You work for a travel booking website aiming to enhance customer experience for delayed flights. The company plans to develop a service that will inform customers about the likelihood of a flight being delayed based on weather conditions before they book a flight to or from the busiest airports in the US.

You are tasked with tackling parts of this problem using machine learning (ML) to determine how likely a flight will be delayed based on available weather data. You have access to the dataset tracking the on-time performance of domestic flights operated by major air carriers. This data can be used to train an ML model to predict whether a flight will be delayed at the busiest airports.

About the dataset

The dataset provided includes scheduled and actual departure and arrival times reported by certified US air carriers, representing at least 1% of domestic scheduled passenger revenues. The Office of Airline Information, Bureau of Transportation Statistics (BTS), collected this data. It covers the date, time, origin, destination, airline, distance, and flight delay status from 2014 to 2018. The data is stored in 60 compressed files, each containing a CSV file with monthly flight details over five years (from 2014 to 2018). The data can be downloaded from this link: [**compressed data**].

Features of the dataset

The dataset used in this assignment was compiled by the Office of Airline Information, Bureau of Transportation Statistics (BTS), Airline On-Time Performance Data, available at the following link: [**dataset attributes**].

Tasks

Part A – Data Science on-premises (60 marks)

In this part, you are expected to:

- Understand the dataset and describe the business problem.
- Document an exploratory data analysis and, whenever possible, conclude the analysis.
- Employ popular graphical modules (matplotlib, seaborn or tableau) to answer questions.
- Implement machine learning techniques to predict whether the flights will be delayed.

You are given a Jupyter notebook named "onpremises.ipynb", which contains starter code and instructions to proceed with this part. You must answer the questions in this notebook and upload your responses for this part.

Part B – Data Science on-cloud (40 marks)

In this section, you are expected to:

- Apply your skills to perform the machine learning pipeline using Amazon SageMaker. *You may use one of the labs on the AWS Academy to do this part.*
- Compare the results of implementing the ML pipeline on-premises versus in the cloud.

You are given a Jupyter notebook called "oncloud.ipynb", which contains starter code and instructions for this section. You need to answer the questions in this notebook and upload your responses accordingly.

Part C – Face-to-Face Presentation and Q&A

- After submitting your project, you will deliver a face-to-face presentation to demonstrate your understanding of the data science pipeline.
- **Requirements**:
  Duration: 7–10 minutes presentation plus 3–5 minutes Q&A
  Format: You will present via your submitted code or prepare a PowerPoint slide summarising the main parts and key findings of your submission. This will be conducted individually, face-to-face.
  Focus:
  - Briefly introduce the business problem and dataset.
  - Summarise your on-premises and on-cloud workflows.
  - Present key findings, model performance, and comparisons.
  - Reflect on challenges and lessons learned.
- **Q&A:**
  After your presentation, you will answer a few questions about your methods, results, and design choices.

Marking:

Marks will be awarded for clarity, technical understanding, justification of approach, and engagement during discussion.

Deliverables

You are required to submit a compressed (e.g. ZIP) file to the Canvas website of the unit with the following files:

1- A Python Jupyter Notebook with the full code and narratives for Part A
2- A Python Jupyter Notebook with the code and narratives for Part B
3- [*Optional*] A PDF document with your reflection on the unit, highlighting what you liked and didn't.
4- Be prepared for a face-to-face presentation and Q&A session as described in Part C.

Good Luck 😊
Ibrahim Radwan