14-813/18-813 Course Project – Option 1

Deadline: November 28th, 11:59pm ET/8:59pm PT

Accept this assignment by accessing GitHub classroom via the following URL:
https://classroom.github.com/a/d0itqaYH

In this project, you will use FIFA Dataset available on Kaggle
https://www.kaggle.com/stefanoleone992/fifa-22-complete-player-dataset/

This dataset contains Soccer player statistics for 2015-2022. In this project, you will need all Men Player Data across all years to conduct the following tasks.

General Expectations

- Follow coding best practices with well-documented code.
- Add your dataset under **data folder** on GitHub repository
- You may choose 1 peer to work with on this project.
- If you choose to work with peer, write the name of your peer in the Canvas submission. If you fail to do so, your peer will not get the grade.
- We will not handle cases where students forget to submit the name of their peers.
- In this dataset, you will see incomplete data for Career Mode and Female Players. For this reason, don't include them in your analyses. The files of interest are named "players_15.csv", "players_16.csv", etc.

Task-I: Build and populate necessary tables (30% of course project grade)

- Ingest the data from all years (2015-2022) into one Postgres Database table.
- Add a new column for the year. Also, ensure every record can be uniquely identified in the database table.
- Identify constraints as needed and document them in your Readme.md file.
- Your tables should be created in schema with the name "fifa".
- In your ReadMe.md, add a description for the features in the dataset.

Task-II: Conduct analytics on your dataset (20% of course project grade)

Develop Python functions that run Spark to answer the following questions (given that x, and y) are user-entered parameters. Core analyses should be conducted via Spark and data should be ingested from Postgres database.

- What are the **X** clubs that have the highest number of players with contracts ending in 2023?
    - Use the players that were listed in the 2022 dataset only for answering this question.
- List the **Y** clubs with highest average number of players that are older than 27 years across all years (i.e. calculate the number of players older than 27 years old for each club in each dataset, calculate the averages and list the Y clubs with highest averages).
- What is the most frequent nation_position in the dataset for each year? (i.e. display the most frequent nation_position for 2015, 2016, etc.).

Task- III Machine Learning Modeling (30% of course project grade)

- Build a machine learning model that can predict the overall value for each player based on their skillsets.
    - Use proper feature engineering principles (including data cleaning and data engineering)
    - Build two versions: one in Spark and the other one in Tensorflow.
    - For each version, choose two different classifiers/regressors (you can use the same two choices for Spark and Tensorflow). For each classifier/regressor, identify a few tunable parameters for your model and use cross-validation to tune the parameters (using proper metric(s)). Then, run the best model obtained from cross-validation on the test data set and record the test accuracy.
    - In your ReadMe file, explain why you chose the classifiers/regressors and provide comments on the impact of the tunable parameters on the accuracy. Also, compare the selected models.

Task- IV Deploy your code to the Cloud: (10% of the course grade)

- Run a version of your code for the three tasks above on the cloud.
- In this version, you may skip the creation of the Database on the cloud (i.e. on the cloud version, you don't need to write data to table for simplicity). You may ingest the data from CSVs directly.
- If you run the PostgreSQL on the cloud: you will receive **10% extra-credit**.

Course Project Checkpoint is planned on **October 27th** and will constitute **10% of the course project grade**. In your course project checkpoint, submit **Task I** and **Task II**.

Submission Guidelines:

- You MUST use the GitHub classroom URL to create your repository. Post your GitHub repository's URL created via GitHub classroom to Canvas. Use the starter code that is provided above as the starter for your code.

- Your GitHub repository should have a ReadMe.md file that lists the “exact” steps on how to get this application working on a new machine (via Docker). I will follow the steps in your ReadMe file and if I can’t get it running on my machine, I will deduct considerable number of points from your project grade.
- **You should record a video demonstrating two elements:**
    1. Code Walkthrough while you are explaining your code changes.
    2. Demoing the running application while you are navigating through EVERY functionality that is working in your application. I will use this video to help assessing your grade. You may lose points for the functionalities that are not demonstrated in the demo.
- Your video size may be large to be uploaded to GitHub. You may use Box to upload the video and add the URL to your ReadMe.md file in your GitHub repository.
    1. Make sure that your video is publicly shared. Private videos won’t be visible to the instructor and TAs and therefore, your project grade will be impacted

Grading Notes:

- Late submissions on Canvas or GitHub: 0% grade (won't be graded)
- **Not submitting the GitHub video (for both code walkthrough and functionality demo): you will get up to 80% of the maximum grade.**
- Not providing clear details in the ReadMe file on how to run the application (or any variables that need to be updated/replaced): **you will get up to 90% of the maximum grade.**

Refer to Course Syllabus for planned course project checkpoints.