Applied MSc in Data Analytics

Applied MSc in Data Science & Artificial Intelligence

Applied MSc in Data Engineering & Artificial Intelligence

Course: Python Machine Learning Labs

Project: Develop an end-to-end Machine Learning Pipeline

Instructor: Hanna Abi Akl

Project Overview:

The aim of the course project is to ensure students are comfortable enough developing an end-to-end pipeline to answer a given problem or use case.

The project is a group project. Since this is a mixed course (DA/DS/DE), students are encouraged to form mixed groups to benefit from the competences of their teammates when working on different components of the pipeline. An ideal group is a group of 3 members: 1 Data Analyst, 1 Data Scientist and 1 Data Engineer. However, since this is not an ideal world, and group formation relies heavily on student distribution, groups of 2-4 are allowed (group diversity is still **highly encouraged**).

Additionally, groups will have the choice to work on 1 of 2 projects:

- Project 1: Book Rating Prediction Model
- Project 2: Free Project

Project 1 is a pre-defined project where students will look to answer a specific use case. The project is developed in the following sections. The advantage of working on Project 1 is that the project framework is pre-defined: use case and project parameters are already fixed and a dataset is provided.

However, students who wish to work on a project closer to their hearts have the alternative of choosing Project 2. The advantage of Project 2 is that students have full control on what they want to work on. The caveat is that students have an additional responsibility to define the project framework, i.e., the use case they are looking to solve and the dataset they wish to use. Students should also explain the choice and method of compilation of their data.

There is no bonus for picking Project 2 other than the satisfaction of subjective creative liberty. Both projects will be graded equally and you do not get a bonus for "being more risky" in selecting Project 2 over Project 1.

Project 1 Summary:

“There is no friend as loyal as a book.” - Ernest Hemingway

Nowadays with so many books available, it can be hard to select the best ones to read. The dataset provided is a curation of Goodreads books based on real user information. It can be used for many tasks like predicting a book’s rating or recommending new books.

Below is the information you have regarding the dataset attributes:

1) **bookID:** A unique identification number for each book.
2) **title:** The name under which the book was published.
3) **authors:** The names of the authors of the book. Multiple authors are delimited by “/”.
4) **average_rating:** The average rating of the book received in total.
5) **isbn:** Another unique number to identify the book, known as the International Standard Book Number.
6) **isbn13:** A 13-digit ISBN to identify the book, instead of the standard 11-digit ISBN.
7) **language_code:** Indicates the primary language of the book. For instance, “eng” is standard for English.
8) **num_pages:** The number of pages the book contains.
9) **ratings_count:** The total number of ratings the book received.
10) **text_reviews_count:** The total number of written text reviews the book received.
11) **publication_date:** The date the book was published.
12) **publisher:** The name of the book publisher.

Project 1 Objectives:

Using the provided dataset, you are asked to train a model that predicts a book’s rating. The project can be submitted as a Jupyter Notebook (at least) and should include exploratory analysis of the data, feature engineering and selection, model training and evaluation and finally, deployment.

You may use additional resources from those that are suggested in the “Project 1 Resources” section or others as you see fit (provided you can justify how they can serve your solution). You can even consult similar solutions from the Internet. **However, this comes with a big responsibility: any submission that is over-plagiarised or does not reflect personal work will not be accepted**.

Project 1 Resources:

Here are additional resources that may be helpful for the project. These resources are not mandatory to use but are meant to give you ideas on enriching the data or analysing the attributes in the dataset.

- Goodreads Datasets
- Recommending Goodreads Books using Data Mining

Project Evaluation:

Both projects will be evaluated using the following rubric. It contains the required items for a complete submission. The grading system is over 5 and the final grade will be transformed to a grade over 100.

- Data analysis (data processing, data cleaning, exploratory analysis, plots of relevant attributes) and feature selection (feature engineering, feature pruning, choice justification) **[1 point]**
- Model training (motivation for selected model, comparison of different models) and evaluation (evaluation metric, results interpretation) **[1 point]**
- Project report (short report explaining the approach and results) **[1 point]**
- Project reproducibility (requirements file with necessary packages, README file for running the project) **[1 point]**
- Project hosting and deployment (Github, Docker, AWS, Heroku or any other method) **[1 point]**

Project Timeline:

The deadline for the project is **1 March 2023**. Additionally, you are free to set a meeting with the instructor to discuss possible approaches, problems or other points pertaining to the project.