EXPLORE || DATA SCIENCE ACADEMY

Unsupervised Predict

Instructions

Your Mission: Find your Own Path to Deliver Value

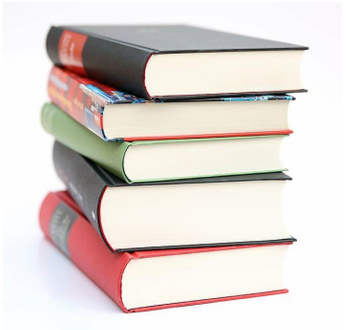
In **previous Sprints**, we were **given all the tools necessary to solve our problems**. We became comfortable with machine learning algorithms covered in the course that could help us make predictions and bring meaning to data.

In the Unsupervised Sprint Predict, we will focus on gaining skills required to **work independently on unseen problems**. This will be valuable experience as you prepare for the Internship phase of the program. It's time for you to become your own teacher!

Once again, in this Predict there are **three main tasks that we will be working on**:

Master various **recommender system algorithms** while competing in a Kaggle-hosted **Hackathon for Movie Reviews**

Deploy your best solutions as a **performant recommender engine** based in **Streamlit**



Perform your own research to solve big-data challenges and find more advanced algorithms.

Task 1) Apply Recommendation Algorithms within a Kaggle Hackathon

Within this sprint, we will once again make use of the fantastic Kaggle platform to apply knowledge we gain in building recommender system algorithms. Here you will compete to accurately predict unseen movie ratings gathered from thousands of users based on their historical preferences.

Unlike previous challenges, however, you will not be provided with an extensive set of tools and techniques to perform this task. Instead, you will be taught some basic concepts and will then be responsible to develop your own knowledge further. This is a vital skill, and will stand you in good stead as a professional Data Scientist one day navigating the uncertain and open-ended paths of the problems you are tasked with.

Task 1) User Movie Rating Prediction - Hackathon Summary

Use this link to enter the competition on Kaggle.

The summary below is taken directly from the competition page:

In today's technology driven world, recommender systems are socially and economically critical for ensuring that individuals can make appropriate choices surrounding the content they engage with on a daily basis. One application where this is especially true surrounds movie content recommendations; where intelligent algorithms can help viewers find great titles from tens of thousands of options.

With this context, EDSA is challenging you to construct a recommendation algorithm based on content or collaborative filtering, capable of accurately predicting how a user will rate a movie they have not yet viewed based on their historical preferences.

Providing an accurate and robust solution to this challenge has immense economic potential, with users of the system being exposed to content they would like to view or purchase - generating revenue and platform affinity.

Task 1) Instructions

1. Enter the EDSA Movie Recommendation Challenge hosted on Kaggle.

2. Register your group as a team for the competition.
   - Format your team name as follows: Team_<team number>_<campus>_<custom-text>
   - Use numerals for numbers, three-letter abbreviations for campus (DBN, CPT, JHB)
   - For example, Team One from Durban could be: “Team_1_DBN_#Recommended_winners”

3. **Create a notebook (kernel) on Kaggle** for you and your team to collaborate on. You may also use other computing platforms such as an AWS EC2 instance for larger compute tasks.

4. Develop, train and validate your recommender algorithm. **Use good practices to version control your experiments** on Comet.

5. **Ensure your notebook can produce a valid submission directly to the competition**. Submit this output to Kaggle to be placed on the Challenge Leaderboard (you can do this multiple times). This is a quick guide on how to read data into your kernel, and here you can learn how to output a file.

6. Forward your notebook link to your supervisor + ensure that your notebook is public at the end of the competition

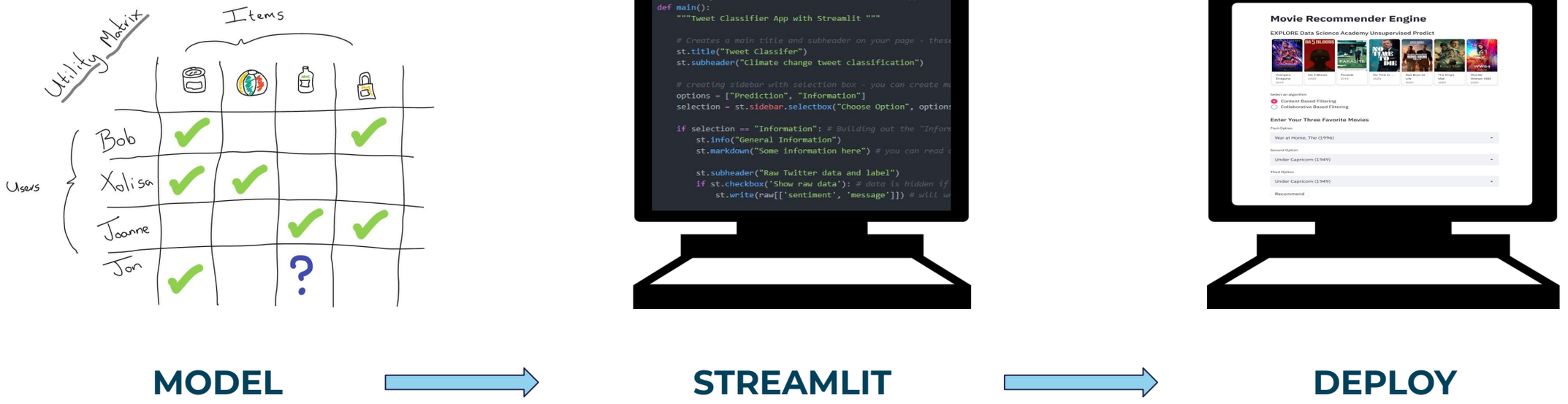
Task 1) Rules

- This project is a **group** project, therefore you are required to enter your submission as a team on Kaggle.
- You are free to share your code with your team members, however, **you are not allowed to share your code, solutions, or submissions with other individuals or teams**.
- Part of your Predict **mark will be based upon your score on the Leaderboard**. Aim to have an RMSE lower than 0.85 on the complete test set.
- The **leaderboard will officially close on 27 July 2020, at 17:00**. No submissions after this point will be accepted.
- You will be **required to prove how you obtained a given submission result**. Teams who cannot will receive a mark of 0 for the predict.

Task 2) Building an Online Recommender Engine

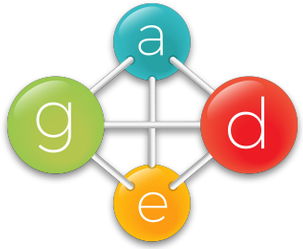
In the previous sprint we became familiar with **Streamlit** as a flexible webserver framework to deliver our solutions to the world.

In the **Unsupervised Predict**, we will make further use of this this capability; deploying our developed algorithms in task 1 as a movie recommender engine hosted on an AWS EC2 instance.



MODEL → **STREAMLIT** → **DEPLOY**

Task 2) Instructions



1. We have created a template repo on GitHub here to as a framework for your recommender engine.

2. In each team, one member needs to **Fork the repo**. As a team, you should then **clone the new repo**, and begin collaborating on its development.

3. Send the URL of your team's new GitHub repo to your supervisor.

4. Carefully **follow the instructions laid out within the repo** around it's modification. You will be tasked with modifying the algorithms it currently implements for improved ones you develop in task 1).

5. As before, the repo contains instructions on how to setup the app on an EC2 instance. You are required to follow these instructions to ensure that your engine runs continuously in the cloud.

6. **We will perform automated testing on your deployed app** on **27 July 2020, after 17:00**. Testing will assess both the suitability of the recommendations your engine is able to make, as well as the latency incurred to produce these suggestions.

Task 3) Communicate your Findings

As a Data Scientist, you need to **continually communicate** your work and findings to various audiences.

Within this Predict, you will be required to **explain your work to a company seeking solutions for their own online recommendation platform**.

You will do this in the following ways:

1. As you **develop your solution notebook for Task 1**, you will need to **ensure that your work is fully documented and is reproducible** by a technical individual.
    - **Have logical structure**, with an introduction, body and conclusion.
    - **Contain essential steps** within your model development process, such as an Exploratory Data Analysis (EDA), data preprocessing, modelling, performance evaluation, and algorithmic analysis.
    - Be supported with **appropriate visuals and metrics**.
    - Separate these sections and **explain your work using Markdown** cells.
    - Contain well written code which is sufficiently commented and **meets best coding practices**.

Task 3) Communicate your Findings

2. Give a **Video Conference Presentation** to your EDSA peers and Supervisors.

   - **Communicate insights to an executive stakeholder** on what trade-offs are afforded by implementing different algorithms in an online setting. Do this **using a slide deck** (or your Streamlit app).
   - Your presentation should **explain your approach, and findings** throughout the entire process, in language that can be understood by non technical people.
   - **Make extensive use of graphs and visuals** to give context to your offered solutions.
   - At the end of your presentation, a **panel of supervisors** and audience members will be given an opportunity to **ask technical and non-technical questions** related to your presented work.
   - Presentations to last **12-15 minutes, followed by 5 min questions**.
   - **Further details around the logistics of the presentation** will be given in due course via **Slack and Athena**.