Project Documentation

1. Our Team

Our team members are Regan, Riley and Jared.

We will first dive deeper into each of our strengths. All 3 of us are majoring in Computer Science, but Regan and Riley have a greater expertise in Machine Learning and Statistics while Jared is more well-versed with Artificial Intelligence and Cloud Services. Upon a deeper analysis on our individual expertise and experiences, we can conclude the following:

1. We have no member of the team which has an expertise in frontend web development
2. All of our members have had experience cleaning and handling data sourced from the web.
3. All of our members have had experience writing code in Python. Regan and Riley created prediction models for schoolwork and in their past internships and Jared has utilized AI to build recommendation systems utilizing Natural Language Processing Tools.

Second, we will take a closer look at our workloads, schedules and educational goals.

- Regan: Will be representing UofT Squash where he will be training 4 times a week with the varsity team. On top of that, Regan has packed his classes into 3 weekdays. He aims to develop further in the aspect of project management in large software companies and write simpler, clear, readable, testable and maintainable code.
- Riley: Has joined several clubs in UofT with the aims of meeting new friends that are of different backgrounds and cultures. Similar to Regan, Riley has packed his classes into 3 weekdays. He aims to refine his abilities as a software developer, and become more well-versed with the software development process in a large software company.
- Jared: Has several hobbies with the goals of exploring Canda during his exchange duration. He aims to learn more about the best practices of project management used in the industry to build good software.

2. Our Technology Stack

With consideration of the aforementioned details about the strengths, schedules and goals of each of our member, we will now provide a description of the chosen tech stack and toolchain that we intend to use to approach this project and how and why we arrive at that design.

A. Programming Language

Our software is written entirely in Python. Below are some of our considerations:

- Python is simple and readable
- Python has great libraries for data analysis and visualization like Pandas, Numpy, Seaborn and Plotly.
- Considered R as R has powerful tools for data analysis and statistical calculations.
- Scala was also considered as it is a very popular functional language, runs on JVM and is ideal for working with high volume datasets.
- We eventually settled on Python as our project is of a smaller scale – we're only creating a data visualization dashboard. Furthermore, Python is something all members of our team are familiar with. This allows for more time to be spent on developing the software instead of learning a new language.

B. Containerisation

Docker for containerisation

- Docker allows for dependencies to be written in code so no matter where it is running the software will run.
- Docker containers are process-isolated and don't require a hardware hypervisor hence it requires less resources than a virtual machine and runs faster.

The Dataset and Problem

Every year, thousands of movies are filmed and produced but not all of them breakeven. An even lesser number of movies make it as a “box-office hit”. In this project, we intend to create a machine learning model that predicts which gross revenue generated by movies based on various features of a movie such as its 1) budget, 2) choice of cast/directors, 3) genre of movie, 4) budget of movie to name a few.

Since there would be many reasons why a movie has high gross revenue, the software will also narrow down to a few selected features which are most important. This enables up and coming directors or filmmakers to understand their audience and what data might suggest about what areas to prioritise when attempting to make a “box-office hit”.

While the machine learning model lives on the back-end, a front end data visualization dashboard will allow for exploratory data visualization on the important features and how they relate to each other. Users are also able to input parameters on given features of the movie and have a predicted gross revenue returned.

The dataset was found on Kaggle and contains 28 columns, 5043 rows