Data Science Career Track - IMDB Image Classification Project Proposal

Problem Statement:
Provide a quick and accurate deep learning neural network image classification model for Family Friendly Productions, movie production company, to predict the movie genre based on IMDB poster image data before their upcoming movies hit the market. This is to give the Marketing team more focused insights to tune their respective marketing methods(Facebook and Instagram Advertising) to target their specific audience members so that Family Friendly Productions will be positive by the end of 2024 financial year.

Context:
Family Friendly Productions, a recent 2 year startup movie production company, saw a loss of profits in movie revenue in the past 2023 financial cycle. In strategizing for the next 2024 financial period, the CEO, CFO, and Marketing Director see this could be due to lingering feelings from the COVID-19 pandemic as well as potentially not capitalizing their marketing efforts quick enough to their avid movie goers and spreading the word on their upcoming movies. Their past strategy was relying heavily on just facebook or instagram advertising but the leadership team deduced that if they could somehow figure out the market genre of their movies prior to release they could capitalize on those specific audience members and be net positive for total revenue in 2024. The leadership team recognized that IMDB is one of the top websites the general public uses for movie watching and is wondering if they can leverage this data to help in producing the genre for their own upcoming movies. As a hired Data Scientist/Engineer, one pitched strategy is to design a deep learning neural network model to predict the movie genre based on the IMDB data movie poster. Having the categorized genre quickly recognized will aid in marketing efforts and help the Marketing team make focused data driven decisions rather than specifically relying on other (Facebook and Instagram) companies hopefully reaching their target audience.

Criteria for Success:
Criteria for success would be the deep learning neural network image classifier has an ROC AUC of above 50% meaning the classifier does better than random guessing in terms of predicting movie genre.

Scope of Solution Space:
Scope of this solution space or the deep learning neural network image classifier would be applied within 2 financial quarters for Family Friendly Productions and then revenue will be assessed based on previous quarters production and revenue margins for assessing model performance before permanence.

Constraints within Solution Space:
Foreseeable constraints include time to be able to implement this reduction plan as leadership wants to deploy a model for 2 financial quarters for the 2024 year. Another is that retrieving data with the IMDB API, the API only allows 250 search results at time. Also the movie constraints based on leadership decision, is to be within the G or PG category as the production company's focus are only movies of this category.

Data Science Career Track - IMDB Image Classification Project Proposal

Stakeholders:
CEO, CFO, Head of Marketing of Family Friendly Productions

DataSources:
Dataset:
Movie data retrieved from IMDB with their API GET requests and their respective poster image were downloaded and saved on a personal machine. IMDB movie dataset saved within .csv format. Image data saved within a .jpg format.
Link to Dataset: https://github.com/tpoozhikala/IMDB_Classification/tree/main/0_Datasets

Description of Dataset:
The dataset retrieved from the IMDB API provides data of different movies filtered to the G or PG rating and stemming from the 1990s to the past year of 2023. Data consists of IMDB title, id, runtime, image data, genre categorization, etc.

Data Science Career Track - IMDB Image Classification Project Proposal

Outline of Problem Solving Steps:

1. Problem Identification:
    a. Develop Project Proposal
    b. Identify Scope and Criteria for Success

2. Data Wrangling:
    a. Access IMDB API for get requests and design get request url
    b. Organize json data from IMDB API to Pandas DataFrame
    c. Determine and drop movies with 'None' values for image data
    d. Determine and drop movies if also no genre given
    e. Get genres of movies with image data and save as target

3. EDA:
    a. Determine class balance of dataset based on genre

4. Modeling:
    a. Read Poster image data
    b. Split image data to train/test (70/30 split)
    c. Design deep learning neural network with layers consisting of CNN, Pooling, FCN (with multi-class labels)
    d. Assess model effectiveness based on confusion matrix and ROC AUC.

5. Documentation:
    a. Create report documenting key insights and model effectiveness
    b. Create a slide deck for Family Friendly Productions Leadership team proving model effectiveness.
    c. Code for the project and steps will be stored in a Github Repository.