MOVIE RATING PREDICTION SYSTEM

By: Manuel Iglesias

Overview

Problem Statement:

- Most movie producers, directors, licensers, and streaming platforms and theaters struggle to know the public rating of a movie prior to its release since it has not been watched yet

Research Goal:

- Develop a system to predict movie ratings that utilizes data storage/cataloging and ETL techniques to identify connections between movies to cater to different user preferences

Users:

- ***Movie Creators*** + ***Licensers*** who need to know the popularity of the movies to know whether they need to have licensing agreements in place
- ***Writers/Actors*** who need to know whether the movie will gain them money or recognition
- ***Theaters/Streaming Services*** who need to know which movies to stream/show and how much platform/theatre space to allocate to them

Benefits:

- Ability to predict movies before spending money on them
- Increases chances of niche movie ideas to be chosen
- Saves money for Movie creators, Theaters, Actors, etc

Data Statistics

Remaining Missing Values: None

Data Shape BEFORE Cleaning: 722,462 rows x 20 columns = 14,449,240

Data Shape AFTER Cleaning: 6,070 rows x 7 columns = 42,490

Data :

Numeric: Budget, Revenue, Runtime, etc

Text: Genre, Title, Overview, Production Company

Dropped Columns: 'id', 'original_language', 'release_date', 'status', 'vote_count', 'credits', 'poster_path', 'backdrop_path', 'recommendations'

Key Data: 'Features' column consists of movie titles, genres, overviews, production companies, taglines, and keywords

Column	Unclean Data	Clean Data
id	0	0
title	4	0
genres	210488	0
original_language	0	0
overview	118341	0
popularity	0	0
production_companies	385187	0
release_date	51847	0
budget	0	0
revenue	0	0
runtime	34363	0
status	0	0
tagline	614121	0
vote_average	0	0
vote_count	0	0
credits	224853	0
keywords	511997	0
poster_path	184729	0
backdrop_path	499531	0
recommendations	687442	0

Data Description

Data Title: Movies Daily Update Dataset

Data Source: Kaggle - https://www.kaggle.com/datasets/akshaypawar7/millions-of-movies

Dataset Description: 700,000 movies with information on cast, crew, plot keywords, budget, revenue, posters, release dates, languages, production companies, countries, TMDB vote counts, vote averages, reviews, recommendations

Data Size: 350.31 MB

Data Time Period: 4/30/1990 - 2/14/2024

Data Types:

- Strings - (Movie Titles, Genre, Movie Overview), Numeric (budget, revenue, runtime), etc

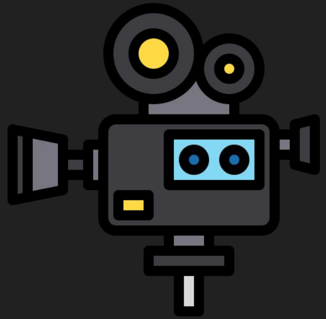
Data Size: 575,340 unique values

Data Storage - Google Cloud Computing

Used Google Cloud Computing Services:

- Created a new project
- Created VM instance
- Updated firewall settings
- Installed Anaconda
- Set up VM Server
- Created bucket
- Uploaded movies.csv file to bucket
- Opened and performed analysis in Jupyter Notebook

Big Data Engineering (Jupyter)



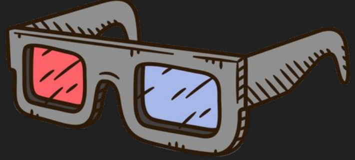
- Jupyter notebook
  - Access Python packages & load movies.csv file
- Perform ETL to clean/transform and reshape our raw data into a format suitable for analysis
- Handled missing values in the dataset
  - Dropped rows containing missing values
- Transformation:
  - Separating text with commas and turning them into strings when necessary
  - Standardized text, tokenize, removed stop words, apply stemming and lemmatization
  - Normalization of numerical values with sklearn MinMaxScaler

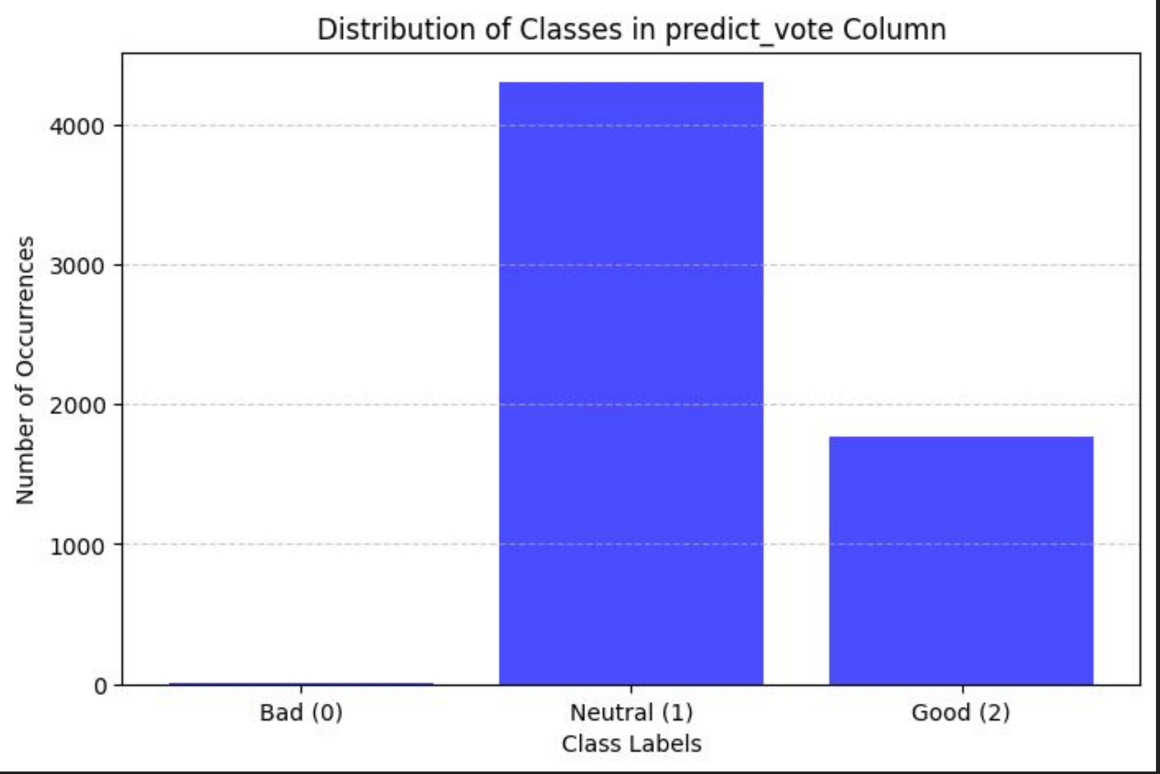
Prediction System

- To create the predict_vote column:
  - vote_average ratings from the original data set are categorized into three classes:
    - Bad (≤3)
    - Neutral (>3 and ≤7)
    - Good (>7).
- The categorization is done using the categorize_score function applied to the vote_average column of the dataset.
  - Each movie's rating is evaluated against these thresholds to determine its category.
- For example: a movie with a vote_average of 2.5 is categorized as 0 (Bad), a vote_average of 5.5 is categorized as 1 (Neutral), and a vote_average of 8.0 is categorized as 2 (Good).
- Predict vote totals: 6,070

	title	genres	overview	popularity	production_companies	budget	revenue	runtime	tagline	vote_average	keywords	features	predict_vote
0	Meg 2: The Trench	Action, Science Fiction, Horror	An exploratory dive into the deepest depths of...	8763.998	Apelles Entertainment-Warner Bros. Pictures-di...	129000000.0	352056482.0	116.0	Back for seconds.	7.079	based on novel or book-sequel-kaiju	meg 2 trench action scienc fiction horror expl...	2
1	The Pope's Exorcist	Horror, Mystery, Thriller	Father Gabriele Amorth Chief Exorcist of the V...	5953.227	Screen Gems-2.0 Entertainment-Jesus & Mary-Wor...	18000000.0	65675816.0	103.0	Inspired by the actual files of Father Gabriel...	7.433	spain-rome italy-vatican-pope-pig-possession-c...	pope exorcist horror mysteri thriller father g...	2
2	Transformers: Rise of the Beasts	Action, Adventure, Science Fiction	When a new threat capable of destroying the en...	5409.104	Skydance-Paramount-di Bonaventura Pictures-Bay...	200000000.0	407045464.0	127.0	Unite or fall.	7.340	peru-alien-end of the world-based on cartoon-b...	transform rise beast action adventur scienc fi...	2
3	Dune: Part Two	Science Fiction, Adventure	Follow the mythic journey of Paul Atreides as ...	4742.163	Legendary Pictures	190000000.0	683813734.0	167.0	Long live the fighters.	8.300	epic-based on novel or book-fight-sandstorm-sa...	dune part two scienc fiction adventur follow m...	2
4	Ant-Man and the Wasp: Quantumania	Action, Adventure, Science Fiction	Super-Hero partners Scott Lang and Hope van Dy...	4425.387	Marvel Studios-Kevin Feige Productions	200000000.0	475766228.0	125.0	Witness the beginning of a new dynasty.	6.507	hero-ant-sequel-superhero-based on comic-famil...	ant man wasp quantumania action adventur scien...	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...
122516	A Rainy Day in New York	Comedy, Romance	Two young people arrive in New York to spend a...	1.577	Gravier Productions-FilmNation Entertainment-P...	25000000.0	23800000.0	92.0	Love In Spring.	6.500	new york city	raini day new york comedi romanc two young peo...	1
134928	X-Men	Adventure, Action, Science Fiction	Two mutants Rogue and Wolverine come to a priv...	1.423	The Donners' Company-Bad Hat Harry Productions...	75000000.0	296339527.0	104.0	Trust a few. Fear the rest.	6.992	mutant-superhero-based on comic-superhuman	x men adventur action scienc fiction two mutan...	1

Data Analysis & Insights

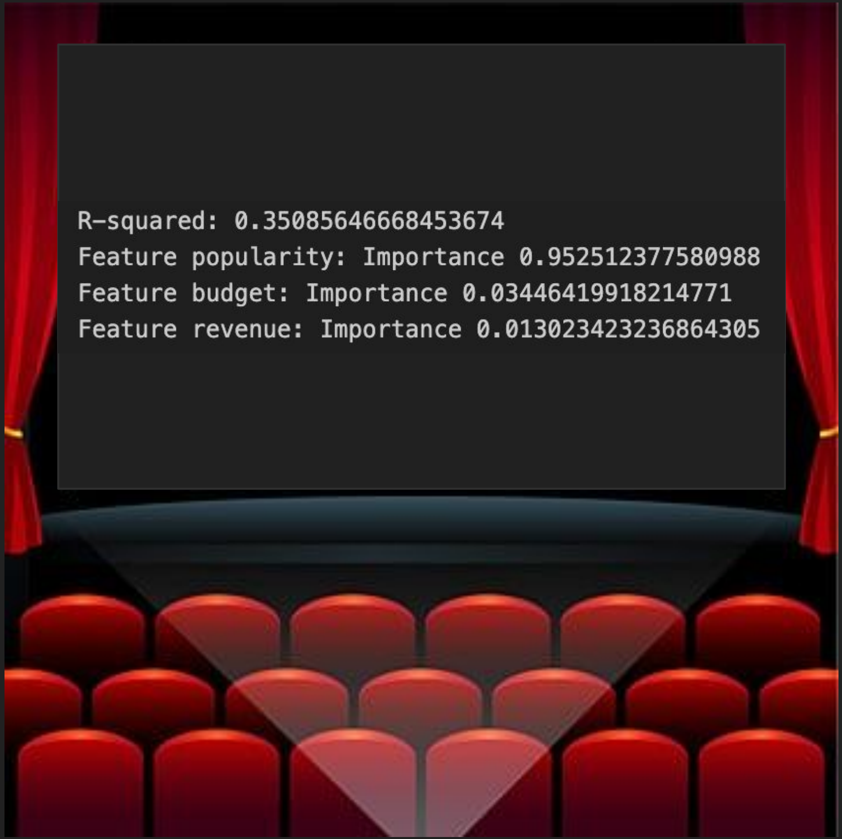




- **Prediction Classes Occurrences:**
  - ***Bad***: 3
  - **Neutral:** 4299
  - **Good:** 1768

- **Class Imbalance:**
  - Too many ratings fall into the "neutral" category and too few are "bad"
  - Either the threshold for "bad" is too low or the model is biased towards "neutral"
  - The absence of "bad" ratings prevents the model from learning how to accurately predict ratings

Random Forest Analysis & Insights



- The R-squared value of 0.35 indicates that the model explains approximately 35% of the variance in the predict_vote.
- Popularity is the most important feature for predicting predict_vote, with an importance score of 0.9525.
- Budget and revenue have relatively low importance scores, suggesting that they have less impact on the predicted predict_vote values compared to popularity.

Next Steps

- **Refine Categorization Thresholds**
  - Instead of manually setting thresholds, use statistical methods like k-means to determine natural breaks in our data distribution to set the cutoffs
- **Experiment with other Modeling strategies**
  - Try using neural network algorithms
  - Use K-fold cross-validation to ensure that each fold of our training and validation sets respects class distribution
- **Feature Engineering**
  - Use NLP techniques to extract features from movie reviews or descriptions (Sentiment analysis, topic modeling, or TF-IDF techniques)