PROJECT REPORT AND PRESENTATION

Project Title: Movie Analysis

Team Members: Jainam Chhadwa
Krupa Shah
Hitaishi Joshi

ORIGINAL WORK STATEMENT

We the undersigned certify that the actual composition of this proposal was done by us and is original work.

	Typed Name	**Signature**
	Jainam Chhadwa	Jainam Nikhil Chhadwa
	Krupa Shah	Krupa Nilesh Shah
	Hitaishi Joshi	Hitaishi Vijay Joshi

Executive Summary:

Our research focused on leveraging PySpark for a comprehensive analysis of a movie's dataset, aiming to predict whether a movie is considered good or bad on IMDB based on its IMDB score. The dataset encompasses various features such as director information, movie duration, cast details, budget, and more. The target variable, IMDB score high, is dichotomized, with movies having a score greater than 7 considered as good.

Our analysis involved both supervised and unsupervised machine learning techniques implemented through PySpark. In the supervised phase, we built predictive models to classify movies as good or bad based on the provided features. The unsupervised phase focused on exploring patterns and relationships within the data without explicit target labels. We used the clustering technique to get the clusters of movies which are similar in nature.

This study serves to provide insights into the factors influencing movie ratings, aiding filmmakers, producers, and industry professionals in understanding the dynamics that contribute to a movie's success on platforms like IMDB. The significance lies in its potential to guide decision-making processes in the film industry, leading to more informed choices in production and marketing strategies. In essence, our findings showcased the feasibility of predicting a movie's performance on platforms like IMDB with a 72% accuracy. This predictive capability empowers filmmakers with valuable foresight, enabling them to gauge the potential reception of their creations even before their debut. Additionally, our investigation illuminated the efficacy of clustering akin movies to personalize recommendations, enhancing user engagement and satisfaction on streaming platforms. Furthermore, our analysis uncovered nuanced relationships between actors, directors, and movies, underscoring the pivotal role of collaboration in a movie's trajectory toward success. Through our rigorous analysis, we have endeavored to discern the genre combinations that yield the highest IMDB ratings for movies, recognizing that a movie often encompasses multiple genres. We gain valuable insights that the combination of animation and drama genre has the highest average IMDB rating of 7.3.

Data Description:

The dataset was obtained from Kaggle, and it includes information on **28** variables for each movie. These variables cover diverse aspects, ranging from the technical details of the movie-making process (e.g., duration, budget) to social media-related metrics (e.g., Facebook likes for actors and directors). **The sample size is (3598) and the number of variables is (28).**

The meta data describing the columns is as follows:

Sr. No.	Column Name	Description	Type
1.	color	Indicates whether the movie is in color	Categorical
2.	director_name	The name of the movie director	String
3.	num_critic_for_reviews	The number of critic reviews for the movie	Numeric
4.	duration	The duration of the movie in minutes	Numeric
5.	director_facebook_likes	The number of Facebook likes for the director's page	Numeric
6.	actor_3_facebook_likes	The number of Facebook likes for the third actor in the cast	Numeric
7.	actor_2_name	The name of the second actor in the cast	String
8.	actor_1_facebook_likes	The number of Facebook likes for the lead actor in the cast	Numeric
9.	gross	The gross revenue generated by the movie	Numeric
10.	genres	The genre(s) of the movie	Categorical
11.	actor_1_name	The name of the lead actor in the cast	String
12.	movie_title	The title of the movie	String
13.	cast_total_facebook_likes	The total number of Facebook likes for the entire cast	Numeric
14.	actor_3_name	The name of the third actor in the cast	String
15.	facenumber_in_poster	The number of faces in the movie poster	Numeric
16.	plot_keywords	Keywords describing the movie's plot	String
17.	movie_imdb_link	The IMDb link for the movie	String
18.	language	The language of the movie	Categorical

19.	country	he country where the movie was produced	Categorical
20.	content_rating	The content rating of the movie	Categorical
21.	budget	The budget of the movie	Numeric
22.	title_year	The year the movie was released	Numeric
23.	actor_2_facebook_likes	The number of Facebook likes for the second actor in the cast	Numeric
24.	aspect_ratio	The aspect ratio of the movie	Numeric
25.	movie_facebook_likes	The number of Facebook likes for the movie's page	Numeric
26.	IMDB_user_reviews	The number of user reviews on IMDb	Numeric
27.	IMDB_user_votes	The number of user votes on IMDb	Numeric
28.	IMDB_score	The IMDb score of the movie	Numeric

Sample Data:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	
1	color	director_n	num_critic	duration	director_fa	actor_3_fa	actor_2_n	actor_1_fa	gross	genres	actor_1_n	movie_title	cast_total	actor_3_n	facenumb	plot_keyw	movie_imd	language	country	co
2	Color	Eric Leight	145	82	0	388	D.B. Swee	1000	1.38E+08	Adventure	Alfre Wooc	Dinosaur	2945	Della Rees	1	egg|iguan	http://www	English	USA	PG
3	Color	Ron Howar	175	110	2000	636	T.J. Thyne	1000	2.6E+08	Comedy|F	Clint Howa	How the G	4146	Molly Shan	0	box office	http://www	English	USA	PG
4	Color	John Woo	237	123	610	653	Dougray Sc	10000	2.15E+08	Action|Ad	Tom Cruis	Mission: In	11930	Richard Rc	0	cure|miss	http://www	English	USA	PG
5	Color	Wolfgang F	231	130	249	461	Mary Elizal	784	1.83E+08	Action|Ad	Karen Aller	The Perfec	2684	Bob Gunto	0	death|fish	http://www	English	USA	PG
6	Color	Roland Em	192	142	776	1000	Adam Bald	13000	1.13E+08	Action|Dr	Heath Led	The Patriot	19454	Tom Wilkir	1	american	http://www	English	USA	R
7	Color	Dominic Se	175	127	57	3000	Angelina Jo	12000	1.02E+08	Action|Cri	Nicolas Ca	Gone in Six	29069	Robert Duv	1	auto theft	http://www	English	USA	PG
8	Color	Ridley Sco	265	171	0	695	Connie Nie	3000	1.88E+08	Action|Dr	Djimon Ho	Gladiator	6521	Oliver Ree	0	battlefield	http://www	English	USA	R
9	Color	Mark Dinda	141	78	10	253	Wendie Ma	558	89296573	Adventure	Eartha Kitt	The Emper	2039	John Fiedle	1	antidote|c	http://www	English	USA	G
10	Color	Bibo Berge	82	89	10	442	Rosie Pere	2000	50802661	Adventure	Frank Well	The Road t	3372	Elton John	1	adventure	http://www	English	USA	PG
11	Color	Robert Zer	185	130	0	568	Amber Vall	11000	1.55E+08	Drama|Fa	Harrison F	What Lies	12890	Miranda O	0	ghost|hau	http://www	English	USA	PG
12	Color	McG	181	94	368	466	LL Cool J	13000	1.25E+08	Action|Ad	Bill Murray	Charlie's A	15419	Kelly Lynch	0	booty shak	http://www	English	USA	PG
13	Color	Paul Verhc	180	119	719	423	Kim Dicker	833	73209340	Action|Ho	Greg Grun	Hollow Ma	2356	Joey Slotni	0	experimen	http://www	English	USA	R

Research Questions:

Let us delve into why we are executing this project. Let us look at some of the research questions that this project answers-

1. The first objective of our research endeavors to illuminate the nuanced relationship between directors, genres, and audience reception by determining the average IMDB rating associated with each director across the spectrum of genres they have contributed to. This investigation seeks to uncover the distinctive impact of directors within specific genres, offering insights into their storytelling prowess and cinematic execution across diverse narrative landscapes. By discerning patterns in directorial influence and genre-specific ratings, our analysis aims to provide valuable guidance for industry stakeholders, enabling informed decisions regarding directorial assignments, genre exploration, and strategic collaborations.
2. The second objective is to delve into the relationship between concatenated genre names and IMDB ratings using SQL queries in PySpark. By concatenating genre names and analyzing their impact on IMDB ratings, we aim to identify genre combinations associated with higher ratings. This analysis provides valuable insights into audience preferences and tastes, informing content creators and filmmakers about the most effective genre combinations to maximize audience engagement and IMDB ratings.
3. The third objective is to predict whether an IMDB rating will be high or not based on various factors. For this we are going to use a supervised learning model. By training a predictive model on features such as director information, movie duration, cast details, budget, and more, we aim to classify movies into categories of high or low IMDB ratings. This analysis enables filmmakers, producers, and industry professionals to anticipate the potential success of a movie based on its characteristics before its release. By understanding the factors that contribute to high IMDB ratings, stakeholders can make informed decisions regarding production strategies, marketing efforts, and resource allocation, ultimately maximizing the likelihood of achieving favorable ratings and audience reception.
4. The fourth objective is to investigate relationships between actors, movies and directors using graph frame analysis. By constructing a graph that connects actors, movies, and directors based on their interactions and collaborations, we aim to uncover patterns and

insights into the dynamics of the film industry. This analysis enables us to identify influential actors and directors, explore collaboration networks within the industry, and understand the impact of these relationships on movie success. Through PySpark's graph processing capabilities, this investigation offers valuable insights into the complex network of relationships within the film industry, empowering stakeholders to make informed decisions and strategic partnerships to enhance movie production and distribution.

5. The fifth objective is to group similar movies together based on their features. By clustering movies into cohesive groups, we aim to facilitate movie recommendations for users who prefer a particular movie. This approach allows for personalized recommendations tailored to individual preferences, enhancing user satisfaction and engagement with movie platforms.

Methodology:

In this project, we initiated a thorough data cleaning and preprocessing phase for the movie dataset, focusing on ensuring data integrity and quality. This involved addressing missing values, eliminating duplicates, and standardizing data formats. Notably, we imputed missing numeric values with their respective mean and median to mitigate potential biases or inaccuracies in the dataset, thereby enhancing the reliability of our subsequent analyses.

Following this, we delved into leveraging SQL using PySpark to glean insights from our dataset. We executed complex queries and extracted insights including identifying the highest average IMDB score for each director based on their movie genres. This allowed us to discern patterns in ratings corresponding to different directors. Additionally, we explored the highest IMDB ratings based on various genre combinations, shedding light on how different genres impact IMDB ratings.

In feature engineering, we employed various techniques to enrich the dataset and bolster model performance. This encompassed creating new features such as "popular actor 1" which was analyzing the percentage of the actor's popularity, providing insights into the significance of lead actor popularity. We also conducted transformations on existing features and encoded categorical variables. Through min-max scaling, one-hot encoding, vector assembly, and string indexing, we categorized our data, capturing complex relationships between features and the target variable, thereby enhancing the predictive power of our models.

For model selection and evaluation in supervised learning, we meticulously devised a robust framework. This entailed splitting the data into training and testing sets, selecting appropriate machine learning algorithms such as decision tree and random forest, and tuning hyperparameters. By evaluating model performance metrics, we identified the best-performing model for our specific problem domain. Furthermore, we planned to explore patterns using unsupervised learning models by creating clusters to discern similarities among movies. This approach enabled us to identify different clusters in which movies fall, providing valuable insights into movie categorization.

Finally, we focused on interpreting and communicating results effectively. Utilizing graph frames, we visualized data based on actors, directors and movies, mapping them with movies to empower stakeholders with actionable insights derived from our analysis.

Results and Finding:

Employing PySpark SQL, we delved into the dataset, uncovering broader insights and discerning patterns in ratings associated with diverse directors and genres. Through this exploration, we aimed to understand the factors influencing IMDB ratings, with a focus on identifying significant variables contributing to high ratings. The decision tree model can predict with a 72.44 % confidence whether a movie is going to have a high rating or not.

To enrich our dataset and enhance predictive power, we employed advanced feature engineering techniques. These methods enabled us to capture intricate relationships between features and the target variable, thereby improving the accuracy of our predictive models.

After thorough analysis, we have identified that movies often encompass multiple genres, and our research has unveiled the combinations that consistently yield favorable outcomes. Our findings reveal that the combination of animation and drama stands out as a noteworthy exemplar, boasting the highest average IMDB rating of 7.36. This insight underscores the potency of genre synergy in captivating audience interest and underscores the potential for strategic genre combinations to enhance a movie's reception and acclaim.

Furthermore, our research extended to analyzing the proficiency of directors across various genres, gauging their average IMDB scores within each genre. This endeavor serves to provide valuable insights not only to the directors themselves but also to producers and audiences alike. By discerning which directors excel in specific genres, stakeholders can make informed decisions regarding directorial assignments and genre selection, thereby increasing the likelihood of movie success and higher IMDB ratings. Lastly, we constructed a comprehensive graph frame that interconnects actors, directors, and movies. This initiative aimed to explore potential patterns of collaboration and preference among directors and actors. By investigating whether directors frequently collaborate with the same actors across multiple projects, we sought to validate the notion that such partnerships contribute to a director's success. Through this analysis, we aimed to provide empirical evidence to substantiate or debunk this widely held belief in the film industry.

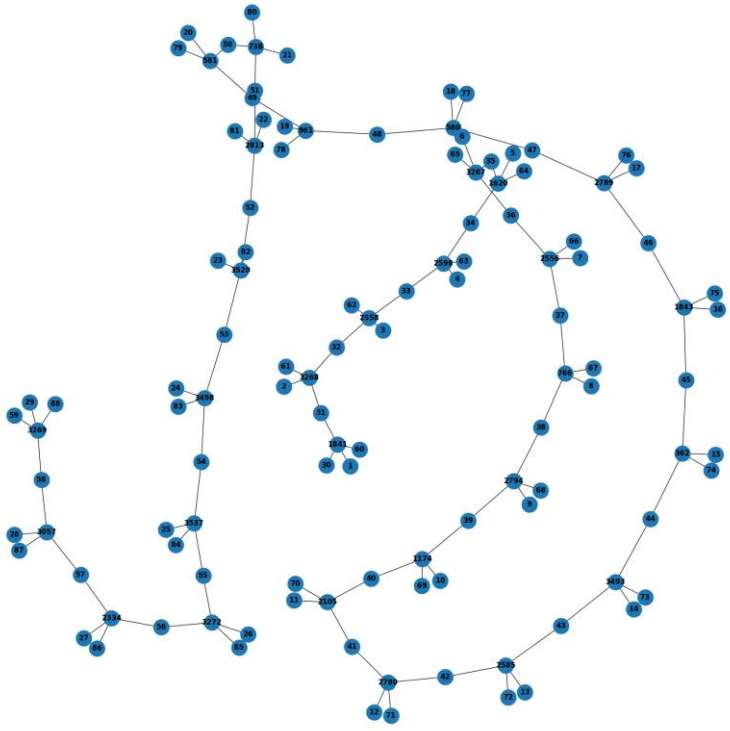
Overall, our project aims to provide valuable insights into the factors influencing movie ratings and to develop a robust predictive model for understanding the likelihood of a movie receiving a high IMDB rating.

director_name	genre	highest_avg_score
Jean-Marie Poiré	Comedy	5.8
Mark Romanek	Drama	7.0
Danny Provenzano	Drama	5.4
Bill Condon	Drama	6.5
Tom Tykwer	Drama	7.5
Nacho Vigalondo	Horror	7.2
Errol Morris	War	7.5
James Kerwin	Drama	5.4
Ian Sharp	Action	6.5
Hugh Hudson	Adventure	5.6
Britt Allcroft	Adventure	3.6
Michael Hoffman	Drama	6.800000000000001
Paul Schrader	Biography	6.6
Jake Kasdan	Comedy	5.833333333333333
Saul Dibb	Biography	6.9
Nora Ephron	Biography	7.0
Ronan Chapalain	Sci-Fi	6.9
Robert Luketic	Comedy	5.95
Martin Scorsese	Drama	8.0
François Ozon	Drama	6.8

only showing top 20 rows

genre_combination	avg_imdb_score
Animation, Drama	7.366666666666666
Animation, Comedy...	7.333333333333333
Action, Adventure...	7.1
Action, Adventure...	6.95
Adventure, Drama	6.936842105263159
Action, Animation	6.933333333333334
Adventure, Animat...	6.88
Drama	6.747131147540984
Action, Adventure...	6.673333333333335
Adventure, Comedy...	6.65
Adventure, Animation	6.648000000000001
Action, Adventure...	6.611111111111111
Comedy, Drama	6.477889447236179
Adventure	6.4604166666666645
Action, Adventure...	6.423529411764707
Action, Drama	6.375000000000004
Adventure, Animat...	6.344
Action, Adventure	6.2632530120481915
Animation, Comedy	6.122222222222223
	6.090882352941176

only showing top 20 rows



Conclusion:

In conclusion, our project represents a comprehensive endeavor to analyze movie data using PySpark, employing a range of methodologies to extract valuable insights into the determinants of IMDB ratings and movie success. Through the application of supervised learning models, we achieved a predictive accuracy of 72.44% in discerning whether a movie's IMDB rating would be high or low, thereby empowering filmmakers and industry professionals to make informed decisions regarding production strategies and resource allocation. Moreover, our exploration of unsupervised learning techniques enabled us to cluster similar movies together, facilitating personalized recommendations and enhancing user engagement on movie platforms.

Furthermore, our investigation into the intricate relationships between actors, directors, and movies uncovered compelling patterns and dynamics within the film industry, shedding light on collaboration networks and their impact on movie success. Leveraging graph frame analysis, we revealed valuable insights into the interconnectedness of industry stakeholders and their contributions to cinematic endeavors. Our research also underscored the significance of genre combinations in shaping audience reception, with the animation and drama genre emerging as a notable exemplar of success. Overall, our study contributes to a deeper understanding of the multifaceted dynamics driving the film industry, offering actionable insights for stakeholders to enhance decision-making processes and maximize the potential for movie success through data-driven strategies.