



FINAL PROJECT
PRESENTATION

PREDICTING MOVIE SUCCESS USING LOGISTIC REGRESSION



PRESENTED BY
PRASHANT KOIRALA
AASKA KOIRALA
AISHMITA YONZAN

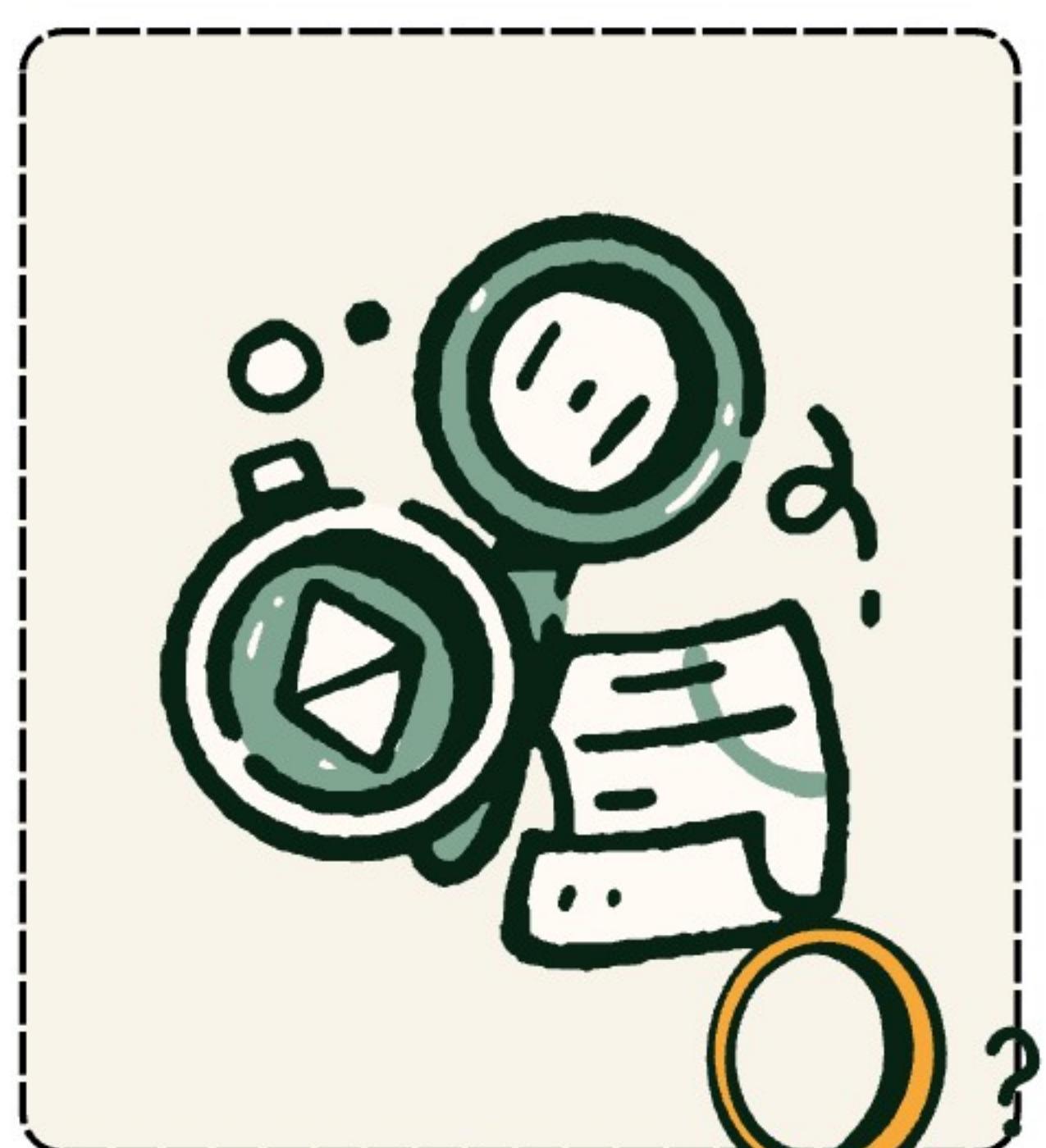


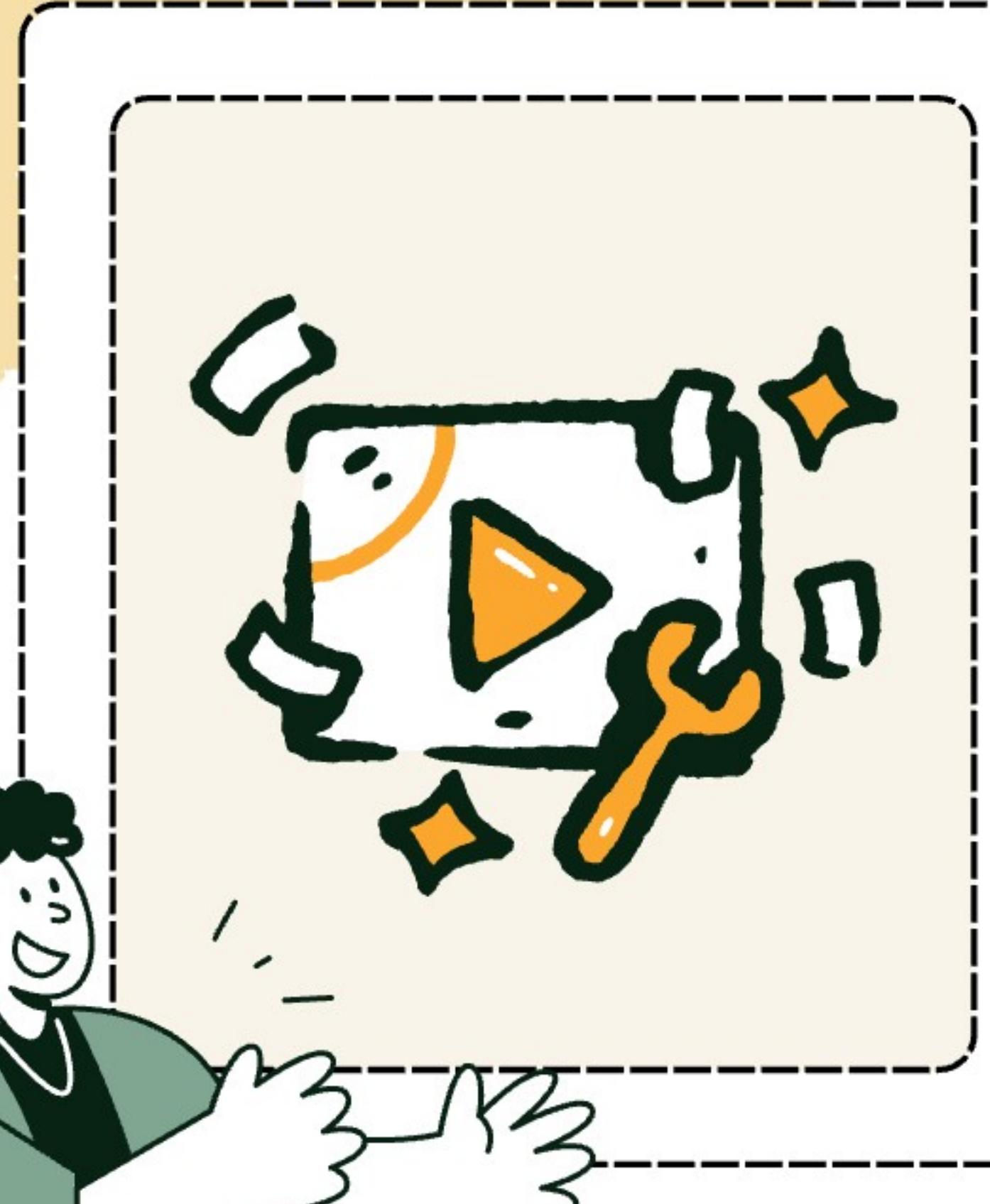


PROBLEM STATEMENT



Can we predict whether a movie will be successful or not based on its metadata (e.g., budget, runtime, genre, and other pre-release attributes)?





TYPE OF PROBLEM & STATISTICAL MODEL

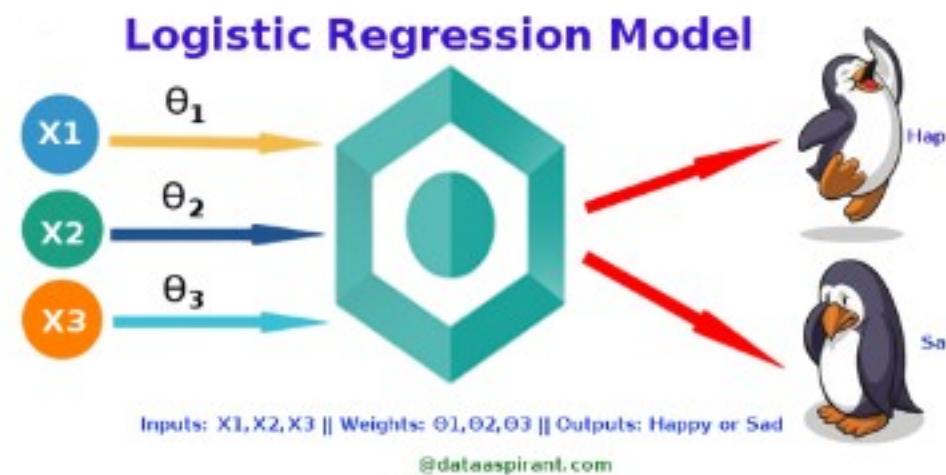
Type of Problem:

- Binary Classification Problem
- Predictive Modeling Task

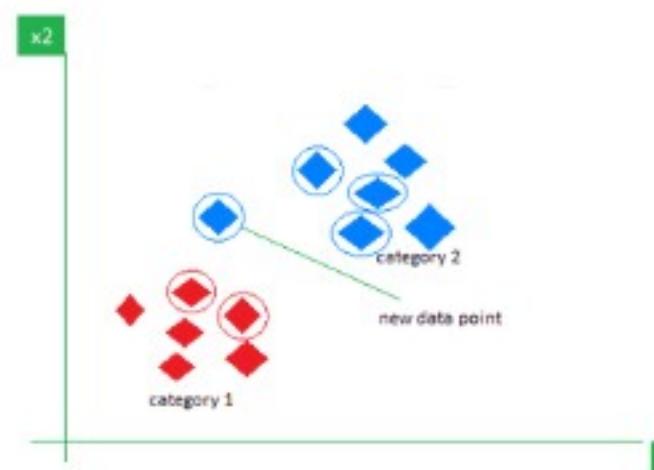
MODEL SELECTION & JUSTIFICATION

Models Evaluated

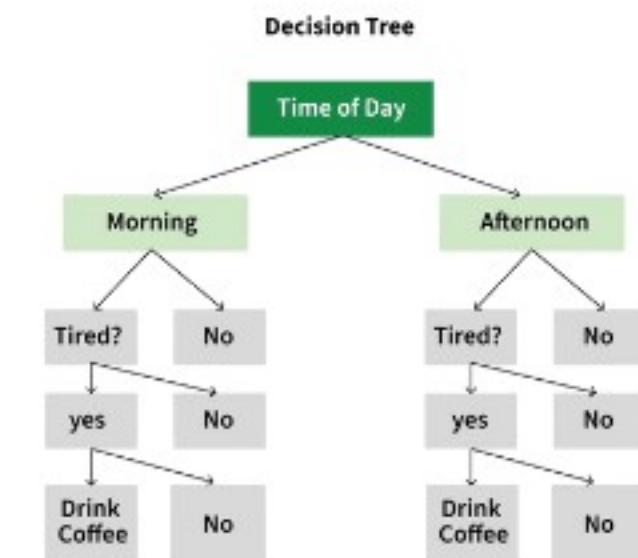
Logistic Regression



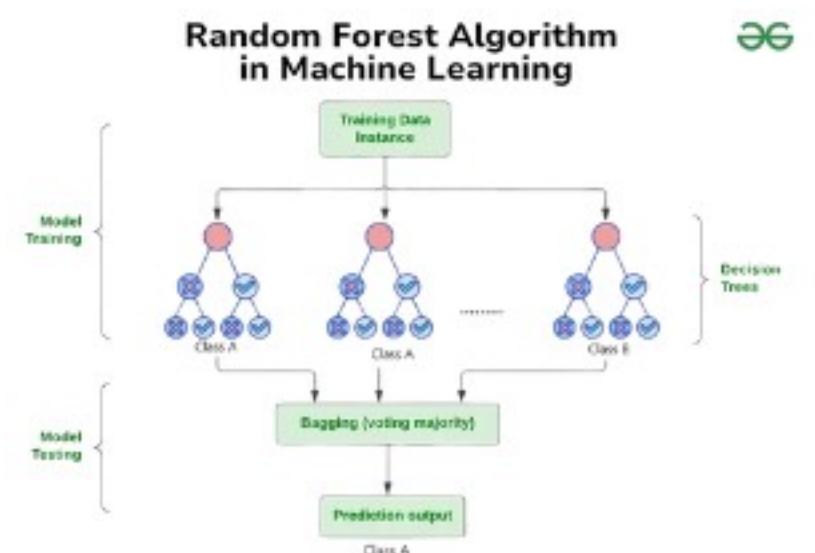
K-Nearest Neighbors



Decision Tree



Random Forest



Chosen Model

Logistic Regression

KEY FEATURES

- It is ideal for binary outcomes (success vs. not success).
- It provides clear coefficients to understand how each feature (e.g., budget) impacts the probability of success.
- Suitable for a student project, as it's straightforward to implement and interpret.



LITERATURE REVIEW & DATASET SELECTION

- DA, O., OM, S., AK, F., O, A., A, O., A, O., A, W., & M, Y. (2021). Movie success prediction using data mining. *British Journal of Computer Networking and Information Technology*, 4(2), 22–30. <https://doi.org/10.52589/bjcnit-cqocirec>
- Zheng, Y. (2024). Predicting movie box office based on machine learning, deep learning, and statistical methods. *Applied and Computational Engineering*, 94(1), 20–32. <https://doi.org/10.54254/2755-2721/94/2024melb0069>
- Subramaniyaswamy, V., Vaibhav, M. V., Prasad, R. V., & Logesh, R. (2017). Predicting movie box office success using multiple regression and SVM. *Journal of Science & Statistics*, 182–186. <https://doi.org/10.1109/iss1.2017.8389394>
- Velingkar, G., Varadarajan, R., Lanka, S., & M, A. K. (2022). Movie Box-Office success Prediction using Machine Learning. 2022 Second International Conference on Power, Control and Computing Technologies (ICPC2T), 9, 1–6. <https://doi.org/10.1109/icpc2t53885.2022.9776798>
- Mr, N. V., Mr, P. M., & Pb, S. B. (2014). Predicting movie success based on IMDB data. *International Journal of Business Intelligents*, 3(2). <https://doi.org/10.20894/ijbi.105.003.002.004>

BRIEF OVERVIEW OF LITERATURE REVIEW

Proceedings Article • 10.1109/ISS1.2017.8389394

[3. Predicting movie box office success using multiple regression and SVM](#)

V. Subramaniyaswamy, M. Viginesh Vaibhav, R. Vishnu Prasad +1 more

1 Dec 2017

66 26 [Request PDF](#) [Podcast](#) [Chat](#)

 66

The paper you referenced is not the same as the one discussed here. This research focuses on predicting movie box office success using multiple regression and SVM, analyzing factors like trailer views, Wikipedia page views, and critic ratings.

Proceedings Article • 10.1109/ICPC2T53885.2022.9776798

[4. Movie Box-Office Success Prediction Using Machine Learning](#)

Gaurang Velingkar, Rakshita Varadarajan, Sidharth Lanka +1 more

1 Mar 2022

66 4 [Request PDF](#) [Podcast](#) [Chat](#)

 66

The paper focuses on predicting movie box office success using machine learning, specifically employing a Random Forest Regression model to analyze factors like genre, budget, and star power, aiding investors in making informed decisions before a movie's release.

 . Proceedings Article • 10.1109/icpc2t53885.2022.9776798

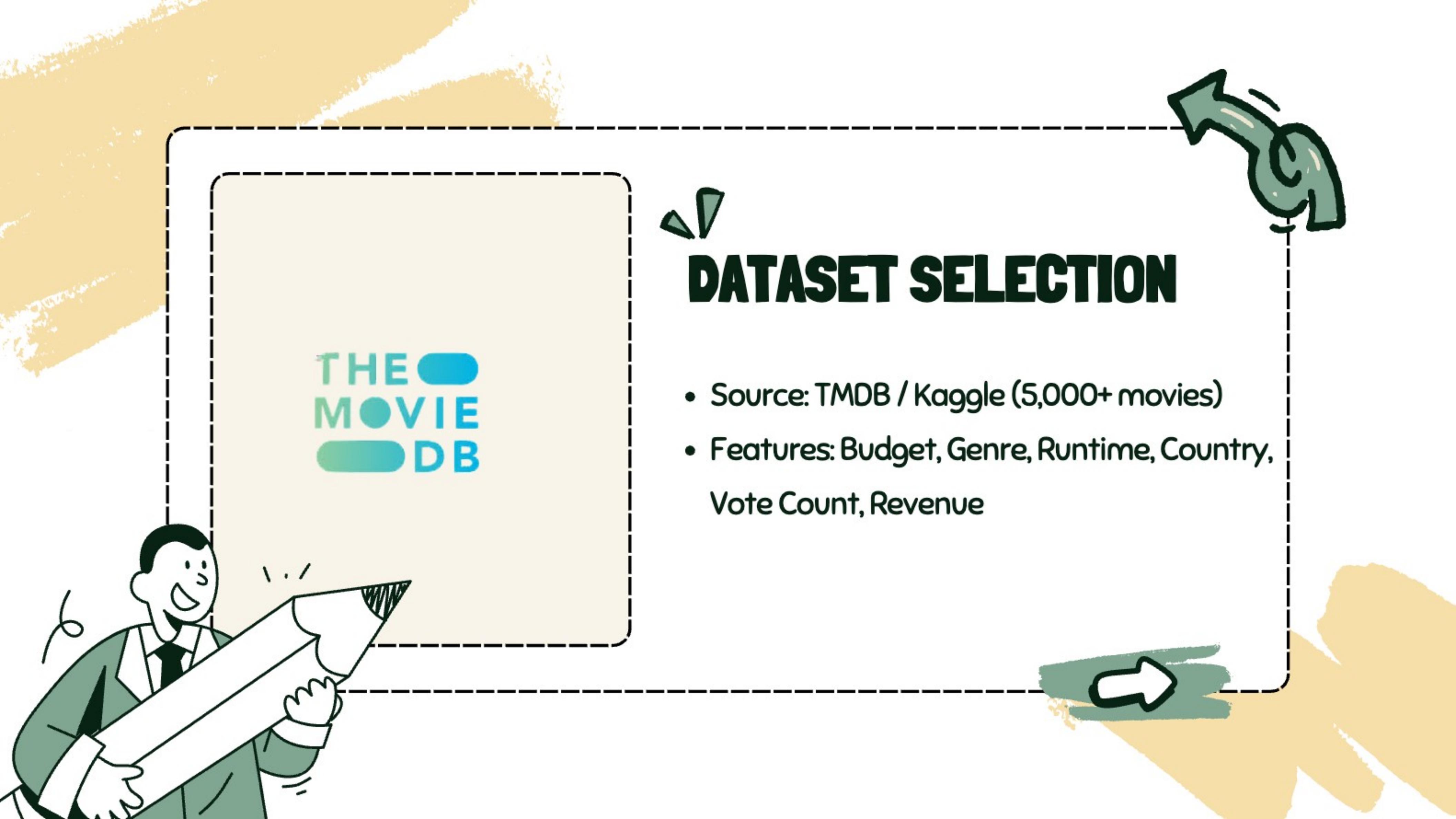
[5. Movie Box-Office Success Prediction Using Machine Learning](#)

1 Mar 2022

66 8 [PDF](#) [Podcast](#) [Chat](#) 

The paper focuses on predicting movie box office success using machine learning, specifically employing a Random Forest Regression model to analyze factors like genre, budget, and star power, aiding investors in making informed decisions before a movie's release.





DATASET SELECTION

- Source: TMDB / Kaggle (5,000+ movies)
- Features: Budget, Genre, Runtime, Country, Vote Count, Revenue



THE
MOVIE
DB

EXPLORATORY DATA ANALYSIS (EDA)

tmdb_5000_movies

Search (Cmd + Ctrl + U)

Home Insert Draw Page Layout Formulas Data Review View Acrobat

Paste Calibri (Body) 12 A A General Conditional Formatting Format as Table Cell Styles Insert Delete Sort & Filter Find & Select Add-ins Create PDF and share link

U1 x ✓ fx

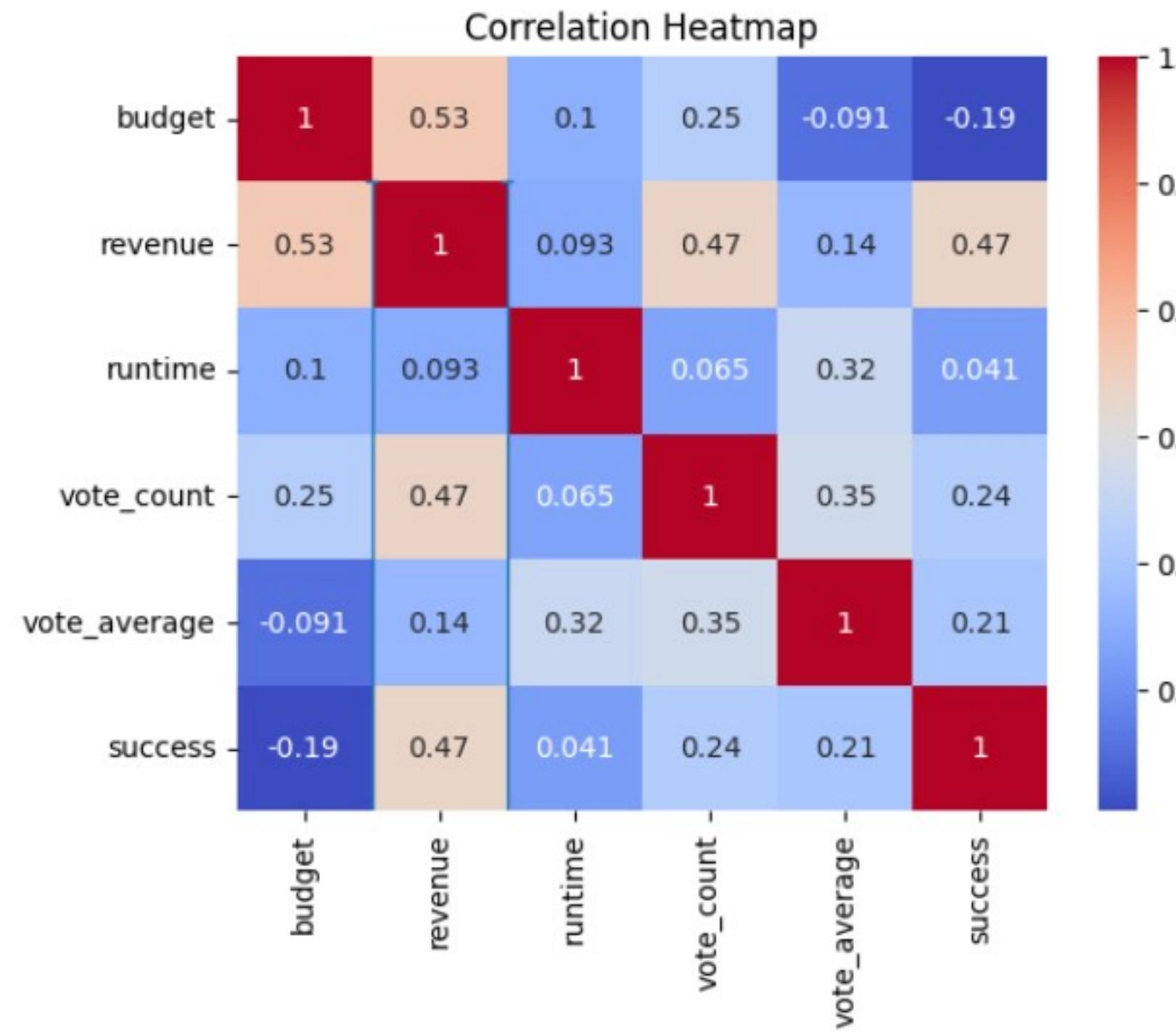
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	budget	genres	homepage	id	keywords	original_lang	original_title	overview	popularity	production_c	production_c	release_date	revenue	runtime	spoken_lang	status	tagline	title	vote_average	vote_count	
2	237000000	[{"id": 28, "n http://www.	19995	[{"id": 1463, "en	Avatar	In the 22nd c	150.437577	[{"name": "Is [{"iso_3166_	12/10/09	2787965087	162	[{"iso_639_1 Released	Enter the W	Avatar	7.2	11800					
3	300000000	[{"id": 12, "n http://disney	285	[{"id": 270, "en	Pirates of th	Captain Barb	139.082615	[{"name": "V [{"iso_3166_	5/19/07	961000000	169	[{"iso_639_1 Released	At the end of Pirates of th		6.9	4500					
4	245000000	[{"id": 28, "n http://www.	206647	[{"id": 470, "en	Spectre	A cryptic me	107.376788	[{"name": "C [{"iso_3166_	10/26/15	880674609	148	[{"iso_639_1 Released	A Plan No Or Spectre		6.3	4466					
5	250000000	[{"id": 28, "n http://www.	49026	[{"id": 849, "en	The Dark Kni	Following th	112.31295	[{"name": "L [{"iso_3166_	7/16/12	1084939099	165	[{"iso_639_1 Released	The Legend E The Dark Kni		7.6	9106					
6	260000000	[{"id": 28, "n http://movie	49529	[{"id": 818, "en	John Carter	John Carter	43.926995	[{"name": "V [{"iso_3166_	3/7/12	284139100	132	[{"iso_639_1 Released	Lost in our w John Carter		6.1	2124					
7	258000000	[{"id": 14, "n http://www.	559	[{"id": 851, "en	Spider-Man	: The seeming	115.699814	[{"name": "C [{"iso_3166_	5/1/07	890871626	139	[{"iso_639_1 Released	The battle w Spider-Man :		5.9	3576					
8	260000000	[{"id": 16, "n http://disney	38757	[{"id": 1562, "en	Tangled	When the kir	48.681969	[{"name": "V [{"iso_3166_	11/24/10	591794936	100	[{"iso_639_1 Released	They're takin Tangled		7.4	3330					
9	280000000	[{"id": 28, "n http://marve	99861	[{"id": 8828, "en	Avengers: Ag	When Tony S	134.279229	[{"name": "N [{"iso_3166_	4/22/15	1405403694	141	[{"iso_639_1 Released	A New Age F Avengers: Ag		7.3	6767					
10	250000000	[{"id": 12, "n http://harryp	767	[{"id": 616, "en	Harry Potter	As Harry beg	98.885637	[{"name": "V [{"iso_3166_	7/7/09	933959197	153	[{"iso_639_1 Released	Dark Secrets Harry Potter		7.4	5293					
11	250000000	[{"id": 28, "n http://www.	209112	[{"id": 849, "en	Batman v Su	Fearing the z	155.790452	[{"name": "C [{"iso_3166_	3/23/16	873260194	151	[{"iso_639_1 Released	Justice or rev Batman v Su		5.7	7004					
12	270000000	[{"id": 12, "n http://www.	1452	[{"id": 83, "en	Superman Ri	Superman re	57.925623	[{"name": "C [{"iso_3166_	6/28/06	391081192	154	[{"iso_639_1 Released	Superman Re		5.4	1400					
13	200000000	[{"id": 12, "n http://www.	10764	[{"id": 627, "en	Quantum of	Quantum of	107.928811	[{"name": "E [{"iso_3166_	10/30/08	586090727	106	[{"iso_639_1 Released	For love, for Quantum of		6.1	2965					
14	200000000	[{"id": 12, "n http://disney	58	[{"id": 616, "en	Pirates of th	Captain Jack	145.847379	[{"name": "V [{"iso_3166_	6/20/06	1065659812	151	[{"iso_639_1 Released	Jack is back! Pirates of th		7	5246					
15	255000000	[{"id": 28, "n http://disney	57201	[{"id": 1556, "en	The Lone Rai	The Texas Ra	49.046956	[{"name": "V [{"iso_3166_	7/3/13	89289910	149	[{"iso_639_1 Released	Never Take C The Lone Rai		5.9	2311					
16	225000000	[{"id": 28, "n http://www.	49521	[{"id": 83, "en	Man of Steel	A young boy	99.398009	[{"name": "L [{"iso_3166_	6/12/13	662845518	143	[{"iso_639_1 Released	You will beli Man of Steel		6.5	6359					
17	225000000	[{"id": 12, "name": "Adver	2454	[{"id": 818, "en	The Chronicl	One year aft	53.978602	[{"name": "V [{"iso_3166_	5/15/08	419651413	150	[{"iso_639_1 Released	Hope has a n The Chronicl		6.3	1630					
18	220000000	[{"id": 878, "n http://marve	24428	[{"id": 242, "en	The Avenger	When an unc	144.448633	[{"name": "P [{"iso_3166_	4/25/12	1519557910	143	[{"iso_639_1 Released	Some assem The Avenger		7.4	11776					
19	380000000	[{"id": 12, "n http://disney	1865	[{"id": 658, "en	Pirates of th	Captain Jack	135.413856	[{"name": "V [{"iso_3166_	5/14/11	1045713802	136	[{"iso_639_1 Released	Live Forever Pirates of th		6.4	4948					
20	225000000	[{"id": 12, "n http://www.	41154	[{"id": 4379, "en	Men in Black	Agents J (Wi	52.035179	[{"name": "A [{"iso_3166_	5/23/12	624026776	106	[{"iso_639_1 Released	They are bac Men in Black		6.2	4160					
21	250000000	[{"id": 12, "n http://www.	122917	[{"id": 417, "en	The Hobbit: 1	Immediately	120.965743	[{"name": "V [{"iso_3166_	12/10/14	956019788	144	[{"iso_639_1 Released	Witness the The Hobbit: 1		7.1	4760					
22	215000000	[{"id": 12, "n http://www.	1930	[{"id": 1872, "en	The Amazing	Peter Parker	89.866276	[{"name": "C [{"iso_3166_	6/27/12	752215857	136	[{"iso_639_1 Released	The untold st The Amazing		6.5	6586					
23	200000000	[{"id": 12, "n http://www.	20662	[{"id": 4147, "en	Robin Hood	When soldie	37.668301	[{"name": "I [{"iso_3166_	5/12/10	310669540	140	[{"iso_639_1 Released	Rise and rise Robin Hood		6.2	1398					
24	250000000	[{"id": 12, "n http://www.	57158	[{"id": 603, "en	The Hobbit: 1	The Dwarves	94.370564	[{"name": "V [{"iso_3166_	12/11/13	958400000	161	[{"iso_639_1 Released	Beyond dark! The Hobbit: 1		7.6	4524					
25	180000000	[{"id": 12, "n http://www.	2268	[{"id": 392, "en	The Golden C	After overha	42.990906	[{"name": "N [{"iso_3166_	12/4/07	372234864	113	[{"iso_639_1 Released	There are w! The Golden C		5.8	1303					
26	207000000	[{"id": 12, "name": "Adver	254	[{"id": 774, "en	King Kong	In 1933 New	61.22601	[{"name": "V [{"iso_3166_	12/14/05	550000000	187	[{"iso_639_1 Released	The eighth w King Kong		6.6	2337					
27	200000000	[{"id": 18, "n http://www.	597	[{"id": 2580, "en	Titanic	84 years late	100.025899	[{"name": "P [{"iso_3166_	11/18/97	1845034188	194	[{"iso_639_1 Released	Nothing on E Titanic		7.5	7562					
28	250000000	[{"id": 12, "n http://marve	271110	[{"id": 393, "en	Captain Ame	Following th	198.372395	[{"name": "S [{"iso_3166_	4/27/16	1153304495	147	[{"iso_639_1 Released	Divided We F Captain Ame		7.1	7241					
29	209000000	[{"id": 53, "name": "Thrill	44833	[{"id": 1721, "en	Battleship	When manki	64.928382	[{"name": "L [{"iso_3166_	4/11/12	303025485	131	[{"iso_639_1 Released	The Battle fc Battleship		5.5	2114					
30	150000000	[{"id": 28, "n http://www.	135397	[{"id": 1299, "en	Jurassic Wor	Twenty-two	418.708552	[{"name": "L [{"iso_3166_	6/9/15	1513528810	124	[{"iso_639_1 Released	The park is o Jurassic Wor		6.5	8662					
31	200000000	[{"id": 28, "n http://www.	37724	[{"id": 470, "en	Skyfall	When Bond's	93.004993	[{"name": "C [{"iso_3166_	10/25/12	1108561013	143	[{"iso_639_1 Released	Think on you Skyfall		6.9	7604					
32	200000000	[{"id": 28, "n http://www.	558	[{"id": 851, "en	Spider-Man	Peter Parker	35.149586	[{"name": "C [{"iso_3166_	6/25/04	783766341	127	[{"iso_639_1 Released	There's a hei Spider-Man :		6.7	4321					
33	200000000	[{"id": 28, "n http://marve	68721	[{"id": 949, "en	Iron Man 3	When Tony S	77.68208	[{"name": "N [{"iso_3166_	4/18/13	1215439994	130	[{"iso_639_1 Released	Unleash the Iron Man 3		6.8	8806					
34	200000000	[{"id": 10751 http://disney	12155	[{"id": 818, "en	Alice in Won	Alice, an unp	78.530105	[{"name": "V [{"iso_3166_	3/3/10	1025491110	1										

EXPLORATORY DATA ANALYSIS (EDA)

Features Used for Analysis

- **Numerical:** budget, revenue, runtime, year, vote_count (and possibly vote_average if not dropped).
- **Categorical:** certification_US, genre, country.
- **Target:** success (True/False, where True means $\text{revenue} \geq 2 \times \text{budget}$).

WHY THOSE FEATURES?



- 1. SUCCESS IS MOST TIED TO REVENUE (0.47) AND VOTE_COUNT (0.24), SO THESE FEATURES MIGHT PREDICT SUCCESS WELL.**
- 2. BUDGET AND REVENUE ARE RELATED (0.53), MEANING BIGGER BUDGETS OFTEN LEAD TO HIGHER REVENUE.**
- 3. RUNTIME HAS LITTLE CONNECTION TO OTHER FEATURES (ALL NEAR 0), SO IT MIGHT NOT BE A STRONG PREDICTOR OF SUCCESS.**

DATA CLEANING

DATA CLEANING PROGRAM

The screenshot shows a code editor window titled "Code" with a dark theme. The main pane displays a Python script named "data cleaning.py". The script is used to clean a dataset from "tmdb_5000_movies.csv". It includes filtering for runtime, revenue, budget, and year, handling missing certification values, replacing missing genres with "None", creating a "year" column from the release date, and adding a success column.

```
1 1 < > X 2 7 > Code 47% 100% 25% Mon 02 Jun 07:35 PM
data cleaning.py 9
Users > prashantkoirala > Desktop > DATA 200 Project > data cleaning.py > ...
9
10
11 df = pd.read_csv('tmdb_5000_movies.csv')
12
13
14 df = df.drop_duplicates()
15
16
17 # remove unusable data
18 df = df[(df['runtime'] > 0) & (df['runtime'].notnull()) &
19 | (df['revenue'] != 0) & (df['revenue'].notnull()) &
20 | (df['budget'] != 0) & (df['budget'].notnull()) &
21 | (df['adult'] == False) &
22 | (df['year'] >= 1970)
23 | (df['vote_count'] < 5)]
24
25
26 # replace missing certification with 'Not Rated'
27 df.loc[df['certification_US'].isnull(), 'certification_US'] = 'NR'
28 df.loc[df['certification_US'] == 'None', 'certification_US'] = 'NR'
29
30
31 # replace missing genre with 'None'
32 df.loc[df['genre'].isnull(), 'genre'] = 'None'
33
34
35 # create year column
36 # convert release_date to have same format
37 for index, row in df.iterrows():
38     try:
39         date = parser.parse(row['release_date'])
40         year = date.year
41         newDate = f'{date.year}-{date.month}-{date.day}'
42     except: # if release_date is empty
43         year = np.nan
44         newDate = np.nan
45
46         df.at[index, 'year'] = year
47         df.at[index, 'release_date'] = newDate
48
49
50 # create success column
```

Bottom status bar: Restricted Mode (X), 3 △ 6, Ln 33, Col 1, Spaces: 4, UTF-8, LF, Python, 47%, 100%, 25%, Mon 02 Jun 07:35 PM.

DATA CLEANING

- 1. Removed duplicates and movies with missing/zero budget, revenue, or runtime.**
- 2. Replaced missing certification_US with “NR”; extracted first genre and country.**
- 3. Filtered movies before 1970 and countries with < 5 occurrences.**
- 4. Created success (True if revenue \geq 2x budget) and year columns.**

AFTER DATA CLEANING

MODEL TRAINING PROCESS

Trained on 5000+ movies



One-hot encoding for categorical features

If `Genre = Action`, it becomes:

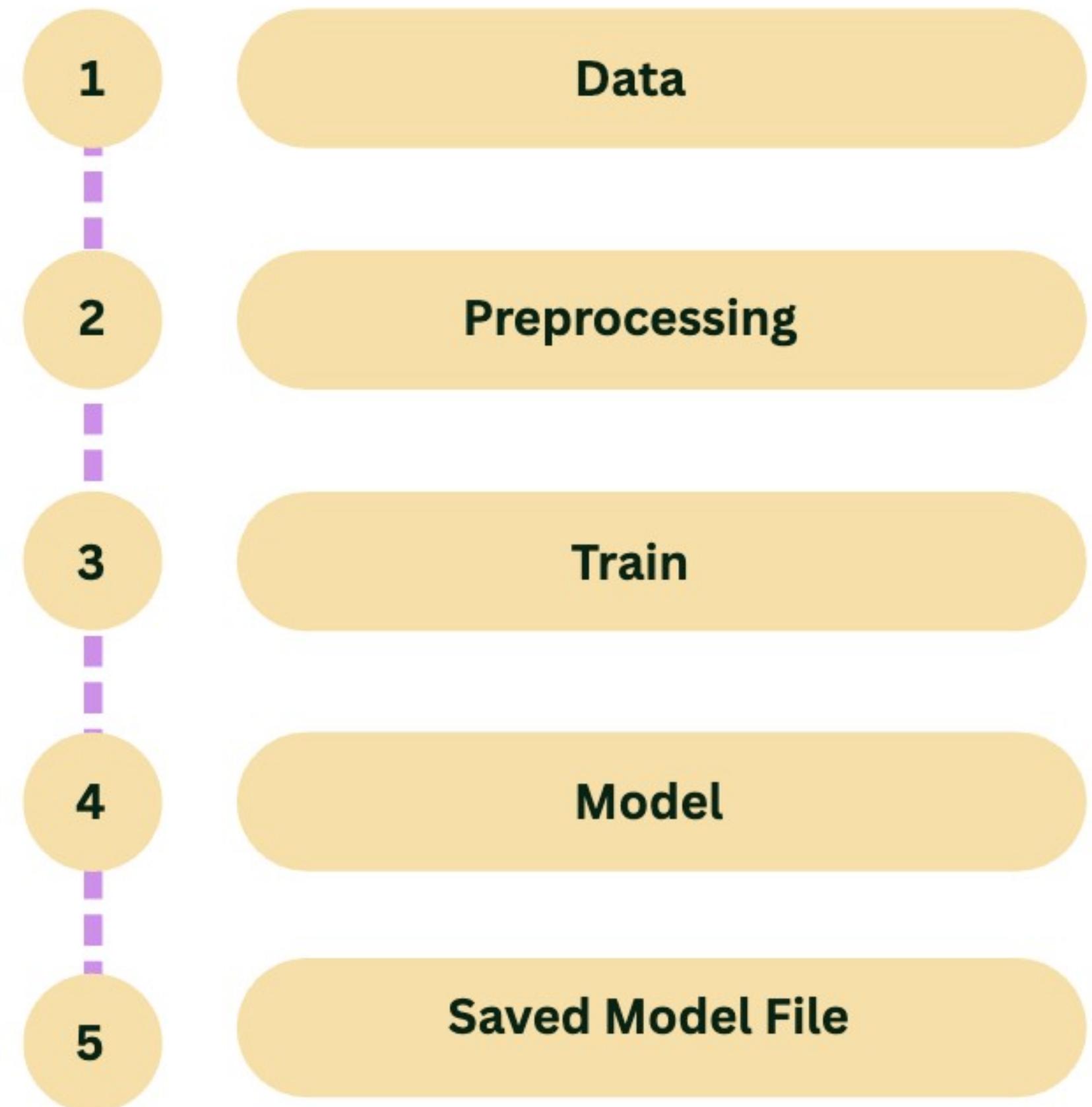
nginx

Action	Comedy	Drama
1	0	0

Model serialized using `joblib`

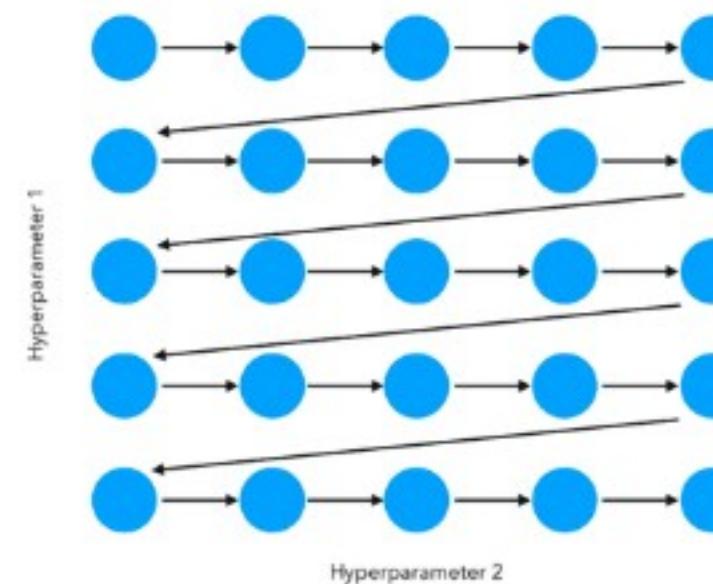


WORKFLOW



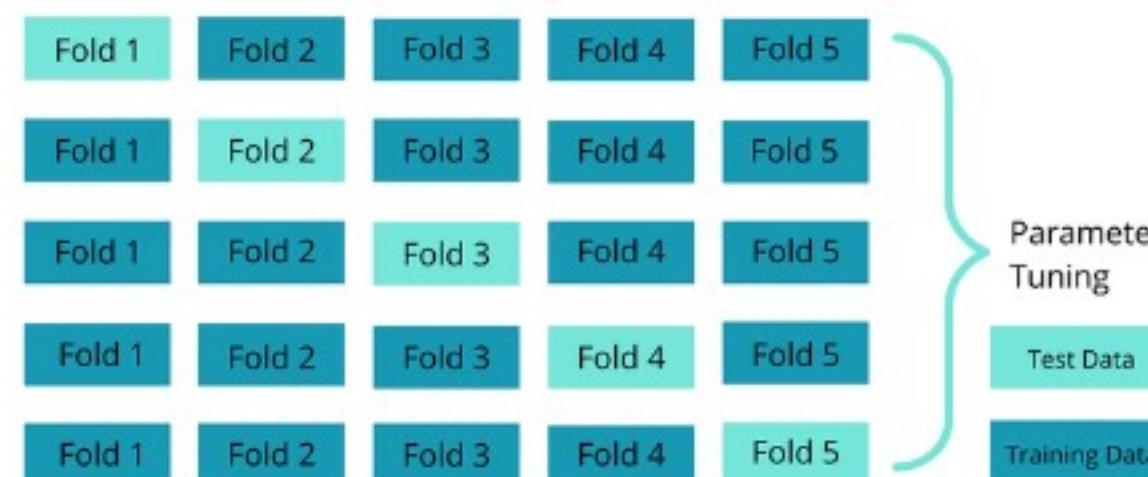
HYPERPARAMETER TUNING

Hyperparameter Tuning (GridSearchCV)



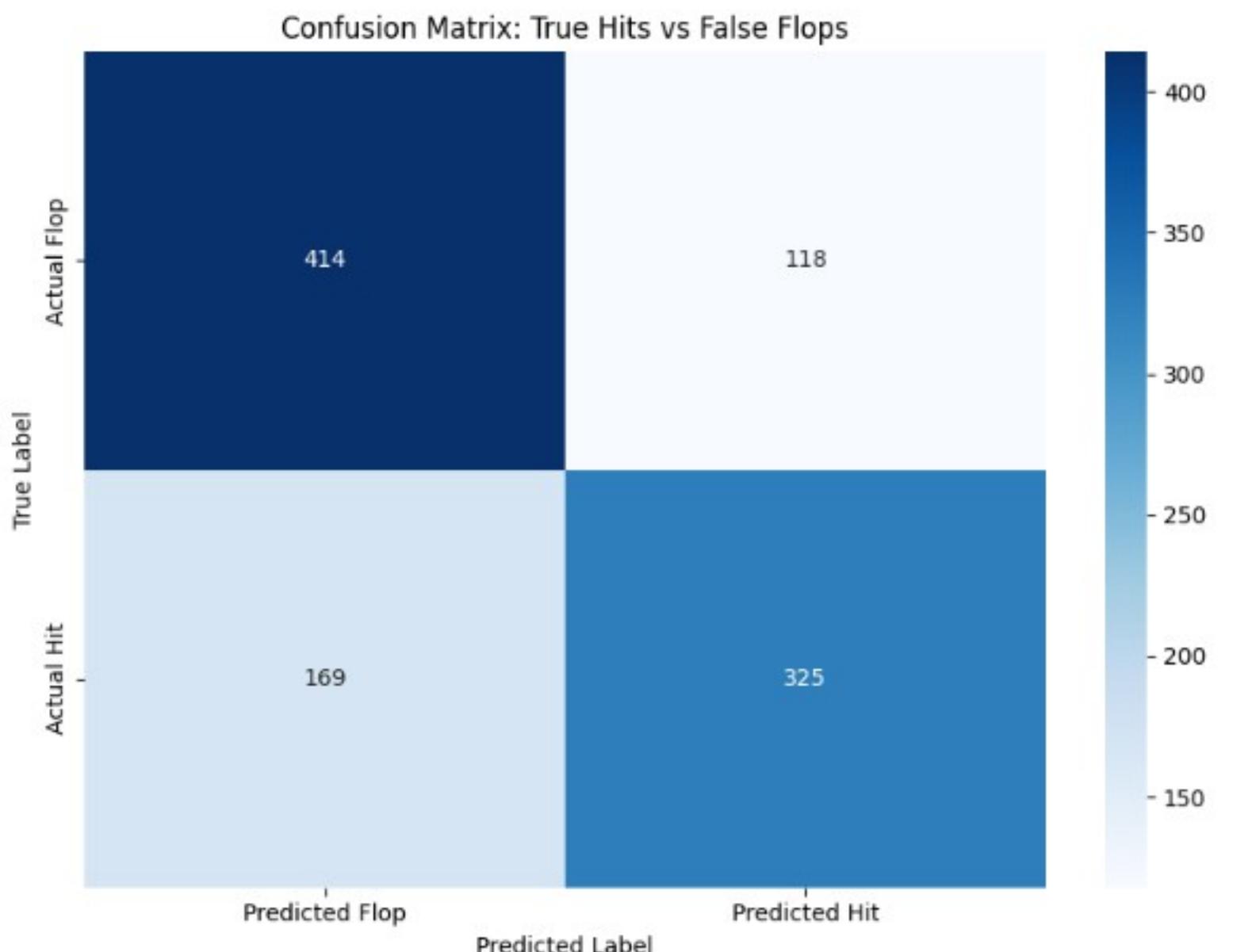
Used GridSearchCV with 10-fold cross-validation

Explored parameters: C, solver



Best Params:
C=1000
solver='newton-cg'

MODEL EVALUATION



MODEL EVALUATION

- 1. Accuracy: 72.03%**
- 2. Precision: 73.36%**
- 3. Recall: 65.79%**
- 4. F1 Score: 69.37%**

— Model Evaluation Results —

Accuracy: 0.7203
Precision: 0.7336
Recall: 0.6579
F1 Score: 0.6937

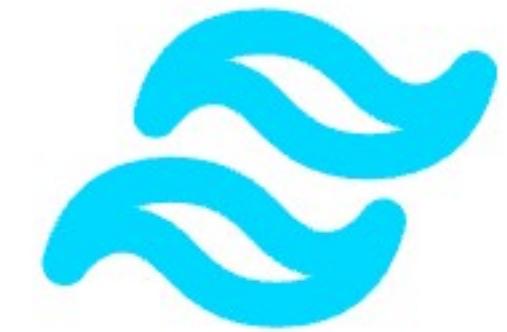
Classification Report:

	precision	recall	f1-score
Flop	0.71	0.78	0.74
Hit	0.73	0.66	0.69
accuracy			0.72
macro avg	0.72	0.72	0.72
weighted avg	0.72	0.72	0.72

ARCHITECTURE



FRONTEND



BACKEND



BACKEND ARCHITECTURE



BACKEND



Built using Flask (Python web framework)

Responsible for:

Serving the REST API

Preprocessing data

Calling the logistic regression model

Returning prediction results

FRONTEND ARCHITECTURE



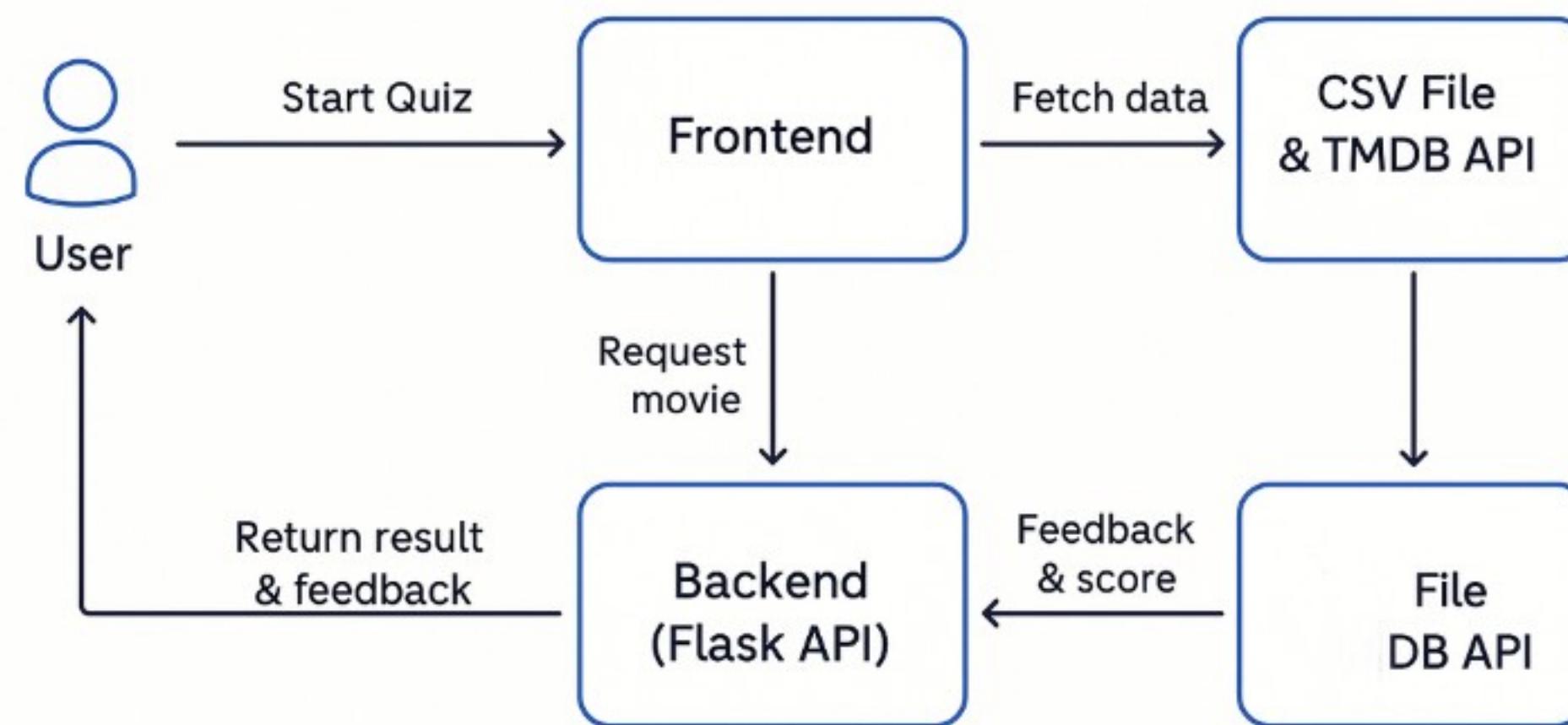
FRONTEND



Key features:

- **Responsive layout**
- **Interactive quiz game**
- **Filters by genre, country, and certification**
- **Animated feedback and score tracking**

SYSTEM INTEGRATION & DATA FLOW



DEMO

CHALLENGES FACED

- Handling missing/inconsistent movie data
- Encoding categorical variables with low frequency
- TMDB Movie Poster Loading
- Ensuring consistent one-hot encoding during training and inference
- Responsive UI across devices



FUTURE ENHANCEMENTS

- Improve prediction with ensemble models (Random Forest, XGBoost)
- Add user authentication and leaderboard
- Support movie trailers using YouTube API
- Enable feedback loop to refine model with real user guesses





CONCLUSION & KEY TAKEAWAYS

- 1. Built an end-to-end movie success predictor**
- 2. Applied logistic regression for real-world classification**
- 3. Designed interactive UI for user engagement**
- 4. Demonstrated machine learning integration with web development**
- 5. Balanced accuracy, explainability, and usability**

REFERENCES

- DA, O., OM, S., AK, F., O, A., A, O., A, O., A, W., & M, Y. (2021). Movie success prediction using data mining. *British Journal of Computer Networking and Information Technology*, 4(2), 22–30. <https://doi.org/10.52589/bjcnit-cqocirec>
- Zheng, Y. (2024). Predicting movie box office based on machine learning, deep learning, and statistical methods. *Applied and Computational Engineering*, 94(1), 20–32. <https://doi.org/10.54254/2755-2721/94/2024melb0069>
- Subramaniyaswamy, V., Vaibhav, M. V., Prasad, R. V., & Logesh, R. (2017). Predicting movie box office success using multiple regression and SVM. *Journal of Science & Statistics*, 182–186. <https://doi.org/10.1109/iss1.2017.8389394>
- Velingkar, G., Varadarajan, R., Lanka, S., & M, A. K. (2022). Movie Box-Office success Prediction using Machine Learning. 2022 Second International Conference on Power, Control and Computing Technologies (ICPC2T), 9, 1–6.
<https://doi.org/10.1109/icpc2t53885.2022.9776798>
- Mr, N. V., Mr, P. M., & Pb, S. B. (2014). Predicting movie success based on IMDB data. *International Journal of Business Intelligents*, 3(2). <https://doi.org/10.20894/ijbi.105.003.002.004>



THANK YOU

