TMDB Box Office Prediction

Predicting the box office revenue for a given movie

Meet the Analysts

Carlos G. Fernandez

Luis Eduardo Leal

Nicole Ardizzi

Why this topic?

- Who doesn’t love movies?
- $41.7 billion industry in 2018,
- The film industry is more popular than ever.
- What if you can make a movie?
- Interesting features to analyze

The Dataset

Source

Kaggle.com

TMDB Box Office Prediction

Can you predict a movie's worldwide box office revenue?

Description

- A list of movies and a variety of metadata
- Metadata from The Movie Database (TMDB)
- Collected from the TMDB Open API

Size

- 3000 rows (movies)
- 23 columns

Dataflow

Postgres → Pandas

- Raw data was stored in Postgres
- Transferred data from Postgres to Pandas using SQLAlchemy
- pd.read_sql( )

Pandas → Tableau

- .to_csv( )

Columns Names & Data Types

String

- Id
- Homepage
- Imbd_id
- Original language
- Original_title
- Overview
- Poster path
- Release_date
- Status
- Tagline
- Title

Numerical

- Budget
- Popularity
- Runtime
- Revenue

Data Structure Format

- Belongs to collection
- genres
- Production Companies
- Production_countries
- Spoken_languages
- Keywords
- Cast
- Crew

Exploratory Data Analysis / Data Processing

Exploratory Analysis

Questions

- How does budget influence revenue?
- What is the average revenue for each genre?
- Revenue of movies with home page vs no homepage?
- Did production companies have a major influence on revenue?
- Which production company had the highest average revenue?

Tools and Technologies Used During EDA

- Pandas
- Numpy
- Matplotlib
- Seaborn
- SqlAlchemy
- Pearsonr

Genres

For loop function:

- Goes through each row
- Looks at each genre list
- Looks at each genre in the genre list

Unpack and encode:

- Unique genre names are converted into columns
- For each genre in the genre list, encode with 1 if column name = genre name.

```
# check new previous vs transformed column
train.loc[:,['genres','new_genre']]
```

	genres	new_genre
0	[{'id': 35, 'name': 'Comedy'}]	[Comedy]
1	[{'id': 35, 'name': 'Comedy'}, {'id': 18, 'nam...	[Comedy, Drama, Family, Romance]
2	[{'id': 18, 'name': 'Drama'}]	[Drama]
3	[{'id': 53, 'name': 'Thriller'}, {'id': 18, 'n...	[Thriller, Drama]
4	[{'id': 28, 'name': 'Action'}, {'id': 53, 'nam...	[Action, Thriller]
...	...	...
2995	[{'id': 35, 'name': 'Comedy'}, {'id': 10749, '...	[Comedy, Romance]
2996	[{'id': 18, 'name': 'Drama'}, {'id': 10402, 'n...	[Drama, Music]
2997	[{'id': 80, 'name': 'Crime'}, {'id': 28, 'name...	[Crime, Action, Mystery, Thriller]
2998	[{'id': 35, 'name': 'Comedy'}, {'id': 10749, '...	[Comedy, Romance]
2999	[{'id': 53, 'name': 'Thriller'}, {'id': 28, 'n...	[Thriller, Action, Mystery]

3000 rows × 2 columns

Cast

- JSON format
- Used regular expressions to clean data
- Created a dictionaries:
  - Actors dictionary
  - Lead actors (1st one credited for a film)
  - Top actors
  - Count of ‘top actors’ per movie. Top actors reference by top 50 actors in the cast

```python
train.iloc[0]['cast']
```

✓ 0.1s Python

"[{'cast_id': 4, 'character': 'Lou', 'credit_id': '52fe4ee7c3a36847f82afae7', 'gender': 2, 'id': 52997, 'name': 'Rob Corddry', 'order': 0, 'profile_path': '/k2zJL0V1nEZuFT08xUdOd3ucfXz.jpg'}, {'cast_id': 5, 'character': 'Nick', 'credit_id': '52fe4ee7c3a36847f82afaeb', 'gender': 2, 'id': 64342, 'name': 'Craig Robinson', 'order': 1, 'profile_path': '/tVaRMkJXOEVhYxtnnFuhqW0Rjzz.jpg'}, {'cast_id': 6, 'character': 'Jacob', 'credit_id': '52fe4ee7c3a36847f82afaef', 'gender': 2, 'id': 54729, 'name': 'Clark Duke', 'order': 2, 'profile_path': '/oNzK0umwm5Wn0wyEbOy6TVJCSBn.jpg'}, {'cast_id': 7, 'character': 'Adam Jr.', 'credit_id': '52fe4ee7c3a36847f82afaf3', 'gender': 2, 'id': 36801, 'name': 'Adam Scott', 'order': 3, 'profile_path': '/5gb65xz8bzd42yjMAl4zwo4cvKw.jpg'}, {'cast_id': 8, 'character': 'Hot Tub Repairman', 'credit_id': '52fe4ee7c3a36847f82afaf7', 'gender': 2, 'id': 54812, 'name': 'Chevy Chase', 'order': 4, 'profile_path': '/svjpyYtPwtjvRxX9IZnOmOkhDOt.jpg'}, {'cast_id': 9, 'character': 'Jill', 'credit_id': '52fe4ee7c3a36847f82afafb', 'gender': 1, 'id': 94098, 'name': 'Gillian Jacobs', 'order': 5, 'profile_path': '/rBnhe5vhNPnhRUdtYahBWx90fJM.jpg'}, {'cast_id': 10, 'character': 'Sophie', 'credit_id': '52fe4ee7c3a36847f82afaff', 'gender': 1, 'id': 1159009, 'name': 'Bianca Haase', 'order': 6, 'profile_path': '/4x3nbtD8q8phAJPmoGWXPvz0iM.jpg'}, {'cast_id': 11, 'character': 'Kelly', 'credit_id': '5524ec51c3a3687df3000dbb', 'gender': 1, 'id': 86624, 'name': 'Collette Wolfe', 'order': 7, 'profile_path': '/aSD4h5379b2eEw3bLou9ByLimmq.jpg'}, {'cast_id': 13, 'character': 'Brad', 'credit_id': '5524ec8ec3a3687ded000d72', 'gender': 2, 'id': 466505, 'name': 'Kumail Nanjiani', 'order': 9, 'profile_path': '/x4nAztHY72SVciRfxEsbhIVTsIu.jpg'}, {'cast_id': 14, 'character': 'Courtney', 'credit_id': '5524ec9bc3a3687df8000d13', 'gender': 1, 'id': 70776, 'name': 'Kellee Stewart', 'order': 10, 'profile_path': '/w3xmsEPmJc1Cf0dQ4aIn8YmlHbk.jpg'}, {'cast_id': 15, 'character': 'Terry', 'credit_id': '5524eca892514171cb008237', 'gender': 2, 'id': 347335, 'name': 'Josh Heald', 'order': 11, 'profile_path': '/pwXJIenrDMrG7t3zNfLvr8w1RGU.jpg'}, {'cast_id': 16, 'character': 'Susan', 'credit_id': '5524ecb7925141720c001116', 'gender': 0, 'id': 1451392, 'name': 'Gretchen Koerner', 'order': 12, 'profile_path': '/muULPexCTJGyJba4yKzxronpD50.jpg'}, {'cast_id': 17, 'character': 'Herself', 'credit_id': '5524ecc3c3a3687ded000d74', 'gender': 1, 'id': 98879, 'name': 'Lisa Loeb', 'order': 13, 'profile_path': '/bGqg58ca0bZR38z9HliUMmeNGE.jpg'}, {'cast_id': 18, 'character': 'Herself', 'credit_id': '5524ecd3c3a3687e11000ed3', 'gender': 1, 'id': 1394648, 'name': 'Jessica Williams', 'order': 14, 'profile_path': '/A4syKjkcYB92wLEhH0c0hC3BCpz.jpg'}, {'cast_id': 19, 'character': 'Himself', 'credit_id': '5524ece6925141718d001009', 'gender':

Crew

- JSON format
- Used regular expressions to clean data
- Created dictionaries:
    - Directors
    - Producers
    - Executive producers
    - Number of male/female crew members

```python
train.iloc[1]['crew']
```

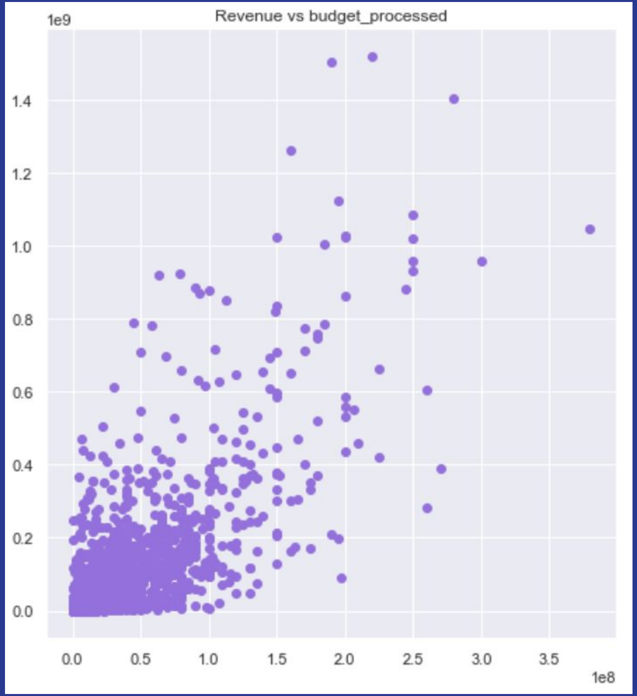
✓ 0.6s Python


```
"[{'credit_id': '52fe43fe9251416c7502563d', 'department': 'Directing', 'gender': 2, 'id': 1201, 'job': 'Director', 'name':
'Garry Marshall', 'profile_path': '/kx77E8p5rnEmKxIhFT0qWCEMEik.jpg'}, {'credit_id': '52fe43fe9251416c75025667', 'department':
'Camera', 'gender': 2, 'id': 1214, 'job': 'Director of Photography', 'name': 'Charles Minsky', 'profile_path': None},
{'credit_id': '52fe43fe9251416c75025661', 'department': 'Sound', 'gender': 2, 'id': 4500, 'job': 'Original Music Composer',
'name': 'John Debney', 'profile_path': '/hTrlvZlDXQk49nfc2BM9sjKfJv.jpg'}, {'credit_id': '52fe43fe9251416c7502564f',
'department': 'Production', 'gender': 1, 'id': 8851, 'job': 'Producer', 'name': 'Whitney Houston', 'profile_path':
'/69ouDnXnmklYPr4sMJXWKYz81AL.jpg'}, {'credit_id': '52fe43fe9251416c7502566d', 'department': 'Editing', 'gender': 0, 'id':
12970, 'job': 'Editor', 'name': 'Bruce Green', 'profile_path': '/yplxWPVgwK1b33AjvbhM9mWX2Aw.jpg'}, {'credit_id':
'52fe43fe9251416c75025655', 'department': 'Production', 'gender': 2, 'id': 38415, 'job': 'Producer', 'name': 'Mario Iscovich',
'profile_path': None}, {'credit_id': '52fe43fe9251416c7502565b', 'department': 'Production', 'gender': 1, 'id': 38416, 'job':
'Executive Producer', 'name': 'Ellen H. Schwartz', 'profile_path': '/6WInjABr1sAYGXaa5qOvSrsHIqP.jpg'}, {'credit_id':
'52fe43fe9251416c75025649', 'department': 'Production', 'gender': 1, 'id': 59973, 'job': 'Producer', 'name': 'Debra Martin
Chase', 'profile_path': None}, {'credit_id': '52fe43fe9251416c75025643', 'department': 'Writing', 'gender': 1, 'id': 25539,
'job': 'Screenplay', 'name': 'Shonda Rhimes', 'profile_path': '/4c77e347InbTAlw9lGvORpZBHV6.jpg'}]"
```

EDA Questions

- **How does budget influence revenue?**
- What is the average revenue for each genre?
- Revenue of movies with home page vs no homepage?
- Did production companies have a major influence on revenue?
- Which production company had the highest average revenue?

Findings:

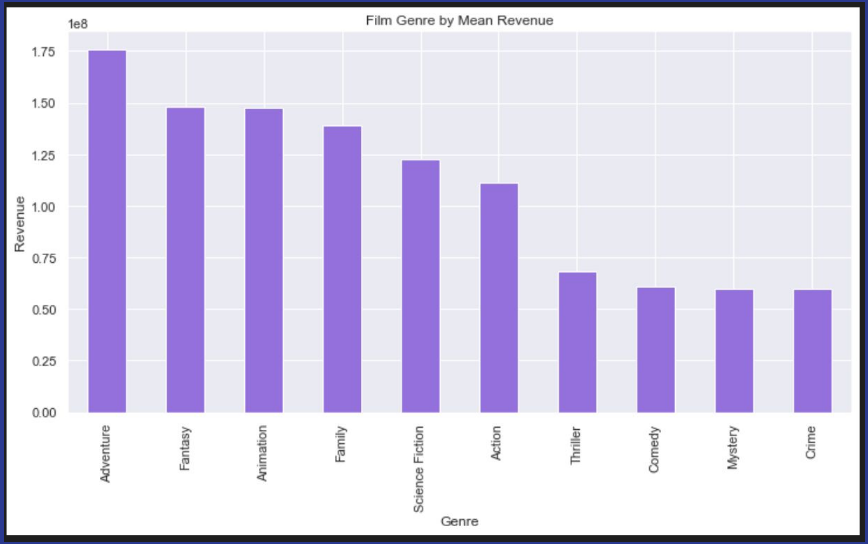


There is a strong correlation

EDA Questions

- How does budget influence revenue?
- **What is the average revenue for each genre?**
- Revenue of movies with home page vs no homepage?
- Did production companies have a major influence on revenue?
- Which production company had the highest average revenue?

Findings:



Adventure films have the highest revenue values

EDA Questions

- How does budget influence revenue?
- What is the average revenue for each genre?
- **Revenue of movies with home page vs no homepage?**
- Did production companies have a major influence on revenue?
- Which production company had the highest average revenue?

Findings:

Mean Revenue

Movies with a homepage:
$120,051,698

No homepage:
$42,165,846

EDA Questions

- How does budget influence revenue?
- What is the average revenue for each genre?
- Revenue of movies with home page vs no homepage?
- **Did production companies have a major influence on revenue?**
- Which production company had the highest average revenue?

Findings:

Based of the Pearson correlation:

- The number of top studios a movie uses in relation to revenue are strongly correlated.
    - r = 0.558

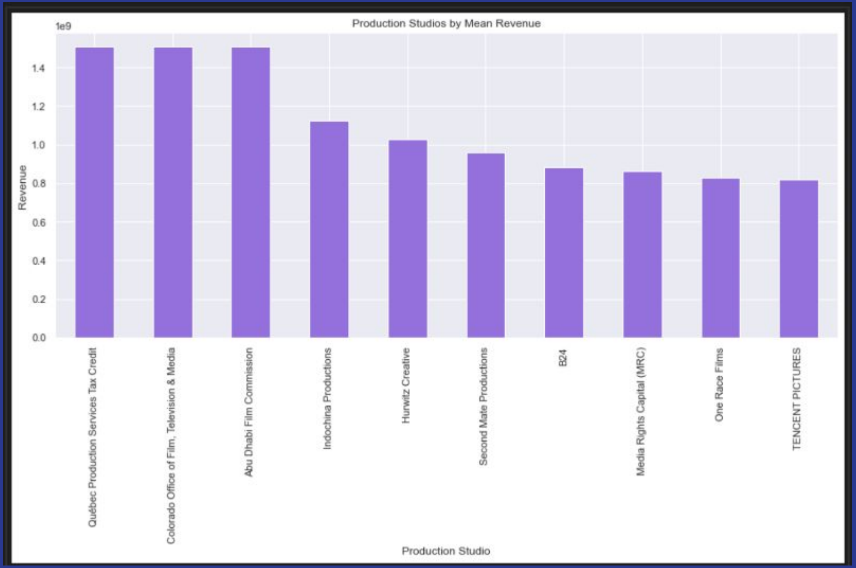
Note:

- Top studios typically have more reputation, funds and resources that benefits the movie overall

EDA Questions

- How does budget influence revenue?
- What is the average revenue for each genre?
- Revenue of movies with home page vs no homepage?
- Did production companies have a major influence on revenue?
- **Which production company had the highest average revenue?**

Findings:

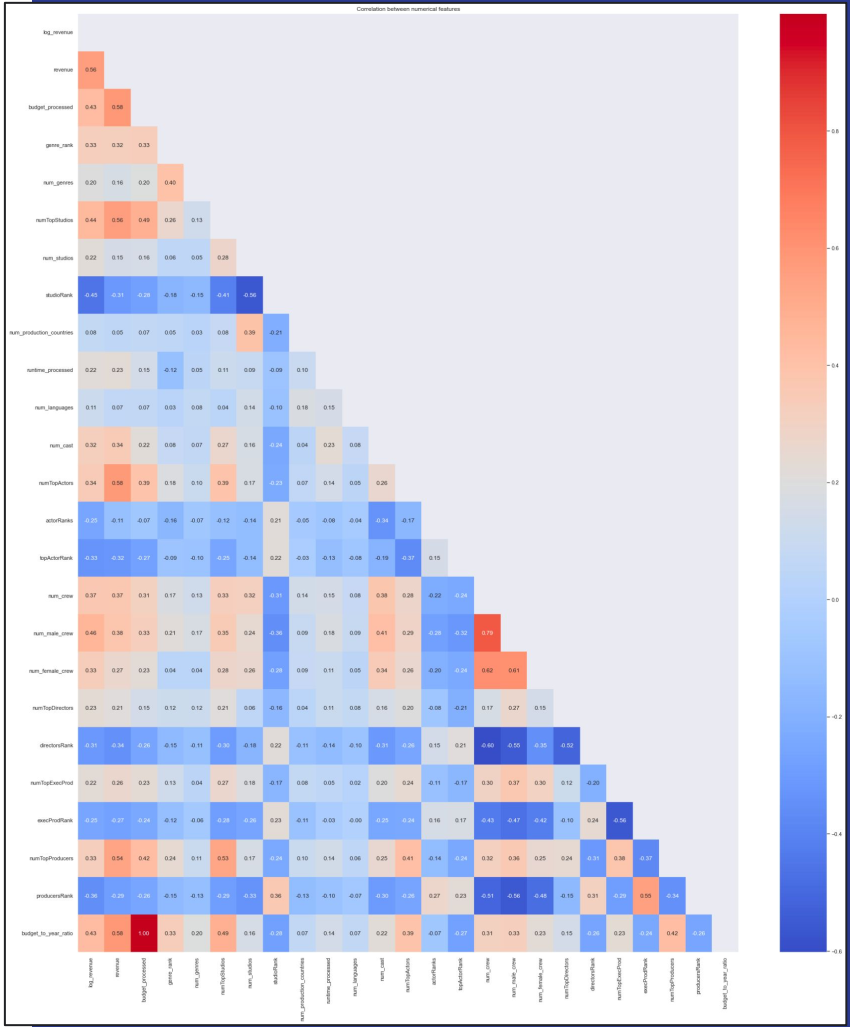


Machine Learning Models

Selecting Features

Numerical Correlation Matrix

- Set 1: Categorical Inputs
  - Genre
  - Studio
  - Collection
  - Transformed into binary
- Set 2: Quantitative Inputs
  - Budget
  - Runtime
  - Popularity
- Set 3: Transforming Features
  - Log Transformation



Feature Importances
Target: Revenue

- Uses feature of importance to rank relevant metrics
- Some of the features that ranked top across the four different models are similar.

```
feature_importances = rfr_best_model.feature_importances_
indices = np.argsort(feature_importances)[::-1]

# Print the feature ranking
print("Feature ranking:")

for f in range(20):
    print("%d. %s (%f)" % (f + 1, X_train.columns[indices[f]], feature_importances[indices[f]]))
```

✓ 0.4s

```
Feature ranking:
1. topStudio (0.361790)
2. usa_produced (0.117028)
3. belongs_to_collection (0.109838)
4. topLeadActor (0.097712)
5. has_homepage (0.060809)
6. fridayRelease (0.028559)
7. topLeadDirector (0.028427)
8. released_in_english (0.024508)
9. originally_english (0.020992)
10. Fall (0.018236)
11. Winter (0.018028)
12. wednesdayRelease (0.018021)
13. thursdayRelease (0.016198)
14. topLeadProducer (0.015913)
15. Summer (0.014740)
16. saturdayRelease (0.014097)
17. Spring (0.011134)
```

Random Forest Regression

Why we chose this model.

- Ensemble method
- It can perform both regression and classification tasks.
- It can handle large datasets efficiently.
- Provides a higher level of accuracy in predicting outcomes over the decision tree algorithm.

Extra Trees

Why we chose this model.

- Faster algorithm in terms of computation

Note:

- Random forest uses bootstrap replicas; it subsamples the input data with replacement, whereas Extra Trees use the whole original sample.

XGBoost Model

Why we chose this model.

- Scalable and highly accurate implementation of gradient boosting for boosted tree algorithms
- Built largely for energizing machine learning model performance and computational speed.

LightGBM

Why we chose this model.

- It’s a distributed and efficient gradient boosting framework that uses tree-based learning.
- Histogram-based algorithm
- Places continuous values into discrete bins
    - Leads to faster training and more efficient memory usage.

Machine Learning Results

Random Tree Forest

Base Model:
Average Error: 1.5369
Accuracy = 89.478%

Model after Tuning:
Average Error: 1.3919
Accuracy = 90.394%
Improvement of 1.02%.

Extra Trees

Base Model:
Average Error: 1.6353
Accuracy = 88.852%

Model after Tuning:
Average Error: 1.4354
Accuracy = 90.139%
Improvement of 1.45%.

XGBoost Model

Base Model:
Average Error: 1.5313
Accuracy = 89.565%

Model after Tuning:
Average Error: 1.3809
Accuracy = 90.497%
Improvement of 1.041%.

LightGBM

Base Model:
Average Error: 1.4221
Accuracy = 90.214%

Model after Tuning:
Average Error: 1.3812
Accuracy = 90.481%
Improvement of 0.30%.

ML Model Conclusion

- XGBoost model performed the best after tuning.
- LightGBM model had the best "base" model of the four.
- The Machine Learning models implemented to predict revenue in this analysis were off by $41M-$43M.

Dashboard Presentation

GitHub

Tableau