Final Project

Predicting Film Profitability / Genre

JOSH STABINSKY | DAT-SF-60 | DECEMBER 2019

Agenda

- Problem statement.
- Metrics and assumptions.
- Approach and process.
- The model(s).
- Performance evaluation.
- Impact of your findings.
- Recommendations / next steps.

CAN WE ACCURATELY PREDICT BOTH THE PROFITABILITY AND GENRE OF A GIVEN FILM GIVEN LIMITED KNOWLEDGE ABOUT THE MOVIE?

Why do we care?

Limited information

Movie studios generally have limited information when funding a film.

Big $$$ involved

We’re talking about potentially massive budgets here, could be a huge competitive advantage.

Increasing competition

More and more studios are funding successful films and there’s increasing competition from streaming services.

Why do we care?

Limited information

Movie studios generally have limited information when funding a film.

Big $$$ involved

We're talking about potentially massive budgets here, could be a huge competitive advantage .

Increasing competition

More and more studios are funding successful films and there's increasing competition from streaming services.

Why do we care?

Limited information

Movie studios generally have limited information when funding a film.

Big $$$ involved

We're talking about potentially massive budgets here, could be a huge competitive advantage.

Increasing competition

More and more studios are funding successful films and there's increasing competition from streaming services.

METRICS & ASSUMPTIONS

Metrics & Assumptions

Data source(s)

Data provided by kaggle & a site on inflation data.

Assumptions

Many variables are unpredictable for an industry like this. We can’t predict how the economy will look at any given point, we can’t predict celebrity scandals, etc.

Dataset(s) description

See next slide.

Metrics & Assumptions

Data source(s)

Data provided by kaggle & a site on inflation data.

Assumptions

Many variables are unpredictable for an industry like this. We can’t predict how the economy will look at any given point, we can’t predict celebrity scandals, etc.

Dataset(s) description

See next slide.

Metrics & Assumptions

Data source(s)

Data provided by kaggle & a site on inflation data.

Assumptions

Many variables are unpredictable for an industry like this. We can’t predict how the economy will look at any given point, we can’t predict celebrity scandals, etc.

Dataset(s) description

See next slide.

Dataset(s)

Kaggle movie data*

- 5,000 films
- as old as 1927
- some fields are super messy
- fields include:
  - budget
  - genres
  - revenue
  - runtime, etc.

Inflation data**

- I needed a way to account for inflation
- This data set is anchored on the year 1990

*https://www.kaggle.com/tmdb/tmdb-movie-metadata
**https://www.in2013dollars.com/us/inflation/1990

Dataset(s)

Kaggle movie data*

- 5,000 films
- as old as 1927
- some fields are super messy
- fields include:
  - budget
  - genres
  - revenue
  - runtime, etc.

Inflation data**

- I needed a way to account for inflation
- This data set is anchored on the year 1990

*https://www.kaggle.com/tmdb/tmdb-movie-metadata
**https://www.in2013dollars.com/us/inflation/1990

APPROACH & PROCESS

Approach & Process

Step 1
Cleaning the data.

Step 2
Linear regression to predict profitability.

Step 3
NLTK model to classify plot summaries into genres. Can we simply take a plot summary and budget and predict box office success?

Approach & Process

Step 1

Cleaning the data.

Step 2

Linear regression to predict profitability.

Step 3

NLTK model to classify plot summaries into genres. Can we simply take a plot summary and budget and predict box office success?

Approach & Process

Step 1

Cleaning the data.

Step 2

Linear regression to predict profitability.

Step 3

NLTK model to classify plot summaries into genres. Can we simply take a plot summary and budget and predict box office success?

Data Cleaning

1. `genre`

<u>ORIGINAL</u>: [{“id”: 28, "name": "Action"}, {"id": 12, "name": "Adventure"}, {"id": 14, “name": "Fantasy"}, {"id": 878, "name": "Science Fiction”}]

<u>JSON</u>: {"28": "Action", "53": "Thriller"}

2. `release_date > 1990`

3. `dropna for genre`

THE MODEL(S)

The Model(s)

Now that our data is clean...

Linear Regression

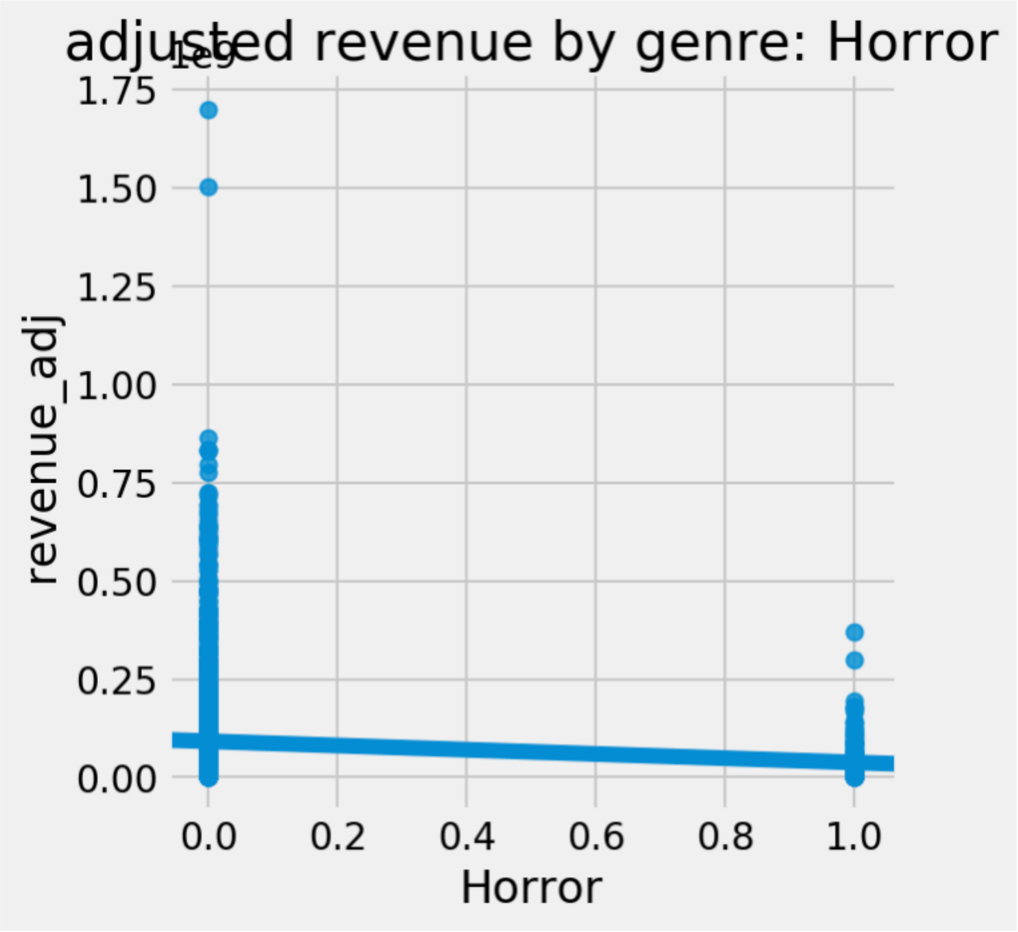
Results on subsequent slide...

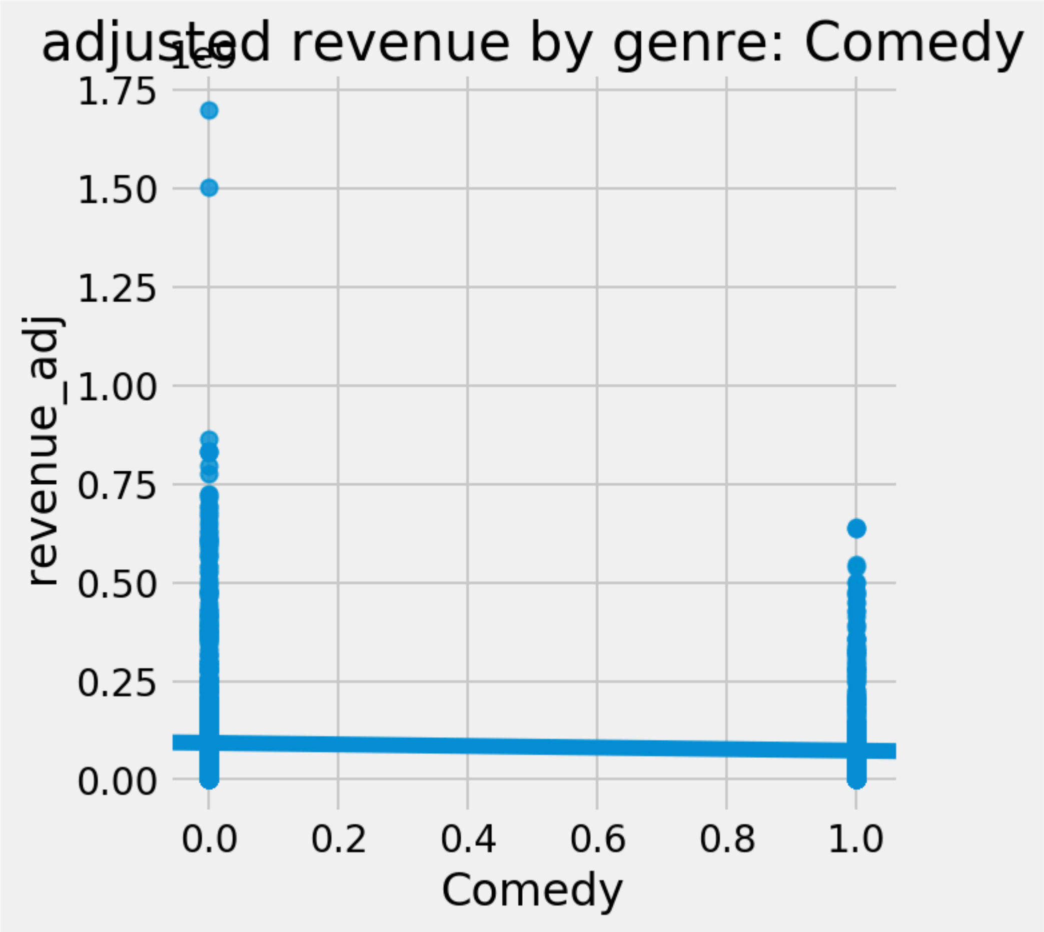
NLTK Model

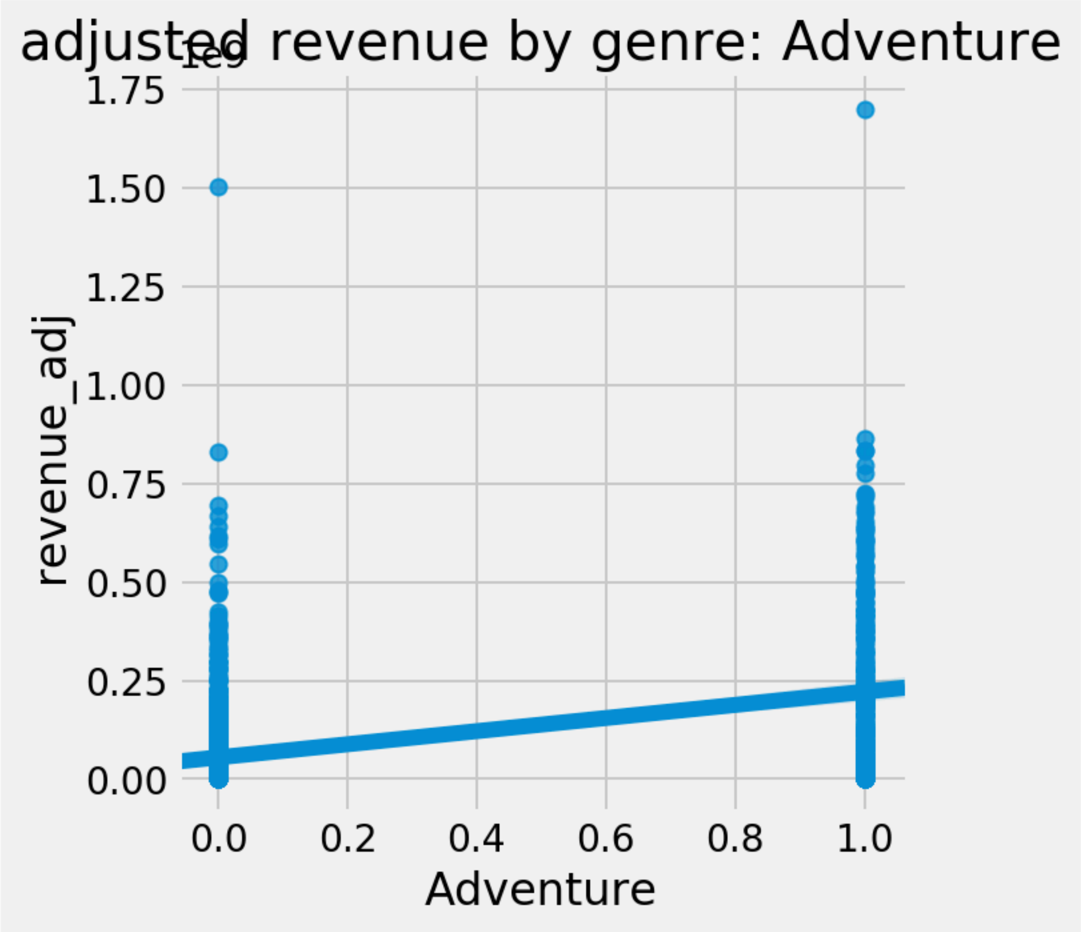
Results on subsequent slide...

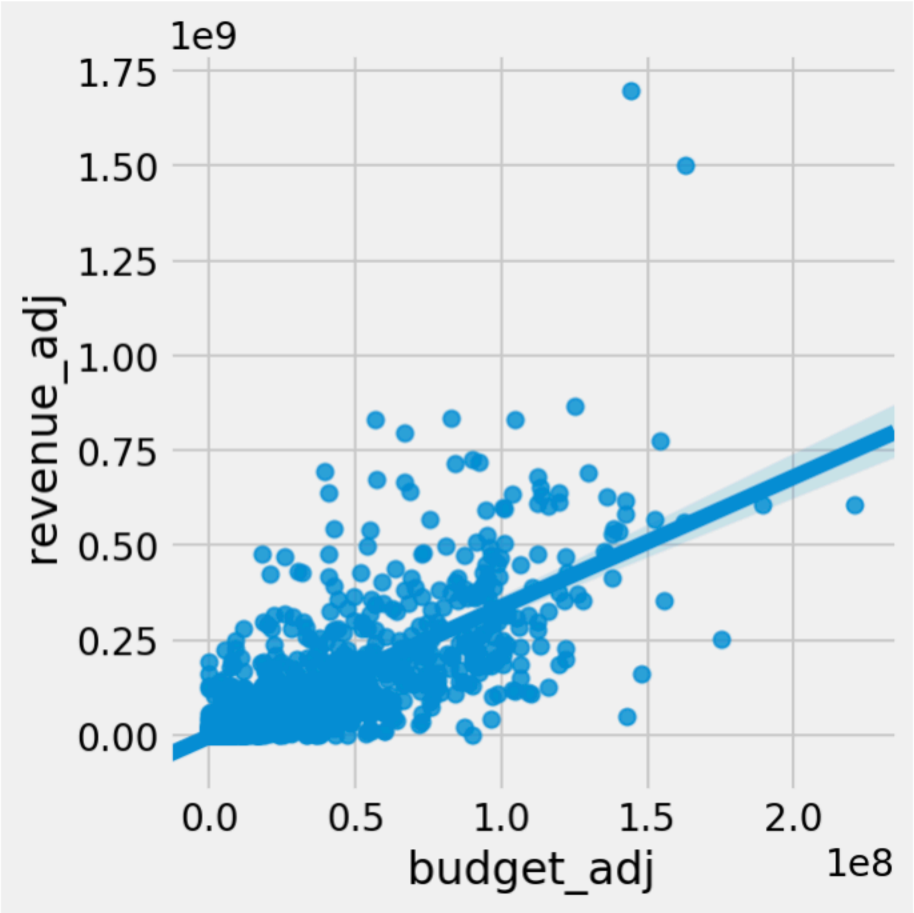
Linear Regression

Correlations









Linear Regression

Final Revenue Model

intercept = 8071901.818 +
(drama * (-6100952.99)) +
(comedy * (-8062230.03)) +
(action * (-12931046.78)) +
(adventure * (22722367.89)) +
(horror * (1082686.49)) +
(crime * (-8279780.70)) +
(thriller * (-6193194.83)) +
(animation * (19905632.67)) +
(fantasy * (-662322.66)) +
(romance * (6267969.47)) +
(science_fiction * (-7565116.035)) +
(documentary * (-3546676.93)) +
(family * (-10783390.91)) +
(mystery * (-5184690.41)) +
(music * (-10828096.09)) +
(western * (-56349763.73)) +
(history * (-24117763.83)) +
(war * (-4336613.48)) +
(tv_movie * (2500026.84)) +
(foreign * (-1505223.62) +
(budget_adj * 3.31))

Final Profit Model

intercept = 8071901.818 +
(drama * (-6100952.99)) +
(comedy * (-8062230.03)) +
(action * (-12931046.78)) +
(adventure * (22722367.89)) +
(horror * (1082686.49)) +
(crime * (-8279780.70)) +
(thriller * (-6193194.83)) +
(animation * (19905632.67)) +
(fantasy * (-662322.66)) +
(romance * (6267969.47)) +
(science_fiction * (-7565116.035)) +
(documentary * (-3546676.93)) +
(family * (-10783390.91)) +
(mystery * (-5184690.41)) +
(music * (-10828096.09)) +
(western * (-56349763.73)) +
(history * (-24117763.83)) +
(war * (-4336613.48)) +
(tv_movie * (2500026.84)) +
(foreign * (-1505223.62) +
(budget_adj * 3.31))
- budget_adj

example

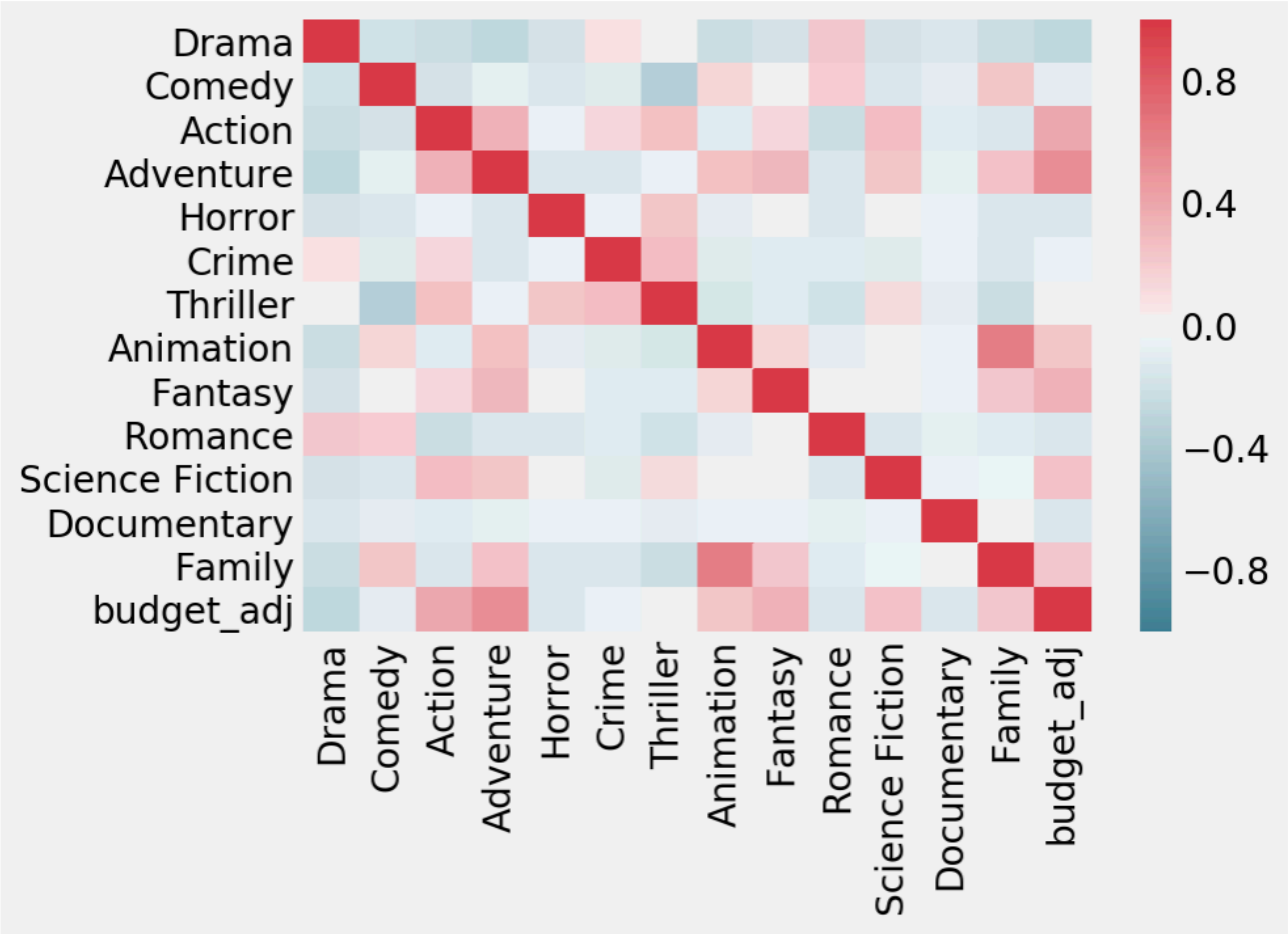
a family animated film with a massive budget (like Frozen):

$8,071,901.82 + (1 * ($19,905,632.67)) + (1 * (-$10,783,390.91)) + ($150,000,000 * 3.31)

```
predicted: $519M
actual: $1.3B
```

Linear Regression

multicollinearity?



RMSE optimized with everything included.

The Model(s)

Now that our data is clean...

Linear Regression

Results on subsequent slide...

NLTK Model

Results on subsequent slide...

NLTK Model

Process

- Ensure data is clean
  - JSON formatting for `genre` field
  - Need to remove stop words from the `plot_summary` field
  - X most frequent words in plot summaries used as features

  - 80/20 split

  - "*OneVsRestClassifier* class to solve this problem as a Binary Relevance or one-vs-all problem"
    - default 50%

https://www.analyticsvidhya.com/blog/2019/04/predicting-movie-genres-nlp-multi-label-classification/

NLTK Model

In [504]:
```
# evaluate performance
f1_score(yval, y_pred, average="micro")
```
Out[504]:
```
0.30541012216404884
```

In [516]:
```
t_list = [.1, .2, .3, .4, .5, .6, .7, .8, .9]

for t_value in t_list:
    t = t_value # threshold value
    y_pred_new = (y_pred_prob >= t).astype(int)
    print('f1 score when threshold =', t_value, '--', f1_score(yval, y_pred_new, average="micro"))
```
```
f1 score when threshold = 0.1 -- 0.42963155583908345
f1 score when threshold = 0.2 -- 0.5448103376406837
f1 score when threshold = 0.3 -- 0.5367215861491205
f1 score when threshold = 0.4 -- 0.4566371681415929
f1 score when threshold = 0.5 -- 0.30541012216404884
f1 score when threshold = 0.6 -- 0.13883299798792756
f1 score when threshold = 0.7 -- 0.0416221985058698
f1 score when threshold = 0.8 -- 0.006535947712418301
f1 score when threshold = 0.9 -- 0.0
```

In [518]:
```
y_pred_new[2]
```
Out[518]:
```
array([0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0])
```

In [519]:
```
y_pred[2]
```
Out[519]:
```
array([0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0])
```

In [521]:
```
yval[2]
```
Out[521]:
```
array([0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0])
```

In [522]:
```
multilabel_binarizer.inverse_transform(yval)[2]
```
Out[522]:
```
('Comedy', 'Drama', 'Family', 'Romance')
```

In [523]:
```
multilabel_binarizer.inverse_transform(y_pred)[2]
```
Out[523]:
```
('Comedy',)
```

In [524]:
```
multilabel_binarizer.inverse_transform(y_pred_new)[2]
```
Out[524]:
```
('Comedy', 'Drama', 'Romance', 'Thriller')
```

https://www.analyticsvidhya.com/blog/2019/04/predicting-movie-genres-nlp-multi-label-classification/

PERFORMANCE EVALUATION

Performance Evaluation

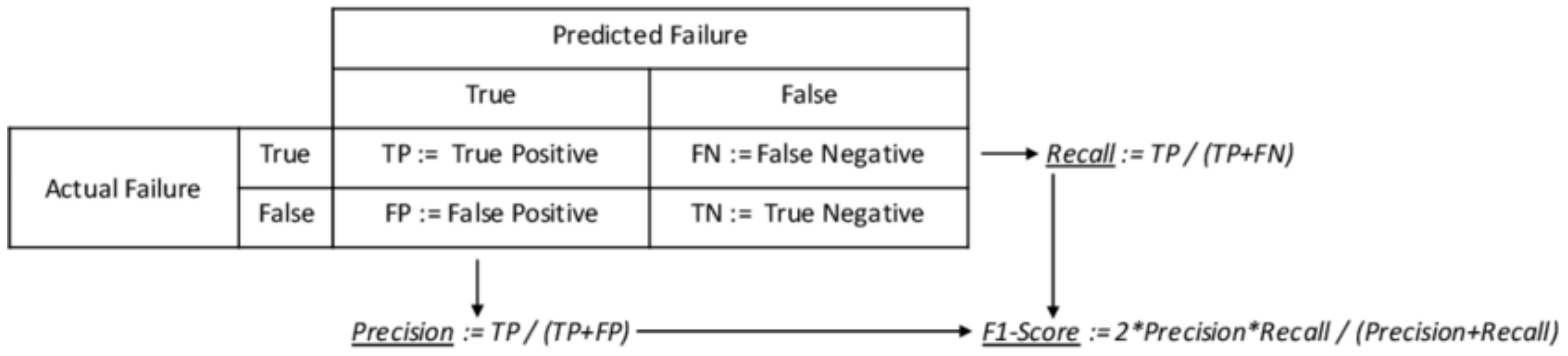
Linear Regression

- not horrible, but inherently flawed
  - (intercept and budjet coefficient)

NLTK Model

- Maxing out at an F1 score of .54
- Precision specifically is really low (.21)
- (too many false positives)
- “Model is fairly accurate, but could be better."

Performance Evaluation



Linear Regression

- not horrible, but inherently flawed
  - (intercept and budjet coefficient)

NLTK Model

- Maxing out at an F1 score of .54
- Precision specifically is really low (.21)
- (too many false positives)
- “Model is fairly accurate, but could be better."

IMPACT

Impact

“Good not great”

Models could be better, no studio in their right mind would use them to make decisions.

Still pretty cool!

Way more accurate than just randomly guessing.

Could be useful in real world

If someone with a ton of data science experience built this model out.

RECOMMENDATIONS / NEXT STEPS

Recommendations / Next Steps

Don’t use this model.

It simply isn’t robust enough for real world decision making.

We need more data.

The data set was just generally too limited and the model was inherently flawed.

I'd look at casts next if I had the time.

But I suspect that’s highly correlated with budget and might not actually help much.

Recommendations / Next Steps

Don’t use this model.

It simply isn’t robust enough for real world decision making.

We need more data.

The data set was just generally too limited and the model was inherently flawed.

I'd look at casts next if I had the time.

But I suspect that’s highly correlated with budget and might not actually help much.

Recommendations / Next Steps

Don’t use this model.

It simply isn’t robust enough for real world decision making.

We need more data.

The data set was just generally too limited and the model was inherently flawed.

I'd look at casts next if I had the time.

But I suspect that’s highly correlated with budget and might not actually help much.

Questions?