Filmmaking Is A Calculated Art

Can You Determine A Movie's Success?

Presented By:
SC1015 - Group 3
- Que An Tran
- Duc Anh Do
- Justin Tan

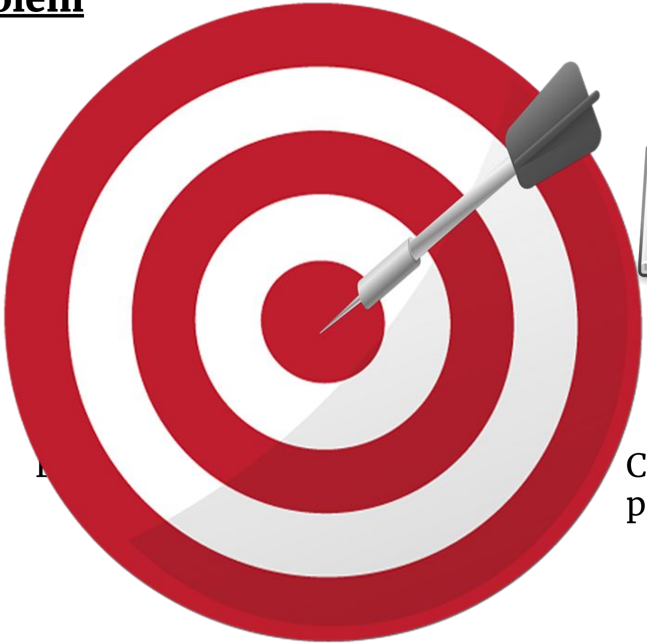
Problem Statement

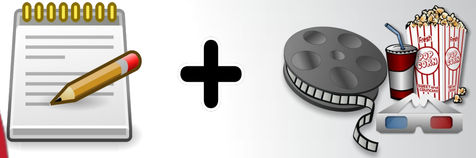
- Concrete evidences of failures and successes in movies & films
- Is it possible to predict the success of a movie based on various parameters?
- We deem the success and failures as:
  - Success → Revenue - Budget = Positive
  - Failure → Revenue - Budget = Negative

The Dataset

- ‘The Movies Dataset’ created by TMDB and GroupLens
- Metadata of 45,000 movies released on or before July 2017
- Ratings from 270,000 users
- Contains 6 CSV files

Breaking Down The Problem





Pre-Element

Pre-production variables

- Budget
- Num_cast
- Num_crew

Pre & Post Elements

Combination of Pre & Post production variables

The Movies Dataset

Credits.csv

Summary:

movie_id : id of the movie
title : title of the movie
cast : information of cast in the movie
crew : information of cast in the movie
There are 4803 records

Movies.csv

Summary:
¶

Budget : budget of the company
genres : show genres of the movie
homapage : link to the homepage
id : id of the movie
keywords : show keywords of the data
original_language : show the language of the movie
original_title : original title of the movie
overview : overview of the movie
popularity : popularity rate of the movie
production_companies : name of production company
production_countries : name of production countries
release_date : date of release
revenue : revenue of the movie
runtime : runtime of the movie
spoken_languages : list of spoken language
status : status of the movie (ex. released)
tagline : tagline of the movie
vote_average : average of vote grade
vote_count : number of vote

Cleaning The Dataset

- Selection of relevant columns
  - Budget, Revenue, Num_cast, Num_crew, Popularity, Vote_count, Vote_average

- Check for missing values
  - Missing values will be filled with default values (Null or 0)

- Perform One-Hot Encoding on categorical variables
  - ```
    genres_d.head()
    ```

	drama	animation	foreign	fantasy	horror	thriller	sciencefiction	tvmovie	comedy	romance	family	music	history	action	documentary	war	western	a
0	0	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0	0	
1	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	
2	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	
3	1	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	
4	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	

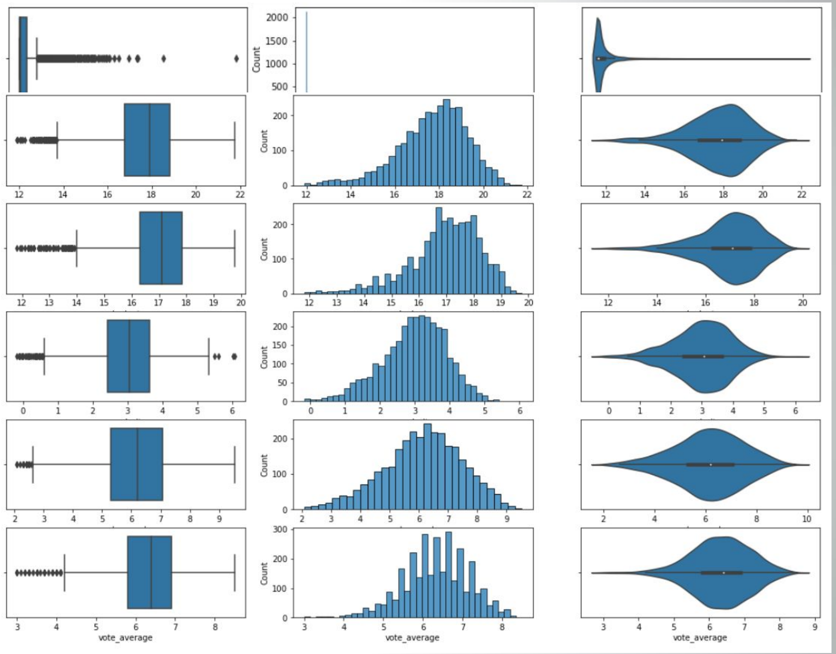
- Concatenate relevant variables into a single dataframe
    ```

Preparing The Dataset

- Normalise the distribution of the variables

- Remove outliers
  - Remove variables containing 0
  - log(0) leads to infinity

- Measure the Skewness & Kurtosis of Data
  - Clearly positively skewed
  - Use log scaling from SKLEARN

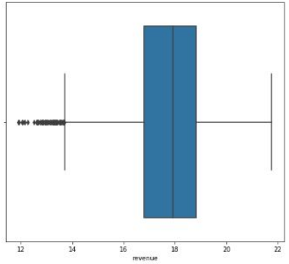


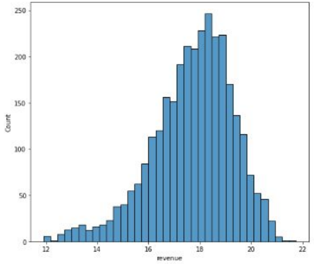
Exploratory Data Analysis

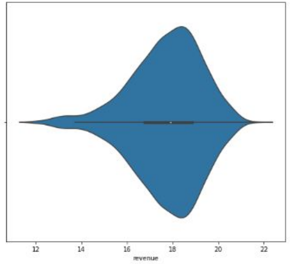
- Statistics of relevant variables
  - Data is normalised
  - Forms a normal distribution

	budget	revenue	num_cast	num_crew	popularity	vote_count	vote_average
count	3098.000000	3098.000000	3098.000000	3098.000000	3098.000000	3098.000000	3098.000000
mean	16.949055	17.714582	26.576501	34.863460	2.986412	6.131337	6.325339
std	1.273891	1.576674	21.665511	35.371232	0.931786	1.344940	0.844225
min	11.805632	11.894112	0.000000	1.000000	-0.179585	2.079442	3.000000
25%	16.300417	16.782571	15.000000	12.000000	2.420102	5.299564	5.800000
50%	17.111347	17.909855	20.000000	21.000000	3.055700	6.210600	6.400000
75%	17.854137	18.832300	31.000000	45.000000	3.640710	7.075809	6.900000
max	19.755682	21.748578	224.000000	435.000000	6.073686	9.528940	8.500000

- Revenue chart





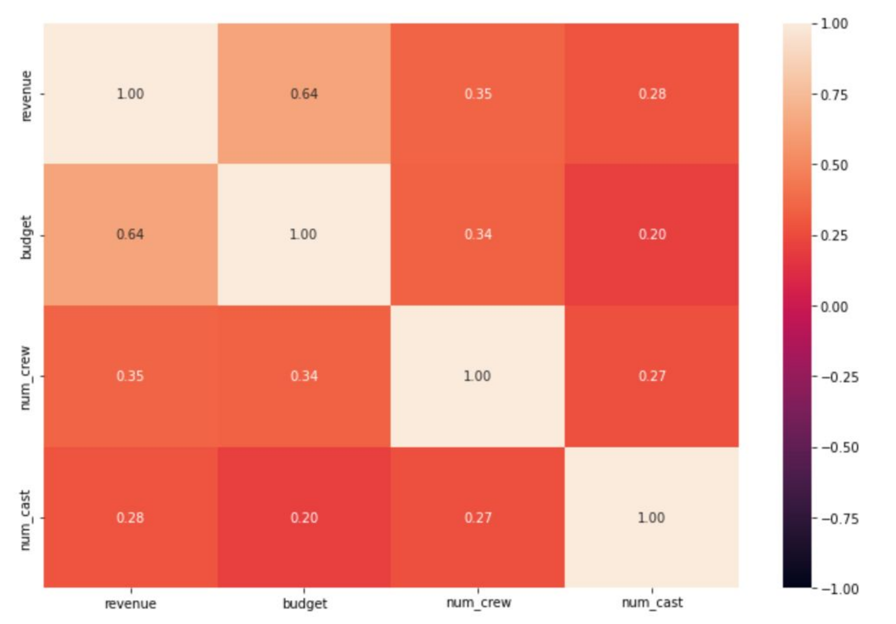


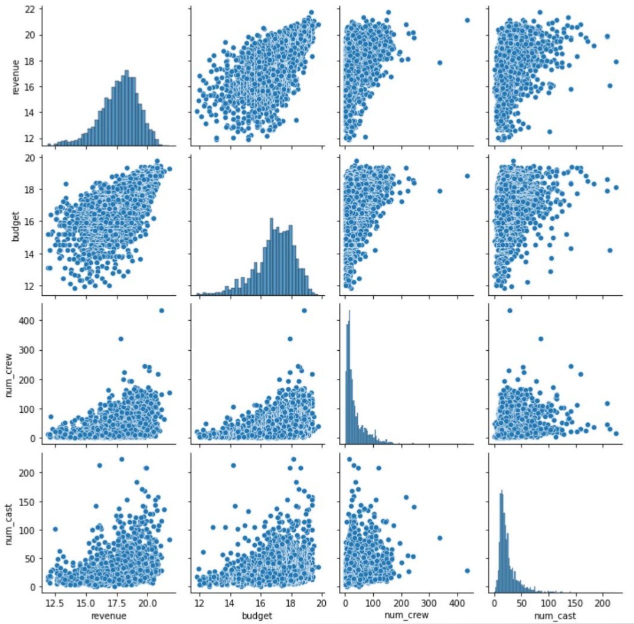
Multivariate Exploratory Analysis

- Presence of correlation with revenue:
  - Pre-elements:
    - Num_cast
    - Num_crew
    - Budget
  - Post-elements:
    - Popularity
    - Vote_count
    - Vote_average

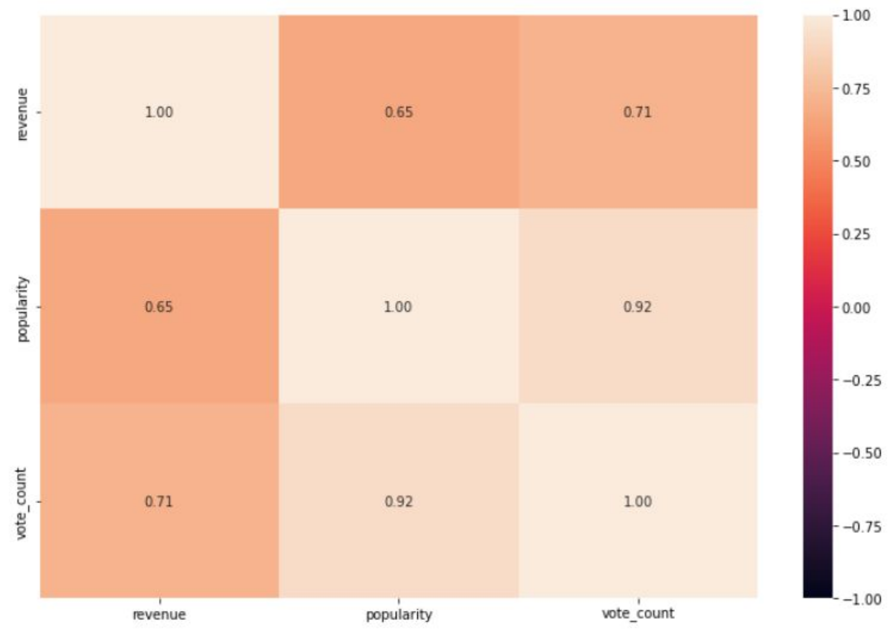
sciencefiction	-0.204297
music	-0.084818
romance	-0.071965
fantasy	-0.070136
mystery	-0.067081
thriller	-0.057001
tvmovie	-0.034417
family	-0.026968
foreign	-0.004931
adventure	-0.004499
animation	-0.002778
war	0.010767
documentary	0.095733
vote_average	0.125916
drama	0.161806
comedy	0.162234
action	0.169732
crime	0.196366
horror	0.262147
num_cast	0.282160
num_crew	0.349847
budget	0.638766
popularity	0.652001
vote_count	0.707789
revenue	1.000000

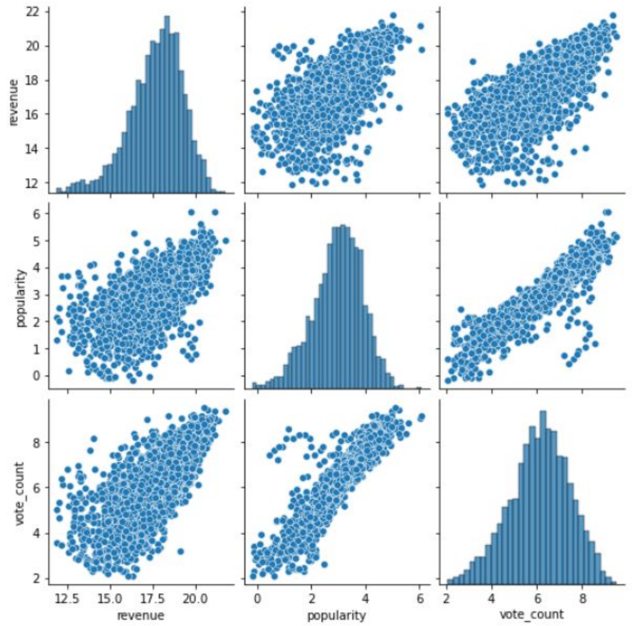
Multivariate Exploratory Analysis: Pre-elements





Multivariate Exploratory Analysis: Post-elements





Building our Models based on Data

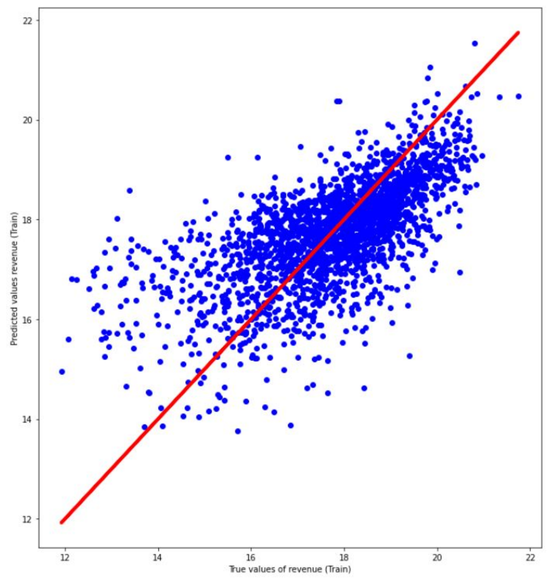
- **Model 1 - Linear Regression**
    - Supervised Learning Model
    - Existence of multiple independent variables
    - Fits a straight line or surface
    - Minimizes discrepancies between predicted and output values

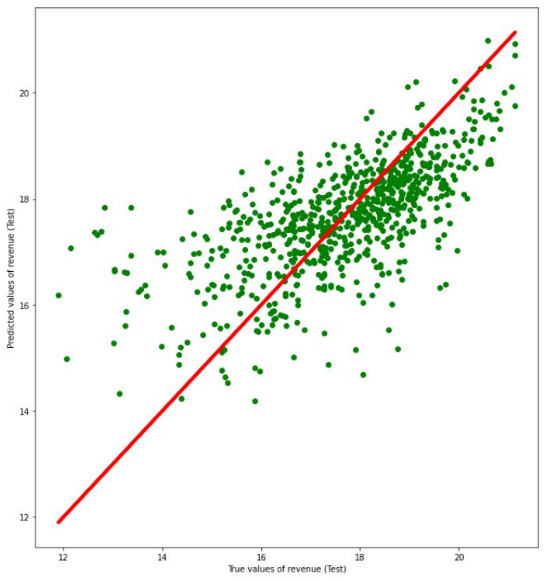
- **Model 2 - Extreme Gradient Boosting (XGBoost)**
    - Improve speed and performance
        - Consists of an ensemble machine learning algorithms
        - Parallelizable and takes advantage of multi-core machines
    - Feasible to train on large datasets

Linear Regression on Pre-Production Elements

Training Data Set

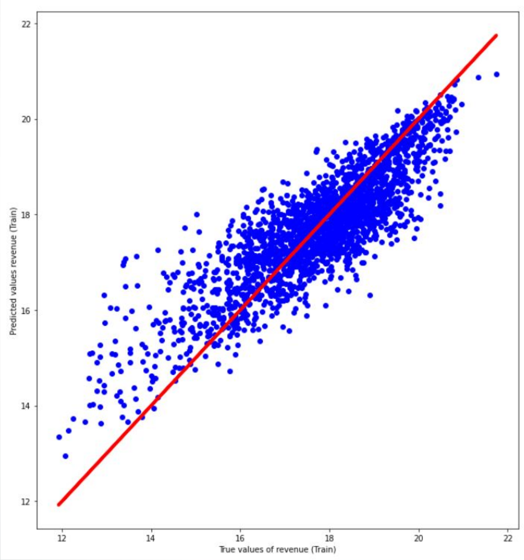
Test Data Set



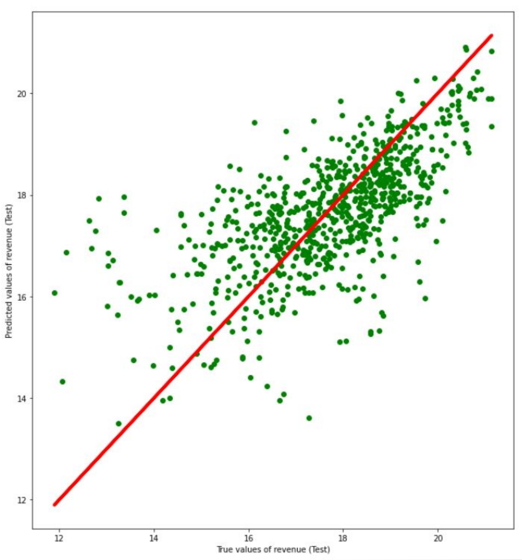


XGBoost on Pre-Production Elements

Training Data Set



Test Data Set



Linear Regression vs XGBoost (Pre-Production Elements)

Linear Regression

```
Goodness of Fit of Model           Train Dataset
Explained Variance (R^2)           : 0.4456528538387208
Mean Squared Error (MSE)           : 1.3372998265889882

Goodness of Fit of Model           Test Dataset
Explained Variance (R^2)           : 0.4474267030299425
Mean Squared Error (MSE)           : 1.4928450364721926
```

XGBoost

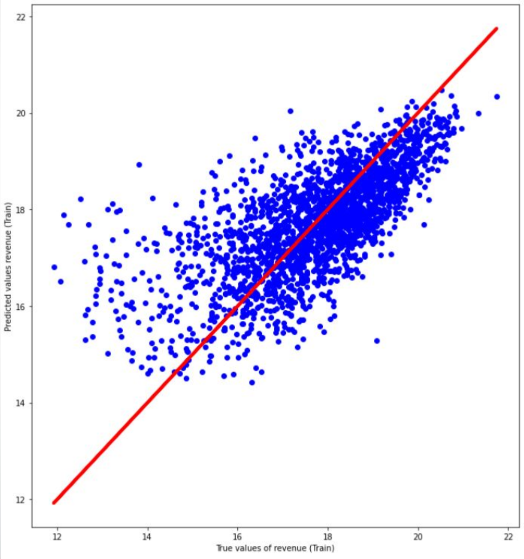
```
Goodness of Fit of Model           Train Dataset
Explained Variance (R^2)           : 0.731856776268491
Mean Squared Error (MSE)           : 0.646865216282422

Goodness of Fit of Model           Test Dataset
Explained Variance (R^2)           : 0.434140502800275
Mean Squared Error (MSE)           : 1.528739348055018
```

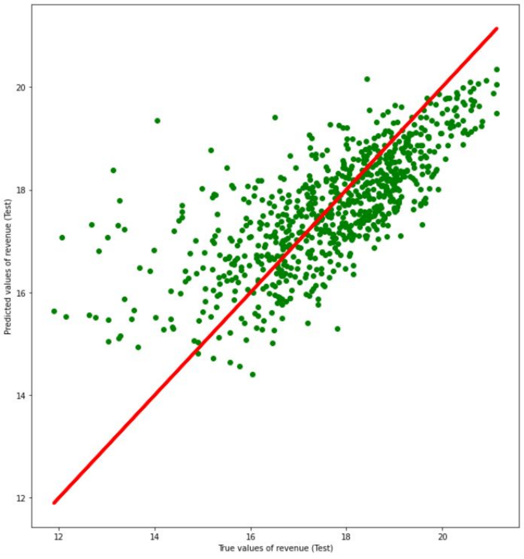
- Visible improvement in performance
- Not a significant improvement
- Not a strong indicator for success

Linear Regression on Post-Production Elements

Training Data Set

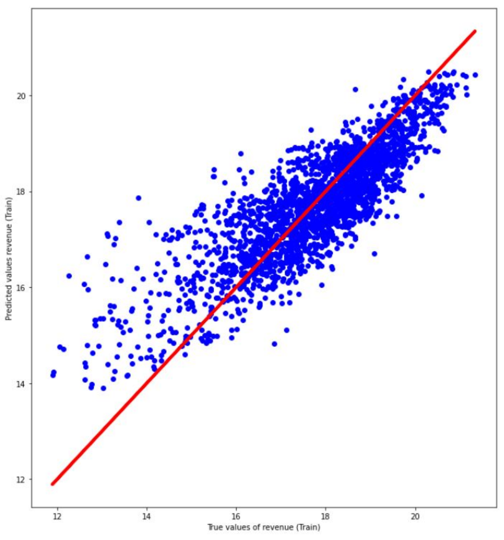


Test Data Set

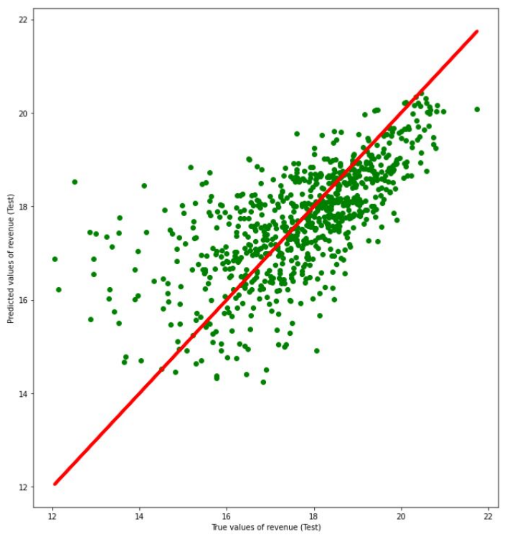


XGBoost on Post-Production Elements

Training Data Set



Test Data Set



Linear Regression vs XGBoost (Post-Production Elements)

Linear Regression

```
Goodness of Fit of Model         Train Dataset
Explained Variance (R^2)         : 0.48844819672590856
Mean Squared Error (MSE)         : 1.234060899468757

Goodness of Fit of Model         Test Dataset
Explained Variance (R^2)         : 0.533097817053046
Mean Squared Error (MSE)         : 1.2613939366095734
```

XGBoost

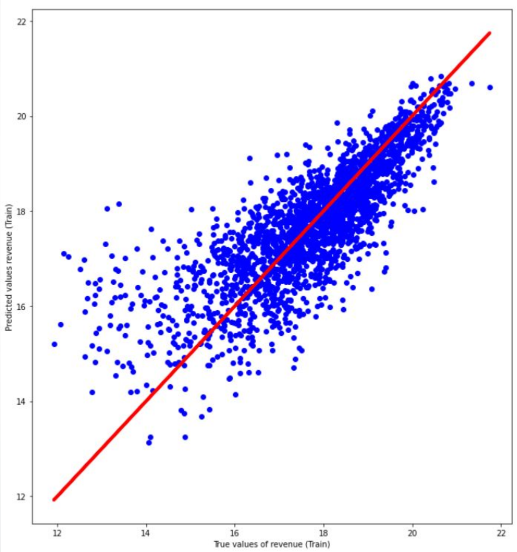
```
Goodness of Fit of Model         Train Dataset
Explained Variance (R^2)         : 0.6832158198980429
Mean Squared Error (MSE)         : 0.7642060251415664

Goodness of Fit of Model         Test Dataset
Explained Variance (R^2)         : 0.4942346588549833
Mean Squared Error (MSE)         : 1.366387560325624
```

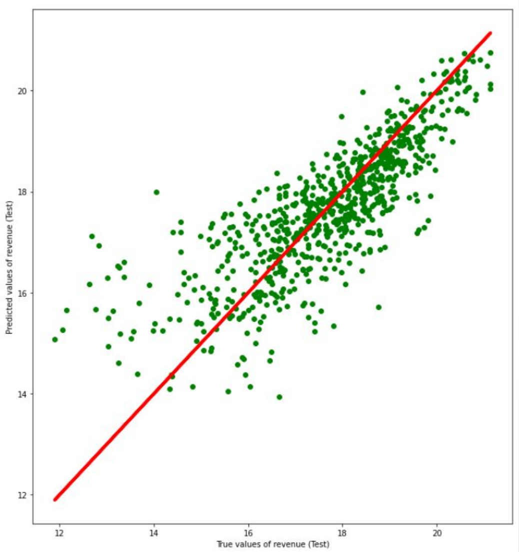
- Better results as compared to pre-production
- Slight improvement overall
- Room for improvement?

Linear Regression Pre + Post Production Elements

Training Data Set

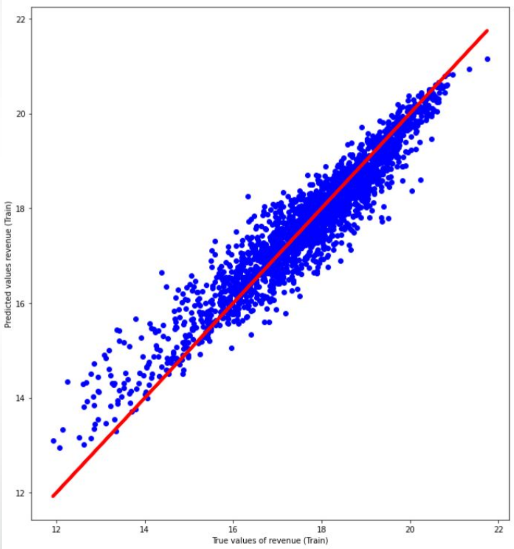


Test Data Set

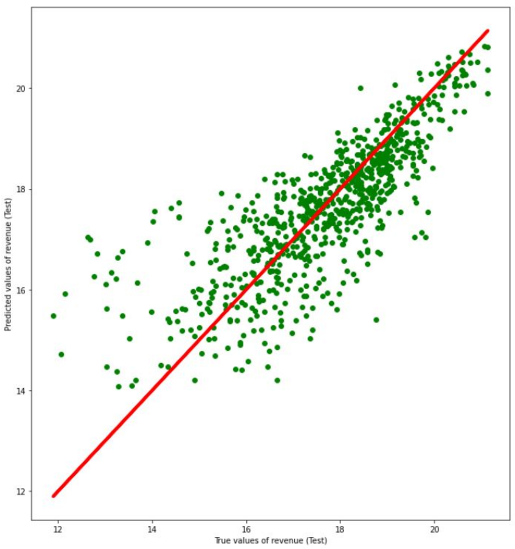


XGBoost on Pre + Post Production Elements

Training Data Set



Test Data Set



Linear Regression vs XGBoost (Pre + Post-Production Elements)

Linear Regression

```
Goodness of Fit of Model 	Train Dataset
Explained Variance (R^2) 	: 0.651533835088266
Mean Squared Error (MSE) 	: 0.8406352321565206

Goodness of Fit of Model 	Test Dataset
Explained Variance (R^2) 	: 0.6639350691672293
Mean Squared Error (MSE) 	: 0.9079209340679699
```

XGBoost

```
Goodness of Fit of Model 	Train Dataset
Explained Variance (R^2) 	: 0.8954700180286357
Mean Squared Error (MSE) 	: 0.25216676541342986

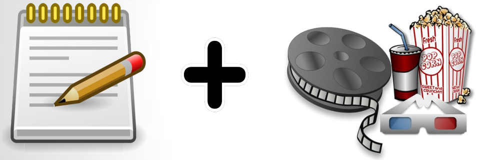
Goodness of Fit of Model 	Test Dataset
Explained Variance (R^2) 	: 0.6432408006176993
Mean Squared Error (MSE) 	: 0.9638290574915717
```

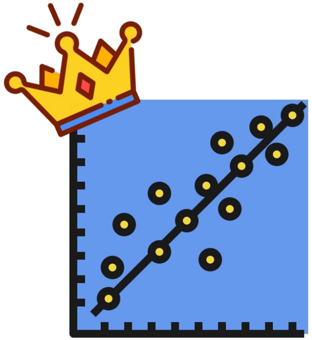
- Best of the 3 parameters
- Visible significant improvement

Conclusion

- "Can a movie's success be determined?"

PERHAPS





dmlc XGBoost

Pre & Post Elements

Combination of Pre & Post production variables

What We Learned

- Data Extraction & Cleaning
- Data Normalization
- Linear Regression
- XGBoost

Outcome Of Project

- Movie producers may predict a movie's success
- Determined by:
    - Marketing to gain popularity
    - Production resources

Bonus Feature (Movie Recommendation)

- director: A list of director in each movie
- name_cast: A list of list of cast in each movie
- name_keywords: The list of list of keywords in each movie
- name_genres: The list of list of genres in each movie

	title	combine
0	Avatar	cultureclash future spacewar action adventure ...
1	Pirates of the Caribbean: At World's End	ocean drugabuse exoticisland adventure fantasy...
2	Spectre	spy basedonnovel secretagent action adventure ...
3	The Dark Knight Rises	dccomics crimefighter terrorist action crime d...
4	John Carter	basedonnovel mars medallion action adventure s...
...	...	...
4798	El Mariachi	unitedstates–mexicobarrier legs arms action cr...
4799	Newlyweds	comedy romance nicklove edwardburns kerrybishé...
4800	Signed, Sealed, Delivered	date loveatfirstsight narration comedy drama r...
4801	Shanghai Calling	asgharfarhadi danielhenney elizacoupe billpaxton
4802	My Date with Drew	obsession camcorder crush documentary justinmo...

4803 rows × 2 columns

Bonus Feature (Movie Recommendation)

```
get_recommendations('Avatar', cosine_sim)
```

```
466                      The Time Machine
26                Captain America: Civil War
47                  Star Trek Into Darkness
94                  Guardians of the Galaxy
206                    Clash of the Titans
10                        Superman Returns
14                            Man of Steel
46                X-Men: Days of Future Past
61                       Jupiter Ascending
85     Captain America: The Winter Soldier
Name: title, dtype: object
```

```
get_recommendations('Batman Begins', cosine_sim)
```

```
3            The Dark Knight Rises
65                 The Dark Knight
4638    Amidst the Devil's Wings
982                  Run All Night
1742               Brick Mansions
3332                  Harry Brown
3603           Lone Wolf McQuade
4099                  Harsh Times
3326               Black November
1986                       Faster
Name: title, dtype: object
```

```
get_recommendations('Romeo Is Bleeding', cosine_sim)
```

```
2154                  Street Kings
3            The Dark Knight Rises
1699           Along Came a Spider
4408                Jimmy and Judy
4638      Amidst the Devil's Wings
1986                        Faster
3359                   In Too Deep
1503                        Takers
2959             Machine Gun McCain
2915                         Trash
Name: title, dtype: object
```