Predictive Modelling of Revenues of Modern American Movies

October 30, 2018

by Stephen Gou

Oct 28, 2018

Student Number: 1000382908

1 Introduction

A movie's box office is the most common metric to gauge its success. A good prediction of the revenue of a movie can guide production companies for building successful movies, and inform investors to pick out the most profitable movies. This project builds a model that predicts a movie's total revenue, given certain traits and facts about the movie. Only movies produced in the United States from 1990 to 2016 are considered, because the entertainment industry and economy changes over time. Movies produced after 2016 are not considered, because have not reached their full total revenue potential. Only movies produced in the U.S are considered, because the market characteristics vary over countries and the modelling of this aspect is beyond the scope of this project. This project aims to provide effective prediction as soon as the movies are released, which means that data like opening weekend box office, IMDb rating, social media sentiments cannot be used as features in the models.

To build an effective predictive model and gain insight, the project first explores and analyzes the major factors that affect a movie's revenue. And then, a model that best suits the case will be selected and trained. Its performance will be analyzed and compared to an alternative model. Last but not least, the model's limitations and potential improvements will be discussed.

2 Data Collection

This project makes use of several sources to collect data for analysis and training. Various types of data are collected that includes movie's revenue, budget, meta-data, cast, crews, rankings of actors and actresses and so on. The detail of all the datasets used is listed below.

1) TMDB 5000 Movies dataset. This is the main dataset which provides budget, revenue, runtime, genre, release-date and production country data. Source: https://www.kaggle.com/tmdb/tmdb-movie-metadata

2) New York Times Review dataset. This dataset includes data like whether a movie was picked by NYT critics, and review summaries. Source: NYT API

3) TMDB 5000 Crew dataset. This dataset has detailed cast and crew information, ranging from actor to writer, for each movie. Source: https://www.kaggle.com/tmdb/tmdb-movie-metadata

4) Top Actors/Actresses Rank. This the list of a Top 1000 Actors/Actresses Ranking released by IMDb. Source: IMDb

5) Top directors Rank. This the list of a Directors Ranking released by IMDb. Source: IMDb

6) Annual CPI. This dataset lists the annual average CPI for U.S. Source: UsInflationCalculator.com

2.1 Cleaning

Movies produced before 1990 and after 2016 are discarded. Movies produced outside of U.S are discarded. Some movies have zero revenue in the dataset, which might be a result of missing data or unreleased movie. These movies are removed.

3 Feature Selection and Mapping

There are a large amount of factors that might affect a movie's revenues ranging from movies' meta-data, to unemployment rate of the release year. Features that will be analyzed and incorporated into the predictive model are selected based on availability, informativeness, unambiguity, and interpretability. According to this criteria, the following features are selected: budget, runtime, critics-pick, genres, MPAA-rating, cast, and director. The following procedures and transformations of data are done to make data representable for modelling and to increase accuracy.

1) The cast of a movie is represented by a popularity score, which is calculated by the following rule. A percentile rank score for each cast is calculated according to the actors rank dataset. Then use 1 - percentage rank as the popularity score for a cast. So 1 is the highest one can get and 0 is the lowest (0 if cast not in the ranking). Then the cast popularity for the movie is calculated as following:

$$Cast\ Popularity\ Score = \sum_{i}^{N} \gamma^{i}(1 - PercentileRank(Cast\ i))$$

where gamma is a decay factor and N is the number of casts.

2) The director is represented by a popularity score, which is calculated by the following rule. A percentile rank score is calculated according to the directors rank dataset. Then use 1 - percentage rank as the popularity score. So 1 is the highest one can get and 0 is the lowest (0 if director not in the ranking).

3) The revenue and budget are adjusted for inflation according to the rule:

$$adjusted = \frac{CPI(2017)}{CPI(release\,year)} * unadjusted.$$

CPI are from the Annual CPI data.

4) Genres are converted by one-hot encoding. Note that a movie can have multiple genres associated with it.

5) MPAA ratings are converted by one-hot encoding.

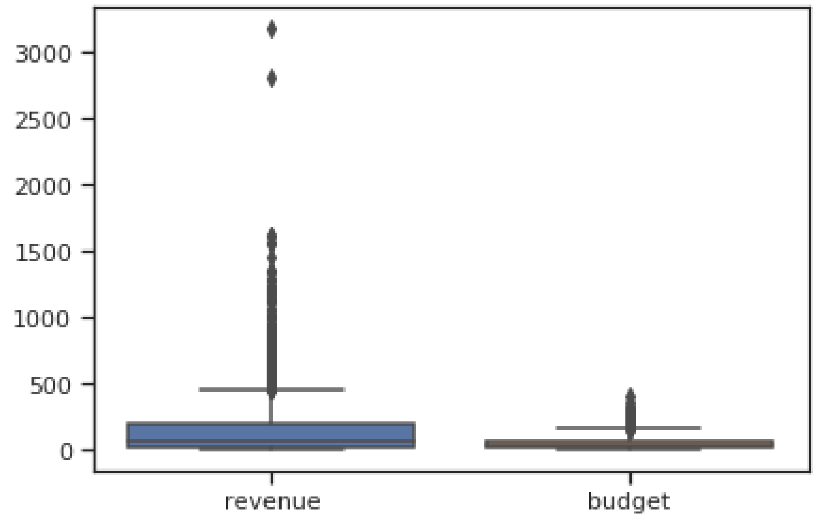
6) Runtime represented by a number and unchanged.

7) Critics pick is represented by 1 or 0 (1 repesents being picked)

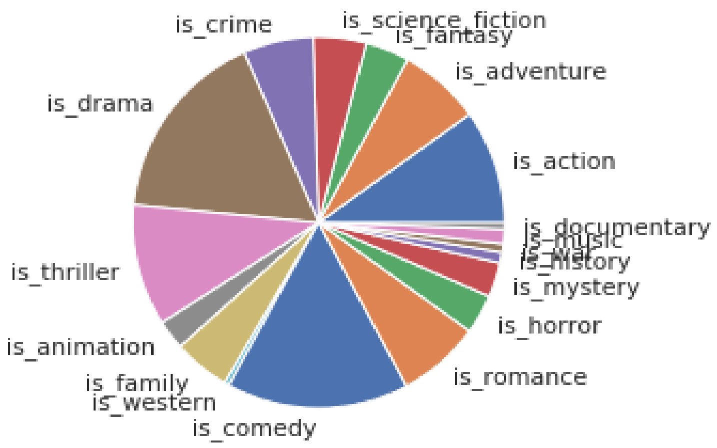
4 Exploratory Data Analysis

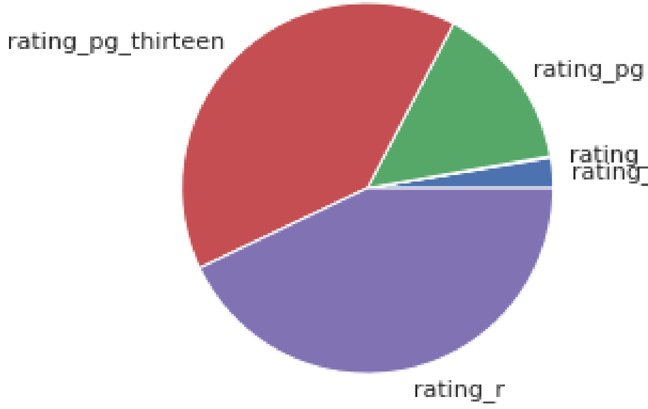
Some observations can be made from the statistics of our wrangled dataset. There are 2,033 movies in our final dataset. 17% of the movies are picked by the critics. Average runtime of a movie is 108 minutes while the lengthiest runs more than 4 hours, the shortest runs 46 mintues.

4.1 Distributions of Data



Revenues and budgets of movies are concentrated in low values, \$ 78m and \$ 37m respectively. There are large number of outliers in both cases. However, revenue has very long-tail towards higher values and outliers with more extreme values.





There are quite diverse and evenly distributed number of genres in the data. And majority of movies are at least PG-13.

4.2 Correlations Between Features

A heatmap of correlation between features is plotted to spot features that have strong relationships with each other, so that redundant features can be discarded to reduce multicollinearity.

Genres and mpaa-rating tend to have strong correlations. From the plot, it's clear that movies that have "family" as a genre is also very likely to have "animation" as a genre as well. Family and animation movies also usually have PG or G rating.

The quality of the cast appear to be uncorrelated with most of the genres of movies except for horror, where quality of cast drops significantly.

Intuitively, the runtime of a movie has correlations with its genres, which is confirmed by the heatmap. The runtime also correlates with budget and quality of director and cast.

Another interesting observation is that New York Time's critics' picks appear to be uncorrelated with most of the features of a movie, meaning that they are not favoring a particular subsets of movies over the others. Action and thriller movies are marginally less likely to be picked, but that could just be a result of noise.

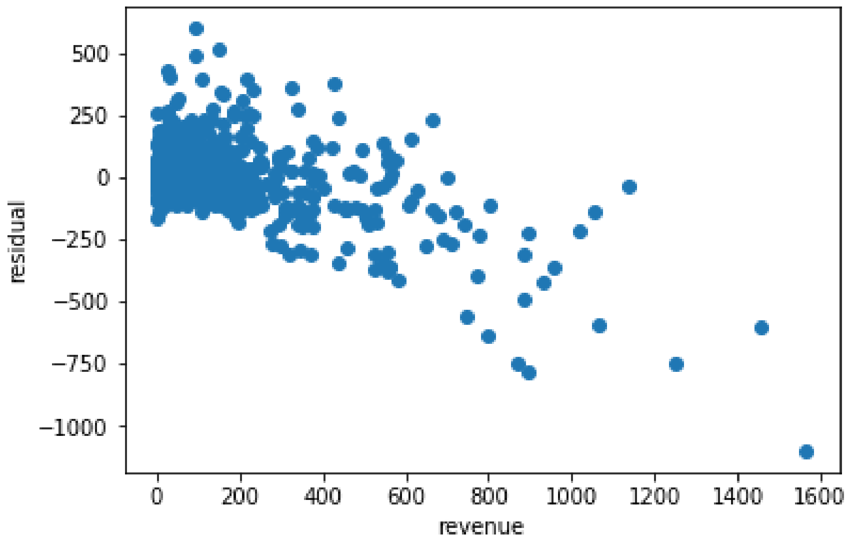
From these observations, runtime and mpaa-rating of a movie could be potentially discarded, because they usually depend on other features of the movie.

5 Analysis and Modelling

Since the goal is to predict revenue, a continuous value over a wide range, regression models are considered. More specifically, OLS regression, Ridge/Lasso regression, Regression Tree, Random Forest regression and Multilayer Perceptrons are the candidate models.

Initially, the models' performance are evaluated based on the R-Squared statistic and the residual plot. An OLS regression model that simply includes all the features without adding higher order terms and interactions is fitted and its result is used as a baseline. Model is trained on training set, which is 70% of the dataset.

It obtained a **R-Squared score of 0.492** on the test set and the residual plot as below:



The other models obtained similar results when fitted with only linear terms. Firstly, the R-Squared statistic is not very informative, given that revenue is an unbounded number and there are large number of outliers in the dataset. Secondly, the residual plot displays a clear linear relationship between residual and the revenues of the movies, meaning that there is a significant pattern of revenues of movies that it is not uncovered yet. However, there is inherently large uncertainly of movie's revenue and given limited information there are about the movies. It's necessary to find a more effective metric to evaluate the models.

The alternative metric is defined as the percentage of predictions that are within 20% error from the true value. It will be referred to as "accuracy". The baseline is **34.8%**, which is obtained by simply use the mean revenue for every prediction.

5.1 OLS Linear Regression

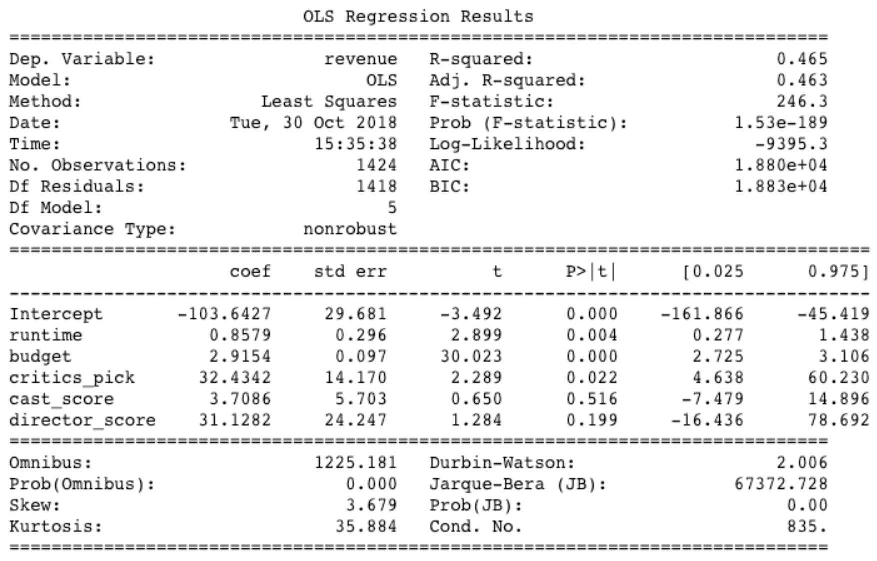
Initially, a brute force model that includes a large number of interactions between between features, and certain second order terms (102 total terms in regression formula) is fitted. It obtained accuracy of **49.6%**.

The following steps are taken to improve the performance, interpretability and reduce overfitting.

- Like suggested in EDA, mpaa-rating is discarded because it depends on other features.
- Genres are discarded as well. OLS regression shows an extremely large condition number (> 10ˆ 10) with genres included, meaning there are strong multi-collinearity. In addition, it caused certain terms to have extremely large weights, even when L1/L2 reguarizers are added. Lastly, in introduced too many potential interactions between each other and other features like directors, actors and budget.
- Naturally removing outliers from dataset was considered, but removing them did not improve any model's performances. Therefore, outliers are kept.

The resulting formula for regression is

$$revenue = Intercept + w_0 * runtime + w_1 * budget + w_2 * criticspick + w_4 * castscore + w_5 * directorscore$$



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                revenue   R-squared:                       0.465
Model:                            OLS   Adj. R-squared:                  0.463
Method:                 Least Squares   F-statistic:                     246.3
Date:                Tue, 30 Oct 2018   Prob (F-statistic):          1.53e-189
Time:                        15:35:38   Log-Likelihood:                -9395.3
No. Observations:                1424   AIC:                         1.880e+04
Df Residuals:                    1418   BIC:                         1.883e+04
Df Model:                           5
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
Intercept      -103.6427     29.681     -3.492      0.000    -161.866     -45.419
runtime           0.8579      0.296      2.899      0.004       0.277       1.438
budget            2.9154      0.097     30.023      0.000       2.725       3.106
critics_pick     32.4342     14.170      2.289      0.022       4.638      60.230
cast_score        3.7086      5.703      0.650      0.516      -7.479      14.896
director_score   31.1282     24.247      1.284      0.199     -16.436      78.692
==============================================================================
Omnibus:                     1225.181   Durbin-Watson:                   2.006
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            67372.728
Skew:                           3.679   Prob(JB):                         0.00
Kurtosis:                      35.884   Cond. No.                         835.
==============================================================================
```

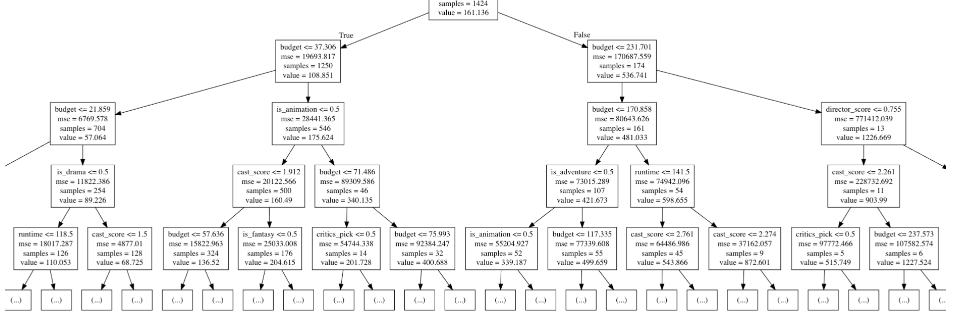
It achieved an accuracy of **55.8%**. A 6% increase comparing to the brute force model. Runtime, budget, and critics pick all have coefficients with lower than 5% p-value, and the coefficents have large values. All else equal, a movie makes \$ 32m more than it's picked by NYT critics. For every million dollar spent on budget, there are \$ 2.9m revenue in return. Actors score and director score have coefficients with large P-value, meaning that it cannot be concluded that a great cast or director will surely drive up revenue.

5.2 Non-Linear Regression Models

Given the large number of potential interactions and non-linear relationship between certain features and revenue, it is extremely hard manually select features. Thus, non-linear models like regression tree and multilayer perceptron (neutral networks with only fully connected hidden layers) are considered. For these models, all available features are included in the input.

1) A MLP with two hidden layers of size 5 and 3 with ReLU activations is trained for 2,000 iterations achieved accuracy scores ranging from **47.0%** to **54.0%**

2) A regression tree with unlimited depth is constructed and achieved accuracy score of **60.7%** The top few layers of the tree is shown as below:


budget <= 117.115 mse = 57782.207 samples = 1424 value = 161.136
True
False
budget <= 37.306 mse = 19693.817 samples = 1250 value = 108.851
budget <= 231.701 mse = 170687.559 samples = 174 value = 536.741
budget <= 21.859 mse = 6769.578 samples = 704 value = 57.064
is_animation <= 0.5 mse = 28441.365 samples = 546 value = 175.624
budget <= 170.858 mse = 80643.626 samples = 161 value = 481.033
director_score <= 0.755 mse = 771412.039 samples = 13 value = 1226.669
is_drama <= 0.5 mse = 11822.386 samples = 254 value = 89.226
cast_score <= 1.912 mse = 20122.566 samples = 500 value = 160.49
budget <= 71.486 mse = 89309.586 samples = 46 value = 340.135
is_adventure <= 0.5 mse = 73015.289 samples = 107 value = 421.673
runtime <= 141.5 mse = 74942.096 samples = 54 value = 598.655
cast_score <= 2.261 mse = 228732.692 samples = 11 value = 903.99
runtime <= 118.5 mse = 18017.287 samples = 126 value = 110.053
cast_score <= 1.5 mse = 4877.01 samples = 128 value = 68.725
budget <= 57.636 mse = 15822.963 samples = 324 value = 136.52
is_fantasy <= 0.5 mse = 25033.008 samples = 176 value = 204.615
critics_pick <= 0.5 mse = 54744.338 samples = 14 value = 201.728
budget <= 75.993 mse = 92384.247 samples = 32 value = 400.688
is_animation <= 0.5 mse = 55204.927 samples = 52 value = 339.187
budget <= 117.335 mse = 77339.608 samples = 55 value = 499.659
cast_score <= 2.761 mse = 64486.986 samples = 45 value = 543.866
cast_score <= 2.274 mse = 37162.057 samples = 9 value = 872.601
critics_pick <= 0.5 mse = 97772.466 samples = 5 value = 515.749
budget <= 237.573 mse = 107582.574 samples = 6 value = 1227.524
(...)

Budget seems to provide the most information gain, as itappear most frequently in top layers. More interestingly, certain genres, seem to have significant effects. For example, in lower budget movies (budget < \$ 117m) animation movies average \$ 340m revenue while the others only average \$ 160m. For high budget movies (more than\$ 230), directors that rank in top 25% average \$ 3000m revenue while others only \$ 903m.

6 Results

This project has constructed a dataset of American movies produced between 1990 and 2016 that include features: budget, runtime, genres, mpaa_rating, cast_score, and director_score. The best linear model achieved **55.8%** accuracy (with 20% error allowed) and the best non-linear model, which is a regression tree achieved **60.7%**. As a reference, predicting the mean everytime will score **34.8%**.

The linear model shows that runtime, budget and critics pick have significant contributions to the gross revenue of a movie. And the decision tree shows that budget provides the most information gain. In addition, the tree shows that genres' effects and director's effect depend on other features, such as budget.

7 Future Work and Improvement

7.1 Find better feature data

Since the residual shows that there is a significant linear pattern for the revenues missed, more relevant data and data that have more explaining power might be explored and incorporated into the models. For example, the number of views of movie's trailers before release, social media influence of casts, writer of the movie, production company and so on. Another possibility is to loosen the assumption so that more post-release data can be incoporated like opening weekend box office, IMDb rating, hashtag counts and so on.

7.2 Improve existing features

There are room for improvements of the features currently used in the models. For example, how the cast and director score is calculated could be improved. Instead of rank based, maybe include more revenue related traits for example, actors' social media following, revenues of past 3 movies, and so on.

7.3 Improve modelling

Since the models are systemically predicting overestimated revenue for low revenue movies and underestimated revenue for high revenue movies, there might be opportunity to take advantage of this observation. For example, use locally weighted regression, kNN or an ensemble of models so that movies in differernt levels can be modelled separately.

7.4 Python Code

```
In [1]: import json
        import requests
        import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt
        import seaborn as sns;
        from pandas.io.json import json_normalize
        import statsmodels.api as sm
        import statsmodels.formula.api as smf
        from sklearn.model_selection import train_test_split
        from sklearn import linear_model
        from sklearn.tree import DecisionTreeRegressor
        from sklearn.ensemble import RandomForestRegressor
        from sklearn.ensemble import RandomForestClassifier
        from sklearn.neural_network import MLPClassifier
        from sklearn.neural_network import MLPRegressor
        from sklearn.ensemble import AdaBoostRegressor
        from sklearn.neighbors import KNeighborsRegressor
        from sklearn import tree
```

7.4.1 Cleaning and Feature Mapping

```
In [2]: github_raw_root = 'https://raw.githubusercontent.com/gouzhen1/Moives-Data-Analysis/mas

        #NY Reviews Dataset
        ny_df = pd.read_csv(github_raw_root + 'NY_movie_reviews.csv')
        ny_df.rename(columns={'display_title':'title'},inplace=True)
        ny_df = ny_df[['title','mpaa_rating','critics_pick']]

        #Wrangle actors and director
        #TMDB Credits Dataset (for cast and director)
        tmdb_credits_df = pd.read_csv(github_raw_root + 'tmdb_5000_credits.csv')
        actors_rank = pd.read_csv(github_raw_root + 'Top_actors_rank.csv')['Name'].tolist()
        directors_rank = pd.read_csv(github_raw_root +'All_time_director_rank.csv')['Name'].to
        total_actors = len(actors_rank)
        total_directors = len(directors_rank)

        def transform_cast(df):
            cast_json = df['cast']
            parsed_cast = json.loads(cast_json)
            score = 0.
            count = 0
            for cast in parsed_cast:
                actor = cast['name']
                if actor in actors_rank:
                    #discounted for later casts
                    score += (0.9 ** count) * (1. - (actors_rank.index(actor)/total_actors))
                count += 1
            return score
        tmdb_credits_df['cast_score'] = tmdb_credits_df.apply(transform_cast, axis = 1)
```

```python
def transform_crew(df):
    crew_json = df['crew']
    parsed_crew = json.loads(crew_json)
    score = 0.
    for crew in parsed_crew:
        if crew['department'] == 'Directing' and crew['job'] == 'Director':
            director = crew['name']
            if director in directors_rank:
                score += (1. - (directors_rank.index(director)/total_directors))
            break
    return score

tmdb_credits_df['director_score'] = tmdb_credits_df.apply(transform_crew, axis = 1)
tmdb_credits_df = tmdb_credits_df[['title','cast_score','director_score']]

#TMDB Main Dataset
main_df = pd.read_csv(github_raw_root + 'tmdb_5000_movies.csv')
main_df['release_date'] = pd.to_datetime(main_df['release_date'])
main_df.drop(main_df[main_df['release_date'].dt.year < 1990].index, inplace=True)
main_df.drop(main_df[main_df['release_date'].dt.year > 2016].index, inplace=True)
main_df = main_df[main_df['revenue'] > 0]
main_df = main_df.merge(ny_df,how='left')

#process and filter countries
def process_country(df):
    country_json = df['production_countries']
    parsed_country = json.loads(country_json)
    if len(parsed_country) > 0:
        return parsed_country[0]['name']
    else:
        return None
main_df['production_countries'] = main_df.apply(process_country, axis = 1)
main_df = main_df[main_df['production_countries'] =='United States of America']
main_df.drop(columns='production_countries',inplace=True)

#wrangle genre
genre_dict = {}
def transform_genre(df):
    genre_json = df['genres']
    parsed_genre = json.loads(genre_json)
    result = []
    for genre in parsed_genre:
        genre_name = genre['name'].replace(' ','_')
        result.append(genre_name)
        if genre_name not in genre_dict:
            genre_dict[genre_name] = 1
        else:
            genre_dict[genre_name] += 1
    return result
main_df['genres'] = main_df.apply(transform_genre, axis = 1)
#drop very low rare genres
del genre_dict['Foreign']
for genre in genre_dict:
    main_df['is_' + genre] = main_df['genres'].transform(lambda x: int(genre in x))
main_df.drop(columns=['genres'],inplace=True)

#map mpaa rating
rating_df = pd.get_dummies(main_df['mpaa_rating'],prefix='rating')
main_df = main_df.merge(rating_df,left_index=True,right_index=True)
main_df.drop(columns=['mpaa_rating','rating_Not Rated'],inplace=True) #drop one catego

#adjust revenue and budget for inflation
cpi_df = pd.read_csv(github_raw_root + 'Annual_CPI.csv')
cpi_df = cpi_df.set_index('DATE')
cpi_dict = cpi_df.to_dict()['CPIAUCSL']
def get_cpi_adjusted_revenue(df):
    year = df['release_date'].year
    revenue = df['revenue']
    return cpi_dict['2017-01-01']/cpi_dict['{}-01-01'.format(year)] * revenue

def get_cpi_adjusted_budget(df):
    year = df['release_date'].year
    budeget = df['budget']
    return cpi_dict['2017-01-01']/cpi_dict['{}-01-01'.format(year)] * budeget

main_df['revenue'] = main_df.apply(get_cpi_adjusted_revenue,axis=1)
main_df['budget'] = main_df.apply(get_cpi_adjusted_budget,axis=1)
main_df['revenue'] = main_df['revenue'] * 0.000001
main_df['budget'] = main_df['budget'] * 0.000001

def cat_revenue(df):
    rev = df['revenue']
    c = min(int(rev/250. * 10.),10)
    return c

#main_df['revenue'] = main_df.apply(cat_revenue,axis=1)

main_df = main_df.drop(columns = ['release_date','original_language','popularity','home
```

```
In [3]: print('wrangled dataset: ' + str(main_df.shape))
        main_df = main_df.merge(tmdb_credits_df,how='left')
        main_df.columns = map(str.lower, main_df.columns)
        main_df.rename(columns={'rating_pg-13':'rating_pg_thirteen','rating_nc-17':'rating_nc_
        main_df['critics_pick'].fillna(0,inplace=True)
        main_df.to_csv('wrangled_dataset.csv')
        main_df.rename(columns={'rating_not rated':'rating_not_rated'},inplace=True)
        main_df.head()
```

```
wrangled dataset: (2033, 28)
```

```
Out[3]:        budget      revenue  runtime                                   title  \
        0  270.771526  3185.238655    162.0                                  Avatar
        1  354.684562  1136.172881    169.0  Pirates of the Caribbean: At World's End
        2  266.936095  1158.437625    165.0                   The Dark Knight Rises
        3  277.613539   303.387927    132.0                             John Carter
        4  305.028724  1053.261376    139.0                            Spider-Man 3

           critics_pick  is_action  is_adventure  is_fantasy  is_science_fiction  \
        0           1.0          1             1           1                   1
        1           0.0          1             1           1                   0
        2           1.0          1             0           0                   0
        3           0.0          1             1           0                   1
        4           0.0          1             1           1                   0

           is_crime       ...        is_war  is_music  is_documentary  rating_g  \
        0         0       ...             0         0               0         0
        1         0       ...             0         0               0         0
        2         1       ...             0         0               0         0
        3         0       ...             0         0               0         0
        4         0       ...             0         0               0         0

           rating_nc_seventeen  rating_pg  rating_pg_thirteen  rating_r  cast_score  \
        0                    0          0                   1         0    1.412578
        1                    0          0                   1         0    2.403865
        2                    0          0                   1         0    4.555502
        3                    0          0                   1         0    1.787542
        4                    0          0                   1         0    2.124545

           director_score
        0        0.836364
        1        0.400000
        2        0.709091
        3        0.000000
        4        0.490909

        [5 rows x 30 columns]
```

7.5 EDA

```
In [4]: main_df.describe()
```

```
Out[4]:             budget      revenue      runtime  critics_pick    is_action  \
        count  2035.000000  2035.000000  2035.000000   2035.000000  2035.000000
```

```
mean     54.405731   162.734096  108.094840      0.138084     0.258968
std      53.526064   238.730589   18.560045      0.345072     0.438176
min       0.000000     0.000015   46.000000      0.000000     0.000000
25%      16.446526    25.164840   95.000000      0.000000     0.000000
50%      37.632211    78.458092  105.000000      0.000000     0.000000
75%      77.706677   198.331272  118.000000      0.000000     1.000000
max     414.154688  3185.238655  254.000000      1.000000     1.000000

       is_adventure   is_fantasy  is_science_fiction     is_crime  \
count   2035.000000  2035.000000         2035.000000  2035.000000
mean       0.186241     0.099754            0.120885     0.158722
std        0.389397     0.299746            0.326073     0.365507
min        0.000000     0.000000            0.000000     0.000000
25%        0.000000     0.000000            0.000000     0.000000
50%        0.000000     0.000000            0.000000     0.000000
75%        0.000000     0.000000            0.000000     0.000000
max        1.000000     1.000000            1.000000     1.000000

          is_drama       ...             is_war     is_music  is_documentary  \
count  2035.000000       ...        2035.000000  2035.000000     2035.000000
mean      0.441769       ...           0.019165     0.032924        0.014742
std       0.496720       ...           0.137137     0.178481        0.120548
min       0.000000       ...           0.000000     0.000000        0.000000
25%       0.000000       ...           0.000000     0.000000        0.000000
50%       0.000000       ...           0.000000     0.000000        0.000000
75%       1.000000       ...           0.000000     0.000000        0.000000
max       1.000000       ...           1.000000     1.000000        1.000000

          rating_g  rating_nc_seventeen    rating_pg  rating_pg_thirteen  \
count  2035.000000          2035.000000  2035.000000         2035.000000
mean      0.019656             0.000983     0.108600            0.296806
std       0.138849             0.031342     0.311213            0.456963
min       0.000000             0.000000     0.000000            0.000000
25%       0.000000             0.000000     0.000000            0.000000
50%       0.000000             0.000000     0.000000            0.000000
75%       0.000000             0.000000     0.000000            1.000000
max       1.000000             1.000000     1.000000            1.000000

          rating_r   cast_score  director_score
count  2035.000000  2035.000000     2035.000000
mean      0.320393     1.320750        0.072986
std       0.466742     0.963622        0.209531
min       0.000000     0.000000        0.000000
25%       0.000000     0.536000        0.000000
50%       0.000000     1.214380        0.000000
75%       1.000000     1.969573        0.000000
max       1.000000     4.644810        1.000000
```