Features Contributing to Gross Box Office Earnings

JOHNS HOPKINS UNIVERSITY

Sambit Panda, Akhil Vasvani, Thomas Keady
Data Mining 550.636

Introduction

- Imagine yourself scrolling through Fandango on a rainy day. While there are several genres of movies playing, you are interested in one. Your friends gave you a good recommendation, but the IMDB score is low. Should you see it? Though countless have faced this very question, our project attempts to solve it via finding the factors that contribute most to a successful box office movie.

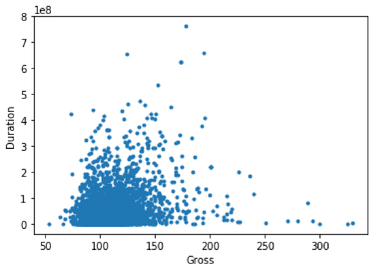
GOAL

- What contributes most to a box office hit? Is it the director, actors, production company, movie genre, or MPAA rating which heavily influence the gross?
- We aim to predict whether you should see any of the upcoming movies for holidays.

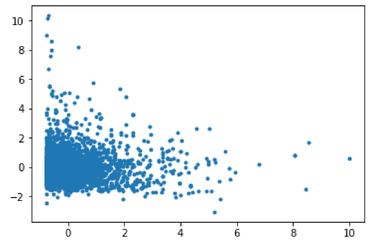
Experimental Design

Pre-Processing

- Categoric data types were mapped from strings to integers ["PG", "PG-13", "R"] to [0, 1, 2]
- We opted not to use PCA to clean our data because both with and without whitening transformation yielded terrible results. Hence, no need to reduce dimensionality



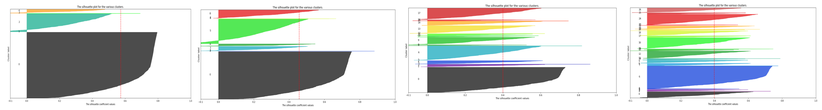
Before PCA



After PCA

Labeling Attempts

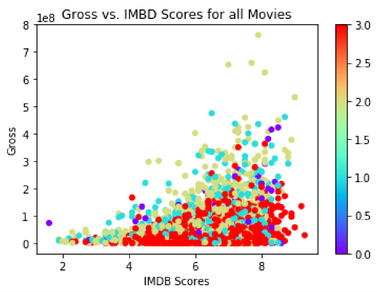
- Tried using genres but multiple for each movie; choosing the first yielded biased results for Action movies (alphabetically first), and choosing random yielded too many labels
- Tried using Good vs Bad movie but very subjective and overlabeled Good movies over bad
- Ended up using PG, PG-13, and R ratings as labels
- Utilize diversity of data mining tools:
  - Classification algorithms for textual data labeled as described above.
  - Regression algorithms used to predict gross earnings
- Clustering algorithms were attempted but the curse of dimenstionality lead to visualization difficulties and large Euclidean distances.



Dataset

- Number of movies contain: 5,044
- Number of Features/movie: 28
- Main variable of interest Gross
- Chose attributes easily identifiable by untrained movie-goer

Gross-Integer	IMDB Rating-Categorical labels	Genre-Categorical Labels
Actor-String	Director-String	Color (Black or White)-String
Positive Critic Reviews-String	Duration-Integer	Actor 3 FB Likes-Integer
Actor 2 Name-String	Actor 1 FB Likes-Integer	Actor 1 Name-String
Movie Title-String	User Voted-Integer	Cast Total-Integer
Director FB likes-Integer	Actor 3 Name-String	Plot Keywords-String
Positive IMDB Reviews-String	Language-String	Country-String
Content Rating-String	Budget-Integer	Title Year-String
Actor 2 FB Likes-Integer	Movie FB Likes-Integer	

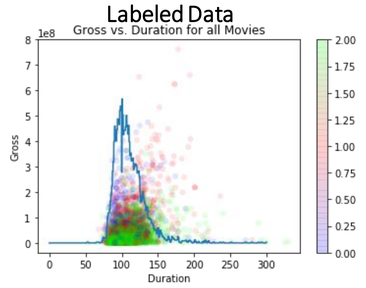


List of all the features in the data set.

Plot of Gross vs. IMDB Scores with the Genres chosen as labels

Results

Classification



Data labeled with MPAA Rating labels overlain with KDE. Density implies high gross for films around 2 hours

We opted to classify our movie data into three categories based on MPAA rating. To best train our model, we plotted the Gross vs Duration for all Movies and then performed KNN to find the nearest neighbor. After which, a naïve Bayes classifier was applied. However, this approach was abandoned because the Naïve Bayes Classifier mislabeled many points. Hence, we shifted to another approach using a decision tree regression.

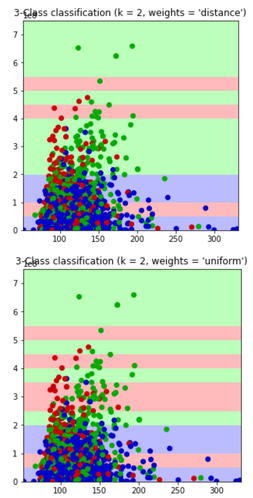
Naïve Bayes Classifier

```
0.0 [1.04934629e+02 7.73519221e+07] [5.91773247e+01 1.06782300e+15]
1.0 [1.12235474e+02 6.89651186e+07] [1.64206402e+02 2.66526254e+15]
2.0 [1.11147059e+02 3.21498388e+07] [2.56893153e+02 7.73492225e+14]

Number of mislabeled points out of a total 3574 points : 1690
```

Naïve Bayes Classifier is unable to label points correctly due to the closeness of the clusters

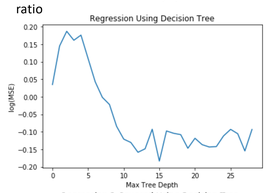
K Nearest Neighbors



KNN classifies better than Naïve Bayes, but still lots of mislabeling

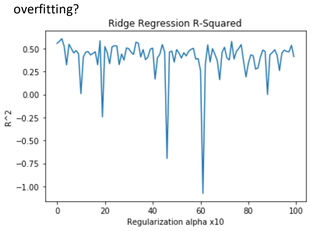
Regression

- Use regression to predict gross box office earnings
- Decision Tree Regression:
  - Depth has (slightly) negative impact past 5 layers
  - Gigantic – max_depth 9 still has over 1500 nodes
  - Training loss levelled off after about depth 10
  - R^2 values seem uncorrelated to depth
    - Possibly overfitting earlier than loss curve suggests? Due to cross validation
- Interpretability - Surprisingly intuitive!
  - Data point most commonly split on: actor 1 Facebook likes
  - Data point least commonly split on: screen aspect ratio



- Reduce cross-validation to see if overfitting reduced
- R^2 values still barely impacted by depth, try different form of regression?

- Attempt Ridge Regression to resist overfitting
- Regularization had no impact on R^2 – linear model likely too simple
- Suggests success from nonlinear models that resist overfitting?



Conclusion

- Low budget directors might find our study useful when planning their filmmaking to maximize profits
- Consider: Actor Facebook likes. Less important: Screen aspect ratio
- Future Work: Investigate neural networks ability to predict gross earnings

Libraries Utilized / Data Set Locations

- Libraries: Scikit-learn, numpy, pandas, matplotlib, sklearn
- Dataset: found on kaggle.com (https://www.kaggle.com/tmdb/tmdb-movie-metadata)
- TMDB 5000 Movie Dataset