BoxOffice Analytics

Date: April 20, 2016

Prepared By: Akansha K, Bharath CS, Chirag M, Parth P

Objectives

- Support production houses for prediction of movie success in early stages

- Aid production decisions
    - Pre-Production phase
        - Prediction intervals for ROI’s

    - Pre-Release Stage
        - Predictive analytics for movies’ revenue
        - Predictive analytics for classification of blockbusters

Data set

Sources

1. Movie Lens
   - 1M user ratings
   - 6040 user demographic information
   - 3952 movies with release year, genre
2. Web Scrapping
   - IMDB: Movie Metadata (Cast, Director, Production House, etc.)
   - Box Office Mojo: Financials (Gross Domestic Revenue, Budget)

Data Preparation

- Aggregated movie ratings for all user demographics
- Assigned binary variables for Sequels, Top 30 actors, Directors & Production house
- Grouped similar genres to ease analyses

Pre Production Phase - KMeans ++

Predictions using meta data only

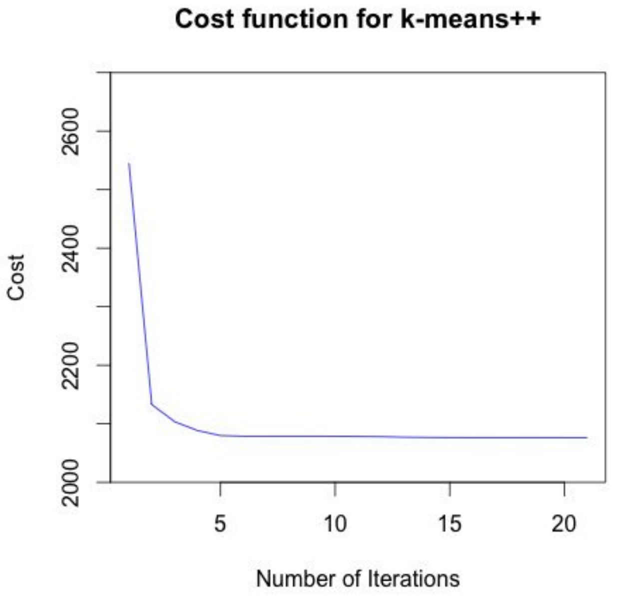
- Feature Vector: 13 Covariates

Movie	Year	Sequel	Budget	Top Actors	Top Actresses	Top Directors	Top Production	Gen 1	Gen 2	Gen 3	Gen 4	Gen 5	Gen 6
Philadelphia	1993	0	2.6e+07	1	0	0	0	0	0	0	0	0	1
The Piano,	1993	0	7.0e+06	0	1	0	1	0	0	0	0	0	1

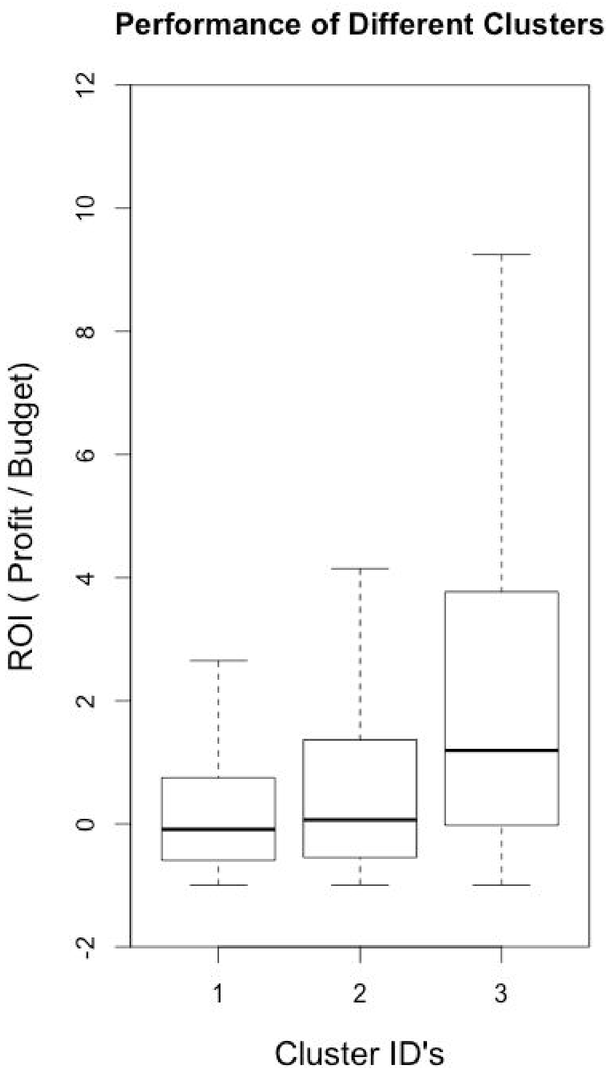
Decision Variables

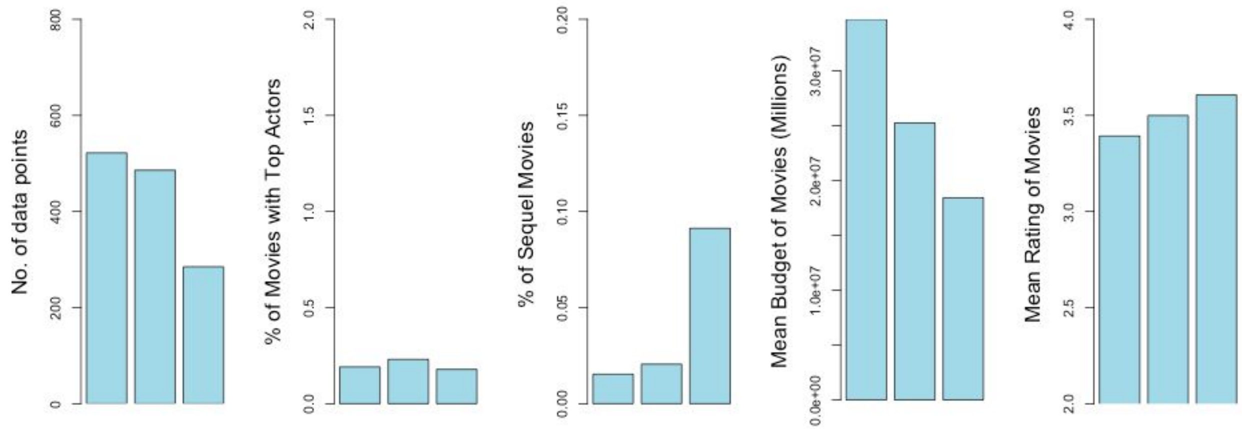
- K: # of Clusters
- Covariates included in Feature Vector
- Similarity Measure

Works well for n >> p !



Value Proposition of Clustering

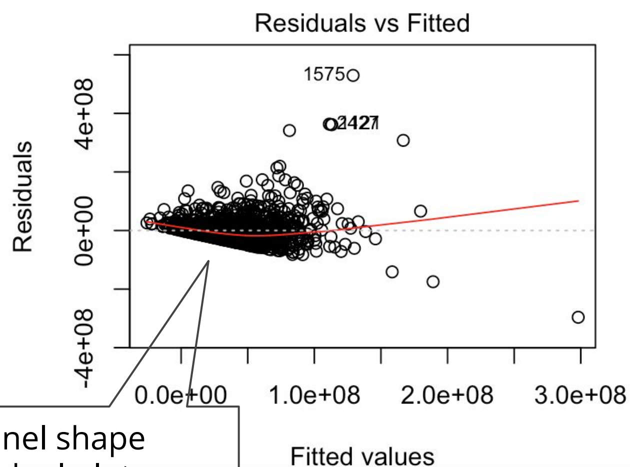




- Distribution of Revenues & Sequels differ across clusters
- Mean Budgets and Ratings weakly correlated with ROI
- % of Movies with Top Actors don’t vary across clusters

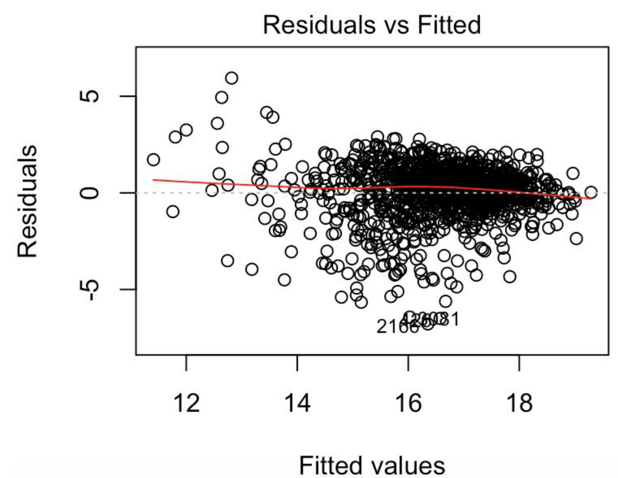
Data Transformations for regression

MovieID	Movie	Year	Sequel	Revenue	Budget	Top.30.Actors	Genre1	MAge1	FAge1
1	Toy Story	1995	0	191796233	30000000	1	1	3.83	4.07
2	Jumanji	1995	0	100475249	50000000	1	1	3.05	3.72



Funnel shape residual plot suggests log transformation

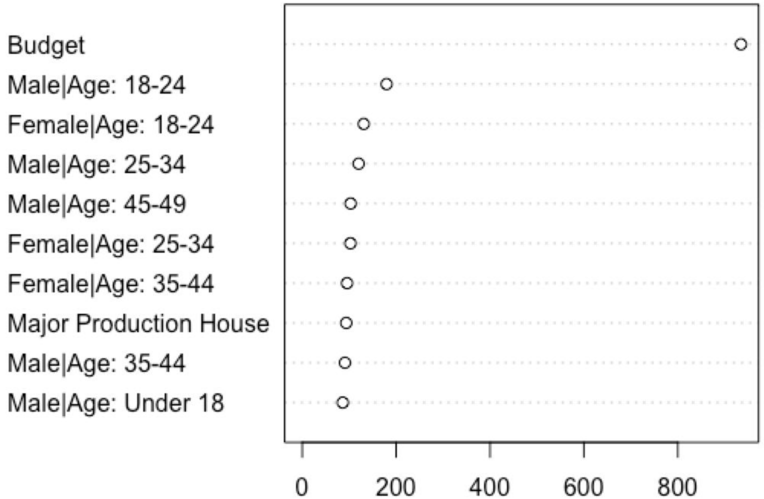
Raw Data



After log transformation of Revenue and Budget

Revenue Prediction - Pre Release Phase

- Entire data set is divided into 75% training and 25% test data sets
- Log(Revenue) is regressed against all other variables
- MSE is calculated by predicting performance on test set



Variable importance based on Random Forest

Method	MSE
Linear Regression	2.01
Subset Selection	1.98
Ridge	1.98
Lasso	1.99
Trees	2.07
Random Forest	1.56

MSE value is based on log transformation → Back-transformed MSE = e^{MSE}
Approximate error in revenue prediction is 3X
Owing to lesser predictability, we explore other ways to classify revenues!

Blockbuster Prediction- Pre Release Phase

- Blockbuster is defined as a movie with *Return on Investment* more than 2X
- A regression is run on to predict if the movie is a blockbuster or not
- Focus on reducing False Positive Rates (FPR) to reduce classification of not-blockbusters as blockbusters

```
      predict
truth   0   1
    0 178  50
    1  59  82
```

Logistic Regression

```
      predict
truth   0   1
    0 172  56
    1  61  80
```

Trees

```
      predict
truth   0   1
    0 188  40
    1  56  85
```

Random Forest

Confusion Matrices

Method	FPR	DR
Logistic Regression	21.93%	58.16%
Trees	24.56%	56.74%
Random Forest	17.10%	60.28%

Takeaways and Limitations

Pre-production phase

- New movie can be clustered to a batch of older movies
- Range for expected ROI can be predicted at some level of statistical significance

Pre-release phase

- Approximate range of revenue prediction can be provided
- Conservatively, movies can be classified with as blockbuster or not with an accuracy of 75%

Limitations

- Lack of user ratings for various demographics
- Limited availability of revenue and budget introduces bias
- Revenues includes only *domestic box office* collection

Future Work

- Use budget allocation data for promoting movies to specific populations based on demographic data
- Use marketing mix data for allocating budgets for promotion in different media for maximizing returns
- Access cleaner data sets to avoid biases in ratings
- Ease analysis of preference of metadata to specific populations
- Predictive clustering of new movies using KMeans++ or EM techniques for quantifiable results

Please reach out in case you turn to Film Production!

THANK YOU!