Analyzing the Commercial Value of Movies

Meng Zhang, Yuntao Lu, Jiaxin Li

Introduction

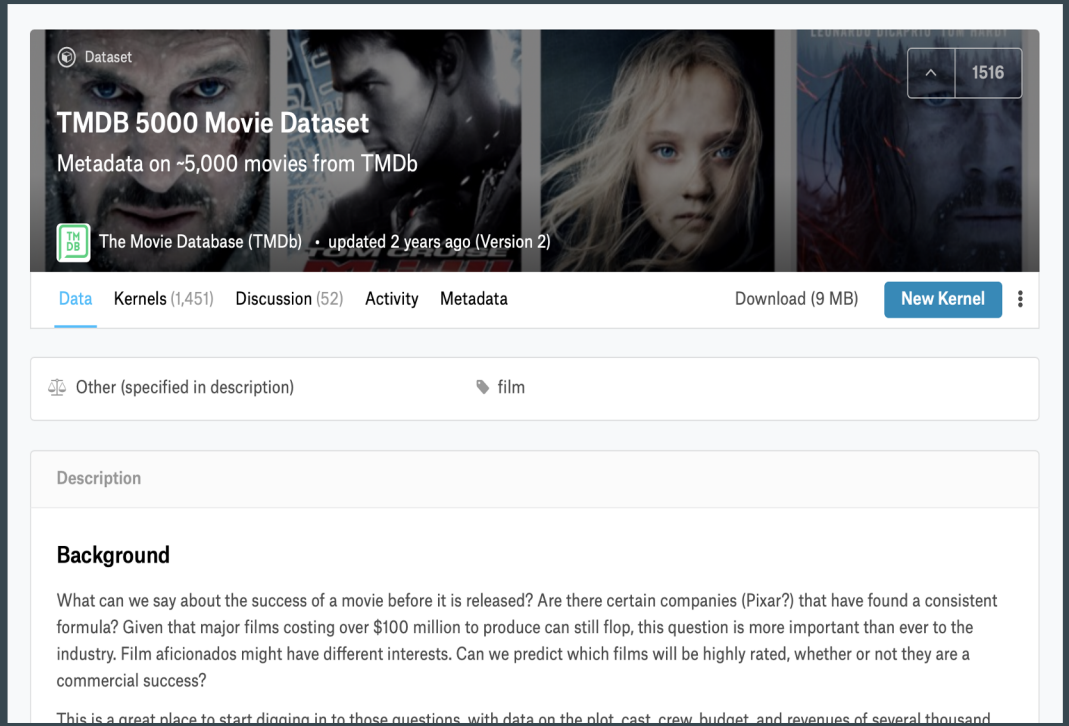
Introduction

- **Box office revenue prediction** is highly valued in the movie industry. Whether a movie will make a profit is closely correlated with important decisions made by producers and investors. Given that movies with tens to hundreds of millions dollars budgets can still flop, the accurate prediction for a movie before it is released will effectively protect producers and investors from high financial risks.

- It is also essential for advertisers to make sure which movies will appeal the audience before placing advertisement before them. The **popularity of a movie** will directly determine the range of people exposed, and consequently affect the performance of advertising campaign correlated with that movie.

Introduction

- **TMDB 5000 Movie Dataset**
- 4803 movies from TMDb
- budget, popularity, revenue, vote_average, vote_count
- genres, keywords, overview, original_language, production_companies



https://www.kaggle.com/tmdb/tmdb-movie-metadata#tmdb_5000_movies.csv

Introduction

- **Research Questions**
- Regression - Which kind of movies are more likely to be a commercial success - the movies with higher box office revenue?
- Classification - How to decide advertisement placement based on the prediction results of popularity?

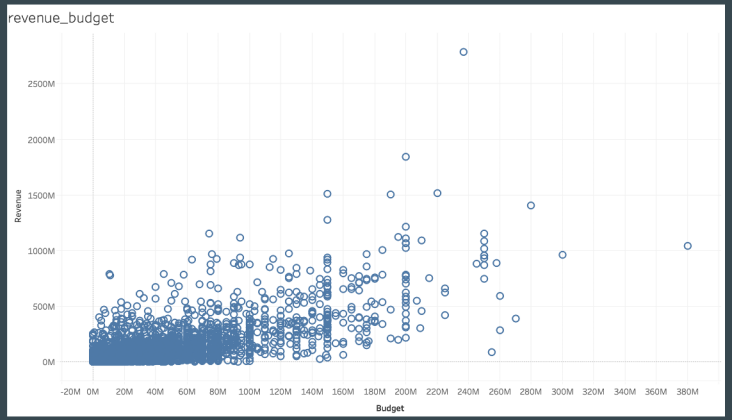
Data Preprocessing

- **Missing values & Dataset split**
  Drop 453 movie samples, 2500 movies as training data.
- **Feature selection**
  Manually drop features that are less useful in statistical analysis.
  homepage, id, original_language, original_title, release_date, runtime, status, tagline
- **Text Analysis**
  Assume that keywords feature, compared with overview feature, is more representative and precise.
  Each unique keyword is encoded with an id.

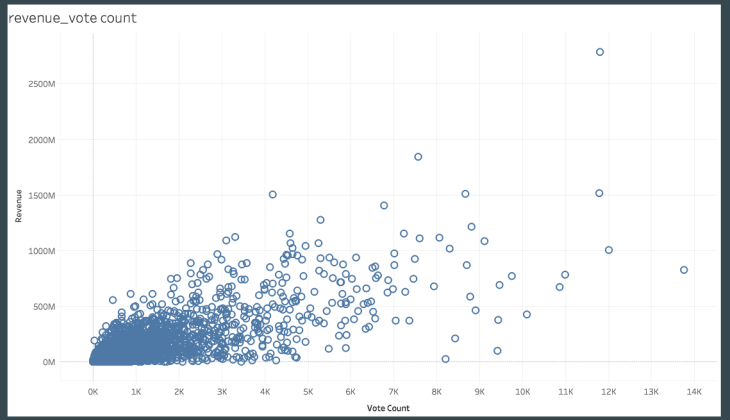
Data Preprocessing

- **Regression - box office revenue prediction**
- Qualitative Predictors: budget, vote_avg, vote_count, popularity.
- Response: revenue
- Revenue of an movie will be higher when it has higher budget, higher popularity, higher vote and more voting people.
- Tableau software - explore the distribution of revenue corresponding to each feature separately in order to figure out whether one predictor is sufficient enough for the prediction.

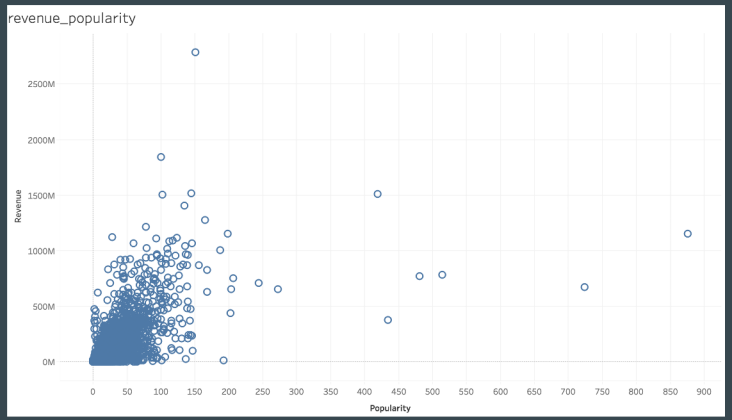
revenue-budget



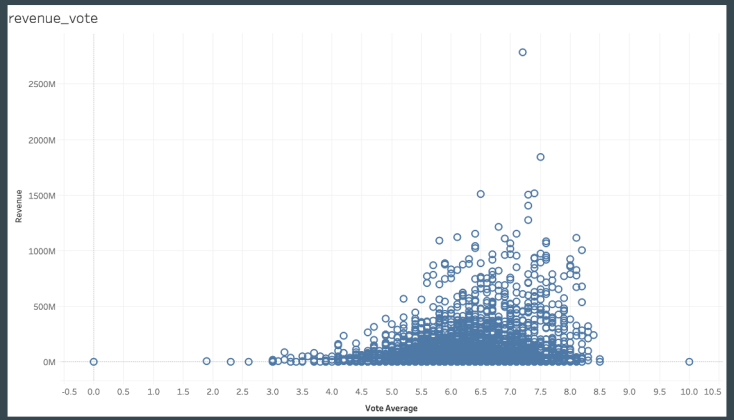
revenue-vote_count



revenue-popularity



revenue-vote_average



Data Preprocessing

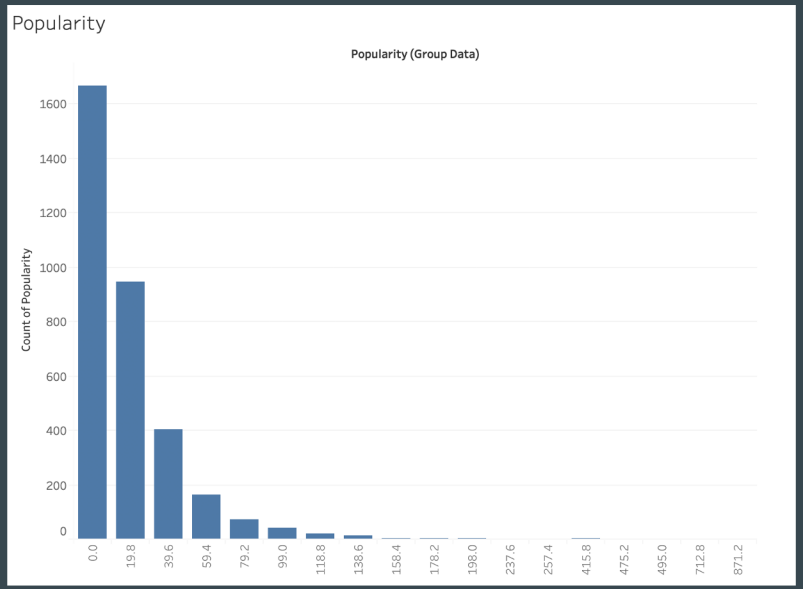
- **Classification - binary classification of popularity**
- Predictors: budget, genres, keywords, production_companies, production_countries, vote_avg, vote_count, and revenue.
- Response: popularity
    - Number of votes for the day
    - Number of views for the day
    - Number of users who marked it as a "favourite" for the day
    - Number of users who added it to their "watchlist" for the day

https://developers.themoviedb.org/3/getting-started/popularity

Data Preprocessing

- **Classification**
- Set the threshold of popularity

- Almost half of the popularity is distributed between 0 and 20.
- Popularity <= 20, no_placement
- Popularity >20, placement



The distribution of popularity

Regression Analysis

Regression Analysis

Purpose: Predicting movie box office revenue

Process: Feature Selection

Regression Model

Feature Selection

Four Quantitative Variables:

- Budget
- Vote_Average
- Vote_Count
- Popularity

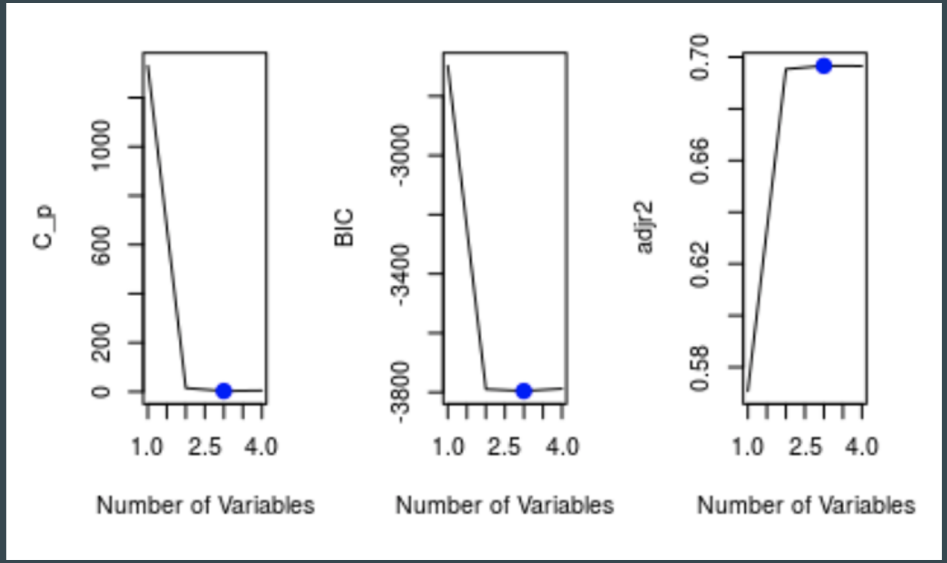
Methods:

- Best Subset Selection
- Forward Stepwise Selection
- Cp, BIC, Adjusted R^2

Feature Selection

Three Predictors:

- Budget
- Vote_Count
- Popularity




```
  (Intercept)        budget    vote_count
-2.129856e+07  1.751140e+00  6.444943e+04
   popularity
 2.868555e+05
```

Regression Analysis

Methods:

- Linear Regression
- Polynomial Regression

```
Call:
lm(formula = revenue ~ budget + vote_count + popularity, data = trainset)

Residuals:
       Min         1Q     Median         3Q
-641038871  -39232964   -2279856   24517855
       Max
1553645149

Coefficients:
              Estimate Std. Error t value
(Intercept) -1.713e+07  3.038e+06  -5.638
budget       1.847e+00  5.624e-02  32.847
vote_count   7.200e+04  2.591e+03  27.788
popularity  -2.394e+05  1.038e+05  -2.307
            Pr(>|t|)
(Intercept) 1.92e-08 ***
budget       < 2e-16 ***
vote_count   < 2e-16 ***
popularity    0.0211 *
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 102200000 on 2401 degrees of freedom
Multiple R-squared:  0.6999,    Adjusted R-squared:  0.6995
F-statistic:  1866 on 3 and 2401 DF,  p-value: < 2.2e-16
```

```{r}
mean((testset$revenue-predict(lm.fit,testset))^2)
```

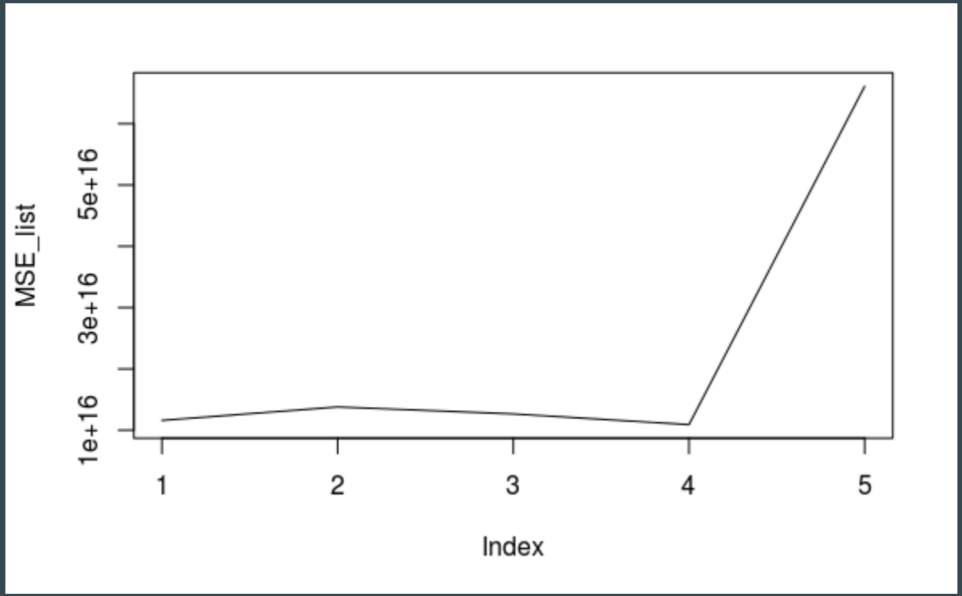
```
[1] 1.160306e+16
```

Regression Analysis

Best Model:

Polynomial Regression

With the Degree of 4




```r
mean((testset$revenue-predict(lm.fit4,testset))^2)
```

```
[1] 1.090603e+16
```

Classification Analysis

Classes & Classification Methods

- Class “0”:

  Popularity < 20

- Class “1”:

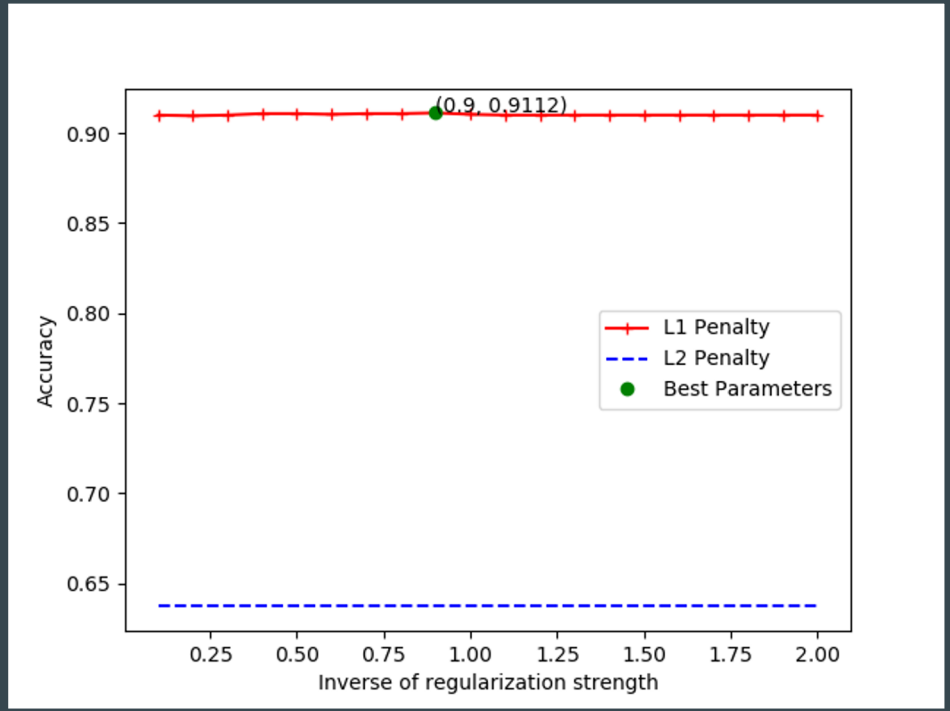
  Popularity >= 20

- Classification Methods
  - Logistic Regression
  - Naive Bayes Classifier
  - Decision Tree Classifier
  - K Neighbors Classifier
  - Random Forest Classifier
  - Boosting Classifier
  - PCA Classifier

Classification Methods

Logistic Regression

- penalty :
  - L1 or L2 penalization.
- C :
  - Inverse of regularization strength.
- Best Model:
  [L1, 0.9]

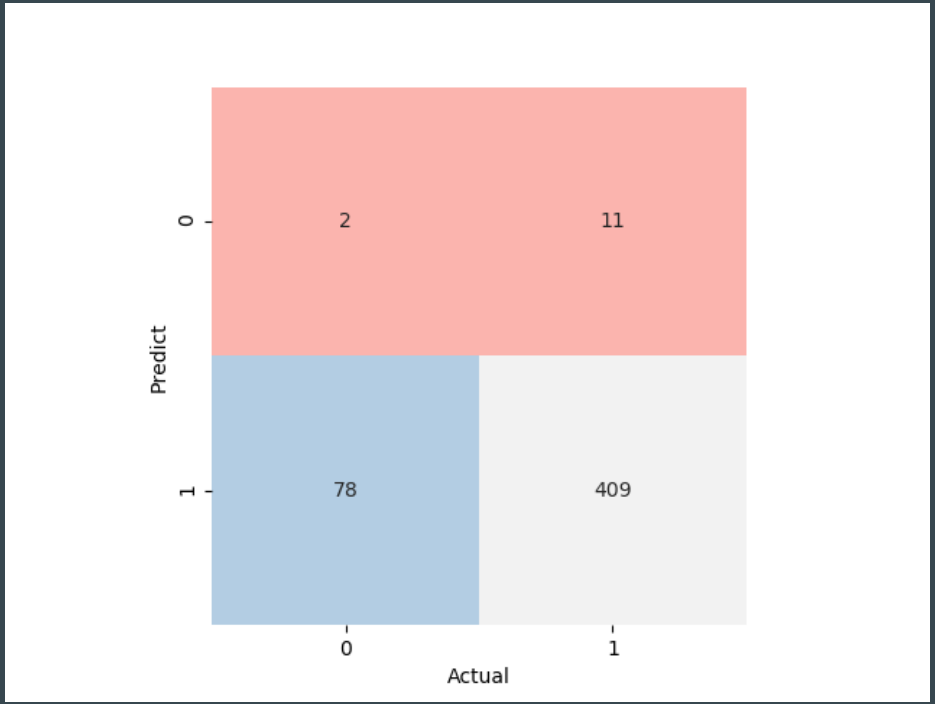


Cross-validation Accuracy	Test Accuracy	Precision Accuracy	Recall Accuracy
0.9112	0.9100	0.9881	0.9121

Classification Methods

Naive Bayes Classifier

- Didn’t tuning parameters



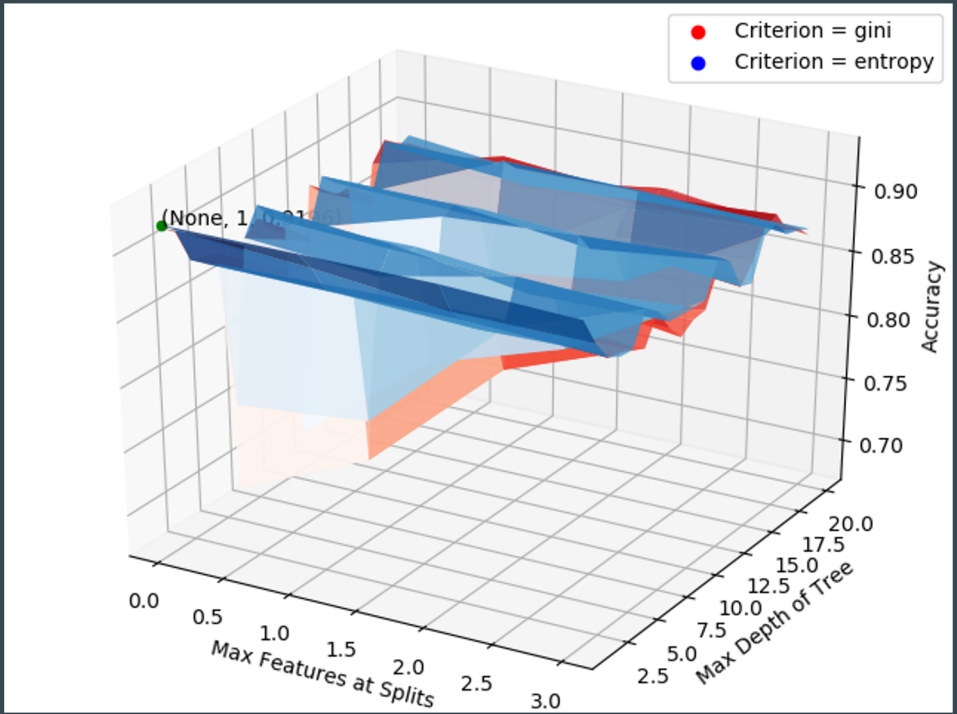
Cross-validation Accuracy	Test Accuracy	Precision Accuracy	Recall Accuracy
-	0.8220	0.9738	0.8398

Classification Methods

Decision Tree Classifier

- criterion:
  - “gini” and “entropy”.
- max_depth:
  - the maximum depth of the tree model.
- max_features:
  - The number of features of the best split.
- Best Model:

  [entropy, 1, None]

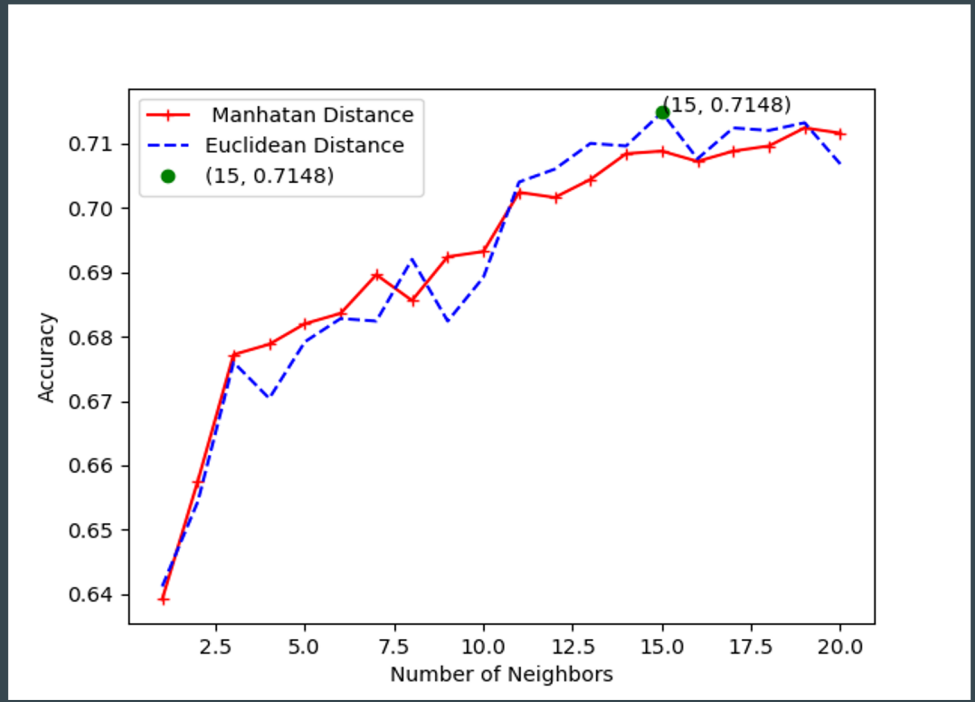


Cross-validation Accuracy	Test Accuracy	Precision Accuracy	Recall Accuracy
0.9196	0.9020	0.9552	0.8989

Classification Methods

K neighbors Classifier

- n_neighbors:
  - number of neighbors to use..
- p:
  - the power of Minkowski metric.
  - p=1, Manhattan distance
  - p=2, Euclidean distance
- Best Model:
  [15, 2]

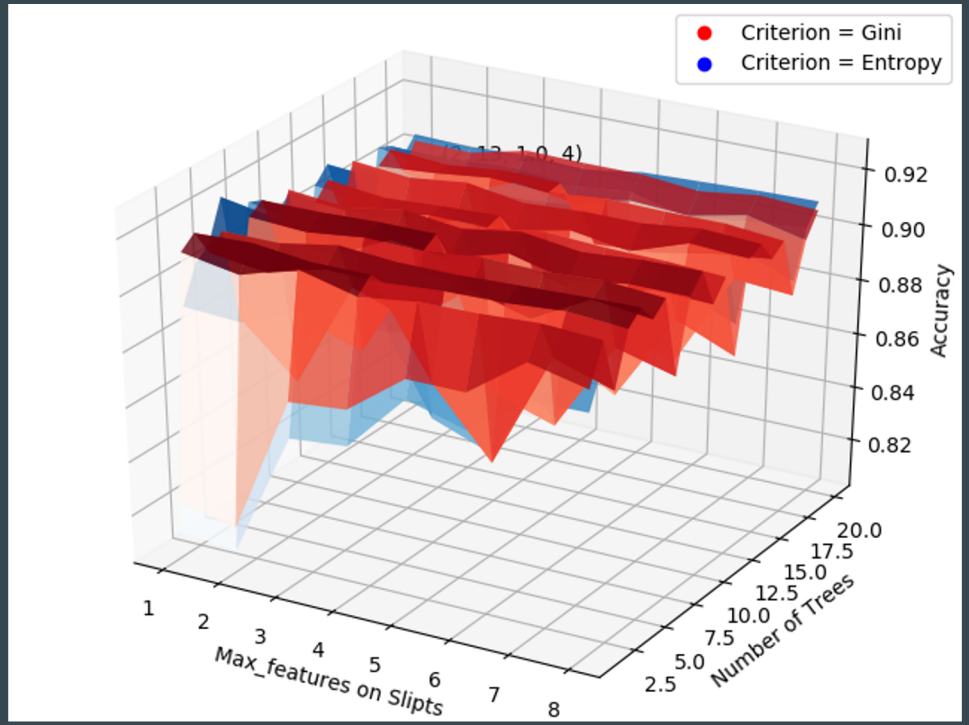


Cross-validation Accuracy	Test Accuracy	Precision Accuracy	Recall Accuracy
0.7148	0.8400	1.0	0.84

Classification Methods

Random Forest Classifier

- n_estimators:
  - number of decision trees in bagging.
- criterion:
  - “gini” and “entropy”
- Max_features:
  - the number of features in each split.
- Best Model:
  [13, entropy, 2]

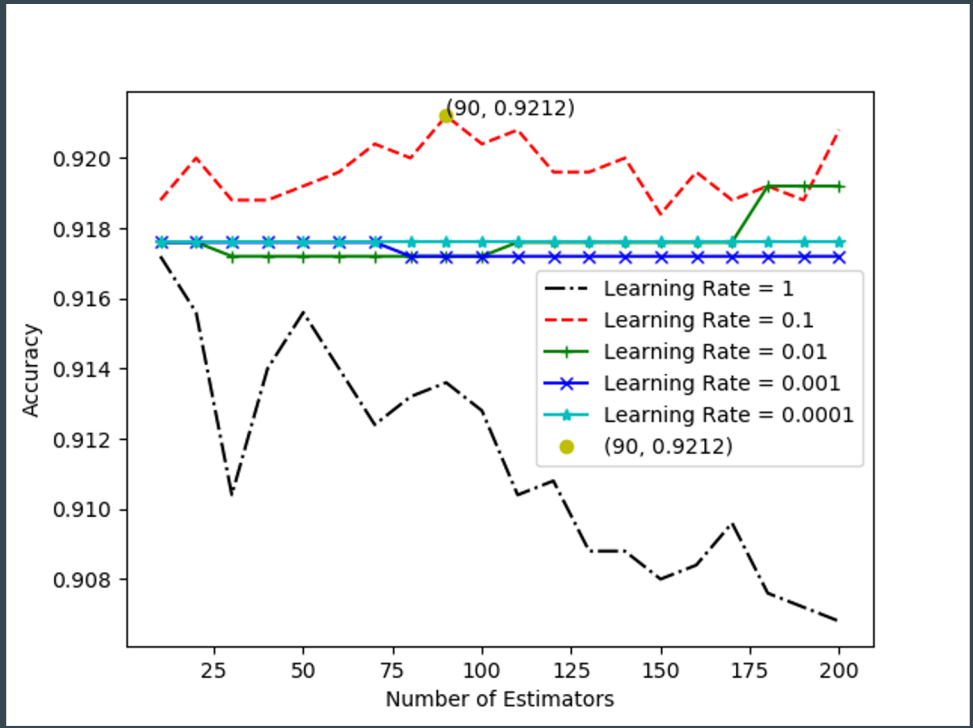


Cross-validation Accuracy	Test Accuracy	Precision Accuracy	Recall Accuracy
0.9224	0.8900	0.9833	0.8959

Classification Methods

Boosting Classifier

- n_estimators:
  - the number of estimators when boosting is terminated
- learning rate:
  - the value shrinks the contribution of each classifier
- Best Model:
  [90, 0.1]

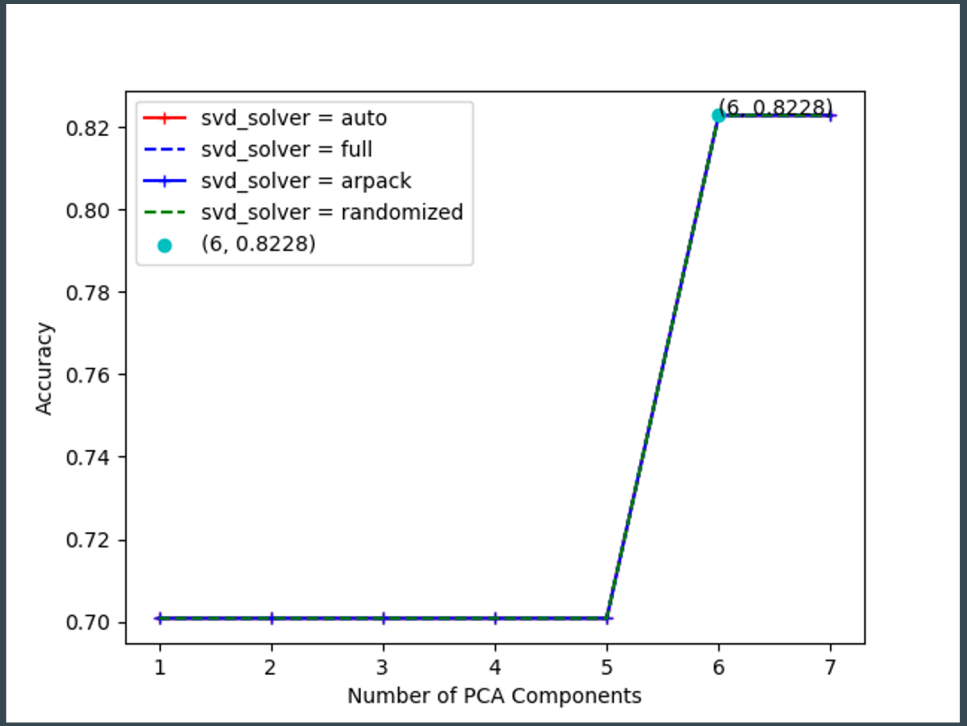


Cross-validation Accuracy	Test Accuracy	Precision Accuracy	Recall Accuracy
0.9112	0.9040	0.9552	0.9009

Classification Methods

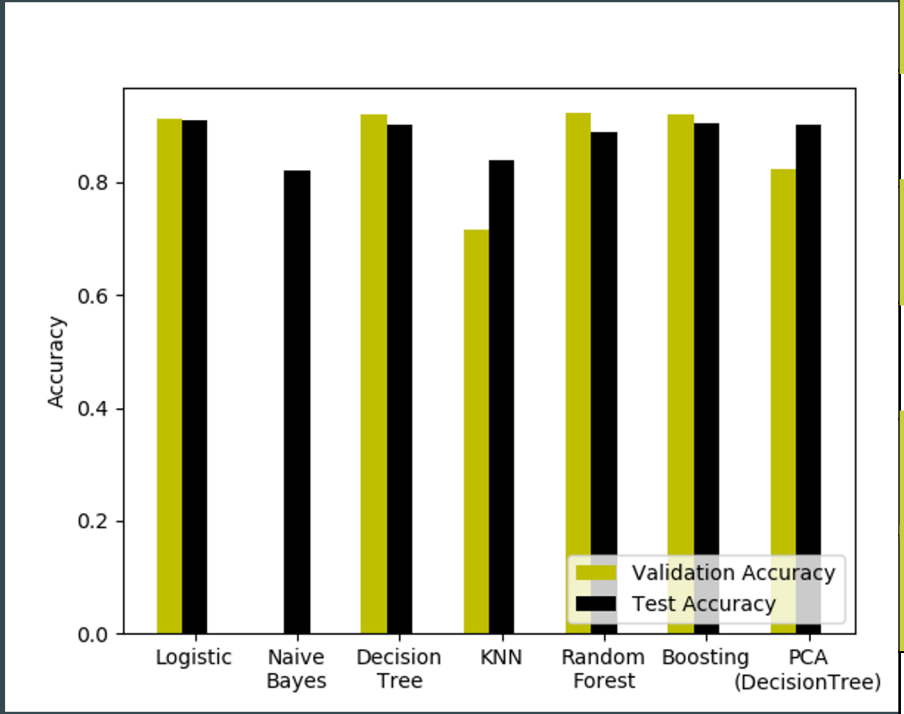
PCA Transform (Decision Tree Classifier)

- n_components:
  - the number of components to use.
- svd_solver:
  - the method SVD calculation.
- Best Model:
  [6, anyone]



Cross-validation Accuracy	Test Accuracy	Precision Accuracy	Recall Accuracy
0.8228	0.9020	0.9952	0.8989

Method Comparison



Classification Method	Validation Accuracy	Test Accuracy
Logistic Regression	0.9112	0.9100
Naive Bayes Classifier	-	0.8220
Decision Tree Classifier	0.9196	0.9020
K Neighbors Classifier	0.7148	0.8400
Random Forest Classifier	0.9224	0.8900
Boosting Classifier	0.9112	0.9040
PCA Classifier	0.8228	0.9020

Limitations & Future Work

Limitations & Future Work

- **Limited size of dataset**
  The TMDB dataset contains less than 5000 movie samples in it. The small size of dataset constrains us from making accurate prediction and are very likely to lead to overfitting problem.
- **Missing values**
  Listwise deletion is simple and avoids inaccurate coefficient estimation.
  Alternative approaches: pairwise deletion, mean substitution, regression imputation, maximum likelihood.
  Wrangling data from different datasets to produce useful, high-quality dataset.

Limitations & Future Work

- **Feature selection method**
  Drop less useful features manually based on our common sense.
  Overlook some potential relationships between certain predictors and response.
  Include some predictors which have strong correlation between them.
  Select useful predictors through subset selection methods.
- **Text analysis**
  Sentimental analysis of movie review is also a critical factor of making prediction for revenue and popularity. Future work on movie data analysis can dive into this direction further with more movie review features are collected.

Q & A