Predicting Movie Ratings Based on Reviews

Team: 3
Project: 9
Mentor: Satyam Mittal
Faculty: Dr. Ravi Kiran

Ananya Mukherjee
Hema Ala
Ritesh
Shivani Sri Varshini

PAPER IMPLEMENTATION

Movie Review Classification Based on a Multiple Classifier

http://aclweb.org/anthology/Y07-1050

”

Movie Review Dataset is created by obtaining the reviews from IMBD site (web scrapping)

“

DATASET CREATION

- Downloaded the kaggle dataset for IMDB.
- Based upon the url for each movie id, reviews are obtained by scrapping data from IMDB site.
- Along with the reviews, ratings are also stored for each review comment of each movie.

DATASET

fn	tid	title	wordsInTit	url	imdbRatin	ratingCour	duration	year
titles01/tt0012349	tt0012349	Der Vagab	der vagabu	http://ww	8.4	40550	3240	1921
titles01/tt0015864	tt0015864	Goldrauscl	goldrausch	http://ww	8.3	45319	5700	1925
titles01/tt0017136	tt0017136	Metropolis	metropolis	http://ww	8.4	81007	9180	1927

SAMPLE REVIEW AFTER SCRAPPING

['Its a great movie undoubtedly and a must watch atlas watch it before you die you may learn something you never wanted to miss Its / for one of the finest silent movies ever', '10']

DATASET MODIFICATION

- Threshold Considered: 7
- For any rating which is above 7 is considered to be positive rating else negative.

MULTIPLE CLASSIFIERS

SVM

Implemented Support Vector Machine binary classifier to classify positive and comments.

ME

Using Maximum Entropy a classifier is implemented using nltk libraries.

SCORING

For scoring we have calculated the polarity scores for each comment and classify based on scores.

INTEGRATING THE CLASSIFIERS

- Naive Voting
- Weighted Voting

  Uses each distance from hyperplanes of each classifier as weights (confidence) of the outputs.

  Scoring: The actual value of the output from the classifier
  SVM: dist(d) x l
  ME : (p(positive,d) - p(negative,d)) x m

PERFORMANCE OF CLASSIFIERS

SVM:	precision	recall	f1-score	support
0	0.91	0.84	0.88	382
1	0.85	0.92	0.89	380
avg / total	0.88	0.88	0.88	762
ME:	precision	recall	f1-score	support
0	0.95	0.84	0.89	382
1	0.86	0.95	0.90	380
avg / total	0.90	0.90	0.90	762
Scoring:	precision	recall	f1-score	support
0	0.86	0.44	0.58	382
1	0.62	0.93	0.75	380
avg / total	0.74	0.68	0.66	762
Weighted Voting				
	precision	recall	f1-score	support
0	0.97	0.79	0.87	382
1	0.82	0.98	0.89	380
avg / total	0.90	0.88	0.88	762

Our Approach

- Extending it as a multi-class problem.
- Predict individual review comment rating.
- Calculate movie rating by taking average of all the ratings.

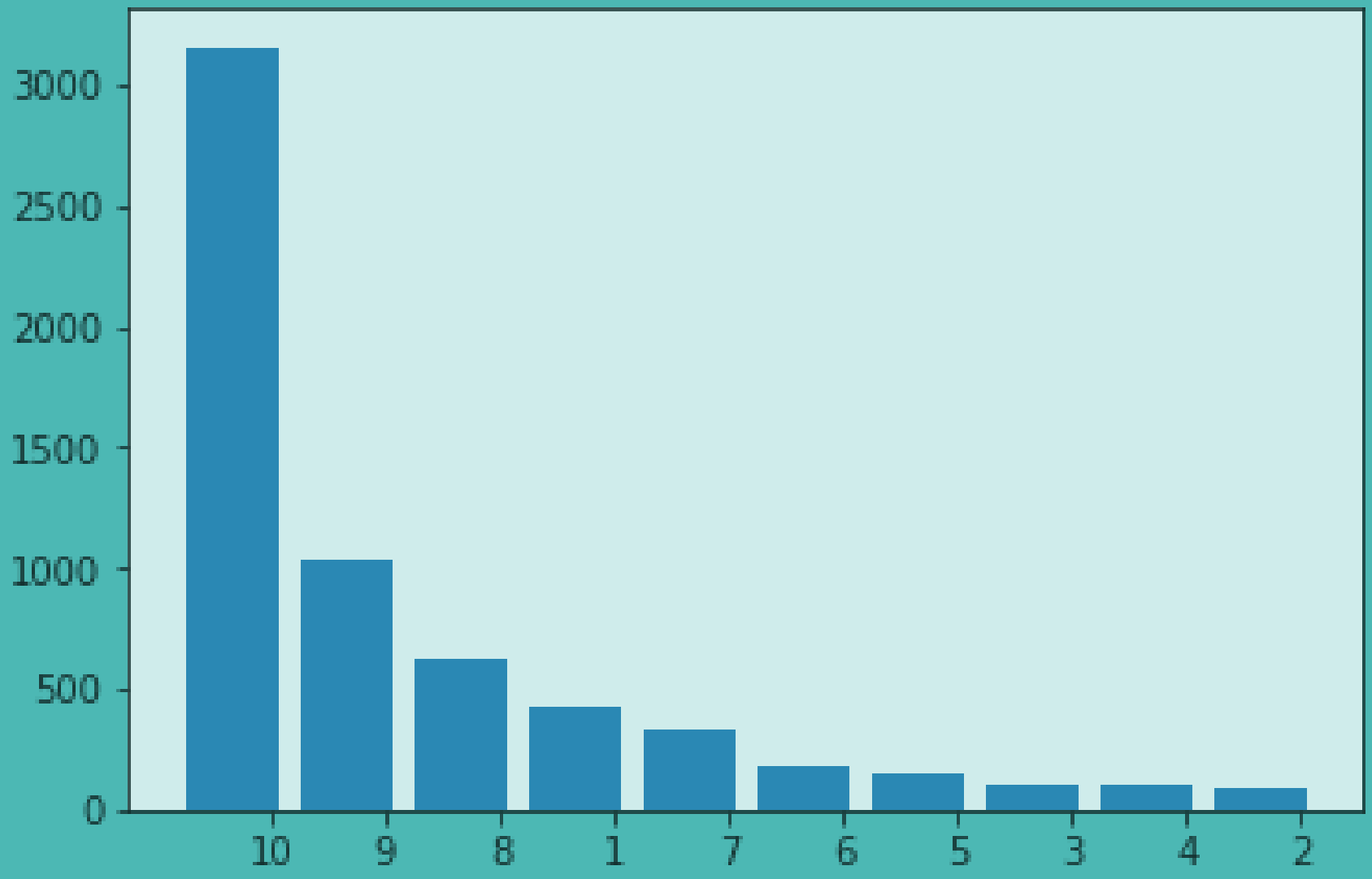
DATASET

- Consider the data-set consisting of reviews and ratings together which was initially obtained.
- Ignored the comments which do not have any rating in the IMDB website.

MODELS

- LSTM
- SVM
- Naive Bayes
- KNN

CLASS IMBALANCE



RESULTS

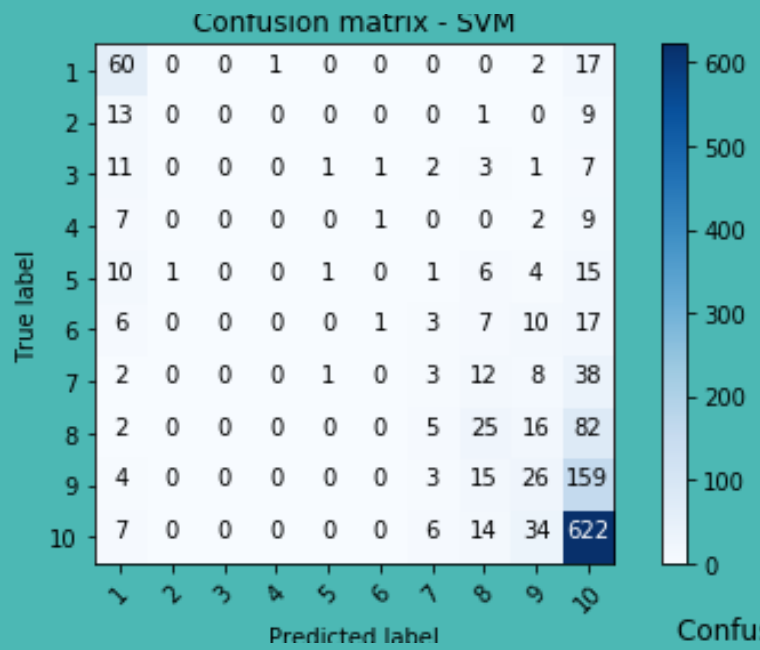
TITLE	Actual	SVM	KNN	NB
Salinui chueok (2003)	8.86	9.5	7.04	10
Vergiss mein nicht (2004)	7.65	9.1	8.73	10
Rang De Basanti (2006)	7.88	9.6	8.68	10
Das Schweigen der Lämmer (1991)	9.77	9.4	8.5	10
Oldboy (2003)	6.77	8.1	9.29	10

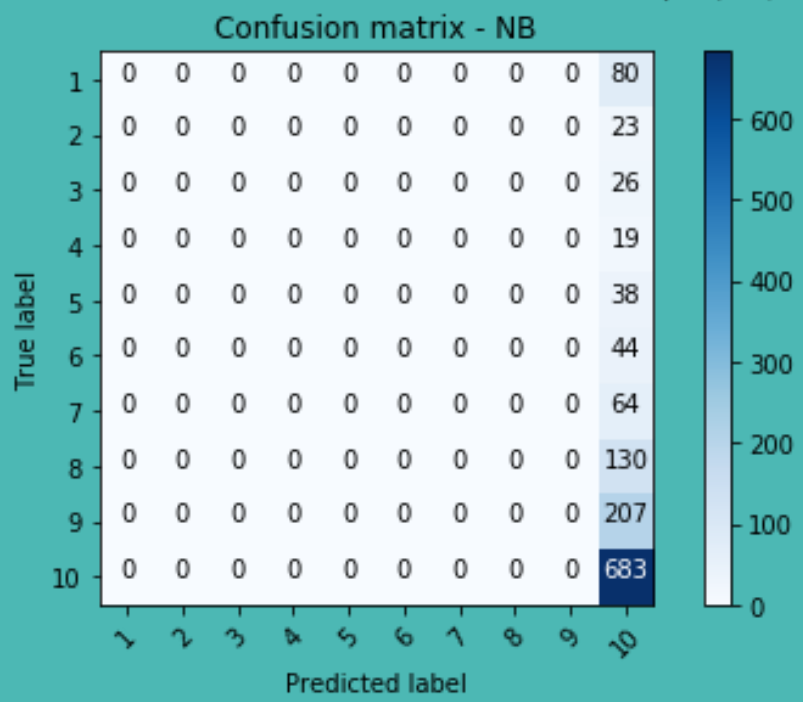
Actual and Predicted movie rating results for various classifiers.

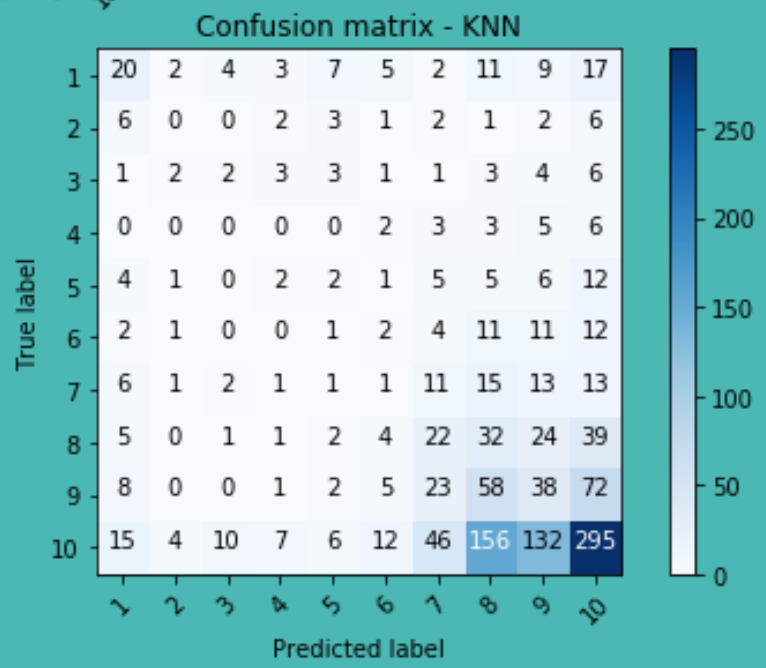
PERFORMANCE

	Train	Test
SVM	87%	56.00%
Naïve Bayes	51%	51%
KNN	76%	30%

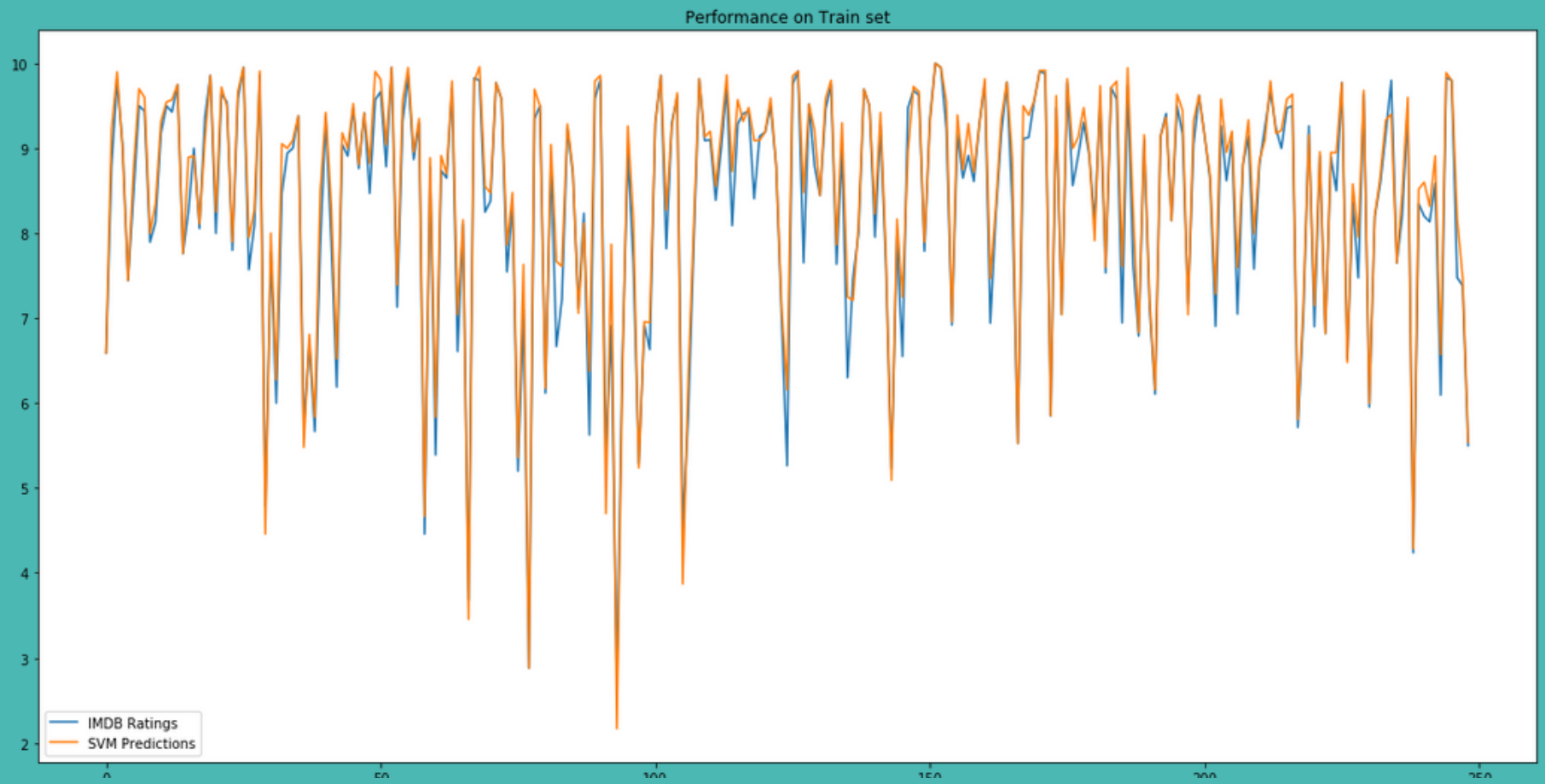
PERFORMANCE



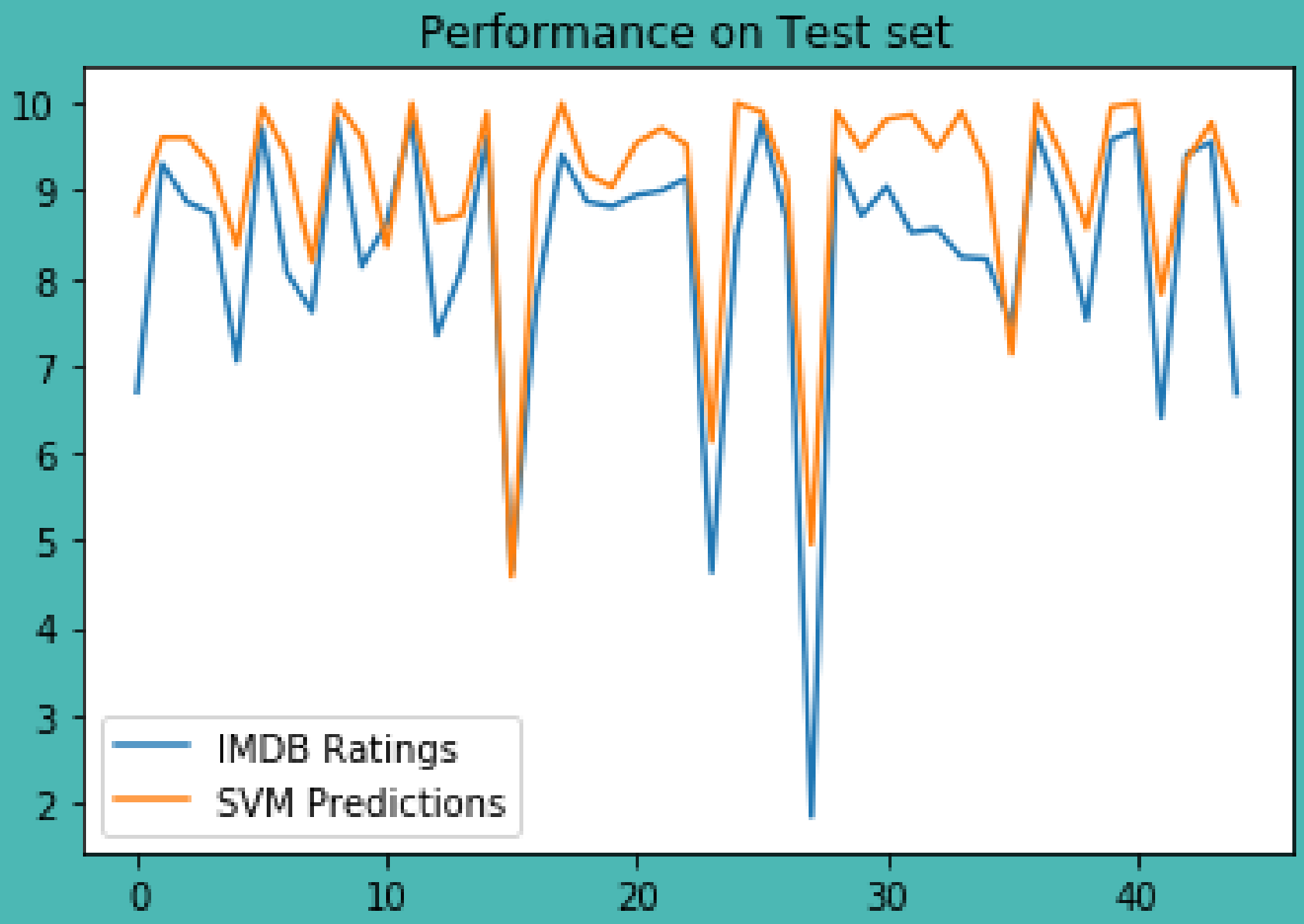




PERFORMANCE



PERFORMANCE



TEST DATA PERFORMANCE

	Accuracy	Precision	Recall	F1-Score
SVM	57%	47%	57%	48%
Naïve Bayes	50%	25%	50%	34%
KNN	33%	40%	33%	35%

BOOK MY SHOW

If we look at the robots.txt file of BookMyShow we'll see the following:

```
User-agent: *
Disallow: /
```

Thank You!