Movie Revenue Prediction

•••

Bairi Sandhya Rani	14CO205
Dokku Tejaswi Surya Sai Gayathri	14CO211

Introduction

- Box office revenue prediction is an important problem in the film industry.
- It governs financial decisions made by producers and investors.
- Basic statistical techniques are commonly used for this purpose.
- But they provide only a coarse estimate of revenue prediction.
- Hence machine learning techniques, specifically supervised learning techniques are now being used along with deep learning.

Dataset

- IMDB 5000 Movie Dataset downloaded from Kaggle
- Scraped from Internet Movie Database
- Consists of 5043 movies with 28 features/variables
- Variables present in the dataset are
    - Color, title of the movie, director’s name, number of critics for reviews, duration, names of actors
    - Facebook likes of movies, directors, actors and total cast
    - gross, genre, number of voted users, number of faces on poster, keywords in the plot, budget, IMDB link of the movie,
    - number of user for reviews, content rating, language, country, title year, IMDB Score, aspect ratio.

- The actual features trained and tested are :
- Director’s Facebook Likes, Number of Critics for Reviews, Actor 3’s Facebook
- Likes, Actor 2’s Facebook Likes, Actor 1’s Facebook Likes, Movie’s Facebook Likes, Duration, Number of voted users, Number of user for reviews, Budget, Year, IMDB Score, Total Facebook Likes.
- These features have continuous values so they are all normalized so that, they have values within range 0 to 1.

- The movie genres is considered as binary feature.
- If a movie has the following genre Action, Adventure, Mystery the feature value will be:

[A,A,F,S,T,C,F,H,W,A,W,R,M,D,D,H,B,M,C]
[1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0]

Action, Adventure, Fantasy, Sci-fi, Thriller, Comedy, Family, Horror, War, Animation, Western, Romance, Musical, Documentary, Drama, History, Biography, Mystery, Crime

Learning Techniques

Algorithms used to predict movie revenue are:

- Support Vector Machine - Linear kernel
- Support Vector Machine - RBF (Gaussian) kernel
- Support Vector Machine - Polynomial kernel
- Naive Bayes
- MultiLayer Perceptron
- Logistic Regression

Hyperparameters

- For SVM linear kernel, C=2 and gamma=2.
- For SVM polynomial kernel, C=2 and gamma=1.
- Here C is a regularization parameter trading the penalty for misclassification, gamma is a free parameter influencing the support vectors of the classes.
- For MLP, 1 hidden layer with 20 neurons is used.
- For activation function, ‘relu’, the rectified linear unit function is used,

$$f(x) = \max(0, x).$$

- For Logistic Regression, C = 1e-5 where C denotes the inverse of regularization strength.

Motivation

- During training, a classifier model learns the features of the training data.
- When tested on the same data, it repeats the same labels correctly but cannot predict when introduced to unseen data.
- This leads to overfitting wherein a function is too closely fit to a limited set of datapoints in a complex model with too many features that it fails to predict unknown labels with good accuracy.
- Holding out part of the training data for testing purpose doesn’t totally remove the risk of overfitting.

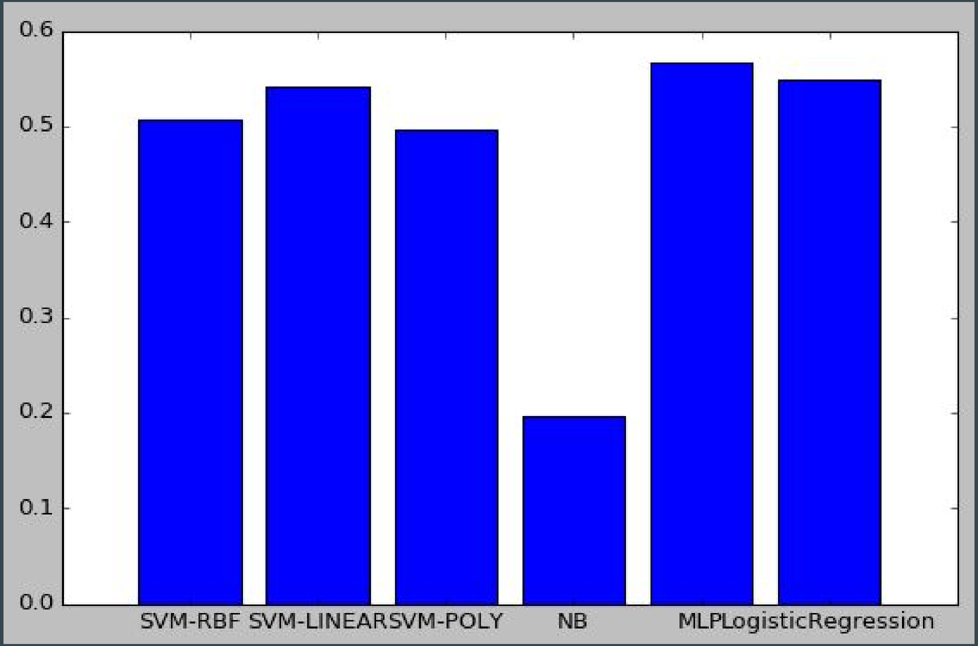
Cross Validation

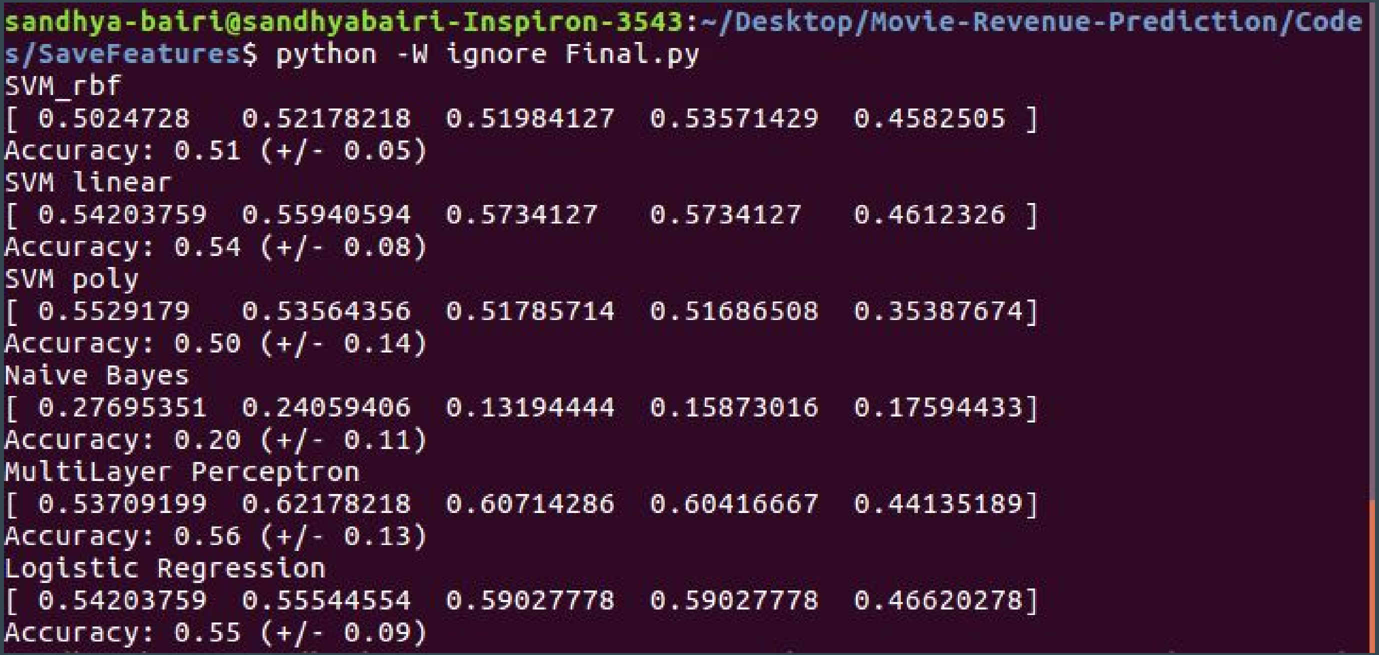
- Holding another part of the dataset as “validation set” so that training proceeds on the training set, then evaluation on the validation set, and when the performance seems optimal, final evaluation can be done on the test set.
- Partitioning the available data into three sets reduces the number of samples which can be used for learning the model.
- Cross-validation is a procedure in which a test set is held out for final evaluation.
- Validation set is no longer needed.

- For a k-fold CV, the training set is split into *k* smaller sets. The following procedure is followed for each of the k folds:
  - A model is trained using (k-1) of the folds as training data;
  - the resulting model is validated on the remaining part of the data
- In this project, 70% of the dataset is used for training purpose and 30% data for testing purpose.
- As this dataset is small and to compare the different models, 5-fold cross validation has been used
- Accuracy is calculated as the mean of the score values returned by each model.

Results

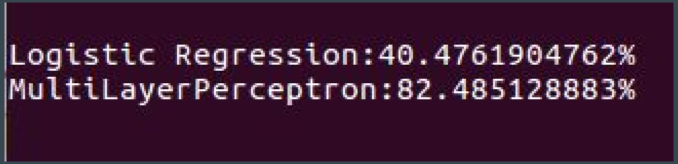
- Following figure shows accuracies of different models based on cross validation.
- Naive Bayes has the least accuracy and MLP has the highest followed by Logistic Regression.



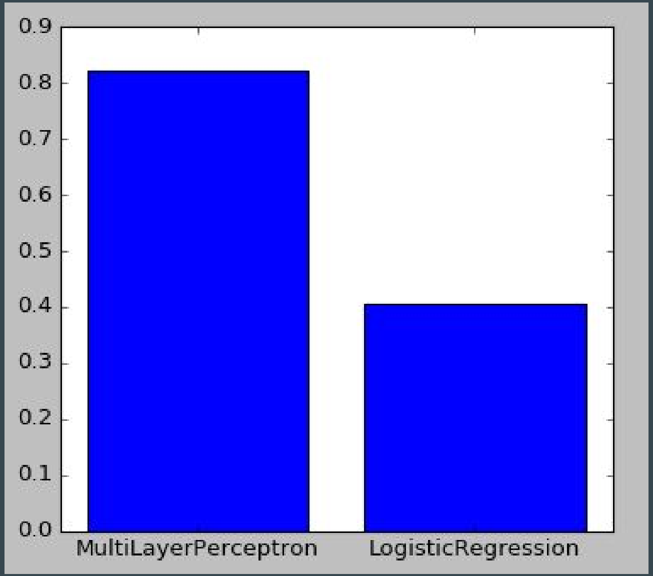


```
sandhya-bairi@sandhyabairi-Inspiron-3543:~/Desktop/Movie-Revenue-Prediction/Code
s/SaveFeatures$ python -W ignore Final.py
SVM_rbf
[ 0.5024728   0.52178218  0.51984127  0.53571429  0.4582505 ]
Accuracy: 0.51 (+/- 0.05)
SVM linear
[ 0.54203759  0.55940594  0.5734127   0.5734127   0.4612326 ]
Accuracy: 0.54 (+/- 0.08)
SVM poly
[ 0.5529179   0.53564356  0.51785714  0.51686508  0.35387674]
Accuracy: 0.50 (+/- 0.14)
Naive Bayes
[ 0.27695351  0.24059406  0.13194444  0.15873016  0.17594433]
Accuracy: 0.20 (+/- 0.11)
MultiLayer Perceptron
[ 0.53709199  0.62178218  0.60714286  0.60416667  0.44135189]
Accuracy: 0.56 (+/- 0.13)
Logistic Regression
[ 0.54203759  0.55544554  0.59027778  0.59027778  0.46620278]
Accuracy: 0.55 (+/- 0.09)
```

- Below figure shows accuracies of MLP and logistic regression in a better way.



```
Logistic Regression:40.4761904762%
MultiLayerPerceptron:82.485128883%
```



Conclusion

- Our approach uses K-fold cross validation to achieve better accuracy of prediction.
- Lower the K value, more is the bias
- Higher value of K is less biased, but can suffer from large variability
- Hence k must be appropriately chosen.
- Multilayer perceptrons method is found to show highest accuracy.
- Hence further work on revenue prediction can be based on neural networks.

THANK YOU