A Project Report on

Prediction of Box Office Revenue

National Institute of Technology Karnataka Surathkal

Department of Computer Engineering

November 2017

Submitted to
Dr. M. Venkatesan

Submitted by

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

1. Summary of the project

The aim of this project is to predict revenue of box office which is done by developing a model based on public data for movies extracted from popular online movie databases and social media.

Movie revenue prediction has been studied in variety of contexts ranging from economics and business to statistics and forecasting. Building a model by analyzing revenues generated by previous movies can help in predicting expected revenue for a movie. Such a prediction is very useful for producers of the movie, investors and theaters to estimate their revenues.

In order to accomplish this task firstly dataset is collected from kaggle and divided as training and testing data. Relevant features from the training data are used in the creation of a K-fold cross validation model using algorithms such as SVM, Naive Bayes, Multilayer perceptron and logistic regression. Accuracies of the algorithms used are compared and results are generated.

2. Approach used

Cross-validation for evaluating performance

Any prediction function when trained, would learn the features of the training data and, when tested on the same data repeats the same labels 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. But holding out part of the training data for testing purpose doesn't totally remove the risk of overfitting.

When different hyperparameters for the models are evaluated, for example, C parameter for SVM is manually set, risk of overfitting still exists on the test set because the parameters can be adjusted until the model becomes optimal. This may leak the test data knowledge which will no longer support a generalized performance. To solve this problem, another part of the dataset can be held out as "validation set" thereby training proceeds on the training set, then evaluation is done on the validation set, and when the performance seems optimal, final evaluation can be done on the test set. However, by partitioning the available data into three sets, number of samples which can be used for learning the model is drastically reduced, and the results can depend on a particular random choice for the pair of (train, validation) sets.

In order to avoid this, cross-validation procedure is used for which a test set is still held out for final evaluation but 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

The cross-validation score can be directly calculated using the cross_val_score helper. Given an estimator (here the prediction model used), the cross-validation object and the input dataset, the cross_val_score splits the data repeatedly into a training and a testing set, trains the estimator using the training set and computes the scores based on the testing set for each iteration of cross-validation.

2.1. Data mining technique

The above cross validation technique is used on several supervised machine learning techniques namely,

- Support Vector Machine - Linear Kernel
- Support Vector Machine - RBF Kernel
- Support Vector Machine - Polynomial Kernel
- Naive Bayes
- MultiLayer Perceptron
- Logistic Regression

Defining Class Label

The revenue generated by movie is a continuous value. But to classify discrete values are required. So to determine the label of a class, log10 of the gross feature values generated by the movie is taken to be the class label. So the class labels were 0 to 8 with 9 classes.

Architecture of Different Models

All the models mentioned earlier were implemented by using the Python Scikit-Learn Machine Learning library.

Technique	*Hyperparameters*
SVM Linear Kernel	C = 2, gamma = 2
SVM polynomial kernel	C = 2, gamma = 1
MLP	Hidden layers = 1, No. of neurons = 20 Activation function : ReLU (Rectified linear unit) f(x) = max(0,x)
Logistic Regression	c= 1e-5

2.2. Data set used

Dataset used: IMDB 5000 movie dataset.

Data set used in this project is taken from kaggle. This has metadata of more than 5000 movies that is scraped from IMDB. Values corresponding to 28 variables of 5043 movies is present in this data set. It has about 2000 unique directors as well as actors and actresses. Variables present in the data set are listed below:

Color, title of the movie, movie's facebook likes, director's name, number of critics for reviews, duration, director's facebook likes, names of actors and their facebook likes, gross, genre, number of voted user's, total facebook likes of the cast, 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 and aspect ratio.

All of the variables mentioned above aren't required for feature extraction. Normalization is made so that continuous values of the features would be converted to values in the range of 0 to 10. The mathematics behind trying to separate the classes in this space just gets drastically more complex, especially as the number of features and observations grow. Thus it is critical to always scale/normalize the data.

Features that are relevant for the revenue prediction are:

- Facebook likes of actors, movies and directors – This data represents the number of users that interested in the actor and the movie which in turn affects the revenue.
- Duration- Many people are particular about duration of a movie because of which duration should be taken into account to predict its revenue.
- Number of user for reviews, number of voted users and total facebook likes – This value represents the actual number of users that are aware of the movie and has some opinion about it, which would in turn affect the revenue made by the movie.
- IMDB score – Many users would prefer to watch a movie with good IMDB score. Hence, a movie with good IMDB score has higher chances of generating more income.
- Budget – The expectations of a movie with higher budget would be high which would result in more number of viewers. Hence this is a factor that affects revenue.
- Genre – Viewers use this as a deciding factor whether to watch the movie or not. If it's an action movie the audience would have more percentage of youth and if it's a mythological movie audience would have more percentage of elders. This factor plays a major role in predicting revenue of a movie.
- Content rating - This represents the viewers eligible to watch a movie which in turn decides the number of people that can watch it, thereby affecting revenue.

In the features mentioned above genre, language, content rating and country are considered as binary features for this project i.e it has either 0 or 1. Genre of the movie has value 1 for categories action, adventure and mystery and 0 for the rest. Content rating has the value of 1 for parental guidance

suggested and 0 for the remaining. Language has value of 1 if it's English and 0 for the remaining. Country feature has a value of 1 if it is USA and 0 for the remaining.

3. Results and Discussions

Several models from sklearn namely, support vector machine using different kernels like rbf (radial basis function, also Gaussian), linear and polynomial are used for testing on the previously mentioned dataset. 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. Figure 1 shows five score values for five folds for different classifier models. In addition to the scores, accuracy is calculated as the mean of the corresponding scores along with their standard deviation (in the braces).

```
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)
```

Figure 1. Accuracies of different prediction models

Figure 2 shows accuracies of different models plotted using matplotlib library. It shows that naive bayes has least accuracy with respect to our dataset and the features that it included. Logistic Regression and MultiLinear Perceptron gave better accuracies.

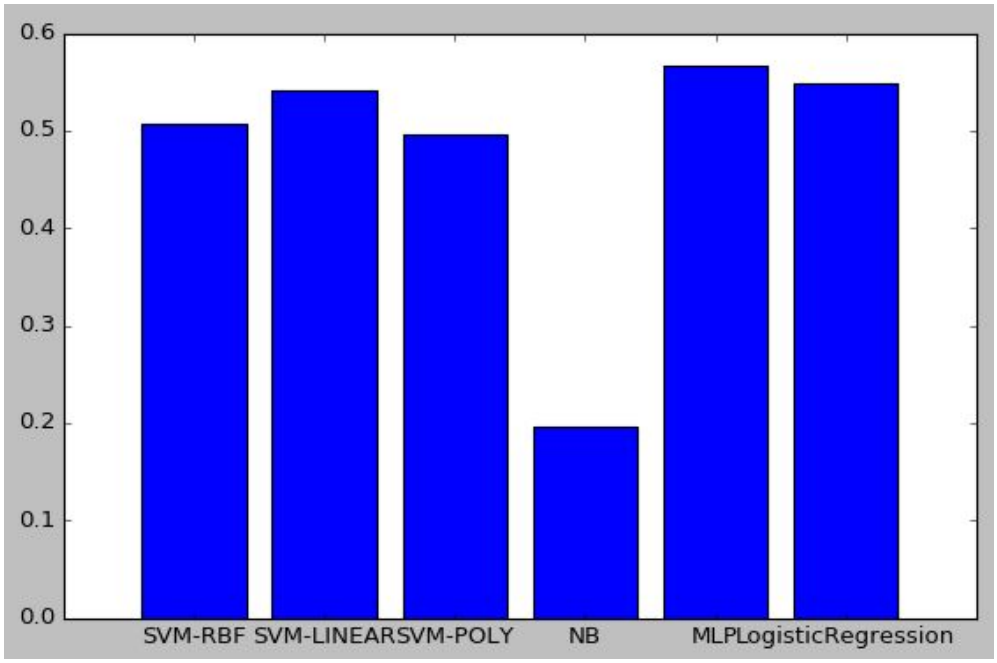
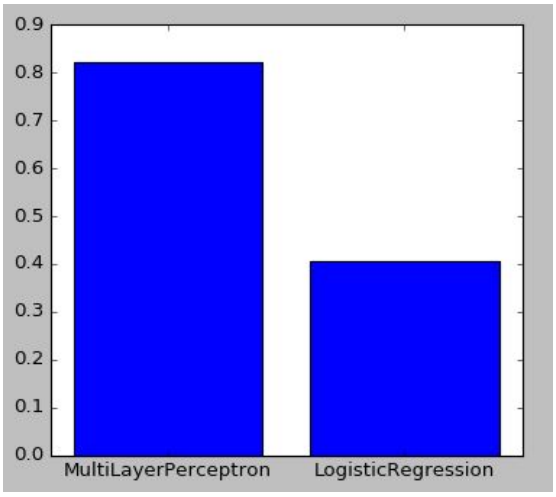


Figure 2. Graphical plot of accuracies of different prediction models

For a better comparison of Logistic regression and MLP, using scikit-learn's model_selection module, the respective models of logistic regression and MLP are fitted and used to make predictions. Figure 3 shows those accuracies.

(a)



(b)

Figure 3. Closer comparison between MLP and Logistic Regression

4. Conclusion

Predicting revenue of box office is a major issue in the film industry which decides financial decisions made by investors and producers. A k-fold cross validation model is built to accomplish the purpose of predicting box office revenue. This model is built using different algorithms among which Multilayer perceptron has the highest accuracy. Hence different neural networks can be used further with more number of hidden layers accounting for more features and bigger datasets to get better performance.

References

[1] Benjamin Flora, Thomas Lampo, and Lili Yang, *Predicting Movie Revenue from Pre-Release Data*, December 12, 2015

[2] Jason van der Merwe, Bridge Eimon, *Predicting Movie Box Office Gross*, December 8, 2013

[3] Ramesh Sharda, Dursun Delen, *Predicting box-office success of motion pictures with neural networks,* February 2006

[4] Darin Im and Minh Thao Nguyen, *Predicting Box-Office Success of the Movies in the U.S. Market,* 2011

[5] Matt Vitelli, *Predicting Box Office Revenue for Movies*

[6] http://scikit-learn.org/stable/

Link for Dataset : https://drive.google.com/open?id=1fD8-W677VDRwVyZEDZnjAZX5N8MlP7iI

Link for Presentation:
https://drive.google.com/open?id=1G3zbu4UpPufrLPWsfdwvyRz-kkbcKrhzJAAsVQJh8U0