Literature Review

Team: The Miner League

SRN: PES2201800025

Name: Manav Agarwal

PAPER2:

V. Subramaniyaswamy ; M. Viginesh Vaibhav ; R. Vishnu Prasad ; R. Logesh, 2017, Predicting Movie Box Office Success using Multiple Regression and SVM, IEEE,

Predicting movie box office success

This paper follows a very similar process to the first paper in performing regression analysis first and then using SVM on the dataset. In this case they have considered the ROI (Return of Investment) as the major contributing factor in the model, along
with various related features to it which included Opening Date, Movie Name ,Budget, Domestic Gross, International Gross, Total Gross, Trailer Views, Studio , Cast and Crew, Genre, Medium (Live action or Film), Trailer Views, Wikipedia Views and Rotten Tomatoes Score. Data pruning is then performed to remove all the junk or unnecessary values from the dataset.

We classified the movies in the dataset as part of 3 categories:

- Low Budget Films (Budget < $50 million)
- Medium Budget Films (Budget between $50 million and $150 million) • Big Budget Films (Budget >$150 million)

Each of them have a separate multiplier for them such as 2x, 2.5x and 3x (can vary). This is to account for the increased marketing costs as the budget increases, meaning the studio needs a larger ROI to break even. A timeframe is then taken and the graph is plotted to see the yearly patterns

for critic scores, wikipedia views, trailers, and total gross. These help in plotting the adjusted ROI.

The first method that is used is the technique of linear regression, first SLR for the individual comparison of a few attributes, but the main technique used is the concept of MLR (Multiple linear regression). They applied multiple regression to predict
the success of the movie, with the independent variables being Trailer Views, Budget, Critic Ratings and Wikipedia Page Views. From finding the t-statistic and the p-value, certain conclusions are drawn such as: We see that budget has the highest t-Stat value of 8.26, and RT scores with lowest t-stat value (3.33) and highest
p-value (0.0011) indicating that it has the least impact on accurate prediction, which is consistent with previous findings.

SVM is the next technique used for regression analysis and classification of the attributes. In SVM, the models consist of points in space, where each point represents an example. Mapping is done in a way that tries to maximize the difference between examples from different classes, so that the examples of a particular class are very different from examples from another class. New examples are then as-signed to a class based on whichever side they are closer to. Making the class or category have high intra-class similarity but low inter-class similarity is a goal of SVM. Here, they have used 10-fold cross-validation on their data.
SVM was used to train multiple variables and the result was an accurate classification rate of 56.52%, which is a higher accuracy than previous attempts at SVM using only a single variable.