EECS349 machine learning
Final report
Lingtao Shui & Xingwei Li

Movies' revenue and popularity prediction

American film studios collectively produce several hundred movies every year, making the United States the third most prolific producer of films in the world. The budget of these movies is of the order of hundreds of millions of dollars, making their box office success is absolutely essential for the survival of the industry. Which film will be highly rated? Whether or not they are a commercial success. Given that major films costing over $100 million to produce still flop, this question arouse our interests.

As a result, we want to predict the revenue and popularity of a movie based on attributes of the movie. Knowing which movies are likely to succeed and which are likely to fail before the release could benefit the production houses greatly as it will enable them to focus their advertising campaigns which itself cost millions of dollars, accordingly.

We used the first 1000 instances of TMDB 5000 Movie Dataset as our training dataset. We divided revenue values into 10 classes and we also divided popularity values into 10 classes. Revenue and popularity were what we want to predict. We chose vote average, runtime, budget, genres (set each genre as an independent attribute), production companies (set each company as an independent attribute) as the attributes. Some of the attributes were saved as dictionaries in the original dataset, which cannot be processed by the Weka. So, we made keys in dictionaries as separate attributes. We used 1000 instances in total. Each instance had 36 attributes. We randomly choose 800 examples as the training set and other 200 examples are used as the test set (10 folds Cross-Validation). Lingtao mainly did the data collection and pre-processing part, and Xingwei mainly did the model training and data analysis part.

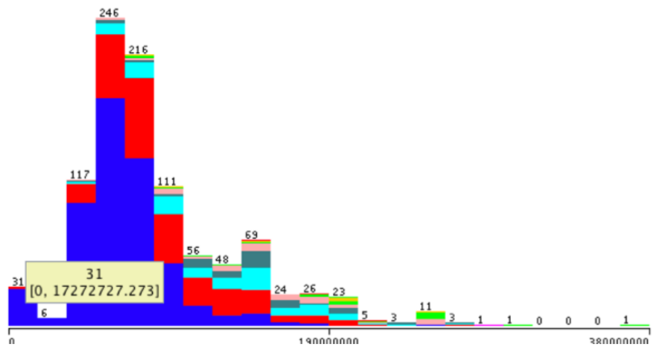

Figure 1. The distribution of budget attribute for each revenue classification

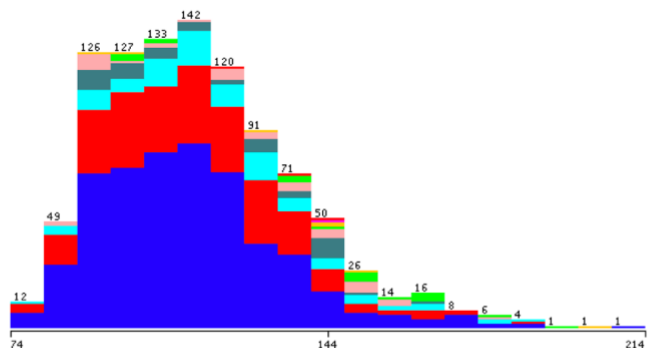

Figure 2. The distribution of runtime attribute for each revenue classification

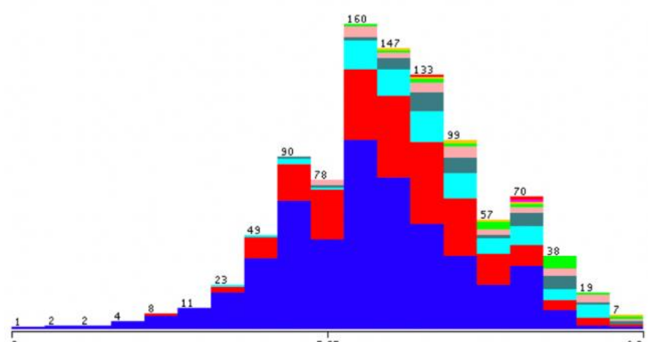

Figure 3. The distribution of vote average attribute for each revenue classification

We started with predicting revenues of movies in our dataset. We used Weka to generate various approximate hypothesis for our prediction function. We began by running ZeroR on our data with 10-fold cross validation and established a baseline performance of 53.7%. We continued to train and evaluate on our dataset with 10-fold cross validation on a wide range of classifiers, however, we were unsuccessful in generating a significantly more accurate prediction function. Our failure to generate a viable revenue prediction algorithm was disappointing. We attempted to modify our dataset by omitting certain attributes or aggregating other attributes. Ultimately, we were unable to come up with more meaningful results. We thought it was because our current attributes do not contain sufficient information to determine a movie's revenue. We believed actors and the director would significantly impact the revenue of a movie.

Classifier	Classifier Type	Accuracy
ZeroR	Rules	53.8%
NaiveBayes	Bayes	60.6%
BayesNet	Bayes	61.1%
IBK	Lazy	49.7%
KStar	Lazy	51.2%
MultilayerPerceptron	Functions	52.2%
Logistic	Functions	58.8%
AdaBoostM1	Meta	58.1%
MulticlassClassifier	Meta	59.7%
DecisionTable	Rules	57.5%
JRip	Rules	56.8%
DecisionStump	Trees	57.8%
J48	Trees	57.4%
RandomTree	Trees	50.2%

Figure 4. Performances of algorithms on predicting movies' revenue

Then, we perform our prediction on movies' popularity. We also used Weka to generate various approximate hypothesis for our prediction function. We began by running ZeroR on our data with 10-fold cross validation and established a baseline performance of 80.4%. We got relatively viable results this time. According to our results shown below, using Multilayer Perceptron can generate the best prediction model for popularity prediction.

Classifier	Classifier Type	Accuracy
ZeroR	Rules	80.4%
NaiveBayes	Bayes	89.6%
BayesNet	Bayes	89.2%
IBK	Lazy	89.7%
KStar	Lazy	90.3%
MultilayerPerceptron	Functions	91.3%
Logistic	Functions	88.5%
AdaBoostM1	Meta	90.4%
MulticlassClassifier	Meta	90.1%
DecisionTable	Rules	91.3%
JRip	Rules	90.6%
DecisionStump	Trees	90.3%
J48	Trees	90.6%
RandomTree	Trees	88.5%

Figure 5. Performances of algorithms on predicting movies' popularity