

Predicting Movie Revenue from IMDb Data

Steven Yoo, Robert Kanter, David Cummings
TA: Andrew Maas

1. Introduction

Given the information known about a movie in the week of its release, can we predict the total gross revenue for that movie? Such information would be useful to marketers, theater operators, and others in the movie industry, but it is a hard problem, even for human beings. We found that, given a set of numeric, text-based, and sentiment features from IMDb, linear regression outperforms class-based logistic regression at predicting gross revenue. However, neither gives sufficiently precise results to be used in practice.

2. Data and Features

We collected data from IMDb for 4052 movies that were released from 1913 to 2011. Among these movies, we only selected the ones that were released in United States and are in English, in the anticipation that we would be able to make more accurate predictions on these movies given that their reviews would also be in English.

The features we used are listed in Table 1 with examples. They are separated into three main sets of features: “simple”, “complex”, and “sentiment” (“simple” is numeric only, “complex” is numeric and text-based, and “sentiment” includes all). Sentiment scores are computed using the subjectivity lexicon obtained from the University of Pittsburgh OpinionFinder project.[4][5] To generate a sentiment score for a movie, we simply take the sum of all sentiment scores for all words in the first week’s worth of reviews for that movie. Some examples of this lexicon are listed in Table 2 with their score.

Feature Categories	Features	Examples
Numeric Features	Days since January 1, 1900	32,872
	Days since January 1 of that year	346
	Duration (minutes)	90
	Budget (\$ USD)	5,000,000
	Aspect-ratio (X:1)	2.1
	Average rating	6.6
	User votes count	56,000
	User review count	531
	Critic review count	93
Text-based Features	MPIAA Rating	G, PG, PG-13, R, unrated/unknown
	Directors	Woody Allen, Steven Spielberg
	Plot keywords	love, murder, friend, police, etc
	Cast	Samuel L, Jackson, Robin Williams, etc
	Genres	comedy, romance, thriller, etc.
Sentiment Feature	Sentiment score	569, -20

Table 1: Features and examples

Sentiment Polarity	Sentiment Score	Examples
Strong Positive	+2	Beautiful
Weak Positive	+1	Trendy
Neutral	0	Finally
Weak Negative	-1	Dwindling
Strong Negative	-2	Awful

Table 2: Sentiment scores and examples

In our experiments, we found that the budget is the strongest individual indicator of a movie’s eventual gross revenue. The X-Y scatter graph of budget and revenue is shown in Figure 1; we found their correlation to be 0.6269. Accordingly, we used this correlation as our baseline result: all reasonable models should be able to achieve at least a correlation of 0.6269.

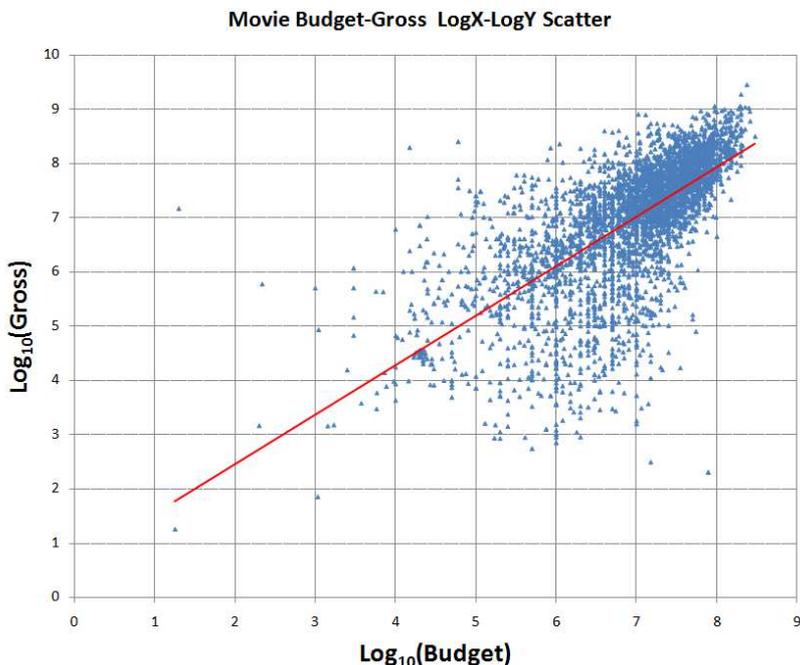


Figure 1: Correlation of budget with actual revenue

3. Model 1: Linear Regression

In our first model, we used a standard least-squares linear regression. To do this, we used stochastic gradient descent, which we implemented in C for efficiency. Once we had trained a set of feature weights, we could then generate gross revenue predictions as follows:

$$Gross = \theta_0 + \theta_1 * F_1 + \theta_2 * F_2 + \dots + \theta_n * F_n$$

where θ_i are the weights, F_i are the features, and n is the number of features

To measure the “goodness” of our results, we looked at the correlation between our predicted revenue values and the actual revenue values, as was done in other papers[1][2]. As alternative measures to interpret our results, we also considered other metrics such as mean absolute percentage error (MAPE) and symmetric mean absolute percentage error (SMAPE).

MAPE is not the best metric for our work because the error is unbounded-- for instance, if we predict revenue of \$1,000,000 for a movie that grossed \$100,000, then error would be 900%, skewing any average we would take over all test examples. In compensation to this, we also tried SMAPE, which returns error values from 0 to 100%. However, neither metric gave consistent, explainable results, and to our best knowledge there are no other papers using these metrics for the same task, so we could not compare the results to others with them.

Below are the correlation results we found for each of our feature sets, using 90% of our data in training and 10% in test:

	Simple features	Complex features	Sentiment features
Test Data Set	0.7218 correlation	0.7480 correlation	0.7479 correlation
Training Data Set	0.7390 correlation	0.8063 correlation	0.8095 correlation

Table 3: Correlation results

We were only able to achieve a correlation of 0.7479 on the test data set; while this is better than the baseline result (i.e. 0.6269), we do not consider it high enough to be useful in practice. Figure 2 shows this result, comparing our predictions to the true gross revenues. Additionally, these results also show that our larger feature sets generally improved performance relative to the smaller feature sets.

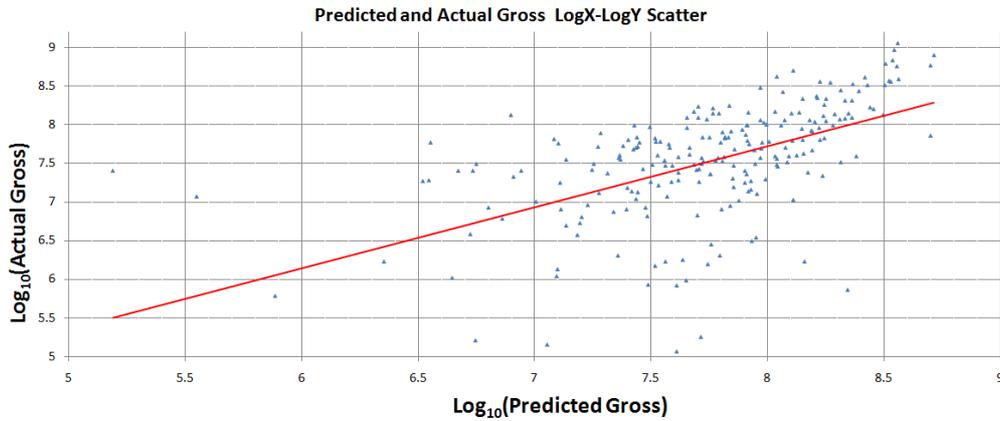


Figure 2: Correlation of predicted with actual revenue

While we observed an increase in correlation with the addition of text-based features in the “complex” feature set, the addition of sentiment scores did not significantly affect the correlation on the test or training data. Two factors can explain this: first, sentiment about a movie is already partially captured by the “rating” feature. Second, the gross for a movie is more directly related to the number of people who watch the movie rather than how good people think the movie actually is. In fact, we saw that features such as the number of user votes on a movie’s rating are actually more important than the average rating itself. For example, Transformers has a low rating (and “bad” reviews), but the actual gross is high because many people watched it.

In an attempt to counteract over-fitting, we terminate our stochastic gradient descent early, having noted at the cross-validation stage that 1000 iterations gave better results on test correlation than waiting until full convergence.

4. Model 2: Classification by Logistic Regression

As a second model, we also tried classification by standard L1-regularized logistic regression. We chose this method because it generated a multi-class model with linear weights, most directly comparable to the feature weights given by linear regression. To define our classes, we drew a histogram of movie revenues to create 5 different buckets for prediction as shown in Table 4. The first bucket includes the lowest 20% of the gross distribution and the last bucket includes the highest 20%.

Buckets (classes)	Bucket 1	Bucket 2	Bucket 3	Bucket 4	Bucket 5
Gross Ranges(\$)	0 to 1.3M	1.3M to 10.4M	10.4M to 28.7M	28.7M to 74.6M	74.6M

Table 4: Bucket ranges for gross classification

The procedure for using logistic regression was fairly similar to that of linear regression; the difference being that we now use labeled buckets as our y-values (instead of real-valued gross revenue numbers) and pass the data to liblinear to build the model for classification. This model gave the following accuracy results on our 10% test set:

	Simple features	Complex features	Sentiment Features
Test Data Set	48.15%	49.14%	48.40%

Table 5: Accuracy results on test data set

In general, none of these accuracy figures were as high as we had hoped, indicating that this kind of classification was not the right approach to the problem.

5. Comparing Performance

Having developed these two different models (linear regression and classification by logistic regression), we needed some way of comparing their results. For this project, we implemented two such methods.

In the first method, we map the results from linear regression into the five bucket classes from logistic regression. To do this, we take the real-valued outputs from our linear regression model, assign labels to them according to the buckets into which they fall, and check whether these correspond to the same buckets as those of the actual gross revenue. For this measure, we generated the following results on our different feature sets:

	Simple features	Complex features	Sentiment Features
Test Data Set	48.38%	45.18%	43.70%

Table 6: Mapped accuracy on test data set

These numbers decrease with additional features, likely because of increased variance (that is, some over-fitting on the training set-- further discussion of this in the following section). However, all are roughly comparable to the 48-49% accuracy achieved by logistic regression on the classification problem, showing that linear regression is almost as good at the task of classification as logistic regression, the algorithm dedicated to classification.

In the second method, instead of mapping from real values to buckets, we map from our five buckets to real values. To do this, we first find the average actual revenue of movies classified into each bucket in logistic regression. Then, instead of generating class labels from logistic regression, we can use the corresponding averages instead, giving us real-valued output from the classifier. The resulting correlation scores are shown below:

	Simple features	Complex features	Sentiment Features
Test Data Set	0.5126 correlation	0.5576 correlation	0.5569 correlation

Table 7: Mapped correlation on test data set

None of these approach the 0.75 range of correlation seen with linear regression, much less the 0.63 baseline correlation using a movie's budget alone. Much of this has to do with the fact that logistic regression can only generate one of five distinct values, so we thought we might experiment with different numbers of buckets. Our expectation was that accuracy would consistently decrease as number of buckets increased, but that correlation would have some optimal point where the (positive) granularity of having smaller-ranged buckets balanced with the (negative) trend toward fewer training examples per bucket. We were surprised to find that while accuracy decreased, correlation remained fairly constant, as shown in Figure 3:

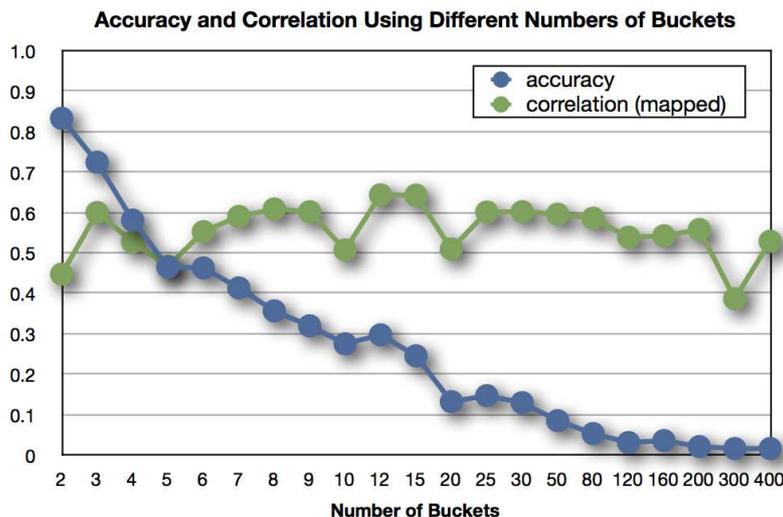


Figure 3: Accuracy and correlation using different numbers of buckets

It appears that, in terms of the correlation measure, having fewer training examples per bucket was evenly offset by the greater granularity of having smaller-ranged buckets.

6. Conclusions

We framed this problem as both a regression and classification problem because we were not sure which would provide a better result; as such, we implemented both and devised methods to compare them. In general, we found that linear regression works almost as well as logistic regression for classification on our data, while having a much better correlation with the actual gross revenues.

In general, we found that the features we used (simple numeric, text, and sentiment features) were insufficient to make strong predictions of gross revenue. Others have had greater success using additional features such as number of theaters, marketing budget, etc., but since IMDb does not contain such data, we were unable to include them. For future work, besides using different feature sets, we might consider using better regularization on linear regression in order to provide a more rigorous safeguard against high-variance models, as we consistently observed decreases in linear regression's test accuracy with increasing numbers of features.

Another, fundamentally different, data set that might be useful in predicting movie revenue would be social graph data: using such data, we could analyze the characteristics of how a movie's popularity propagates through social networks, as well as characteristics of the propagation tree, such as its speed and extent over time. The propagation speed of a movie would represent people's expectation to see that movie, which we expect will be directly related to its gross revenue.

7. Epilogue

For fun, we looked at our feature weights for text-based features, and from them extracted the highest- and lowest-weighted features in MPAA rating, director, plot keywords, cast, and genre. For a guaranteed hit film, we recommend a G-rated animated family adventure about 'escape', 'prince', and 'vampire', directed by Steven Spielberg and starring Bill Paxton and Tom Cruise. For a guaranteed flop, we recommend an R-rated mystery sci-fi musical about 'Vietnam', 'computer', and 'train', directed by Stephen King and starring Giovanni Ribisi and Robin Wright.

References

- [1] M. Joshi, D. Das, K. Gimpel, and N. A. Smith. Movie reviews and revenues: An experiment in text regression. In Proceedings of NAACL-HLT, 2010.
- [2] W. Zhang and S. Skiena. Improving movie gross prediction through news analysis. Web Intelligence and Intelligent Agent Technology, IEEE/WIC/ACM International Conference on, 1:301–304, 2009.
This paper use AMAPE (Adjusted Mean Absolute Percentage/Relative Error) for their measurement
- [3] Simonoff, Jeffrey S. & Ilana R. Sparrow. Predicting Movie Grosses: Winners and Losers, Blockbusters and Sleepers. Chance Magazine, 13 (3), 15-24, 2000
- [4] Brendan O'Connor, et al. "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series." Proceedings of the International AAAI Conference on Weblogs and Social Media, 2011
- [5] Leonid Velikovic, et al. "The viability of webderived polarity lexicons." NAACL, 2010.