Predicting Box-office Earnings with Data on the Web

Yu-Han Chen yuhan@mit.edu 913279965
https://github.com/yuhan210/box-office-predictor

Introduction

In the United States, the film industry is a multi-billion dollar market, generating more than 10 billion dollars of revenue annually [1]. As profitable as it may seem, not all movies become the blockbuster. In fact, it is estimated that only 6% of films account for 80% of the industry's profits over the past decade; 78% of movies lost money over that period [2]. On the other hand, with the accessibility of movie information on the Internet, it becomes possible to mine the determinant of a blockbuster.

In this project, we use information presented on the web to understand the factors that influence the box-office earnings, the indicators that have strong predictive power on the earnings, and devise a novel regression tree algorithm to predict the box-office earnings. Comparing to a naïve linear regression using all the features, our regression algorithm improves the prediction accuracy from 27.7% to 66%.

1. Data Collection and Features

In this section, we explain how we collect our dataset for the analysis and the features used in the experiment.

Since there is no single website that provides all the movie information that we need, we crawl and parse data from multiple sources. First, we download the box-office earning information for the top-500 movies between 2014 and 2015 on BoxOffice®. Then, given the movie title, we use the IMDb API to retrieve detailed information related to the movie such as the genre, director, MPAA rating, and the budget. Considering the averaged director box-office earnings might be useful for our analysis, we write scripts which compose the IMDb web address for each director, download the webpage, and parse the source file to obtain his past movies. Then, based on the movie title, we download the movie webpage on IMDb and extract the total box-office earnings. Table 1 shows the features used in our analysis:

Features	Sources	Percentage of missing values (%)	Data types
Total box-office earning	BoxOffice®	0	Numerical
Total theaters	BoxOffice®	0.2	Numerical
Opening earning	BoxOffice®	2.8	Numerical
Opening theaters	BoxOffice®	2.6	Numerical
Genre	IMDb	3	Categorical (24 categories, not mutually exclusive)
Runtime	IMDb	5	Numerical
MPAA rating	IMDb	0	Categorical (4 categories)
Budget	BoxOffice®	62.6	Numerical
Averaged director box-office earning history	IMDb (derived)	63.4	Numerical

Table 1. Features used in the analysis and their source, missing percentage, and data type.

The total box-office earning is the number we predict. Total theaters indicate the number of theaters that the movie got released. Opening earning and opening theaters indicate the earning and the number of theaters that the movie got released in the first opening weekend. The genre describes the type of the movie, such as action, adventure, drama, and romance; the runtime is the length of the movie. The MPAA ratings consist of four categories: R, PG, PG-13, and Unrated. These ratings help to assess the degree of sexual content, violence, and adult language for any movie before its theatrical release. The budget an the number for the production plus advertising budget. It is an estimated number since the budgets are usually industry trade secret. And the averaged director box-office earning history is a derived feature, which computes the averaged total box-office earnings of movies a director has directed in the past. In our implementation, for each director, we use his last 10 movies (at most 10), and do not include the earnings of the movies appeared in our dataset.

2. Exploring the Data

It is crucial to understand the property of the data before we start analyzing it. In this section, we explore our data and discover two challenges that we need to address while designing our algorithm – missing features and high dimensional categorical features.

Missing features: The data collected from multiple sources is imperfect and missing feature is quite prevalent. To ensure our following analysis is not biased due to this issue, we quantify the severity of missing values (as shown in the third column in Table 1) for each of the features. We find that the first seven features listed in the table have low missing percentage, and the missing values occur at random; however, the last two features (budget and the averaged director earnings) have very high missing percentage, and less well-known movies seem to be the ones that do not have that information.

Therefore, for features with low missing percentage (<= 5%), we remove data entries without that feature. And for features with high missing percentage, we *embrace* it as a fact and design our algorithm to adapt to the set of features that the movie has (explained in Sec. 4(c)). After dropping movie entries without features with low missing percentage (first 7 features in Table 1), we have 459 movies for our analysis.

High dimensional and non-mutually exclusive categorical features: We have two categorical features – the MPAA rating and the genre. One common way to handle categorical data is to create dummy variables, and for each categorical feature with *k* categories (k > 2), we create *k-1* dummy variables. This approach works for the MPAA rating since each movie only has one rating; however, the genre of the movie contains non-mutually exclusive values (for example, the genres of the American Sniper are Action, Biography, Thriller, and War), and there are 24 (k = 24) genres in total. Figure 1 shows the histogram of the number of genre types for each movie. Most of the movies belong to at least 2 genres. To specify the genre combination, we have to use *k* variables instead of *k-1* to present all the possible combinations. Besides, encoding one feature (i.e., genre) with 24 variables creates redundancy and increases the number of parameters. We analyze the relationships between genre types using association rule mining and explain how we reduce the feature dimension in Sec. 4(b).

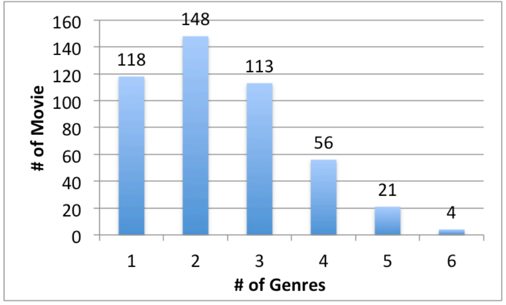

Figure 1. The histogram of number of genre types in our dataset.

3. Evaluation Metrics

We introduce the two evaluation metrics that we will be using throughout the analysis.

(a) **Root-mean-square error (RMSE).** RMSE is a frequently used measure of the differences between values predicted by a model and the values actually reported. We use RMSE to measure the differences of the estimated box-office earnings from the reported earnings.

(b) **Classification error.** Since the ranges of box-office earnings spread widely (from 10^4 to 10^8), the RMSE might be dominated by the blockbuster earnings. Besides, in reality, a movie producer might not need to know the exact number of the total earnings but the possible range of the earnings. Therefore, similar to [3], we split the earnings into buckets from 'flop' to 'blockbuster', and compute the classification error. Table 2 shows the buckets/classes.

Classes/Buckets	1	2	3	4	5	6	7	8	9
Range (in Millions)	e < 1	1 < e < 10	10 < e < 20	20 < e < 40	40 < e < 65	65 < e < 100	100 < e < 150	150 < e < 200	200 < e

Table 2. The buckets/classes used to evaluate the prediction accuracy. e indicated the estimated/actual total box-office earning.

4. Predicting Box-office Earnings Using Movie Features

In Sec. 2, we mentioned that missing features and high dimensional categorical features are the main issues in our dataset. Following the observation, we categorize all the features into three categories based on its missing percentage and the data type. In this section, we explain how we handle each of the categories and construct our final regression tree algorithm.

Note that we split the data into training and testing dataset using a 60%: 40% ratio. We freeze the testing data and do all the analysis on the training set.

(a) Handling numerical features with low missing percentage

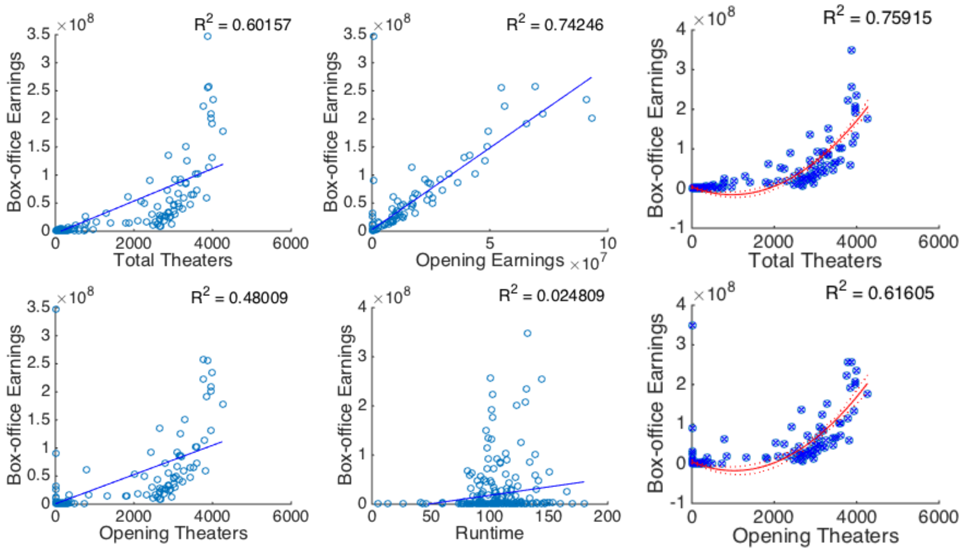

Figure 2 (a) Figure 2(b)

Figure 2(a). The scatterplots of numerical features against box-office earnings and the R-squared value using linear regression. Figure 2(b) Linear regression with quadratic terms on the number of total/opening theaters against the total box-office earnings.

We first analyze the predictive power of the numerical features with low missing percentage on the box-office earnings. There are four such features: The total number of theater, the opening box-office earnings, the total number of opening theaters, and the runtime of the movie. Figure 2(a) shows the scatterplots for these features against the total box-office earnings. Based on the fitted results and the R^2 value, we make three observations: (1) Movie runtime has little predictive power ($R^2 = 0.02$) on the box-office earnings. (2) The relationship between the opening earning ($R^2 = 0.74$) and the total earning is strong. (3) The total number of theaters has stronger predictive power than the number of opening theaters. And the scatterplots concave up indicating their relationships might not be linear but quadratic. Indeed, when we use a quadratic model to fit the number of total and opening theaters (as shown in Figure 2(b)), the R^2 value increases to 0.76 (from 0.6) and 0.62 (from 0.48), respectively. It seems worthwhile to introduce quadratic terms in the model. Therefore, we discard the runtime feature and use a quadratic model to fit the number of total and the number of opening theaters. The following table shows the impact of these optimizations. Removing runtime reduces the number of parameters by 1 and slightly improves the prediction error; adding quadratic terms reduces the classification error by 6% and adds two more parameters.

Model	**Classification Error (%)**	**RMSE (10^6)**	**# of Parameters**
Model(opening earning + total theater + opening theaters + runtime)	72.87	17.99	5
Model(opening earning + total theater + opening theaters)	71.28	17.85	4
Model(opening earning + total theaters**^2** + opening theaters**^2**)	65.43	13.86	6

(b) **Handling categorical features with low missing percentage**

We have two categorical features: the MPAA rating and the genre. Both of them have low missing percentages and missing feature is not an issue. However, as mentioned earlier, the genre is non-mutually exclusive and might introduce redundant dummy variables. In this section, we explain how we process these two categorical features.

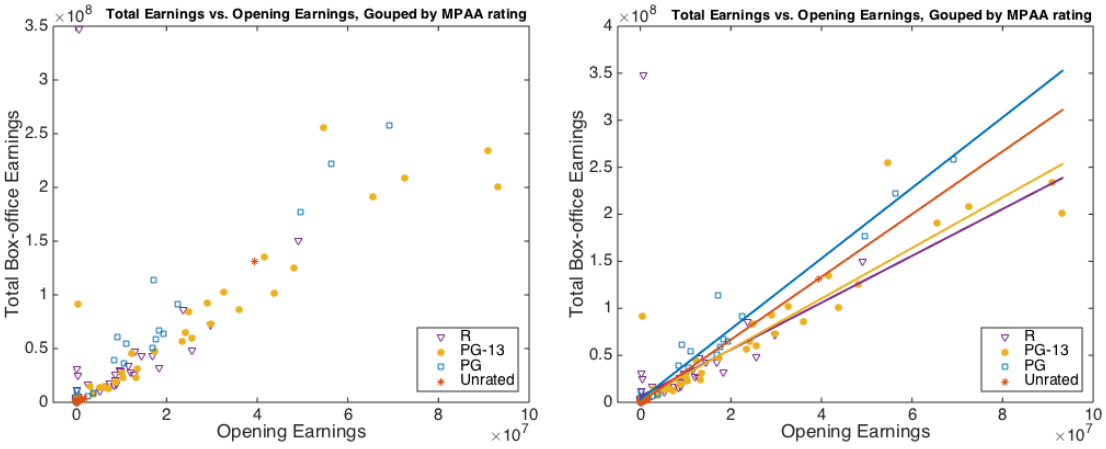

Figure 3(a) Figure 3(b)

Figure 3(a). The scatterplot of total earnings against opening earnings, grouped by MPAA rating. Figure 3(b) The slope of opening earnings against total earnings differ for each MPAA rating.

Mutually exclusive categorical features – MPAA rating: Each movie has exactly one MPAA rating: PG, PG-13, R, or unrated. Figure 3(a) shows the scatterplot of total earnings against opening earnings (the feature obtaining the highest R^2 value in Sec. 4(a)), grouped by MPAA rating. The scatter plot suggests that *the slope of total earnings against the opening earnings might differ for each MPAA rating*. For example, the slope is smaller for movies rated PG than movies rated PG-13. Besides, the total earnings for the unrated movies are small.

To assess this, we include opening earnings-rating interaction terms and compare it with using the opening earnings alone to predict the total earnings. The proposed model is:

$$\begin{aligned} Total\ earnings = {} & \beta_0 + \beta_1 OE + \beta_2 I[\text{PG13}] + \beta_3 I[\text{PG}] + \beta_4 I[\text{Unrated}] \\ & + \beta_5\ OE{\times}I[\text{PG13}] + \beta_6\ OE{\times}I[\text{PG}] + \beta_7 OE{\times}I[\text{Unrated}] \end{aligned}$$

where OE indicates opening earnings and $I[\text{PG13}]$, $I[\text{PG}]$, and $I[\text{Unrated}]$ are dummy variables indicating the MPAA ratings PF-13, PG, and Unrated. $I[\text{PG13}]$ takes the value 1 if the MPAA rating is PG-13 and takes the value 0 if it is not. In this model, rate R is the reference rating. We first test if there are significant differences between the slopes. This is equivalent to testing the hypothesis:

$$\text{H}_0: \ \beta_5 = \beta_6 = \beta_7 = 0$$
$$\text{H}_\text{A}: \ \beta_i \ \neq 0 \ \text{ for at least one } i.$$

Running the anova test on the proposed model, we obtain the following results:

```
                        SumSq     DF     MeanSq       F        pValue
                      _________   __   _________    _____    _________
OE                    3.863e+17    1   3.863e+17   679.57   7.3059e-75
MPAA_rating          4.7413e+15    3  1.5804e+15   2.7803     0.041548
OE:MPAA_rating       8.7996e+15    3  2.9332e+15   5.1601    0.0017575
Error                 1.495e+17  263  5.6844e+14
```

This output shows that the p-value for the test is 0.0018, so the null hypothesis is rejected at the 0.05 significance level, and there is sufficient evidence that the slopes are not equal for all ratings (as shown in Figure 3(b)). Comparing to using opening earnings alone, when taking the interaction terms into consideration, the R^2 value increases from 0.743 to 0.764. The following table shows adding interactive terms reduces the classification error by 14% and adds 6 parameters.

Model	Classification Error (%)	RMSE (10^6)	# of Parameters
Model(opening earning + total theaters^2 + opening theaters^2)	65.43	13.86	6
Model(opening earning + total theaters^2 + opening theaters^2 + **opening earning:MPAA rating**)	51.06	13.54	12

Non-mutually exclusive categorical features – Genre: As we mentioned earlier, a movie might belong to one or more genres, and there are 24 genres in total (i.e., Action, Biography, Thriller, War, Adventure, Sci-Fi, Animation, Comedy, Family, Fantasy, Romance, Drama, Crime, Mystery, Musical, Sport, Horror, History, Music, Western, Documentary, Short, Talk-Show, and News). Encoding all 24 genres with 24 dummy variables increases the feature dimension from 1 to 24. As we prefer low dimensional features as it introduces fewer parameters, we seek for ways to reduce the feature dimension.

For the following analysis, we create dummy variables and use 24 1-0 values to encode the 24 genres. A 1-0 value is set to be 1 if the movie is of that genre.

Collinearity: We first check whether representing the categorical data with dummy variables causes collinearity using the variance inflation factor (VIF). The VIF value is large when the variation of a feature is largely explained by a linear combination of other features, and a VIF smaller than 5 is preferred. The following table shows the VIF values for all the genres. In this case, all VIF values are smaller than 5, indicating there is no concern for collinearity. Besides, the Documentary has the highest VIF (2.15), but it does not jump out from the rest of the genres.

Genre	Action	Biography	Thriller	War	Adventure	Sci-Fi	Animation	Comedy	Family	Fantasy
VIF	1.44	1.24	1.54	1.20	1.54	1.20	1.64	1.53	1.89	1.25
Genre	Romance	Drama	Crime	Mystery	Musical	Sport	Horror	History	Music	Western
VIF	1.13	1.93	1.13	1.22	1.10	1.16	1.16	1.16	1.11	1.04
Genre	Documentary	Short	Talk-Show	News						
VIF	2.15	1.08	1.02	1.26						

Feature transformation using Principal Component Analysis (PCA): We then use PCA to drop less descriptive genre variables and reduce the feature dimension. Figure 4 shows the orthonormal principal component coefficients for each genre and the principal component value for each observation. All 24 genres are presented in the bi-plot by a vector, and the direction and the length of the vector indicate how each genre contributes to the first two principal components. Looking at the direction of each genre, we make an interesting observation: genres having the same sign for a PCA component coefficient tend to be used together to describe a movie. For example, movies that are of the genre “romance” are likely to be of the genre “drama.” On the other hand, genres having different signs for a PCA component coefficient are seldom used together. For example, crime/history/biography movies are seldom comedies and sport/action/documentary movies are seldom romance.

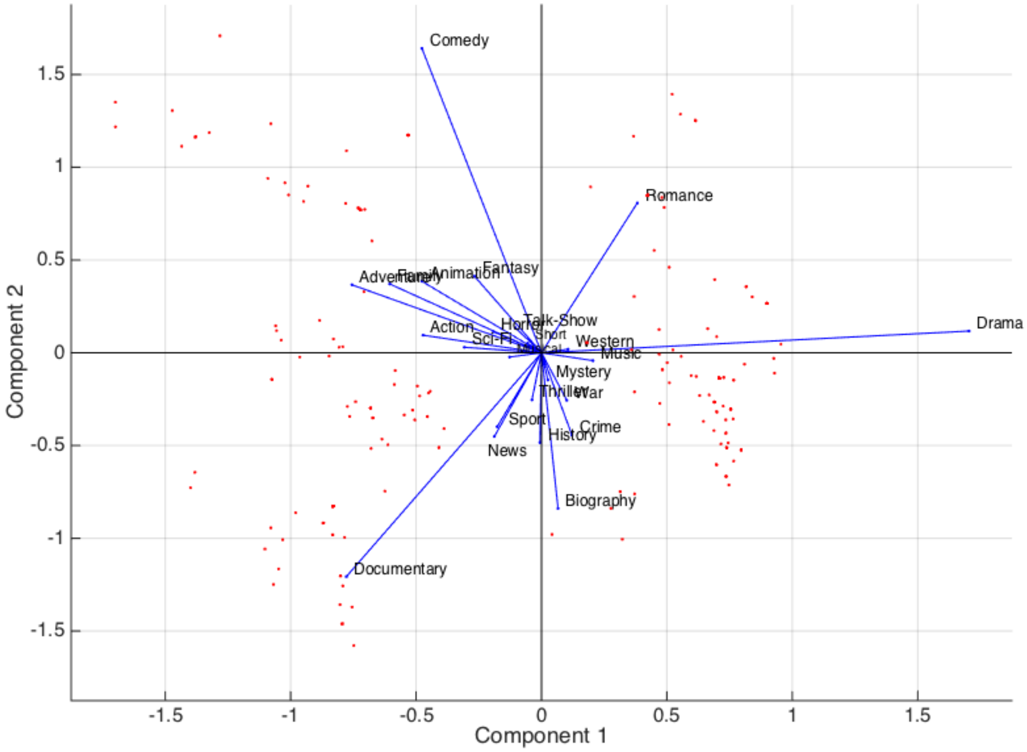

Figure 4. The bi-plot of the first two PCA components and the orthonormal principal component coefficients for each genre.

Running an association rule mining algorithm confirms our observation. Using the Apriori algorithm, which finds the frequent association genre sets, we obtain the following rules:

```
Rule: ('War') ==> ('Drama'), 0.667
Rule: ('Romance') ==> ('Drama'), 0.692
Rule: ('Music') ==> ('Drama'), 0.783
Rule: ('Animation') ==> ('Family'), 0.630
Rule: ('Biography') ==> ('Drama'), 0.627
Rule: ('Comedy', 'Adventure') ==> ('Family'), 0.667
Rule: ('Comedy', 'Family') ==> ('Adventure'), 0.750
Rule: ('Adventure', 'Family') ==> ('Comedy'), 0.706
```

The rules present genres that are often used together to describe a movie. For example, the rule ("War")-> ("Drama") indicates that if a movie is of genre "War", it is likely to also of the genre "Drama". Comparing the rules with the PCA component coefficients, for genres appearing in the same rule, their first PCA component coefficient always have the same sign.

To determine the number of PCA components to include in our model, we make a screen plot of the percent variability explained by each principal component (shown in Figure 5). The screen plot shows the first 10 components that explain 80% of the total variance. The first component explains around 20% of the variance, and the biggest drop happens after the third component. We use the first 5 principal components since they explain 60%of the total variability.

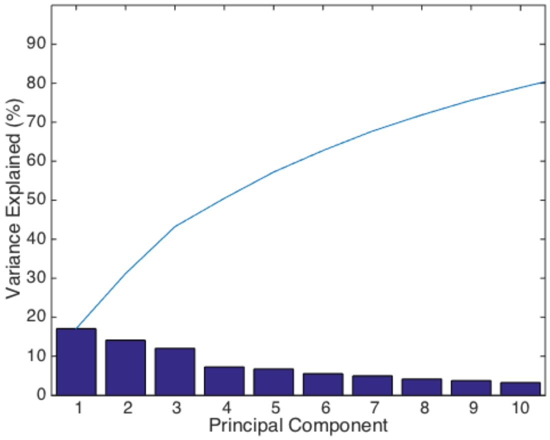

Figure 5. The screen plot of the PCA components showing the variance explained by the number of principal components.

(c) Handling numerical features with high missing percentage

The budget and the averaged director total earnings are numerical feature with high missing percentages. We first understand the predictive power of these two features and then explain how we make our algorithm adaptive to the availability of these two features.

Analyzing the predictive power of budget and averaged director box-office earnings: Using movies with budget or averaged director earnings, Figure 6(a)(b) shows the scatterplots visualizing the relationships between budget and total earnings and the relationships between the averaged director earnings and the total earnings.

The budget (Figure 6(a)) has a better predictive power on the total earnings when the budget is more than 100 million. When the budget is under 100 million, some movies can still be blockbuster films. What is surprising is that the averaged director earnings have low predictive power on the total earnings ($R^2 = 0.23$). Digging into the data, it seems to be true that there's no guarantee that a director will always make blockbuster films. For example, in Figure 5(b), the point which has the biggest (total earning/averaged director earnings) ratio is the American Sniper, and the director is Clint Eastwood. Looking at the movies he directed, they weren't so popular and famous (e.g., J. Edgar, Hereafter, etc.). And the movie has the smallest (total earning/averaged director earnings) ratio is the Chef, directed by Jon Favreau, who is famous for directing the Iron man I, Iron man II, and the Elf.

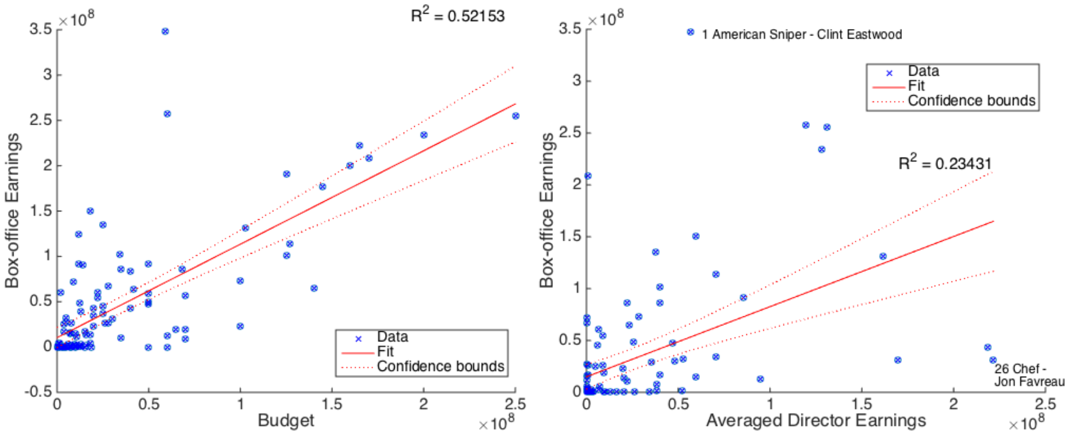

Figure 6(a). Figure 6(b)

Figure 6(a). The relationships between the budget and the total earnings. Figure 6(b). The relationships between the averaged director box-office earnings and the total earnings.

Embracing missing features with the regression tree: We *embrace* missing features as a fact, and design our algorithm to adapt to the availability of features. Our solution is to construct a regression tree *based the set of features that the movie entry has*, and we fit a regression model for each scenario. Figure 6 illustrates the regression tree structure.

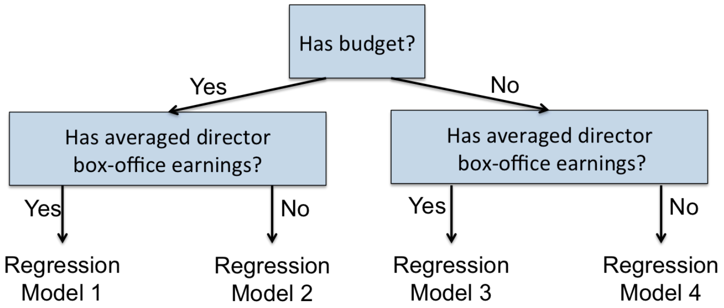

Figure 7. Regression tree structure. It checks whether the movie entry has the budget feature and the averaged director box-office earning feature, and determines the regression model to use based on the availability of those features.

The regression tree has four nodes, and we fit a model at each node only using the features available at that node. The regression tree design improves our testing prediction accuracy from 45.7% to 66% (see Sec. 5). The reason why it provides such a large improvement is because *whether or not the movie has that feature also says something about its total box-office earnings*. Movies with small box-office earnings often do not have the budget and director box-office earnings information because (1) the director's past movies are not so famous and the information is missing. (2) The casts are not well-known actors so the production budget information undetermined.

(3) The movies did not have much advertisement so the advertising budget is unknown. Therefore, using a separate regression model for each scenario does help predict the total box-office earnings.

(d) Feature selection using stepwise regression

Lastly, we use stepwise regression method to eliminate extra features. For each regression model at the regression tree node, we apply stepwise regression method to select the best set of features to include in the model. The stepwise algorithm uses forward and backward stepwise regression to determine a model. At each step, the algorithm searches for features to add or to remove from the model based on a criterion metric (e.g., AIC, R^2, and p-value, etc.). In our experiment, we tried AIC, R^2, and p-value, and obtain the best performance when using the p-value metrics with the threshold set to be 0.005 (at a 0.005 significance level).

5. Results

We evaluate the performance of algorithm in three parts. We first examine the importance of each optimization; we then investigate when our algorithm performs well using the confusion matrix.

(a) The influence of each optimization

The following table shows the influence of each optimization on the testing error, and the number of parameters used in each model. The baseline model fits a linear model on all the variables and the classification error is around 72.34%. The second row shows that introducing quadratic terms reduces the prediction error, and it outperforms the baseline model and uses fewer parameters. The third row shows adding interactive terms further reduces the classification error. Adding the first 5 PCA components increases the classification error but reduces the RMSE. The fifth row shows the regression tree reduces the classification error by 20% but increases the RMSE by 1 million. Our conjecture is that the model starts giving up on accurately predicting the blockbusters and aims to categorize them in the correct earning range. Using the stepwise regression reduces the number of parameters without sacrificing the classification error (but the RMSE error). This indicates that depending on the error metrics one wishes to optimize, he/she might wish to use different models.

Models	Classification Error (%)	RMSE (10^6)	# of Parameters
Linear regression on all variables	72.34	18.39	32
Model (opening earning + total theaters**^2** + opening theaters**^2**)	65.43	13.86	6
Model(opening earning + total theaters^2 + opening theaters^2 + **opening earning:MPAA rating**)	51.06	13.54	12
Model(opening earning + total theaters^2 + opening theaters^2 + opening earing:MPAA rating + **PCA genre**)	54.26	13.52	17
Regression tree based on feature availability	**34.03**	**14.77**	**(17, 18, 18, 19)**
Stepwise Regression on each node regression model (p-value > 0.005)	34.04	27.61	(14, 18, 18, 17)

(b) Confusion matrix

The following four tables show the classification matrix at each of the regression nodes. The results match our expectation: movies do not have both of the budget and the director box-earnings are usually movies with small earnings (all under class 5). And the movies that do not have the budget information but the director's box-office earnings are also movies with small earnings (all under class 6). Estimating the box-office earnings in a small range makes the problem easier, and these are the two cases where our regression model achieves the lowest classification error (26.7% and 20%, respectively). In cases where we have the budget feature, the box-office earnings spread across all 9 classes, and it becomes harder to make precise predictions, thereby the error increases (46.3% and 48.5%).

		Predicted Class								
		Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7	Class 8	Class 9
Actual Class	Class 1	50	5	0	0	0	0	0	0	0
	Class 2	2	13	0	0	0	0	0	0	0
	Class 3	1	2	1	0	0	0	0	0	0
	Class 4	1	0	0	0	0	0	0	0	0
	Class 5	0	0	0	0	0	0	0	0	0
	Class 6	0	0	0	0	0	0	0	0	0
	Class 7	0	0	0	0	0	0	0	0	0
	Class 8	0	0	0	0	0	0	0	0	0
	Class 9	0	0	0	0	0	0	0	0	0

Movies do not have the budget feature and averaged director box-office earnings features. (Total classification error = 26.7%)

		Predicted Class								
		Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7	Class 8	Class 9
Actual Class	Class 1	22	0	0	0	0	0	0	0	0
	Class 2	5	4	0	0	0	0	0	0	0
	Class 3	0	1	0	0	0	0	0	0	0
	Class 4	0	0	0	1	1	0	0	0	0
	Class 5	0	0	0	0	1	0	0	0	0
	Class 6	0	0	0	0	0	0	0	0	0
	Class 7	0	0	0	0	0	0	0	0	0
	Class 8	0	0	0	0	0	0	0	0	0
	Class 9	0	0	0	0	0	0	0	0	0

Movies do not have the budget feature but have the averaged director box-office earnings. (Total classification error = 20%)

		Predicted Class								
		Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7	Class 8	Class 9
Actual Class	Class 1	7	4	0	0	0	0	0	0	0
	Class 2	3	3	1	0	0	0	0	0	0
	Class 3	1	0	0	2	0	0	0	0	0
	Class 4	0	0	1	3	0	0	0	0	0
	Class 5	0	0	0	1	6	0	1	0	0
	Class 6	0	0	0	0	2	1	0	0	0
	Class 7	0	0	0	0	0	0	1	0	0
	Class 8	0	0	0	0	0	0	0	0	0
	Class 9	0	0	0	0	0	0	0	0	4

Movies have the budget feature but do not have the averaged director box-office earnings**.** (Total classification error = 46.3%)

		Predicted Class								
		Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7	Class 8	Class 9
Actual Class	Class 1	1	0	1	0	0	0	0	0	0
	Class 2	1	1	0	2	1	0	0	0	0
	Class 3	0	1	1	0	0	0	0	0	0
	Class 4	0	1	1	1	2	0	0	0	0
	Class 5	0	0	0	0	5	1	1	0	0
	Class 6	0	0	0	0	0	4	0	0	0
	Class 7	0	0	0	0	0	1	4	1	0
	Class 8	0	0	0	0	0	0	3	1	0
	Class 9	0	0	0	0	0	0	0	0	2

Movies have both budget feature and the averaged director box-office earnings. (Total classification error = 48.65%)

Our results verify that using a regression tree based on the availability of high missing percentage features does help improve the classification accuracy by dividing the problem into smaller and easier sub-problems.

6. Related Work

[1] http://en.wikipedia.org/wiki/Cinema_of_the_United_States#Modern_cinema

[2] http://sloanreview.mit.edu/article/what-people-want-and-how-to-predict-it/

[3] Predicting box-office success of motion pictures with neural networks, Expert Systems with Applications, 2006.