Predicting Opening Week Box Office Performance of Hollywood Movies

Analysis Summary, Insights & Methodology

Deepshikha Yadav
Vibha Naryan
Udayan Dasgupta
Santosh P N
Shanawaz Janmohamed

12/24/2010

Table of Contents

Executive Summary

Movie industry is a highly dynamic industry. The uncertainty involved due to the involvement of various factors in determining the box office success makes it even more ambiguous. There are high variations in the strategies followed by successful movies. For example, a movie like Blair Witch Project which is produced with just a budget of $35,000 could earn huge box office revenue while big budget movies at times might fail.

Unlike other products, the shelf-life of a movie is very less. The box office return during the first weekend largely determines the success of the movie. For this reason we have build a model to predict the first weekend box office return of movies based on various factors like release time, budget, presence of Oscar actors etc.

In this model we have analyzed movies of major motion pictures.We find that the following factors are critical in determining whether a movie will break even or not in the opening week and therby the opening week box office performance:

1. **Genre and Distributors**
   - Key genres that drive revenue: Adventure, Drama and Horror
   - Key distributors that drive revenue: Paramount, Warner Brothers, DreamWorks, Miramax, 20th Century Fox
2. **Content :** Sexual content tend to be positively correlated with higher ROI while presence of profanity and violence is negatively correlated with ROI.Movies with MPAA rating as R, generally earn very high ROI.
3. **Release timing :** A summer release is a strong contributor to a film's success in breaking even within the opening week
4. **Screens and Budget**
5. **Presence of Oscar Actor/Director/Producer** has no significant impact on the probability of the movie to break-even in the first weekend.

Inspite of using an ensemble model, where we combined all the 5 models outputs' by assigning optimal weights (Non Linear optimization using Lindo Software), the error was 30%. Certain other variables like Marketing Budget and channels can definitely reduce this error percentage.

Problem Description

The original problem of our project was to predict the opening week box office gross for Hollywood movies. After running both a multi-linear regression as well as a regression tree, we soon re-assessed our problem statement and reframed the question as to predict whether a Hollywood movie would break even in its opening week.

Business Application of this Model :

The prediction model could be used in conjunction with the market research data to develop the overall strategy of movie at the pre-release stage.The total budget to be allocated, the release timing, market positioning, the type of content the movie should have (in terms of sexual content, violence and profanity etc) could be found by combining the two data. The distributors, at the pre-launch stage, could use it to determine the number of screens the movie should be screened into.

Existing Ways of Addressing the Issue

Before launching any Hollywood movie, extensive market research is carried out to gauge public sentiments and to figure out the positioning strategy for the movie. Market research process is a costly and time consuming process. Various insights derived from our model could substitute the market research insights to some extent.

In terms of predicting revenue, very few Studios actually use prediction models. Moreover those predictions are based on custom scripts, with no BI tools involved. Such methods usually have accuracy of less than 40%. WarnerBros however, recently used business intelligence software to predict the sales of DVDs of Harry Potter’s latest series.Moreover movie studios such as MGM and Lions Gate Entertainment use information gleaned from the The Hollywood Stock Exchange market, an online game where players make box office predictions for thousands of upcoming pictures, to help make advertising and promotion decisions.But this is based more on understanding the moods of people rather than on statistical evidence.

Data

Description

The data consisted of approximately 1800 data points collected from the following websites:

- http://www.imdb.com
- http://www.imdb.com/title/ tt0116191
- http://www.imdb.com/title/tt0116191/parentalguide# certification
- http://www.kids-in-mind.com/
- http://www.1728.com/page8.htm
- www.the-numbers.com
- www.leesmovieinfo.com
- www.boxofficemojo.com

The data itself comprises of Title (Categorical Variable), Distributor name (Categorical Variable), genre (Categorical Variable), screens (Numerical Variable), opening date (week/month/year) (Categorical Variable), Box office (opening week sales) (Numerical Variable, Budget (Numeric Variable) , MPAA rating (Categorical Variable), KIM_sex (Numeric Variable) , KIM_violence(Numeric Variable), KIM_profanity (Numeric Variable) (KIM – Kids in Mind). Binary variables indicating whether the movie has an Oscar actor and/or Oscar director have been included in the data set. Another binary variable indicating whether the movie is a sequel or has a sequel has also been included in the data set.

Bollywood data set -We have even tried to collect additional data points for Bollywood movies from sites like http://www.bollywoodtrade.com/box-office/movies-domestic.htm, http://en.wikipedia.org/wiki/Bollywood_films_of_2010. However, the data collection for Bollywood movies from such sites has been difficult due to lack of information on key variables.

Data Pre-Processing

Through a correlation matrix, we were able to identify certain relationships (Appendix A):

1. High Budget movies are generally associated with Screening in more theatres.

2. Opening gross box office collections has a high positive correlation with number of screens the movie is screened on.

Pre-processing of our data followed a four stage process which included coding, creating dummy variables, binning of variables and deleting irrelevant or unintuitive data. For a detailed description of the data pre-processing, please refer to Appendix B.

Data Exploration

We explored the data first using SPOTFIRE (Appendix C). We could observe the following things which a movie producer can use in pre production stage:

1. Presence of Oscar Actor/Director/Producer has no significant impact on the success of the movie or on the probability of the movie to break-even in the first weekend.

2. **Budget:** On an average the movies with genre-black comedy and horror could be produced in low budget.However Action, Adventure, Sci-Fi typically involve huge budget (exceeding $65000000).As is obvious the budget of a movie increases with the presence of Oscar actor/director/producer. Our data further confirms it.

3. **Type of Content:** ROI for a movie is negatively correlated to the presence of violence and profanity. However, ROI has a positive correlation with presence of sexual content.

4. **MPAA Ratings**:The movies with MPAA rating as R, generally earn very high ROI.

5. **Release timing:** It is observed that before 2001, in the month of October and December, distributors usually released movies of the genre-drama. This could be because October and December, being the holiday seasons , would lead to an increase in demand for more family oriented movies as opposed to action or sci-fi genres. But between 2001 and 2005, we observed that movies of drama, comedy and suspense genres were released in October and December.

So just by looking at the data, we could recommend to the producer what genre of movie to produce, What content should the movie have to get the right MPAA ratings,What content he can have based on his budget considerations,What type of actors,directors, producers to avoid and when to release a movie based on the content of the movie.

Model Results

As stated above, the initial objective of our project was to predict the opening week box office gross of Hollywood films. After running both a regression tree and a multi-linear regression, we concluded that our predictions were far too erroneous to present to a producer and thus attempted a series of classification methods.

The classification tree, Naïve Based Classification and logistic regression performed better than the regression models.The improvement on our accuracy using these models (error range from 37% to 43% on the various Models) was significant in comparison to the naïve rule (error of 52.75% predicting using the majority).

Building of a Unique Ensemble model for Prediction

We then built a unique ensemble model combining the predictive power all the models used so far.We combined the outputs of all the models by assigning weights to the outputs of all the models. We did a non linear optimization using the Lindo Software to assign the weights to different models. The output and screenshots of this optimization can be found in Appendix E. A full description of our analysis and resulting screenshots can be found in Appendix D.

Thus we were able to build a very powerful model which combined the strengths of all the models and the predictive accuracy of this model was 70% .

Key Insights & Conclusion

The following are the key insights that we found regarding prediction of whether a Hollywood movie will breakeven in its opening week:

1.Both genre and distributors have an impact on a films ability to breakeven in the opening week

- Key genres that drive revenue: Adventure, Drama and Horror
- Key distributors that drive revenue: Paramount, Warner Brothers, DreamWorks, Miramax, 20th Century Fox

2.Type of content plays an important role in determining the success of a Hollywood movie in its first week of release

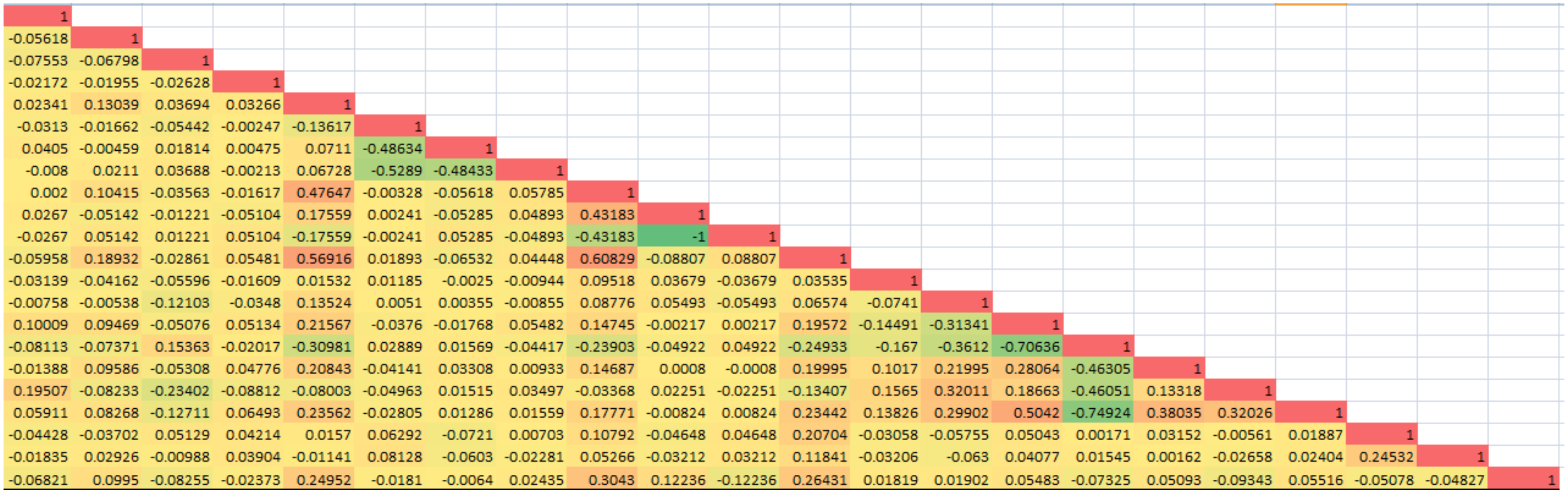
- Sexual content tend to be positively correlated with higher ROI (Appendix C)

3. A summer release is a strong contributor to a film’s success in breaking even within the opening week

4.Screens and Budget are the most significant variables in predicting whether a film will break even in the opening week.

Though the model does not have the type of accuracy of the Hollywood exchange (96%), inclusion of other key variables like Marketing budget and channels and running on model on a bigger database will definitely improve its accuracy.

Appendix A – Screen Clipping of Correlation Matrix of Input Variables



1																					
-0.05618	1																				
-0.07553	-0.06798	1																			
-0.02172	-0.01955	-0.02628	1																		
0.02341	0.13039	0.03694	0.03266	1																	
-0.0313	-0.01662	-0.05442	-0.00247	-0.13617	1																
0.0405	-0.00459	0.01814	0.00475	0.0711	-0.48634	1															
-0.008	0.0211	0.03688	-0.00213	0.06728	-0.5289	-0.48433	1														
0.002	0.10415	-0.03563	-0.01617	0.47647	-0.00328	-0.05618	0.05785	1													
0.0267	-0.05142	-0.01221	-0.05104	0.17559	0.00241	-0.05285	0.04893	0.43183	1												
-0.0267	0.05142	0.01221	0.05104	-0.17559	-0.00241	0.05285	-0.04893	-0.43183	-1	1											
-0.05958	0.18932	-0.02861	0.05481	0.56916	0.01893	-0.06532	0.04448	0.60829	-0.08807	0.08807	1										
-0.03139	-0.04162	-0.05596	-0.01609	0.01532	0.01185	-0.0025	-0.00944	0.09518	0.03679	-0.03679	0.03535	1									
-0.00758	-0.00538	-0.12103	-0.0348	0.13524	0.0051	0.00355	-0.00855	0.08776	0.05493	-0.05493	0.06574	-0.0741	1								
0.10009	0.09469	-0.05076	0.05134	0.21567	-0.0376	-0.01768	0.05482	0.14745	-0.00217	0.00217	0.19572	-0.14491	-0.31341	1							
-0.08113	-0.07371	0.15363	-0.02017	-0.30981	0.02889	0.01569	-0.04417	-0.23903	-0.04922	0.04922	-0.24933	-0.167	-0.3612	-0.70636	1						
-0.01388	0.09586	-0.05308	0.04776	0.20843	-0.04141	0.03308	0.00933	0.14687	0.0008	-0.0008	0.19995	0.1017	0.21995	0.28064	-0.46305	1					
0.19507	-0.08233	-0.23402	-0.08812	-0.08003	-0.04963	0.01515	0.03497	-0.03368	0.02251	-0.02251	-0.13407	0.1565	0.32011	0.18663	-0.46051	0.13318	1				
0.05911	0.08268	-0.12711	0.06493	0.23562	-0.02805	0.01286	0.01559	0.17771	-0.00824	0.00824	0.23442	0.13826	0.29902	0.5042	-0.74924	0.38035	0.32026	1			
-0.04428	-0.03702	0.05129	0.04214	0.0157	0.06292	-0.0721	0.00703	0.10792	-0.04648	0.04648	0.20704	-0.03058	-0.05755	0.05043	0.00171	0.03152	-0.00561	0.01887	1		
-0.01835	0.02926	-0.00988	0.03904	-0.01141	0.08128	-0.0603	-0.02281	0.05266	-0.03212	0.03212	0.11841	-0.03206	-0.063	0.04077	0.01545	0.00162	-0.02658	0.02404	0.24532	1	
-0.06821	0.0995	-0.08255	-0.02373	0.24952	-0.0181	-0.0064	0.02435	0.3043	0.12236	-0.12236	0.26431	0.01819	0.01902	0.05483	-0.07325	0.05093	-0.09343	0.05516	-0.05078	-0.04827	1

Appendix B – Data Pre-Processing Procedure

1. *Coding*

In order to prepare the data for processing using categorical techniques, we coded the opening week to include three distinct categories; Holiday (November, December, January, February), Off Season (March, April, May, June) and Summer (July, August, September, October). Second, we created a new variable, "Break-Even". This variable allowed us to determine whether the film was able to achieve an opening week gross higher than its budget (1) or an opening week gross lower than its budget (0).

2. *Dummy Variables*

To simplify the modeling process, we created dummy variables for five of our factors; Distributor coded, Genre, Opening Week (Coded), MPAA Rating, and Break Even. Doing so allowed us to use the data for processing with both continuous and categorical response variables.

3. *Binning*

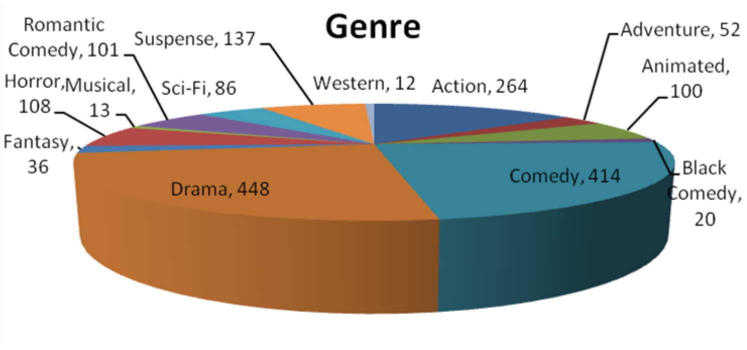
Two variables were binned as separate data for processing under Naïve Bayes. The Screens variable was binned into 5 equal interval bins and Budget was binned into 15 equal interval bins. The output can be found in Appendix XX.

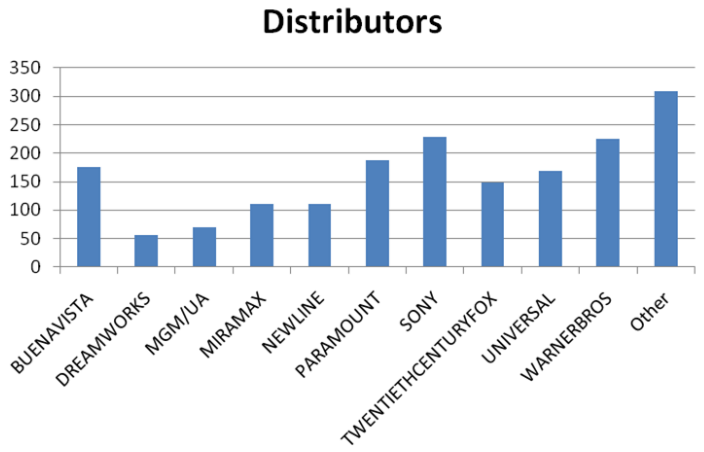
4. *Deletion of Irrelevant Variables*

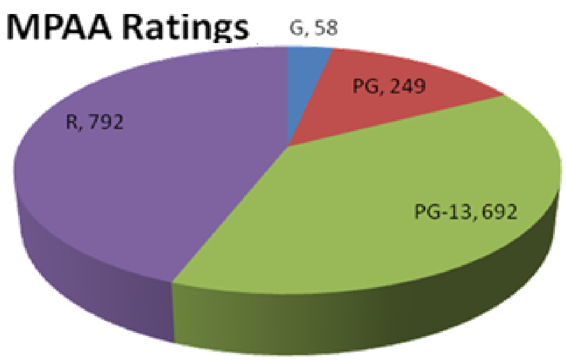
Our data included many other variables which were either deleted or ignored due to their irrelevance in predicting opening week box office gross. These variables include title, whether the film is a US production or not and foreign box office gross. The three variables kids in mind (sex), kids in mind (profanity) and kids in mind (violence) were ignored in all models because these three ratings are inputs into determining the film's MPAA rating.

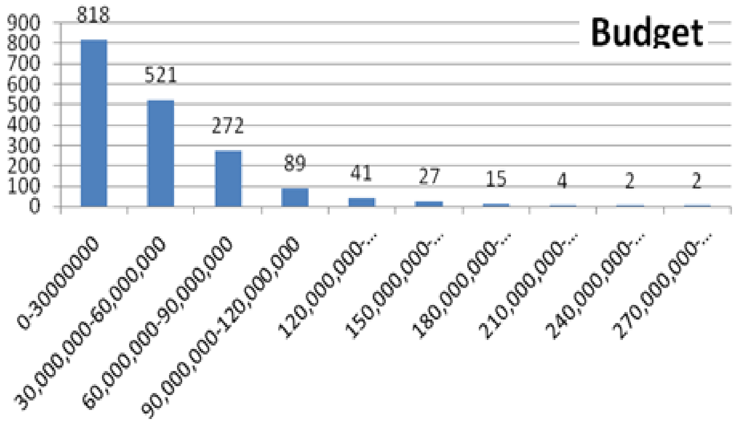
5. We created a dummy variable called Breakeven. If the opening week gross box office collections was greater than budget it was coded as 1.

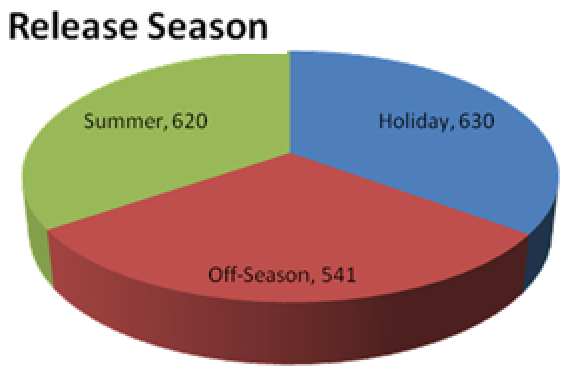
Apendix C – Snapshot of Data

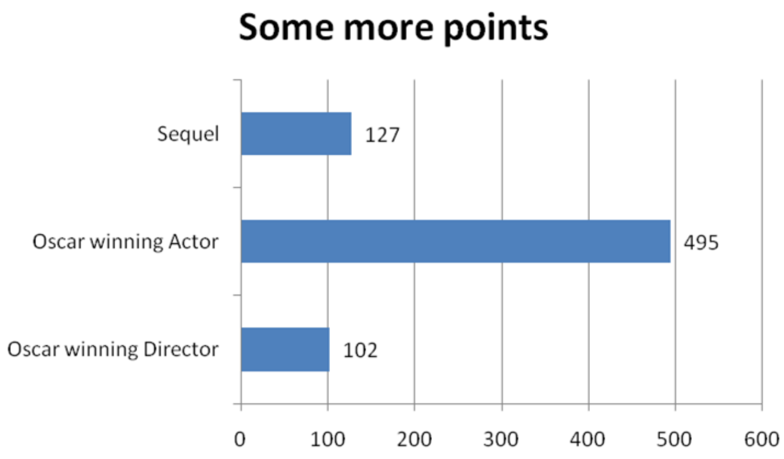




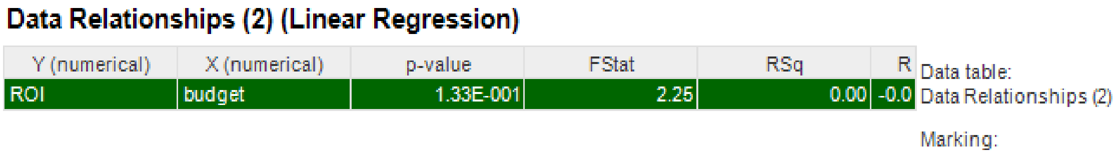








ROI vs Budget Scatter Plot

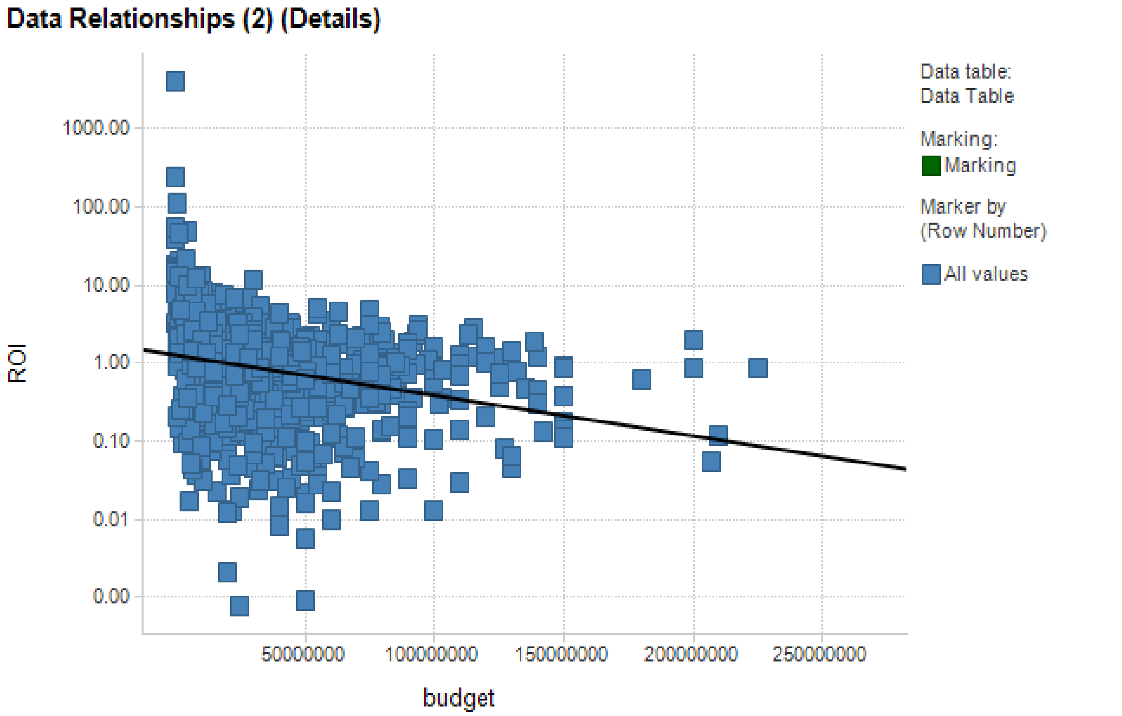


Data Relationships (2) (Linear Regression)

Y (numerical)	X (numerical)	p-value	FStat	RSq	R
ROI	budget	1.33E-001	2.25	0.00	-0.0

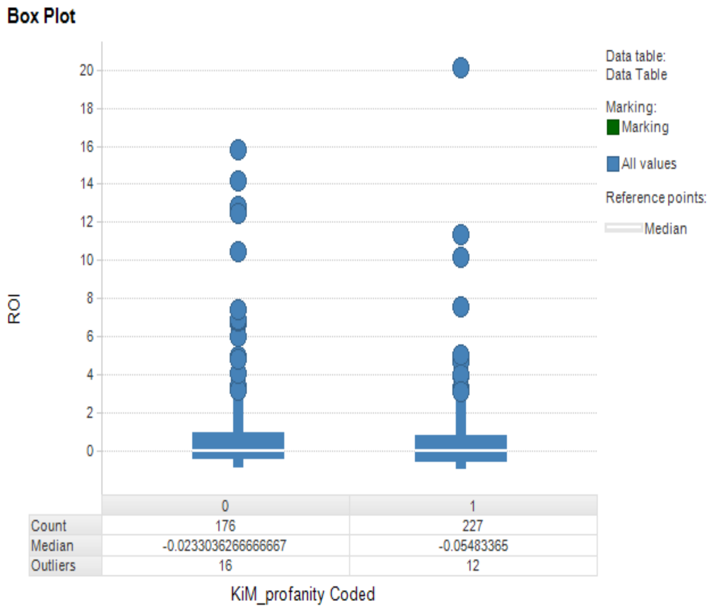
Data table:
Data Relationships (2)

Marking:

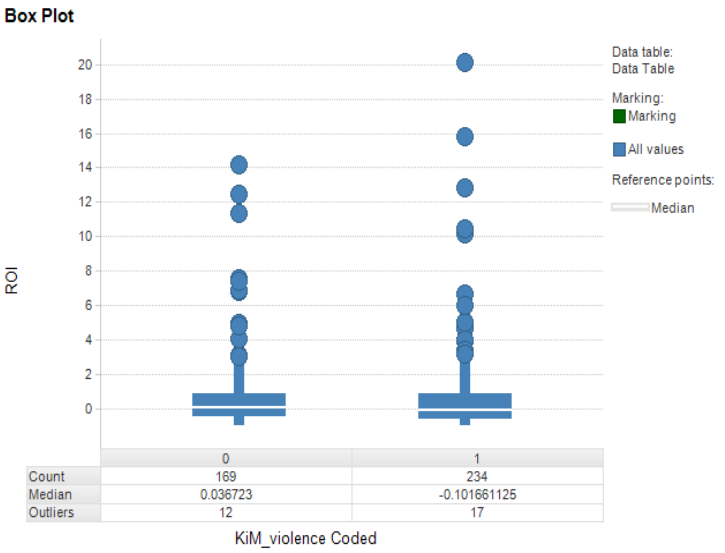


Data Relationships (2) (Details)

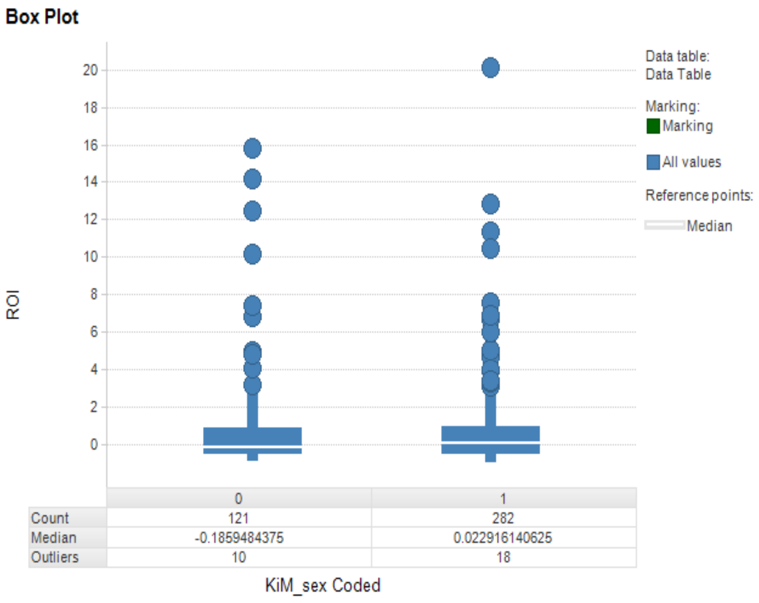
Box Plot Of ROI Vs KIM Ratings



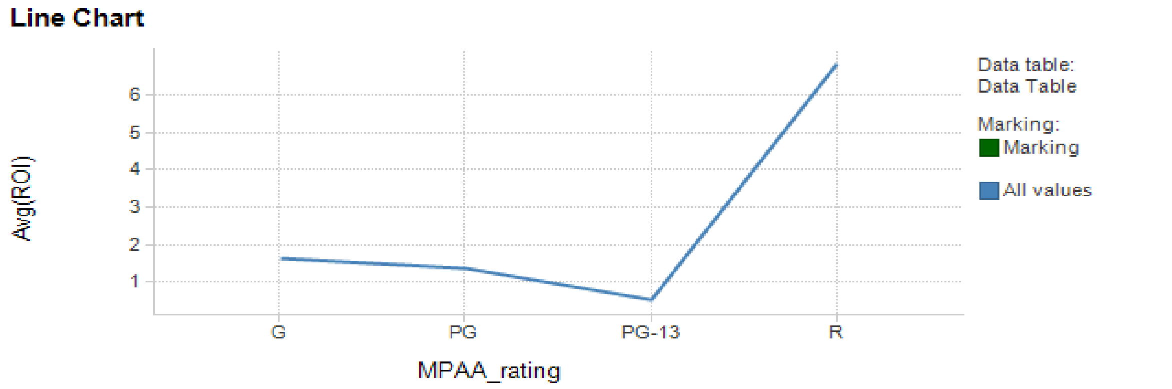
	0	1
Count	176	227
Median	-0.0233036266666667	-0.05483365
Outliers	16	12

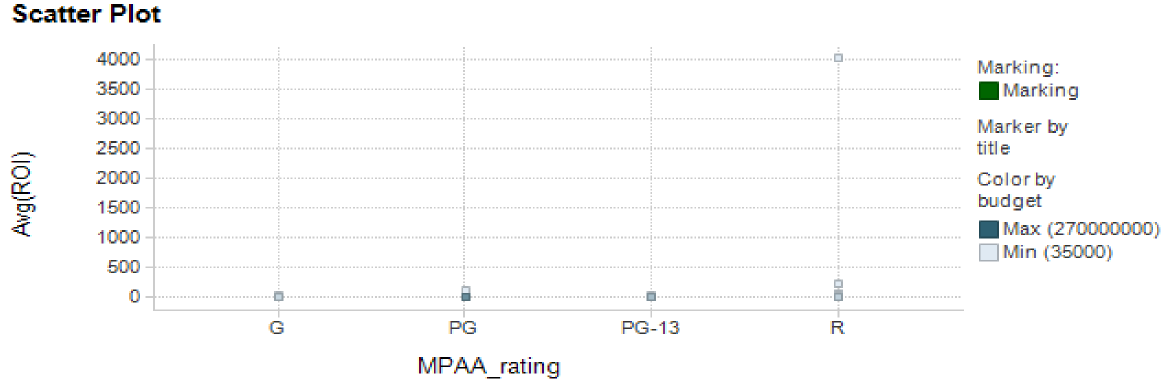


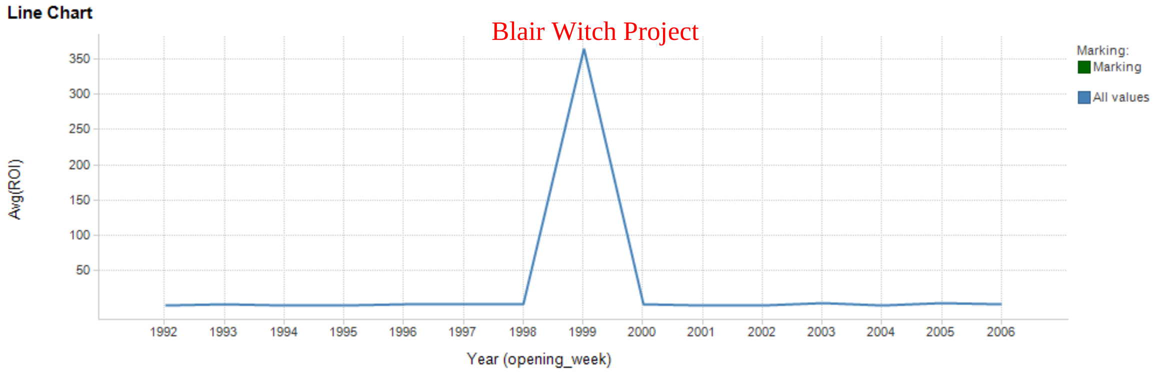
	0	1
Count	169	234
Median	0.036723	-0.101661125
Outliers	12	17

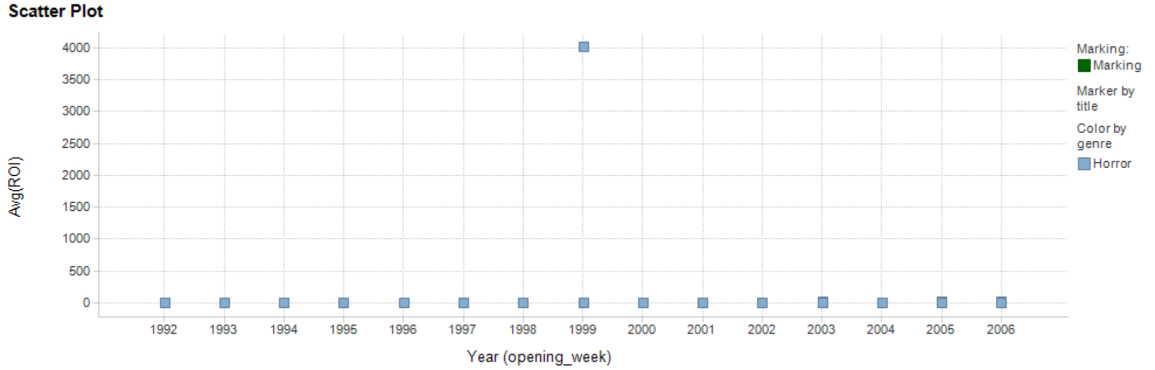


	0	1
Count	121	282
Median	-0.1859484375	0.022916140625
Outliers	10	18






Blair Witch Project



Appendix D – Description of Methodology & Model Outputs

Data Exploration

- **Presence of Oscar Actor/Director/Producer** has no significant impact on the success of the movie or on the probability of the movie to break-even in the first weekend.

  Recommendation – It is not necessary to include an Oscar actor/director/producer while coming with a new movie.

- **Budget:** On an average the movies with genre-black comedy and horror could be produced in low budget however Action, Adventure, Sci-Fi typically involve huge budget (exceeding $65000000).As is obvious the budget of a movie increases with the presence of Oscar actor/director/producer. Our data further confirms it.

  Recommendation – Budget plays a role in choice of the movie genre. Availability of high budget can be utilized for producing action, adventure and Sci-Fi movies.

- **Type of Content:** ROI for a movie is negatively correlated to the presence of violence and profanity. However, the data shows that ROI has a positive trend in relation to presence of sexual content.

  Recommendation- Content heavy on violence and profanity should be avoided as it is negatively correlated to ROI.

- **Release timing:** It is observed that, in the month of October and December, distributors usually release movies of the genre-drama. This could be because October and December, being the holiday seasons would lead to an increase in demand of more family oriented movies as opposed to action or sci-fi genres. Moreover, from year 2001-2005, we observed that over the years the distributors produce combination of genre - drama, comedy, suspense. This means that in October customers usually prefer to view movies in these genres. The observation could be further verified by marketing research.

Recommendation – If producing a movie in holiday season, especially in the month of October and December, it is recommended to produce a movie of the genre drama and family movies.

- **Anomalies/outliers:**

  - Some outliers were found during the data analysis. In the genre 'Horror', the movie 'THE BLAIR WITCH PROJECT" stands out as an outlier with an extremely high ROI as compared to other movies of the same genre. This behavior was due to the extremely low budget of the movie (35000$) as compared to other movies. The marketing for the movie was done using online medium and social media marketing which saved a lot of cost.

  - In the Suspense Genre, movies like 'OPEN WATER' and "SAW" stand out as outliers as compared to other movies of the same genre. Open Water had a low budget with decent earnings which led to an extremely high ROI .On the other hand, SAW was a medium budget movie but earned huge revenues and high ROI in turn.

Recommendation - Budget along with strategy adopted, online/offline marketing techniques play an important role in determining the ROI. Adopting online, social media marketing techniques lowers the cost of the movie.

Prediction Process

First we performed the regression tree on the data. The regression tree gave us the critical parameters which could be used for subsequent multi Linear regression. We found that Budget and Number of Screens were the two key variables which determined the Opening box office gross of the movies.

The RMS error of the Validation data was 53498702.35$ and the average error was 1362057 $. The RMS error of the Test data was 42992718$ and the average error was -710374$. The rms error was very large compared to the average budget of the movie. We were thus very skeptical of the resulting predictions from this model.

Next we performed the Multi Linear Regression on the given Data. We again found that the following variables were very important:

a) Genre Horror
b) Screens
c) Open_ Weekend Coded Summer
d) Budget
e) Kim_Sex
f) Kim_Violence
g) Kim_Profanity

Again, the errors were very large in comparison to the average value of the data. So we concluded that we would not be able to predict the opening box office gross of the movies using these models due to the poor accuracy and predictive performance.

The opening gross office performance is an indicator of the overall gross collections of a movie. This has been empirically proved by Simonoff model:

Log domestic = -0.164+1.09 Log 1st weekend

We have seen that when a movie breaks even in the first week it becomes a blockbuster. For example: Passion of the Christ, The Dark Knight, Harry Potter 6.etc. So we decided to build a model to predict whether the opening box office performance would be greater than the budget of a movie. This will help him to plan further marketing activities specifically with respect to promotion and distribution.

We therefore recoded the predicted outputs from the regression model and the regression tree; The records (in validation and tested data) whose predicted opening box office collections were greater than the budget of the movie were coded as 1 whereas the others were coded as zero.

Misclassification error % can be seen below:

a) Validation data in Regression tree : 43.2%
b) Test Data in Regression Tree : 42.9%
c) Validation Data in Multi Linear Regression model: 53.4%

d) Test Data in Multi Linear Regression Model: 53.91%

The next step; therefore was to develop models where the response variable was a categorical value. We ran a classification tree on the data set and found that the following variables were key in predicting whether the movie would break even or not in the first week.

a) Budget
b) Screens
c) Genre Action
d) Distributor Sony and Distributor Universal.
e) MPAA rating PG-13
f) Sequel

To avoid over fitting, we pruned the tree using validation data and tested the model using test data. The misclassification error in the validation data was 37.43% and in test data, 37.68%. There was thus a significant improvement in the prediction accuracy using this model in comparison to our earlier models.

We wanted to further improve the predictive accuracy and therefore ran a Naïve Bayes Classification on the data. In this simulation, the predictor variables must be categorical. We therefore binned all the continuous variables as outlined in the data pre procession section. The misclassification error in the validation data using Naïve Bayes was 43.02% and 42.32% in the test data. The predictive power of this model was much less than the predictive power of classification tree model.

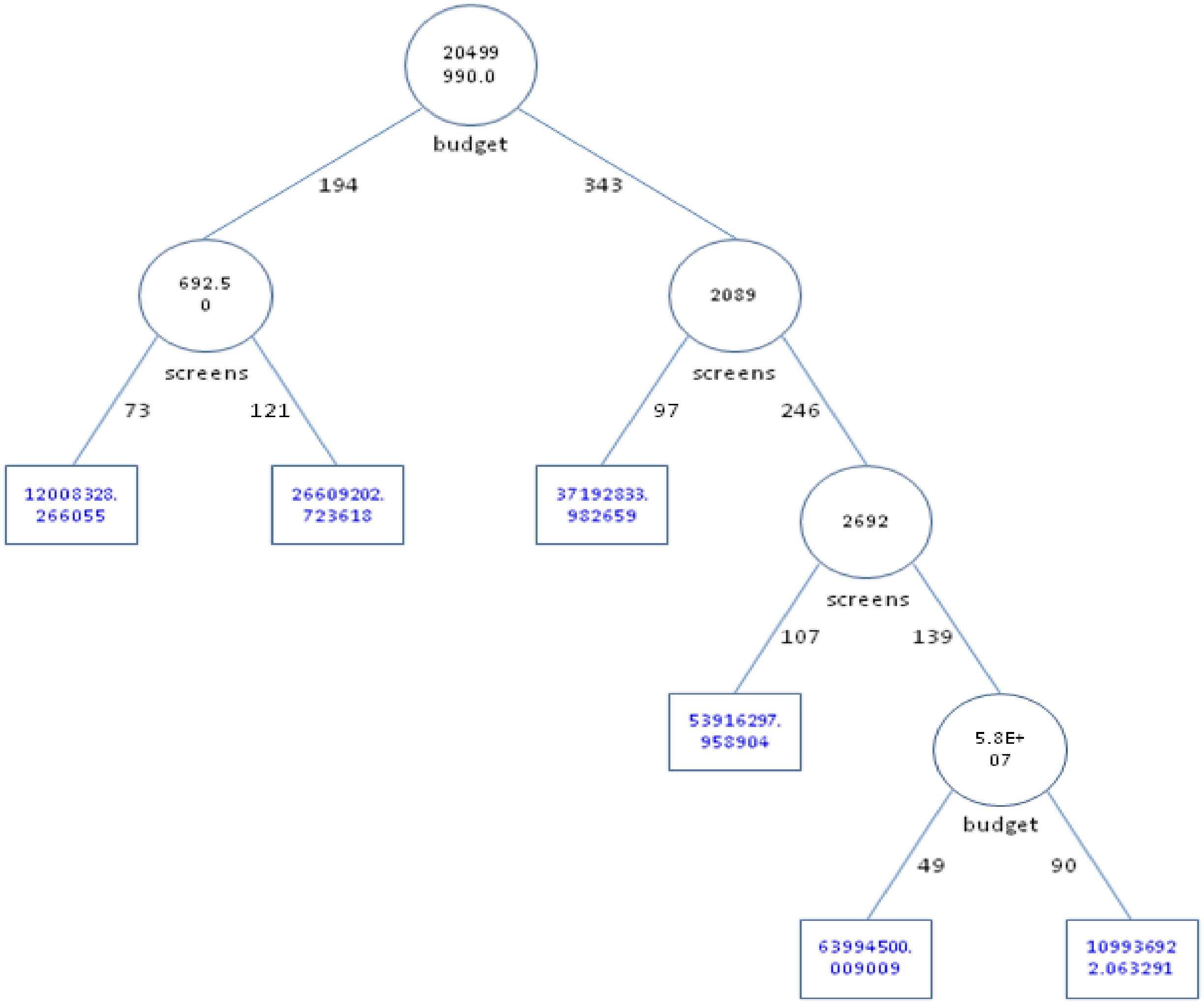
We finally ran a logistic regression (Using Exhaustive and Sequential search Methods) on the data. The key variables that resulted were:

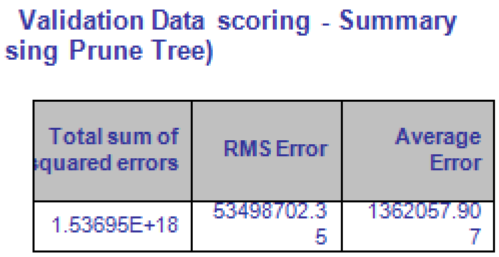
a) Budget
b) Screens
c) Sequel
d) Distributors: Paramount, Warner Brothers, DreamWorks, Miramax, 20th Century Fox
e) Genres: Adventure, Drama and Horror.
f) Summer release

The misclassification error in the validation data using the logistic regression was 40.41% and in 36.59% in the test data. So we found that the predictive power of this model was fairly similar to that of the classification tree model.

To ensure we had built the most accurate model given our data, we decided to weight each of our models and combine them in order to optimize our prediction accuracy. The weights were calculated to minimize the error on the validation test results of all the models using "Lindo" Software. The naïve rule (classifying films as break-even or not based on the majority rule) results in an error percentage of 52.75%, and our optimized model was able to reduce this error to 29.0%

Appendix D1 – Regression Tree Output

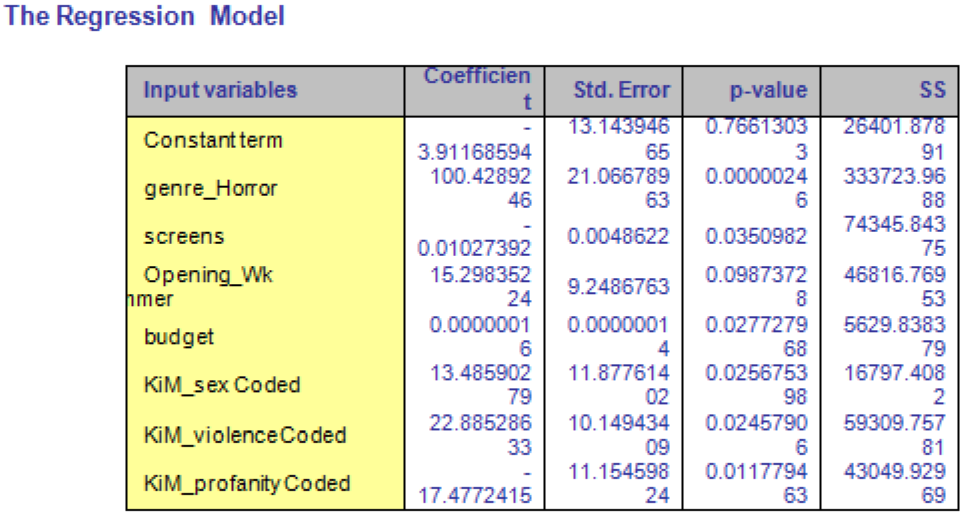




Validation Data scoring - Summary
sing Prune Tree)

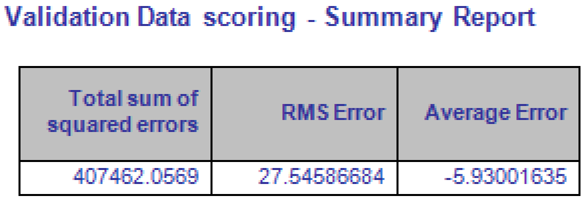
Total sum of quared errors	RMS Error	Average Error
1.53695E+18	53498702.3 5	1362057.90 7

Appendix D2 – Multi-Linear Regression Output



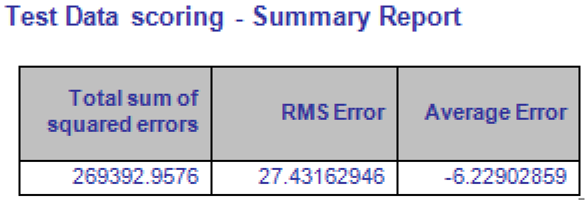
The Regression Model

Input variables	Coefficient	Std. Error	p-value	SS
Constant term	-3.91168594	13.14394665	0.76613033	26401.87891
genre_Horror	100.4289246	21.06678963	0.00000246	333723.9688
screens	-0.01027392	0.0048622	0.0350982	74345.84375
Opening_Wk mer	15.29835224	9.2486763	0.09873728	46816.76953
budget	0.00000016	0.00000014	0.027727968	5629.838379
KiM_sex Coded	13.48590279	11.87761402	0.025675398	16797.4082
KiM_violence Coded	22.88528633	10.14943409	0.02457906	59309.75781
KiM_profanity Coded	-17.4772415	11.15459824	0.011779463	43049.92969



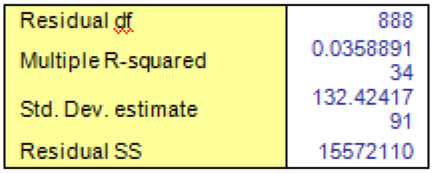
Validation Data scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
407462.0569	27.54586684	-5.93001635



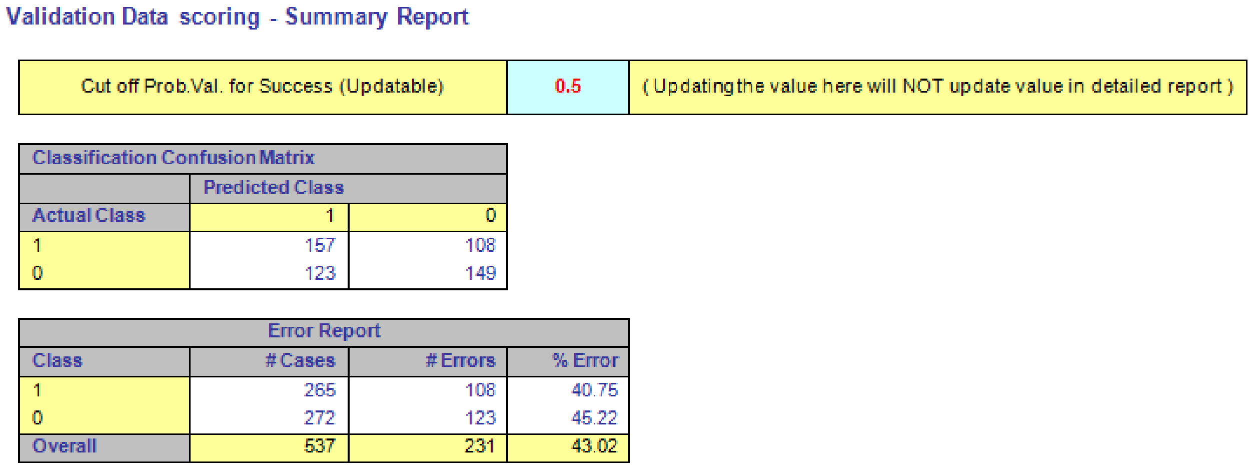
Test Data scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
269392.9576	27.43162946	-6.22902859



Residual df	888
Multiple R-squared	0.035889134
Std. Dev. estimate	132.4241791
Residual SS	15572110

Appendix D3 – Naïve Bayes Output

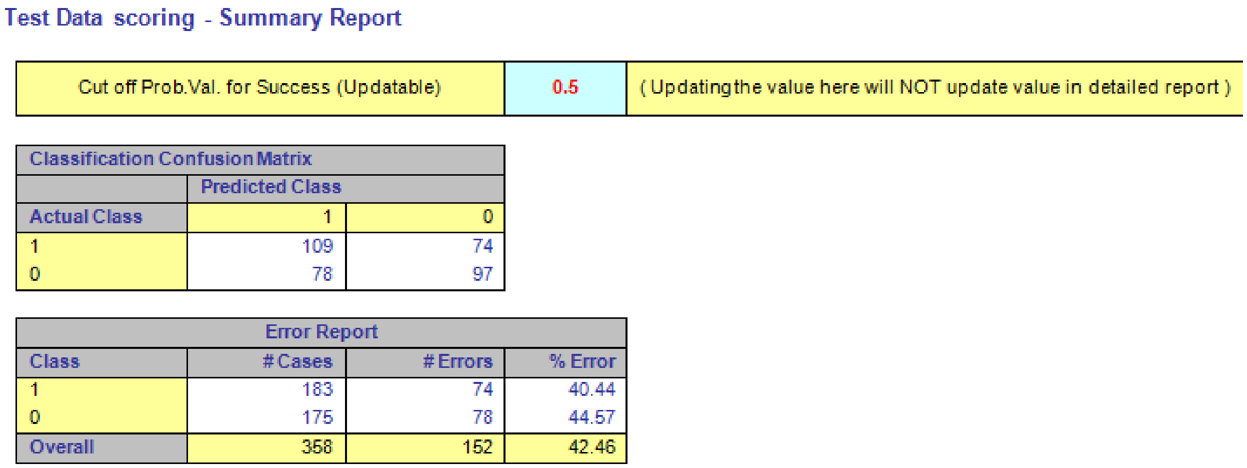


Validation Data scoring - Summary Report

Cut off Prob.Val. for Success (Updatable)	0.5	(Updating the value here will NOT update value in detailed report)

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	157	108
0	123	149

Error Report			
Class	# Cases	# Errors	% Error
1	265	108	40.75
0	272	123	45.22
Overall	537	231	43.02



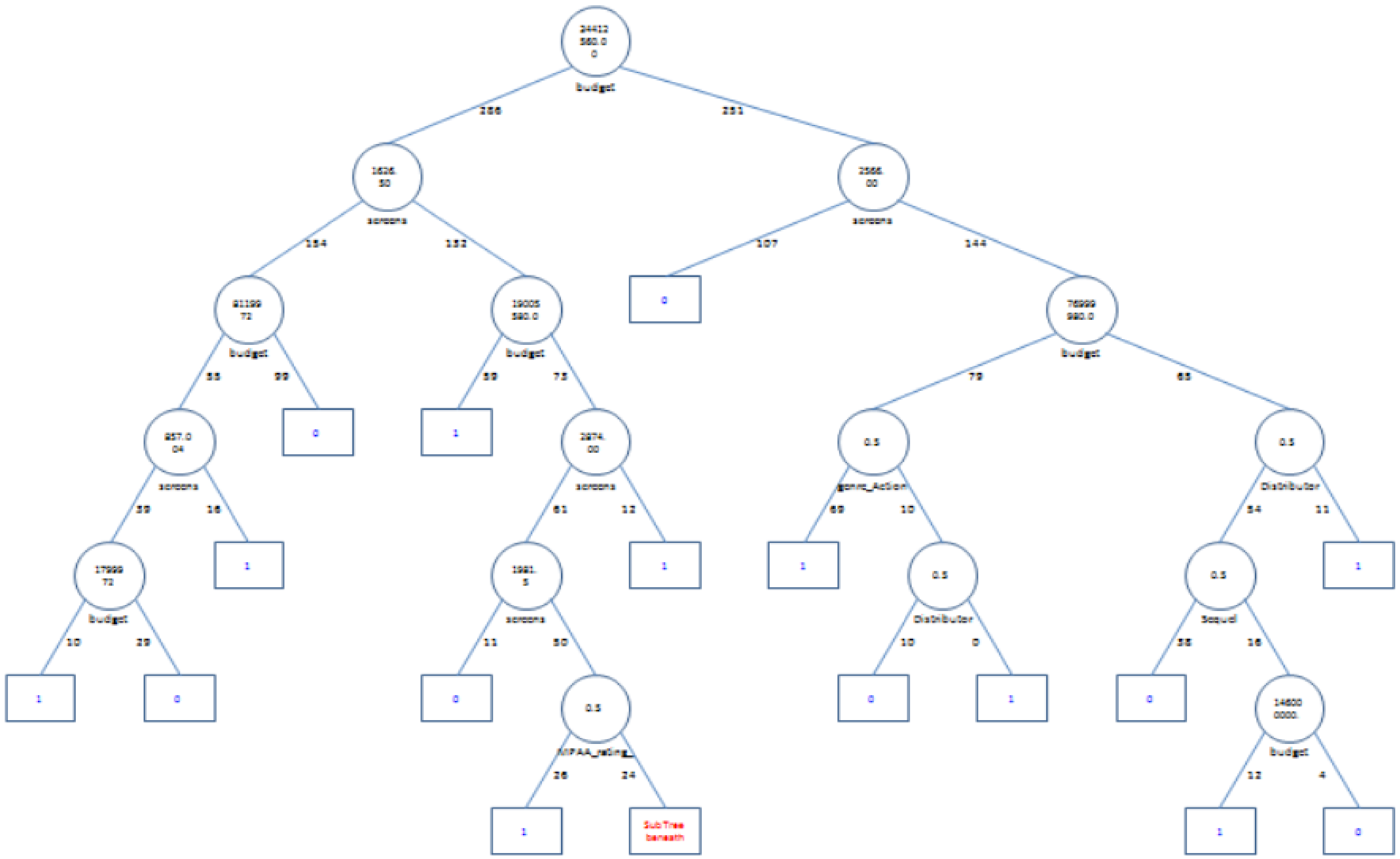
Test Data scoring - Summary Report

Cut off Prob.Val. for Success (Updatable)	0.5	(Updating the value here will NOT update value in detailed report)

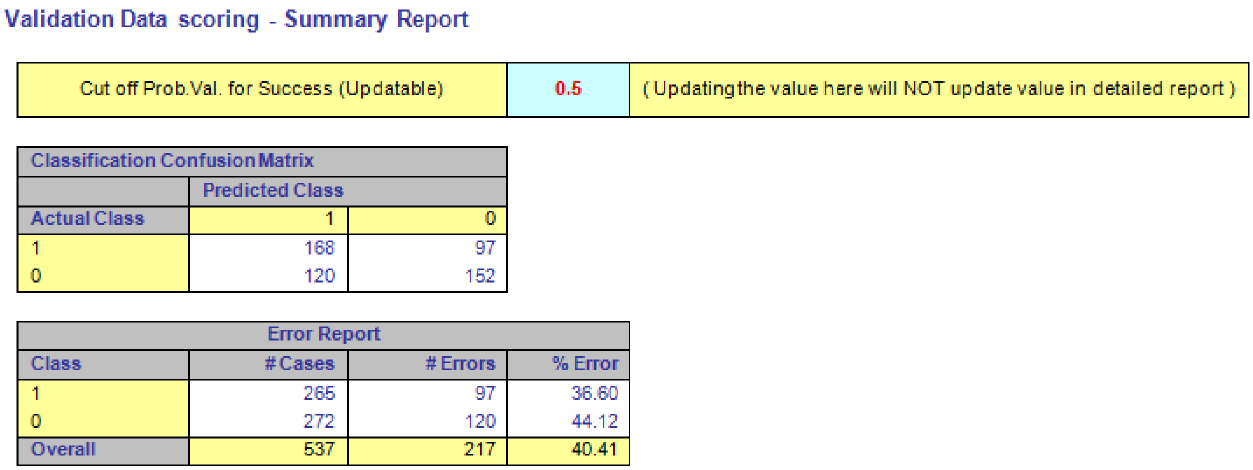
Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	109	74
0	78	97

Error Report			
Class	# Cases	# Errors	% Error
1	183	74	40.44
0	175	78	44.57
Overall	358	152	42.46

Appendix D4 – Classification Tree Output



Appendix D5 – Logistic Regression Output

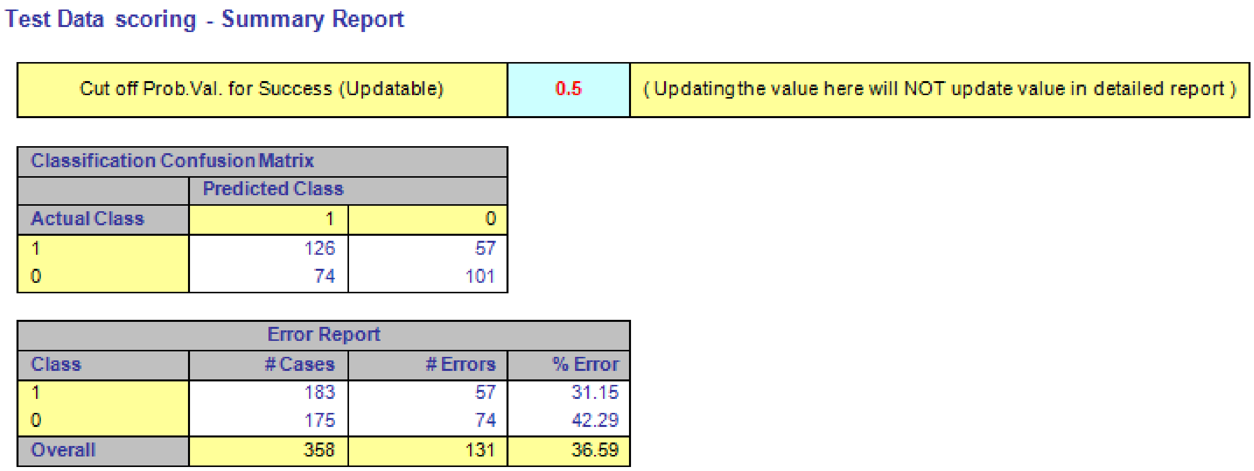


Validation Data scoring - Summary Report

Cut off Prob.Val. for Success (Updatable)	0.5	(Updating the value here will NOT update value in detailed report)

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	168	97
0	120	152

Error Report			
Class	# Cases	# Errors	% Error
1	265	97	36.60
0	272	120	44.12
Overall	537	217	40.41



Test Data scoring - Summary Report

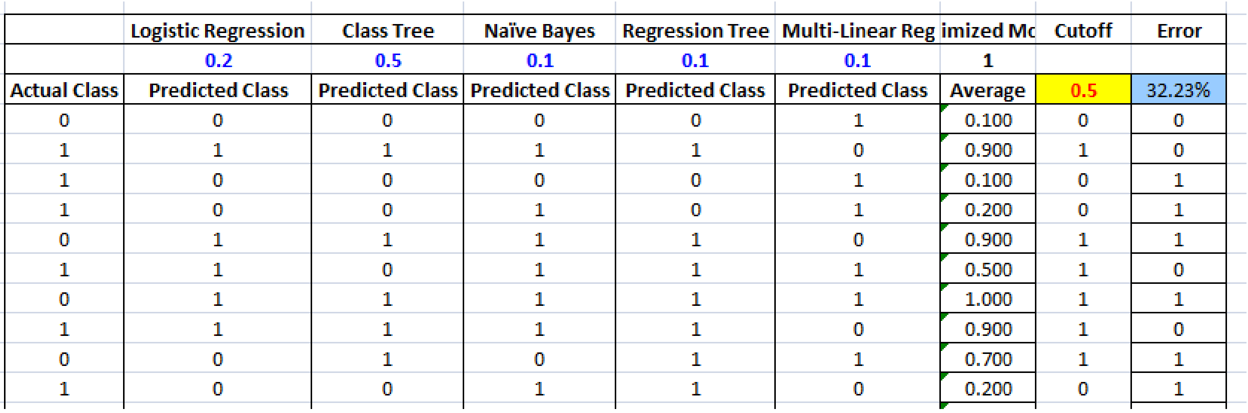
Cut off Prob.Val. for Success (Updatable)	0.5	(Updating the value here will NOT update value in detailed report)

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	126	57
0	74	101

Error Report			
Class	# Cases	# Errors	% Error
1	183	57	31.15
0	175	74	42.29
Overall	358	131	36.59

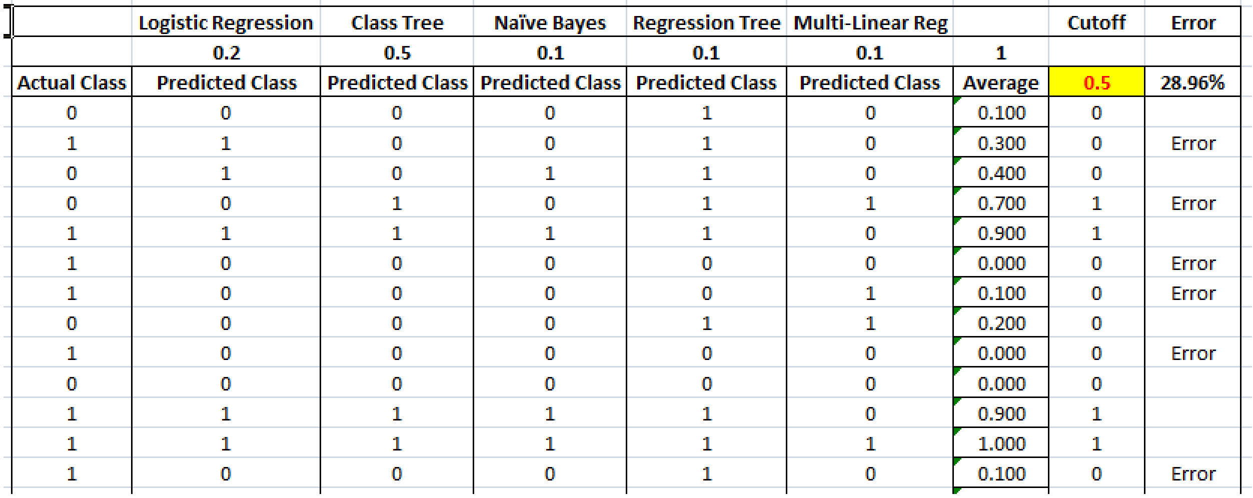
Appendix D6 – Ensemble: Optimized Model Output

Validation Score



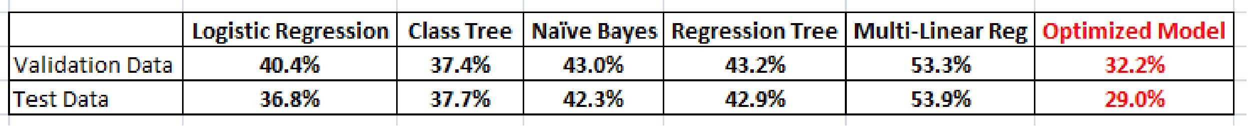
	Logistic Regression	Class Tree	Naïve Bayes	Regression Tree	Multi-Linear Reg	imized Mc	Cutoff	Error
	0.2	0.5	0.1	0.1	0.1	1		
Actual Class	Predicted Class	Predicted Class	Predicted Class	Predicted Class	Predicted Class	Average	0.5	32.23%
0	0	0	0	0	1	0.100	0	0
1	1	1	1	1	0	0.900	1	0
1	0	0	0	0	1	0.100	0	1
1	0	0	1	0	1	0.200	0	1
0	1	1	1	1	0	0.900	1	1
1	1	0	1	1	1	0.500	1	0
0	1	1	1	1	1	1.000	1	1
1	1	1	1	1	0	0.900	1	0
0	0	1	0	1	1	0.700	1	1
1	0	0	1	1	0	0.200	0	1

Test Score



	Logistic Regression	Class Tree	Naïve Bayes	Regression Tree	Multi-Linear Reg		Cutoff	Error
	0.2	0.5	0.1	0.1	0.1	1		
Actual Class	Predicted Class	Predicted Class	Predicted Class	Predicted Class	Predicted Class	Average	0.5	28.96%
0	0	0	0	1	0	0.100	0	
1	1	0	0	1	0	0.300	0	Error
0	1	0	1	1	0	0.400	0	
0	0	1	0	1	1	0.700	1	Error
1	1	1	1	1	0	0.900	1	
1	0	0	0	0	0	0.000	0	Error
1	0	0	0	0	1	0.100	0	Error
0	0	0	0	1	1	0.200	0	
1	0	0	0	0	0	0.000	0	Error
0	0	0	0	0	0	0.000	0	
1	1	1	1	1	0	0.900	1	
1	1	1	1	1	1	1.000	1	
1	0	0	0	1	0	0.100	0	Error

Optimized Model Summary



	Logistic Regression	Class Tree	Naïve Bayes	Regression Tree	Multi-Linear Reg	Optimized Model
Validation Data	40.4%	37.4%	43.0%	43.2%	53.3%	32.2%
Test Data	36.8%	37.7%	42.3%	42.9%	53.9%	29.0%