Presentation to:

Data Science

Predicting Box Office Results and Movie Sentiment Using Supervised Learning Methods

March 2, 2016

Project Objective and Questions

Objective(s)

- Predict Box Office results based on select movie attributes, ratings, and reviews

- Predict movie sentiment based on Box Office results and select movie attributes

- Predict movie sentiment based on plot summary

Questions

- How do certain movie attributes determine Box Office success?

- How do movie reviews and ratings determine Box Office success?

- Which classifier is most accurate in determining movie sentiment?

- Which features are most important in determining movie sentiment?

- Is movie plot predictive of movie sentiment?

Description of Data and Pre-Processing Steps

Dataset	Select Var.	Definition
Box Office Mojo	Title	▪ Title of the movie
	Total Gross	▪ Total domestic gross earnings
	Total Theatres	▪ No. of locations playing movie
	Opening Date	▪ Opening date of movie
IMDb	Title	▪ Title of the movie
	Year	▪ Year of movie release
	Rating	▪ Film rating (e.g. PG, PG-13, R)
	Runtime	▪ Length of the movie
	Genre	▪ Genre(s) of the movie
	Director	▪ Director(s) of the movie
	Cast	▪ Notable cast members in the movie
	Metacritic	▪ "Metascore" from critics (1 – 100)
	IMDb Rating	▪ Wtd. avg. of IMDb user votes (1 – 10)
	IMDb Votes	▪ No. of IMDb registered user votes
	Plot	▪ Plot summary
Rotten Tomatoes	Sentiment	▪ Fresh or rotten based on T Rating
	T Rating	▪ % of positive critic reviews (1 – 100)
	User Meter	▪ % positive user ratings (1 – 100)
	User Rating	▪ Avg. user rating (1 – 5)
	User Reviews	▪ No. of user reviews

Pre-Processing Steps

1 ▪ Web-scraped Box Office results from Box Office Mojo using "out-of-the-box" tool Import.io

2 ▪ Cleaned the data:
  – Removed $, *, spaces, and forward dates
  – Converted certain string columns to numeric and date format data types
  – Replaced "&" with "and" in movie titles
  – Decoded strings from Latin to UTF equivalents
  – Lowercased all strings
  – Imputed and removed certain null values

3 ▪ Filtered movie results from 2000 – 2015, given Box Office tracking and relevance of Rotten Tomatoes

4 ▪ Created Title_Year column for Box Office Mojo and IMDb data

5 ▪ Joined tables to create two new datasets:
  – IMDb data and Rotten Tomatoes joined on ID
  – Box Office Mojo and IMDb data joined on Title_Year, joined on ID with Rotten Tomatoes

6 ▪ Mapping numbers to categorical variables Rating and Sentiment (i.e. Fresh or Rotten)
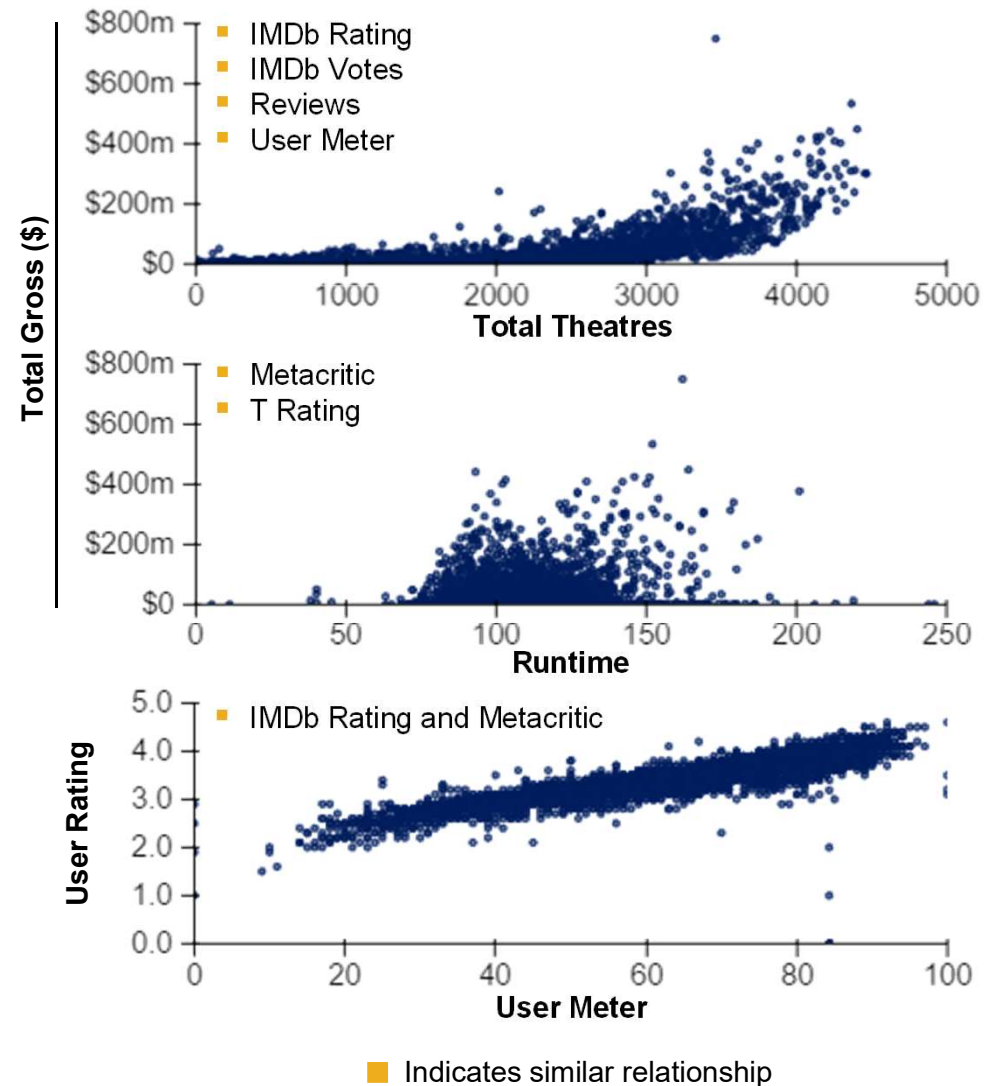
Predicting Box Office Results
Regression Variables Exhibit Similar Relationships

Description of Analysis

- Used visualizations to identify key relationships between Total Gross and independent variables

- Created three datasets based on five-year time periods: 2000 – 2004, 2005 – 2009, and 2010 – 2014

- Reserved 2015 movie data for out-of-sample testing

- Performed Linear Regression and reviewed Adj. R-Squared, P-values, Cond. No., and Prediction results

- Selected 2010 – 2014 data to train the model, based on Adj. R-Squared and prediction performance

- Removed high P-value variables and those reflecting collinear relationships

- Performed Linear Regression on out-of-sample data and captured prediction and prediction error results

- Compared test and out-of-sample results to demonstrate model generalization

Key Variable Relationships

Predicting Box Office Results
Recursive Selection of Significant Features

Regression Results by Time Period

2010 - 2014

Variables	Coefficient	P-Value	Adj. R-Squared	Condition No.
Total Theatres	32,280,000	0.000		
Runtime	1,451,000	0.194		
Metacritic	3,569,000	0.206		
IMDb Rating	(7,199,000)	0.000		
IMDb Votes	27,280,000	0.000		
T Rating	20,460	0.996		
Meter	1,881,000	0.588	0.751	1.82E+16
Reviews	(3,703,000)	0.001		
Fresh	(409,600)	0.783		
Rotten	(6,791,000)	0.000		
User Meter	(16,750,000)	0.000		
User Rating	22,870,000	0.000		
User Reviews	9,925,000	0.000		

	2000 - 2004	2005 - 2009
Adj. R-Squared	0.707	0.708
Condition No.	9.69E+15	8.99E+15

Regression Results with Select Features

2010 - 2014

Variables	Coefficient	P-Value	Adj. R-Squared	Condition No.
Total Theatres	31,050,000	0.000		
IMDb Rating	(6,926,000)	0.000		
IMDb Votes	26,160,000	0.000	0.748	4.35
Rotten	(8,045,000)	0.000		
User Rating	9,058,000	0.000		
User Reviews	9,658,000	0.000		

Test Sample Metrics

Avg Prediction ($)	$35,733,508
Avg Prediction Error ($)	$2,213,967
Avg Prediction Error (%)	6.60%
Min Prediction ($)	($29,149,846)
Max Prediction ($)	$274,829,881
Avg Actual ($)	$33,519,541
Min Actual ($)	$935
Max Actual ($)	$400,738,009

Predicting Box Office Results

Model does not generalize well for significant proportion of data due to high bias / low variance nature of Linear Regression

Distribution of Prediction Error

Legend:
- Test Sample Data
- Out of Sample Data

Bar chart values (Prediction Error):
- ≤ 10%: 60.9% (Test), 64.5% (Out of Sample)
- 10% - 25%: 7.9% (Test), 5.9% (Out of Sample)
- 25% - 50%: 6.7% (Test), 3.6% (Out of Sample)
- 50% - 75%: 2.0% (Test), 4.1% (Out of Sample)
- 75% - 100%: 2.8% (Test), 3.6% (Out of Sample)
- > 100%: 19.8% (Test), 18.3% (Out of Sample)

~30%
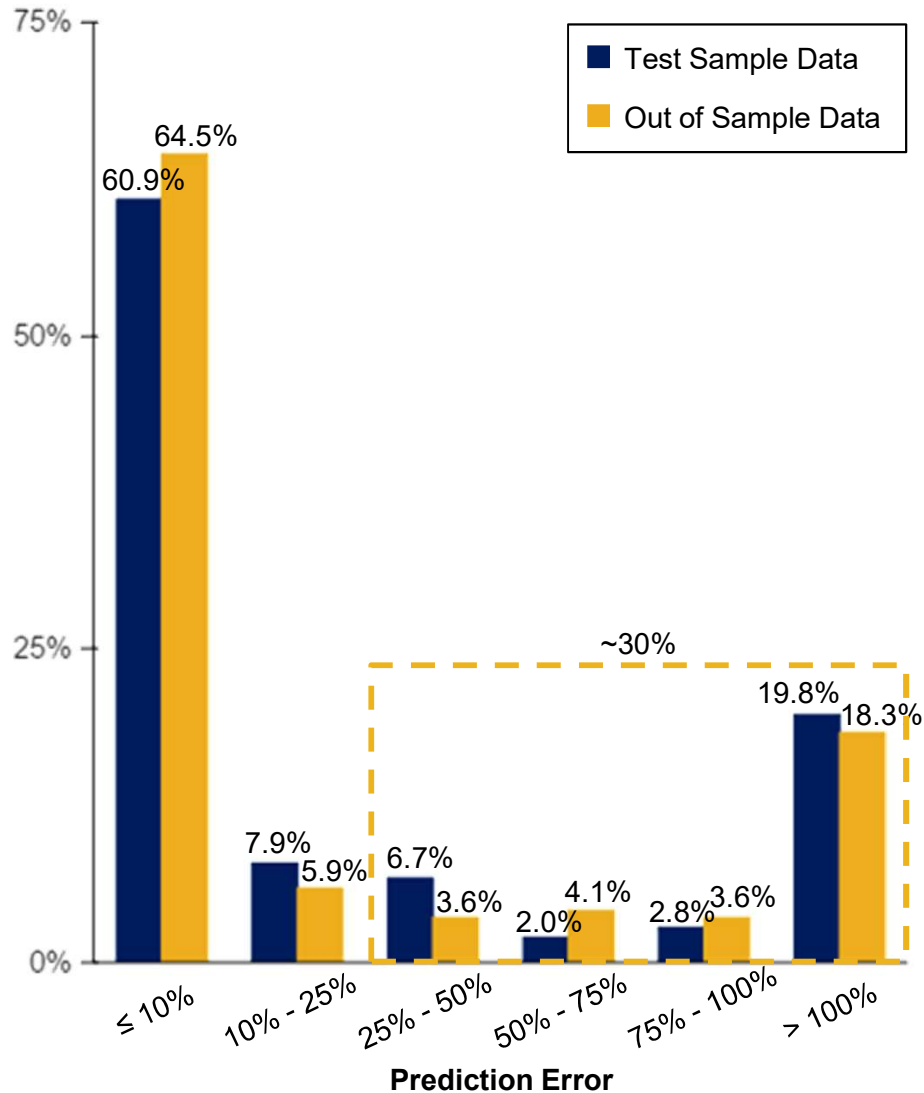
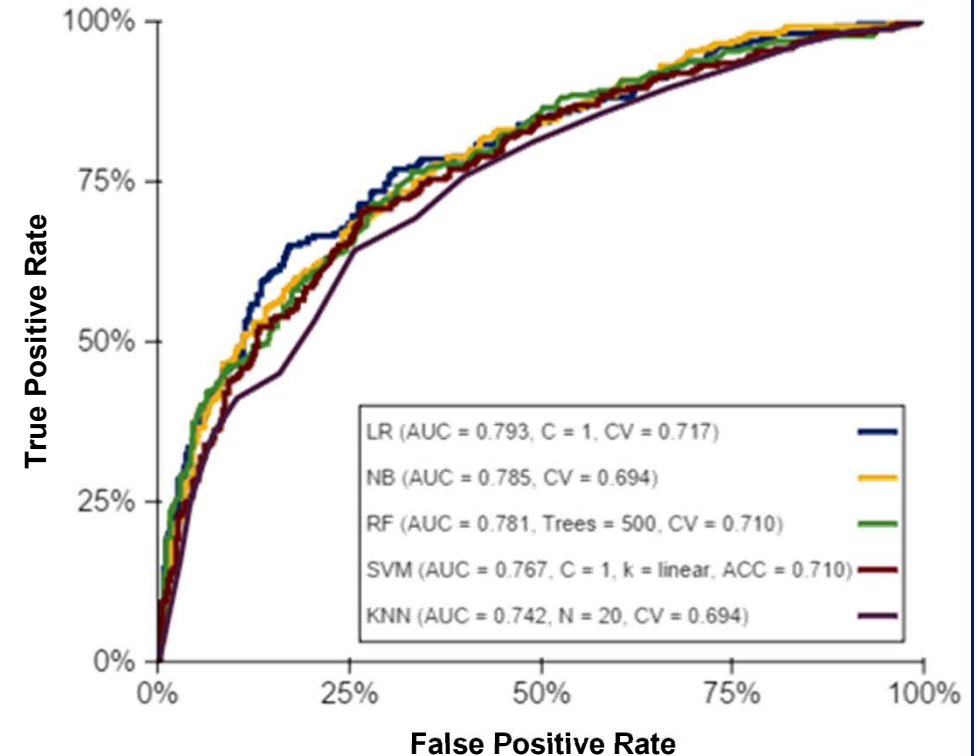|		Prediction	Actual	Error
The Good		$32,382,471	$33,078, 266	2.1%
		$74,769,715	$71,038,190	5.3%
The Bad		$185,830,779	$356,461,711	47.9%
		$72,009,359	$46,458,288	55.0%
The Ugly		$11,147,832	$1,903	585,703%
		($11,475,291)	$1,116	1,000,000%+

6

Predicting Movie Sentiment
Classifier Performance Varies Between Test Set and OOS Data

Description of Analysis

- Movie sentiment based on fresh or rotten rating received from Rotten Tomatoes

- Feature space included Total Gross, Rating, Runtime, Genre, Director, and Cast

- Separated string of multiple names and one-hot encoded Rating, Genre, Director, and Cast
  - Example: Action, Adventure, Fantasy → Genre_Action, Genre_Adventure, Genre_Fantasy

- For Logistic Regression and Random Forest classifiers:
  - Determined initial parameters for C-value and number of trees resulting in optimal CV-score
  - Used initial model to predict out-of-sample data
  - Plotted feature importance and selected cutoff range for coefficient and importance attributes
  - Tested multiple points within cutoff range to create drop list of features for model
  - Selected model producing the max CV-score to predict out-of-sample data

- SVM classifier chosen with set parameters for kernel and C-value producing optimal test set score

- KNN classifier chosen based on K-neighbors producing optimal CV-score

- No tuning performed on Naïve Bayes classifier

ROC Curves – Initial Parameters

True Positive Rate vs False Positive Rate

- LR (AUC = 0.793, C = 1, CV = 0.717)
- NB (AUC = 0.785, CV = 0.694)
- RF (AUC = 0.781, Trees = 500, CV = 0.710)
- SVM (AUC = 0.767, C = 1, k = linear, ACC = 0.710)
- KNN (AUC = 0.742, N = 20, CV = 0.694)

n = 144   Condition

	Condition +			Condition −			OOS Score
Test +	52	53	63	17	16	6	68.1%
		49	52		20	17	72.9%
							71.5%
Test −	29	23	35	46	52	40	68.1%
		26	38		49	37	61.8%

Predicting Movie Sentiment
Recursive Selection of Important Features

Logistic Regression

Random Forest

No. of Features
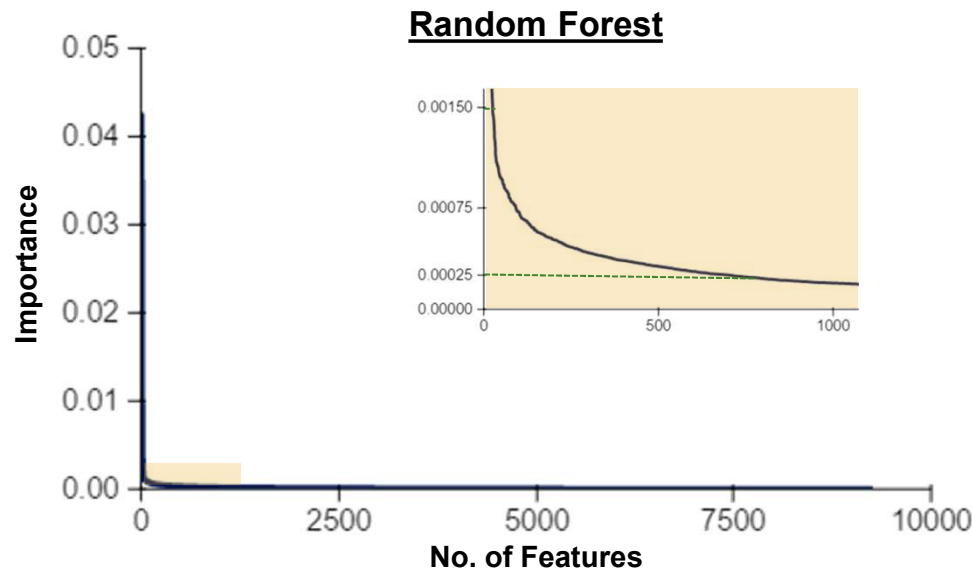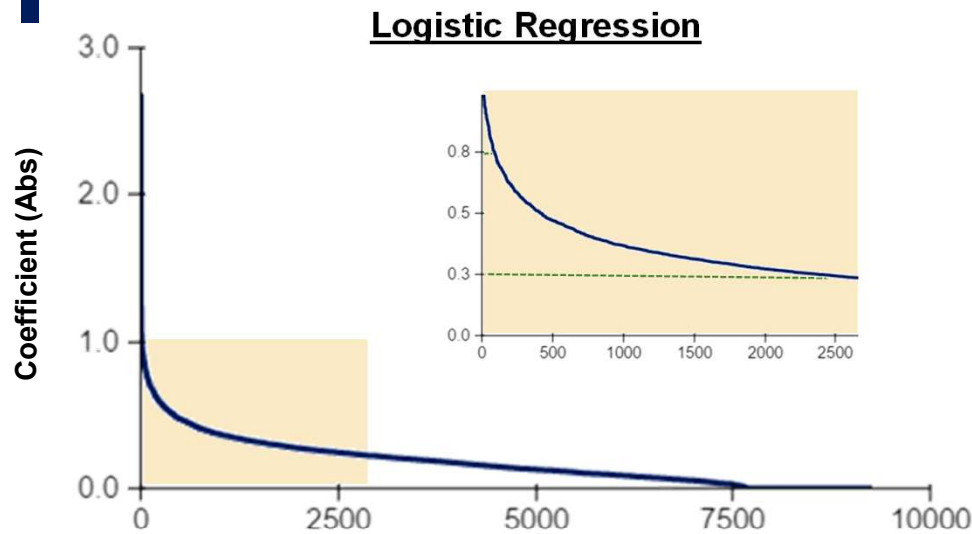
LR Top Features / Coefficients	
Documentary	2.69
Paul Giamatti	1.49
PG-13	(1.44)
PG	(1.38)
Michael Bay	(1.29)
Robert Rodriguez	1.14
Penelope Cruz	(1.14)
Ewan McGregor	1.11

RF Top Features / Importance	
Total Gross	0.042
Runtime	0.035
Documentary	0.029
Comedy	0.014
Drama	0.014
PG-13	0.011
Biography	0.011
Action	0.010

4X

Predicting Movie Sentiment

Logistic Regression CV-score suggests lower generalization error, though Naïve Bayes and Random Forest score higher on OOS data

ROC Curves – "Optimized" Parameters

LR (AUC = 0.793, C = 1, CV = 0.717)
LR* (AUC = 0.774, C = 10, CV = 0.784)
RF (AUC = 0.781, Trees = 500, CV = 0.710)
RF* (AUC = 0.770, Trees = 500, CV = 0.719)
SVM (AUC = 0.767, C = 1, k = linear, ACC = 0.710)
SVM* (AUC = 0.769, C = 1, k = linear, ACC = 0.719)
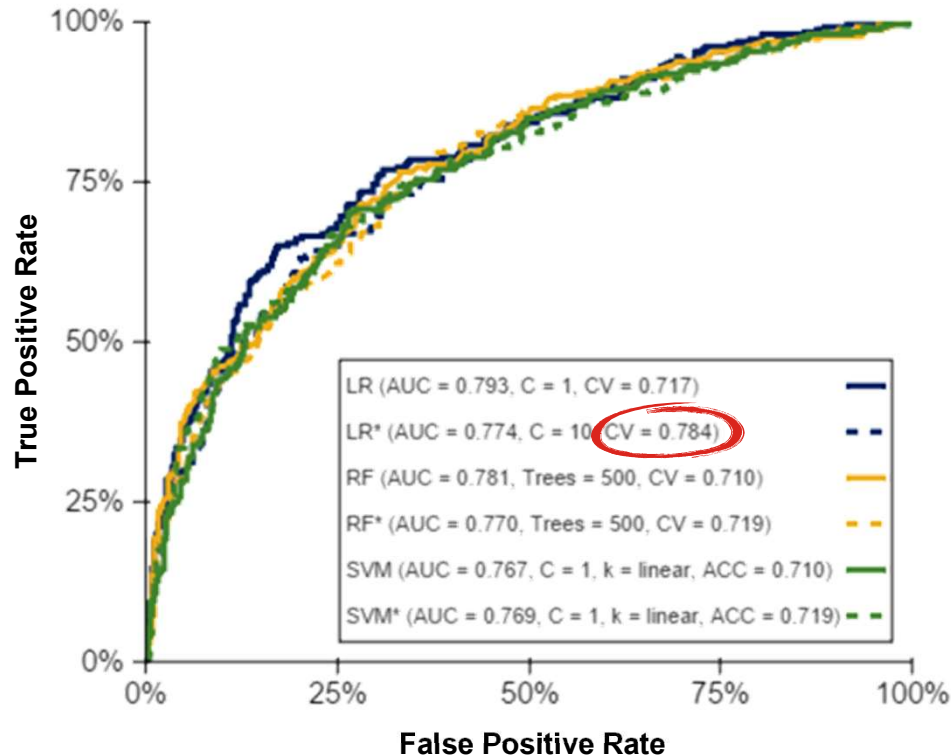
n = 144	✚		Condition	▬			OOS Score
✚	57	59	52	12	10	17	**70.1%**
Test							**70.8%**
▬	31	32	28	44	43	47	**68.8%**

LR Top Features / Coefficients	
Biography	4.85
Donald Rice	(4.15)
Jeremy Brock	4.14
Andrew Rossi	3.86
Bill Ross IV	(3.67)
Alison Klayman	3.66
Bruce Beresford	(3.55)
Aiyana Elliott	3.55

RF Top Features / Importance	
Total Gross	0.137
Runtime	0.110
Documentary	0.042
Drama	0.026
Comedy	0.024
PG-13	0.019
Action	0.017
Biography	0.016

Predicting Movie Sentiment
Plot Summary Not Highly Indicative of Performance

Top / Bottom Token Ratios (OOS)

Token	Fresh	Rotten	Ratio
Age	7	1	7:1
Park	6	1	6:1
Border	6	1	6:1
Known	6	1	6:1
Revolut	6	1	6:1
Feel	1	5	1:5
Rescu	1	6	1:6
Maria	1	6	1:6
Assassin	1	6	1:6
Special	1	6	1:6

n = 357

	Condition +	Condition −	OOS Score
Test +	98 96	66 68	60.5% (RF)
Test −	75 70	118 123	61.3% (NB)

	Plot	Prediction	Actual
Age	A widow and former songstress discovers that life can begin anew at any **age**.	(splat)	93%
Age	An **aged**, retired Sherlock Holmes deals with early dementia as he tries to remember both his final case and a mysterious woman whose memory haunts him…	(tomato)	87%
Age	En route to meet his estranged daughter and attempting to revive his dwindling career, a broken, middle-**aged** comedian plays a string of dead-end shows in the Mojave desert.	(tomato)	80%
Special	A teenage **special** ops agent coveting a "normal" adolescence fakes her own death and enrolls in a suburban high school. She quickly learns that surviving the treacherous waters…	(splat)	22%
Special	As a small-town girl catapults from underground video sensation to global superstar, she and her three sisters begin a journey of discovering that some talents are too **special** to keep hidden.	(splat)	19%
Special	In the south of France, former **special**-ops mercenary Frank Martin enters into a game of chess with a femme-fatale and her three sidekicks who are looking for revenge against a sinister Russian kingpin.	(splat)	16%

Summary of Key Learnings

- Linear Regression model does not capture likely non-linear nature of Box Office results
  - Visualizations of target and feature variables exhibit linear-to-exponential relationships
- Feature space probably too narrow to achieve "slightly better" Box Office predictions
  - More parameters expand the hypothesis space and may better represent the data
  - Number of theatres and certain ratings and reviews found to be somewhat predictive but many features were not contemplated (e.g. movie ratings, genre, cast, production budget)
- Logistic Regression, Random Forest, and SVM classifiers demonstrated best ability to generalize, though Naïve Bayes performed well on OOS data
- Precision and Specificity may be the most important measures of performance given nature of classification problem – see good movies and avoid bad ones
  - Random Forest significantly above other classifiers at 91.3% and 87.0%, with optimized Logistic Regression next highest at 82.6% and 78.6%, respectively
- Random Forest feature ranking remains similar when used to select a subset of features, due to recursive nature of determining feature importance based on impurity reduction
- Logistic Regression classifier most accurate, though Random Forest classifier most appropriate – see good movies and avoid bad ones, but maybe miss a few good ones
- Lots of work needed on plot-based sentiment predictions!
- Moving forward, go with depth not breadth