THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

An Analysis of the Movie Industry

Data Science Project 2

Presented by:

Tran Hieu Le, Totyana Hill, Fahim Ishrak, Zhilin Wang

Contents

Introduction

- The American film industry generated \$43.4 billion in revenue last year, increasing in each of the past five years at an annualized rate of just 2.2%.
- According to the IBISWorld report, even with stagnant box office revenue, the industry's revenue will continue to increase.
- Most films shown at the domestic box office make up less than a quarter of the revenue they generate.
- If box office predictions are more accurate, then studios and investors would save millions and appropriately allocate the necessary funds.

Questions

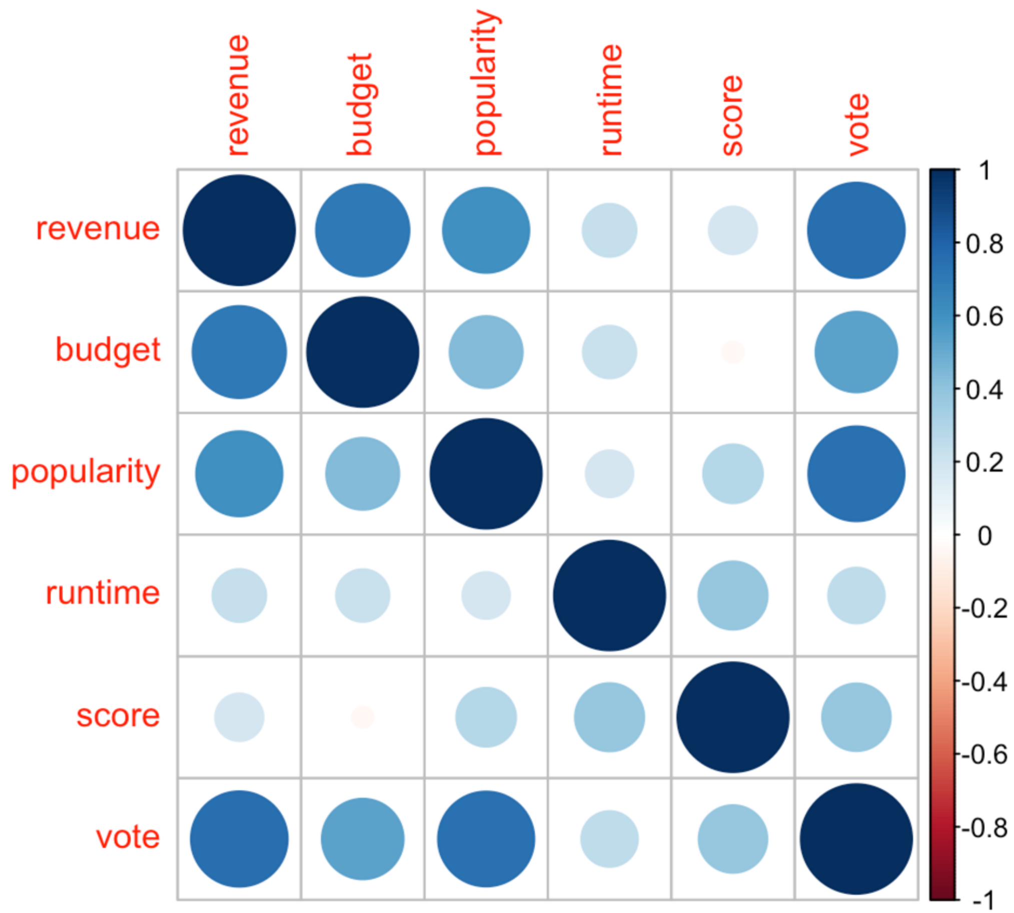
- ➢ Which factors will best determine a movie's revenue performance and profitability?
- ➢ How does seasonality impact revenue generated?

Goals

- ➢ Construct and evaluate different kinds of models to predict a movie's revenue and profitability.
- ➢ Evaluate models and find the best predictive model.

Revenue Prediction

Correlation Matrix



Anova Test

Predictor	Company	Season	Genres
p-value	1.13e-28	5.29e-10	2.58e-80

- All the p-values are smaller than 0.05 level.
- The overall effect of company, season and genre on revenue are statistically significant.

Data Summary

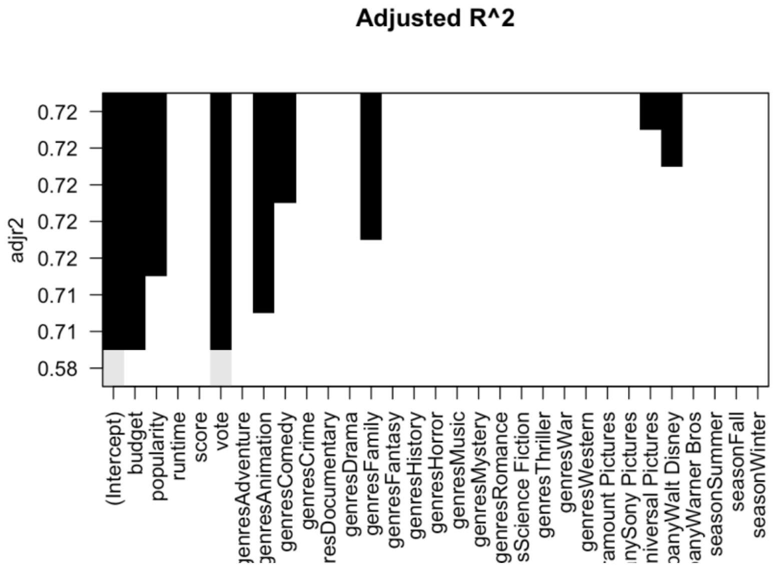
	revenue	budget	vote	score	popularity	runtime
Min.	5.00e+00	1.0e+00	1	2.30	0	41
1st Qu.	1.71e+07	1.05e+07	179	5.80	10	96
Median	5.52e+07	2.50e+07	471	6.30	20	107
Mean	1.21e+08	4.07e+07	978	6.31	29	111
3rd Qu.	1.46e+08	5.50e+07	1148	6.90	37	121
Max	2.79e+09	3.80e_08	13,752	8.50	876	338

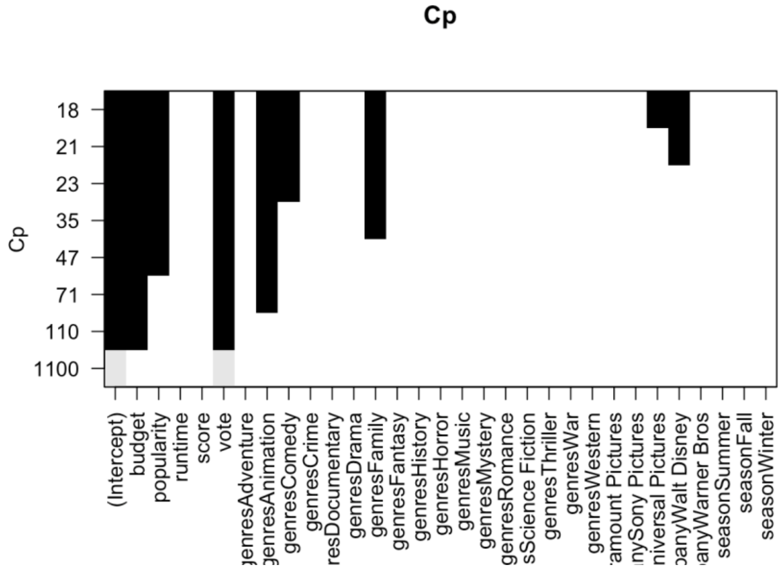
Linear Regression

Formula: revenue ~. , data = training data			
R-squared	Adj R-squared	F-statistic	p-value
0.725	0.721	188	<2e-16

- ➢ All vifs are below 3.
- ➢ Score, runtime are not statistically significant.
- ➢ The effects of four seasons are not statistically significant.

Feature Selection





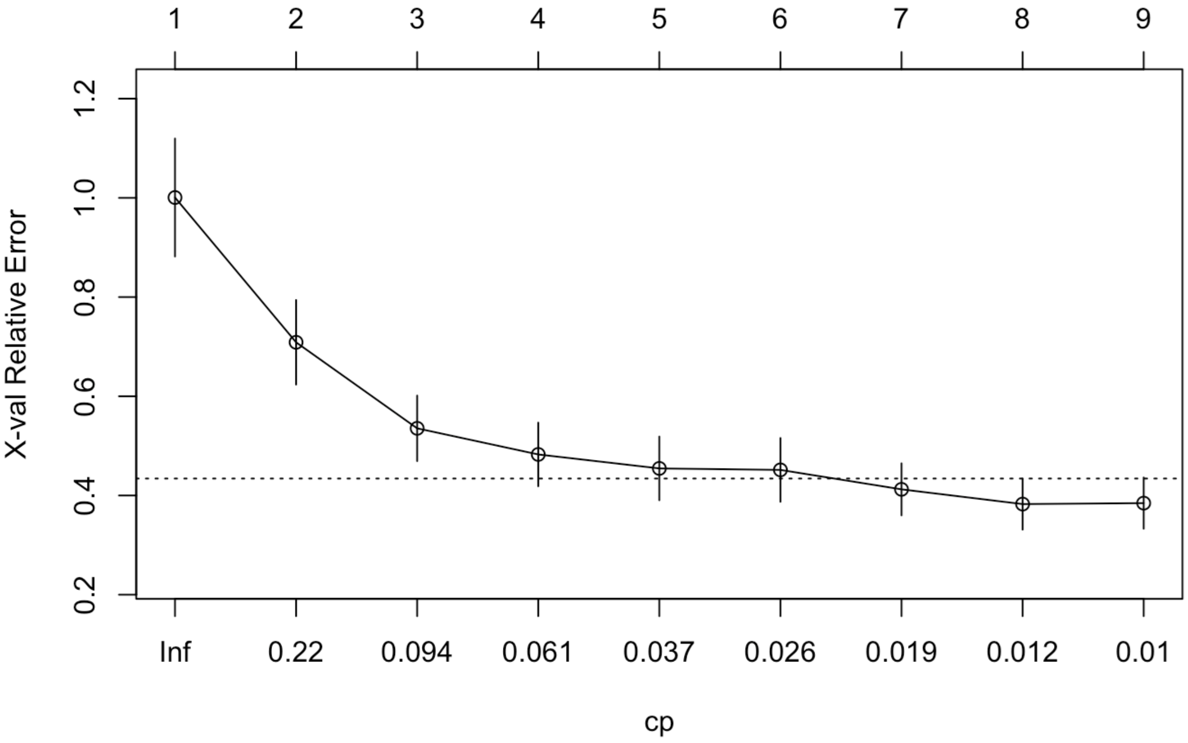
THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Linear Regression

revenue ~ vote + popularity + budget + company + genres			
R-squared	Adj R-squared	F-statistic	p-value
0.725	0.721	255	<2e-16

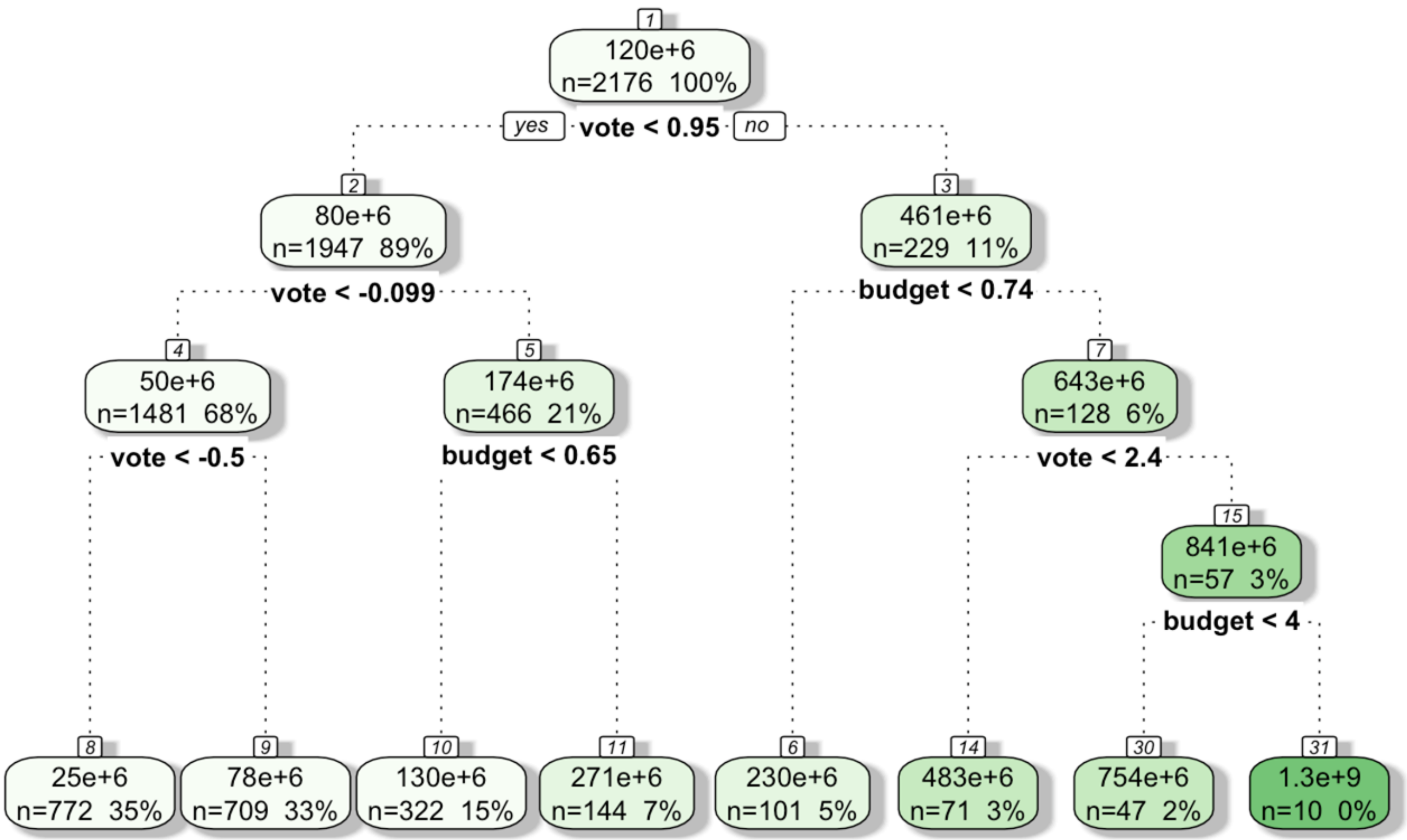
- The results are the same as the model with full predictors.
- All coefficients of numerical variables are statistically significant.

Regression Tree



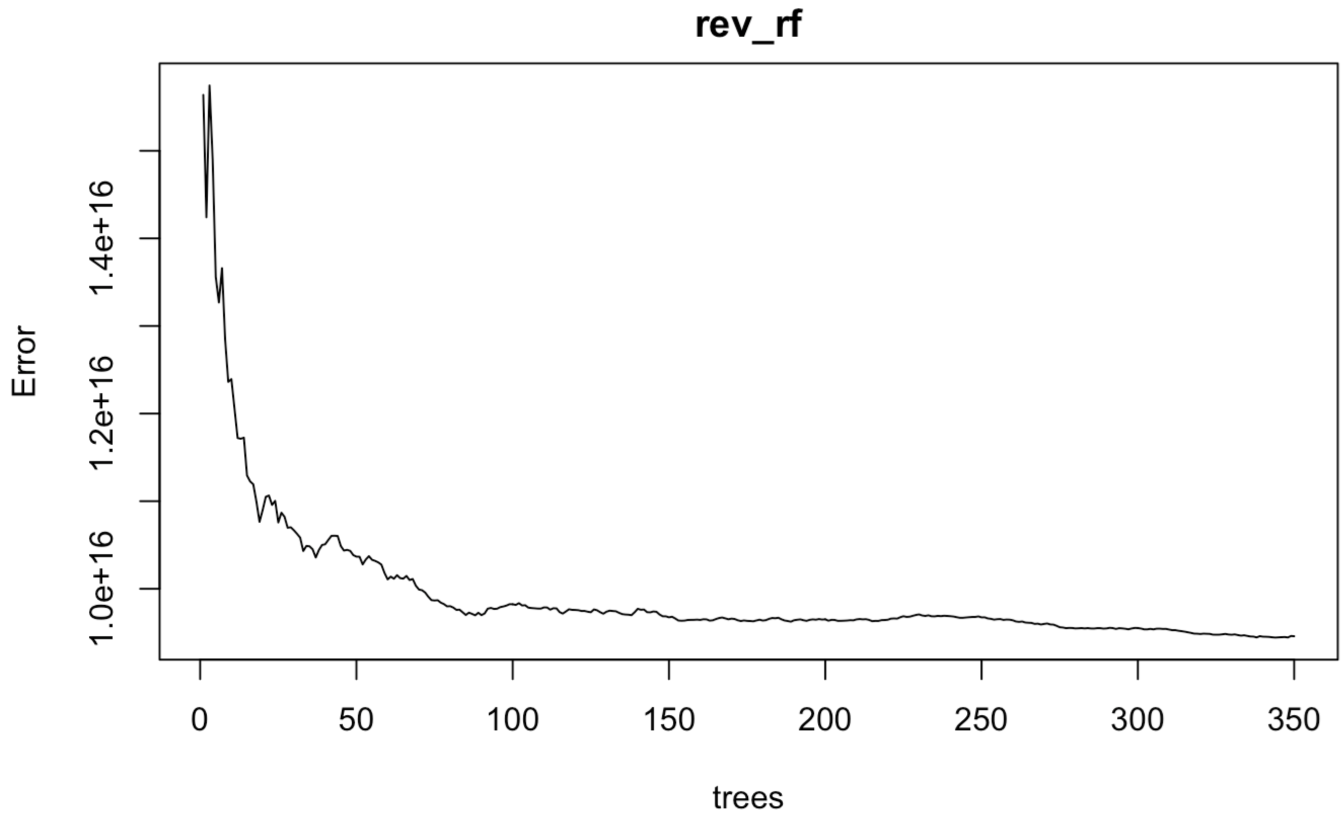
CP	Rel Error
1	1.000
2	0.709
3	0.535
4	0.483
5	0.455
6	0.452
7	0.412
8	0.383 (min)
9	0.385

Pruned Tree



1
120e+6
n=2176 100%
yes
vote < 0.95
no
2
80e+6
n=1947 89%
vote < -0.099
3
461e+6
n=229 11%
budget < 0.74
4
50e+6
n=1481 68%
vote < -0.5
5
174e+6
n=466 21%
budget < 0.65
7
643e+6
n=128 6%
vote < 2.4
15
841e+6
n=57 3%
budget < 4
8
25e+6
n=772 35%
9
78e+6
n=709 33%
10
130e+6
n=322 15%
11
271e+6
n=144 7%
6
230e+6
n=101 5%
14
483e+6
n=71 3%
30
754e+6
n=47 2%
31
1.3e+9
n=10 0%

Random Forest



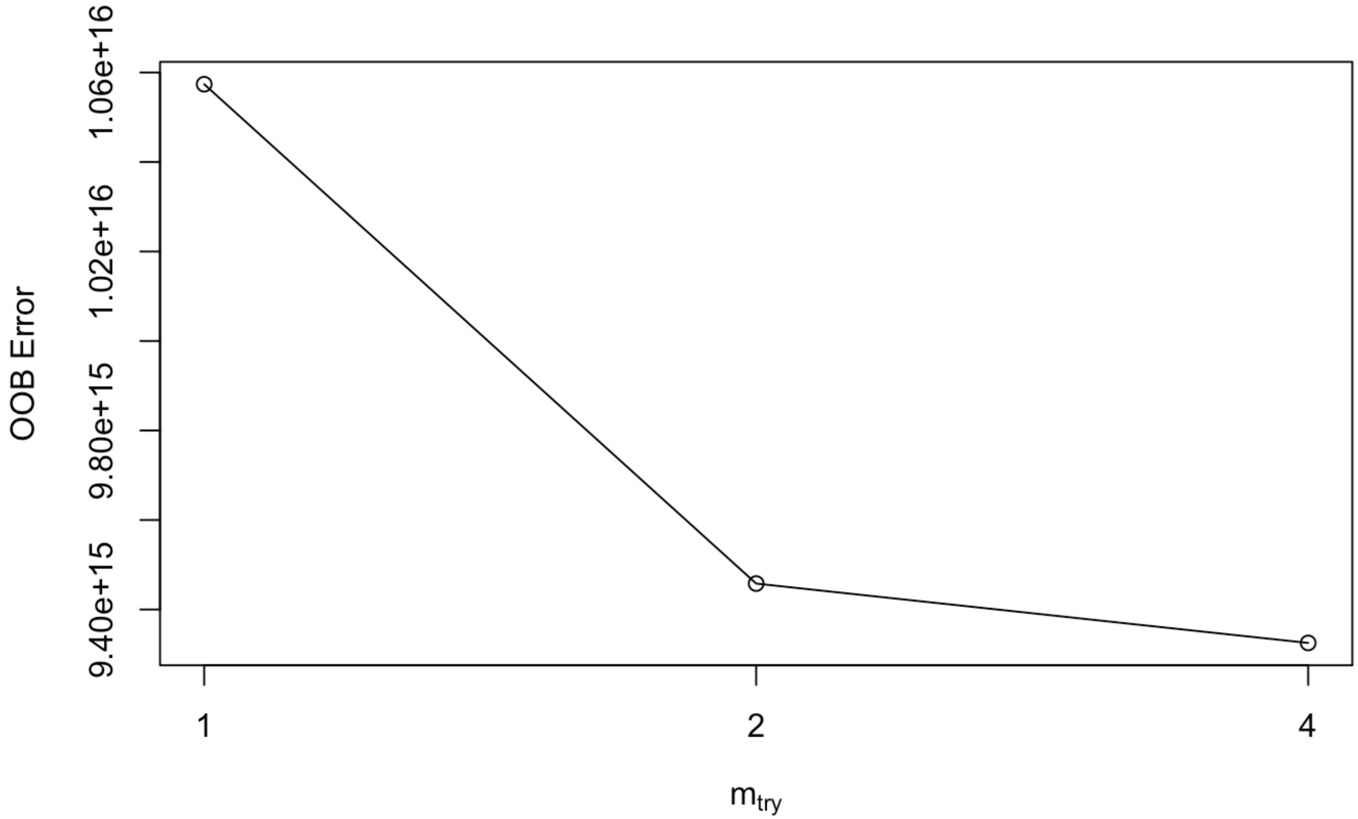
Random Forest

Formula: revenue ~. , ntree = 350, data = training		
Pseudo R-squared	OOB Error[350]	mtry
0.732	9.46e+15	2

- Number of variables at a split is 2.
- The R-squared in RF is higher than in LM.

Variable	Importance
budget	1.90e+19
popularity	1.43e+19
runtime	4.15e+18
score	3.39e+18
vote	2.33e+19
genres	5.58e+18
company	2.60e+18
Season	1.49e+18

Hyperparameter Tuning



Tuned Random Forest

revenue ~. , ntree = 350, mtry = 4	
Pseudo R-squared	OOB Error[350]
0.740	9.17e+15

- ➢ Pseudo R-squared improves from 0.732 to 0.740.
- ➢ Lower OOB Error[350] / MSE.
- ➢ After tuning, we receive better results.

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Model Comparison

Metrics		Linear Model	Regression Tree	Random Forest
R-squared		0.721	0.711	0.741
MAE	Train	5.8e+07	6.0e+07	5.2e+07
	Test	6.2e+07 (+6.9%)	6.6e+07 (+10%)	5.5e+07 (+5.7%)
RMSE	Train	9.9e+07	1.0e+08	9.6e+07
	Test	1.0e+08 (+1%)	1.1e+08 (+10%)	9.7e+07 (+1%)

➢ Random Forest has the best performance.

Profit Prediction

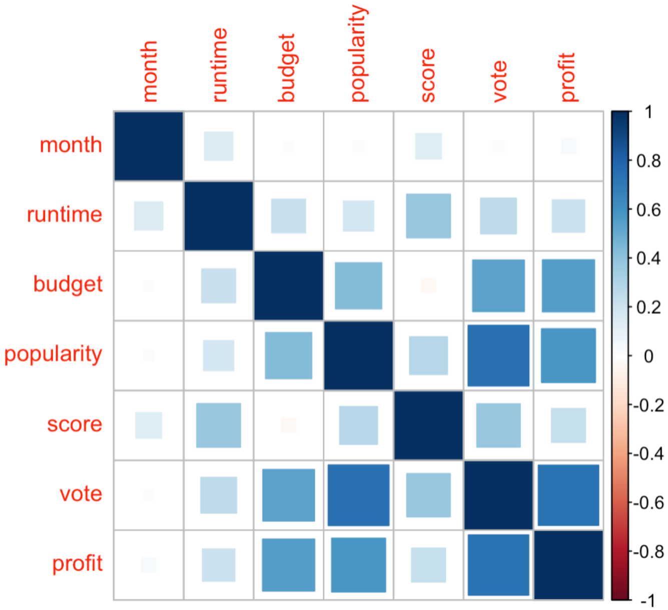
Profit

Profit Analysis

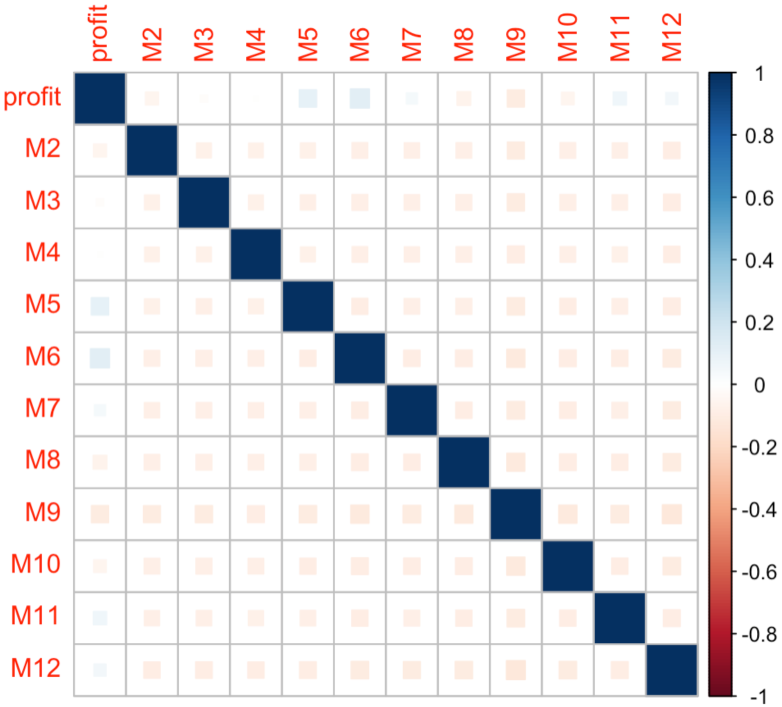
1. Briefly examine the correlations between profit and the other variables
2. Build a Linear Regression model by feature selection
3. Attempt to train the data with Ridge Regression and Lasso Regression to see if it might be overfit
4. Build random forest regression
5. Try to predict the profit of some movies in the future

Profit

Correlation Matrix



Profit vs. numerical variables



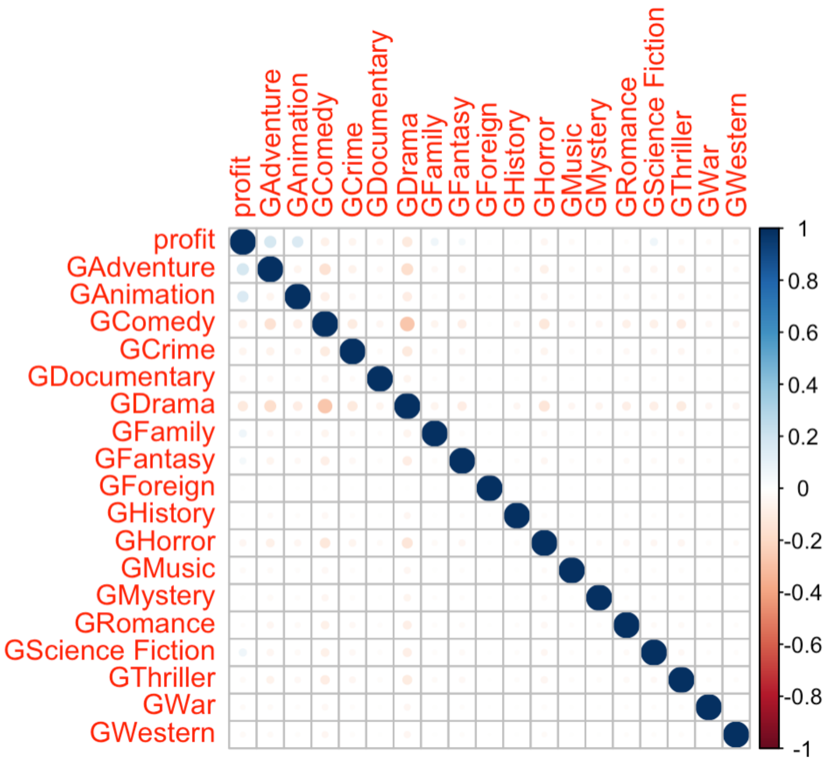
Profit vs. Month

Profit

Correlation Matrix Cont'd

	profit	Paramount	Sony	Universal	Disney	Warner Bro
profit	1.00000	0.0438	0.00886	0.0952	0.0979	-0.00719
Paramount	0.04375	1.0000	-0.08979	-0.1002	-0.1309	-0.07964
Sony	0.00886	-0.0898	1.00000	-0.1048	-0.1370	-0.08332
Universal	0.09521	-0.1002	-0.10485	1.0000	-0.1529	-0.09300
Disney	0.09791	-0.1309	-0.13696	-0.1529	1.0000	-0.12148
Warner Bro	-0.00719	-0.0796	-0.08332	-0.0930	-0.1215	1.00000

- Company does not have strong correlation with profit overall.
- Different companies have significantly different impacts on profit



Profit vs. Genre

THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

Profit

Linear Regression

```
Residual standard error: 0.632 on 3186 degrees of freedom
Multiple R-squared:  0.606,     Adjusted R-squared:  0.601
F-statistic:  125 on 39 and 3186 DF,  p-value: <2e-16
```

- Pick relevant and useful variables.
- Then take all the variables into account.
- The r-squared is 0.601

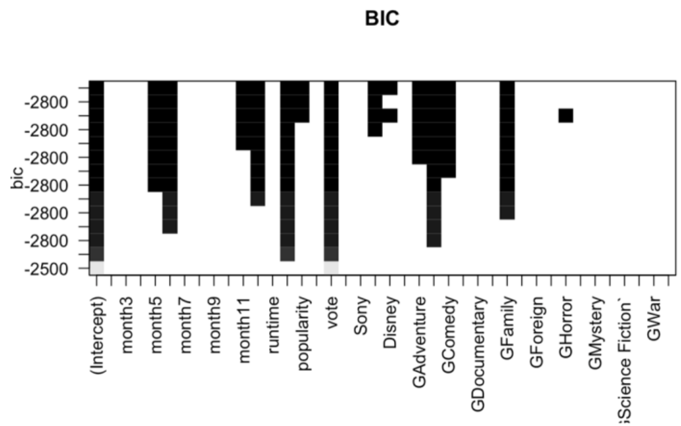
```
Residual standard error: 93200000 on 3160 degrees of freedom
Multiple R-squared:  0.66,      Adjusted R-squared:  0.653
F-statistic: 94.3 on 65 and 3160 DF,  p-value: <2e-16
```

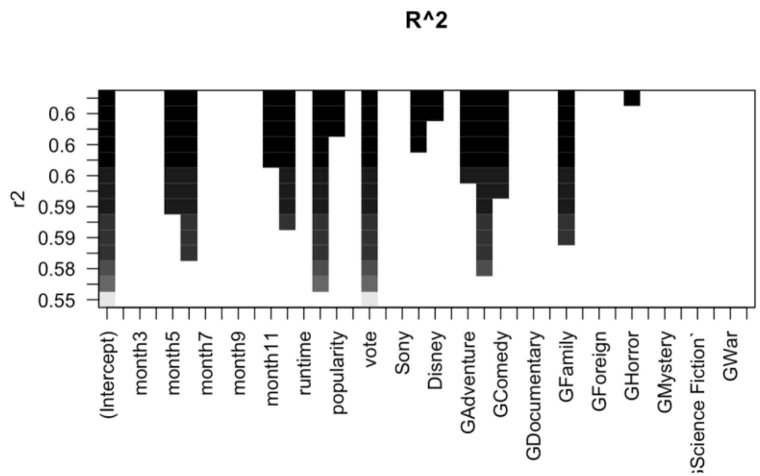
- The model can be improved by adding some polynomials
- The amount of predictors becomes increasingly large

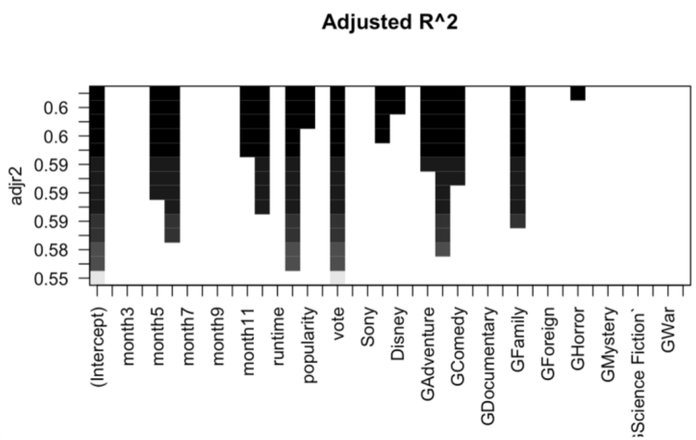
THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

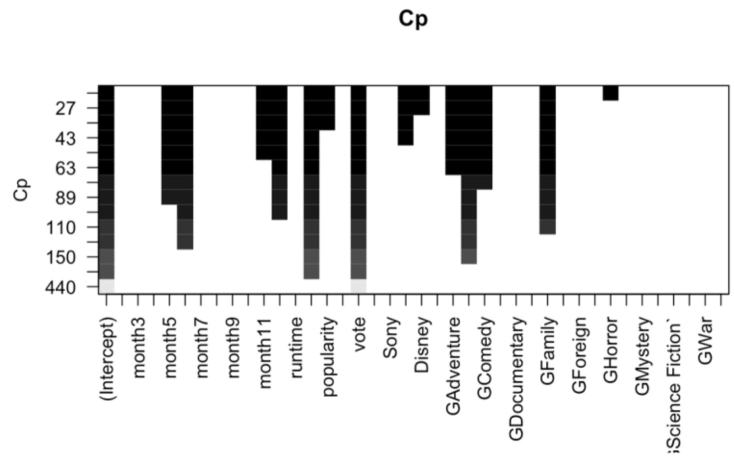
Profit

Feature Selection









THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

Profit

Linear Regression w/ Feature Selection

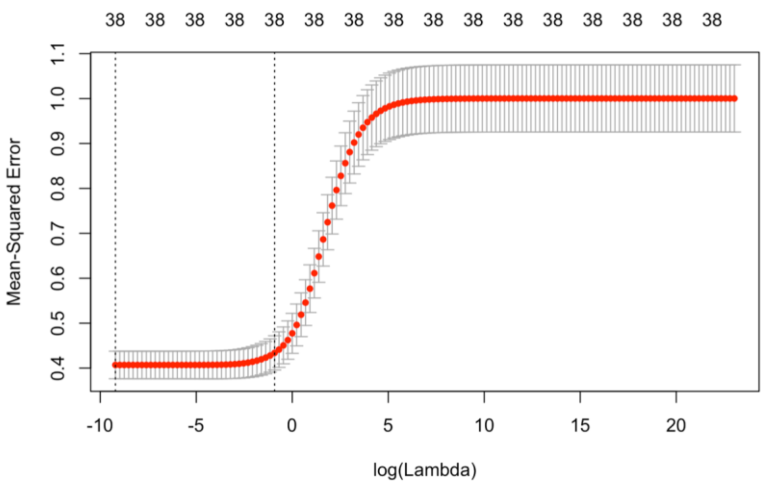
```
Residual standard error: 0.64 on 3218 degrees of freedom
Multiple R-squared:  0.592,	Adjusted R-squared:  0.591
F-statistic:  666 on 7 and 3218 DF,  p-value: <2e-16
```

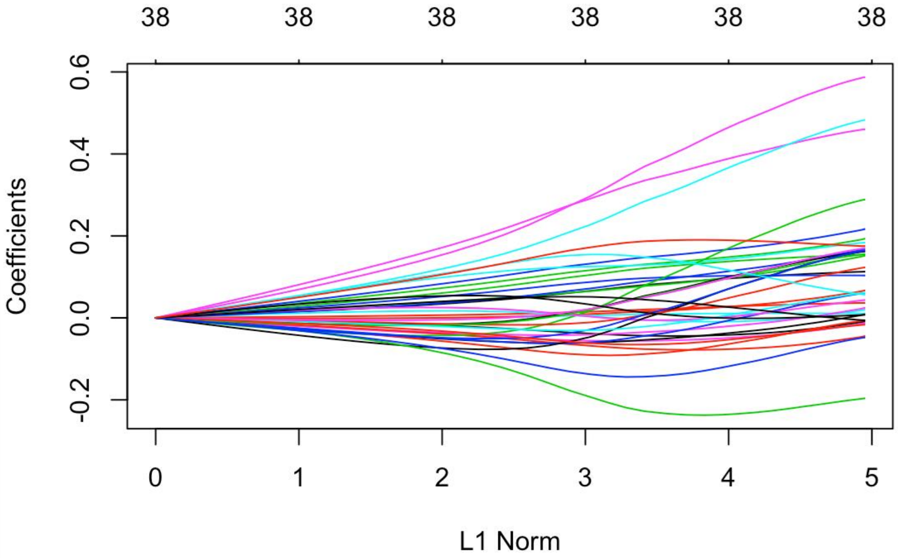
- Originally, we have 39 features as predictors, and the r-square is 0.601. To achieve a higher r-square of 0.653, we may have around 60 features
- With feature selection, we reduce the amount of predictors to 7: (budget, vote, May, June, December, genre Animation, and genre family)
- The new r-square value is 0.591, pretty close to 0.601

Profit

Ridge Regression

- ➢ Ridge Regression
  - ○ best lambda: 0.0398
  - ○ the new r-square: 0.604

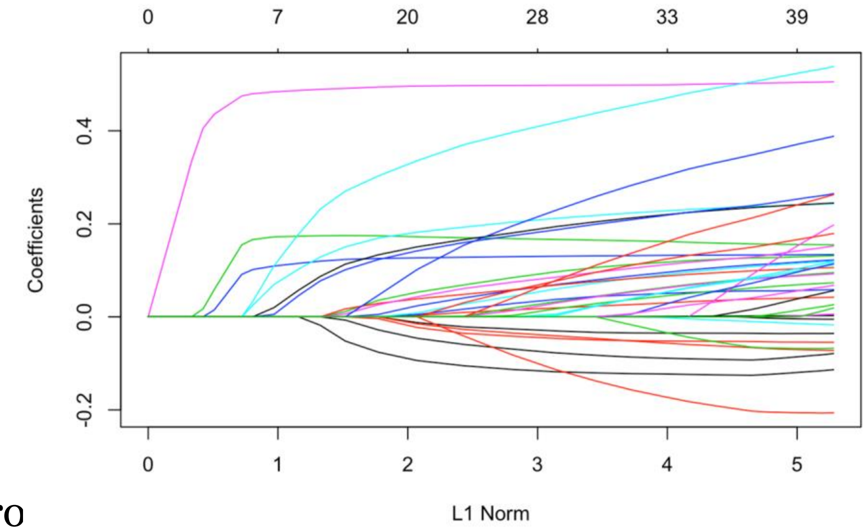




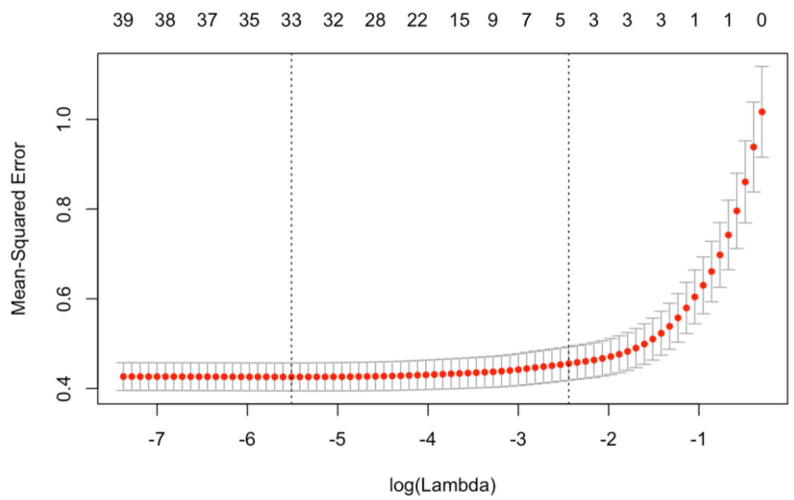
Profit

Lasso Regression

- Lasso Regression
  - best lambda: 0.00501
  - the new r-square: 0.604
  - 29 of the coefficients become zero

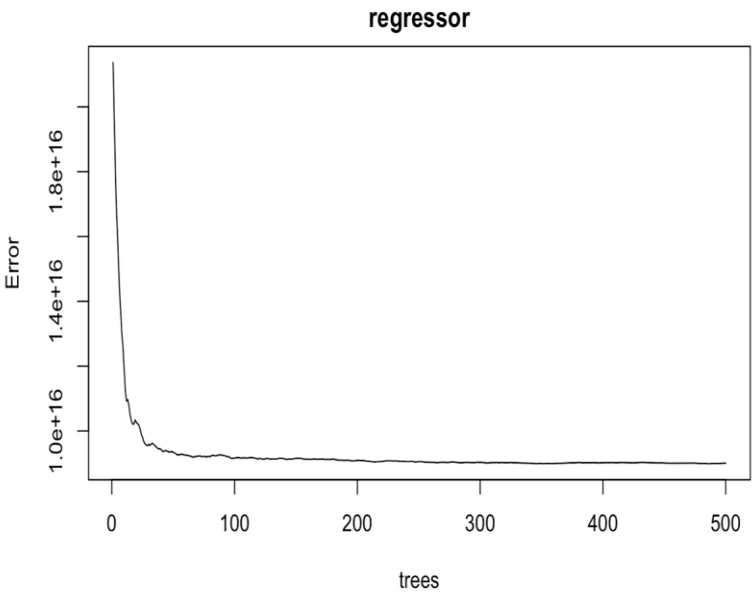



```
companyParamount Pictures    companySony Pictures    companyWarner Bros          month2
                        0                       0                     0               0
                   month3                  month4                month5          month7
                        0                       0                     0               0
                   month8                  month9               month10         month11
                        0                       0                     0               0
                  month12             genreComedy            genreCrime genreDocumentary
                        0                       0                     0               0
               genreDrama            genreFantasy          genreForeign    genreHistory
                        0                       0                     0               0
              genreHorror              genreMusic          genreMystery    genreRomance
                        0                       0                     0               0
    genreScience Fiction           genreThriller              genreWar    genreWestern
                        0                       0                     0               0
                    score
                        0
```



Profit

Random Forest Regression

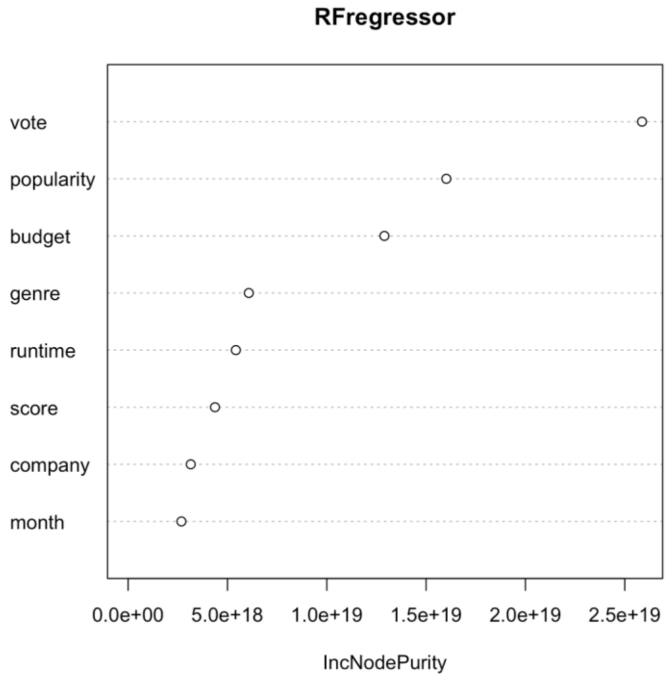



```
Call:
 randomForest(x = movies_rf[-9], y = movies_rf$profit, ntree = 500)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 2

          Mean of squared residuals: 9e+15
                    % Var explained: 64
```

Profit

Random Forest Regression



Variable Importance

From high to low:

Vote

popularity

budget

genre

runtime

score

company

month

Profit

Actual Predictions

Frozen II (2019)

- Release in Nov.
- Carries a $33 million budget
- Animation
- Runtime 1h44min
- etc.

The predict profit would be:

$521,197,933 -with RF

$590,351,271 -with LM

Profit

Extension

Box Office: 'Frozen 2' Breaks More Box Records And Reaches $739 Million Worldwide

- However, 'Frozen II' already earns a lot more than we expect
- Factors that affect profit are complicated and diverse. Like brand influence and the rivals.

and-counting). That said, it's entirely possible that *Frozen II* won't drop dead after Thanksgiving weekend, since it's the only biggie between now and *Jumanji: The Next Level* on December 13. It's both the big fantasy adventure and the big toon in the marketplace.

Principal Component Analysis

Principal Component Analysis

Purposes

- Dimensional reduction.
- Data storage optimization.
- Computational speed optimization.

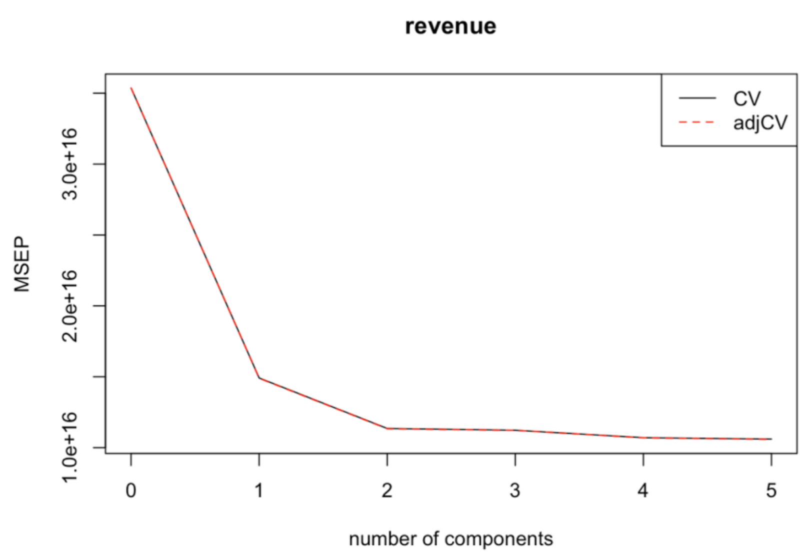
Expectation

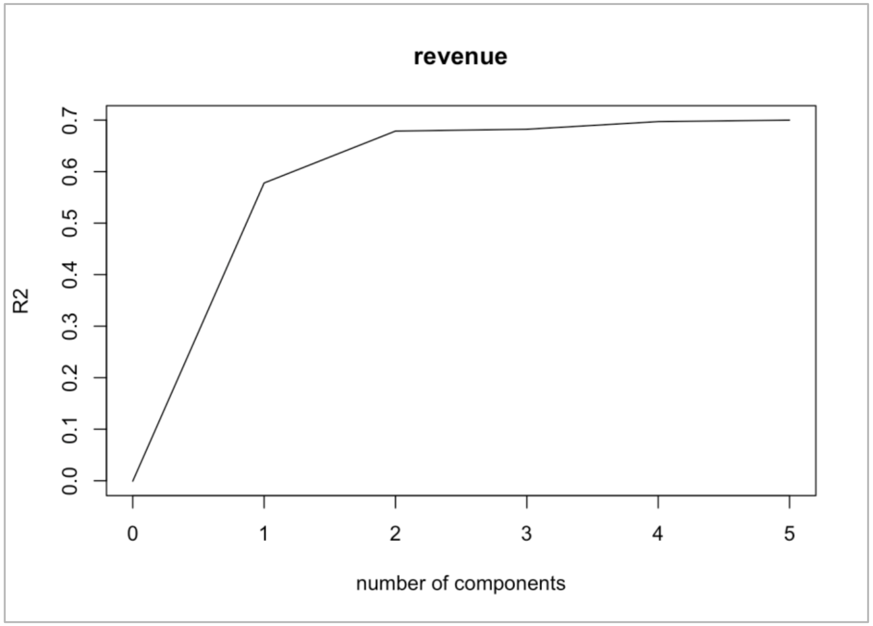
- Not losing much predictive power.
- Remove multicollinearity between predictors

Model

- revenue ~ budget + score + vote + popularity + runtime

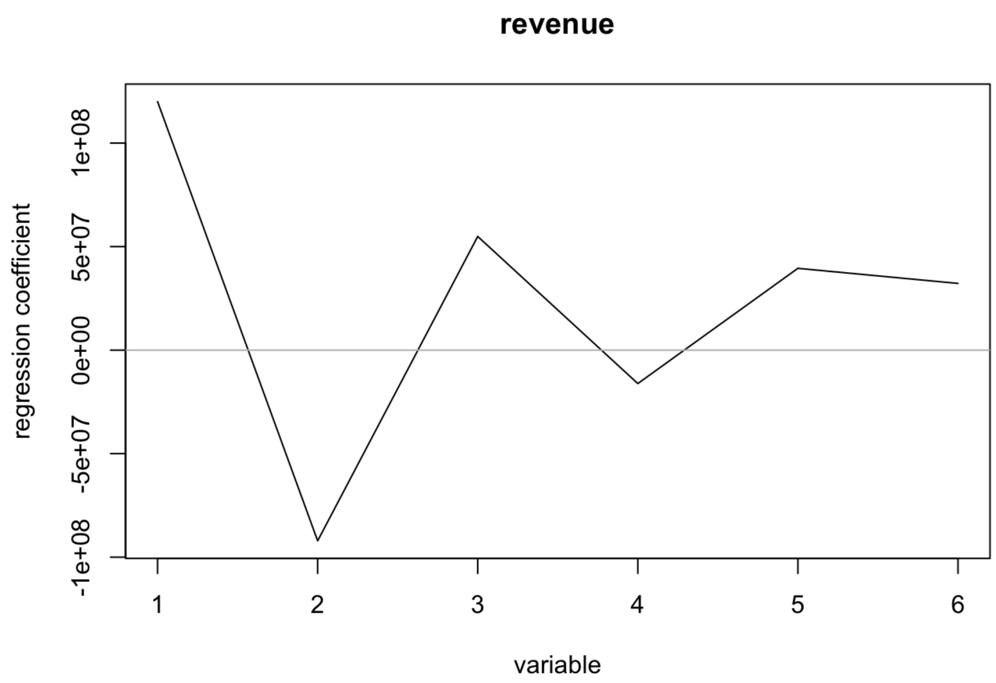
Principal Component Analysis





2 components are enough to capture 95% of the MSEP and R-squared.

Principal Component Coefficients



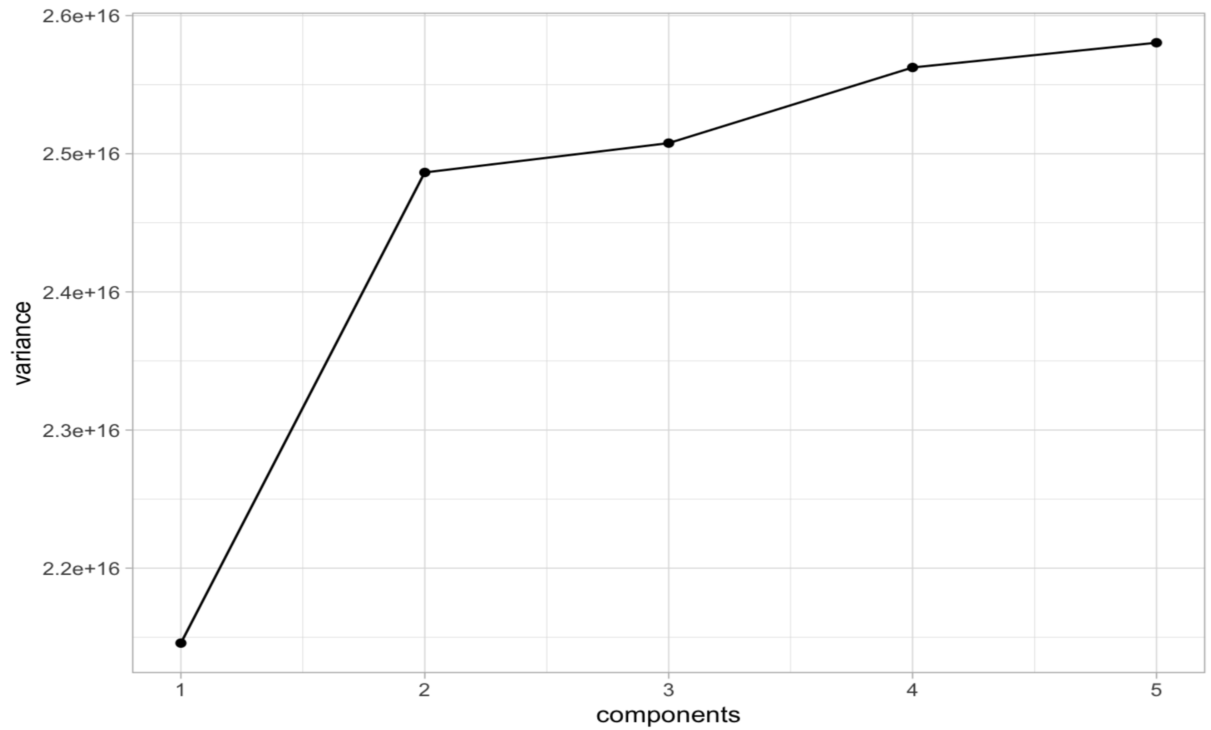
After the second components, the differences of coefficients are not drastic.

Principal Component Regression

Training	% var explained				
X	48.50	71.69	87.41	95.61	100.00
revenue	58.21	68.09	68.67	70.49	71.13

- ➢ To capture 80% of the variance of predictors, we need 3 components.
- ➢ To explain the response revenue, 2 components are enough.

PCR Validation on Testing Data



- With 2 components, we still capture 95% of the variance.
- Our PCR seems to perform properly on the testing data.

Linear Model with PCs

Metrics	5 Original Variables	2 Principal Components
Adj R-squared	0.711	0.681
RMSE - train	1.01e+08	1.06e+08
RMSE - test	1.06e+08	1.55e+08 (+46%)

- The R-squared does not reduce a lot when we using PCA.
- The predictive power seems to be influenced as RMSE in testing set increases by 46%.
- Potential overfitting when using PCs to build linear model.

Profitability Prediction

Model Overview

- ➢ Response: y – binary output (0,1).
  - ○ 0 – the movie does not gain profit (revenue < budget).
  - ○ 1 – the movie gains profit (revenue > budget).
- ➢ Predictors:
  - ○ Numerical variables: budget, runtime, vote, popularity, score.
  - ○ Categorical variables: company, season, genres.

Logistic Regression

Formula: y ~ . , data = training data		
Null deviance	Residual deviance	AIC
2409.2	1805.0	1867

We observe that some predictors are not statistically significant.

Wald Test

Variables	Company	Season	Genres
p-value	8.4e-09	0.045	0.12

- The overall effect of company is clearly statistically significant.
- The overall effect of season is not statistically significant.
- Genres can be removed from the model formula.

Model Selection - AIC

budget	popularity	runtime	score	vote	genres	company	season	criterion
true	false	false	true	true	false	true	true	1855
true	true	false	true	true	false	true	true	1856
true	false	true	true	true	false	true	true	1857
true	true	true	true	true	false	true	true	1858
true	false	false	true	true	false	true	false	1858

Best logit model formula: y ~ budget + score + vote + company + season

Predicted Probability & Accuracy

Predicted Probability Cut-off	Accuracy	Kappa
0.5	81.9%	0.425 (moderate)
0.6	80.7%	0.486 (moderate)
0.7	77.4%	0.476 (moderate)
0.8	69.0%	0.379 (fair)
0.9	59.2%	0.280 (fair)

- Predicted Probability > cut-off, y = 1; otherwise y = 0.
- The results show that the cut-off 0.5 gives the best prediction on testing set, while the cut-off 0.6 gives the highest interrater reliability.

Exponentiated Coefficients

Variables	exp(coef)
budget	0.99999999817 (~ 1)
score	1.234
vote	1.003
Paramount Pictures	2.623
Universal Pictures	2.322
Sony Pictures	1.729
Walt Disney	2.399
Warner Bros	2.200
Summer	1.469
Fall	0.993
Winter	1.055

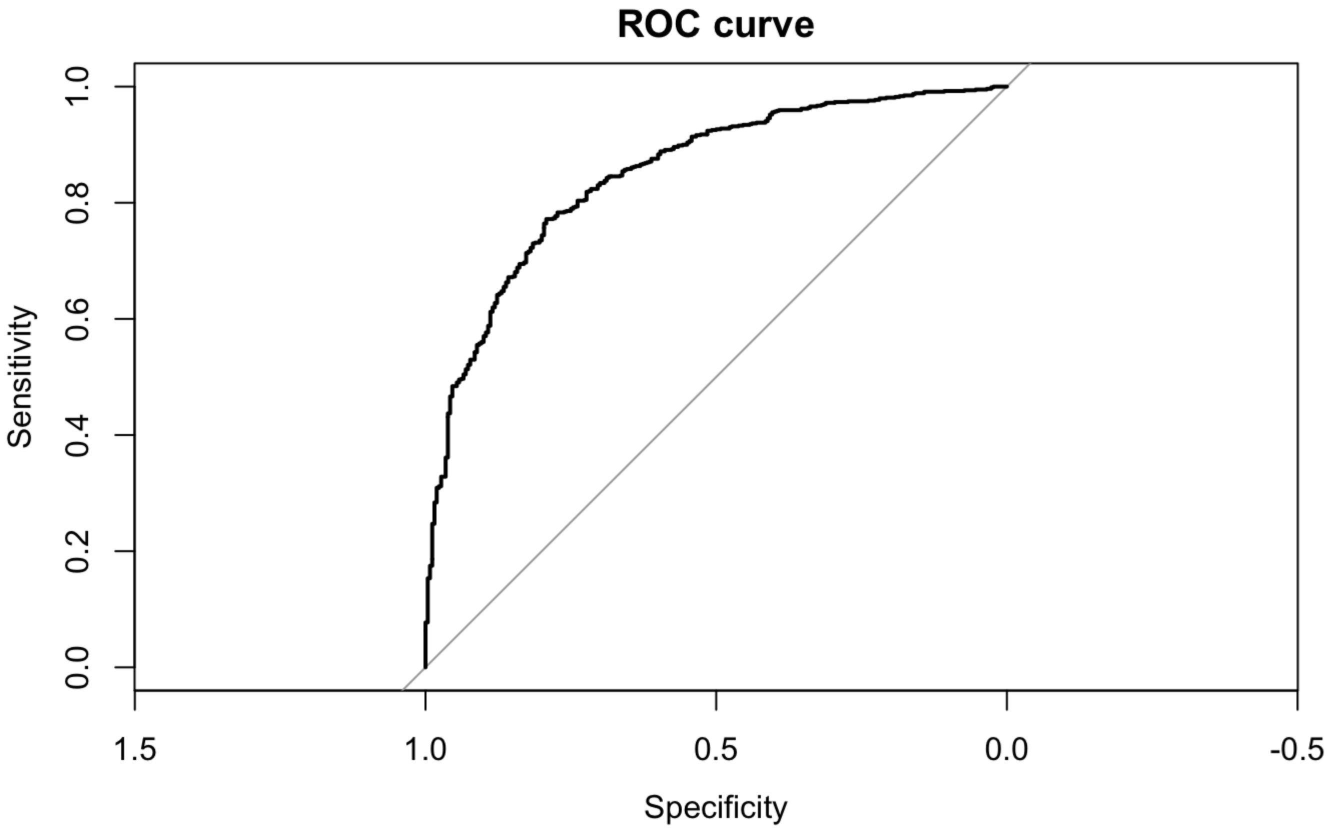
THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

Model Evaluation

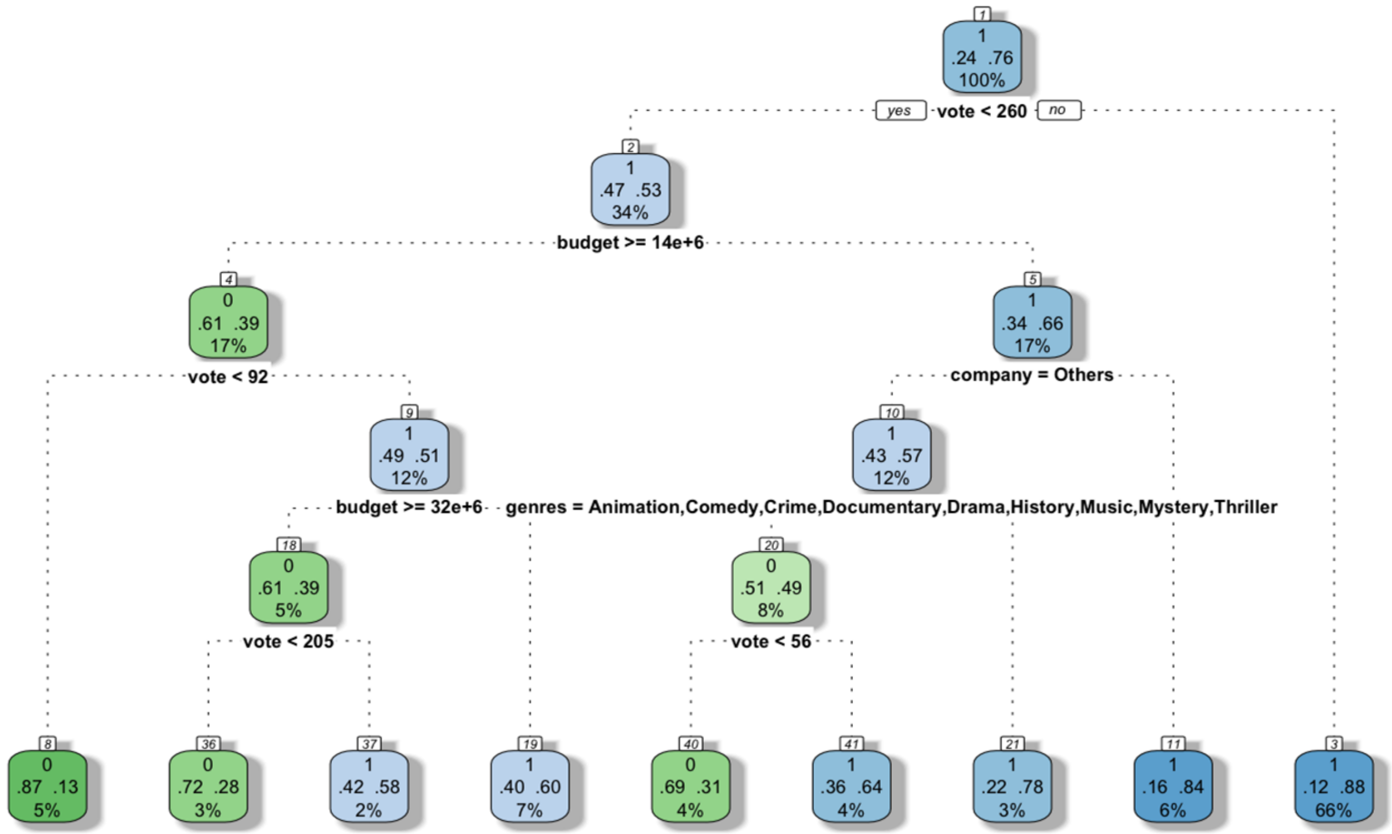
- Hosmer and Lemeshow goodness of fit:
  - Since p-value <2e-16, our logit model seems to be a good fit.
- McFadden:
  - 23.9% variations in y is explained by explanatory predictors in the model.

ROC Curve and AUC



AUC = 0.849

Classification Tree

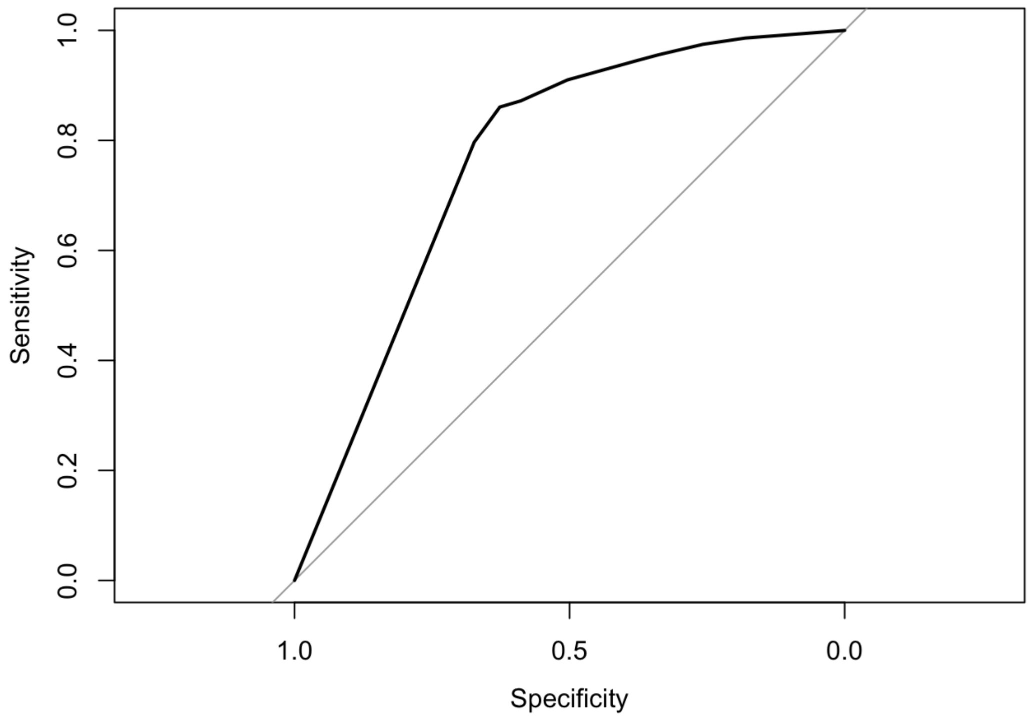


Confusion Matrix - Decision Tree

actual / prediction	0	1
0	88 (True Negative)	35 (False Positive)
1	172 (False Negative)	754 (True Positive)

- accuracy = 80.3%
- kappa = 0.357 (fair)

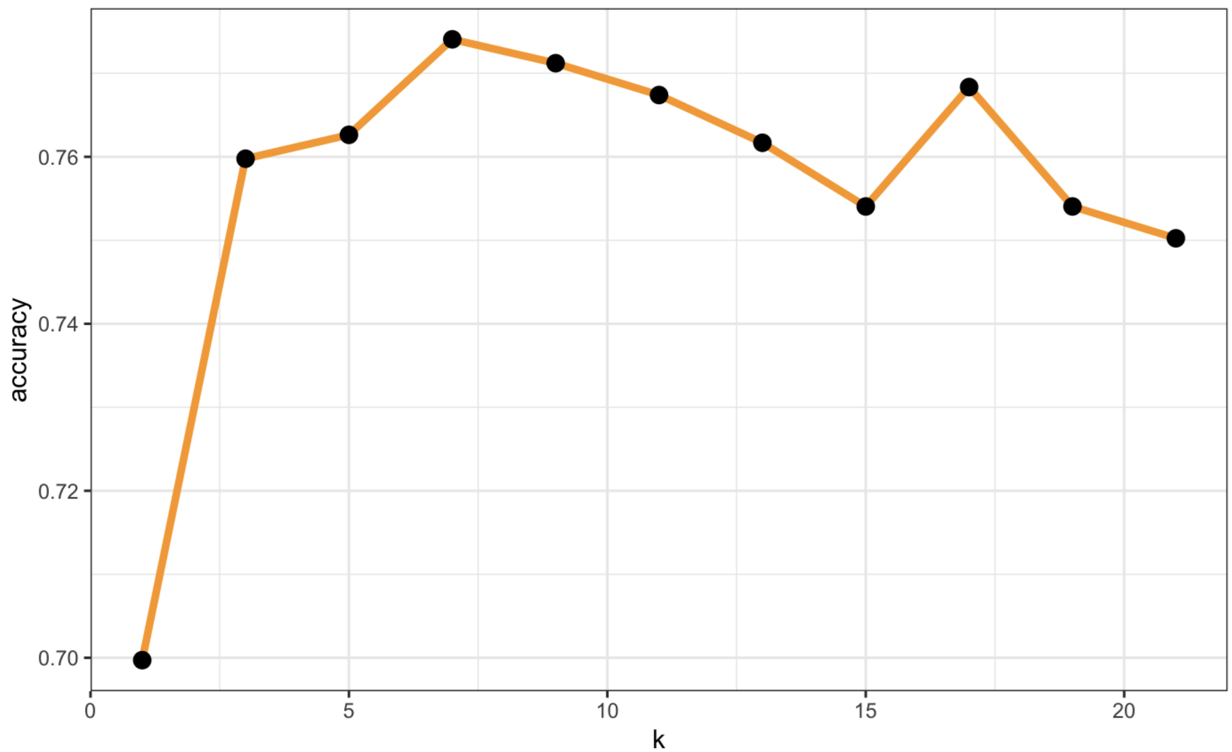
ROC Curve and AUC



AUC = 0.764

Logit Model is better than Classification Tree in our case.

KNN Model



➢ Best accuracy at k = 7.

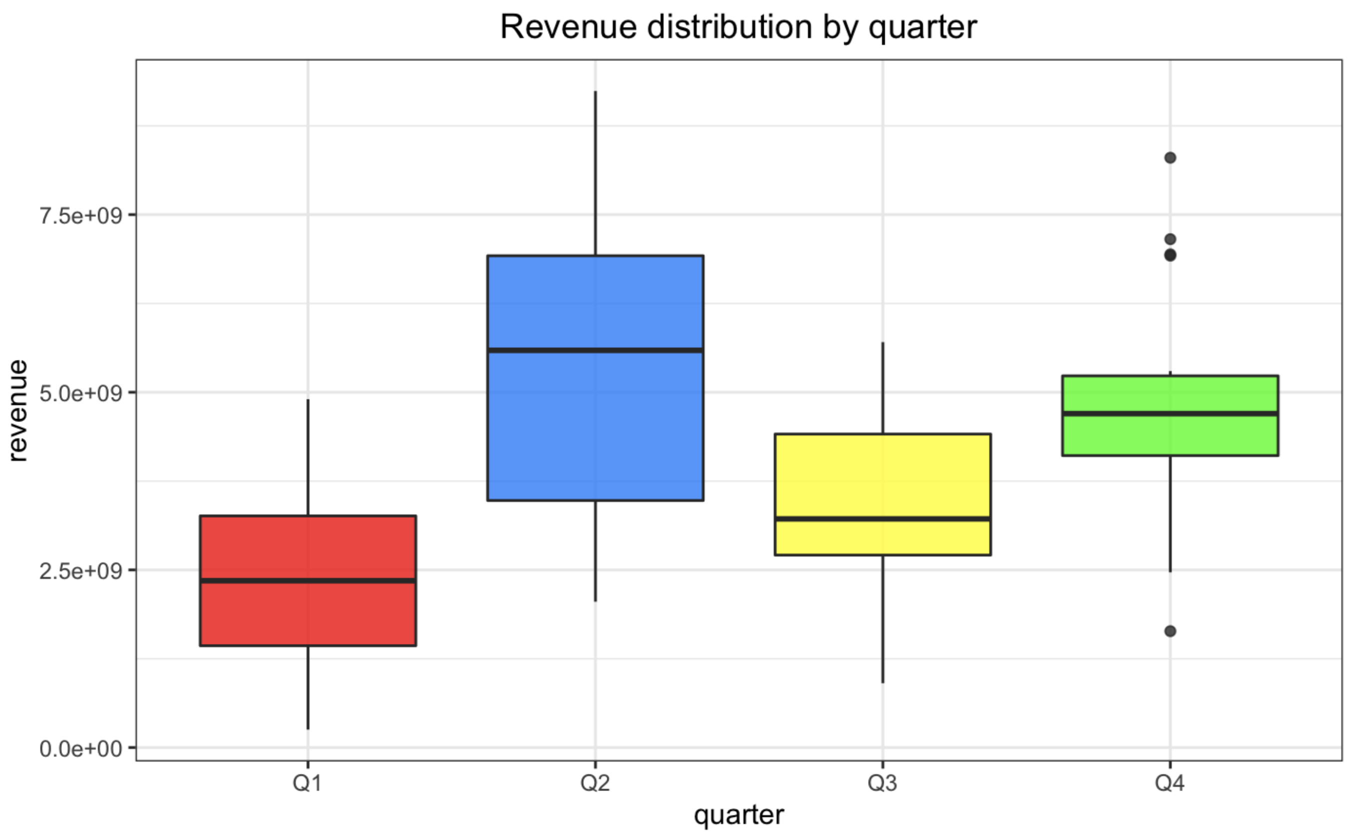
Confusion Matrix - KNN

actual / prediction	0	1
0	90 (True Negative)	67 (False Positive)
1	170 (False Negative)	722 (True Positive)

- accuracy = 77.4%
- kappa = 0.301

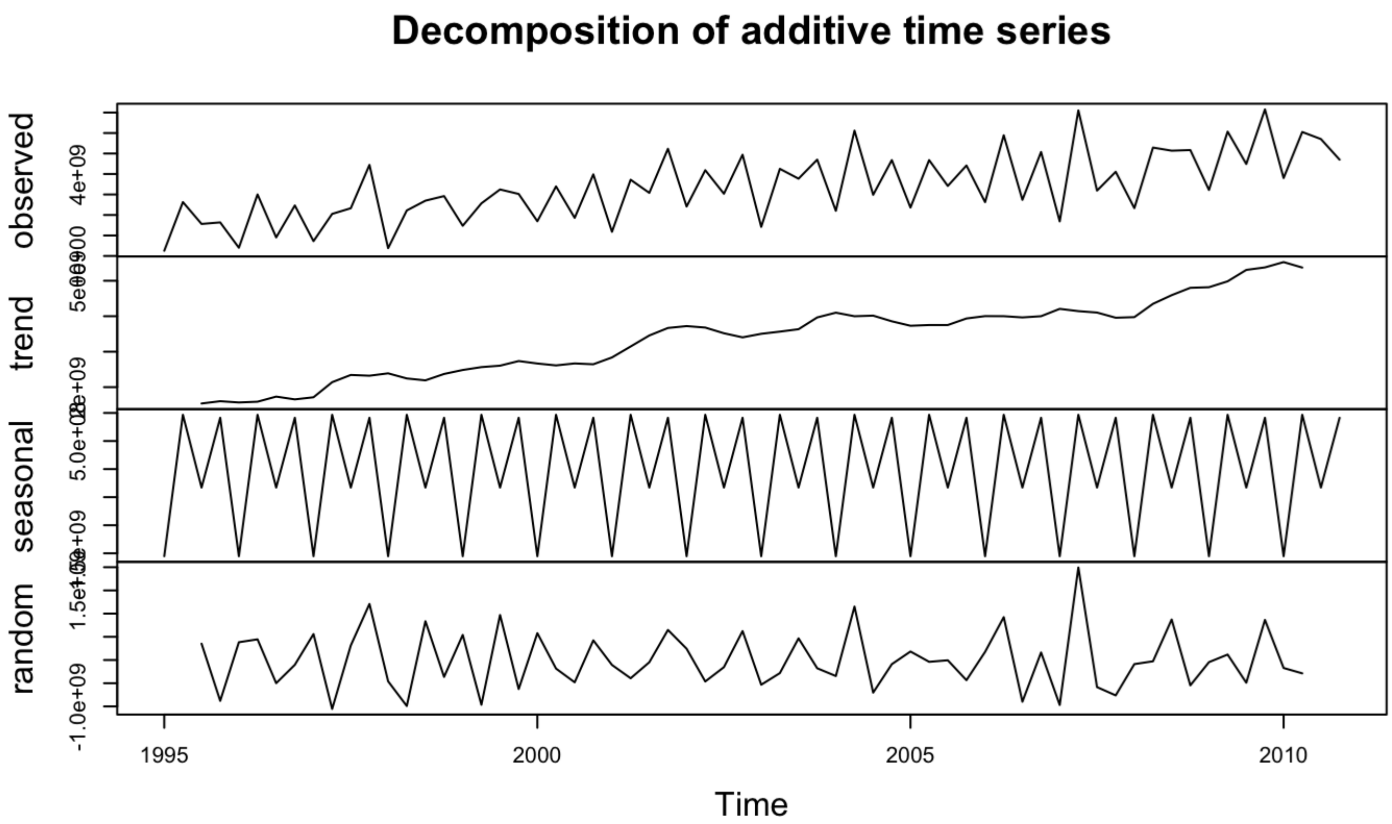
Seasonality Trends

Revenue Distribution by Quarter



THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

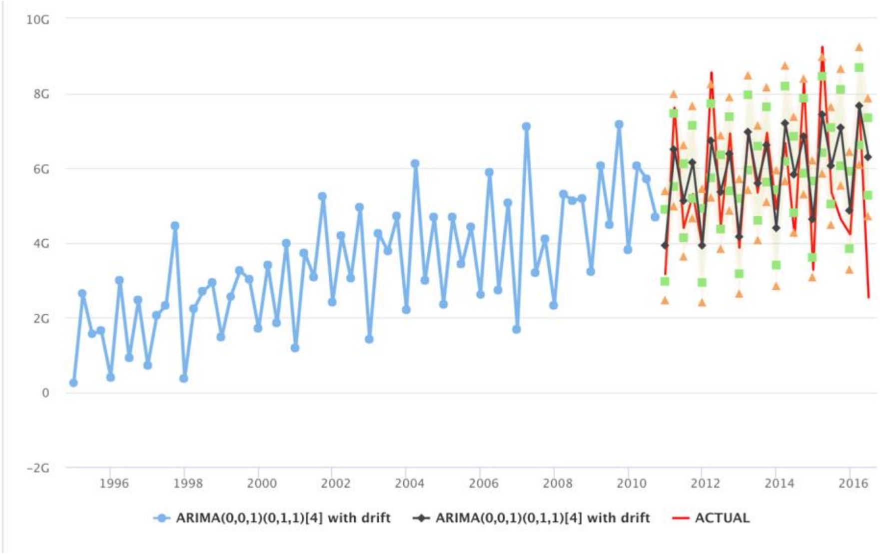
Time Series



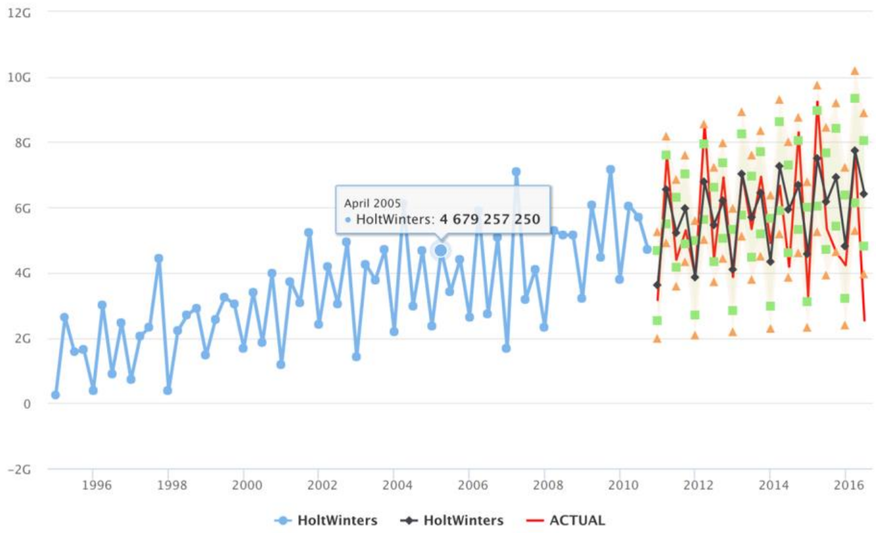
THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Time Series

ARIMA

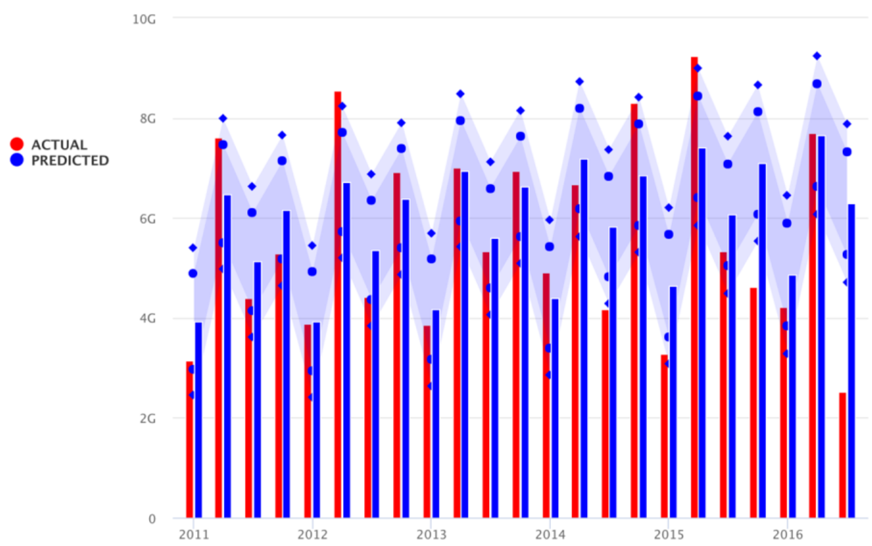


HoltWinters

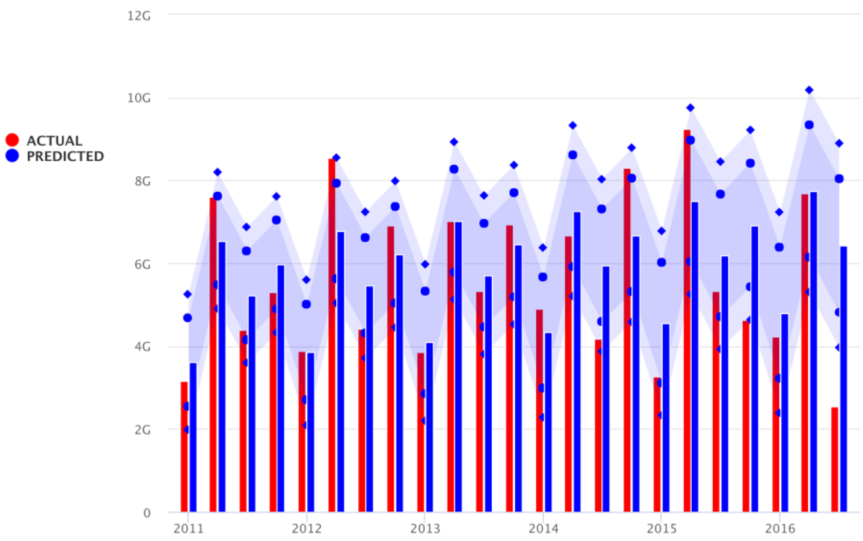


Detailed Visualization

ARIMA



HoltWinters



THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

Model Comparison

Model	RMSE		MAE	
	Train	Test	Train	Test
ARIMA	7.06e+08	1.31e+09	5.23e+08	9.88e+08
HoltWinters	8.33e+08	1.32e+09	6.51e+08	9.99e+08

- ➢ ARIMA performs better on the training data.
- ➢ However, there is no significance between two models when predicting the testing data.

Conclusion

- When predicting revenue and profit, Random Forest yields the best performance model.
- When predicting profitability, Logistic Regression yields the best performance model.
- The most important features that a movie studio/investor should consider for box office success are budget, vote and popularity.
- Movies released in April, May and June tend to generate larger revenues.
- Movie released in June tend to generate larger profit.

References

https://towardsdatascience.com/

https://stats.idre.ucla.edu/r/dae/

https://stackoverflow.com/

https://www.kaggle.com/learn-forum

https://www.r-bloggers.com/

https://www.datacamp.com/tracks/r-programming

https://www.forbes.com/sites/scottmendelson/2019/12/01/box-office-walt-disney-frozen-2-starring-idina-menzel-and-kristen-bell-breaks-thanksgiving-records-and-tops-738-million-worldwide/#2f2c0c956d80