**Exploring Movie Box Office Predictions:**

**An Analysis Using Principal Component Analysis & Traditional Regression Models**

Group 7: Gurjas Singh Chawla, Cat Kim, Khanh Huynh Bao Nguyen, Sriram Srinivasan

The Jindal School of Management, The University of Texas at Dallas

BUAN 6383/ MIS 6386: Modeling for Business Analytics

Professor Syam Menon, Ph.D.

September 8, 2024

Table of Contents

[Evaluating the Impact of “Buzz” Variables on Model Performance 3](#_Toc176641190)

[#7. Compare the models developed so far - which of these would you choose, and why? 3](#_Toc176641191)

[Examining the Effect of Principal Components on Model Predictions: 4](#_Toc176641192)

[#9. What can you say about this model vis-à-vis the other models so far? 4](#_Toc176641193)

[#10. Recommending Models Regression Models with Principal Components 5](#_Toc176641194)

[Evaluation of All Regression Models and Variables 7](#_Toc176641195)

[#12. Compare and recommend regression models using the number of principal components based on Kaiser's Rule and “explained variances” 7](#_Toc176641196)

[#13. Are the buzz variables or PCA helping to build a better model? 8](#_Toc176641197)

[Conclusion 10](#_Toc176641198)

[# 14. A) Did you learn anything surprising while doing these analyses? 10](#_Toc176641199)

[# 14. B) Managerial Takeaways & Decisions 10](#_Toc176641200)

# Evaluating the Impact of “Buzz” Variables on Model Performance

## #7. Compare the models developed so far - which of these would you choose, and why?

The model in Question 6 (linear regression using the significant variables of all independent variables, including “buzz” variables) would be recommended because:

* The adjusted R2 is the highest among 4 models from Q2-Q6.
  + R2 may not be the best measure because R2 increases as the number of variables used in the linear regression model increases.
  + Adjusted R2 accounts for the number of variables.
* This model also has the lowest AIC and BIC.
  + Like adjusted R2, AIC and BIC better reflect the model quality by having a penalty term for the number of variables.
* This model has a reasonable number of variables (5), and 4 out of 5 variables are statistically significant (Based on the t-statistics and associated probabilities (p > |t|)). Simpler models are easier to interpret and avoid overfitting.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Q | Model | R2 | Adjusted R2 | AIC / BIC | Significant Predictors at 0.10 Level\* |
| 2 | Traditional variables (without “buzz”) | 0.342 | 0.214 | 163.7 / 187.1 | log\_budget  sequel  horror |
| 3 | Only significant traditional variables | 0.291 | 0.254 | 154.4 / 162.9 | log\_budget  sequel  horror |
| 5 | All independent variables (including “buzz”) | 0.624 | 0.512 | 137.0 / 168.9 | PG action  animated  cntwait3  log\_addict  log\_fandango |
| 6 | Only significant independent variables (including “buzz”) | 0.558 | 0.519 | 129.0 / 141.8 | action  animated  cntwait3  log\_addict  log\_fandango |

Table 1: Comparison of regression models using traditional and all independent variables with/without significant variables (significance of .10 using p> |t|)

# Examining the Effect of Principal Components on Model Predictions:

## #9. What can you say about this model vis-à-vis the other models so far?

Of all 4 PCs, only PC1 is statistically significant.

* The model in Q9 uses “traditional” independent variables (excluding 4 “buzz” variables) + principal components (PCs) of the 4 “buzz” variables after applying PCA.
  + The other models in Q2-Q6 solely use the original variables. Q2 and Q3 use only the “traditional” variables, while Q5-Q6 use the “traditional” and “buzz” variables (not PCs).
* R2 / Adjusted R2: Sameas Q5 which has all variables (including the original “buzz” variables). Since each PC is a linear combination of the 4 “buzz” variables, using all 4 PCs altogether only rearranges the coefficient estimates on the 4 original “buzz” variables but does not change the model prediction and complexity.

Figure 1: Explained Variance Plot of 4 "Buzz" Variables

A graph with lines and numbers

Description automatically generated

## #10. Recommending Models Regression Models with Principal Components

Although PC1 is the most statistically significant and shows the same number of variables, the recommendation is choosing the model using the “traditional” variables and the first 2 PCs of the 4 “buzz” variables.

* Adjusted R2 is the highest among 4 models involving PCs of the “buzz” variables” and AIC is the lowest.
* The first 2 PCs explain ~80% of the variability of the “buzz” variables.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | R2 | Adjusted R2 | AIC / BIC | Significant Variables at 0.10 Level\* |
| Traditional variables + 4 PCs of “buzz” variables | 0.624 | 0.512 | 137.0 / 168.9 | PG action  animated  PC1 |
| Traditional variables + 3 PCs of “buzz” variables | 0.609 | 0.503 | 137.4 / 167.2 | PG action  animated  PC1 |
| Traditional variables + 2 PCs of “buzz” variables | 0.609 | 0.513 | 135.5 / 163.1 | PG action  animated  PC1 |
| Traditional variables + 1 PC of “buzz” variables | 0.589 | 0.498 | 136.6 / 162.1 | PG action  animated  PC1 |

Table 2: Comparison of the regression models with the PCs (including four components, from Q9)

\* Based on the t-statistics and associated probabilities (p > |t|)

Figure 2: Explained Variance Plot 4 "Buzz" Variables, Budget (Log-Transformed), and Star Power

A graph with a line and a blue line

Description automatically generated with medium confidence

# Evaluation of All Regression Models and Variables

## #12. Compare and recommend regression models using the number of principal components based on Kaiser's Rule and “explained variances”

Recommend the regression model using 3 PCs with the “buzz” variables + “starpowr” + “budget” based on Kaiser’s Rule and “explained variance” thresholds of 70%/80%.

* Coverage of 4 significant predictors at .10 level (p > |t|): “fandango” (log-transformed), “cntwait3”, PC1, PC3
* The model is in a “sweet spot” of dimension reduction and interpretability of too simple & not enough coverage (2 PCs) vs. too complex with too many predictors (4 PCs).

However, it would not be recommended to use any of these regression models due to signs of multicollinearity, indicating redundant variables. The R2 and the Adjusted R2 stay at .475 and .418 respectively and no changes in the AIC/BIC despite changing the number of PCs. Only the number of significant predictors at the .10 level changes, which doesn’t change the prediction or enhance the model quality.

|  |  |
| --- | --- |
| Method of Choosing ‘K’ | Recommended Number of PCs |
| Kaiser’s Rule | 2 |
| “Explained variance” threshold: 60% | 2 |
| “Explained variance” threshold: 70% | 3 |
| “Explained variance” threshold: 80% | 3 |
| “Explained variance” threshold: 90% | 4 |

Table 3: Number of principal components (PCs) selected using (i) Kaiser's Rule, and using “explained variance" thresholds of (ii) 60%, (iii) 70%, (iv) 80% and (v) 90% from “buzz” + “starpwr” + budget (log-transformed)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | R2 | Adjusted R2 | AIC / BIC | Significant Predictors at 0.10 Level\* |
| “Buzz” variables + “budget” + “starpowr” + 4 PCs from Q11 | 0.475 | 0.418 | 141.7 / 156.6 | log\_addict  log\_fandango  cntwait3  PC1  PC3  PC4 |
| “Buzz” variables + “budget” + “starpowr” + 3 PCs from Q11 | 0.475 | 0.418 | 141.7 / 156.6 | log\_fandango  cntwait3  PC1  PC3 |
| “Buzz” variables + “budget” + “starpowr” + 2 PCs from Q11 | 0.475 | 0.418 | 141.7 / 156.6 | cntwait3  PC1  PC2 |

Table 4: Comparison of the regression models of "buzz" and continuous variables with the PCs

\* Based on the t-statistics and associated probabilities (p > |t|)

## #13. Are the buzz variables or PCA helping to build a better model?

Both the “buzz” variables and PCA help build a more quality model and effective fit, resulting in higher Adjusted R2, lower AIC/BIC values, and interpretable quantity of indicators.

* Regressions using only the traditional variables (without “buzz” variables) have the lowest Adjusted R2 compared to models that include the “buzz” variables.
* Regressions using PCs and “buzz” variables consistently retain significant predictors of “PG”, “action”, “animated”. This reinforces the model’s interpretability and reliability while also highlighting effectiveness in dimensionality reduction while keeping high levels of variability.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Q | Model | R2 | Adjusted R2 | AIC / BIC | Significant Predictors at 0.10 Level\* |
| 2 | Traditional variables (without “buzz”) | 0.342 | 0.214 | 163.7 / 187.1 | log\_budget  sequel  horror |
| 3 | Only significant traditional variables | 0.291 | 0.254 | 154.4 / 162.9 | log\_budget  sequel  horror |
| 5 | All independent variables (including “buzz”) | 0.624 | 0.512 | 137.0 / 168.9 | PG action  animated  cntwait3  log\_addict  log\_fandango |
| 6 | Only significant independent variables (including “buzz”) | 0.558 | 0.519 | 129.0 / 141.8 | action  animated  cntwait3  log\_addict  log\_fandango |
| 9 | Traditional variables + 4 PCs of “buzz” variables | 0.624 | 0.512 | 137.0 / 168.9 | PG action  animated  PC1 |
| 10 | Traditional variables + 3 PCs of “buzz” variables | 0.609 | 0.503 | 137.4 / 167.2 | PG action  animated  PC1 |
| 10 | Traditional variables + 2 PCs of “buzz” variables | 0.609 | 0.513 | 135.5 / 163.1 | PG action  animated  PC1 |
| 10 | Traditional variables + 1 PC of “buzz” variables | 0.589 | 0.498 | 136.6 / 162.1 | PG action  animated  PC1 |
| 11 | “Buzz” variables + “budget” + “starpowr” + 4 PCs from Q11 | 0.475 | 0.418 | 141.7 / 156.6 | log\_addict  log\_fandango  cntwait3  PC1  PC3  PC4 |
| 11 | “Buzz” variables + “budget” + “starpowr” + 3 PCs from Q11 | 0.475 | 0.418 | 141.7 / 156.6 | log\_fandango  cntwait3  PC1  PC3 |
| 11 | “Buzz” variables + “budget” + “starpowr” + 2 PCs from Q11 | 0.475 | 0.418 | 141.7 / 156.6 | cntwait3  PC1  PC2 |

Table 5: Comparison of All Regression Models

\* Based on the t-statistics and associated probabilities (p > |t|)

# Conclusion

## # 14. A) Did you learn anything surprising while doing these analyses?

* Budget is something that is usually thought of as the reason for low box performance, but in all the models that include “buzz” variables it is not a significant predictor.
* If “buzz” variables were ignored, the significant variables are completely different vs. if “buzz” variables are included.

## # 14. B) Managerial Takeaways & Decisions

* “Buzz” variables play a crucial role in predicting the domestic opening box office revenues. By producing quality movie trailers and effective marketing strategies to drive public interest while staying up-to-date on the potential turnout using fandango.
* The model can suggest that movies with a PG MPAA rating and are action/animated are significant predictors of box office performance. Recommend conducting additional analysis to compare box office earnings using MPAA ratings and movie type to determine which movies will produce higher earnings.
* Future consideration: For more specific indicators, break down the movie types of action/animation into more specific categories. (I.e. Action: Superhero, adventure, martial arts, crime; Animation: family animation like *Frozen*, anime, fantasy, superhero, comedy)
* Future consideration: For better interpretation and focus on significant variables, run PCA on non-significant variables using all variables including “buzz”. Most of the models that we ran so far show the same variables and are not significant.