Predicting Box Office Success

Jake Weinberg, Aubrey Winger, Bella Samper

Problem Description

How can movie studios and executives determine if a film is going to be a smash hit or if it's destined to flop? Insight into that question would not only enable decision makers to better choose which movies to produce, but it would also inform the entire promotion cycle. Budgeting, marketing, and release decisions would be made more intelligently. We attempted to find a Bayesian answer to this question by crafting a regression model that would predict the box office revenue for a movie. The goal of this model is twofold. First, it can be used to help these studios determine what types of movies are successful, so that they can create more films with similar topics and generate more revenue. Second, this model will aid companies who specialize in production and distribution. By incorporating features related to the initial performances of movies at festivals and at critic viewings, the model can help companies determine whether they should purchase the rights and distribute a film to a more general audience.

Probability Model

While we included several conventional, accessible predictors in our regression model, such as budget, we knew that revenue models that only included these variables would be commonplace. To garner additional insight, we also extracted information from critic and audience reviews, as we suspected they contained underlying themes that would help determine a film's financial success. Leveraging Latent Dirichlet Allocation (LDA), we engineered ten additional predictors from latent topics that we derived from Rotten Tomatoes movie reviews. We hypothesized that a Bayesian regression model, where we could model predictor coefficients as random variables, combined with our latent topic predictors would effectively predict movie revenue. We used Gaussian priors for the features in the regression model, as well as the target revenue variable (Figure 1).

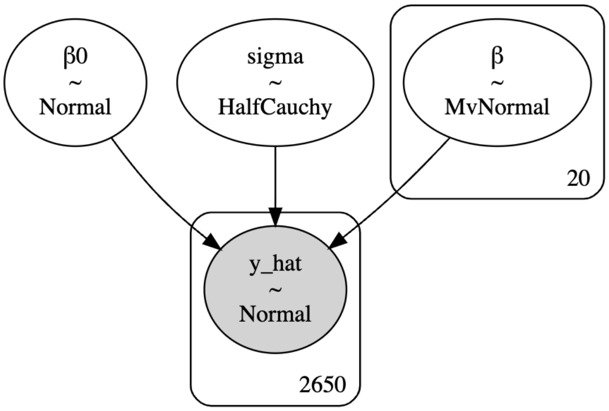


Figure 1. Graphical Visualization of Bayesian Linear Regression Model

Approach

We started with three different datasets: a catalog of over one million critic reviews from Rotten Tomatoes, another Rotten Tomatoes dataset containing additional information related to reviews, and a final table from Kaggle with movie financials – including revenue and budget. Because there were several reviews for each movie in the dataset, we needed to condense the review information into one row's worth of variables before we could join the tables. We used LDA to determine ten latent topics from the reviews (Appendix A). Then, we predicted the topics that would be most likely to be associated with each movie in our data. This was done by averaging the probability that a review fell into each topic for all reviews belonging to a given movie. For example, if Movie A had some reviews that had high probabilities for Topic 1 and other reviews that had high probabilities for Topic 2, the resulting overall score for Movie A would reflect both Topics 1 and 2 as more likely than the rest. The result of this process was ten new variables, one for each topic, that summed to 1 for each movie. These variables are read as the probability that a movie is associated with the given topic.

With our topic variables calculated, we could merge our three datasets so that we had one row for each movie. Missing data was imputed with the mean of each feature, and movies with zero values for revenue or budget were filtered out. We decided to remove all the review variables that had to do with general audience score, since the goal of our project is to help film executives make decisions before the movie is released to the general public. The final variables in our dataset included our response variable (revenue), our ten topic predictors, and ten other visible predictors.

The first visible predictor we decided on was tomatometer_rating, which is the critic rating of the movie on a scale from 0 to 100. Next, we included tomatometer_top_critics_count and tomatometer_fresh_critics_count, which are the number of top critics that reviewed the movie and the number of critics that gave the movie a fresh (positive) rating, respectively. We also included budget and movie runtime. The last five variables were dummy encodings of the content rating of the movie, which could be NR, G, PG, PG-13, R, or NC-17. NC-17 was removed so that these encoded variables were not collinear. These last five variables were binary, so if a movie was rated G, it would have a 1 in the G content rating column and a 0 in all the other columns.

The final equation for our model is as follows, where Y was the revenue, the first 10 predictors were the LDA hidden variables, and the second ten predictors were our visible variables:

$$Y = \beta_0 + \beta_{1...10}\, x + \beta_{11...N}\, x + \boldsymbol{\varepsilon}\ ,$$

In order to determine Baysian intervals of uncertainty, Variational Approximation (VA) was used to estimate the coefficients. This was initially chosen over sampling methods because our dataset is large, and VA scales better than sampling. Additionally, if this model was used in the future with a larger movie dataset, VA would be more efficient for the purposes of the client. An ELBO plot was generated, and it verified the strong convergence of VA (Appendix B).

Results

After running initial test models, we found that the performance of our model was very poor, with a loss from VA in the tens of thousands. Additionally, almost every single variable had an interval of uncertainty straddling zero, meaning they were insignificant in model predictions. Thinking about the data we had, this made sense. Movies in our dataset had an incredibly large range of revenue, and they were released over the span of several decades. To improve our model, we focused on only predicting movies in the middle range of revenue. We removed movies with revenue below the 25th percentile and above the 75th percentile. This reduced our sample size to 2,056 movies. Lastly, as our final transformation to our model, we took the log of revenue and of budget, which further reduced the variability in these variables. This improved the predictive power of our model, because log scaling normalizes the revenue and scales it to smaller values (Appendix C). This greatly reduced the model loss, improving its predictive performance. We also experimented with adding interaction terms to the model, but this did not result in a decrease in loss, and we did not find significant interaction terms, so we determined that it was better to have a simpler, more interpretable model.

The forest plot (Figure 2) from the full model shows that several of the coefficients for the predictors were significant. The topic predictors had wider levels of uncertainty than the visible predictors, with the first five visible predictors having extremely small intervals. Of these, only movie revenue was significant; logically, budget and revenue are likely connected, so this result made sense. However, it was surprising to see that the three variables connected to critic reviews did not have any impact on the model. We thought they would be a gauge of popularity and quality of the movie, but they were not. Five out of the ten LDA coefficients were significant in the model, showing that latent topics can be a credible addition. This partially proved our hypothesis, since it shows that generated topics can add predictive modeling power. The trace plot for the variables in the full model is shown in (Appendix D).

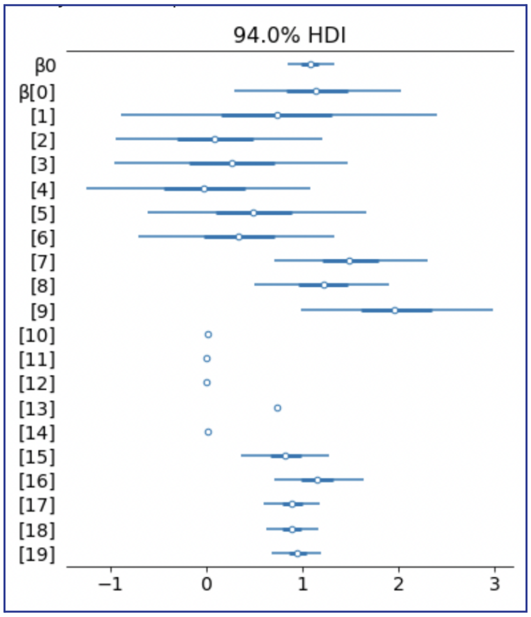


Figure 2. Forest Plot of Full Model

We also created a reduced version of our model to further examine the impacts of the LDA topic variables. The reduced model consisted of only the visible variables: tomatometer_rating, tomatometer_top_critics_count, tomatometer_fresh_critics_count, budget, runtime, and the dummy variables of content rating – G, PG, NR, PG-13, and R. The ELBO plot we created again showed strong convergence (Appendix E).

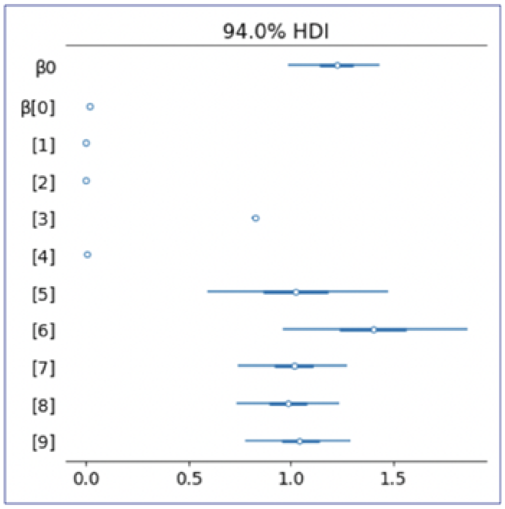


Figure 3. Forest Plot of Reduced Model

We also drew a forest plot of the reduced model's predictors to show uncertainty estimates (Figure 3). Similar to the full model, the first five predictors had very small credible intervals, while the content rating dummy variables had fairly wide measures of uncertainty. Again, only the budget out of the first five variables was not close to zero, indicating that it was the only variable with a significant effect on predicting revenue. The three review variables and runtime were exactly at zero in both models; we concluded that they would not help us in our solution. The other six predictors, however, did not bound zero and were significant in both models. Since budget has little to no variability, it can help inform us of our solution with less uncertainty than the other predictors.

After analyzing each one separately, we used Bayesian techniques to directly compare the models. We used WAIC and LOO to determine which model worked better in predicting revenue (Appendix F). We used deviance to measure both, therefore a higher WAIC and LOO estimate represents a better model. For both measures, the reduced model had a higher WAIC and LOO estimate but only by a narrow gap. This went against our original hypothesis that the full model would do better because of the LDA topics. However, the full model may have received a more favorable score because WAIC penalizes models with high complexity. We are comparing a model with 10 predictors to one with 20 predictors, so the full model was penalized more. Another possible reason for the reduced model scoring higher could have been that some of the LDA topics' confidence intervals cover zero. This would have resulted in a lower fit component in the WAIC.

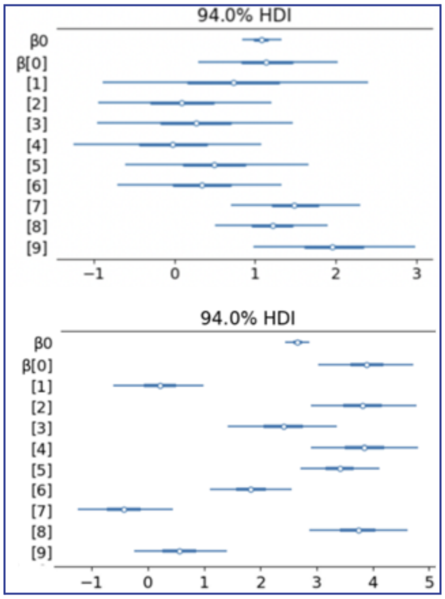


Figure 4. Separate Forest Plots of LDA Topics

Conclusions

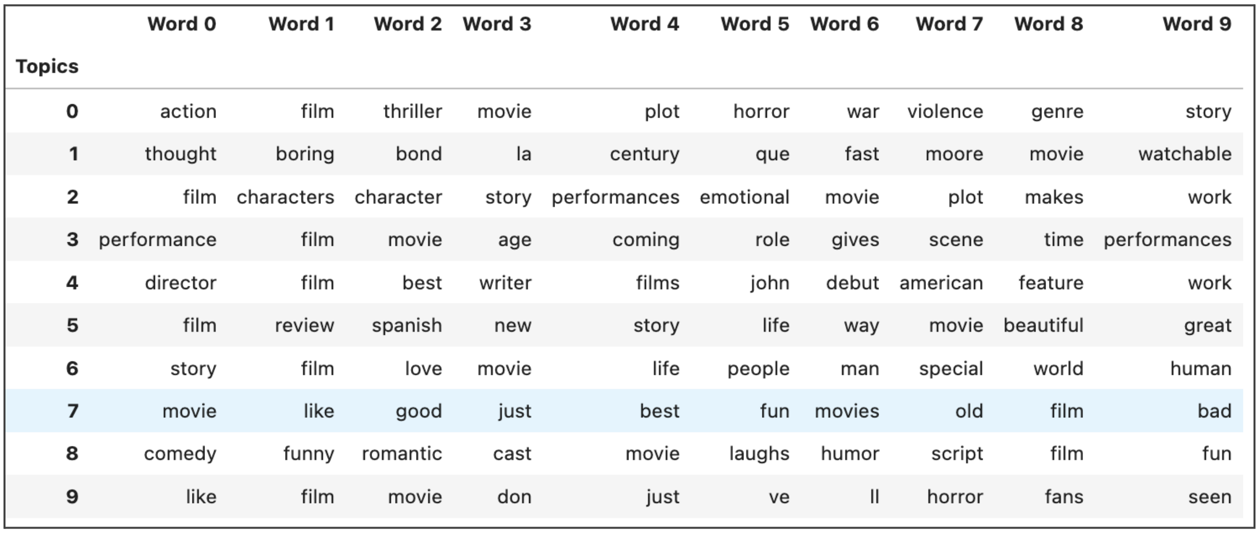
Our analysis led us to believe that LDA topics can be a helpful, new insight into predicting movie revenue before a movie is shown to the public. However, generating a reliable model that incorporates latent topics would require additional analysis. A limitation of our project was that every time LDA was run, the topics generated changed. Figure 4 shows the dramatically different forest plots for two runs of LDA. We would strive to expand the LDA component of our work to derive a stable and predictively powerful set of topics for our best model. Another limitation in our project was that out of our ten visible variables, five of them were dummy encoded (binary), which limits the information in our dataset. This model could be expanded by adding additional datasets into the analysis funnel to explore more possibilities. Despite these challenges, we were encouraged by the potential of our analysis and are confident that incorporating LDA into Bayesian regression is an insightful and powerful methodology for predicting box office success.

References

N, L. (2019, March 9). *IMDB dataset of 50K movie reviews*. Kaggle. Retrieved December 11, 2022, from https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews

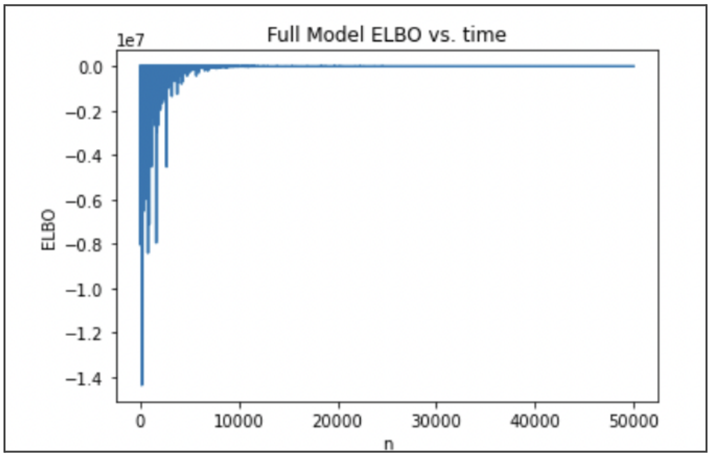
Leone, S. (2020, November 4). *Rotten tomatoes movies and critic Reviews Dataset*. Kaggle. Retrieved December 11, 2022, from https://www.kaggle.com/datasets/stefanoleone992/rotten-tomatoes-movies-and-critic-reviews-dataset

Appendix A. Topics generated by LDA analysis

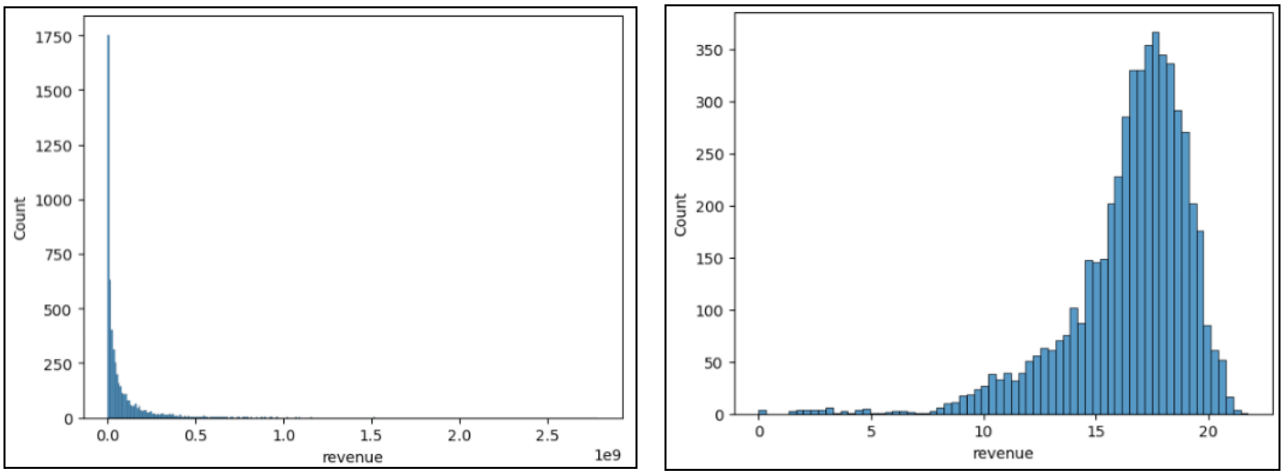


Topics	Word 0	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6	Word 7	Word 8	Word 9
0	action	film	thriller	movie	plot	horror	war	violence	genre	story
1	thought	boring	bond	la	century	que	fast	moore	movie	watchable
2	film	characters	character	story	performances	emotional	movie	plot	makes	work
3	performance	film	movie	age	coming	role	gives	scene	time	performances
4	director	film	best	writer	films	john	debut	american	feature	work
5	film	review	spanish	new	story	life	way	movie	beautiful	great
6	story	film	love	movie	life	people	man	special	world	human
7	movie	like	good	just	best	fun	movies	old	film	bad
8	comedy	funny	romantic	cast	movie	laughs	humor	script	film	fun
9	like	film	movie	don	just	ve	ll	horror	fans	seen

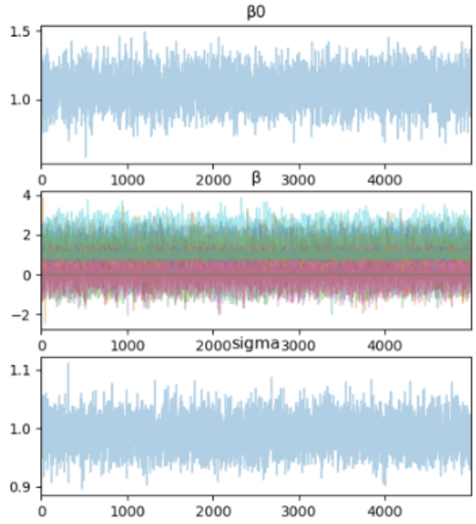
Appendix B. ELBO plot shows strong convergence of Variational Approximation for determining the Coefficients of the Full Bayesian Linear Regression model.



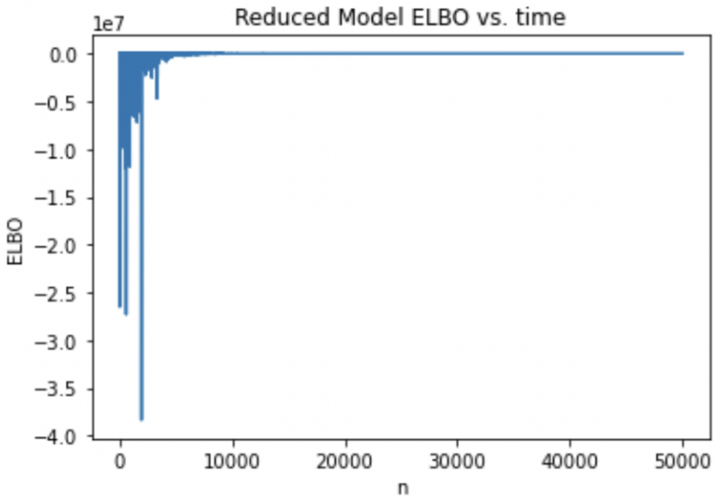
Appendix C. Revenue variable before and after log adjustment



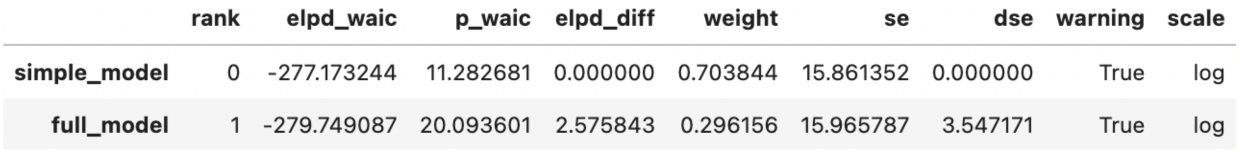
Appendix D. Trace plot for full model



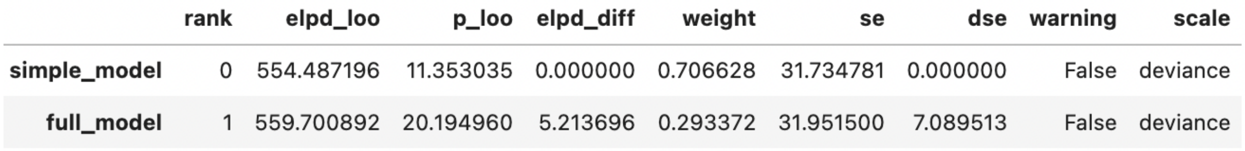
Appendix E. ELBO plot shows strong convergence for Variational Approximation for the Reduced Linear Regression model.



Appendix F. Results for WAIC and LOO Comparing the Full and Reduced Model



	rank	elpd_waic	p_waic	elpd_diff	weight	se	dse	warning	scale
simple_model	0	-277.173244	11.282681	0.000000	0.703844	15.861352	0.000000	True	log
full_model	1	-279.749087	20.093601	2.575843	0.296156	15.965787	3.547171	True	log



	rank	elpd_loo	p_loo	elpd_diff	weight	se	dse	warning	scale
simple_model	0	554.487196	11.353035	0.000000	0.706628	31.734781	0.000000	False	deviance
full_model	1	559.700892	20.194960	5.213696	0.293372	31.951500	7.089513	False	deviance