Topic: Do critic reviews have a causal effect on movie revenue?

Movie Data Ninjas

Deja Bond

Ravi Choudhary

Davida Kollmar

Tatenda Ndambakuwa

December 8, 2018

Introduction to Data Science

Business Understanding

Movies are an integral part of modern culture, generating over billions of dollars each year and delivering rich, intricately crafted stories to a worldwide audience. Movies are shared with these audiences by movie distribution companies. The movie distribution companies determine the number of copies of a film to provide to theaters, the best way to advertise and release a film such as theatrical release, television, home video, etc. The goal of the distributor is to maximize the revenue earned by the movie, which can be done by maximizing the audience size for the film and understanding the various ways the audience would like to receive the movies.

When deciding to watch a movie, many moviegoers read critic reviews to see whether it is worth their time. There are two main functions of critics' reviews. First, they help increase the awareness of movies. They can inform readers about which movies are being released, and thus function as advertisements. Second, critics' reviews also inform the public of the quality of movies. They praise the movies that they consider to be of high quality, while criticizing those they consider to be of a lower caliber. Overall, moviegoers can be encouraged or discouraged to watch a movie based on critic reviews.

The question of impact of critic revenue has been analyzed by many. Jehoshua Eliashberg and Steven M Shugan[1] ran a regression model on the number of positive critic ratings, number of negative critic ratings and the revenue of the movie per week including the overall revenue. They found a statistically significant correlation between the number of positive critic reviews and the overall revenue as well as number of negative critic reviews and the overall revenue. As we

[1] Eliashburg, Jehoshua, and Steven M. Shugan. "Film critics: Influencers or predictors?" *Journal of Marketing.* Vol 61,no. 2 (1997). https://faculty.wharton.upenn.edu/wp-content/uploads/2012/04/Film-Critics.pdf.

know that correlation does not imply causation, we want to further analyze if and by how much critic ratings can have causal effect on the movie revenue. Distribution companies can use this data to help their decision making. For example, if we determine that there is in fact a causation, then if a movie scores good reviews from critics in pre-screenings, the distribution companies can decide to screen the film in more theaters, whereas if the reviews are negative, they may choose to send the film to fewer theaters and focus more on the home release. If we find that there is little causation between critic reviews and movie revenue, then the distribution companies can choose to use other metrics when deciding how to distribute the movie, such as the budget of the film. In addition to analyzing the dataset as a whole to determine whether there is causation, we also examined whether budget or genre play a role in how much of a causal effect reviews have.

Data Understanding

Using the world film industry as our empirical setting, we examine the effects of reviews on movie revenue. We combined data from *The Movie Database (TMDB)*, a database for TV and movie data, and *The Numbers*, a website that specializes in movie business data, especially financial. Specifically, we used the TMDB 5000 dataset available on Kaggle because it was the largest movie dataset available. This dataset comes in two parts. The first has information such as title, genre, production house, and release date; and the second has information on the cast and crew. *The Numbers* dataset was used to obtain budget and revenue; this dataset had 5632 instances.

Metascore from *metacritic.com* was chosen as a single number to reflect all the critic reviews. We chose it because it reflects a normal distribution more closely than comparable

Tomatometer score from *Rotten Tomatoes*.[2] Also Metacritic claims that Metascore predict Box Office Performance.[3] We obtained the Metascore by scraping the Metacritic website, and added the Metascore as a feature on the TMDB dataset. We couldn't get metascore for movies with aliases like "Harry Potter and the Philosopher's Stone" and also movies with very common names like "Epic." This reduced our data from ~5000 to ~4000 instances.

Selection Bias: TMDB is a crowd-sourced website, therefore it mostly only has movies that many people are interested in. This might bias the movies to the ones which have been seen by many people. Also, we used the TMDB 5000 dataset directly, so we are not sure about the actual bias involved in selecting the movies. Our analysis was done on movies with metascore, therefore majority of Direct to Videos, TV movies and movies produced by some independent movie houses may not be included as they are not critically reviewed and don't have metascore.

Another consideration when analyzing our data, though not a selection bias, is that the movies in our dataset come from different years. Critic reviews may have different levels of influence in different eras, and inflation would affect the revenue of films.

Data Preparation

Once we had gathered our different datasets, we merged them into one dataset. We combined the two *TMDB* datasets based on TMDB's "movie id," and the *TMDB* and *The Numbers* datasets based on title. We removed rows where the release year from the *TMDB* and *The Numbers* datasets was more than five years apart, since this indicated to us that we had two

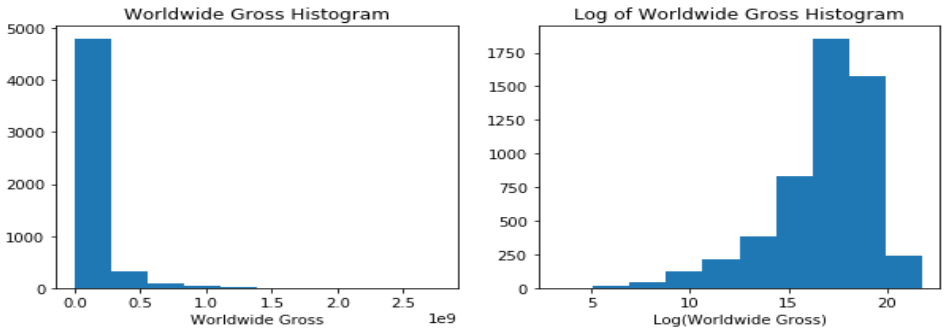
[2] Olteanu, Alex. "Whose ratings should you trust? IMDB, Rotten Tomatoes, Metacritic, or Fandango?" freeCodeCamp. Apr 10, 2017. Accessed Dec 7, 2018. https://medium.freecodecamp.org/whose-reviews-should-you-trust-imdb-rotten-tomatoes-metacritic-or-fandango-7d1010c6cf19

[3] Dietz, Jason. "Can Metascores Predict Box Office Performance?" Metacritic. Apr 12, 2016. Accessed December 7, 2018. https://www.metacritic.com/feature/film-quality-vs-box-office-grosses.

different movies with the same name. We kept the instance if the release years were less than five years apart, because sometimes the discrepancy was based on the theatrical release of a movie vs. its video release, but was the same film. We also removed rows with no revenue data. Our final dataset had 3316 instances.

After finalizing our dataset, we were ready for feature engineering. We deleted features which we thought would be irrelevant to our research, such as the movies' homepages. For the categorical features, such as production house, genre, and cast/crew, we used one hot encoding to indicate which production house created the film.

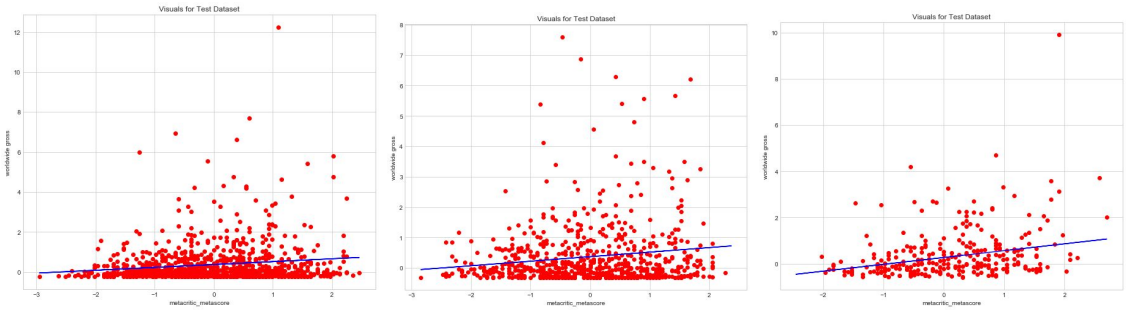
Once we had created our dataset, we did some exploratory analysis. First, we made a histogram of the revenue data (the column labeled "Worldwide Gross"). From the histogram, we could see that our data was skewed, so we decided to add a column to the dataset with the log transformation of Worldwide Gross as a target variable in order to reduce the order of magnitude of the target variable. Therefore in the current form, the target variable is the scale of the revenue of the movie. We will further refer to this target variable as Order of Magnitude of revenue or scale of revenue.



We did not remove any outliers from the data since we believe that the distribution is natural and removing outliers might bias our analysis.

Modeling & Evaluation:

We did an exploratory analysis to see if there was a correlation between Metascore and log of the revenue. First, we calculated Pearson's Correlation Coefficient, getting a value of 0.07687, which shows a positive correlation between the two. We then graphed the relationship between "log Worldwide Gross" and Metascore using sklearn's Linear Regression algorithm. We repeated this analysis using only small budget i.e., movies with budgets less than or equal to $40 million, and using only large budget i.e., movies with budgets greater than $40 million. $40 million was the median budget in our data.



Linear regression results: worldwide gross revenue vs metacritic metascore whole movie dataset. (left to right: whole dataset, small budget, big budget)

	All Movies	Big Budget Movies	Small Budget Movies
Training score, R^2	0.01143	0.07940	0.007104
Testing score, R^2	0.01991	0.1003	0.02175
Linear Regression RMSE	0.9691	1.0552	0.9842

We obtained low training and testing scores (R^2) for big budget movies, small budget movies and the combined budget movie dataset. From the graphs obtained, however, we do see a positive slope, indicating that at least to some extent higher Metacritic scores indicate higher revenue. Next we will move to the Causal Analysis.

Causal effect is usually measured through the random experiments where a sample is randomly allotted to a treatment or control group and the average treatment effect is calculated based on difference in test and control groups. However, our problem is not this simple. In this project, our problem is an observational one - we have not randomly chosen a control group and a treatment group for the positive and negative reviews, but rather we have observed movies which have already received reviews. A further challenge is that the assignment of reviews is not randomized, but rather is based on confounding factors such as genre, budget, actors, etc. This means that the average treatment effect of good reviews on movie revenue cannot be calculated directly by subtracting the expected value for revenue of movies with bad reviews from the expected value for revenue of movies with good with bad reviews. Instead, we need to approximate the average treatment effect by artificially creating a control, such that we can compare two movies with similar characteristics but with different reviews, so we can see how the reviews influence the revenue.

In general, causal analysis on observational data requires two main assumptions. The first is strong ignorability assumption, i.e., the choice of assignment to treatment or control group can be assumed to be effectively random when the data is conditioned on the confounders/reasons for assignment. The second assumption is overlap, i.e., some data with similar confounders are present in both the treatment group and the control group. Our dataset has many confounders, both measured and unmeasured. We believe measured confounders include the production house, the budget of the movie, the lead cast, the director, the writer and the genre. These measured confounders in turn can also act as latent variables for some unmeasured confounding variables, such as the script of the movie and the quality of its visual effects. There are a few

ways to handle causality inference on observational data, as discussed by David Sontag and Uri Shalit in their ICML tutorial[4]. We have identified propensity score matching as a potential way to check for statistical significance of the effect of different types of review (positive, mixed and negative) on the movie revenue. We chose this method because it is easy to interpret and implement. In order to implement the propensity score matching, we first converted Metascore in our dataset to a classification label indicating positive, negative, or neutral review, based on Metacritic's classification of which numeric scores fall into each category[5]. We then divided the data into a training and test set. Using two categories at a time, the training set was used to model for propensity, i.e., probability of a movie to get good reviews, based on the confounding factors we had identified. We have chosen logistic regression for the probability score prediction as it can work very well on very small datasets. Also Logistic Regression is the most preferred propensity score modelling technique[6]. The parameter selection for Logistic regression was done based on Log-Loss since the propensity score is supposed to be an estimate of the true probability distribution. We then used 1-Nearest-Neighbor on test set to match movies with similar propensity scores where one was labeled positive and one was labeled negative. By doing this, we could then compare their revenues, to estimate what the positively-reviewed movie would have earned if it had negative reviews, and what the negatively-reviewed movie would have earned if it had positive reviews. We chose 1-nearest neighbor based on propensity score, as suggested in the ICML tutorial and in "Matching methods for causal inference: A review and

[4] Shalit, Uri, and David Sontag. "Causal inference for observational studies." NYU Computer Science. June 2016. https://cs.nyu.edu/~shalit/slides.pdf.
[5] "How We Create the Metascore Magic." Metacritic. https://www.metacritic.com/about-metascores.
[6] Weitzen S, Lapane KL, Toledano AY, Hume AL, Mor V."Principles for modeling propensity scores in medical research." *Pharmacoepidemiol Drug Safety*. Vol. 13 no. 12 (2004). pp. 841-53. https://www.ncbi.nlm.nih.gov/pubmed/15386709.

a look forward" by Elizabeth A Stuart[7]. Paired T-test was done to test for statistical significance of difference in revenue due to different reviews as suggested by Peter C Austin[8]. The results based on T-test are mentioned in Appendix. The T-tests were only done on groups with at least 30 samples to get stable results. T-test results show that there is statistically significant average difference in scale of movie revenues when they are assigned positive reviews vs any other type of reviews.

The 1-nearest neighbor algorithm can bring a lot of variability in the model, and Propensity Score Matching as a technique requires that the probability estimate be very accurate for the estimate to be unbiased. Therefore we decide to use a doubly robust estimator[9] over Propensity Score Matching to approximate the average effect of critic reviews on movie revenues. Doubly Robust Estimator relies on 2 different models. The first model is the propensity score model, and the second model is a regression model to estimate the revenue of the movie conditioned on one type of reviews and confounders. The 2 models are used together to estimate the difference in revenue due to positive and negative reviews through the following formula:

$$\begin{aligned}\widehat{\Delta}_{DR} &= n^{-1}\sum_{i=1}^{n}\left[\frac{Z_i Y_i}{e(\boldsymbol{X}_i,\widehat{\boldsymbol{\beta}})} - \frac{\{Z_i - e(\boldsymbol{X}_i,\widehat{\boldsymbol{\beta}})\}}{e(\boldsymbol{X}_i,\widehat{\boldsymbol{\beta}})} m_1(\boldsymbol{X}_i,\widehat{\boldsymbol{\alpha}}_1)\right] \\ &\quad -n^{-1}\sum_{i=1}^{n}\left[\frac{(1-Z_i)Y_i}{1-e(\boldsymbol{X}_i,\widehat{\boldsymbol{\beta}})} + \frac{\{Z_i - e(\boldsymbol{X}_i,\widehat{\boldsymbol{\beta}})\}}{1-e(\boldsymbol{X}_i,\widehat{\boldsymbol{\beta}})} m_0(\boldsymbol{X}_i,\widehat{\boldsymbol{\alpha}}_0)\right]\end{aligned}$$

[7] Stuart, Elizabeth. "Matching methods for causal inference: A review and a Look Forward." Stat Sci. Vol 25 no. 1 (2010). pp. 1-21. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2943670/.

[8] Austin, Peter C. "An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies." *Multivariate Behavioral Research.* Vol. 46, no. 3 (2011). Pp. 399-424. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3144483/.

[9] Davidian, Marie. "Double Robustness in Estimation of Causal Treatment Effects." NCSU Statistics. 2007. https://www4.stat.ncsu.edu/~davidian/double.pdf.

$\hat{\Delta}_{DR}$ is the doubly robust estimate of the average difference in revenue due to positive vs negative reviews

n is total number of movies with positive and negative reviews

Z_i is the a binary variable indicating the type of reviews for the i^{th} movie. For comparing positive and negative reviews, it is set to 1 for positive reviews and 0 for negative reviews and

Y_i is the actual revenue of the i^{th} movie.

$e\left(X_i, \hat{\beta}\right)$ is the propensity score of i^{th} movie - probability of i^{th} movie to get positive reviews

$m_1\left(X_i, \widehat{\alpha_1}\right)$ is the Expected revenue of the i^{th} movie if it receives positive reviews

$m_0\left(X_i, \widehat{\alpha_0}\right)$ is the Expected revenue of the i^{th} movie if it receives negative reviews

The benefit of doubly robust estimator is that even if one of the regression or the classification model is incorrect, it gives an unbiased estimate[10]. This removes the strict condition on the reliability of the estimate from the propensity score model. This model can help us estimate the difference in revenue for the movie if it has negative, mixed or positive reviews.

The training data was used to model 3 separate revenue prediction models for movies with positive reviews, movies with mixed reviews and movies with negative reviews. We had a problem of underdetermined system as we had 13K features and 1k training samples for each type of review, so we decided to use regularised linear regression instead of just the linear regression to avoid overfitting. We tried Lasso Regression, ElasticNet Regression and Ridge Regression with mean square error loss. With respect to scale of movie revenues, the prediction of few millions as revenue for a billion dollar movie is more accurate than prediction of few

[10] Funk, Michele Jonsson, Daniel Westreich, Chris Wiesen, Til Stürmer, M. Alan Brookhart, and Marie Davidian. "Doubly Robust Estimation of Causal Effects." American Journal of Epidemiology." Vol 173 no. 7 (2011). pp. 761-767. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3070495/.

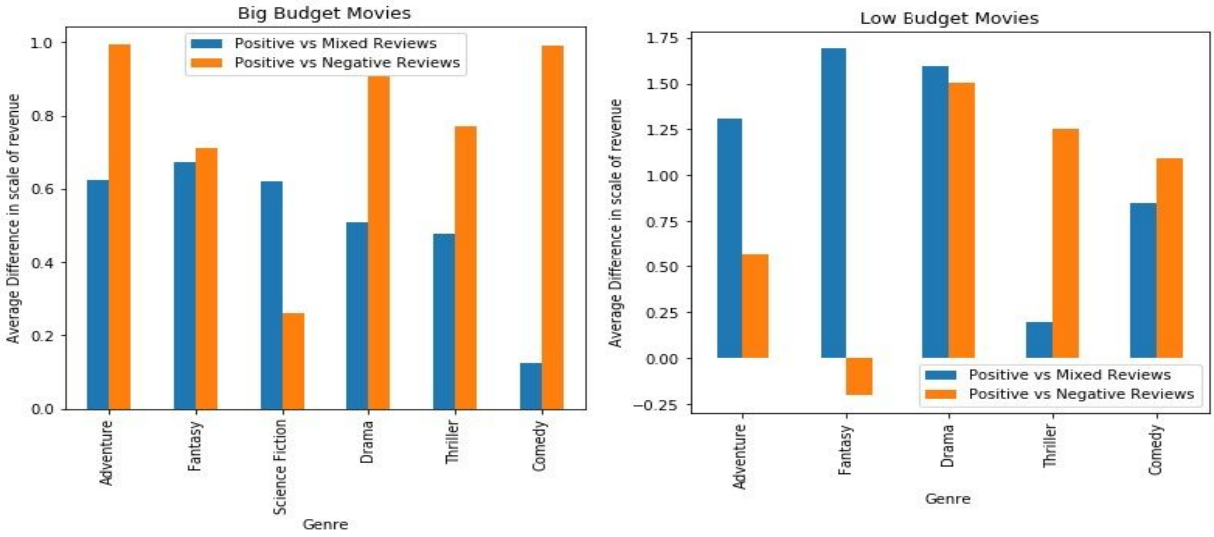
hundreds as revenue for the same movie. We chose MSE as loss function because it is very useful in this regard as it penalizes large differences much more than small differences. Hyper-parameter tuning was done using grid search and 5-fold cross-validation using MSE as the scoring metric. The MSE results of each model based on the 5 fold cross validation for the positive reviews only is given below. Similar results were received on other types of reviews. Ridge regression was selected based on the metric.

Model	**5-Fold Cross-Validation MSE**
Lasso Regression	5.47
ElasticNet Regression	3.65
Ridge Regression	2.19
Decision Tree Regression	2.87
Random Forest Regression	2.71

We decided to try Decision Tree regressor to improve our the results from Ridge Regression. We chose a decision tree based models as it is more robust to outliers than Linear Regression, and the revenue data has a lot of outliers. Also, Decision Tree based models can capture unspecified nonlinear interaction between variables. Random Forest regressor was also chosen because it has another advantage over decision trees of being more robust to overfitting. Parameter tuning for Random Forest was done through RandomizedSearchCV.

Both the decision tree and the Random Forest models failed to improve over the baseline Ridge regression model. Once we determined that Ridge regression was the best model, we then divided the dataset based on genre and on budget and generated the doubly robust estimate for

the expected value of difference in scale of revenue due to positive reviews compared to mixed and negative reviews. We have shown the results in a bar graph.



The results show a difference of almost one order of magnitude in revenue of a big budget movie if it gets positive vs negative reviews. The low budget movies show a even higher difference in order of magnitude of revenue. The difference in revenue due to positive vs negative reviews of low budget fantasy movie can be ignored as it is statistically insignificant by the t-tests.

Deployment:

This model will not need to be deployed for a real time prediction but can be used to identify and assess the impact on the revenue of a movie that potential reviews can have. Moreover the propensity score model can be used to identify the type of reviews with most likelihood for the movie. The Distribution Company can use the estimates and the most likely reviews to optimize the marketing budget and the number of theaters the movie will be released in. The company would also need to periodically update the model through updation of training

set to take care of any concept drift i.e., change in average effect of reviews on revenue to make sure that the estimates are correctly aligned. Also any unmeasured confounder at present that can be measured in the future, should also be included in the model for better estimates.

Considerations:

The analysis was done on movies with metascore therefore it limits our inference to only the movies with metascore. There might be some bias involved due to selection of Metascore as a single score representing critic reviews, as metascore is a weighted representation of each critic's reviews[6] and the weights are not publicly disclosed.

Ethical risks:

If we know that critic ratings affect movie revenue, movie producers might start paying critics to influence them to give better reviews. Movie producers could also pay critics to target their competitor's movies on the same release schedule. Independent production houses might suffer very easily as they don't have same level of monetary influence as big budget production houses.

Bibliography:

1. Austin, Peter C. "An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies." *Multivariate Behavioral Research.* Vol. 46, no. 3 (2011). pp. 399-424. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3144483/.
2. Davidian, Marie. "Double Robustness in Estimation of Causal Treatment Effects." NCSU Statistics. 2007. https://www4.stat.ncsu.edu/~davidian/double.pdf.
3. Dietz, Jason. "Can Metascores Predict Box Office Performance?" Metacritic. Apr 12, 2016. https://www.metacritic.com/feature/film-quality-vs-box-office-grosses.
4. Eliashburg, Jehoshua, and Steven M. Shugan. "Film critics: Influencers or predictors?" *Journal of Marketing.* Vol 61,no. 2 (1997). https://faculty.wharton.upenn.edu/wp-content/uploads/2012/04/Film-Critics.pdf.
5. Funk, Michele Jonsson, Daniel Westreich, Chris Wiesen, Til Stürmer, M. Alan Brookhart, and Marie Davidian. "Doubly Robust Estimation of Causal Effects." American Journal of Epidemiology." Vol 173 no. 7 (2011). pp. 761-767. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3070495/.

6. "How We Create the Metascore Magic." Metacritic. https://www.metacritic.com/about-metascores.
7. Olteanu, Alex. "Whose ratings should you trust? IMDB, Rotten Tomatoes, Metacritic, or Fandango?" freeCodeCamp. Apr 10, 2017. https://medium.freecodecamp.org/whose-reviews-should-you-trust-imdb-rotten-tomatoes-metacritic-or-fandango-7d1010c6cf19
8. Shalit, Uri, and David Sontag. "Causal inference for observational studies." NYU Computer Science. June 2016. https://cs.nyu.edu/~shalit/slides.pdf.
9. Stuart, Elizabeth. "Matching methods for causal inference: A review and a Look Forward." Stat Sci. Vol 25 no. 1 (2010). pp. 1-21. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2943670/.
10. Weitzen S, Lapane KL, Toledano AY, Hume AL, Mor V."Principles for modeling propensity scores in medical research." *Pharmacoepidemiol Drug Safety*. Vol. 13 no. 12 (2004). Pp. 841-53. https://www.ncbi.nlm.nih.gov/pubmed/15386709.

Appendix:

Contributions:

We all contributed to the data collection, data clean-up, and the write-up.

Deja worked on the reduction of the data and analysis for the Random Forest Regressor and its optimization with the RandomSearchCV.

Ravi wrote the code for scraping metacritic.com for metascore, one-hot encoding of categorical variables, propensity score matching and doubly robust estimator.

Davida wrote the code to merge the datasets. She also helped debug the causal inference code and did the analysis for linear regression and causal inference by genre.

Tatenda started working on removing 'outliers' in the data. She also worked on linear regression on the overall data and divided movie budgets into small budget and large budget movies to perform the linear regressions. She also explored working on GBDT but came across challenges like the runtime due to large number of variables, thus the idea was dropped.

Results from Paired T-tests from Propensity Score Matching

P values for hypothesis: Difference in revenue of high budget movies is zero	Genre						
	All Genre	Adventure	Fantasy	Sci-Fi	Thriller	Comedy	Drama
Positive and Mixed Reviews	0.05	10^{-4}	0.05	0.84	0.05	0.001	0.008
Negative and Mixed Reviews	0.13	0.42	0.06	0.13	10^{-5}	0.16	0.04
Positive and Negative Reviews	0.03	0.005	0.002	0.001	0.01	0.002	0.002

P values for hypothesis: Difference in revenue of high budget movies is zero	Genre						
	All Genre	Adventure	Fantasy	Sci-Fi	Thriller	Comedy	Drama
Positive and Mixed Reviews	0.0008	0.02	0.08	0.10	0.28	0.005	10^{-5}
Negative and Mixed Reviews	0.68	0.48	0.43	0.30	0.30	0.85	0.55
Positive and Negative Reviews	10^{-9}	0.99	0.42	0.06	0.49	10^{-6}	10^{-6}

All of the codes used in our paper can be found in our Github repository.