University of Toronto
Faculty of Applied Science and Engineering
MIE368 - Analytics in Action
Final Report - Team 17

Name	Student Number	Email
H.E.	—	—
I.C.	—	—
J.L.	—	—
A.K.	—	—

Introduction

The film industry involves substantial financial investments such as production costs and marketing expenses. Annually, billions are invested in the film industry by corporations. For example, between 2015 and 2021, it was estimated that over $64 billion was spent on expenses [1]. With substantial financial commitments, it is imperative to implement data-driven solutions for the meticulous management of expenses. Employing such solutions not only enhances operational efficiency but also ensures a strategic approach to financial control. Thus, distributors, studios, and other stakeholders must be aware of what factors are most relevant for determining whether a proposed movie concept would be successful before going forward with investing in production. Predicting the success of movies is often difficult due to being affected by a multitude of factors from different sources, from more impacting internal factors such as writing and screenplay to less conspicuous external ones like competitors, and even the weather.

With this rationale acting as a spur, the aim is to establish a robust framework that benefits stakeholders in the film industry through means that can be controlled. The cast and directors have such a direct influence on the revenue since movies with a well-known cast and director tend to help guarantee financial investments and production development, contributing to a greater box office revenue. Taking into account the more impactful controllable factors and showcasing the effect of the cast and director on the success of a movie, the current iteration of the model aims to answer a variation of the question: “Given a set of attributes for an upcoming movie, what is the box office produced for the best combination of leading actors and the most suitable director for a movie fitting the given attributes?”

Thus, the objective of the model is to determine the optimal set of leading actors and directors through calculated utility values for each representing their star power and bankability, thereby enabling the prediction of box office revenue across various categories in motion pictures. After implementation, the next aim is to find any insights that may indicate whether the model is strong and reliable for usage, having valuable and non-trivial results, and whether it is worthwhile for stakeholders to invest time and money into.

Data

All relevant data has been extracted from the comprehensive film database "The Numbers." The Python web scraping library BeautifulSoup was heavily employed to retrieve information from selected web pages, in conjunction with the pandas and NumPy libraries to clean and process the results. As exploratory data analysis and to form the final datasets, the three main entities of actors, directors, and movies define the bulk of the information to act as the input to the proposed algorithms. They are classified into the following categories as shown in Table 1, each having their own respective types (see Appendix A):

Genre	Rating	Creative Type	Distributor	Production Method	Source Material

Table 1. Categories taken into account for actors, directors, and movies.

Movies have been scrapped from the years 1995 to 2023 in the domestic market to prevent scope creep. For optimization, tables have been scraped for actors and directors in all of the mentioned categories containing the columns of domestic box office earnings, the number of movies participated in, and the average box office earnings per movie. Through these, a column of utility values was calculated (see Appendix B). In addition, these columns were utilized to approximate salaries, which were calculated for each actor and director to function as constraints on the model. A crew utility value column was added to the final movie dataset to represent an aggregate of the actor and director values. Thus, with all data cleaning finalized, the final dataset resulted in containing 2566 samples (see Appendix C). This final dataset was then partitioned from the years 1995 to 2020 and 2021 to 2023 to form the training and testing sets respectively.

Methods

The analysis of the data collected will be conducted through the two main methods of optimization and linear regression, which subsequently employ different techniques to arrive at solutions. This model implements the approach of optimizing and then predicting in which method stacking is utilized, inputting the result of the optimization into the linear regression.

Due to the model being centralized around the financial aspects of the film industry and profitability, the utility values were calculated to be a measure of how bankable an actor or director is, based on their career earnings. Given that the top-grossing stars earn orders of magnitude more than the lowest-tiered individuals, there would be a large variance and disparity in the utilities of the upper and lower percentiles. Thus to mitigate this, the utility values were scaled logarithmically, making the distribution more compact and simpler for analysis. Additionally, to account for high-grossing individuals whose earnings mainly came from a few hit movies but were relatively low standing previously or early in their career, the number of overall movies participated in was used to measure their experience in the industry. The degree to which experience factors into the utility value is higher than raw bankability since actors and directors with more movies tend to be more consistent and generally have higher total earnings over their careers. Thus with this rationale, the formula for the value v_i for an actor or director i is expressed as follows:

$$v_i = ln(a_i \cdot n_i^{\lambda}),$$

Where a_i, n_i, and λ are the average domestic box office earnings per movie representing bankability, the number of movies representing experience, and the order of correlation for experience, respectively. A value for λ too high or too low would provide heavily skewed results and as such through tuning the parameter and analyzing the sensitivity until the quantities reflected that which is observed in the film industry, the order of correlation has been set to $\lambda = 2$ for the purposes of this model (see Appendix D).

Just as the utility values are a core element of the objective function, the salaries of the actors and directors impose key constraints. These earnings are derived directly from the budget of the film, allocating a portion of it to the salaries to be paid. However, depending on the status and popularity of the individual, the majority of their salary would come from residual earnings, being a portion of the domestic box office earnings rather than the budget of the film [2]. Thus, the total earnings from a movie would be the sum of the percentage cut received from the budget as well as the residuals. Via this reasoning, the approximate salary s_{ij} of actor/director i for movie j can be expressed through the following:

$$s_{ij} = B_j \cdot \alpha_i + G_j \cdot \gamma_i,$$

Where B_j and G_j denote the budget and gross domestic earnings of movie j, whereas α_i and γ_i indicate percentage cut received from the budget and gross domestic earnings for

actor/director i, respectively. As s_{ij} is the salary for each movie, grouping the salaries by name and then averaging the rows resulted in the mean salary per movie s_i of actor/director i.

With the variance of actors and directors being high on each end of the spectrum in terms of status and popularity, the percentage cut they receive from the budget and gross domestic earnings also vary significantly. In the film industry, A-List actors and directors tend to get a relatively large portion of the budget and a cut from the gross domestic earnings as residuals while lower-tiered individuals get less of a cut from the budget and very little to no residuals at all [2][4]. Through this logic, a distribution of salary cuts can be defined as follows in Table 2, where the tiers are partitioned correspondingly on ranges of utility values:

Status	Utility Value Range	Percent Cut
A-List	≥ 22.0	$\alpha_i = 0.10$
		$\gamma_i = 0.02$
B-List	$20.5 \geq$ and < 22.0	$\alpha_i = 0.07$
		$\gamma_i \approx 0$
C-List	< 20.5	$\alpha_i = 0.05$
		$\gamma_i = 0$

Table 2. Salary cut distribution for actors and directors based on status and value.

Given that one of the main objectives of the model aims to determine the best set of leading actors and directors, a means of optimization can be defined to obtain the best possible solution. The aim of optimization is to select a set of actors and a director for the cast in accordance to the utility value. To do as such, a knapsack program can be utilized, to choose an optimal set of casts, and to maximize actor/director utility. With this rationale, the objective function for the value v_i and actor/director x_i is as follows:

$$Maximize \quad \sum_{i=1}^{n} v_i x_i \;\; \forall\, i = 1, 2, \ldots, n$$

$$Subject\ to \quad \sum_{i=1}^{n} x_i \leq C \;\; \forall\, i = 1, 2, \ldots, n$$

$$\sum_{i=1}^{n} s_i x_i \leq \rho B \;\; \forall\, i = 1, 2, \ldots, n$$

$$x_i \in \{0,1\} \;\; \forall\, i = 1, 2, \ldots, n$$

The objective function aims to maximize actor/director utility, where x_i indicates whether actor/director i is selected and subject to the constraint that the number of optimal casts cannot exceed user-defined cast size C. When selecting actors and directors, a distributor must allocate a percent of the budget towards actor payment, where cast salary cannot exceed this parameter. Since a portion of the budget is dedicated to salaries, the parameter ρ defines that percentage, which has been set to *0.3* for actors and *0.1* for directors, in the general case [2][3]. Optimizing casts and directors returns values for each, which is then aggregated to generate an overall crew value as the sum of cast and director values for a movie.

Through model stacking, the results of the optimization are input into a linear regression model for box office prediction, by using the same categories as the movies dataset (see Appendix C), as well as the release date, production budget and a newly optimized crew value. To account for categorical variables, one-hot encoding is utilized to map these variables with binary values. Additionally, categories occupying 5% to 10% of the entries and highly correlated categories are combined to reduce multicollinearity and to prevent overfitting, regularization techniques such as Ridge and Lasso regression are employed.

Results

By evaluating the model, a training and testing score of *0.65095* and *0.629036* were obtained for linear regression respectively, with Ridge and Lasso regularizations outputting slightly lower scores. Additionally, the beta coefficients of significant attributes such as crew value and production budget were obtained as *1.57e+6* and *6.09e+6*, respectively.

Discussion

From the results from the linear regression, the benchmark statistic used to evaluate the validity of the optimization model is the crew value beta coefficient. This is due to the fact that the beta coefficient indicates the dependence of the model on the variable. That is, if the model returns a high beta value with respect to other betas, it will indicate that the optimization model is valuable for usage.

Additionally, the strength of the optimization model needs to be considered to determine if it is worth investing time in using. The benchmark statistic used for this evaluation is the difference in box office returns between the model with existing crew values inputted and the model with the optimized crew values. Such a difference in domestic gross returns could suggest that the optimization model is strong. Based on the results of the regression, a beta coefficient value of approximately $\beta = 1.57e+6$ for crew value suggests that the optimization model used is viable. While the production budget value had a much higher value for beta with $\beta = 6.09e+6$, the crew value beta was the second most correlated, outperforming the other features. For the overall strength of the optimization model, the sample box office predictions outperform the original domestic gross to where it is worth using logistically. Taking the sample prediction example of the movie "Avengers: Endgame," the most valuable of the first eight common members of the original and optimized cast are in Table 3 with the last row for the director:

Original	Optimized
Robert Downey Jr.	Chris Pratt
Chris Hemsworth	Robert Downey Jr.
Scarlett Johansson	Zoe Saldana
Chris Pratt	Scarlett Johansson
Zoe Saldana	Chris Hemsworth
Samuel L. Jackson	Vin Diesel
Chris Evans	Samuel L. Jackson
Mark Ruffalo	Jason Statham
Joe Russo & Anthony Russo	Michael Bay

Table 3. The original and optimized casts and directors for "Avengers: Endgame."

The value for the original crew is approximately *48.47* and *49.06* for the optimized crew. Since the difference is noticeable on a logarithmic scale, it indicates that the optimization model is reliable in producing valuable results. For this specific case, it shows that if the movie had the optimized crew, then the box office earnings would have been greater than or equal to that of the original. This resulted in being true since the original had a box office of $858,373,000 and the optimized version had a box office of approximately $901,653,103, showing that the model performs well in this regard.

The linear regression model's most important factor for its evaluation is the goodness of fit, to determine if the chosen variables correlate with domestic gross. For the linear regression

model, the train score and test score were used as benchmark statistics to evaluate the fit of the variables. Given the train score and test score results, it can be concluded that the regression model's fit is subpar and not very reliable. This is likely due to the fact that a majority of the features being categorical and one-hot encoded, which allows for the quantitative features to essentially dominate in terms of correlation. What the regression does account for, however, is features that do correlate with some statistical significance. To address the issue of having an excess of one-hot encoded features, techniques such as grouping categories with low statistical power and grouping categories based on similar features can be implemented to a greater extent. One further intervention that may aid in bettering the regression is doing further exploratory data analysis to search for more quantifiable features that may have a correlation with the domestic box office earnings, thereby improving the overall score for both the training and testing sets.

With respect to possible weaknesses the optimization may have, the knapsack integer program at the crux of the model is linear, meaning that any combination of actors and directors directly provides a linearly aggregated utility sum. With this, the possible synergies and various dynamics that groups of certain actors and directors may have between each other are not accounted for and thus, would not be as reflective of movies that do contain such dynamic combinations. Furthermore, additional constraints can be implemented within the program that are external to the budget but still may have an impact on the results, which would thereby correspond more to how it is in reality.

To scale for market trends and inflation, the budget and domestic box office earnings were adjusted for inflation via the division of yearly ticket prices. This would therefore reduce the variance of the monetary data along with the skew towards later years since there was an increasing trend overall in the prices of tickets annually. However, after adjusting the model for this, there was not a significant improvement to the scores as was expected, indicating that the method of adjusting was not effective. In the future to improve this, more sophisticated methods for calculating and adjusting for inflation can be employed rather than doing so via yearly ticket prices.

Likewise, the quantitative data was standardized so that the values were scaled down to more meaningful proportions between *0* and *1*, but made no changes to the model results.

In terms of future directions, the optimization model can be expanded to account for broader features such as physical traits, ethnicity, and age. Additionally, due to the lack of data for actor and director value calculations apart from financial data, further research should be conducted for datasets containing data on other aspects that can reflect important factors such as popularity, bankability, and salary over a continuous period. As mentioned previously, the optimization model can also be changed to account for synergies between different actors and directors. This would imply that certain groupings of individuals would yield larger utility values when chosen together. A possible implementation could take the form of a nonlinear integer program in which specific combinations of actors and directors would be scaled by some factor or order of magnitude, thus making the model more reflective of real occurrences in the film industry. Additionally, after addressing all of the mentioned weaknesses, the scope of the project can be expanded to accommodate domestic and international films as well.

Conclusion

The epicenter of the project aims to select a set of optimal leading actors and directors, with the inclusion of predicting the box office performance. The model first attempted to select a set of optimal actors and directors to maximize crew utility, by accounting for other important factors such as the percentage of the production budget associated with actor salaries, and cast size, Crew utility is a significant parameter that takes into account crew experience and bankability, both of which are of high importance to movie production companies.

Through model stacking, the results of the optimization were used as inputs to a linear regression model to predict the box office revenue. From the overall result of linear regression, the benchmark statistics used to evaluate the validity of optimization suggested that the optimization model is viable, due to the beta coefficients of crew value and production method outperforming other features. By running the model on various movies, it was observed that the model was successful at providing viable leading casts and predicting box office revenues, which would often be better than that of the original movies.

Overall, the model provides valuable insights for a movie production company regarding selecting the best combination of actors and directors for a movie, considering factors that can make or break going forward with production, potentially saving or earning a fortune.

References

[1] Statista. (2023, January 5). *Expenses of motion picture & video industries in the U.S. 2007-2021*. https://www.statista.com/statistics/185312/estimated-expenses-of-us-motion-picture-and-video-industry-since-2005/

[2] *A Guide to SAG Residuals*. (2023, March 18). Backstage. Retrieved December 6, 2023, from https://www.backstage.com/magazine/article/calculating-sag-residuals-17706/

[3] *How Much Money Do Actors Make?* (2023, October 10). Backstage. Retrieved December 6, 2023, from https://www.backstage.com/magazine/article/how-much-money-do-actors-make-75180/

[4] Provost, R. (2022, November 21). *How much do directors make — DGA rates for film & TV*. StudioBinder. https://www.studiobinder.com/blog/how-much-do-directors-make/#:~:text=This%20

Appendices

Appendix A: The specific types of each category used in the proposed model, stored in a nested Python dictionary for iteration during the web scraping process.

```
categories = {
    'genres': {
        'adventure': 1601,
        'action': 1301,
        'comedy': 2801,
        'drama': 3701,
        'thriller-or-suspense': 1401,
        'horror': 1301,
        'romantic-comedy': 701,
        'musical': 201
    },
    'creative-types': {
        'contemporary-fiction': 6401,
        'science-fiction': 1201,
        'kids-fiction': 901,
        'historical-fiction': 1500,
        'fantasy': 901,
        'super-hero': 301,
        'dramatization': 1201,
        'factual': 101
    },
    'production-methods': {
        'live-action': 8701,
        'animation-and-live-action': 501,
        'digital-animation': 601,
        'hand-animation': 101,
        'stop-motion-animation': 1,
        'rotoscoping': 1,
        'multiple-production-methods': 1
    },
    'sources': {
        'original-screenplay': 6901,
        'based-on-fictional-book-or-short-story': 2101,
        'based-on-comic-or-graphic-novel': 501,
        'remake': 501,
        'based-on-tv': 401,
        'based-on-real-life-events': 1101,
        'based-on-factual-book-or-article': 401,
        'spin-off': 101
    },
    'mpaa-ratings': {
        'pg-13-(us)': 2801,
        'r-(us)': 3901,
        'pg-(us)': 1901,
        'g-(us)': 301,
        'not-rated-(us)': 2801,
```

```
        'gp-(us)': 1,
        'nc-17-(us)': 1,
        'm-pg': 1
    },
    'theatrical-distributors': {
        'walt-disney': 801,
        'warner-bros': 1101,
        'sony-pictures': 901,
        'universal': 1001,
        'paramount-pictures': 801,
        '20th-century-fox': 801,
        'lionsgate': 601,
        'new-line': 201,
        'dreamworks-skg': 1,
        'mgm': 301,
        'miramax': 201,
        'fox-searchlight': 201,
        'focus-features': 301,
        'weinstein-co': 201,
        'summit-entertainment': 1,
        '20th-century-studios': 1,
        'sony-pictures-classics': 401,
        'stx-entertainment': 101,
        'miramax-dimension': 1,
        'relativity': 101,
        'open-road': 101,
        'united-artists': 101,
        'a24': 201,
        'roadside-attractions': 201,
        'newmarket-films': 101
    }
}
```

Appendix B: Dataset of actor utility values for the corresponding movies they played in.

	Name	Movie	Value
0	Tom Hanks	Toy Story	23.428941
1	Tim Allen	Toy Story	23.483320
2	Mel Gibson	Pocahontas	20.320583
3	Denzel Washington	Crimson Tide	21.998387
4	Gene Hackman	Crimson Tide	23.785503
...	...	...	...
8323	Samuel L. Jackson	The Last Full Measure	20.582953
8324	Peter Fonda	The Last Full Measure	21.293618
8325	Jeremy Irvine	The Last Full Measure	14.722839
8326	Charlie Plummer	Words on Bathroom Walls	15.207038
8327	Taylor Russell	Words on Bathroom Walls	13.782108

Appendix C: Final dataset of movies incorporating crew values for the cast and director.

	Movie	genres	mpaa-ratings	theatrical-distributors	Production Budget	Domestic Gross	Release Date	creative-types	production-methods	sources	Crew Value
0	Toy Story	adventure	g-(us)	walt-disney	30000000.0	192523233.0	November, 1995	kids-fiction	digital-animation	original-screenplay	46.021657
1	Crimson Tide	action	r-(us)	walt-disney	55000000.0	91387195.0	May, 1995	contemporary-fiction	live-action	original-screenplay	46.475760
2	Judge Dredd	action	r-(us)	walt-disney	85000000.0	34687912.0	June, 1995	science-fiction	live-action	based-on-comic-or-graphic-novel	40.164674
3	The Jungle Book	adventure	pg-(us)	walt-disney	175000000.0	364001123.0	April, 2016	fantasy	animation-and-live-action	based-on-fictional-book-or-short-story	28.066436
4	The Lion King	adventure	g-(us)	walt-disney	260000000.0	543638043.0	July, 2019	kids-fiction	animation-and-live-action	remake	35.609616
...	...	...	...	...	...	...	...	...	...	...	...
2503	The New Mutants	horror	pg-13-(us)	sony-pictures-classics	67000000.0	23855569.0	August, 2020	super-hero	live-action	based-on-comic-or-graphic-novel	34.361398
2504	The Gentlemen	action	r-(us)	miramax-dimension	22000000.0	36471796.0	December, 2019	contemporary-fiction	live-action	original-screenplay	40.861833
2505	Honest Thief	thriller-or-suspense	pg-13-(us)	united-artists	30000000.0	14163574.0	October, 2020	contemporary-fiction	live-action	original-screenplay	35.646974
2506	The Last Full Measure	drama	r-(us)	walt-disney	20000000.0	2949212.0	January, 2020	dramatization	live-action	based-on-real-life-events	34.623697
2507	Words on Bathroom Walls	drama	pg-13-(us)	walt-disney	9300000.0	2542518.0	August, 2020	contemporary-fiction	live-action	based-on-fictional-book-or-short-story	35.035282

Appendix D: Parameter tuning and sensitivity analysis.

The tuning conducted for the order of correlation was based on actors that clearly rank higher than others but have lower utility. One of the actors whose values were not reflective of their status in the action category was Robert Downey Jr., who had a lower utility value than Paul Bettany. This was due to the latter having a larger average domestic box office earnings and less movies participated in than the former. Thus, λ was adjusted such that this discrepancy was handled accordingly with Robert Downey Jr. being higher ranked than Paul Bettany. This sensitivity analysis was done for various actors and directors having the same issue and as such, the parameter that corresponded to the most optimal rankings was $\lambda = 2$.