

Increasing Revenues for Glen Art Theater through Analytics



Final Report

Prepared By: FilMetrics

Delivered On: May 28, 2019

Increasing Revenues for Glen Art Theater Through Analytics

Contents

| | |
|---|-----------|
| Problem Overview | 2 |
| Description of Data | 3 |
| Data Overview | 5 |
| Data Transformations and Feature Engineering | 9 |
| Modeling and Findings | 10 |
| Optimization | 13 |
| Analytic Dashboard and Mobile Application | 14 |
| Dashboard | 14 |
| Analytics | 15 |
| Predictive | 15 |
| Historical | 16 |
| Raw Data | 16 |
| Optimization | 17 |
| Conclusions and Recommendations | 17 |
| What's Next? | 18 |
| Project Status | 18 |
| Appendix 1 - Project Plan | 20 |
| Appendix 2 - Data Dictionary | 23 |
| Appendix 3 - EDA Results | 25 |
| Appendix 4 - Modeling Methods | 32 |
| Appendix 5 - Optimization Algorithm | 34 |



Problem Overview

The film industry has seen a marked expansion over the past decade, with box office revenues increasing by approximately 25%. However, due to the established economics of displaying films, theater owners are unable to fully capitalize on the expanding market as studios and distributors command fees tied to the week of a given engagement that are front loaded for the most lucrative weeks of a given film's run, as illustrated in Table 1 below.

| Week of Release | Box Office Net Percentage | Box Office Gross Percentage |
|-----------------|---------------------------|-----------------------------|
| 1 | 95% | 70% |
| 2 | 95% | 70% |
| 3 | 90% | 60% |
| 4 | 85% | 50% |

Table 1 *Distributor's Due Proportion of Ticket Sales*

With this model in play, it is not unusual for theaters to lose money while showing even the most popular films available at any given time. This is especially true during the second week of release when most films see sharp declines of between 40-60% in box office revenues but require the same percentage of the week's revenue to be paid to the distributor. Among 4,796 films measured to date, only 212 (4%), have seen second week box office increases¹, meaning that the best case scenario is to show films that will have the slowest decay in revenues during their engagement. Identifying films that will have momentum beyond their opening weekend is of prime importance for theater owners since this enables them to increase the possibility that they will see profits from the box office, especially as many films are spending less time in theaters in order to accelerate their release to the direct-to-consumer market. Paired with already existing alternate revenue streams from concessions the goal is to maximize theater profits without heavily relying on a loss leader philosophy.

Description of Data

Data has been drawn from a number of open sources including imdb.com, themoviedb.org, and movielens.org. Table 2 below summarizes the curated list of selected files that contain data

¹ Smallest Second Weekend Drops 1982 - Present, <https://www.boxofficemojo.com>



Increasing Revenues for Glen Art Theater Through Analytics

that is believed to be what will be most helpful in determining projected revenues to be the most assistance to Glen Art Theater.

| Category | Description | Additional Information |
|------------------------------|---|--|
| File Name | credits.csv | |
| File Format | Comma delimited | Header row |
| Number of Records | 45,504 | See Appendix 2 for data dictionary |
| Number of Fields | 10 | Cast and crew information for indexed film IDs |
| Potentially Useful Variables | Name Character Department Gender ID | Can provide insight into: <ul style="list-style-type: none"> • Individual star power • Impact of cast/crew combinations • Gender bias |
| File Name | keywords.csv | |
| File Format | Comma delimited | Header row |
| Number of Records | 46,419 | See Appendix 2 for data dictionary |
| Number of Fields | 3 | List of keywords associated with indexed film IDs |
| Potentially Useful Variables | Name Id | May correlate keywords with box office success |
| File Name | links.csv | |
| File Format | Comma delimited | Header row |
| Number of Records | 45,843 | See Appendix 2 for data dictionary |
| Number of Fields | 3 | X reference info for imdb.com and themoviedb.org |
| Potentially Useful Variables | imdbid tmdbid | Strictly functional usefulness to link different data sets |

Increasing Revenues for Glen Art Theater Through Analytics

| | | together |
|------------------------------|--|---|
| Category | Description | Additional Information |
| File Name | movies_metadata.csv | |
| File Format | Comma delimited | Header row |
| Number of Records | 45,466 | See Appendix 2 for data dictionary |
| Number of Fields | 24 | Key information associated with indexed film IDs |
| Potentially Useful Variables | Budget Genres Production_Companies Revenue Popularity Director_Rank | Can provide insight into: <ul style="list-style-type: none"> • Correlation of budget to success • Correlation of genre to success • Do production companies differ in success • Does fan "buzz" result in box office success • Are certain directors more likely to have financial success |
| File Name | ratings.csv | |
| File Format | Comma delimited | Header row |
| Number of Records | | See Appendix 2 for data dictionary |
| Number of Fields | 4 | General public ratings for indexed film IDs |
| Potentially Useful Variables | Rating | Aggregated rating information for unreleased films could be an indicator for future box office success |

Table 2 *List of Files Currently in Use for Analysis*

Increasing Revenues for Glen Art Theater Through Analytics

In addition to these data, the incorporation of weekly revenue information per film available on boxofficemojo.com aided in identifying films that will have the smallest week over week revenue losses. Films that meet those criteria would then be the most attractive to theater owners since these films run the least risk of resulting in a net weekly loss after fulfilling the weekly distribution fee requirements.

Data Overview

After collecting the various data sources, the next step the FilMetrics team took was to examine the collected data to ensure what was collected was complete, did not contain any unusual or unexplained results, and check for any missing data. To do so, the team first compiled the discrete data sources into a composite data set that included 46,622 individual records with 13 fields. As part of the compilation of this data, categorical variables were created to properly model individual genre values which brought the field total up to 34 before any additional data transformations or feature engineering took place. Five additional fields were added during the feature engineering process, which will be discussed in more detail below, bringing the current working data set up to 39 unique fields.

With the data set ready for examination, the first step taken was to clean it of any incomplete records in order to maximize the utility of any subsequent modeling. In doing so, records that were missing budget or revenue values were identified for exclusion which reduced the initial data set rather significantly to 5,459 individual records. While potentially concerning, as will be discussed in more detail below, this data reduction does not appear to have any negative impact on the results.

From here, a number of explorations of the data were undertaken and recorded to ensure the records did not contain any concerning outliers. A complete record of these graphical outputs can be found in Appendix 3. As should hopefully be obvious, there are outliers that exist at the top of the revenue and budget ranges that are tied to high production cost and high grossing films (see Figure 1 below).

For the purposes of this analysis and the underlying goals of the project, these outliers will remain in the data set as a means to attempt to model for similar future success. It should also be noted that the data was collected prior to the release of *Avengers: Endgame* and so the box office totals associated with that film are excluded from the process.



Increasing Revenues for Glen Art Theater Through Analytics

Highest Grossing Movies

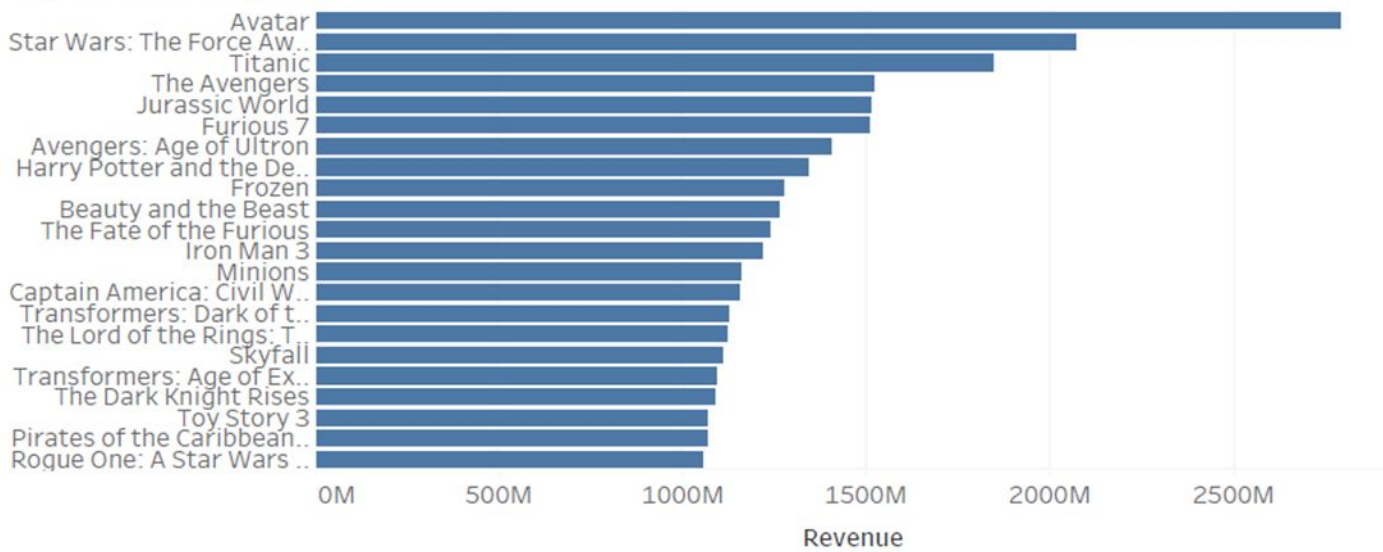


Figure 1 Highest Grossing Films in dataset

Looking at the number of movies by genre, the most number of movies is Drama, followed by Comedy and Action (Figure 2).

Number of Movies by Genre

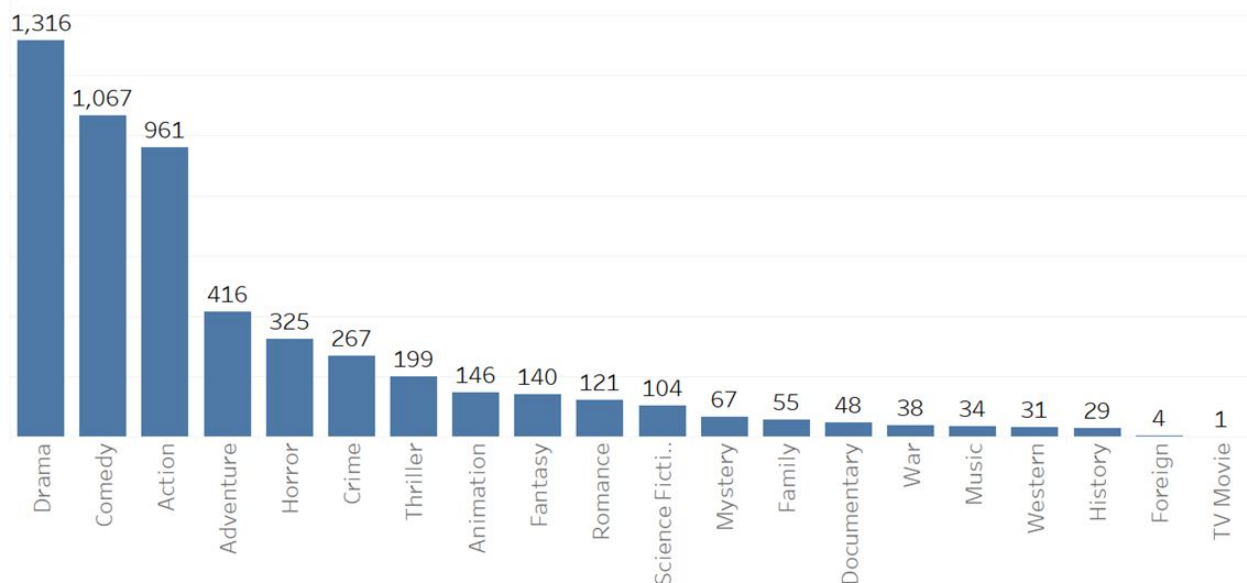


Figure 2 Number of movies by Genre

Increasing Revenues for Glen Art Theater Through Analytics

The other significant effort undertaken as part of the exploration of the data was to identify which field or fields correlate most closely with the target revenue field. To assist with this a heat map, identified as Figure 3, of the correlations between all of the available features was created in order to graphically depict which features held the most potential for predictive efficacy. In the heat map, the intersection of each pair of features is colored to correspond to the correlation, or connection, calculated to exist between them. In this instance, the lighter the intersection point is, the stronger the relationship is between features, where the darker the intersection point is the weaker the relationship is. With that understanding, the upper left quadrant contains the vast majority of the high correlations. In particular, looking at what is most highly correlated with revenues, two features stand out as being most appropriate for predicting revenue: budget and vote_count. Budget is self-explanatory, while vote_count is the number of votes associated with online user ratings associated with a given film.

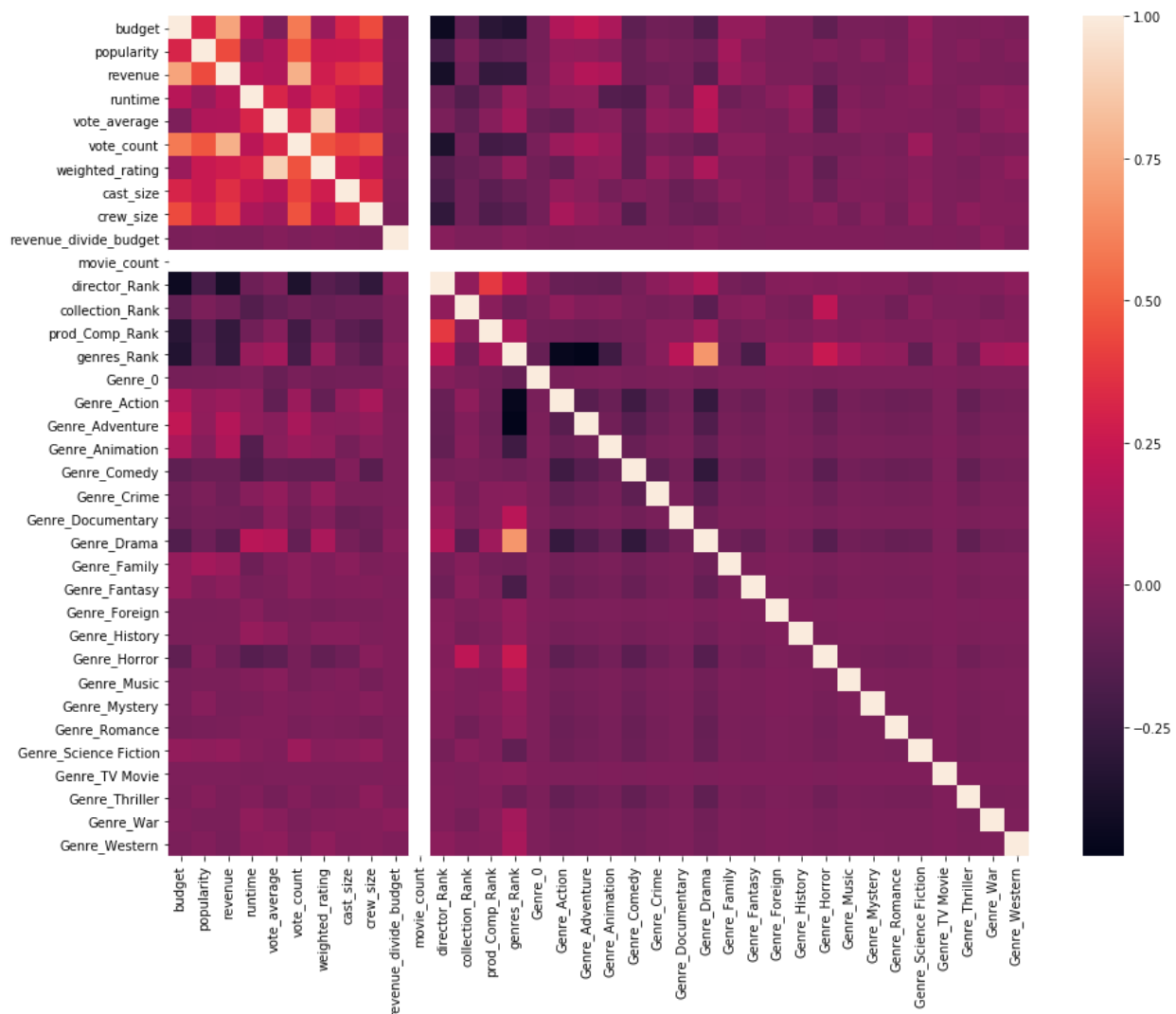


Figure 3 Heat Map of Feature Correlations

Increasing Revenues for Glen Art Theater Through Analytics

While the correlations are not as strong at the individual film level, there are strong correlations at the genre level between revenue vs popularity (Figure 4) and revenue vs weighted rating (Figure 5).

Revenue vs Popularity by Genre

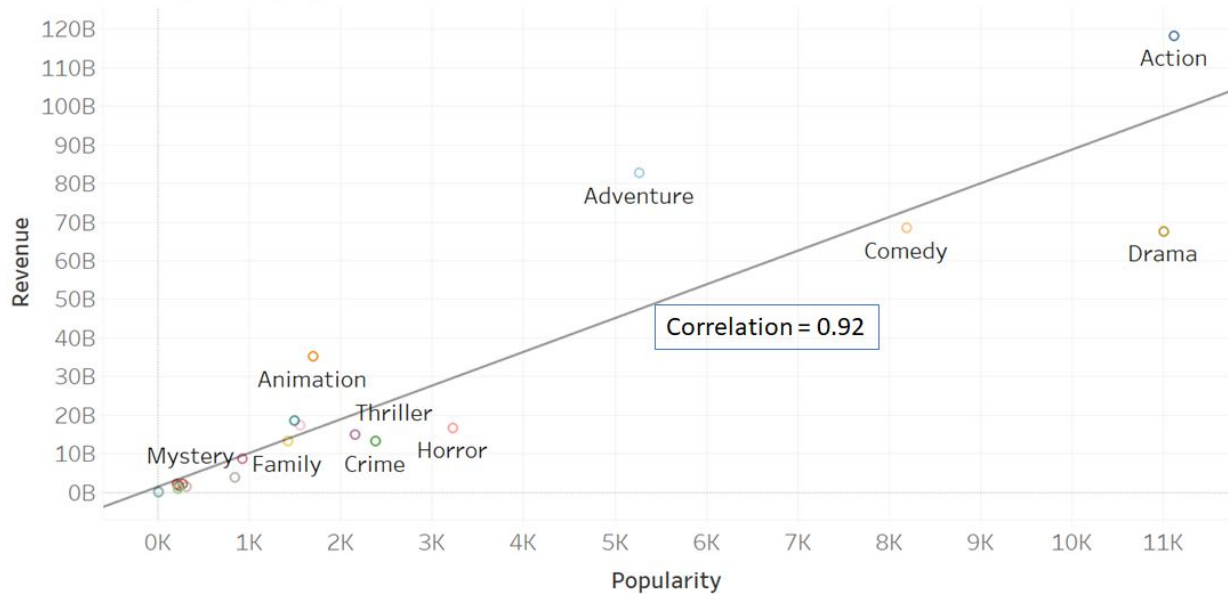


Figure 4 *Correlation of Revenue vs Popularity*

Revenue vs Weighted Rating by Genre

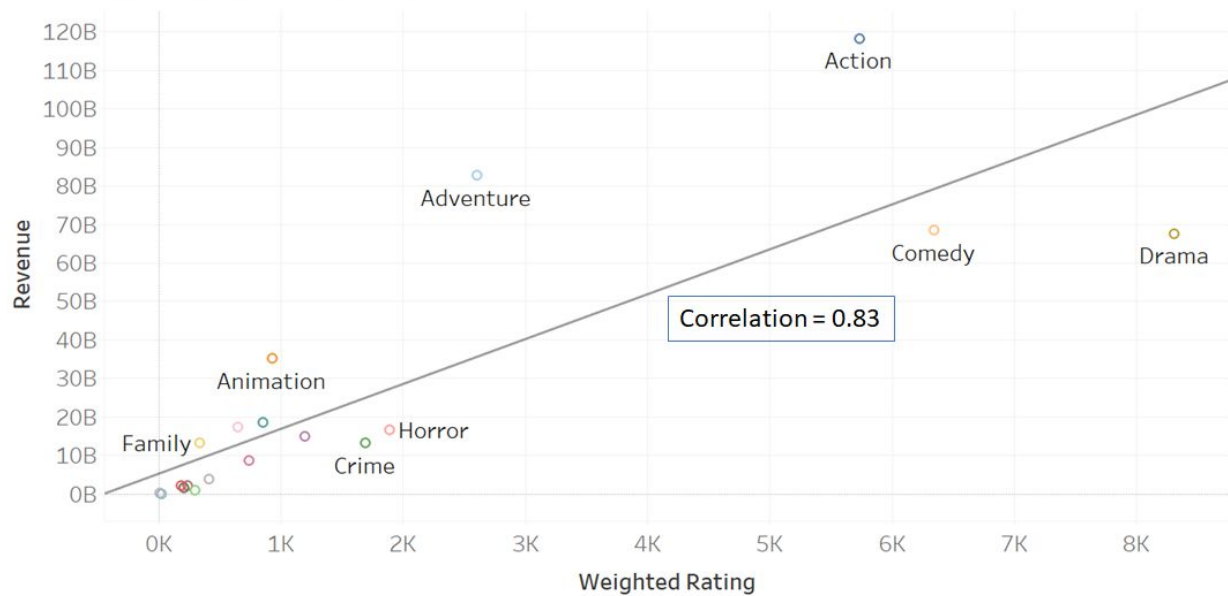


Figure 5 *Correlation of Revenue vs Weighted Rating*

Data Transformations & Feature Engineering

As noted above, five features were created in order to test their potential predictive efficacy. The first, Weighted Rating, was built using a version of IMDB.com's weighted rating formula² but using ratings information from themoviedb.org:

$$\text{Weighted Rating (WR)} = \left(\frac{v}{v+m} * R\right) + \left(\frac{m}{v+m} * C\right)$$

where,

- v is the number of votes for the film
- m is the minimum votes required to be listed in the data
- R is the average rating of the film
- C is the mean vote across the full data set

With this formula in hand, the next step was to determine an appropriate value for m , or the minimum votes required to be listed in the chart. The decision was made to use the 95th percentile as the cutoff. In other words, for a film to feature in the results, it must have more votes than at least 95% of the films in the list. Being able to weight the ratings of the films allows for a more equitable assessment of each film's quality based on user reviews.

In addition to the weighted rating information, a ranking algorithm was developed to rank different features in order to test if these rankings could have predictive influence. These rankings are primarily based on the revenues associated with each film. The individual features created are:

- **Director Ranking:** Aggregated revenue generated by the director's movies and then rank the directors based on total revenue generated by that director
- **Collection Ranking:** First, grouped films together into collections such as *The Lord of the Rings Trilogy*, *Star Wars*, or *Jason Bourne* and then rank those collection based on the aggregated revenue across the collection
- **Production Company Ranking:** Ranking production companies by the revenue generated by each
- **Genre Ranking:** Ranking genre by revenue

Modeling and Findings

Initial modeling had been undertaken with a 'less is more' philosophy. As such, models were initially built either using minimal predictor variables, primarily budget and/or weighted rating,

² <http://answers.google.com/answers/threadview/id/507508.html>



Increasing Revenues for Glen Art Theater Through Analytics

or used all available ratings to test model performance. Initial testing revealed that there were a core set of features that had the most influence over the outcomes of the modeling. Using this information, as derived from the Random Forest Decision Tree as seen in Figure 6, the following nine features were selected to be used for training of all models: *vote_count*, *budget*, *weighted_rating*, *runtime*, *cast_size*, *director_Rank*, *crew_size*, *prod_Comp_Rank*, and *vote_average*

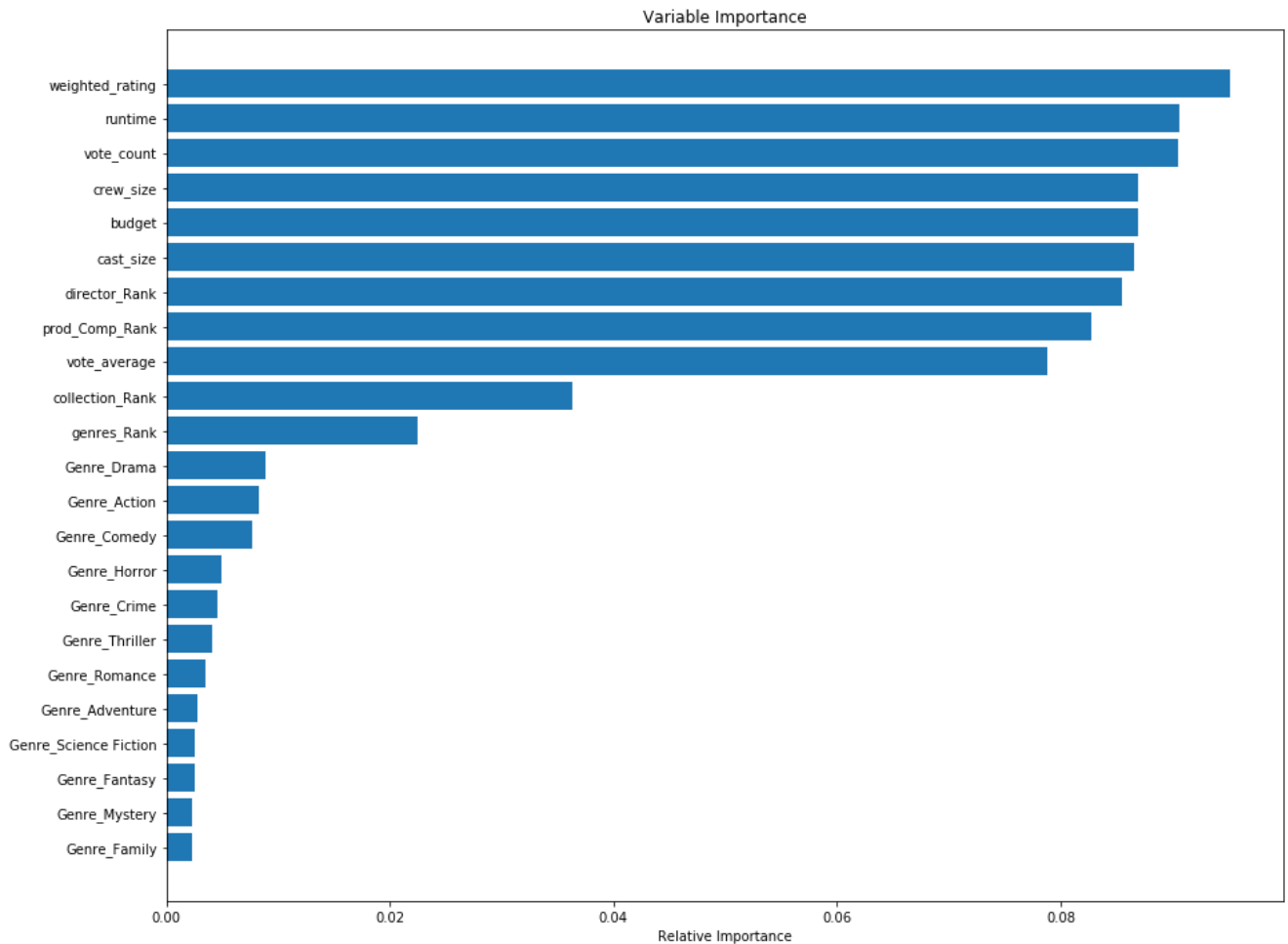


Figure 6 Ranked Variable Importance

With this condensed feature list in hand updated testing and analysis was completed in order to identify the model best suited to provide predictive accuracy for the goal of predicting future film box office success. Multiple modeling techniques were tested using these features including: linear regression, ordinary least squares (OLS) regression, random forest classification, polynomial regression, lasso regression, ridge regression, support vector machines (SVM), and a deep learning AI approach. Unfortunately unforeseen issues were encountered when testing the SVM model and as such results are currently unavailable for this model. For more information about each of these modeling methods please see Appendix 4.

Increasing Revenues for Glen Art Theater Through Analytics

The interesting outcome is that nearly all modeling approaches achieved similar results, particularly when looking at the R^2 statistic, used to measure the amount of variance that is explained by the model based on values that range from 0 to 1. An R^2 closer to 1 indicates that the model “fits” the data more perfectly and so can account for the inherent variance present in the data, while an R^2 closer to 0 indicates less of the variance is explained and so “fits” the data poorly. Table 3 below provides an accounting for the calculated R^2 for each model. Unless otherwise noted all models used the nine features outlined above for model training.

| Model | R^2 Result |
|---|--------------|
| Ordinary Least Squares (OLS) 1: <i>budget</i> | 8.862 |
| Ordinary Least Squares (OLS) 2: <i>budget and weighted rating</i> | 8.858 |
| Linear Regression | 0.71 |
| Random Forest Decision Tree | 0.52 |
| 2 nd Degree Polynomial Regression | 0.79 |
| 3 rd Degree Polynomial Regression | 0.73 |
| 4 th Degree Polynomial Regression | 0.51 |
| Lasso Regression | 0.72 |
| Ridge Regression | 0.72 |
| Deep Learning AI | 0.73 |
| Support Vector Machines (SVM) | TBD |

Table 3 R^2 Summary

Note: For the OLS models, the R^2 values over 1 are indicative of a model that fits the data particularly poorly making the use of these models inadvisable.

On the whole these results are better than initially anticipated, especially with six different model variants having R^2 over 0.70. Absent any further refinement or the ability to successfully model with SVM the 2nd Degree Polynomial Regression model is what has been identified as being production ready and is what all subsequent efforts have leveraged.

With a model in place to predict any given film’s anticipated total gross, the next step to deliver insights that are truly valuable to Glen Art Theater the anticipated weekly breakdown of that



Increasing Revenues for Glen Art Theater Through Analytics

total gross is needed. Leveraging historic film weekly revenues the FilMetrics team was able to determine the average anticipated percentage of overall gross per week through the first ten weeks of release, as illustrated in Table 4 below.

| Week of Release | % of Total Gross |
|-----------------|------------------|
| Week 1 | 30.4% |
| Week 2 | 18.8% |
| Week 3 | 12.8% |
| Week 4 | 9.4% |
| Week 5 | 7.1% |
| Week 6 | 6.1% |
| Week 7 | 5.1% |
| Week 8 | 4.4% |
| Week 9 | 4.3% |
| Week 10 | 3.6% |

Table 4 *Anticipated Weekly Portion of Total Film Grosses*

For the initial delivery of insights these averages will be utilized to assist in identification of films to display as well as factor into the screen optimization tool. Admittedly, further refinement of this model is required in order to become more individual film specific, but a significant portion of that refinement will be reliant on actual POS data collected from Glen Art Theater on a going forward basis in order to better account for location specific attendance, demographic, and seasonal traffic patterns that are currently unavailable. Likewise, other factors such as correlation between individual films and concession sales can also be captured on a going forward basis in order to further refine the model to better serve the needs of Glen Art Theater beyond what can be ascertained via the national data currently in use.

Optimization

In parallel with the predictive modeling activities, the FilMetrics team developed an optimization algorithm to aid theater management in the scheduling of selected films within operational constraints. The algorithm draws inspiration from previous work which combined a



Increasing Revenues for Glen Art Theater Through Analytics

movie demand forecast output with a linear programming optimization model to produce a weekly movie schedule for a large theater in the Netherlands.³ While many of the underlying principles from their methodology apply to the Glen Art Theater scenario, there are enough nuances and differences between the two that a different formulation for our solution was warranted. For an in-depth review of our optimization algorithm, please see Appendix 5.

The FilMetrics team has chosen a problem formulation from the field of dynamic programming known as weighted interval scheduling (WIS). The goal of a WIS problem is to find the maximum-weight subset of non-overlapping jobs, given a set of all possible jobs that have associated weights with them. Each job has a start time, a finish time and a weight. In the Glen Art Theater scenario, jobs are replaced by potential showtimes for a movie. We must consider every possible start/end time for each selected film such that it falls within the allotted screening window. The weight then becomes the forecasted demand or revenue should a specific film be shown at a specific time.

The allotted screening window is determined by Glen Art Theater management. The selected films all have associated run times, to which theater management also adds time for advertisements, trailers, and cleaning. Our predictive models supply the demand/revenue information. With this, we have enough information to optimize the schedule for a single screen on a given day.

Additional attributes or constraints beyond run time and the others listed above can also be incorporated into the optimization protocol. For instance, the scheduling of G- or PG-rated films can be de-emphasized after 8:00 PM while R-rated films may be prioritized for showtimes starting after 6:00 PM. This process promises to be an improvement over existing scheduling methodologies in two primary ways: (1) the theater is no longer bound by the conventional restriction of assigning one film to one screen and simply fitting as many screenings into the available screening window as possible, and (2) significant time is saved by automating a routine task.

Analytic Dashboard & Mobile Application

In order to provide real time updates on both Glen Art Theater operations as well as up to the minute film projections a dashboard and mobile application has been designed and deployed for use by theater management. These visualizations can be accessed via any device connected to the internet at: <https://filmetricscapstone.shinyapps.io/FilMetricsCapstone/>

³ Eliashberg, J., Hegie, Q., Ho, J., Huisman, D., Miller, S.J., Swami, S., Weinberg, C.B., & Wierenga, B. (2009). Demand-driven scheduling of movies in a multiplex. *International Journal of Research in Marketing*, 26, 75-88.



Increasing Revenues for Glen Art Theater Through Analytics

As an overview of the available information, the site is broken down into three main components:

Dashboard

As the landing page for the mobile application this section contains at a glance information about Glen Art Theater's operational status, including status of meeting annual sales targets, performance of Glen Art Theater vs averages, and top grossing films for the current week, month, and quarter.

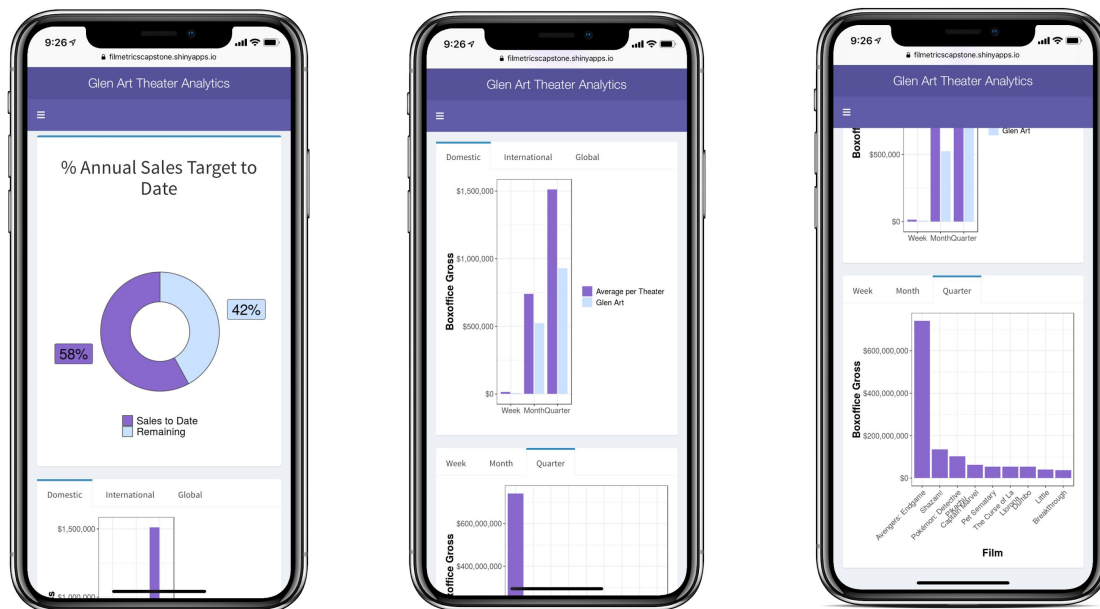


Figure 7 *Sample dashboard views*

As with some of the modeling efforts, the addition of POS information from Glen Art Theater on a going forward basis will aid in increasing the accuracy and usefulness of these reports. For instance, incorporating concession sales into the annual sales to date will be a simple exercise if and when that data is both available and interest is expressed.

Analytics

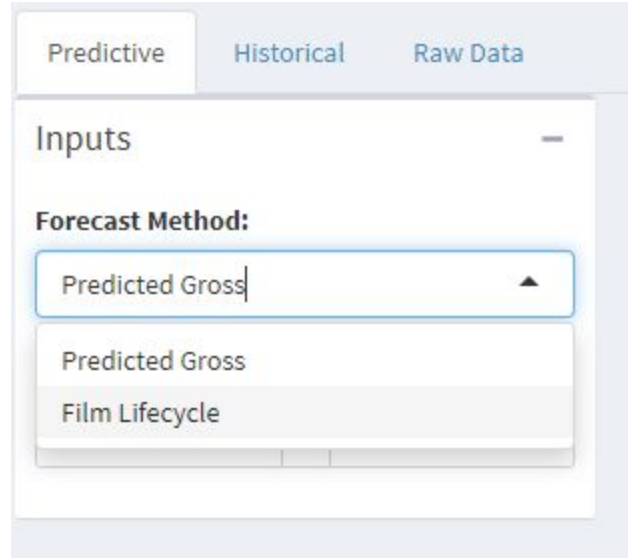
The analytics section is designed to enable exploration of additional insights based on Glen Art Theater's management interest or need for further information about a variety of categories. By providing multiple query and filter options selected to maximize utility with minimal input from users visualizations and tables can be created to answer virtually any question that management can imagine.

Increasing Revenues for Glen Art Theater Through Analytics

Predictive

Using model outputs paired with films selected Glen Art Theater management can view the predicted gross associated with the booked slate for a selected date range or view the anticipated fluctuations in daily grosses for one or more films scheduled to release in the next two quarters. Simple drop down menus and calendar selections help guide the user to select what they need intuitively.

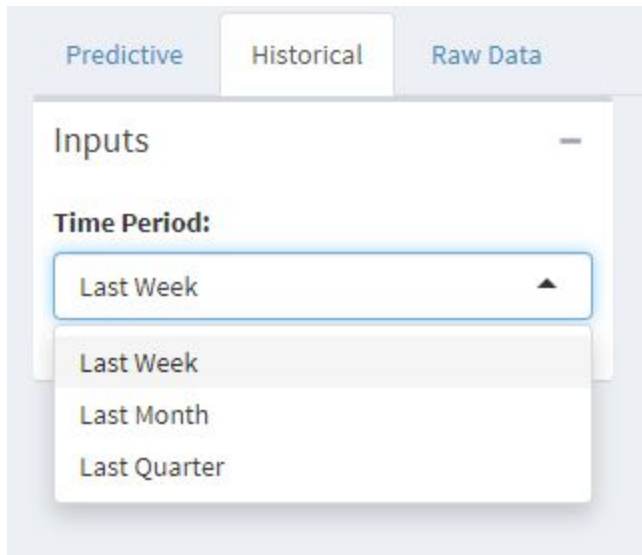
For future releases, a potentially muddled upcoming slate of films is limited to only those scheduled for wide release and are further broken down by scheduled month of release for simple identification and selection.



The screenshot shows the 'Predictive' tab selected. Under the 'Inputs' section, the 'Forecast Method:' dropdown menu is open, showing 'Predicted Gross' as the selected option. Below it, 'Film Lifecycle' is also visible in the dropdown list.

Historical

Provides a look back at actual film performance vs projected for last week, last month, and last quarter to track and gain confidence about the overall quality of projections as well as to confirm when they failed to anticipate potential losses. This will serve as an excellent means to promote an ongoing dialogue about potential future improvements in the modeling as the pilot program progresses.



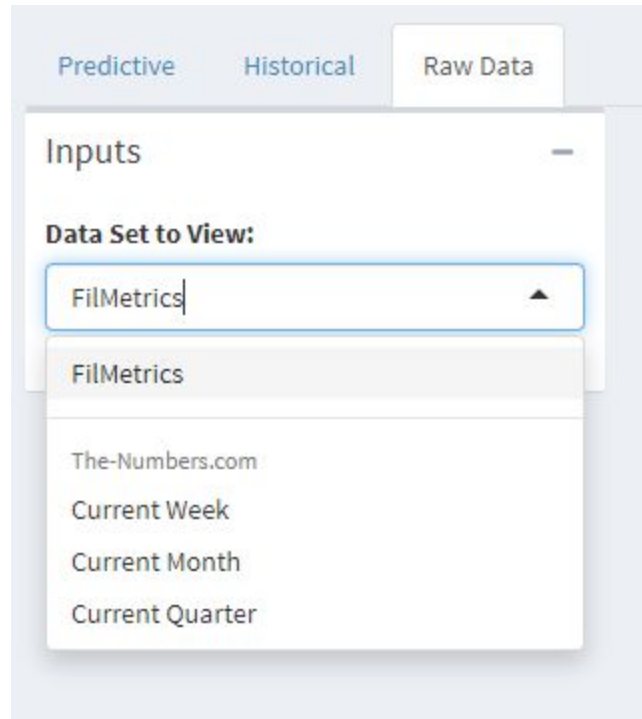
The screenshot shows the 'Historical' tab selected. Under the 'Inputs' section, the 'Time Period:' dropdown menu is open, showing 'Last Week' as the selected option. Below it, 'Last Month' and 'Last Quarter' are also visible in the dropdown list.

Increasing Revenues for Glen Art Theater Through Analytics

Raw Data

When the numbers can speak for themselves the option to explore them is made available to management. While FilMetrics is confident in the choices made for which features to use in the projections provided, if a more granular view of either the training data or the real time feed of film data is required it can be accessed quickly and easily.

Once a set of data is selected, the resulting table is easily sorted and scanned to find information of interest easily and without significant effort.



Optimization

Leveraging the optimization algorithm developed in tandem with the predictive outputs of the selected model Glen Art Theater management can select constraints for a given day and then generate an optimized schedule within milliseconds expressly designed to increase Glen Art Theater revenues. Operating hours, enforced minimum screening counts, and intervals designed to allow patrons to fill a theater can all be incorporated based on the theater's needs. Likewise, should any future adjustments to the Glen Art Theater physical space be made the tool can adjust for the number of screens as well.

Inputs

Date to Schedule:

Available Films

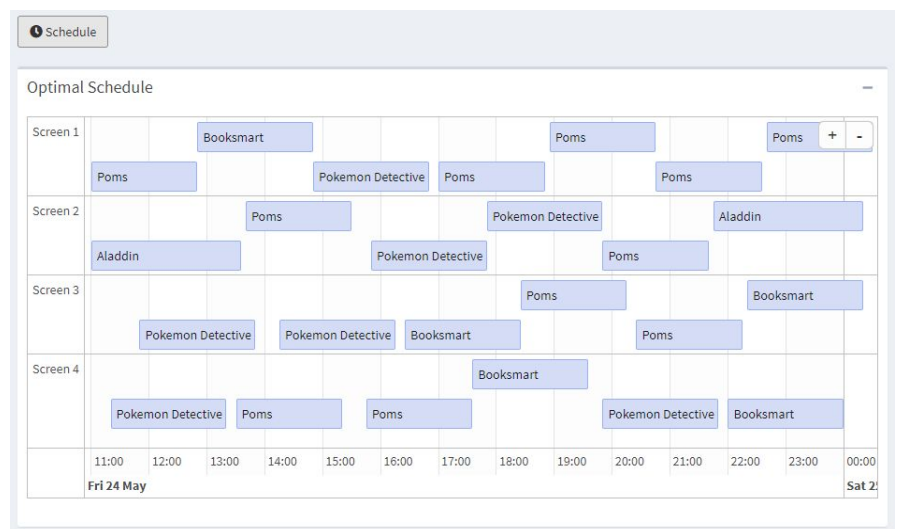
Earliest Start Time:

Latest Finish Time:

Interval Between Start Times (min)

Available Screens

Minimum # Showings per Film



Conclusions and Recommendations

As mentioned previously, the modeling and deliverables provided have met the initial requirements outlined at the beginning of this project. Using the 2nd Degree Polynomial Regression model just under 80% of the variance present in the available data is accounted for. FilMetrics does plan to continue to explore why the Support Vector Machine (SVM) model failed in order to verify that this approach can not outperform the selected model. Otherwise, in real world terms, this is an excellent outcome and suggests that the initial projections for total grosses through the end of 2019 should be highly accurate, with the understanding that a certain percentage of projections will not successfully mirror the real world outcomes. Having

Increasing Revenues for Glen Art Theater Through Analytics

this information in hand will enable Glen Art Theater to make better informed purchasing decisions for what films to display during the third and fourth quarters of 2019.

With those purchasing commitments made, the optimization tool will allow management to maximize the use of the available screens at the theater based on not just run time, but also by doing things like de-emphasizing showings of G or PG rated films after 8PM, promoting the showing of R rated films after 6PM, accounting for positive or negative buzz around a given film, and factoring in where a given film is in its life cycle in order to use the available screens in the most revenue focused way possible.

What's Next?

With the completion of the initial proof of concept to Glen Art Theater the work is really just beginning. FilMetrics is committed to not only supporting the use of the mobile application provided, but also in providing refinement and improvement via feedback received from Glen Art Theater. As a pilot program and proof of concept FilMetrics expects that there will be bugs to work out and that through use of the mobile app management may identify additional information and feedback they would like to use.

As mentioned above, one area already anticipated as being of interest, but that was outside the scope of the original project plan, is the incorporation of concession sales and other POS data to further refine the predictions, dashboard, and other reporting to account for what makes the Glen Art Theater experience unique. All of the input data currently in use represents national totals and averages which are more than likely out of sync with the realities of Glen Art Theater to some degree. On the one hand, the wonderful thing about the film industry is that so much of the information associated with it is data driven and publicly available , so the real question is about how best to leverage that public information in conjunction with theater specific data. While underestimating the total gross for a blockbuster is an acceptable error for a studio or distributor since it will only lead to unanticipated revenue growth, for a theater owner the same error could lead to under ordering of a film and missing out on the prospective financial benefits.⁴

Project Status

At this time the original goals of this project project are complete with the exception of the presentation to the CEO during the week of June 3. We will follow-up with proposed times to conduct that meeting this week. The project plan in Appendix 1 has been reverted to illustrate all work completed throughout the project.

⁴

<https://www.theringer.com/movies/2019/5/7/18534880/marvel-cinematic-universe-mcu-box-office-projections-a-vengers-endgame-spiderman-far-home>



Increasing Revenues for Glen Art Theater Through Analytics

Going forward, as mentioned earlier, FilMetrics would like to have, at minimum, monthly follow-up meetings for the next three months to gather feedback from the Glen Art Theater team in order to further optimize the performance of all aspects of the mobile application for the standalone Glen Art Theater deliverable as well as for the larger beta test group scheduled to go live during Q4.

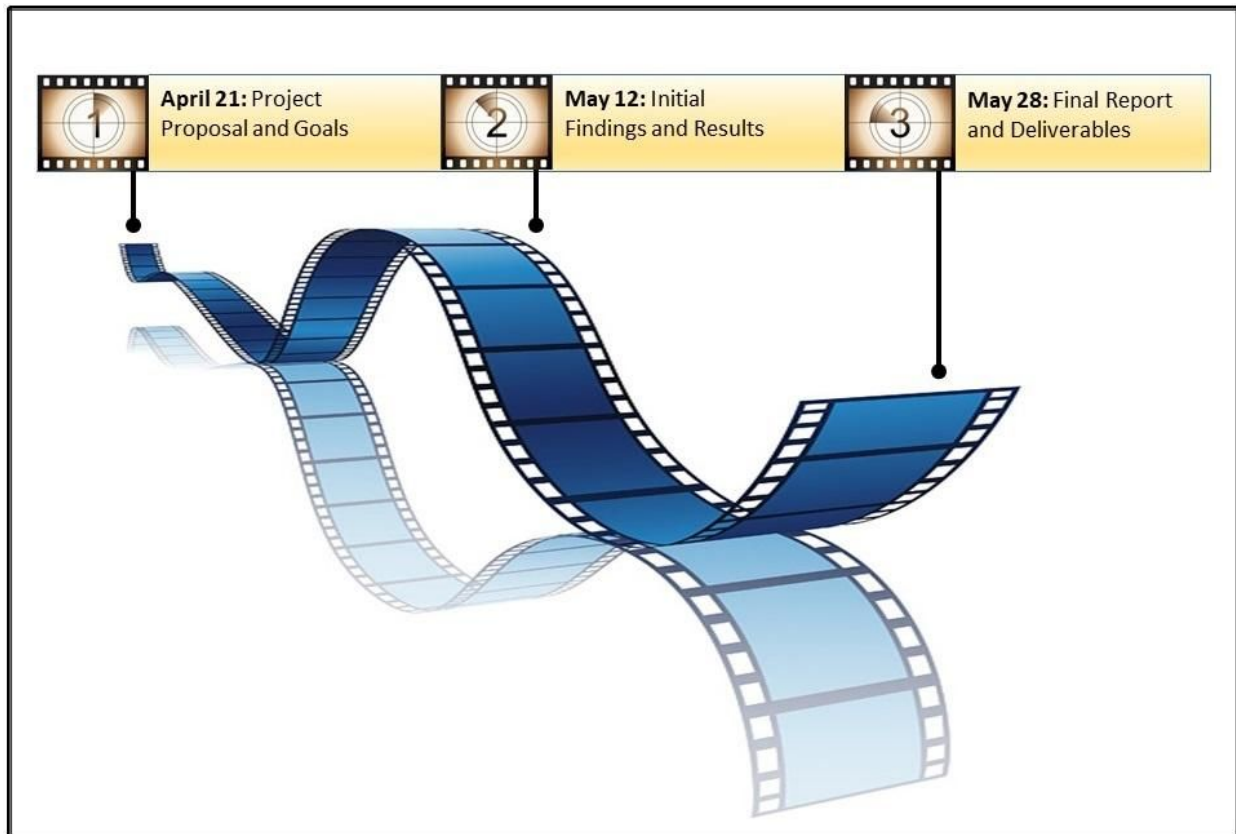


Figure 7. Major Project Milestones

Appendix 1 - Project Plan

| Glen Art Theater Project Timeline | Wed | Thu | Fri | Sat | Sun | Mon | Tue | Wed | Thu |
|--|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| Task | 17-Apr | 18-Apr | 19-Apr | 20-Apr | 21-Apr | 22-Apr | 23-Apr | 24-Apr | 25-Apr |
| Complete first draft of project goals | | | | | | | | | |
| Team meeting | | | | | | | | | |
| Edit project goals draft | | | | | | | | | |
| Finalize goals document and submit for review | | | | | | | | | |
| Identify and finalize data sources | | | | | | | | | |
| Consolidate data to be used for modeling and analysis | | | | | | | | | |
| EDA, data cleaning, etc. | | | | | | | | | |
| Circulate EDA findings | | | | | | | | | |
| Build data dictionary | | | | | | | | | |
| Team meeting | | | | | | | | | |
| Team meeting | | | | | | | | | |
| Prepare EDA documentation and visualizations | | | | | | | | | |
| Draft initial findings report & executive summary | | | | | | | | | |
| Model building and testing | | | | | | | | | |
| Prepare model build documentation for appendix | | | | | | | | | |
| Team meeting | | | | | | | | | |
| Edit initial findings report & executive summary | | | | | | | | | |
| Finalize initial findings report & executive summary and submit for review | | | | | | | | | |
| Select model(s) to finalize and validate results | | | | | | | | | |
| Select slate of films to predict on | | | | | | | | | |
| Team meeting | | | | | | | | | |
| Refresh data for selected slate (release dates/YouTube views/rating/run time/etc.) | | | | | | | | | |
| Test and validate selected models with updated data | | | | | | | | | |
| Develop optimization model based on selected theater(s) screen count and operating hours | | | | | | | | | |
| Develop dashboard design | | | | | | | | | |
| Develop mobile app | | | | | | | | | |
| Team meeting | | | | | | | | | |
| Mobile app testing | | | | | | | | | |
| Prepare draft of final report & executive summary | | | | | | | | | |
| Edit final report & executive summary | | | | | | | | | |
| Finalize and submit final report & executive summary | | | | | | | | | |
| Schedule or record oral report | | | | | | | | | |
| Team meeting | | | | | | | | | |
| Submit or present oral report | | | | | | | | | |

Increasing Revenues for Glen Art Theater Through Analytics

| Glen Art Theater Project Timeline | Fri | Sat | Sun | Mon | Tue | Wed | Thu | Fri | Sa |
|--|--------|--------|--------|--------|--------|--------|--------|--------|----|
| Task | 10-May | 11-May | 12-May | 13-May | 14-May | 15-May | 16-May | 17-May | 18 |
| Complete first draft of project goals | | | | | | | | | |
| Team meeting | | | | | | | | | |
| Edit project goals draft | | | | | | | | | |
| Finalize goals document and submit for review | | | | | | | | | |
| Identify and finalize data sources | | | | | | | | | |
| Consolidate data to be used for modeling and analysis | | | | | | | | | |
| EDA, data cleaning, etc. | | | | | | | | | |
| Circulate EDA findings | | | | | | | | | |
| Build data dictionary | | | | | | | | | |
| Team meeting | | | | | | | | | |
| Team meeting | | | | | | | | | |
| Prepare EDA documentation and visualizations | | | | | | | | | |
| Draft initial findings report & executive summary | | | | | | | | | |
| Model building and testing | | | | | | | | | |
| Prepare model build documentation for appendix | | | | | | | | | |
| Team meeting | | | | | | | | | |
| Edit initial findings report & executive summary | | | | | | | | | |
| Finalize initial findings report & executive summary and submit for review | | | | | | | | | |
| Select model(s) to finalize and validate results | | | | | | | | | |
| Select slate of films to predict on | | | | | | | | | |
| Team meeting | | | | | | | | | |
| Refresh data for selected slate (release dates/YouTube views/rating/run time/etc.) | | | | | | | | | |
| Test and validate selected models with updated data | | | | | | | | | |
| Develop optimization model based on selected theater(s) screen count and operating hours | | | | | | | | | |
| Develop dashboard design | | | | | | | | | |
| Develop mobile app | | | | | | | | | |
| Team meeting | | | | | | | | | |
| Mobile app testing | | | | | | | | | |
| Prepare draft of final report & executive summary | | | | | | | | | |
| Edit final report & executive summary | | | | | | | | | |
| Finalize and submit final report & executive summary | | | | | | | | | |
| Schedule or record oral report | | | | | | | | | |
| Team meeting | | | | | | | | | |
| Submit or present oral report | | | | | | | | | |



Increasing Revenues for Glen Art Theater Through Analytics

| Glen Art Theater Project Timeline | Sat | Sun | Mon | Tue | Wed | Thu | Fri | Sat | Sun |
|--|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| Task | 1-Jun | 2-Jun | 3-Jun | 4-Jun | 5-Jun | 6-Jun | 7-Jun | 8-Jun | 9-Jun |
| Complete first draft of project goals | | | | | | | | | |
| Team meeting | | | | | | | | | |
| Edit project goals draft | | | | | | | | | |
| Finalize goals document and submit for review | | | | | | | | | |
| Identify and finalize data sources | | | | | | | | | |
| Consolidate data to be used for modeling and analysis | | | | | | | | | |
| EDA, data cleaning, etc. | | | | | | | | | |
| Circulate EDA findings | | | | | | | | | |
| Build data dictionary | | | | | | | | | |
| Team meeting | | | | | | | | | |
| Team meeting | | | | | | | | | |
| Prepare EDA documentation and visualizations | | | | | | | | | |
| Draft initial findings report & executive summary | | | | | | | | | |
| Model building and testing | | | | | | | | | |
| Prepare model build documentation for appendix | | | | | | | | | |
| Team meeting | | | | | | | | | |
| Edit initial findings report & executive summary | | | | | | | | | |
| Finalize initial findings report & executive summary and submit for review | | | | | | | | | |
| Select model(s) to finalize and validate results | | | | | | | | | |
| Select slate of films to predict on | | | | | | | | | |
| Team meeting | | | | | | | | | |
| Refresh data for selected slate (release dates/YouTube views/rating/run time/etc.) | | | | | | | | | |
| Test and validate selected models with updated data | | | | | | | | | |
| Develop optimization model based on selected theater(s) screen count and operating hours | | | | | | | | | |
| Develop dashboard design | | | | | | | | | |
| Develop mobile app | | | | | | | | | |
| Team meeting | | | | | | | | | |
| Mobile app testing | | | | | | | | | |
| Prepare draft of final report & executive summary | | | | | | | | | |
| Edit final report & executive summary | | | | | | | | | |
| Finalize and submit final report & executive summary | | | | | | | | | |
| Schedule or record oral report | | | | | | | | | |
| Team meeting | | | | | | | | | |
| Submit or present oral report | | | | | | | | | |

Appendix 2 - Data Dictionary

| Field Name | Description of Field Information |
|-----------------------|--|
| (unnamed) | Key value for full entry |
| Budget | Dollar value for production of film |
| Id | ID value assigned to film |
| original_title | Title at time of original release (i.e. Star Wars, not Star Wars: Episode IV: A New Hope) |
| popularity | |
| revenue | Dollar value of tickets sold internationally, unadjusted |
| runtime | Total minutes |
| vote_average | Average rating provided by reviewers |
| vote_count | Total ratings provided |
| weighted_rating | Adjusted rating based on IMDB calculation |
| cast_size | Number of credited on screen contributors |
| crew_size | Number of credited technical contributors |
| revenue_divide_budget | Calculated value of the actual revenue divided by the actual budget |
| movie_count | |
| director_Rank | Rank of director based on lifetime gross attributed to associated films |
| collection_Rank | Rank of film collection based on lifetime gross attributed to associated film collection (i.e. <i>Star Wars</i> , <i>Lord of the Rings</i> , <i>Jason Bourne</i>) |
| prod_Comp_Rank | Rank of production company based on lifetime gross attributed to associated films |
| genres_Rank | Collected genre categorical variable values |
| Genre_0 | Categorical variable used to identify films with no revenue or undefined genre |
| Genre_Action | Categorical variable used to identify action films |
| Genre_Adventure | Categorical variable used to identify adventure films |
| Genre_Animation | Categorical variable used to identify animation films |
| Genre_Comedy | Categorical variable used to identify comedy films |



Increasing Revenues for Glen Art Theater Through Analytics

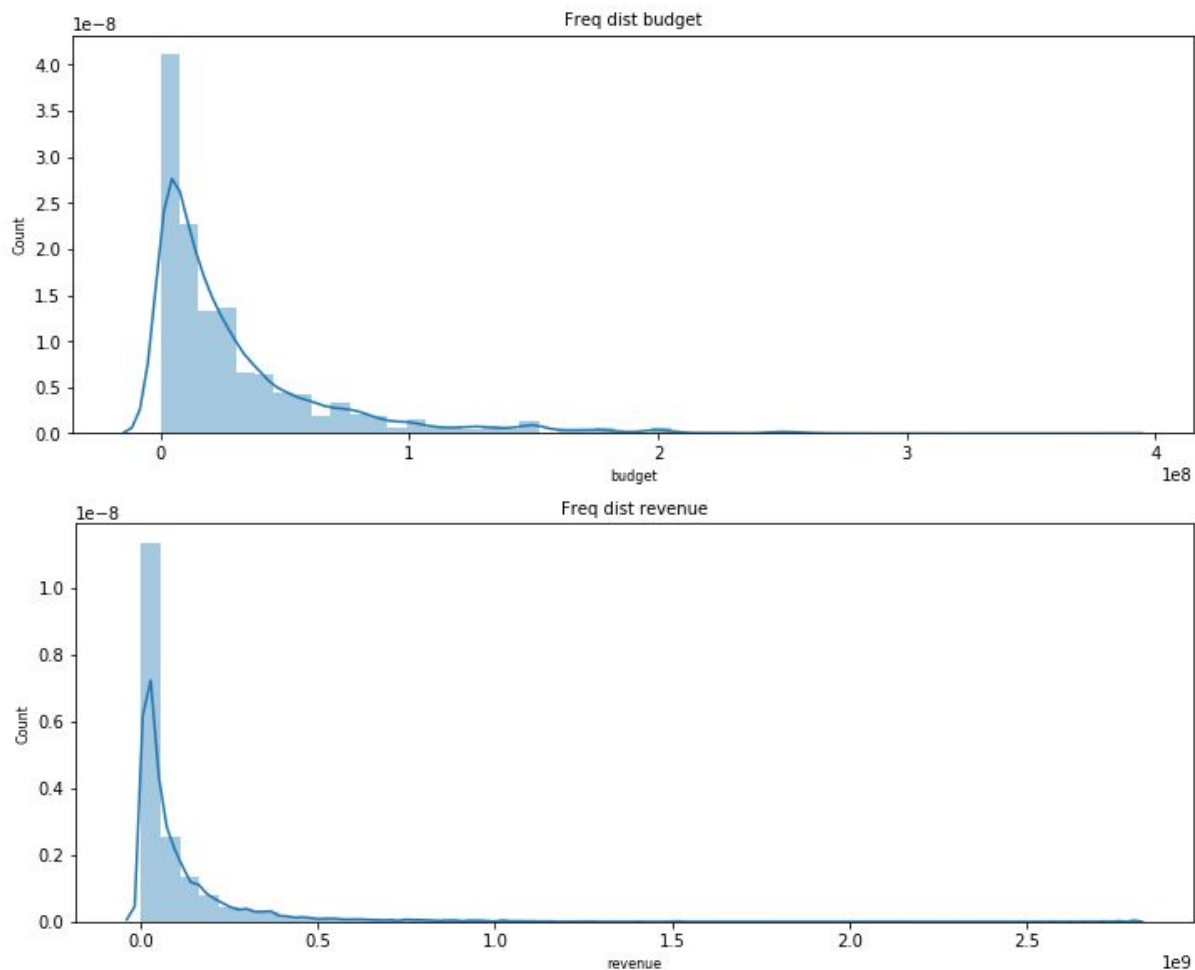
| Field Name | Description of Field Information |
|-----------------------|---|
| Genre_Crime | Categorical variable used to identify crime films |
| Genre_Documentary | Categorical variable used to identify documentary films |
| Genre_Drama | Categorical variable used to identify drama films |
| Genre_Family | Categorical variable used to identify family films |
| Genre_Fantasy | Categorical variable used to identify fantasy films |
| Genre_Foreign | Categorical variable used to identify foreign films |
| Genre_History | Categorical variable used to identify history films |
| Genre_Horror | Categorical variable used to identify horror films |
| Genre_Music | Categorical variable used to identify music films |
| Genre_Mystery | Categorical variable used to identify mystery films |
| Genre_Romance | Categorical variable used to identify romance films |
| Genre_Science Fiction | Categorical variable used to identify science fiction films |
| Genre_TV Movie | Categorical variable used to identify TV movies |
| Genre_Thriller | Categorical variable used to identify thriller films |
| Genre_War | Categorical variable used to identify war films |
| Genre_Western | Categorical variable used to identify western films |



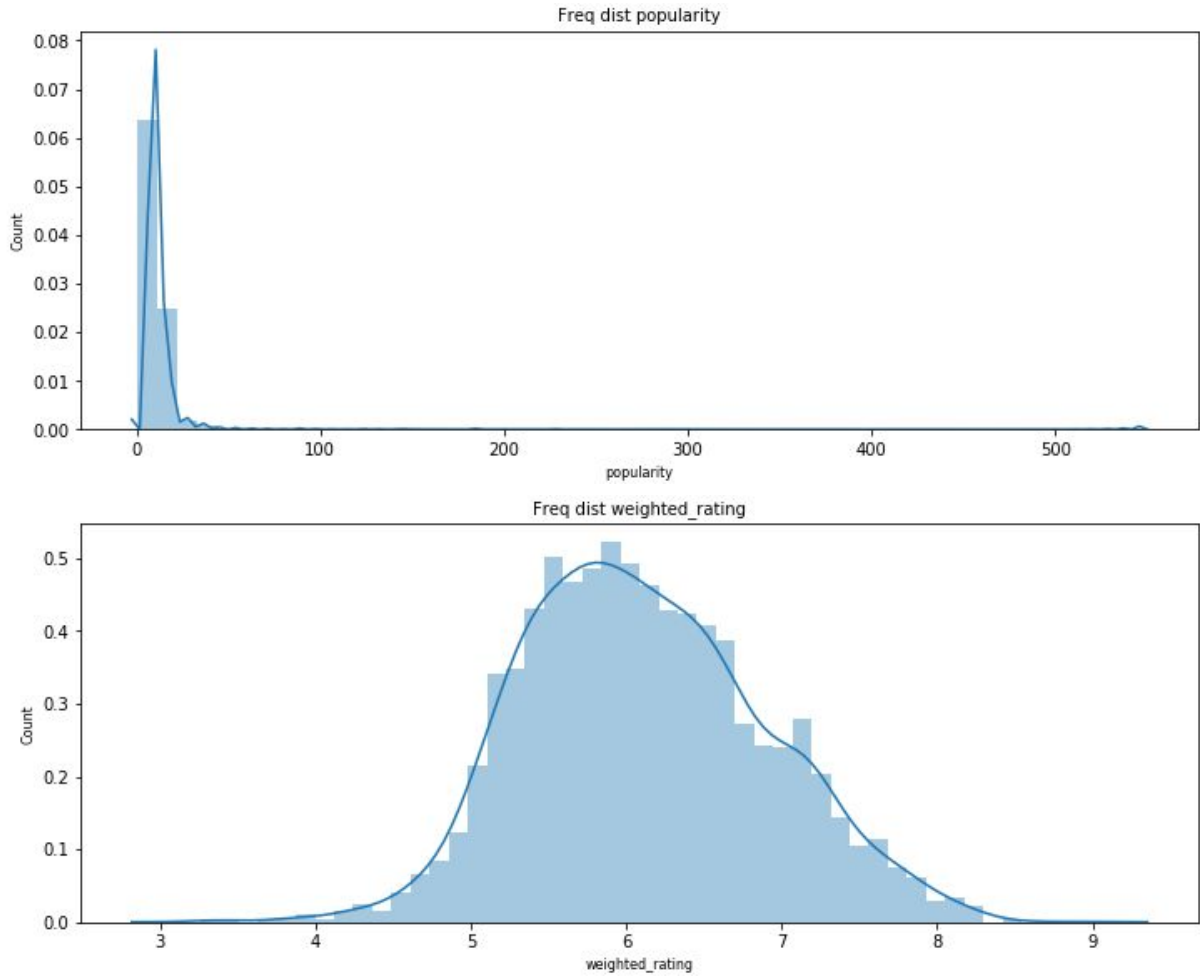
Appendix 3 - EDA Results

Below are a series of data visualizations created during the data analysis process. All are designed to examine the distribution of the available data to ensure there are not any unexpected or unexplained outliers as well as to aid in the decision making process about which combinations of features may provide the most assistance in predicting an individual film's revenue.

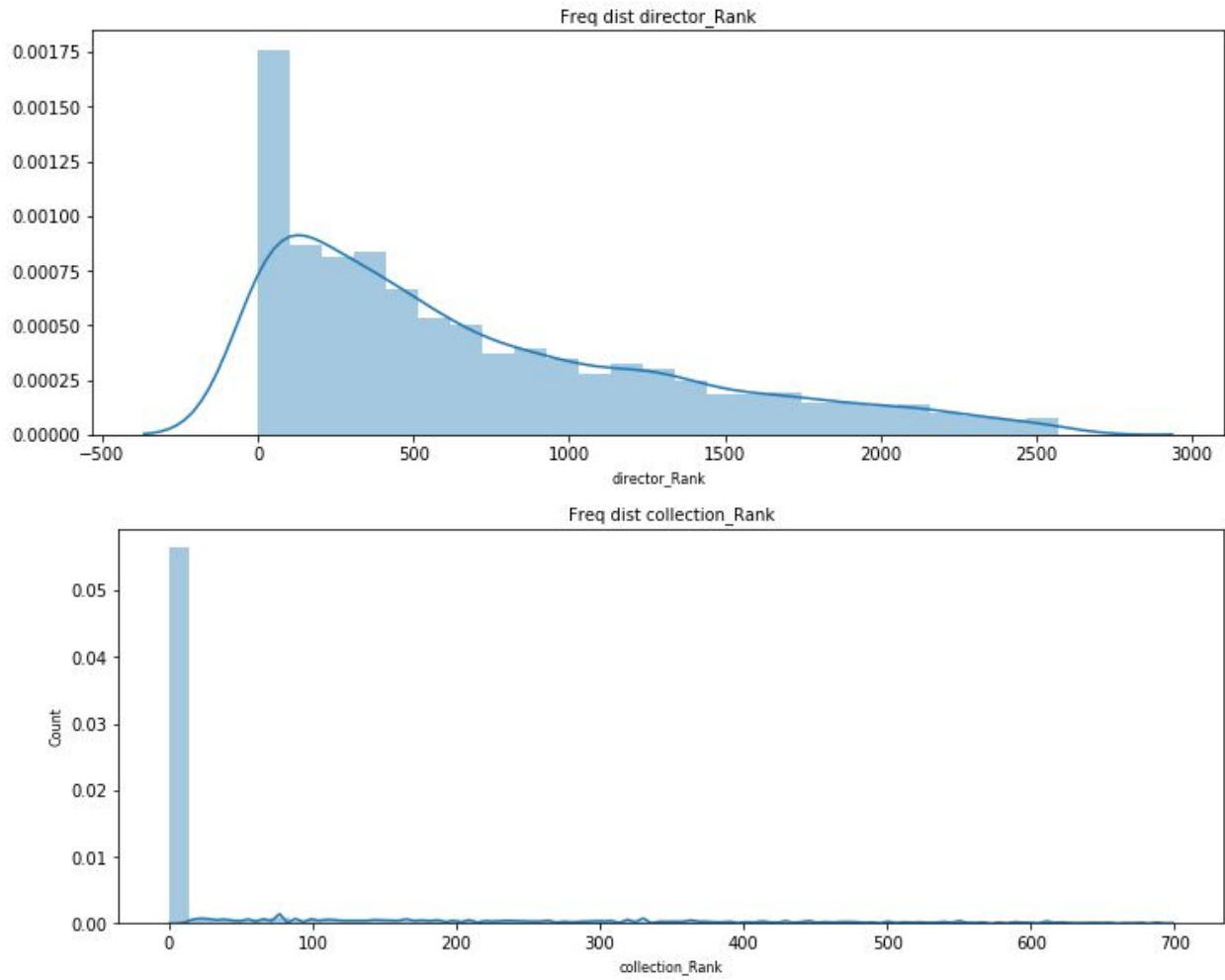
First are a series of distribution charts with a smoothed line overlaid to help determine if the data is normally distributed. In most instances the data is normally distributed but with a right skew. The exceptions are the weighted rank, which was specifically engineered to be as close to a normal distribution as possible, and the genre rank which is nearly a flat line but for the lowest values which is indicative of relatively small number of high grossing film collections.



Increasing Revenues for Glen Art Theater Through Analytics

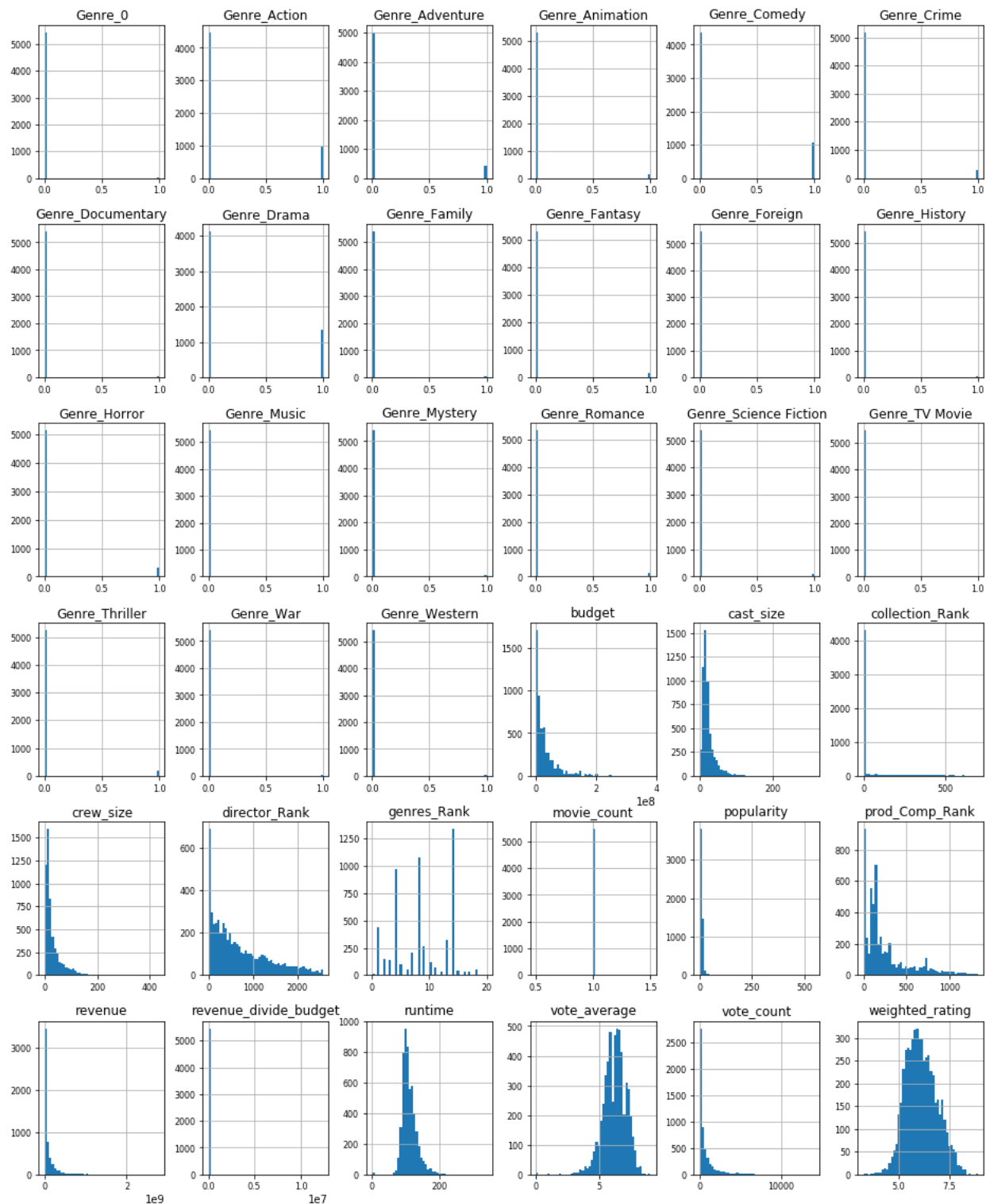


Increasing Revenues for Glen Art Theater Through Analytics



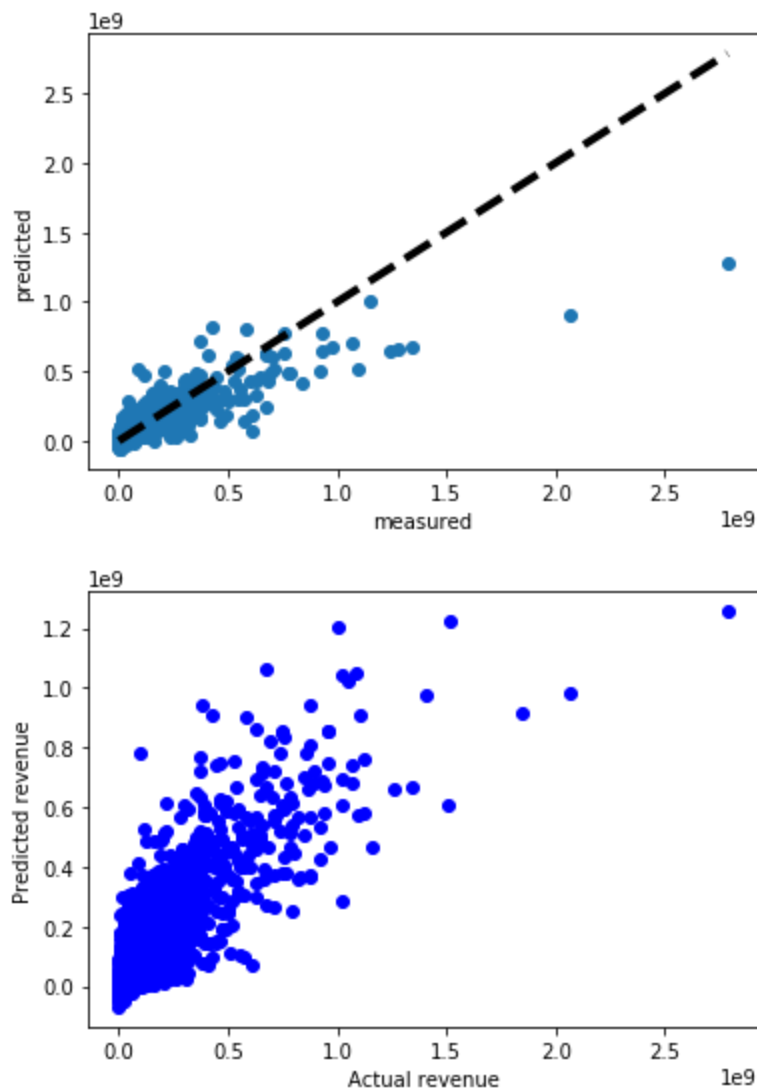
Increasing Revenues for Glen Art Theater Through Analytics

The next set of charts are histograms of all of the in scope features. These are again used to visually confirm the normal distribution of each feature. The obvious exceptions that do not have normal distributions are the genre specific categorical variables which are binary data points indicating either a film is or is not categorized to a particular genre.



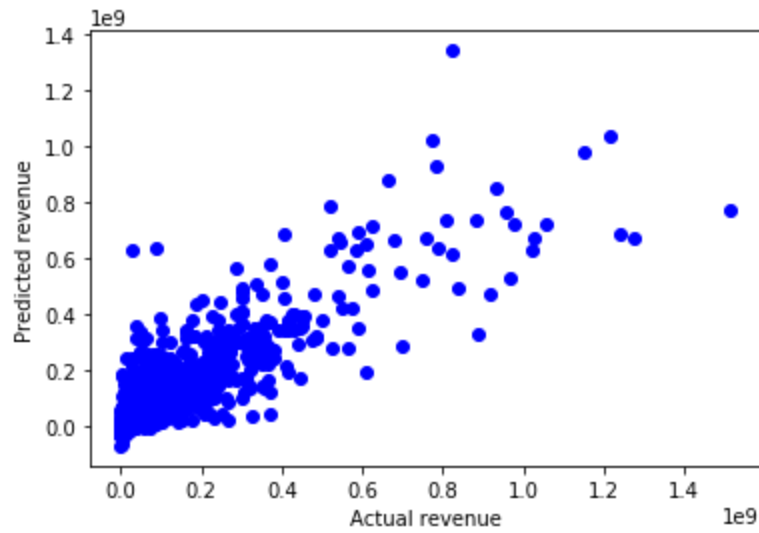
Increasing Revenues for Glen Art Theater Through Analytics

A set of scatter plots for the intersection of different variables was created in order to examine specific relationships and how well they are distributed as well.



Actual vs predicted revenue based on linear regression model using Tensorflow

Increasing Revenues for Glen Art Theater Through Analytics



Increasing Revenues for Glen Art Theater Through Analytics

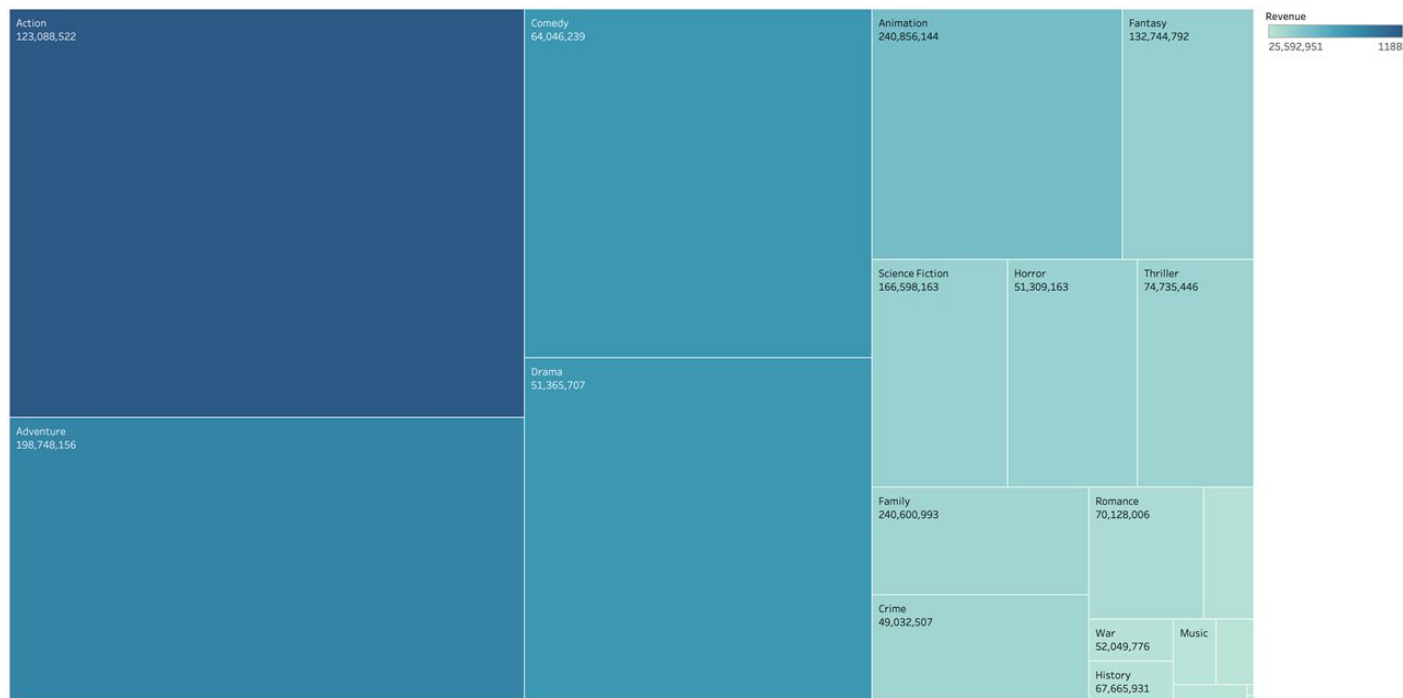
Finally, a set of visualizations was used to further examine the attributes of specific fields such as average profit, revenue, average ratings, and weighted rating.

Bubble by Avg Profit



Genre and average of Revenue. Color shows details about Clusters. Size shows average of Profit. The marks are labeled by Genre and average of Revenue. Details are shown for Genre.

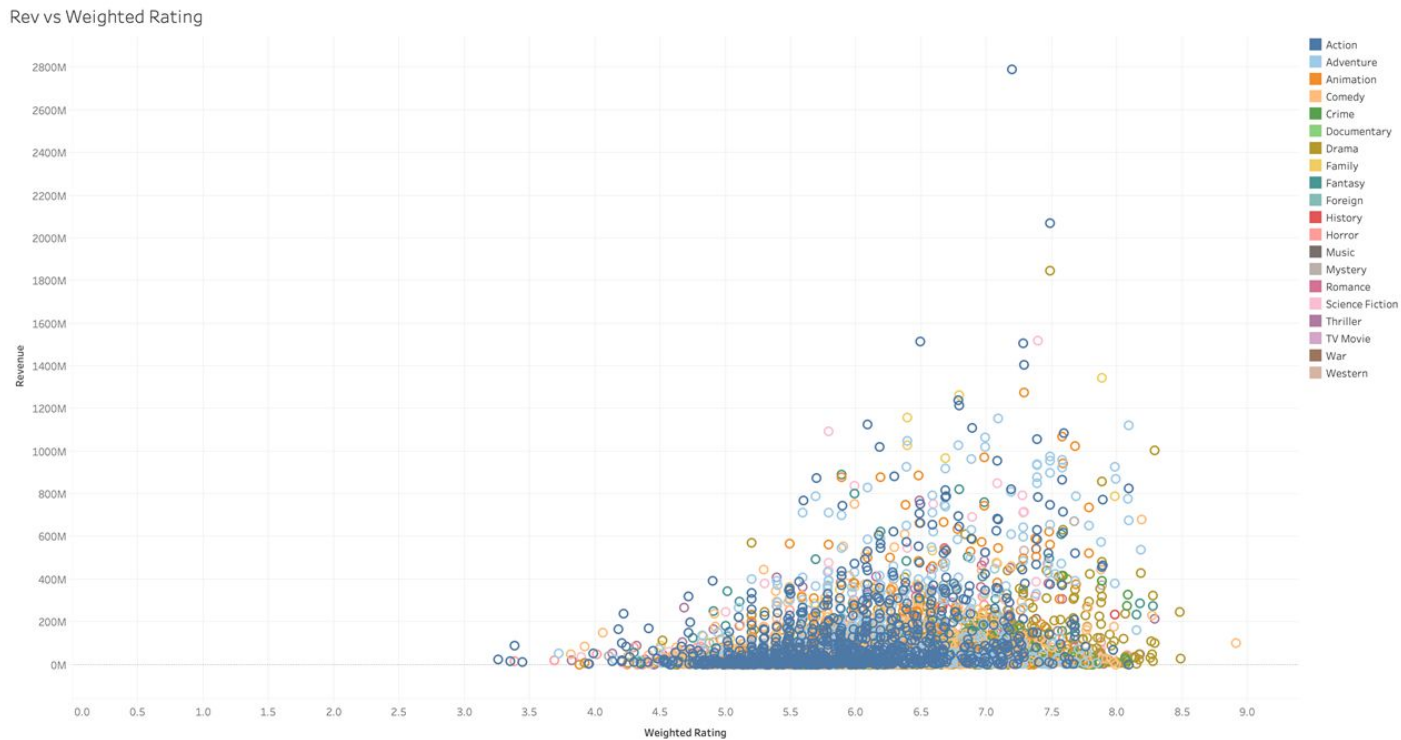
Treemap sorted on total Revenue



Genre and average of Revenue. Color shows sum of Revenue. Size shows sum of Revenue. The marks are labeled by Genre and average of Revenue.



Increasing Revenues for Glen Art Theater Through Analytics



best rev movies (2)



Title, sum of Revenue Divide Budget and Genre. Color shows sum of Revenue Divide Budget. Size shows sum of Revenue. The marks are labeled by Title, sum of Revenue Divide Budget and Genre. The view is filtered on Title, which has multiple members selected.



Increasing Revenues for Glen Art Theater Through Analytics



Appendix 4 - Modeling Methods

While all machine learning methods attempt to provide a prediction or classification for the desired output, they all approach the task in slightly different ways. Below is a description of the modeling methods used to date.

Ordinary Least Squares (OLS) Regression

Using a regression line that minimizes the sum of the squared values of the errors associated with actual values vs regression predicted values. This aggregated regression line can then provide a prediction for what average revenue value would result using any known or estimated film budget.

Linear Regression

The most common predictive tool used, the idea here is similar to OLS in that a regression line is generated to represent the most likely solution based on explanatory variable inputs to generate a response, or dependent, variable. If more than one explanatory variable is used, then the methodology is called multiple linear regression.

Polynomial Regression

When distribution of data is more complex than a straight line, it can be considered to be non-linear. Polynomial regression is a method to adapt linear regression methods to non-linear data in an effort to improve the fit of the model. Through the use of different model based transformations of the underlying data, the curve of the resulting regression line is manipulated in an effort to find a better fit for the data. A 2nd degree transformation would add a single curve to the regression line, 3rd degree adds a second curve, 4th a third, and so on.

Least Absolute Shrinkage and Selection Operator (Lasso) Regression

Lasso regression is another form of linear regression where the model attempts to shrink the data values in such a way as to simplify the coefficients used in the modeling process in order to make a complex set of data more sparse and therefore potentially achieve better results. The shrinkage that is applied is akin to enabling the model to perform optimized feature selection and is typically most useful when there is significant multicollinearity that exists within a given data set.

Ridge Regression

Another linear regression option when multicollinearity exists within a data set, ridge regression uses a regularization function to better fit the available data. By applying a consistent cost function to all features the goal is to optimize the space that exists



Increasing Revenues for Glen Art Theater Through Analytics

between a standard linear regression line and a line representing the data's mean in order to circumvent any influence that the existing multicollinearity may have on the regression model. The cost function is only applied during the training of the model and its effectiveness is then measured against unregularized prediction data via measurements such as recall, precision, or R^2 .

Support Vector Machines (SVM)

Support Vector Machines (SVM) are a powerful machine learning tool that is optimized for use when data sets are small to mid-sized yet complex. SVM models are capable of completing linear or non-linear classifications, regressions, and outlier detection. When classifying data, the model determines an optimal hyperplane, or dividing line, to separate the data into distinct classes. The optimum hyperplane will not just separate the data, but will find the point that the resulting classes exist as far away from that dividing line as possible. Similarly, if using SVM for regression, the model will attempt to adapt the regression line in order to fit as many data points as possible on or close to it in order to reduce predictive error.

Random Forest Classifier

Random Forest Classifiers are an ensemble set of decision trees meant to create an orderly classification mechanism for target data. Using this method ensures that with every branch created in the decision tree, all of the predictor variables are re-considered for the next decision point. This enables a computationally sleek mechanism to generalize the data based on the available data.

Deep Learning AI via TensorFlow

Using TensorFlow as the machine learning framework, deep learning is a method for the model to parse through and test variations of, in this instance, linear regression models based on a large number of features. Through the manipulation of any number of hyperparameters the AI process can then arrive at a result that is optimized based on the established hyperparameters. Hyperparameters can include the number of layers, or rounds of learning, nodes, and tensors that the algorithm process through in each epoch. While not the same as automatic feature selection, through an appropriate number of nodes, for instance, unimportant or covariant features can be minimized without wholly excluding them.



Appendix 5 - Optimization Algorithm

As a complement to the revenue forecasting models, FilMetrics is developing a screen scheduling algorithm to assist Glen Art Theater in maximizing expected profits given operational constraints. The forecasting model answers the question: *If a specific movie were shown on a specific day at a particular time, how many tickets would be sold?* This scheduling algorithm answers the question: *Given this forecast, what is the optimal lineup of show times for multiple movies on multiple screens that is still compatible with theater operations?* The remainder of this appendix provides the details of our solution algorithm.

Weighted interval scheduling (WIS) belongs to a category of dynamic programming algorithms which seek to break a larger problem down into several smaller subproblems, solve these subproblems recursively, and then combine their solutions to get the final solution. Specifically, in the weighted interval scheduling problem, we want to find the maximum-weight subset of non-overlapping jobs given a set of jobs that have weights associated with them. Each job has a start time, a finish time and a weight. In the Glen Art Theater scenario, jobs are replaced by potential showtimes i for a given movie selected from the set of available movie showtimes I . We must consider every possible start time s_i and end time f_i for each selected film such that it falls within the allotted screening window. The weight then becomes the forecasted demand or revenue d_i should a specific film be shown at a specific time. This optimal subset O of non-overlapping show times with the maximum possible demand can be written mathematically as,

$$O = \max_{O \subset I; \forall i \in O: f_{i-1} \leq s_i} \sum_{i \in O} d_i$$

To illustrate this algorithm optimizing the schedule for a single screen on a given day, we will make the following assumptions. Assume that Glen Art Theater begins showing movies at 12:00 pm and management decides that all showings must be finished by 12:00 am in order for all cleaning and closing processes to be completed by 12:30 am. Additionally, movies can potentially start every five minutes (i.e., no showings start at a time other than XX:X0 or XX:X5). Glen Art Theater has three movies to show which include *Avengers: Endgame*, *Amazing Grace* and *The Mustang*. The run times for these films are 181 minutes, 87 minutes and 96 minutes, respectively. The time allotted each film for advertisements, trailers and cleaning, as well as the total durations are 30 minutes, 20 minutes and 20 minutes, and 211 minutes, 107 minutes and 116 minutes, respectively.

The possible start and end times for each film can be obtained by converting these total durations into five-minute time periods and running a sliding window of this length through the screening window. The pseudo-code to this point is shown below.



Increasing Revenues for Glen Art Theater Through Analytics

```
# Run times
runtimes <- c(181, 87, 96)

# Advertisements, trailers and cleaning times
addon <- c(30, 20, 20)

# Total durations (round up)
durations <- ceiling((runtimes + addon)/5)

# Possible start and end times for each movie
showtimes <- c()
for (i in 1:length(durations)) {
  s <- 1:(length(screeningWindow) - durations[i] + 1)
  f <- s + durations[i] - 1
  x <- cbind(s,f)
  rownames(x) <- rep(ifelse(i == 1, "Avengers.Endgame",
                             ifelse(i == 2, "Amazing.Grace", "The.Mustang")),
                    nrow(x))
  showtimes <- rbind(showtimes, x)
}
```

The result is 349 possible showtimes between the three films for the day. The next piece that is needed is the demand forecast for each one of these 349 show times, which is obtained via the predictive modeling work FilMetrics is also undertaking. Then, we order each of these show times from earliest to latest finish time.

The WIS solution algorithm begins by initializing both a demand vector equal to the demand column from the show time matrix as well as an index list with elements corresponding to each row index from the show time matrix. We iterate using nested *for* loops, beginning with $i=2$ and $j=1$ to determine if show time i and j are compatible or not. If they overlap, we increase j until we reach $j=i-1$, then we increase i by 1 and start over with $j=1$. If the show times do not overlap, we store the value $\max(d[i], d[j] + \text{showtimes}[i, "d"])$ in place of $d[i]$, where d is the initialized demand vector and $\text{showtimes}[i, "d"]$ is the forecasted demand of show time i . We use the initialized index list to store the indices of which non-overlapping show times together produce the largest demand. The algorithm pseudo-code is shown below.

Increasing Revenues for Glen Art Theater Through Analytics

```
# Sort by finish time
showtimes <- showtimes[order(showtimes[, "f"]),]

# Initialize maximum demand vector
d <- unname(showtimes[, "d"])

# Initialize list of movie start times to get demands in d
ind <- as.list(1:nrow(showtimes))

# Calculate maximum demand
for (i in 2:nrow(showtimes)) {
  d_i <- d[i]
  ind_j <- c()
  for (j in 1:(i-1)) {
    if (showtimes[j, "f"] <= showtimes[i, "s"]) {
      if (max(d[i], d[j] + showtimes[i, "d"]) > d_i) {
        d_i <- max(d[i], d[j] + showtimes[i, "d"])
        ind_j <- ind[[j]]
      }
    }
  }
  d[i] <- d_i
  ind[[i]] <- sort(c(ind[[i]], ind_j))
}
```

The optimal schedule is then found by looking up the largest value from our initialized demand vector, which has now had its values replaced by maximum demands from combined showtimes. We can find this same location in our index list, and grab those show times from our show time matrix. A sample optimal schedule for a single screen based on simulated demands is shown in the table below. (Note: start time and end times include advertisements, trailers and cleaning times.)

| Film | Start Time | End Time |
|--------------------------|------------|----------|
| <i>Amazing Grace</i> | 12:00 | 13:45 |
| <i>Amazing Grace</i> | 13:50 | 15:35 |
| <i>The Mustang</i> | 16:05 | 18:00 |
| <i>Avengers: Endgame</i> | 18:00 | 21:30 |
| <i>Amazing Grace</i> | 21:50 | 23:35 |

We see that the algorithm has produced a schedule that starts and ends within the allotted screening window. We can also observe that movies do not always immediately begin at the soonest available start time. For example, *The Mustang* does not begin until 16:05 when



Increasing Revenues for Glen Art Theater Through Analytics

Amazing Grace finished at 15:35. This is because Glen Art Theater can capitalize on higher forecasted demand by waiting. Even a trained film scheduler would have immense difficulty tracking all of the possible movie combinations and associated demands, highlighting the power with which this algorithm supplies theater management.

