Predicting Movie Box Office and Virtual Stock Market Price

Olivia Fan (Runtime: ~ 1 min)

1. Introduction

With the advent of rapid digitization, the movie industry has encountered an explosive growth of greater than 1000 movies produced per year; consequently, it becomes a crucial concern to investors whether the movie succeeds (Bhave et al. 2015). This zealous growth subsequently gives rise to virtual stock markets (VSMs), the world's largest of which, established in 1996 is the Hollywood Stock Exchange where unlimited number of consumers can trade thousands of entertainment securities (Karniouchina 2011). It comes to the fact that not only the success of the movie itself is at stake, but the inextricable relations of the movie's success with the VSM stock price could give rise to numerous hedging implications that would allow investors to make statistically informed decisions.

While extensive literature has constructed models predicting movie box office, the assessment of which in light of virtual market proves to be a fairly understudied domain recently. Older studies have found that despite arbitrage opportunities in VSMs, the predictive power of HSX is quite high (Karniouchina 2011). Within the movie box office models, a considerate amount of work relies on methods that lack interpretability, such as multi-layer back propogation neural network and ensemble learning (Lee et al. 2018). Researchers (Bhave et al. 2015) point out that accuracy can be improved by incorporating social factors on various online platforms, in addition to classical intrinsic factors of the movie itself. Therefore, this study aims to gauge insights into significant predictors of this multi-layered relationship with recent data (~2020) via statistical methods with greater interpretability such as ARIMA, Baysian Model Averaging and decision tree.

Study Design

The aim of this multifaceted study is three-fold, which contains three inextricably intertwined complex objectives. First, we would like to analyze factors that predict movie box office. Secondly, we would like to analyze factors that predict virtual market stock prices. Finally, we would also like to assess whether virtual markets are efficient predictors of new product success, with manifestation in box office. On top of this hierarchy of research questions, the nature of this time series data set lends itself to diverse methods such as the ARIMA (autoregressive integrated moving average) model, exponential smoothing, etc. The two sets of predictor variables of interest are (1) movie budget, genre, distributor, release date, number of theaters, MPAA rating, (2) trading volume, total volume held long, total volume held short, and IPO date. The response variables are domestic, international and worldwide box office, and stock price of the 9380 movies.

2. Data & Data Processing

The data in this data set were scraped from two websites ("Hollywood Stock Exchange & Box Office Data" 2021): (1) Hollywood Stock Exchange (HSX.com), the world's leading virtual entertainment market which provides information on movie stock prices, (2) BoxOfficeMojo.com which tracks box-office revenue in a systematic way and provides the information on movie box office. The former HSX data source contains 325,640 daily domestic box office results (1995-2020) which includes the number of theaters exhibiting the movie release on this date and identifier of movie release; it also contains 16,968 movie releases, its identifier, budget, distributor name, domestic gross to date, international gross to date, worldwide gross to date, release date, widest release, genre and MPAA rating. The latter BoxOfficeMojo data source contains master movie data on 9,380 movies from HSX.com., i.e. genre, stock IPO date, release date, delist date, MPAA rating, number of theaters and distributor; it also contains 12,677,219 hourly movie stock prices (1997-2020) from HSX.com, along with total number of shares held short, shares held long and trading volume at the time stamp.

In data processing, we first filtered out extraneous information irrelevant to our analysis, such as the old BoxOfficeMojo id, the BoxOfficeMojo symbol, synopsis of the movie and the BoxOfficeMojo url. To facilitate further analysis, we transformed the dates from characters to the correct format. Because many movies are attributed multiple genres, in order to analyze the impact of genre on the response, we separated the list of genres into separate rows so that each contains one category Then we filtered out missing information such as phase (with over 96% missing) and release pattern (98% missing) in HSX data, as well as domestic opening (every row contains the identical 0) in

Table 1: Top Movies by Box Office to Budget Ratio

title	ratio	genres
Paranormal Activity	12890.3867	Horror, Mystery, Thriller
The Blair Witch Project	4143.9850	Horror, Mystery
Tarnation	2690.9727	Biography, Dcoumentary
The Gallows	429.6441	Horror, Mystery, Thriller

BoxOfficeMojo data. In order to investigate the relationship between distributor name and movie box office, we filtered out distributors with fewer than 10 movie releases and obtained 15,112 data records for 3,004 movie releases from 107 unique distributors. Likewise, we filtered out the News genre since it has only 1 observations, and obtained 3,003 movie releases from 21 genres.

3. Exploratory Data Analysis

Objective 1: Predicting Movie Box Office

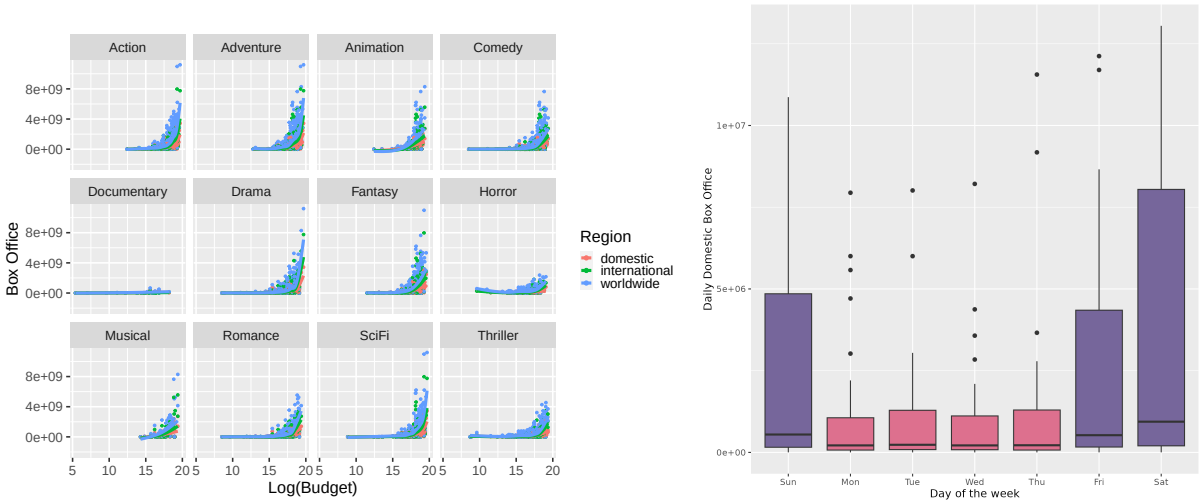


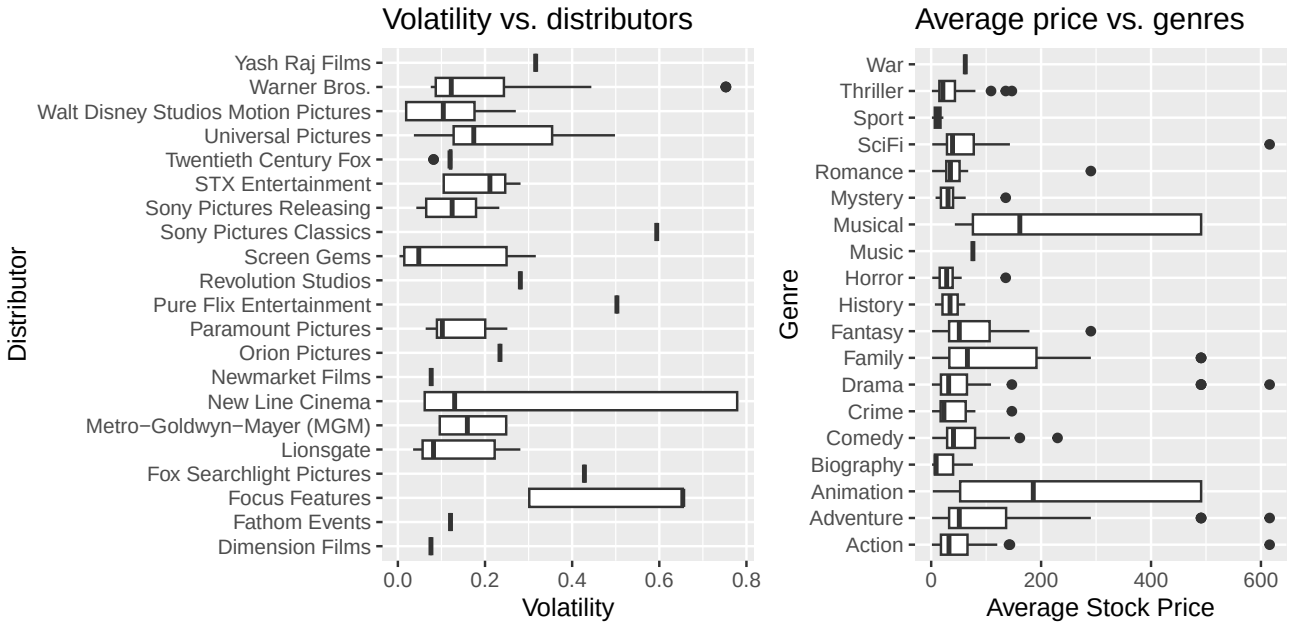
Figure 1: (a) Relationship between Budget and Box Office by Region across Genres, (b) Box Office by Day of the Week

We observe that the relationship between budget and box office (Figure 1(a)) is vastly different across genres: Action, adventure, drama and sci-fi movies have a steep slope and generally high budget spans, with outliers which have exceedingly high budget and high box office. On the other hand, genres such as horror, thriller and romance have a much flatter slope, which corresponds to the industry knowledge that certain genres are more conducive to low-budget film making than others. According to New Review of Film and Television Studies (2011), horror and thriller movies are typically associated with such framework, and our further analysis corroborates this insight. By calculating the box office to budget ratio (Table 1), we found that three out of the top 4 movie in terms of this cost effectiveness are horror or thriller movies.

We also observe in Figure 1(b), the daily domestic movie box office for the movie Titanic as an example, that the day in the week has a significant effect on the box office, with the weekend (Friday, Saturday and Monday) associating with noticeably higher box office.

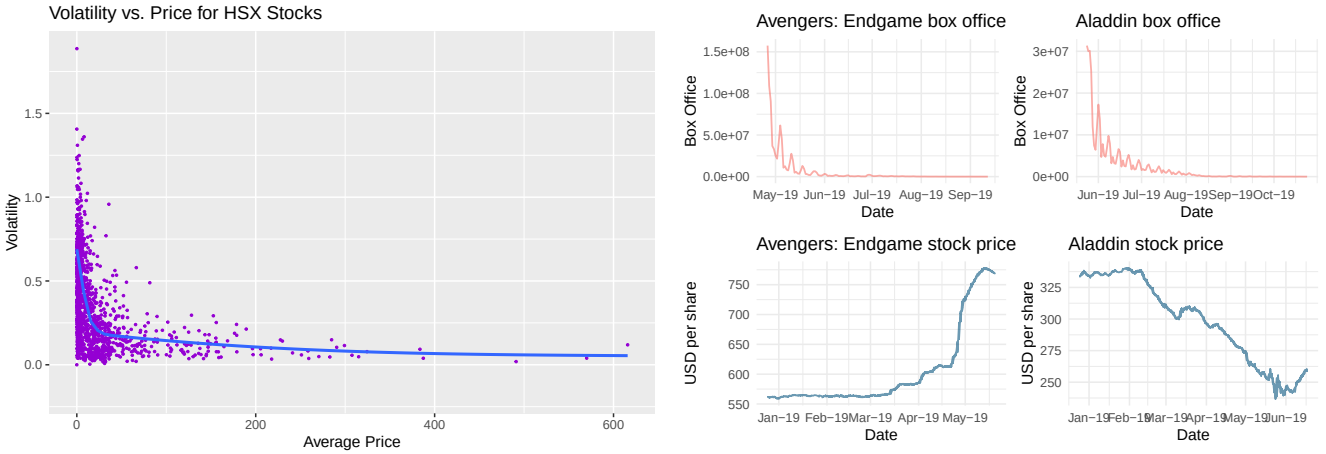
Objective 2. Predicting Stock Price

The second objective is again two fold. We first want to predict (1) the price, and (2) the volatility of the stock in order to both gauge insights into its average performance, and assess the risk in the investment. We measure the stock price as the average over time. To measure volatility, we use the standard deviation of its prices over time to quantify the rate of fluctuations, as suggested by the Corporate Finance Institute (CFI 2023).



We observe distributor's significant impact on volatility: As we might have expected, the stock of the "big name" production companies such as Walt Disney, Sony Pictures, Fox and Warner Brothers tend to have low volatility and tight range; Whereas, New Line Cinema has exceedingly variable volatility spanning from under 0.1 to around 0.8. Similarly we observe genre's impact on stock price: Musical and Animation tend to have the highest stock prices albeit a wide range, whereas history, sport and horror movies have consistently low average stock prices.

Objective 3. Relating virtual markets to product success



We observe an inverse association between average price and volatility. Comparing the time series movement of box office and stock price for movies Avenger and Aladdin, we notice that for both movies, significant movements in stock price (albeit different directions, Avenger rose in stock price while Aladdin declined) preceded significant declines in box office, hinting the predictive power of VSM on box office.

4. Aims & Hypotheses

Aim 1 What are the factors that affect the fluctuations of movie box office over time? Specifically, does the daily theater count or widest release correlate more strongly with the oscillations in movie box office - in other words, does a movie achieve success through continuous rapport or does a transient success suffice? How are budget, genre, runtime and distributor associated with a movie's box office? Do daily theater count correlate with, or wildest release?

Aim 2 What are the factors that affect HSX stock average price, and volatility over the span of time?

Aim 3 To what extent does the HSX stock prices predict the movie box office?

Primary Hypothesis The daily theater count correlates more strongly with daily movie box office than the widest release. While genre, and distributor have a significant effect on the box office, the budget and runtime do not have significant effect.

Secondary Hypothesis The HSX stock average price correlates negatively with the volatility over the span of time, and HSX stock prices move synchronously with fluctuations in box office.

5. Baseline Univariate Model: ARIMA

To establish a baseline model for movie box office over time independent of the covariates, we first examine a Auto-Regressive Integrated Moving Average (ARIMA) model, also known as Box-Jenkins approach (Kotu and Deshpande 2019). As a combination of two models, the auto-regressive and the moving average models, the ARIMA model helps us predict the future forecast via lagged observations and an integrated moving average. We take the time series box office of the movie Deep Sea as an example, and visualize the time series data:

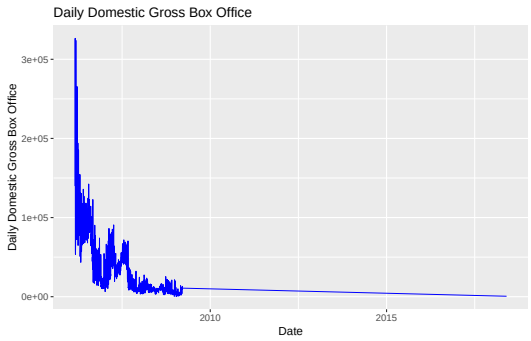


Figure 2: Time Series Visualization for Deep Sea Box Office

In order to perform any successive modeling, by model assumption, requires data to be stationary. That is, the mean, variance, and covariance of the series should be constant with respect to time, and there should not be white noise. Therefore, we take the difference between the log value of the daily domestic gross box office to stationarize the data, and demonstrate later through the Dickey Fuller Test that the data meets model assumptions (in the section below). By the same token, we can stationarize the HSX stock price data before fitting ARIMA to forecast future prices and perform the model assessment, diagnostics and sensitivity analysis in the following sections.

5.5. ARIMA: Model Assumptions, Sensitivity Analysis & Validation

5.5.1 Model Assumptions

5.5.1.1 Stationary: Dickey-Fuller test We conduct the Dickey-Fuller test to assess the stationary principle: The Dickey-Fuller test returns a p-value of 0.01, resulting in the rejection of the null hypothesis and accepting the alternate, that the data is stationary.

By by stationary it means that the properties of the series doesn't depend on the time when it is captured. A white noise series and series with cyclic behavior can also be considered as stationary series.

5.5.1.2 Univariate We assess box office as the univariate response variable, which aligns with ARIMA's assumptions that data should be univariate, since ARIMA works on a single variable.

5.5.2 Sensitivity Analysis: ACF/PACF There are primarily two hyperparameters in the model that we can tune to perform sensitivity analysis, MA (moving-average) and AR (auto-gression) coefficients. The ACF (Auto-Correlation Function) gives us values of any auto-correlation with its lagged values which will help us determine the number of MA coefficients in our ARIMA model, while the PACF (Partial Auto-Correlation Function) finds correlation of the residuals with the next lag value which helps us identify the number of AR coefficients in our ARIMA model. In the ACF graph below, the curve drops significantly after the first lag, which indicates a moving average component of MA(1). We can tune the MA and AR coefficients to achieve sensitivity analysis.

The standard ARIMA models expect as input parameters 3 arguments, p which standards for the number of lag observations, d which is the degree of differencing, as well as q which is the size of the moving average window. This study will tune the parameters via cross validation with a 70%-30% split, as well as sensitivity analysis.

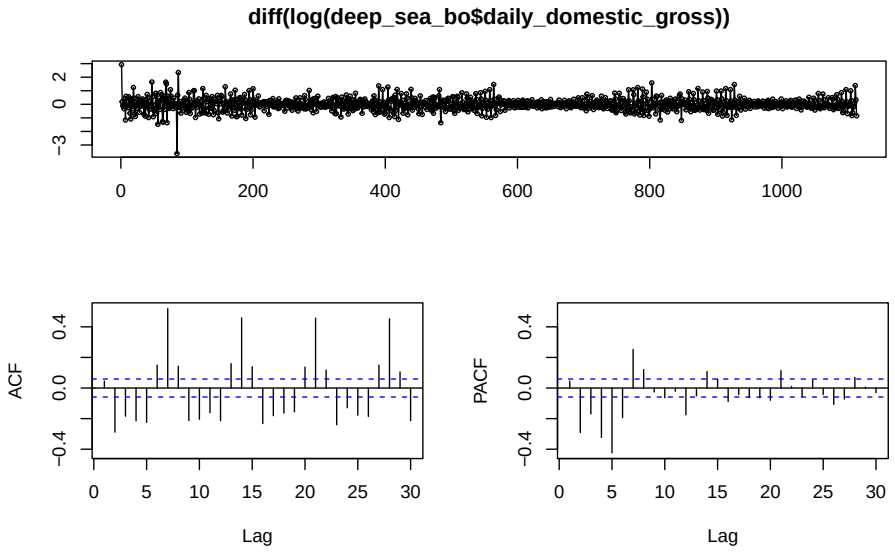


Figure 3: ACF and PACF for the Deep Sea Box Office Model

Additionally, the study also aims to perform sensitivity analysis based on seasonality, which compares the robustness of the model over the seasonal span.

5.5.3 Preliminary Results & Validation As preliminary results, we obtained a with p (AR coefficient) of 5, d (Integrated value) of 0, and q (MA) value of 2 which obtains an AIC value of 759.91, and BIC value of 800.03. The graph above demonstrates that the model is a close fit to the training data. Splitting past data into a training set (pseudo future data), we can examine performance on this pseudo future data to achieve cross validation.

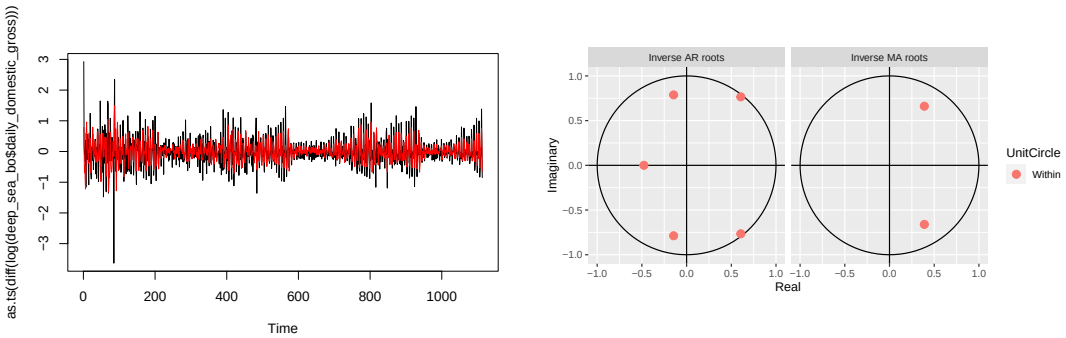


Figure 4: (a) Training Fit for ARIMA Model, (b) Diagnostic Plot for ARIMA Model

5.5.4 Model Diagnostics Plotting the characteristic roots for the model fitted, we see that they are all inside the unit circle, as we would expect because R ensures the fitted model is both stationary and invertible.

6. Multivariate Model: Vector Autoregression (VAR)

While the baseline ARIMA model gauges insights into the changes into the fluctuations of the box office over the span of time in and of itself, since we are essentially interested in the the factors that predict the box office or stock price, we resort to the VAR (Vector Autoregression) model which is essentially a generalization of the univariate autoregressive ARIMA model. A VAR model is a type of multivariate time series model that can capture the

dynamic interactions between multiple time series variables via the assumption that each variable is a function of its own past values as well as the past values of other variables in the system (Stock and Watson 2001). For movie box office, we are primarily interested in genre, distributor, MPAA rating, budget, daily theater count, widest release and runtime which we plan to include in the model. For the former three categorical variables, we plan to use one-hot encoding to convert them to factor levels. Taking the box office data of the movie Titanic as an example, we explore the time series visualization of the quantitative variables below:

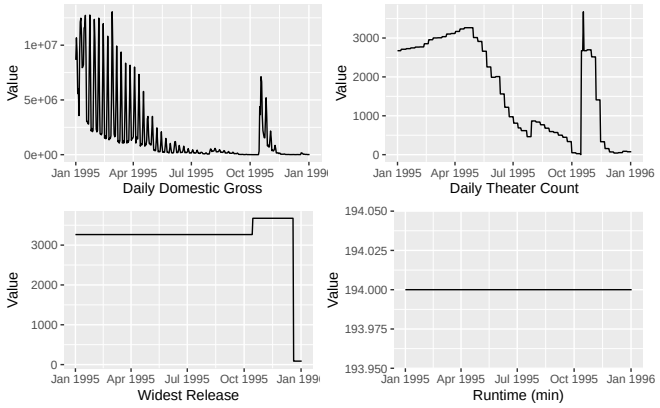


Figure 5: Time Series Plots of Quantitative Variables for Titanic

By the same token, for HSX stock prices, we are primarily interested in genre, distributor, number of shares short, number of shares long, total trading volume which we plan to include in the model. For the former categorical variables, we plan to use one-hot encoding to convert them to factor levels.

6.5 VAR: Model Assumptions, Sensitivity Analysis & Validation

6.5.1 Model Assumptions

6.5.1.1 Stationary Principle In the same token as ARIMA, since VAR is essentially a generalization of ARIMA in the multivariate case, we would also like to assess whether the variables under study are stationary. We use the Philips Perron test to assess the stationary principle, which finds that the response variable (daily domestic gross box office) along with all the predictor variables of interest above (daily theater count, wildest release, and runtime in minutes) having p values of 0.01, 0.05, 0.018 and 0.01 respectively. Therefore, we reject the null hypothesis which suggests that the data is stationary.

6.5.2 Preliminary Results & Validation We fit a preliminary model using the daily theater count, widest release and daily domestic gross box office which obtained an adjusted R^2 value of 0.93. After this, we will select the optimal lag order behind the VAR we will be using, which is 8 from the model output. Lastly, we will run diagnostics tests for autocorrelation, heteroscedasticity, normality and stability. By the same token, we plan to fit the VAR model for stock price using the number of shares long, the number of shares short and the trading volumne as predictors. Splitting past data into training set to create a pseudo future dataset from the dataset given, we plan to examine performance on future data via cross validation.

6.5.4 Diagnostics

6.5.4.1 Non-autocorrelated Residuals We first assess whether our model meets the assumption that the residuals should be non-autocorrelated, based on our assumption that the residuals are white noise and thus uncorrelated with the previous periods. We run the Breusch-Godfrey test for serially correlated errors to obtain a p value of 0.01, therefore see that the residuals do not show signs of autocorrelation. However, in case that is a chance that if we change the maximum lag order, there could be a sign of autocorrelation. Therefore, this study aims to experiment with multiple lag orders which we will confirm through sensitivity analysis.

6.5.4.2 ARCH Effects: Heteroscedasticity Another aspect to consider is the presence of heteroscedasticity, essentially clustered volatility areas in a time series dataset known as ARCH effects, which is common is time series

data such as stock prices where massive rises or declines could be seen (Stock and Watson 2001). Through the ARCH test, we obtain a p value of less than $2.2e^{-16}$ under degrees of freedom of 540, which signifies no degree of heteroscedasticity as we reject the null hypothesis.

6.5.4.3 Normality The VAR normality test has three components: the Jarque-Bera test, the Kurtosis test, and the Skewness test. All of the three tests give us a p value of less than $2.2e^{-16}$. Therefore, based on all the three results, it appears that the residuals of this particular model are normally distributed.

6.5.4.4 Stability Finally, we perform the stability test through the CUSUM test which assesses the stability of the covariates in the time series VAR model via a plot of the sum of recursive residuals (Stock and Watson 2001). The diagnostic plot indicates structural breaks if at any point in the graph, the sum goes out of the red critical bounds. As we can see from the diagnostic plot below, while neither daily theater count nor widest release presents a structural break, the daily domestic gross box office slightly exceeds the critical bounds.

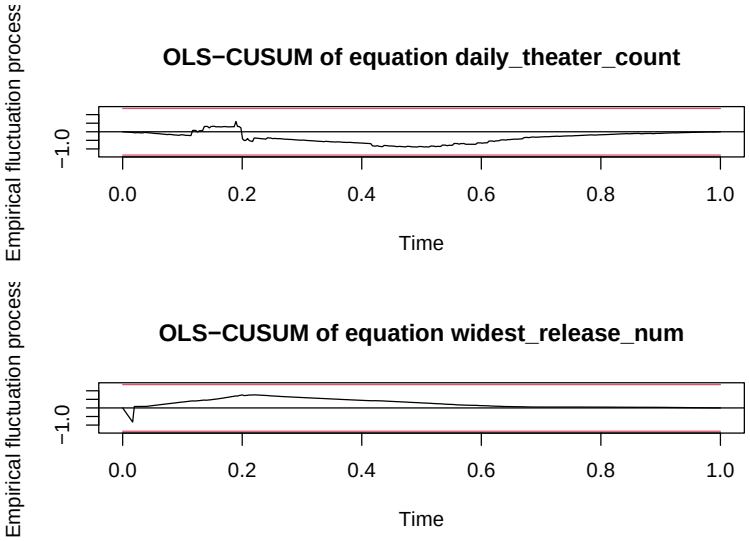


Figure 6: CUCUM Test for VAR Model

6.5.5 Sensitivity Analysis: Lag Structure Apart from the aforementioned multiple lag orders, this study also aims to perform sensitivity analysis by varying the lag structure in the VAR model. According to previous study (Hafer and Sheehan 1989) which examines the effect of lag structure on forecasting accuracy, the forecasting accuracy of the VAR model varies dramatically across simple ad hoc rules, versus statistical criteria such as mean square error and Bayesian rules.

7. Linear Mixed Effects Models

Given the complex nature of the time-series data which also involves non-varying predictors as well as within-group correlation for each movie release, we model the daily movie box office and HSX stock price respectively with linear mixed effects models that incorporate both fixed effects and random effects. Fixed effects enable us to model non-varying predictor variables that are constant over time and specific to each movie, such as the genre of the movie, budget, MPAA rating, distributor and run time under the assumption that they have constant effects on the daily box office across all movies. Temporal variables, on the other hand, are variables that change over time and are not specific to each movie, such as the day of the week, the time of the year, daily theater count for movies, as well as the number of shares long, the number of shares short and the trading volume for HSX stock price. To account for the fact that the effect of the temporal variables vary across different movies, we specify a random intercept structure which accounts for correlation within movies constant across time as well as across-movie variability. This model has the general form, which we apply to our box office and HSX stock price response variables:

$$Y_{i,j} = X_{i,j}^t \beta + u_{i,j}^t \gamma_i + \epsilon_{i,j}, \qquad \gamma \sim N_q(0, D), \qquad \epsilon_i := \begin{pmatrix} \epsilon_{i1} \\ ... \\ \epsilon_{in_i} \end{pmatrix} \sim N_{n_i}(0, \Sigma_i)$$

7.1 Movie Box Office

We fit a linear mixed effects model with the following model specification to gauge insights into how fixed effects, specifically the distributor, runtime, widest release, MPAA rating, budget, genre associate with the daily movie box office, as well as how temporal variables, namely the daily theater count and day in the week associate with the response. Due to the large number of 196 distributors, we separate out the top 10 distributor with most movie releases as individual categories, while grouping the rest as one single "Others" category. While the original data consists of multiple genres for a single movie release, we extracted the first genre listed as the genre variable in the model, while also creating another binary variable to account for whether the movie has multiple genres. Likewise,due to the presence of 27 different genres, we separate out the top 10 genres with the most movie releases as individual categories, while grouping the rest as the "Others" category. With the aid of software package, we extracted the day of the week from the original date variable using Monday as the baseline, and fit a significant interaction term between week of the day and the daily theater count. We mean center the daily theater count in order for interpretation of the main effects in terms of the interaction term.

7.1.2 Sensitivity Analysis & Validation

The model includes a random intercept term for each movie, which captures the movie-specific variation in the response variable that is not explained by the fixed effects, for which the identifier variable is used to group observations by movie. The model also includes a temporal correlation structure specified, conditional on the movie identifier variable and the date variable, which models the correlation between the residuals of the response variable over time within each movie. We perform sensitivity analysis on the specification of correlation structure, and evaluate the model via cross validation using in-sample MSE, out-of-sample MSE, AIC, BIC and Log-likelihood metrics.

Table 2: Different Correlation Structures for Movie Box Office

	corAR1	corARMA	corCAR1
In-Sample MSE	4875060000000	3.573365000000	4875041000000
Out-of-Sample MSE	4503241000000	3144479000000	4503221000000
AIC	6286188	6286188	6286188
BIC	6286659	6286659	6286659
Log-likelihood	-3143048	-3143048	-3143048

We fit the model using autoregressive process of order 1 (`corAR1`), autoregressive moving average process (`corARMA`)and continuous autoregressive process (`corCAR1`) structures to derive the model with `corARMA` as the final model with lowest in-sample and out-of-sample MSE, which has the following specification (Table 2).

We observe that while the daily theater count, runtime in minutes, budget, whether the movie has multiple genres, day in the week have significant associations with daily movie box office with p-values all smaller than 0.05, the distributor has a weaker effect with some levels having p-values smaller than 0.05, and MPAA rating as well as genre do not have a significant effect with large p-values. It is within our expectation that the day in a week has a significant effect as we observe in the EDA: At mean level for daily theater count of 904, for a movie aired on the weekend (Friday, Saturday or Sunday), it is expected to have significantly higher box office than if it is aired on Monday, holding all else constant. Within the weekend, Saturday has the highest box office: A movie aired on Saturday is expected to make 86928 dollars more per day than if it is aired on Monday, conditional on all other predictors. Out of all the days in the week, Wednesday seems to be the trough that underperforms Monday by -101395, holding all else constant. For distributor, a movie produced by Walt Disney Studios Motion Pictures is expected to make 955,359 more per day than if it is produced by non-top 10 distributors in the "Others" category; conversely, a movie produced by Paramount Pictures is expected to make 499892 less per day than if it is produced by distributors in the "Others" category. For every million increase in the budget of a movie, it is expected to make 15,689 dollars more per day holding all else constant. Daily theater count directly has a positive impact on box office: for each additional theater that airs a movie, the box office is expected to 255 dollars more per day holding all else constant.

The interaction term reveals that daily theater count has different effects on box office for different days in the week: Every additional theater that airs a movie corresponds to 1,275.36 higher box office, if the movie is aired on Saturday compared to Monday. We obtain a phi1 value of 0.969, which suggests strong temporal autocorrelation.

Table 3: Model Output for Daily Box Office

	Value	Std.Error	DF	t-value	p-value
(Intercept)	254852.025	546835.492	201979	0.466	0.641
daily theater count	875.151	13.840	201979	63.231	0.000
distributorFox Searchlight Pictures	169256.374	189172.845	4609	0.895	0.371
distributorLionsgate	43400.673	163826.299	4609	0.265	0.791
distributorParamount Pictures	-499891.992	170971.057	4609	-2.924	0.003
distributorRoadside Attractions	354678.449	208536.093	4609	1.701	0.089
distributorSony Pictures Releasing	-448703.576	163553.522	4609	-2.743	0.006
distributorThe Weinstein Company	34481.846	218800.621	4609	0.158	0.875
distributorTwentieth Century Fox	-381369.700	151006.821	4609	-2.526	0.012
distributorUniversal Pictures	-35489.983	148228.118	4609	-0.239	0.811
distributorWalt Disney Studios Motion Pictures	955359.724	179289.543	4609	5.329	0.000
distributorWarner Bros.	-175954.759	134097.680	4609	-1.312	0.190
runtime minutes	6013.345	1713.403	4609	3.510	0.000
widest release num	9.067	40.754	4609	0.222	0.824
mpaa ratingM/PG	624986.806	1786476.704	4609	0.350	0.726
mpaa ratingNC-17	377710.595	984083.724	4609	0.384	0.701
mpaa ratingPG	-153721.468	320944.016	4609	-0.479	0.632
mpaa ratingPG-13	296813.833	325473.757	4609	0.912	0.362
mpaa ratingR	187977.394	328625.536	4609	0.572	0.567
budget	0.016	0.001	4609	15.089	0.000
genreAction	-284394.853	432240.658	4609	-0.658	0.511
genreAdventure	-244570.632	440112.715	4609	-0.556	0.578
genreAnimation	-51299.614	528694.913	4609	-0.097	0.923
genreBiography	-413809.600	440158.501	4609	-0.940	0.347
genreComedy	-472778.118	427130.958	4609	-1.107	0.268
genreCrime	-458137.910	452307.830	4609	-1.013	0.311
genreDocumentary	1524.457	476916.122	4609	0.003	0.997
genreDrama	-426466.624	428998.813	4609	-0.994	0.320
genreFantasy	-523388.443	758112.261	4609	-0.690	0.490
genreHorror	-79142.304	460634.544	4609	-0.172	0.864
multiple genresTRUE	-224811.448	126427.663	4609	-1.778	0.075
dayFri	583971.610	6362.638	201979	91.781	0.000
daySat	869280.329	5884.726	201979	147.718	0.000
daySun	483353.851	4917.504	201979	98.293	0.000
dayThu	1135.166	6372.123	201979	0.178	0.859
dayTue	13829.841	4915.950	201979	2.813	0.005
dayWed	-10139.555	5910.092	201979	-1.716	0.086
daily theater count:dayFri	849.938	5.551	201979	153.105	0.000
daily theater count:daySat	1275.358	5.129	201979	248.671	0.000
daily theater count:daySun	743.241	4.289	201979	173.303	0.000
daily theater count:dayThu	101.210	5.524	201979	18.320	0.000
daily theater count:dayTue	67.092	4.274	201979	15.698	0.000
daily theater count:dayWed	52.698	5.135	201979	10.262	0.000

7.2 HSX Stock Prices

We fit a linear mixed effects model to predict the raw values of the stock price, and a ridge regression to predict its volatility as defined by standard deviation detailed in Section 4.

7.2.1 Stock Price: LME Model

Likewise, we fit a linear mixed effects models with the number of shares sold long, the number of shares sold short, and the trading volume to investigate the temporal movement of the stock price, after conducting sensitivity analysis with the three different correlation structure specifications (`corAR1`, `corARMA`, and `corCAR1`). We select the `corAR1` model as the final model which results in the smallest in-sample MSE, out-of-sample MSE, AIC and BIC. This model has the following specification:

$$\hat{Price} = 0.0000006 \times shares\_long - 0.0000007 \times shares\_short + 0.0000000 \times trading\_vol + \epsilon$$

Table 4: Model Evaluation Under Different Correlation Structures for Stock Price

	corAR1	**corARMA**	**corCAR1**
In-sample MSE	247.7842	287.2423	150.9847
Out-of-sample MSE	224.2619	259.2329	137.1212
AIC	-414.9137	-414.9137	1359.702
BIC	-386.8348	-386.8348	1387.781
Log-likelihood	214.4568	214.4568	-672.8509

	Value	Std.Error	DF	t-value	p-value
(Intercept)	-8.06416018	14.16632827	587	-0.5692484	0.5694053
Number of Shares Long	0.00000057	0.00000002	587	33.9356542	0.0000000
Number of Shares Short	-0.00000067	0.00000004	587	-16.4100003	0.0000000
Trading Volume	-0.00000001	0.00000001	587	-1.3794890	0.1682693

We observe that both the number of shares sold long, and the number of shares short have significant effects on the HSX stock price, with p-values smaller than 0.5, while the trading volume does not have a significant effect. For every 1,000,000 additional stocks sold long, the HSX stock price is expected to increase by 6 dollars per stock per day, holding all else constant. For every 1,000,000 additional stocks sold short, the HSX stock price is expected to increase by 7 dollars per stock per day, holding all else constant.

7.2.2 Stock Volatility: Ridge Regression Model

To investigate the properties that correlate with higher stock volatility, we fit a ridge regression model with genre, MPAA Rating, distributor and IPO decade as predictors. We perform 10-fold cross validation to tune the hyperparameter λ and obtained the optimal value of 30.54648.

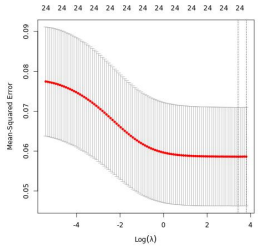


Figure 7: Ridge Regression Cross Validation

We observe through the ridge regression model that HSX stocks issued by different distributors have great variability in volatility: For a stock issued by Lionsgate, it is expected to have 6.46e-04 greater volatility than if the stock is issued by a non-top 10 distributor in the "Others" category. Similarly, genre and MPAA rating also have diverging effects on the HSX stock volatility.

Table 5: Volatility model output

predictor	coefficient	predictor	coefficient
genreAction	1.360963e-03	mpaa ratingr	5.096444e-04
genreAnimated	-1.181718e-03	distributor20th century fox	-1.492772e-04
genreComedy	-5.359766e-04	distributora24	8.646593e-04
genreDocumentary	-2.177618e-05	distributorifc films	-1.071686e-05
genreDrama	-3.490515e-04	distributorlionsgate	6.468972e-04
genreFantasy	-9.217823e-04	distributorparamount	2.762789e-04
genreHorror	1.574434e-04	distributorsony	-3.527879e-04
genreRomance	-6.722932e-05	distributoruniversal	1.154761e-04
genreSci	4.211737e-04	distributorwalt disney	-3.009726e-04
genreThriller	8.635522e-04	distributorwarner bros.	-3.616480e-04
mpaa ratingpg	-3.653962e-04	IPO in 2009 decade	-4.660983e-04
mpaa ratingpg-13	-4.100696e-04	IPO in 2019 decade	6.853935e-04

7.3 LME Model Diagnostics

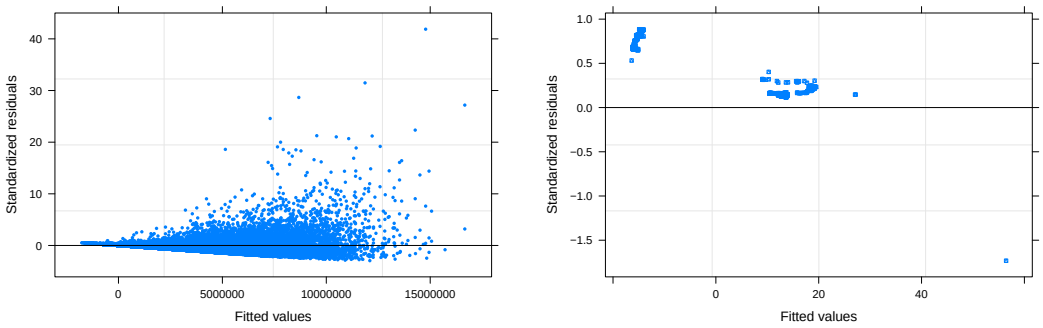


Figure 8: Residual Diagnostic Plot for Daily Box Office (a), and HSX Stock Price (b) LME Models

We check the linearity and independence assumptions of the linear mixed effects models. Linearity of the predictors is satisfied by checking the residual plot: The box office model virtually has no systematic shape, while the HSX price model shows more deviation from the assumption mainly possibly due to the small sample size after truncating minute-wise temporal data points to daily prices. The independence assumption is satisfied since we assume that the box office and HSX stock price observations are independent across different movies.

8. Predictability between Two Time Series: CCF

Given the relationship between two time series, we decide to use the cross correlation function (CCF) model to quantify the degree of resemblance between box office and HSX stock prices, and identify lags of the fluctuations in HSX stock prices that might be useful predictors of movie box office. Since the CCF model gives us informative information on the order of prediction between movie box office and HSX stock prices through the set of sample correlations, we can also identify which variable is leading and which is lagging. The essential structure of the model lies in the assumption that given two processes X_{1t} and X_{2t}, $\rho(X_{1t}, x_{2t+k})$ is the cross-correlation between X_{1t} and X_{1t} at lag k, while $\rho(X_{2t}, x_{1t+k})$ is the cross-correlation between X_{2t} and X_{3t} (Usoro et al. 2015), where the definition of the autocorrelation between times s and t is $\rho(s,t) = \frac{E[(X_t-\mu_t)(X_s-\mu_s)]}{\sigma_t \sigma_s}$

We come to an interesting observation that for unsuccessful movies (with worldwide gross box office under the threshold of 500 million), the HSX stock price and box office do not seem to predict each other, whereas for successful movies (with worldwide gross box office over the threshold of 500 million), the HSX stock price tend to correlate more strongly with the box office. Take the unsuccessful movie Three Peaks as an example (Figure 9(1)), we first identify the two inflection points at lag -7 and lag 2, and then calculate the significance threshold $\frac{2}{\sqrt{n-|k|}}$ where n is the number of observations and k is the lag, to derive thresholds of 0.417 and 0.535 respectively (Usoro et al. 2015). We compare the correlation to the threshold to observe that the CCF does not exceed the threshold, therefore the

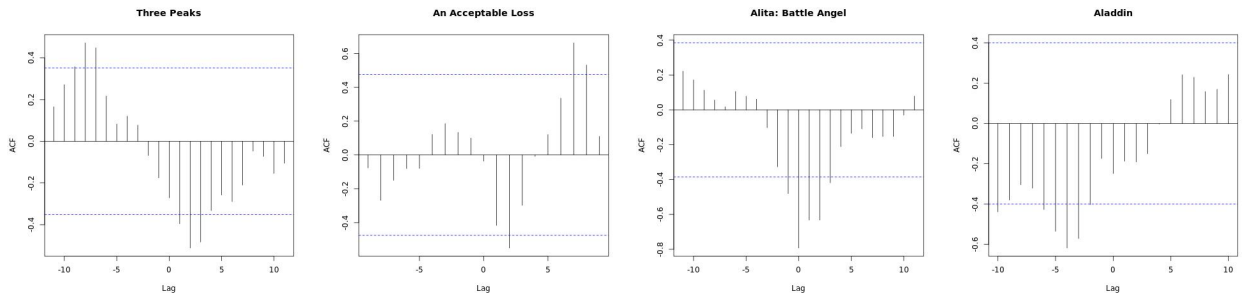


Figure 9: CCF for two unsuccessful movies (1-2) and two successful movies (3-4)

time signals do not seem to predict each other. Likewise, we repeat the process for the other movies (shown in Figure 9) and reach the conclusion that while for the latter two commercially successful movies (Alita and Aladdin), stock price correlate more strongly with box office, for the former two unsuccessful movies (Three Peaks and An Acceptable Loss), the two time series do not correlate strongly with each other. We can take a further look into the CCF of the Alita movie, which is well below 0 for the most part, which suggests that the HSX stock price leads box office since CCF is negative. On the other hand, for Aladdin, the CCF is above 0 for the majority, which is to say that the HSX stock price lags box office since CCF is positive (Usoro et al. 2015).

9. Conclusions

In this study, we investigate time series box office and HSX stock price separately, and their relationship in light of each other. To answer the first question on box office, we fit linear mixed effects models with daily theater count, distributor, runtime, widest release, MPAA rating, budget, genre and day of the week as predictors. We find that while the widest release does not have a significant effect, the temporal daily theater count does have a significant positive effect. Within our expectation, Saturday correlates with highest box office while Wednesday correlates with the lowest. To our surprise to a certain extent, runtime does have a significant positive effect that can be explained by studies which have found better IMDB rating if runtime is longer (Kumar, Mehta, and Pal 2019), since longer movies typically have more content to attract audiences with a more immersive and in-depth movie experience. It is also worth noting that the typical "big name" top 10 distributors do not necessarily correlate with higher box office: Paramount, Sony and Universal pictures correspond to a negative effect on box office so that a movie production investor might be mindful of over-reliance on these distributor companies when making decisions.

To answer the second question, we first fit a linear mixed effects model to find positive effect of shares sold long, and negative effect of shares sold short on stock price. This can be explained by the economic intuition that long selling signifies faith in a company, whereas short selling creates a negative perception of a company in the market, as short sellers are essentially betting against the company's success. We then fit a ridge model to find that stocks issued by A24 or Lionsgate, in the 2019 decade, with rating R tend to associate with higher risk and volatility.

To answer the third question, we fit CCF to measure the similarity between box office and stock price as a function of the time delay between them, and found that while for successful movies the two time signals tend to correlate, for unsuccessful movies they do not. Within the successful movie subgroup, we found that the two time signals have diverging lagging or leading effects, depending on the individual movie.

10. Limitations & Future Work

From the data's perspective, one greatest limitation is that the movie box office and the HSX stock price data are scraped from two different websites, so that joining the two data sets for the CCF analysis greatly reduces the sample size which could affect the generalizability of the model. Moreover, even though the stock price data includes 12,677,219 data points, most of them are recorded close to each other with minutes apart so that the stock price does not fluctuate noticeably, reducing the effective size of the data sample. From the model's perspective, while the linear mixed models enable us to account for between-movie variability to include the complete dataset, the CCF does not incorporate random effects, so that we are unable to use the complete data and select a few of the movies that might not be representative. While our model demonstrated good predictive performance on the test data, there are several avenues for future work that could improve the accuracy and generalizability of the model. One potential direction is to explore additional fixed and random effects that could be included in the model, for example, variables such as star power of actors, or prior box office performance of directors as fixed effects.

11. Citations

Bhave, Anand, Himanshu Kulkarni, Vinay Biramane, and Pranali Kosamkar. 2015. "Role of Different Factors in Predicting Movie Success." In *2015 International Conference on Pervasive Computing (ICPC)*, 1–4. https://doi.org/10.1109/PERVASIVE.2015.7087152.

CFI. 2023. "Volatility: A Measure of the Rate of Fluctuations in the Prices of a Security over Time," January. https://corporatefinanceinstitute.com/resources/capital-markets/volatility-vol/.

Hafer, R. W., and Richard G. Sheehan. 1989. "The Sensitivity of VAR Forecasts to Alternative Lag Structures." *International Journal of Forecasting* 5 (3): 399–408. https://doi.org/https://doi.org/10.1016/0169-2070(89)90043-5.

"Hollywood Stock Exchange & Box Office Data." 2021. https://www.kaggle.com/datasets/zeegerman/hollywood-stock-exchange-box-office-data.

Karniouchina, Ekaterina V. 2011. "Are Virtual Markets Efficient Predictors of New Product Success? The Case of the Hollywood Stock Exchange*." *Journal of Product Innovation Management* 28 (4): 470–84. https://doi.org/https://doi.org/10.1111/j.1540-5885.2011.00820.x.

Kotu, Vijay, and Bala Deshpande. 2019. "Chapter 12 - Time Series Forecasting." In *Data Science (Second Edition)*, edited by Vijay Kotu and Bala Deshpande, Second Edition, 395–445. Morgan Kaufmann. https://doi.org/https://doi.org/10.1016/B978-0-12-814761-0.00012-5.

Kumar, Saurabh, Avinay Mehta, and Joy Pal. 2019. "Movie Success Prediction Using Data Mining." *Vellore Institute of Technology*.

Lee, Kyuhan, Jinsoo Park, Iljoo Kim, and Youngseok Choi. 2018. "Predicting Movie Success with Machine Learning Techniques: Ways to Improve Accuracy." In *Information Systems Frontiers*, 20:1–1. 577-588. https://doi.org/10.1007/s10796-016-9689-z.

Stock, James H., and Mark W. Watson. 2001. "Vector Autoregressions." *Journal of Economic Perspectives* 15 (4): 101–15. https://doi.org/10.1257/jep.15.4.101.

Usoro, Anthony E et al. 2015. "Some Basic Properties of Cross-Correlation Functions of n-Dimensional Vector Time Series." *Journal of Statistical and Econometric Methods* 4 (1): 63–71.