Imperial College
London

MEng Individual Project

Imperial College London

Department of Computing

Machine Learning for the Analysis and Prediction of Film Performance

Author:
Padmanaba Srinivasan

Supervisor:
Prof William Knottenbelt

Second Marker:
Prof Peter Harrison

June 17, 2020

Abstract

When was the last time you watched a film?

Probably not too long ago. In watching a film you have taken part in a tradition that, since the early 20th century, has had a profound influence on politics, fashion and economics and is no less significant an influence today. The proliferation of the Internet has made films more visible and accessible to infinitely more people than before and has seen the industry evolve and diversify. Nowadays, films tap many sources of revenue from the traditional and familiar theatrical release, to DVD and more recently, Video on Demand. The ability to predict a movie's performance before and during the early days of its release can significantly reduce the financial risk. However, a film is not just about its story: the actors, the director(s) and many other factors also need to be just right to create a financial success. In this project, we predict film performance using parameters such as these.

Our contributions are twofold. Firstly, we look into predicting film performance prior to its release. We begin with a classification problem − how to determine whether a film will be profitable in advance of release and identify the key factors that contribute to profitability. We then propose a regression problem: one where we design Machine Learning models to make concrete predictions of revenue generated by films in several regions. This is followed by performing model interpretation as an investigation into the dependencies and relationships between the performance of regions, which sheds light on the interconnectedness of film markets and the level to which our models leverage diverse features. We find that geographically close regions, such as Germany and France, exhibit a large amount of interdependence, with films that perform well in one country also performing well in the other. We also find more latent relationships, such the relationship between Germany and Japan where film performance is correlated despite the countries being geographically distant and not sharing a common language, thus suggesting that there is an inherent similarity in their box office markets and the expectations and preferences of their audiences. We then analyse the performance of our models and explain the observations we have made. We show that our models record a performance improvement over existing models, with our Neural Network model achieving a Coefficient of Determination of 0.851 for predictions on final box office gross.

Secondly, we model how to predict film performance after the release of a film. We propose thinking of a film release as similar to an epidemic using Ordinary Differential Equations, and develop models to use data as it is collected to fit model parameters. We also explore how Machine Learning can be applied to the problem, using time-series data along with other features that describe a film. With these models, we show how they can be applied to forecast film performance, firstly for one week ahead at a time and secondly, for several weeks in the future. Finally, we evaluate the merits of each model, finding a Mean Forecast Error of approximately \$842 000 for the best performing model. We identify similarities in the performance of the two models containing LSTM units and investigate why, showing that most of the information required to forecast revenue is obtained from a few trailing weeks of information thus more complex models offer a lower improvement in performance.

Acknowledgements

I would like to express my gratitude to the following people:

- My supervisor, Prof William Knottenbelt, for his boundless enthusiasm, insight inducing questions and concise advice. I've enjoyed our conversations regarding this project as much our discussions about the sport of tennis.
- My family for their encouragement, curiosity and support.
- The *FilmChain* team for sharing their knowledge of the film industry and for giving me a better understanding of what the industry needs. I would like to thank Liv, for her insight into the industry and for making excellent suggestions through the course project.

Contents

List of Figures

Chapter 1

Introduction

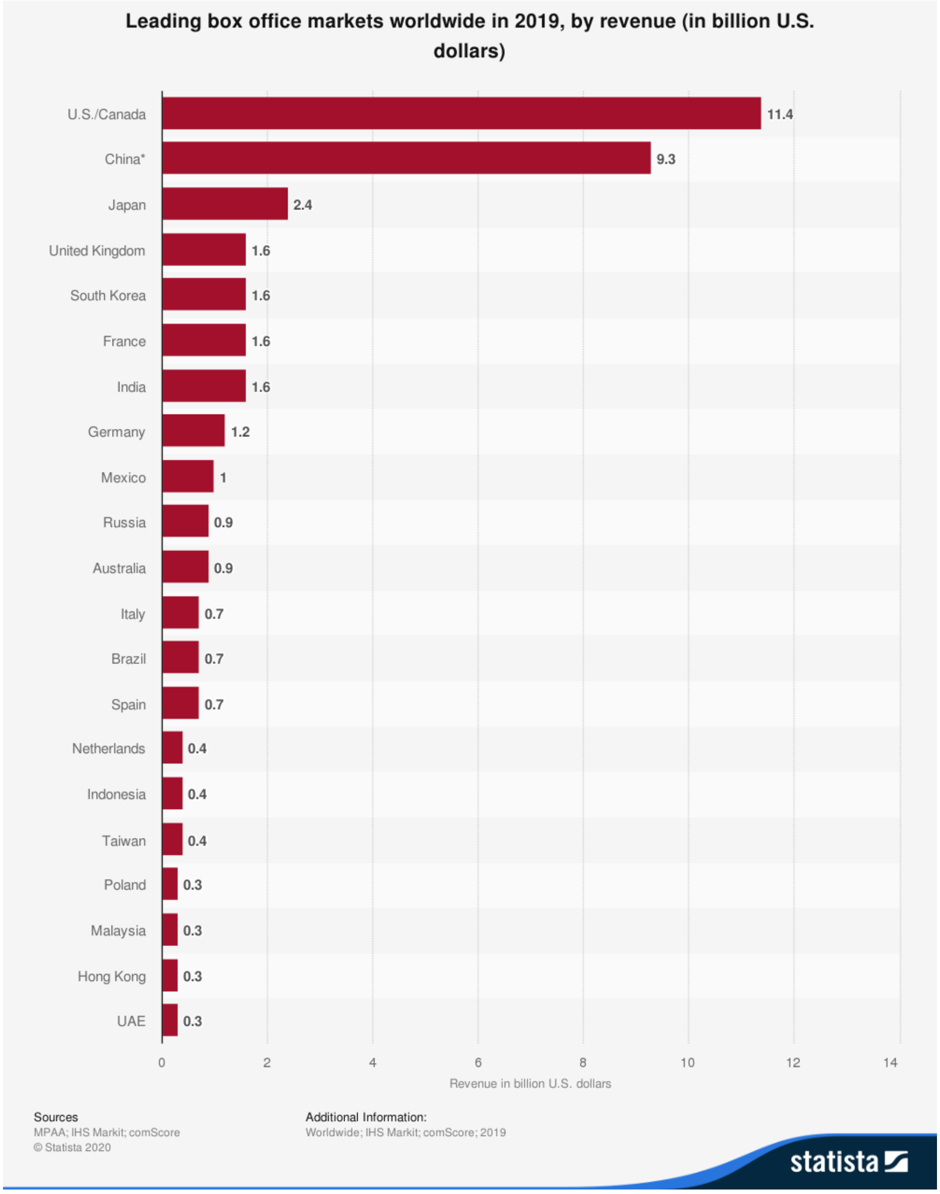
> I don't know any other business that tells you not to go in and buy their product.
>
> *Jack Valenti. Former President of the MPAA on the film rating system, 1985*

1.1 Motivation

Figure 1.1: *Back to the Future*, the 'Parking Lot' scene. Credit: YouTube

Films constitute a high profile, multi-billion dollar industry that sees its origins in the late 19th century with the invention of the kinetiscope by Thomas Edison and William Dickson [57]. Innovation propelled cinema technologies forward, and by 1915 an entire industry had developed, being centred in a suburb of Los Angeles where land was cheap and the weather favourable: Hollywood. The industry grew in size and influence, entering its Golden Age and becoming a champion of creativity, culture, and during periods of war, a tool for propaganda. The success of Hollywood spawned industries in several other countries, which in 2018 were estimated to be worth a staggering $41.1 billion [1] worldwide.

Figure 1.2: Leading Box Office Markets by Revenue, 2019. Credit: Statista

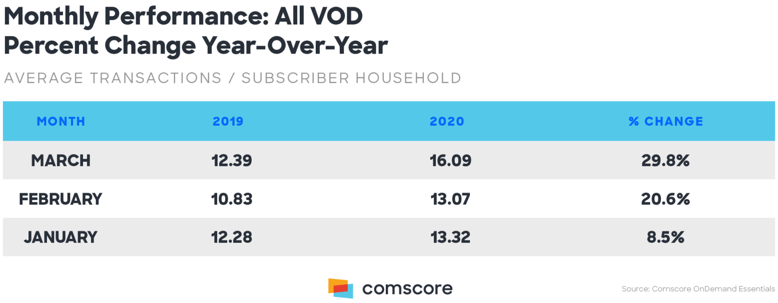


With such large amounts of money at stake, people involved in making films place high value on a predictive model that gives them an idea of the returns they will receive. In practice, profit sharing amidst people involved in the film process is quite complex. A Rental Contract includes the duration an exhibitor will show a film, an agreement on how revenue will be shared between exhibitor and distributor, as well as potential exclusivity clauses. The first few weeks of screening is when a film makes the most money with the distributor taking the lion's share (60-70%) of revenue. For the remaining period, the split is often reversed with the exhibitor taking the larger cut. Adding further complexity to the matter is the house nut – if the film makes more money in a given week than a predetermined amount, then 90% of this goes to the distributor. In this project, we ignore the potential variables of such contracts and focus instead on how much raw box office gross a film will make. After all, it is film performance we are interested in and the revenue (or gross, used interchangeably), which is what we predict and, which is what the distributor receives after the exhibitor has taken their share.

Films can make money beyond the confines of a cinema – Video on Demand (VoD), TV and DVD are all supplementary sources of revenue for films. VoD, in particular, is one of the fastest growing methods of viewing content, recording a 161% growth between 2011 and 2015 [2] in Europe alone. The VoD market itself is split into subsections to include Transactional VOD (TVoD), Subscription

VoD (SVoD) and Advertising VoD (AVoD), with many players in the market operating on both regional and international scales. Examples of market players in VoD are Netflix (SVoD) and iTunes(TVoD). On a more recent note, COVID-19 has spurred the growth of the VoD market as a whole; Comscore reports a significant increase (Figure 1.3) in the number of transactions in the VoD market corresponding to the imposition of lockdown in many countries.

Figure 1.3: Average Growth of VoD Transactions. Credit: Comscore



Monthly Performance: All VOD
Percent Change Year-Over-Year

AVERAGE TRANSACTIONS / SUBSCRIBER HOUSEHOLD

MONTH	2019	2020	% CHANGE
MARCH	12.39	16.09	29.8%
FEBRUARY	10.83	13.07	20.6%
JANUARY	12.28	13.32	8.5%

comscore

Source: Comscore OnDemand Essentials

It is thought that this is only the beginning of a new trend for films to be released far earlier into VoD markets, such as was done for the film *Trolls World Tour*, which Universal Pictures released into theatres and the VoD market simultaneously in April 2020. The film went on to generate around \$100 million in rental fees alone within the first three weeks (source: The Wall Street Journal). However, VoD providers do not release details of sales or revenue generated by films and as a result, we consider only the theatrical and DVD/Blu-ray revenues in our investigation into box office performance.

Of the 37 472 films made in the US between 1999 and 2018 only 5.7% were made by the largest five studios and 90.3% of films in this period saw no theatrical release, yet only 3.4% of film ventures reported a profit. Of the films that made a loss, the majority had one thing in common: they were films made by independent studios [3]. Clearly, the industry would benefit from a tool that can provide insight into film profitability.

1.2 Objectives

The goals for this project are as follows:

- Understand how to model the box office performance of films and the Key Performance Indicators of success;
- Develop models to predict the revenue of films from various sources on a per-country basis at strategically informative points in time;
- Produce and evaluate models for use both prior to a film's release as well as during, using live data;
- Apply models to data and characterise the uncertainty in the fit of the model.

This project is run in collaboration with *FilmChain*, a startup that manages end-to-end financial transactions for creative industries and as a service, delivers insights into film revenue using Machine Learning.

1.3 Contributions

Key Performance Indicators Prior to Release We explore how to predict whether or not a film will be profitable in advance of its release and extract the Key Performance Indicators to gain an insight into what drives profitability.

Granular Regression of Performance Prior to Release Having good quality estimates of how a film will fare revenue wise is invaluable to the beneficiaries of its revenue. Previous research tackles the problem of predicting only final gross, but having information on the structure of that gross, that is to say, the cumulative gross at various points in time will allow more accurate calculation of how much of the proceeds individuals are likely to receive according to terms of distribution agreements. To make this more relevant and informative, we predict these grosses on a country by country basis allowing even more accurate estimates following the terms of individual country agreements.

Ordinary Differential Equations for Revenue Prediction We develop a set of coupled Ordinary Differential Equations explaining the intuition behind them and explain how to and apply these to the modelling of films both prior to release as well as during release.

Revenue Forecasting During Release Performing regression during the pre-release stage alone is insufficient. Hence, we also explore how to use data available after a film's release as it is collected to provide even more granular and up to date predictions of future performance. This means considering a film not as a static entity, but as a dynamic object for which the future outcome is affected by current events, many of which are caused by the whims and fancies of the general film-loving populace.

1.4 Project Outline

The remainder of this dissertation is organised as follows:

Chapter 2 performs a literature survey, introducing the types of models developed till now and introduces the techniques and models we use in the remainder of the project.

Chapter 3 introduces the data used in our modelling. We comment on where and how the data was collected, how features are related to one another and how this influences the way in which we perform feature engineering on the data. We also comment on the nature and characteristics of the final datasets and highlight some aspects of the integrity of the data.

Chapter 4 poses the pre-release problem, first of all tackling the question of profitability before moving on to perform regression. Firstly, we investigate the factors that affect profitability and evaluate these profitability models against each other and against models in literature. We then introduce the regression problem and the models we use to solve it, show what predictions are like and how the models interpret features for each regressand and identify key relationships between features. Finally, we evaluate these models and comment on the reasons behind the observed performance.

Chapter 5 explores the problem of modelling post-release performance at the box office using a set of coupled Ordinary Differential Equations, presenting the development and intuition behind these equations as well as seeing the capabilities of this modelling technique. We base our work on the approach taken by Edwards and Buckmire, developing our own original model and investigating the uncertainty associated with it.

Chapter 6 develops Machine Learning models to model post-release at box office as a time-series forecasting task. We explore different methods of forecasting and show performance on individual samples before formally evaluating the performance and commenting on reasons behind observations.

Chapter 7 concludes this project, with a summary of work, reflection on the achievements as well as directions for future work.

Chapter 2

Background

Elementary, my dear Watson.

The Adventures of Sherlock Holmes, 1939

The problem of quantitatively modelling box office dynamics has spawned interesting and fruitful research. Smith and Smith [4] in *Applied Economics* described cinema as "one of the best examples of the differentiated product envisioned in conventional models of monopolistic competition". Most existing research explores cinema as a problem to be solved with either probabilistic and statistical approaches or as a Machine Learning problem.

2.1 Ordinary Differential Equation Models

This section lays down the basis for modelling the box office performance of a film using an Ordinary Differential Equation (ODE).

ODE Models have been used extensively in various fields for modelling purposes, such as in epidemics where Kermack and McKendrick [10] suggested a method of using ODEs to model such phenomena, developing the notable Susceptible, Infected, Recovered (SIR) model.

2.1.1 Bass Model

One of the earliest consumer behaviour models that can be applied to modelling box office performance of a film is through the Bass Model [5]. The Bass Model employs an Ordinary Differential Equation which relates how new products are adopted in a population:

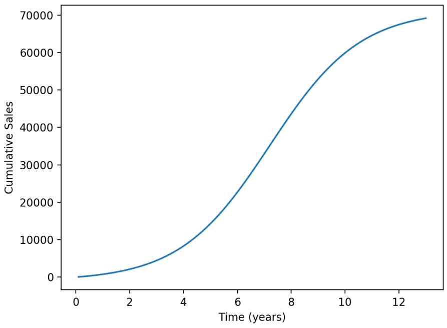
$$\frac{f(t)}{1-F(t)} = p + qF(t) \tag{2.1}$$

Where

- $f(t)$ is the rate of change of the installed base fraction
- $F(t)$ is the installed base fraction
- p is the coefficient of innovation
- q is the coefficient of limitation

The solution to this system yields the famous S-curve (Figure 2.1) which represents the rate of adoption of a product over time. This model has been one of the most influential and frequently cited papers in the history of *Management Science*.

Figure 2.1: S-Curve showing Cumulative Sales over Time.



2.1.2 Behaviour Models

Sawney and Eliashberg [6] presented a model utilising a two parameter exponential distribution and three parameter gamma distribution which modelled the process by which a person would decide to see a movie in two stages − firstly, as the time taken to decide whether or not to see a movie, and secondly, to model the time taken to act on that decision.

In further work, Eliashberg et al. [7] designed a program to predict box office gross for a film using only pre-release data. Specifically, it is able to predict the box office for a film based by modelling the behaviour of consumers using Interactive Markov Chains [8] where the transition probabilities depend on the number of people already in the other states. In MOVIEMOD, the program that Eliashberg et al. develop, they consider the effect of positive and negative word of mouth, the duration of this effect and the frequency at which people talk about their opinions of a film.

The primary difficulty in using probabilistic approaches − especially those that attempt to model behaviour − is the problem of choosing an appropriate distribution for the process being modelled.

2.1.3 Edwards Buckmire Model

The success of the Bass Model motivated the use of ODEs to model film box office performance. Edwards and Buckmire [9] presented a model developed using the same tools, in the *IMA Journal of Management Mathematics.*

Edwards and Buckmire developed a system of coupled ODEs to model the box office performance on a film after release. The Edwards Buckmire (EB) model takes a deterministic approach, employing governing equations with the rate of change of gross at time t, initially given by:

$$\frac{dG}{dt} = \tilde{S}\tilde{A} \tag{2.2}$$

$$\frac{d\tilde{A}}{dt} = -\alpha_A \tilde{A} \tag{2.3}$$

$$\frac{d\tilde{S}}{dt} = -\alpha_S \tilde{S} \tag{2.4}$$

where

$\tilde{G}$, $\tilde{A}$ and $\tilde{S}$ are the gross, amount of money earned per screen per week and the number of screens on which the film is presented, respectively. P is the ticket price. Conditions include $\tilde{G}(0) = 0$ and $\tilde{G}(\infty) = \int_0^\infty \tilde{S}\tilde{A}\ dt$.

The EB model is then further developed, attempting to model negative human response ($\tilde{H}$, with the percentage of people who hate a film given by $H_{\%}$) to a film governed by:

$$\frac{d\tilde{H}}{dt} = \frac{H_{\%}\tilde{S}\tilde{A}}{P} \tag{2.5}$$

This assumes previous knowledge of the total number of people who will watch the movie and knowledge of the number of people who dislike it − which is impractical in the real world. The EB model is further developed allowing for people to watch a film multiple times, adding parameters for consideration of genre, the amount spent on advertising and the effectiveness of advertising. As they note, many of these parameters must be estimated in real-time.

The EB model yielded promising results for a number of movies such as the graphs for the parameters in Figures 2.2, 2.3 and 2.4.

Figure 2.2: Gross for *At First Sight.* Prediction Curve [9]

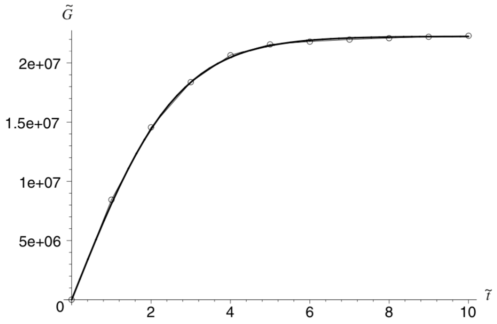


Figure 2.3: Attendance for *At First Sight.* Prediction Curve [9]

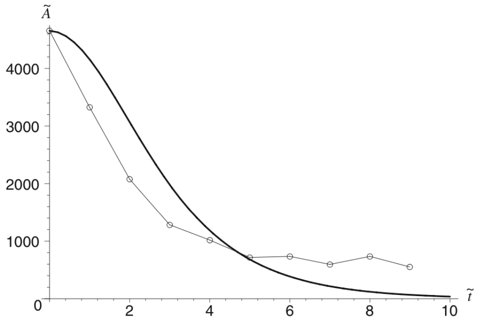
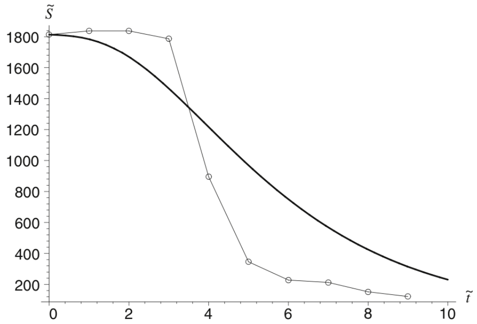


Figure 2.4: Screens for *At First Sight.* Prediction Curve [9]

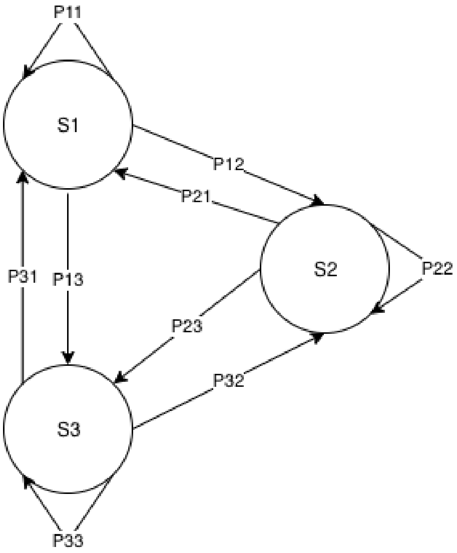


2.1.4 Stochastic Approximation Methods

The box office performance of films is a stochastic process. The gross of a film is determined by the number of theatres that are exhibiting the film and the number of people that go to watch it. There is randomness in this and stochastic models attempt to simulate this randomness.

Stochastic processes can be modelled using Markov Chains where the next step is dependent only on the present state. A Markov Process is memoryless $-$ it has no knowledge of how the present state was reached or at what time it was reached at. For example, given n states $S = \{s_1, s_2, ..., s_n\}$, the Markov process can transition from one state to another with the probability P_{ij} (in a discrete-time process). An illustration of a Markov Chain is in Figure 2.5.

Figure 2.5: A Three State Markov Chain



The Markov Chain can model the revenue behaviour of individual theatres but translating this to a macroscopic view of the entire system can be challenging. Modelling several individuals using Markov Chains results in the state space explosion problem [11] which can make this approach unviable.

We can approach this from a deterministic point of view, modelling the stochastic process using ODEs [12, 13]. This means we lose the discrete states of a Markov Chain model, but, we gain a model that is far easier to simulate than several Markov processes. However, ignoring the stochastic effects can mean that in some cases the ODEs generate a significant error [14, 15].

As the step size of jumps in a stochastic simulation tends to zero, the impact of oscillations due to the stochastic nature of the process decrease as well [16]. A visualisation in Figure 2.6 of this shows the effect of halving the time step and doubling population from an original graph.

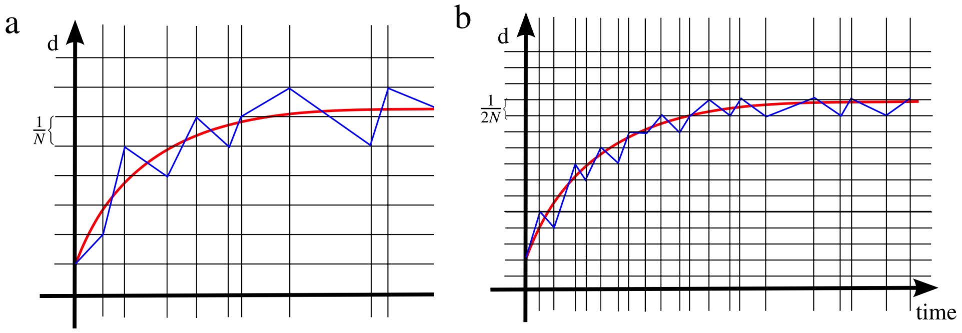


Figure 2.6: Visualisation of Deterministic Approximation Theorems with Large (a) and Small (b) Step Size [16]

2.2 Machine Learning Models

Machine Learning (ML) is a subset of the field of Artificial Intelligence (AI) that develops algorithms which can learn from data. Supervised ML involves a model learning a function from labelled training data.

In the case of predicting film performance, historical data containing information such as budget, language etc. can be used to train samples. Feature Engineering of data is a crucial part of developing an effective ML model.

The problem posed here is one of regression where the output data will be a real-valued number, such as the total box office gross of a film.

Predicting box office gross has been tackled as a classification problem with some success [28] and is potentially a sensible approach to the problem if regression is unable to yields good results.

The techniques presented here are algorithms which have either been applied to similar problems, or whose properties are particularly interesting for the modelling this problem.

2.2.1 Linear Regression

Linear Regression analysis is an approach to modelling the relationship between independent variables and dependent variables.

Given a data set $\{y_i, x_{i1}, x_{i2}, ..., x_{in}\}_{i=1}^{n}$ a model can be defined that takes the form:

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_n x_{in} + \epsilon = x_i^T \beta + \epsilon \tag{2.6}$$

Where ϵ is a noise term. This can be rewritten in matrix form as:

$$y = X\beta + \epsilon \tag{2.7}$$

In Maximum Likelihood Estimation [23] we attempt to find Θ that minimises the log likelihood:

$$\arg\min_{\theta} \; -\log \; L(\theta) \tag{2.8}$$

$$= \arg\min_{\theta} \; -\log \; P(y|X, \Theta) \tag{2.9}$$

Where X is a design matrix. A closed form solution can be found as:

$$\Theta_{ML} = (X^T X)^{-1} X^T y \tag{2.10}$$

If the derivative cannot be calculated then an iterative solution can be found through Gradient Descent, or when using a large data set through Stochastic Gradient Descent [23, 24].

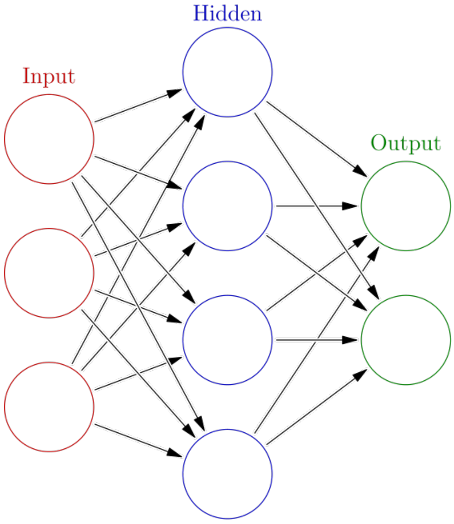
Gopinath et al. [25] applied regression to a data set of film information (e.g. indicators such as the number of theatres a film is released in, advertising budget etc.) to find the relationship between pre-release factors on opening box office performance, and both pre and post release factors on performance after one month. Their work shed light on how films could be released to ensure exposure to the top US local area markets.

Almadi [26] too applied Linear Regression to predict the box office of films and found important post-release relationships between post-release media and gross, and awards and gross.

2.2.2 Neural Networks

A Neural Network (NN) is a system inspired by structure of biological neural networks. NN's can be structured to have several layers with each neuron connected to each neuron in the following layer (except for the input layer).

Figure 2.7: Illustration of a Three Layer Neural Network (Credit: Dataiku)



Each neuron has associated with it a weight, which the neuron uses along with its inputs to calculate an output value. An activation function can be applied to the sum of the weighted inputs from the previous layer. The activation function that is applied is a hyperparameter than must be chosen. Neural Networks are fitted to a training dataset using the back propagation algorithm which can use optimisation algorithms such as Stochastic Gradient Descent which present further parameters to optimise including learning rate and batch size.

Dropout

Dropout [51] is a regularisation technique where every neuron in a given layer has a probability, p, of being temporarily removed during training. This results in successive layers in the NN model having to compensate for neurons that are missing at random thus forcing layers to consider data from a wider variety of neurons in the previous layer. Dropout is typically not used in the output layer as this is where predictions are generated from.

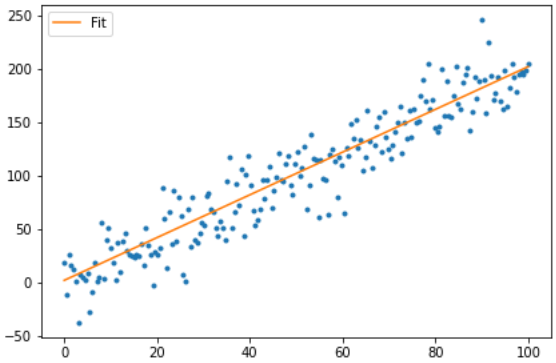
Monte Carlo Dropout for Uncertainty Estimation

Neural Networks are an effective and flexible method of modelling complex functions with data that is both high dimensional and numerous. However, they are prone to being overconfident [53] in their predictions which leads to the key question: just how uncertain are they?

Before we explore how to quantify uncertainty, we must understand what uncertainty is in the first place. In modelling, uncertainty is of two types: aleatoric uncertainty and epistemic uncertainty.

Aleatoric uncertainty, visualised in Figure 2.8, refers to the uncertainty in the data itself. For example, when predicting house price using floorspace only, there may be multiple prices for houses of the same size.

Figure 2.8: Example of Aleatoric Uncertainty



Epistemic uncertainty is the uncertainty of the model itself, often caused by a lack of data. Epistemic uncertainty can be reduced by increasing the amount of data available and by optimising model parameters. However, in many cases the epistemic uncertainty is ignored and instead models are fitted by minimising/maximising a performance metric. Nevertheless, having a means of obtaining the model's uncertainty would be a key informative piece of information.

Dropout, as described previously, is a technique used to perform regularisation and reduce overfitting. Gal and Gahramani (2015) [52, 59] introduced a method of using dropout during inference to estimate the model's uncertainty. They use a Bayesian interpretation of dropout to show that using dropout in training and prediction is sufficient to approximate a deep Gaussian Process.

We begin from maximising the Evidence Lower Bound (ELBO) (see Section 2.2.3 for more information) during Variational Inference (VI):

$$L_{VI}(\omega) = \int q_\omega(\theta) \log(P(Y|X,\theta))d\theta - KL(q_\omega(\theta)||p(\theta)) \tag{2.11}$$

This integral is not tractable for almost all q, therefore we perform Monte Carlo integration to yield an the approximation in Equation 2.12 where $\hat{L}$ is an unbiased estimator of L. To optimise $\hat{L}$, we sample one $\hat{\theta}$ from $q_\omega(\theta)$, then perform the optimisation step on ω, which is then repeated.

$$\hat{L}_{VI}(\omega) = \log(\frac{q_\omega(\theta)}{P(Y|X,\hat{\theta})}) - KL(q_\omega(\theta)||p(\theta)) \tag{2.12}$$

Each layer of the Neural Network with layers $[1, ..., L]$ has weight matrix W_i and we define $q_\omega(\theta) = \sum_{i=1}^{L} p_{M_i}(W_i)$, where M_i is the mean weight matrix for layer i.

$$M_i = mean(W_i) \tag{2.13}$$
$$q_{M_i} = M_i \; diag([z_{ij}]) \tag{2.14}$$
$$z_{ij} \sim Bernoulli(p_i) \tag{2.15}$$
$$W_i \sim q_{M_j}(W_i) \tag{2.16}$$

Sampling the diagonals of z, which is Bernoulli distributed, means that columns of M_i are randomly set to zero vectors, which is analogous to dropout where outputs of random neurons are set to zero.

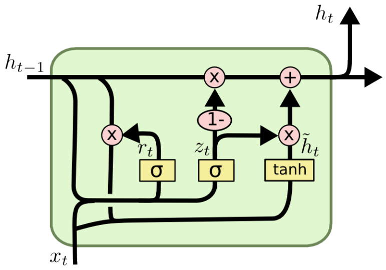
Performing multiple (hundreds) of inferences on the same input with dropout enabled means we obtain several samples. The mean and standard deviation of these samples can be calculated yielding an approximation of a Gaussian distributed, predictive posterior. This procedure is known as Monte Carlo (MC) dropout.

Long Short-Term Memory

Neural Networks are powerful tools capable of modelling complex equations. But what happens when the data being modelling is temporally related? One solution is to use a Recurrent Neural Network (RNN). The brain is an RNN – it closely resembles layers of neurons with feedback connections that can exploit temporal relationships between data. Consequently, RNN's are inherently better suited to modelling temporally related data than feed forward Neural Networks in areas such as Natural Language Processing and time-series prediction. The primary disadvantage of RNN's is the inability to use from a long time ago, so during back propagation earlier layers may only receive a small gradient update, effectively preventing further learning.

Long Short-Term Memory (LSTM) is a solution created to solve this. LSTMs have the ability to retain or forget information thus allowing earlier information to make it to later stages if it is useful. An illustration of an LSTM is in Figure 2.9.

Figure 2.9: Illustration of an LSTM Cell [54]



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$
$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$
$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$
$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

Estimating model uncertainty using dropout is possible with LSTMs as well, as demonstrated by researchers, Zhu & Laptev [55] at Uber.

Applications of Neural Networks

Sharda and Delen [28] used a NN, treating the problem as one of classification, classifying movies into nine different, ranged categories based on budget. An illustration of their model is in Figure 2.10. On their data set, the model achieved a 37% correct classification with 75% of classifications being within one category of the correct one. They compared the results (Figure 2.11) of this to other Machine Learning methods, and demonstrated a performance improvement when using Neural Networks.

Figure 2.10: Sharda and Delen Model [28]

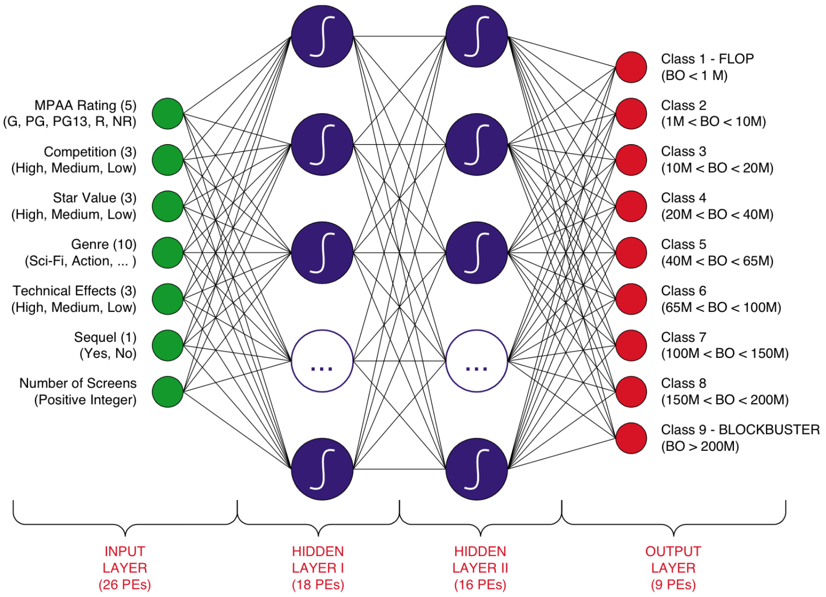
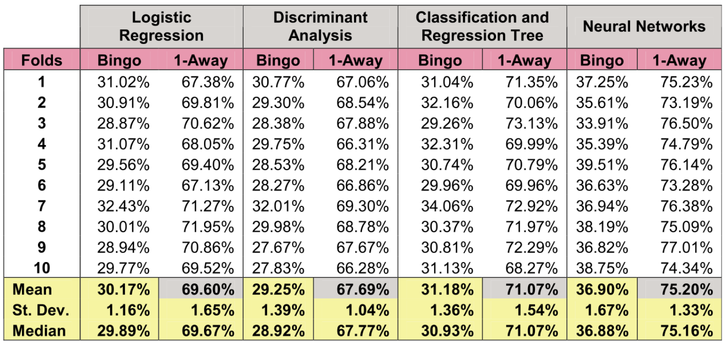


Figure 2.11: Sharda and Delen comparison of Various ML Techniques [28]



	Logistic Regression		Discriminant Analysis		Classification and Regression Tree		Neural Networks	
Folds	**Bingo**	**1-Away**	**Bingo**	**1-Away**	**Bingo**	**1-Away**	**Bingo**	**1-Away**
1	31.02%	67.38%	30.77%	67.06%	31.04%	71.35%	37.25%	75.23%
2	30.91%	69.81%	29.30%	68.54%	32.16%	70.06%	35.61%	73.19%
3	28.87%	70.62%	28.38%	67.88%	29.26%	73.13%	33.91%	76.50%
4	31.07%	68.05%	29.75%	66.31%	32.31%	69.99%	35.39%	74.79%
5	29.56%	69.40%	28.53%	68.21%	30.74%	70.79%	39.51%	76.14%
6	29.11%	67.13%	28.27%	66.86%	29.96%	69.96%	36.63%	73.28%
7	32.43%	71.27%	32.01%	69.30%	34.06%	72.92%	36.94%	76.38%
8	30.01%	71.95%	29.98%	68.78%	30.37%	71.97%	38.19%	75.09%
9	28.94%	70.86%	27.67%	67.67%	30.81%	72.29%	36.82%	77.01%
10	29.77%	69.52%	27.83%	66.28%	31.13%	68.27%	38.75%	74.34%
Mean	**30.17%**	**69.60%**	**29.25%**	**67.69%**	**31.18%**	**71.07%**	**36.90%**	**75.20%**
St. Dev.	**1.16%**	**1.65%**	**1.39%**	**1.04%**	**1.36%**	**1.54%**	**1.67%**	**1.33%**
Median	**29.89%**	**69.67%**	**28.92%**	**67.77%**	**30.93%**	**71.07%**	**36.88%**	**75.16%**

2.2.3 Gaussian Processes

A Gaussian Process (GP) is a class of model that defines a distribution over functions, with predictions taking the form of a full predictive posterior [31]. This allows them to be used to perform non-parametric regression as well as classification tasks while also providing true model uncertainty.

At their core, Gaussian Processes rely on Bayes Theorem (Equation 2.17), using Bayesian Inference to generate a predictive posterior distribution.

$$p(f|D) = \frac{p(D|f)\ p(f)}{p(D)} \tag{2.17}$$

Gaussian Processes have been shown to be equivalent to a one layer neural network with infinite width [33] and that the function approximated by a NN is a function drawn from a Gaussian Process [34]. In other words, a Neural Network is a realisation of a Gaussian Process.

Definition of a Gaussian Process

A GP is a continuous stochastic process with a set of inputs $\boldsymbol{X} = \{x_1, x_2, ..., x_N\}$ and a corresponding set of functions $\boldsymbol{f} = \{f_1, f_2, ..., f_N\}$ where any set of random function variables are distributed as multivariate Gaussian:

$$p(\boldsymbol{f}|\boldsymbol{X}) = N(\boldsymbol{\mu}, \boldsymbol{K}) \tag{2.18}$$

The marginal distributions are expected to be consistent:

$$p(f_1) = \int p(f_1, f_2) df_2 \tag{2.19}$$

Given a dataset, X, Y, a GP looks to predict the posterior distribution that captures the most probable parameters given the observed data.

$$p(\theta|X, Y) = \frac{p(Y|\theta, X)\ p(\theta)}{p(Y|X)} \tag{2.20}$$

The denominator term, $p(Y|X)$ is the marginal likelihood which normalises the posterior distribution. It is computed by integrating over all parameters weighted by their probabilities:

$$p(Y|X) = \int p(Y|\theta, X) p(\theta) d\theta \tag{2.21}$$

Gaussian Process Covariance

The covariance (or kernel) matrix, $\boldsymbol{K}$, is necessary when defining a GP prior, along with the mean (usually assumed to be zero), that defines the GP model. The covariance function is a crucial parameter that determines how well the model generalises.

A covariance function, $K(\boldsymbol{x}, \boldsymbol{x}^{'})$, describes the correlations between different points in the process. The covariance matrix must be positive semi-definite. There are several options for covariance functions such as Squared Exponential (SE) covariance, Polynomial covariance and Neural Network covariance. Covariance functions can be combined via summing, multiplication and convolution, to form more complex kernels [32].

Sparse Gaussian Process

Several considerations must be made when using a Gaussian Process:

- **Choice of Kernel.** The choice of kernel is a crucial hyperparameter as it determines how well the model learns useful features of data.
- **Computational complexity.** GP's require matrix inverses to be computed for the covariance matrix which has asymptotic complexity of $O(N^3)$ which makes dealing with large datasets impractical. Methods exist to perform approximate inference [35, 36].

A Sparse Gaussian Process (Sparse GP) is a class of model that enables GP's to be trained on large amounts of data. Initial forays into scaling GP's to work on larger datasets involved selecting subsets of a dataset [37] to approximate the true covariance matrix and then proceed to train the GP. Snelson and Ghahramani (2006) [36] proposed using 'pseudo-inputs' or inducing points where selected samples did not necessarily have to be part of the training data.

Titsias, 2009 [38] proposed constructing a variational approximation of the posterior by minimising the Kullback-Leibler (KL) divergence (Equation 2.22). This procedure is called Variational Inference (VI). During VI we approximate the posterior [60] from Equation 2.20 using a simple distribution, $q_\omega(\theta)$, and minimise the KL divergence between this distribution and the true posterior with respect to the parameters ω.

$$KL(q_\omega(\theta)||p(\theta|X,Y)) = \int q_\omega(\theta)\log(\frac{q_\omega(\theta)}{p(\theta|X,Y)})d\theta = E_q\left[\log(q_\omega(\theta)) - \log(p(\theta|X,Y))\right] \tag{2.22}$$

Minimising KL divergence maximises the Evidence Lower Bound (ELBO), a lower bound on the log marginal likelihood with respect to ω. The best ω that maximises the ELBO is given by Equation 2.23.

$$L_{VI}(\omega) = \int q_\omega(\theta)\log(\frac{q_\omega(\theta)}{P(Y|X,\theta)})d\theta - KL(q_\omega(\theta)||p(\theta)) \tag{2.23}$$

Most importantly, Sparse GPs have a much reduced computational complexity of $O(NM^2)$ where M is the number of inducing points, a hyperparameter that can be chosen.

Deep Kernel Processes

A Deep Kernel Process (Deep KP) combines the expressiveness of a Neural Network with the non-parametric flexibility of a Gaussian Process. Structurally, a Deep KP consists of a Neural Network input, to whose outputs a kernel function can be applied following which it is fed into a layer Gaussian Process that produces the final prediction. This means that rather than computing the covariance of raw data, the covariance is calculated on the latent representation output of a Neural Network. A Neural Net kernel can be applied to a Sparse Gaussian Process as well, when using large amounts of training data.

2.3 Model Fitting

Fitting a model to data is an important procedure that forms the basis of modern Machine Learning. The goal is to find a vector of weights for each parameter that enables the model to conform to the observations as best as possible. The extent of the fit can be defined in many ways and can include other parameters to penalise certain qualities of the weights vector such as LASSO Regression [17], the Akaike Information Criterion [18] (AIC) or the Bayesian Information Criterion [19] (BIC).

2.3.1 Least Squares Method

The approach that Least Squares takes is to minimise the sum of the squared errors between the model estimates and the observed data.

Least Squares can be represented as an optimisation problem with solution found by solving:

$$\underset{\theta}{\arg\min} \sum (y_i - f(x_i,\theta))^2 \tag{2.24}$$

Where y_i is the ith observed value and $f(x_i,\theta)$ is the ith prediction by the model. This is the same as using a Mean Squared Error (MSE) function.

Implementing and performing the computation for this is simple and allows us to find solutions for a system with more unknowns than variables. Least Squares provides the maximum likelihood solution and if Gauss-Markov Conditions apply, then it yields the best unbiased estimator.

Least Squares is especially sensitive to outliers as these can end up making a much larger contribution to the penalty. Least Squares also tends to overfit data, and using techniques such as LASSO [17] or Ridge [20] regression can reduce this.

2.3.2 Nelder-Mead Optimisation

Nelder-Mead [22] (NM) is a multidimensional global optimisation algorithm that can be applied to nonlinear optimisation problems where the derivatives may not be known.

NM works by creating a simplex (an n dimensional version of a triangle) containing $n+1$ vertices in an n dimensional problem. At every stage, NM moves the simplex towards an optimal region in the domain. In the last few iterations, this would shrink to the optimal point inside the simplex.

The stages in each iteration can be summarised as follows:

1. Sort points from worst to best and note the indices for the worst, second worst and best points.
2. Compute centroid for all but the worst point.
3. Transform the simplex as a reflection, expansion or contraction.
4. Redefine the entire simplex keeping only the best point, thereby shrinking the simplex.

The advantage of NM is that at each step it needs at most two function evaluations, therefore making it efficient compared to many other n-dimensional optimisation algorithms.

2.3.3 Confidence Intervals

Confidence Intervals are an important part of modelling as it indicates how reliable the estimate for a given parameter is. Normal approximation confidence intervals [29] provide a good indication of the probability that an interval contains the true value of the parameter:

$$estimate \;\pm\; (percentile \times SE(estimate)) \tag{2.25}$$

where SE is the standard error, $\frac{\sigma}{\sqrt{n}}$, with σ, the standard deviation, and n the number of samples.

2.4 Analysis, Interpretation and Evaluation

2.4.1 Coefficient of Determination

The Coefficient of Determination, R^2, is the proportion of the variance in the dependent variable that can be attributed to the independent variables. R^2 is commonly used as a measure of how well regression models manage to fit data and typically range from 0 to 1 with 1 indicating a perfect fit, although it can be negative in cases where $SS_{res} > SS_{tot}$. R^2 is defined as:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{2.26}$$

where

$$SS_{res} = \sum_i (y_i - f_i)^2 \tag{2.27}$$

$$SS_{tot} = \sum_i (y_i - \bar{y_i})^2 \tag{2.28}$$

2.4.2 Shapley Values

In game theory, Shapley values are a method for distributing costs and gains to players of a game in situations where contributions are not equal but where players work in cooperation to achieve the same end. This can be extended to find the average marginal contribution of a feature value across all coalitions and so find to what extent and how a feature contributes to the outcome.

The Shapley value of a feature represents its contribution to the prediction, weighted and summed [39, 40] over all feature combinations:

$$\phi_j(val) = \sum_{S \subseteq \{x_1, x_2, \ldots, x_p\} \setminus \{x_j\}} \frac{|S|!(p - |S| - 1)!}{p!} (val(S \cup \{x_j\}) - val(S)) \tag{2.29}$$

where S is a subset of the features used in the model, x is a vector of the feature values with p features. $val(S)$ is the the prediction for feature values in set S that are marginalised over features that are not included in the set:

$$val(S) = \int \hat{f}(x_1, ..., x_p) dP_{x \notin S} - E_x(\hat{f}(X)) \tag{2.30}$$

In practice, we have to exclude each possible feature from S and hence end up computing multiple integrals for each excluded feature. Performing this for even medium sized datasets is computationally expensive as there are 2^k possible combinations of feature values. We can estimate the Shapley value of each feature by performing Monte Carlo sampling [40, 41] and limiting the number of iterations by using a smaller subset of data to compute Shapley values.

2.4.3 Analysis of Variance

Analysis of Variance (ANOVA) is a collection of statistical models and their associated estimation procedures that can be used to analyse the difference in means of groups. One-way ANOVA is a version of ANOVA that is used to determine whether there are any statistically significant differences between the means of three or more independent groups. This means we can use ANOVA to check whether features separated by a categorical variable have means different enough to be significant.

As with many other hypothesis tests, a Null Hypothesis, H_0, is defined as:

$$H_0 : \mu_1 = \mu_2 ... = \mu_n \tag{2.31}$$

for n groups.

ANOVA makes several assumptions:

1. The dependent variable is normally distributed within each group.
2. Variances are homogeneous ie. the variances of each population group is the same
3. Observations in each population are independent

In practice, ANOVA is relatively robust against violations of normality.

Chapter 3

Feature Extraction

> Gentlemen, you can't fight in here! This is the war room!
>
> *Dr. Strangelove or: How I Learned to Stop Worrying and Love the Bomb, 1964*

3.1 Sourcing Data

Films have several defining features: some of the most well known features include the genre, the cast and crew and the budget. IMDb is a popular repository of film data, containing information on the core features of a film and maintaining a set of ratings where users are allowed to rate films on a scale of 1 to 10. Box Office Mojo, owned by IMDb, is another collection of film information that focuses on quantitative data on releases of a film in various regions/territories including the release date, the number of theatres showing the film and revenue on a more granular level − usually at weekly intervals.

IMDb provides open datasets [42] containing basic information on films, their initial release year and cast and crew names. However, this alone is insufficient: more data is needed on releases, revenues and theatres on a country by country basis, and data on revenue from non-theatrical sources.

Using films from the IMDb datasets, we scraped information from Box Office Mojo, TMDb and JP Box-Office using Python packages *Scrapy* and *Selenium*, ensuring that all prices were in USD. We then assembled the data into the pre-release and post-release datasets. We also collected several datasets containing history on actors, directors and production companies from The Numbers, which we use when feature engineering. In total, we use data containing information on 30 770 films. We summarise our data sources in Table 3.1.

Table 3.1: Summary of Data Sources

Data Type	Source
Genre, Director, Actors, IMDb Rating	IMDb
Region, Language, Runtime	IMDb, TMDb
Production Companies, Budget	TMDb
Revenue, Theatres, Release Dates	Box Office Mojo
Physical Media	JP Box-Office
Extra Information on Actors, Directors and Production Companies	The Numbers

3.2 Availability of Data

Revenues streams for films include the well known theatrical means, as well as income from sales of copies of DVDs and Blu-rays. Other means include rental or digital sales revenue from TVoD, as well as broadcasting rights from SVoD and TV. Data pertaining to revenue from VoD is not released by the market players nor is it available from third party, paid data sources. Similarly, data for revenue from TV rights sales too is not available. As a result, we use the data that is available to us: theatrical gross for several countries and sales of units and revenue from physical media.

3.3 Which Features are Considered?

We are faced with two distinct tasks: firstly, to make predictions on performance prior to a film's release; and secondly, to use data as it is collected post-release and make short term predictions on performance.

In both tasks we consider performance across several countries or regions and as a result the data used must be free of country-specific characteristics and nomenclature. For example, the use of an Motion Picture Association of America (MPAA) rating would be inappropriate and redundant as it is an organisation that (optionally) provides ratings for films released in the USA. Other countries have their own rating bodies, such as British Board of Film Classification in the UK and the Australian Classification Board for Australia. Rating criteria for each of these bodies is unique for each institution and no equivalent conversions between these exist. Furthermore, we consider many films that do not see release in countries with reported ratings and hence we ensure our models perform inference independent of such age/content ratings. We consider the core features described in Table 3.2.

Table 3.2: Overview of Core Features

Feature name	Feature description
Genre	The genres of the film, limited to the top three
Region	Region(s) in which the film originates
Language	The languages the film is released in
Director	The director(s) of the film
Actor 1	The lead actor in the film
Actor 2	The secondary actor in the film
Actor 3	The third leading actor in the film
Budget	The film's budget
Production Companies	The production companies involved in the film
Runtime	The total length of the film in its primary country of release
Country Release Date	Release date in each country
Country Beginning Theatres	Number of theatres showing the film per country
IMDb Rating	Community rating between 1 and 10 on the quality of a film

3.4 Data Visualisation and Analysis

In this section we analyse the relationship between some key factors and the total box office gross of films. We explore the importance of strategic relationships and implications of data.

3.4.1 Budget

Budget is a reasonable indicator of the quality of production of a film as well as the extent and scale of its marketing. The marketing budget does not form part of the film's budget, rather it is a separate expenditure that can sometimes be as large as the film's budget itself. A film's true budget can be hard to come by as many studios do not release this information and reported budgets are often just estimates. The marketing budget is even less available and although potentially informative, is not included in this analysis.

Figure 3.1: Budget vs Gross Plot

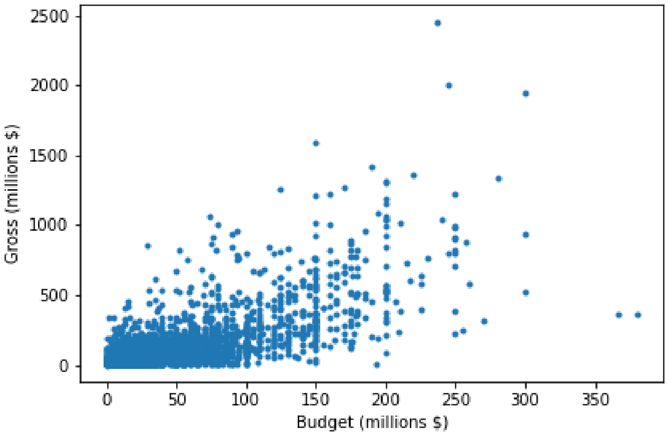


Figure 3.2: Budget Distribution

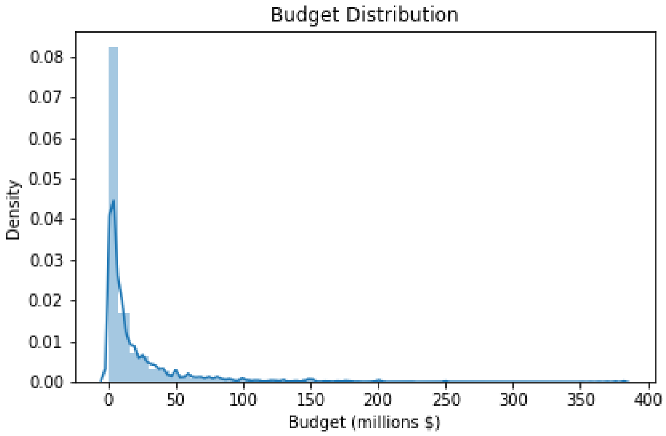


Figure 3.1 shows the plot of budget against the gross. There is a clear positive, although non-linear, relationship between the two. Looking at the distribution of budgets in Figure 3.2 we see that there are very few films with large budgets of over $100 million and the vast majority of films have a budget of less than $50 million which skews the distribution. Despite it being representative of real world film budgets, the skewness could result in there being very few samples of high budget film to model.

3.4.2 Genre

We now turn our attention to genre; Figure 3.3 shows the mean and standard deviation of the gross by genre. Adventure films and Sci-Fi films tend to gross more on average as compared with some genres such as Documentaries which see very small revenues. But what are budgets like for each genre? From the bar plot of genre and budget in Figure 3.4 we see that in general, the budget for each genre is proportional to the gross it generates on average.

Figure 3.3: Average Gross by Genre

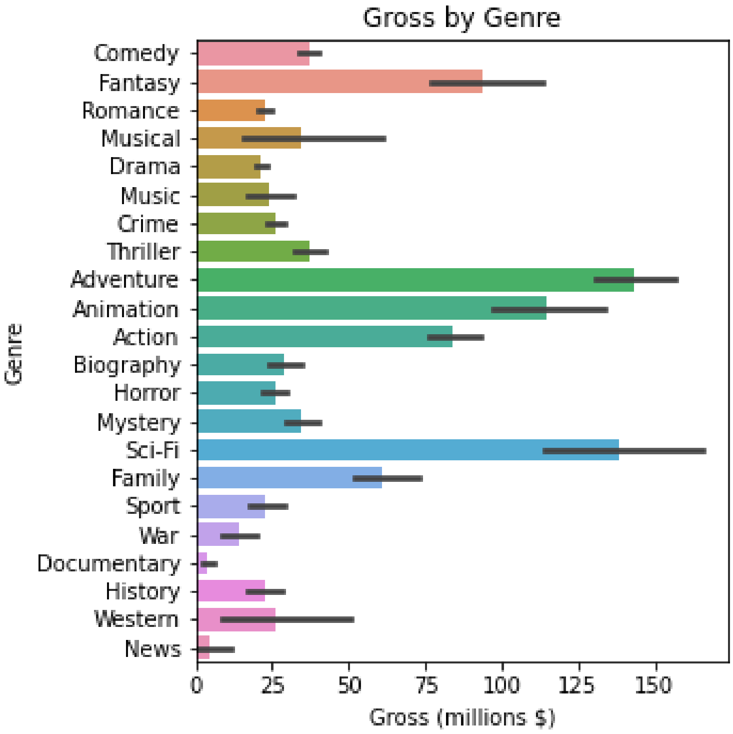
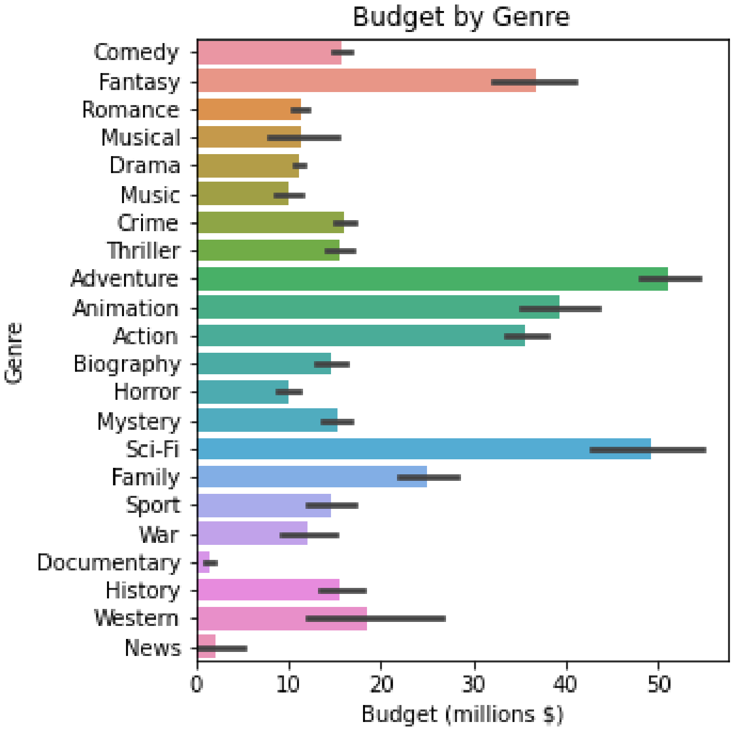


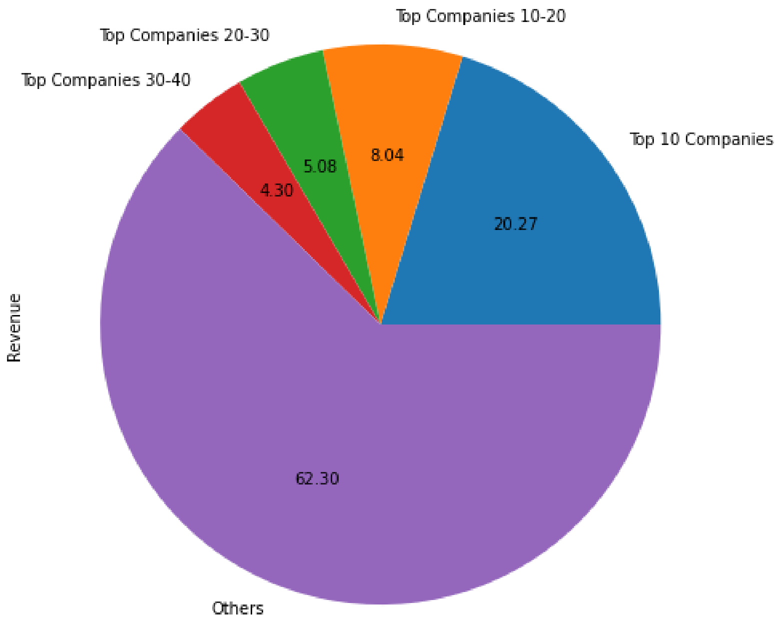
Figure 3.4: Average Budget by Genre



3.4.3 Production Company

Production companies are the powerhouses behind films, providing management, oversight, crew and then marketing the film to exhibitors. Involvement of a major studio is a reasonable indicator that a film will have a high production value and is expected to be successful. Looking at the market share of the top production companies in Figure 3.5 we see that the top ten production companies account for a disproportionate 20% of market share with the top thirty companies combined taking nearly a third of market share. This suggests that from this set of 10 340 companies, the involvement of a top production company is indicative of a high grossing film.

Figure 3.5: Market Share of Top Production Companies



Looking into the regions in which production companies are based, of the top thirty production companies, 90% are based in the USA and correspondingly films originating in the USA tend to have a higher average revenue.

We see similar trends with directors and actors too. Similarly, among 7296 unique directors there are a few, high performers that account for a much larger market share than the others. It is clear that the entities involved in the making of the film are a good indicator of financial success and incorporating a measure of past performance of these entities is likely to be a powerful and informative feature to predict box office success.

3.4.4 Theatres

As one of the feature inputs to the pre-release model, we provide the number of screens a film is exhibited in, in each country. Looking at how the initial number of theatres affects gross (Figure 3.6) we don't initially see any pattern. There are films that have made a lot of money despite releasing in relatively few theatres, just as there are films that have released in around 5000 theatres with some high grosses. There are also films that have released in around 25 000 theatres and yet made comparatively little! This occurs because we are looking at releases from all countries; some countries such as China and India have a large number of theatres but the final gross of these films may be low compared to the revenues of Hollywood films. We see peaks of profitability with a few hundreds of theatres as well with around 5000 theatres because of the nature of releases: **Wide** releases indicate high budget films that are exhibited in many screens at release and stand to make a lot of money and hence we observe a range of grosses, many high for releases of this nature. There are **Limited** releases where a film potentially has a lower budget and is screened in fewer theatres. A combination of the film's popularity and competition with other releases can result in it being more widely screened after release. Films that become more popular after release generate more revenue but are also rarities.

Next, we inspect the correlations between the number of theatres and the final gross per country in Figure 3.7 which displays correlations for a subset of countries. There is a positive correlation between the number of theatres a film is released in and the final gross of a film in that country. Also, notice that the grosses for some countries are highly correlated, such as between Italy and Germany. We learn more about these relationships and how they relate to modelling in Section 4.3.6.

Figure 3.6: Initial Number of Theatres vs Gross

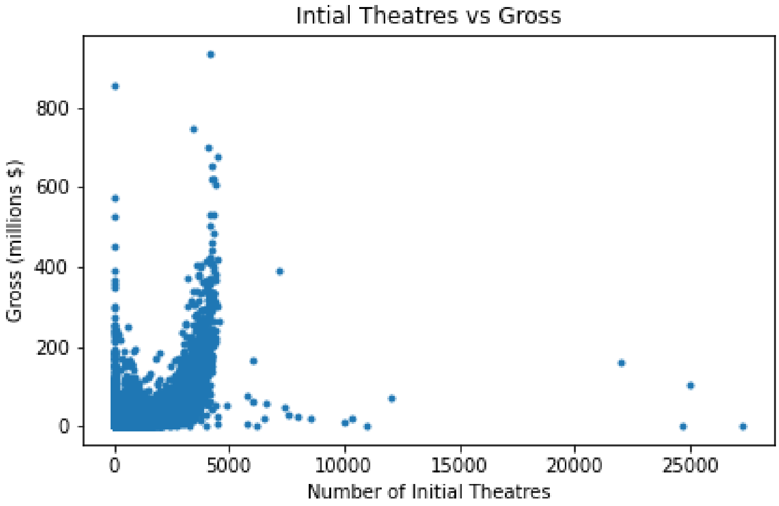
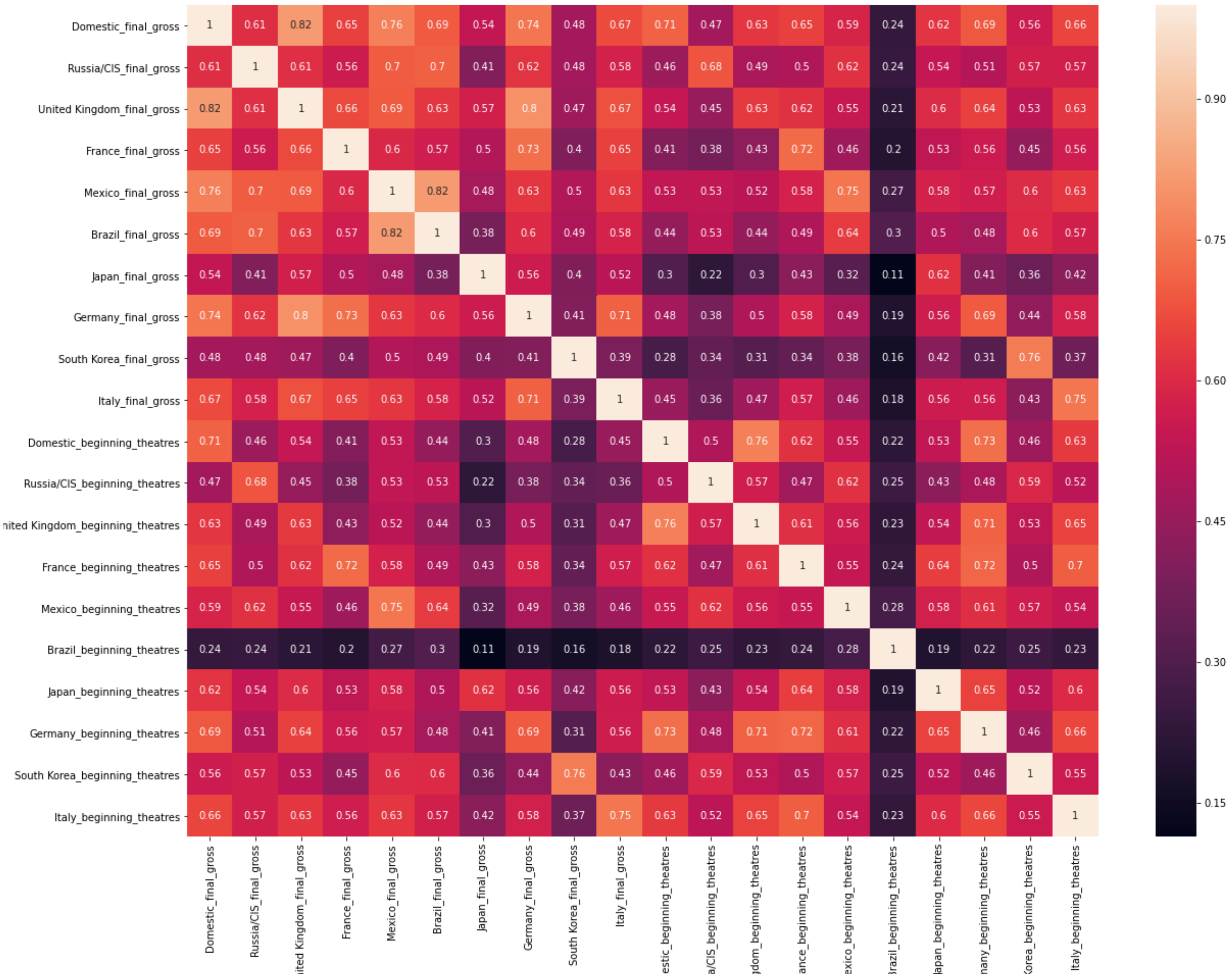


Figure 3.7: Starting Theatres and Final Gross Pearson Correlation



3.4.5 Year of Release

Included in the unprocessed dataset is the year of release of each film. We do not want to use year of release in our modelling as models should be able to perform inference independently of the year in which film was released. But before we remove this feature, we use ANOVA to explore the relationships between year of release and the grosses of films released in that year.

To perform ANOVA, we first check whether the assumption that groups are normally distributed is met. At a 5% significance level both the D'Agostino and Pearson, and the Jarque-Bera tests report that none of the groups are normally distributed. The distribution for the year 2001 is in Figure 3.8. Looking at the quantile-quantile (Q-Q) plot for 2001 (Figure 3.9), we see significant skew away from the mean. Applying a log transformation to gross and conducting the normality tests shows normality for almost all the years. An inspection of the distribution and Q-Q plots (Figures 3.10, 3.11) suggests a far more normal looking distribution.

Figure 3.8: Gross Distribution for 2001

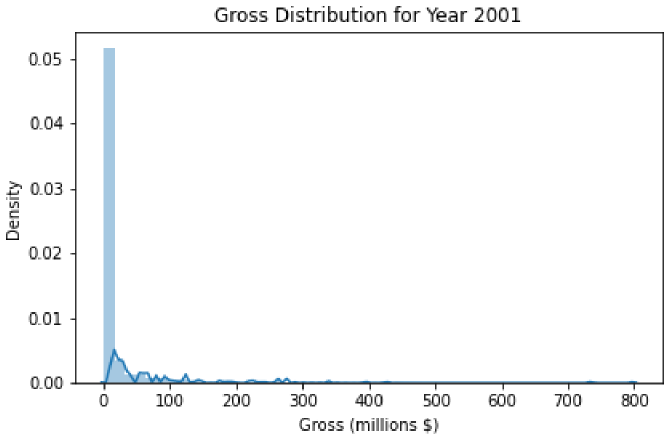


Figure 3.9: Q-Q Plot for Gross for 2001

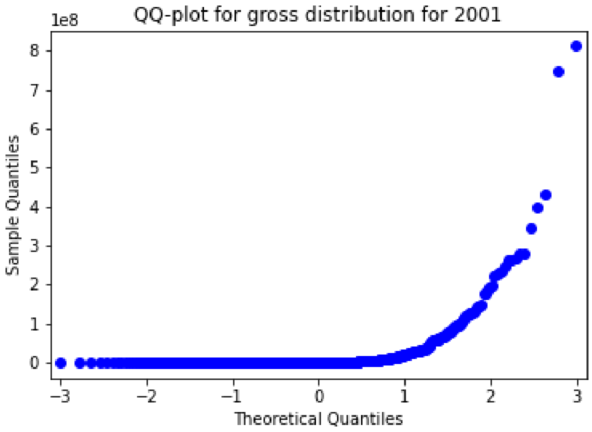


Figure 3.10: Log Gross Distribution for 2001

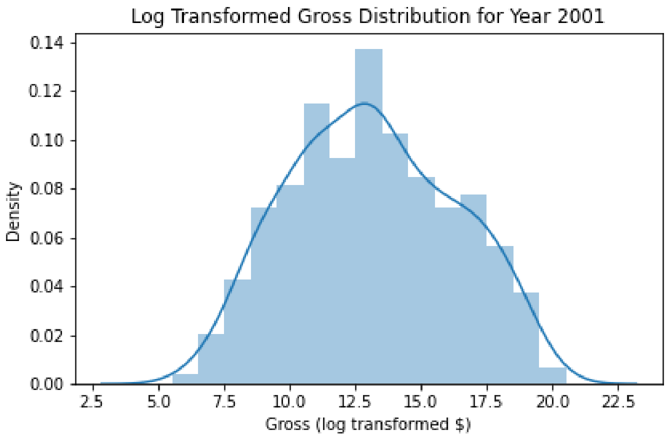
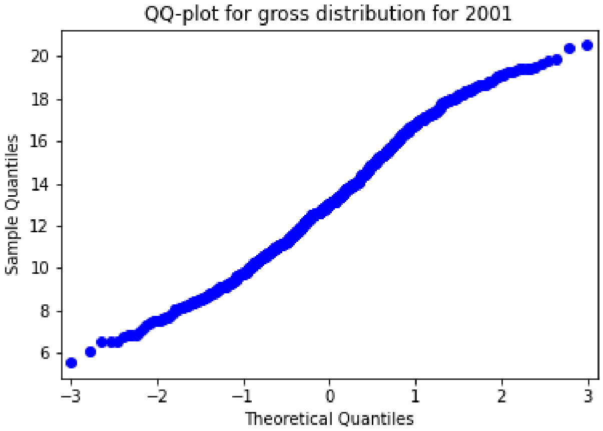


Figure 3.11: Q-Q plot for Log Transformed Gross for 2001



With the assumptions of ANOVA satisfied, we run the one way ANOVA test with a 5% significance level yielding an F-statistic of 20.314. The corresponding p-value is less than the 5% significance hence, we reject the null hypothesis and the means are not equal.

If the means are not equal then there must external factors that cause the mean to shift year on year. The most likely reason is inflation; inflation can be adjusted by discounting or compounding prices so that they are comparable.

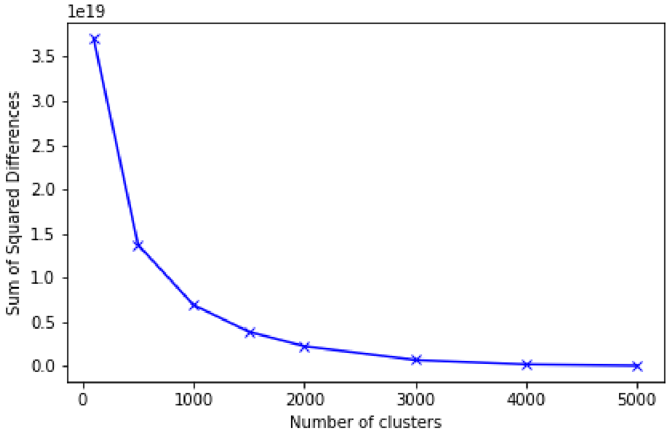
Discounting for inflation as described in Section 3.5.2 and repeating the one tail ANOVA test still results in a rejection of the null hypothesis which suggests that factors other than inflation are at play. These factors may come from films earning revenue from different markets and in different currencies. Currencies can have rates of inflation that are not the same as for USD leading to non-uniform changes in exchange rates. Growth of other film markets too can have an effect, such as the rapid growth of the Chinese Box Office [47] which has grown disproportionately compared to the US box office market.

3.4.6 Clustering Analysis

In Clustering Analysis, we pose an unsupervised learning problem and attempt to group sets of samples using clustering algorithms. Clustering analysis can provide an insight into the data as well as explore the feasibility of a clustering based approach to missing value imputation.

We minimise the Sum of Squared Distances (SSD) between samples and centroids using the K-Means algorithm using a subset of approximately 10 000 samples with non-zero budgets. Using the Elbow method, we see in Figure 3.12 that SSD decreases as the number of centroids increases. The optimal number of centroids seems to be at around the 2 000 mark where there will be an average of five samples per cluster. There are likely too few samples per centroid to perform missing value imputation. This also suggests that samples in this dataset do not form well populated groups.

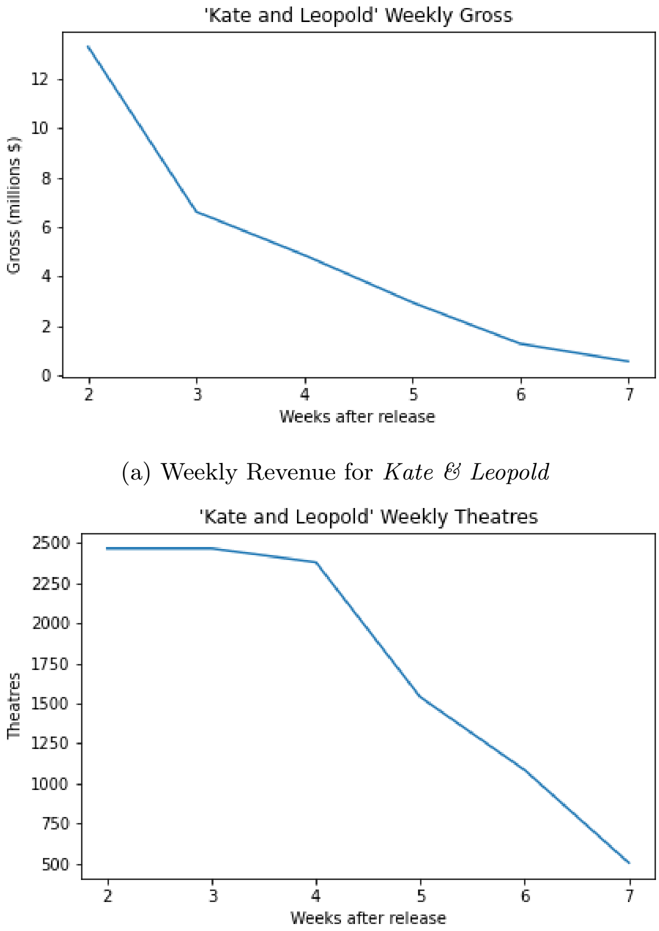
Figure 3.12: Clustering Sum of Squared Distances for K clusters



3.4.7 Post-release Time Series

When modelling post-release revenue and theatres exhibiting a film, we need to understand the nature of the series being forecast.

Film releases are generally of two types: wide releases and limited releases. A film with a wide release is released into a large number of theatres and typically is a high budget film with extensive marketing. *Kate & Leopold* is characteristic of this type of release with both the number of theatres and revenue per week again surpassing the release week's (Figure 3.13) numbers.

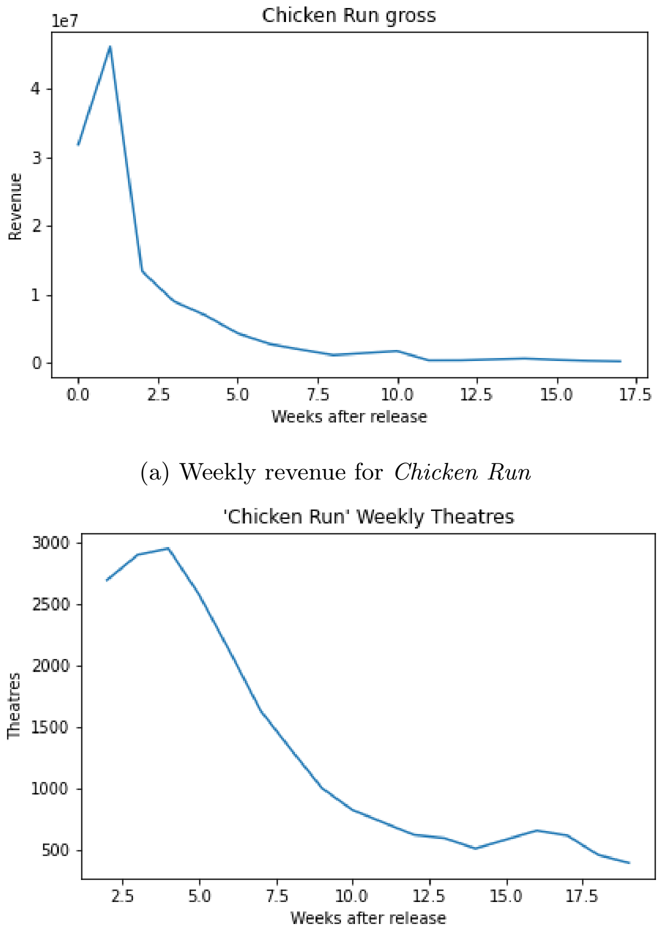


(a) Weekly Revenue for *Kate & Leopold*

(b) Theatres Exhibiting *Kate & Leopold*

Figure 3.13: Theatres and Revenue for *Kate & Leopold* in *North America*

A limited release, on the other hand, describes films released into a small number of theatres. A positive response and promise of high gross are likely to result in exhibitors increasing the number of theatres showing the film resulting in both the number of displaying theatres and revenue collected peaking after the initial release. *Chicken Run* (Figure 3.14) is an example of this type of release.



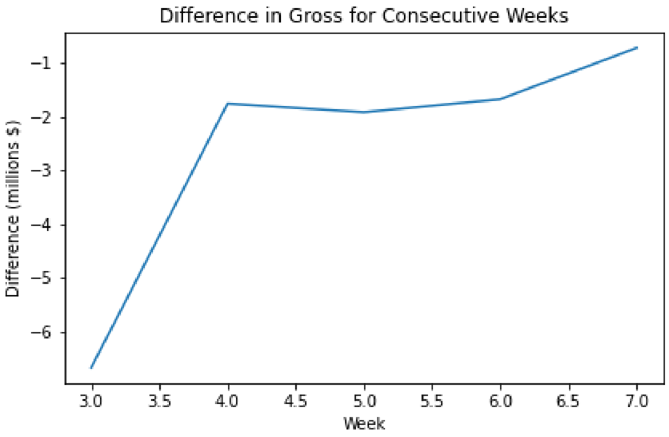
(a) Weekly revenue for *Chicken Run*

(b) Theatres showing *Chicken Run*

Figure 3.14: Theatres and Revenue for *Chicken Run* in *North America*

Neither of these release types are stationary time-series; even when applying the conditions for a Wide Sense Stationary (WSS) process, where we demand stationarity in the mean and variance only the raw time series does not satisfy. A common technique to obtain a WSS time-series is to compute the differences between raw values. But once again, the differences in Figure 3.15 are not WSS. This is, in fact, the case for all revenue and theatre sequences. Hence we conclude that we model only non-stationary time-series.

Figure 3.15: *Kate & Leopold* Differenced Weekly Revenue



3.5 Data Preprocessing

3.5.1 Dealing with Categorical Variables

In this dataset there are two types of categorical variables. Firstly, there are variables with a fixed upper limit on the number of possible states e.g. countries, languages etc. Variables such as these can be One Hot Encoded (OHE) as it is known beforehand the number of vectors needed to represent them.

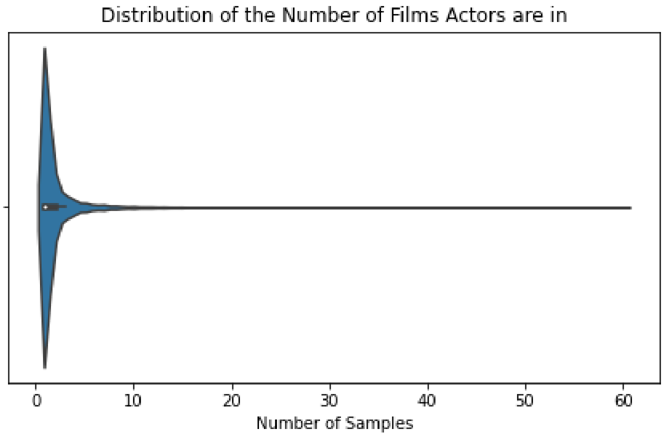
There are also variables with no such upper limit on the number of unique values such as the production companies, directors and actors. The number of unique values for these variables are summarised in Table 3.3.

Table 3.3: Number of Unique Values for Variables with No Upper Limit

Feature	Number of unique values
Production Companies	18 635
Directors	18 351
Actors	49 016

Using OHE this type of data could result in large, sparsely encoded matrices. An alternative technique, Label Encoding, does not cause large and sparse matrices. However, Label Encoding does not allow the model to learn relationships between features. Furthermore, these encoding methods often assume a large number of samples for each unique value which is not true for actors and directors where individuals may have only a few credits. We confirm this by looking at the Violin plot for credits for all actors in Figure 3.16 where the majority of actors, credited as either actor 1, 2 or 3, have been in fewer than ten films.

Figure 3.16: Violin Plot of Number of Acting Credits for all Actors



Multiple techniques have been proposed for handling encoding when faced with this type of problem [43, 44] such as encoding only the top n categories uniquely, with one more category dedicated to any values that do not fall into the top n. Another method proposed by Galvão & Henriques (2018) [44] is to split actors into 'stars' and 'non-stars' and encode actors as binary variables. This is a broad generalisation which ignores characteristics of individuals and does not allow the model to learn the characteristics of an actor.

An alternative to OHE for is to represent them with a measure of *star power* [45] or social media popularity [46] as a means of quantifying their appeal and cinematic performance. IMDb maintains an up to date *Starmeter* ranking system for actors crew going back over twenty years, but this data lacks granularity when looking at rankings from a long time ago.

We propose providing raw metrics to the model and let it form its own latent representation of how these metrics interact. We use secondary datasets collected from The Numbers, which contains information on the number of films and total box office gross for actors, directors and production companies. For actors and directors, we encode each actor and director with the number of films they have starred in or directed, respectively, as well as the average revenue generated per film. For Production Companies − for which there are often multiple companies involved per film − we encode each film using the total number of films released by all companies involved as well as the average revenue per film generated by all the companies. Examples of this encoding relationship are in Table 3.4. We indicate any actors or directors not found in the secondary dataset by a further column such as *actor_ 1_ is_ experienced* which is set to zero when an actor or director has no valid history.

The advantage of using this representation is that it does not result in the loss of information − classes of actors can still be recognised by models. More importantly, it does not have the disadvantages of an OHE approach. As actors are defined by their past information, and not uniquely without any measure past performance as with OHE, we do not face the penalty of large and sparse matrices or a lack of training samples for individual actors. Rather we expect a model to construct a latent representation of actors using their essential statistics. This technique can also be extended to suggest ideal actors from a selection of similar actors which we discuss in Section 7.3.2.

Table 3.4: Variable Encoding for Actors, Directors and Production Companies

Name	Total number of films	Average Box Office Gross ($)
Tom Cruise	42	238 449 585
Christopher Nolan	12	394 311 461
Warner Bros.	249	176 630 668

3.5.2 Adjusting for Inflation

Inflation is defined as *a continuing rise in the general price level usually attributed to an increase in the volume of money and credit relative to available goods and services* (Credit: Merriam Webster). The average annual rate of inflation of the US Dollar between 2000 and 2019 is shown in Figure 3.17.

Figure 3.17: Annual US Dollar Inflation Rate 2000-2019

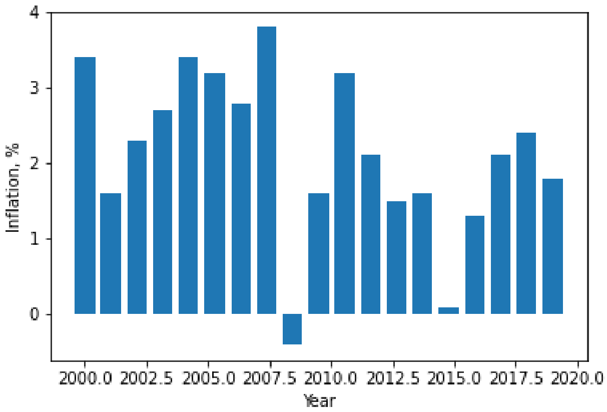
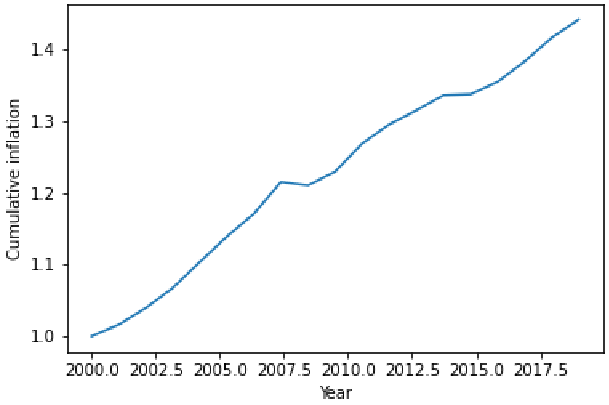


Figure 3.18: Cumulative US Dollar Inflation Rate 2000-2019



Our dataset contains films released from the year 2000. To ensure prices are comparable, we discount all monetary values to 2000's prices. We do this by calculating the cumulative rate of inflation using 2000 as the baseline price. The cumulative rate of inflation is shown in Figure 3.18. Prices can be corrected to 2000's prices by discounting with the formula:

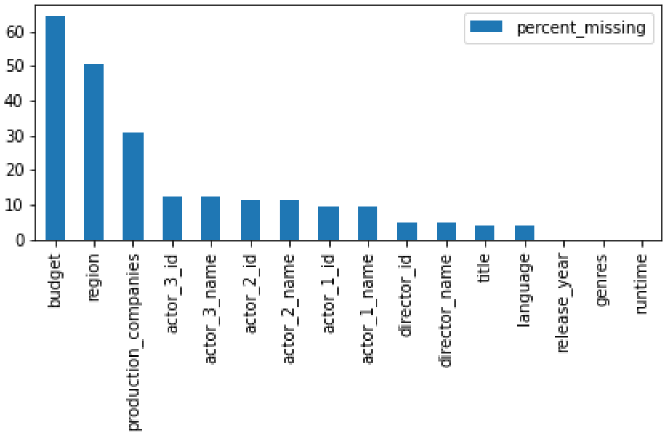
$$Corrected\ price = \frac{Price}{Cumulative\ inflation_{year}} \tag{3.1}$$

3.5.3 Missing Data

Missing Core Characteristics

Of the entire dataset, which contains 29 839 samples there are a few columns that have missing data. Figure 3.19 shows values missing as a percentage of the total number of values per column. The *region* and *production_companies* columns contain the second and third highest proportion of missing values. Missing values in these columns are inherently related as production companies produce films for the country they are headquartered in and as a result, we can impute some missing values in the *region* column.

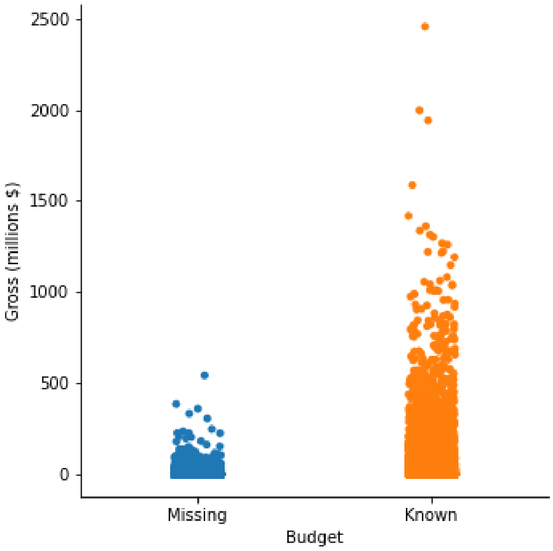
Figure 3.19: Percentage of Values Missing per Feature



Dealing with Missing Budgets

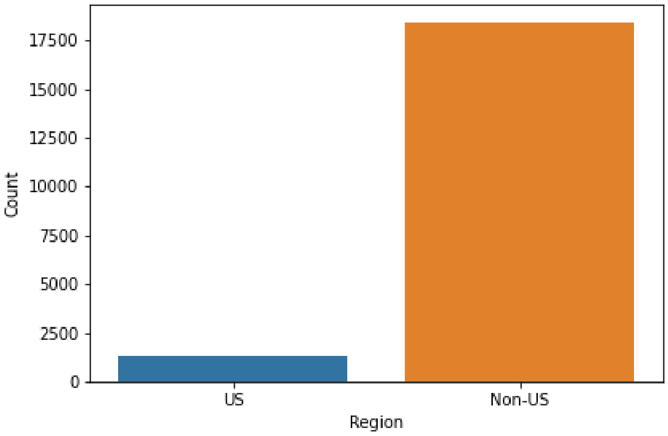
The *budget* column has the largest proportion of missing values – removing all rows with missing budgets would reduce the number of available samples by two-thirds. Removing all samples with missing budgets would also mean excluding an important subset of data. Figure 3.20 shows that the grosses of films with missing budget tend to be low − excluding these would mean ignoring a large and important demographic of the dataset.

Figure 3.20: Gross of Films with Missing and Known Budgets



Inspecting the missing budgets by region in Figure 3.21, we see that the majority of samples with missing budgets films not made in the USA. The aim of the project is to predict grosses for films generated from several countries and removing this data would lead to the loss of valuable information on non-US made films. Following the results of the clustering analysis in Section 3.4.6, we should not impute missing values using clustering either. Instead we zero impute these values and introduce an indicator binary variable, *budget_known*, which informs the model whether the budget is known for a particular sample or not.

Figure 3.21: The Regions of Films with Missing Budgets



Missing Physical Sales Data

Part of the total revenue generated by a film includes income from sales of DVDs and Blu-Rays. Information on the revenue from this income source, as well as the total (estimated) sales of units in North America was obtained from JP Box-Office. However, of all the samples in the dataset only 1100 samples have valid data for physical media sales. As a very small subset of samples contain this information, imputing sales data for over 95% of samples is not possible. Therefore, we introduce a binary variable, a feature that indicates whether a physical release occurred for a particular film. We only expect models to make predictions for samples which indicate a physical release.

3.5.4 Final Datasets and Notes

Using the feature engineering techniques outlined previously, we can create the pre-release and post-release datasets. We only model performance for a subset of countries (summarised in Table 3.5) for which we have sufficient data.

Table 3.5: Countries used in Modelling

Country	Pre-release	Post-release
North America	Yes (as *Domestic*)	Yes (as *Domestic*)
Russia/CIS	Yes	Yes
United Kingdom	Yes	Yes
France	Yes	No
Mexico	Yes	Yes
Brazil	Yes	Yes
Japan	Yes	No
Germany	Yes	Yes
South Korea	Yes	Yes
Italy	Yes	Yes
Turkey	Yes	Yes
Netherlands	Yes	No
Poland	Yes	No
Romania	Yes	No
Ukraine	Yes	No
Czech Republic	Yes	No
Slovakia	Yes	No
Norway	Yes	No
New Zealand	Yes	Yes
South Africa	Yes	No
Portugal	Yes	Yes
Bulgaria	Yes	Yes
Lithuania	Yes	No
Iceland	Yes	No
Slovenia	Yes	No
Australia	Yes	No
Spain	Yes	Yes
Taiwan	Yes	No
Belgium	Yes	Yes
Denmark	Yes	No
Sweden	Yes	No
Colombia	Yes	No
United Arab Emirates	Yes	No
Hong Kong	Yes	No
Hungary	Yes	No
Peru	Yes	No
Argentina	Yes	Yes
Finland	Yes	No
Austria	Yes	Yes
Greece	Yes	No
Singapore	Yes	No
Thailand	Yes	No
Chile	Yes	No
Malaysia	Yes	No
Lebanon	Yes	No
Bolivia	Yes	No
Uruguay	Yes	No

Pre-release dataset

We create two types of datasets – one containing raw prices and the other containing log transformed prices. Each dataset contains 412 features which we summarise in Table 3.6.

Table 3.6: Summary of Pre-release Data Features

Feature Name	Description	Scaling/Transformation
Budget	The film's budget (+ indicator)	S, LTS
Runtime	Film length in minutes	S
Region	Region of origin, 89 unique values	OHE
Language	Languages of release, 95 unique values	OHE
Country Release	Indicates release in a country	OHE
Country Beginning Theatres	Initial theatres into per country	S
Country Release Day	Day of week of release in country	S
Country Release Month	Month of release in country	S
Actor 1 Num Films	Actor 1 number of films	S
Actor 1 Average Gross	Actor 1 average gross	S, LTS
Actor 2 Num Films	Actor 2 number of films	S
Actor 2 Average Gross	Actor 2 average gross	S, LTS
Actor 3 Num Films	Actor 3 number of films	S
Actor 3 Average Gross	Actor 3 average gross	S, LTS
Director Num Films	Director number of films	S
Director Average Gross	Director average gross	S, LTS
Studio Films	Production Company number of films	S
Studio Gross	Production Company average gross	S, LTS

S = Scaled

LTS = Log Transformed then Scaled

OHE = One Hot Encoded

Post-release dataset

The post-release dataset shares many features with the pre-release one. Only features that are most relevant after release are present, such as the languages, budget, runtime and the actor, director and studio encoded variables. Each sample also contains the previous week's revenue, number of theatres, rank in the country and the IMDb rating for the film. The country for which the sample corresponds to is indicated by a One Hot Encoded set of variables. This results in 231 input features which we summarise in Table 3.7.

The time-series data for many countries contains missing values. As a result, several countries with significant amounts of missing data are removed. When a small amount of data is missing, we fill in the gaps using linear imputation.

Table 3.7: Summary of Post-release Data Features

Feature Name	Description	Scaling/Transformation
Budget	The film's budget (+ indicator)	S
Runtime	Film length in minutes	S
Country	Country to which current sample corresponds	OHE
Region	Region of origin, 89 unique values	OHE
Language	Languages of release, 95 unique values	OHE
Actor 1 Num Films	Actor 1 number of films	S
Actor 1 Average Gross	Actor 1 average gross	S
Actor 2 Num Films	Actor 2 number of films	S
Actor 2 Average Gross	Actor 2 average gross	S
Actor 3 Num Films	Actor 3 number of films	S
Actor 3 Average Gross	Actor 3 average gross	S
Director Num Films	Director number of films	S
Director Average Gross	Director average gross	S
Studio Films	Production Company number of films	S
Studio Gross	Production Company average gross	S
Previous Week Gross	Gross in the previous week	S
Previous Week Theatres	Number of screens in previous week	S

S = Scaled

OHE = One Hot Encoded

Quality of Data

Most of the data collected is from Box Office Mojo and IMDb where data for North American and European films is generally well reported and detailed. However, for films made in countries outside of these regions, data is far harder to come by and in some cases is incomplete. For example, we expect the Indian film *Ashoka the Great*[1] to have its largest market in India. However, only data for North America and the UK is reported and some data clearly is incorrect − for the third weekend of the theatrical run in the UK, the number of theatres showing the film is reported as 731 698. This is impossible because as of 2018, the UK had fewer than 800 film theatres (source: Statista) in the entire country! We filter out some values that are clearly anomalous during preprocessing.

Poor reporting of data for non North American and European countries is a common issue as some of these countries/regions do not have official box office tracking bodies. As a result, some of the largest film markets such as China and India are not included in final datasets.

[1] https://www.boxofficemojo.com/releasegroup/gr3742913029/

Chapter 4

Pre-release Modelling

> Pay no attention to that man behind the curtain!
>
> *The Wizard of Oz, 1939*

4.1 Motivation

Having discussed some modelling techniques and the nature of the data we are considering, we now begin modelling how a film will perform in advance of its release. Prior to release, there is no way of knowing how critics and film fans will react to a film and to further complicate matters, positive response from one group doesn't necessarily translate to a positive response from the other. We begin by exploring the profitability of films as a concept in and of itself before modelling performance at a granular level.

4.2 What Determines Film Profitability?

4.2.1 Introduction

The first question we ask ourselves is this: what makes a film profitable? Films need to earn approximately twice their budget at the box office [3] in order to be profitable. As discussed in Section 3.5.3, the pre-release dataset contains several samples with an apparent zero budget and so, for this task we consider films with non-zero budgets only.

Before considering the full regression problem, we tackle a classification task − predicting whether a film will be profitable. Most importantly, we look at identifying the Key Performance Indicators (KPI) of profitability.

4.2.2 Dataset Division

We define a profitable film as one that earns at least 1.8 times its original budget. Samples that satisfy this condition, the profitable films, are denoted with a 1, and unprofitable films are denoted by a zero. This results in a binary classification problem.

We shuffle and separate the dataset using 90% of data to create the training set, with the remaining 10% functioning as a test set. This results in approximately 9000 training samples and 1000 test samples.

Cross validation is a technique that can be used to assess how well a model generalises, or to determine optimal hyperparameters. It works by separating a dataset into k batches of which one batch is used as a test set with the remaining k*-1* batches used to train a model. The performance on the test batch is recorded and the procedure repeated where every other batch is used as a test set exactly once. This technique has a crucial disadvantage; k-fold cross-validation is computationally expensive as the model has to be retrained for every fold before testing. As a result, we do not use cross validation in this modelling task.

4.2.3 Modelling Profitability & Key Performance Indicators

Description of the Models

We consider two models: Logistic Regression and *XGBoost*.

Logistic Regression optimises a vector of $n+1$ parameters, taking in n inputs and uses one further linear offset parameter, β_0, to map the given features to a probability of profitability:

$$P(\mathit{profitable}) = \sigma(z) \tag{4.1}$$

where $\sigma(z)$ is the *sigmoid* function:

$$\sigma(z) = \frac{1}{1+e^{-z}} \tag{4.2}$$

and z is:

$$z = \beta_0 + \beta_1 x_1 ... + \beta_n x_n \tag{4.3}$$

where β_i is the feature weight for feature i.

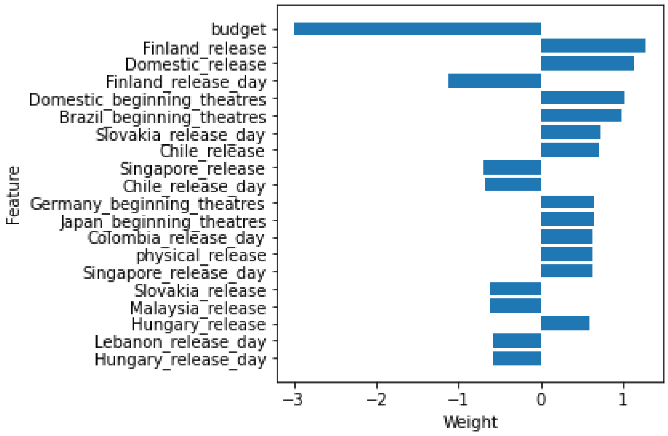
XGBoost is a tree based model used for supervised learning problems where the task is one of classification or univariate regression. *XGBoost* uses the concept of *boosting*, where several models are trained in sequence to fit on the error of the previous model and weighted sums of predictions of all models are used to generate an overall prediction.

Key Performance Indicators

In this section we are primarily concerned with gaining an insight into the factors that indicate profitability. We value simplicity of interpretation and this is a key reason for using transparent models such as logistic regression that, despite their often lower accuracy compared to more complex tree based models, offer clarity and interpretability.

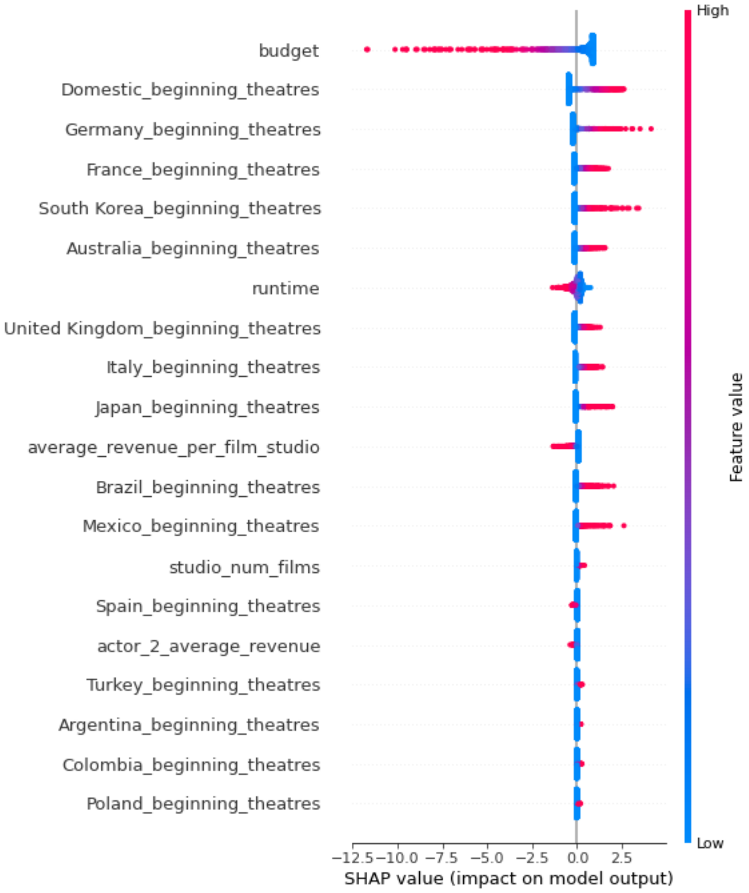
For the logistic regression model, the plot of the top twenty feature weights in Figure 4.1 show that *budget*, by far, has the largest impact on profitability. Many of the other important features are binary variables which indicate whether a film is released in a particular country, the most significant of which is the *Domestic_release* feature which indicates release in *North America*. This suggests that films stand to make a lot of revenue from the *North American* market and that this release here is a key consideration in profitability. Also considered important are the number of theatres the film is exhibited in initially in *North America*, *Brazil*, *Germany* and *Japan*. It comes as a surprise that the release indicator variables see a more significant weighting than the profiles of directors, production companies and actors. To try and explain this, we look to the field of Modern Portfolio Theory: Markowitz Minimum Variance portfolios can achieve returns with the lowest possible risk [56] and in the same vein, release into several markets indicates diversification that results in more consistent profitability.

Figure 4.1: Logistic Regression Feature Weights



The primary weakness of looking only at the coefficients of a logistic regression model is the lack of information on how magnitude affects the outcome − the feature weights would suggest that a *budget* of any magnitude is detrimental to profitability. We require a more granular interpretation of the model and for that we calculate Shapley values for the models.

Figure 4.2: Logistic Regression Input Feature Shapley Values

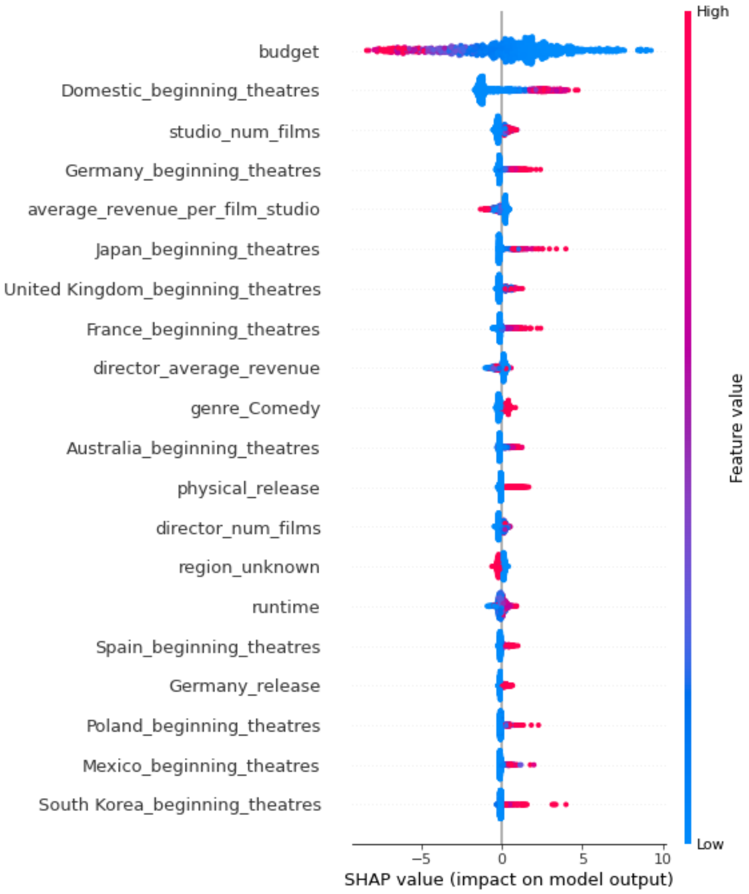


As we observed when inspecting the weights of the logistic regression model, *budget* has the most impact. On a more granular level we see that high budgets have have a large negative impact on profitability; this seems counter-intuitive as generally high budget films have more production value, more marketing, wider releases and make a lot of money at the box office. However, films with large budgets of over $100 million generate a higher proportion of revenue from non-theatrical sources[1]. As we fail to capture a large proportion of alternative revenue streams for large films, this can lead to the model believing that high budget films are less likely to be profitable.

The initial number of theatres a film is screened in, in key markets such as *North America* has a positive impact on profitability. We attribute this to the wide release effect (see Section 3.4.4) wherein in a film receives high exposure prior to release. Particularly surprising is the effect of the *average_revenue_per_film_studio* feature which suggests that films made by studios with large average revenues have a negative impact on profitability. This is likely linked to the high budget where such films often generate large portions of revenue from non-theatrical sources.

We shift our attention to the *XGBoost* model for which Shapley values are shown in Figure 4.3 which shows striking similarity with the Shapley values for the logistic regression model. Looking at the *budget* feature we see much more nuance – only extremely high budgets have a large negative impact on profitability. The *XGBoost* model also seems to understand the effect of wide and limited releases, with a more pronounced penalty for releasing in very few theatres in *North America.*

Figure 4.3: *XGBoost* Input Feature Shapley Values



4.2.4 Evaluation

We now answer the question: how well can our models predict the profitability of a film? We explore in detail how the models perform on unseen test data and compare them to models from previous work in the field.

[1] https://stephenfollows.com/how-movies-make-money-hollywood-blockbusters/

Comparing Logistic Regression and *XGBoost*

To compare logistic regression and *XGBoost* we consider two performance metrics: Logarithmic-Loss (log-loss) and Classification Rate (or Accuracy, used interchangeably). Figure 4.4 shows these performance metrics for the two models. *XGBoost* achieves a lower log-loss and a higher classification rate than logistic regression on unseen test data. Inspection of confusion matrices for the logistic regression and *XGBoost* models in Figures 4.5 and 4.6, shows the raw number of correct classifications and mis-classifications using which we calculate detailed performance metrics for each model in Figure 4.7. The recall rate for both models for the *Profitable* class is low suggesting that many films predicted to be profitable are in fact not so. This in turn results in the corresponding F1 ratio being low. Both model suffer from this problem which indicates that the models face difficulty in determining profitability from pre-release data alone.

Figure 4.4: Logistic Regression and *XGBoost* Model Performance

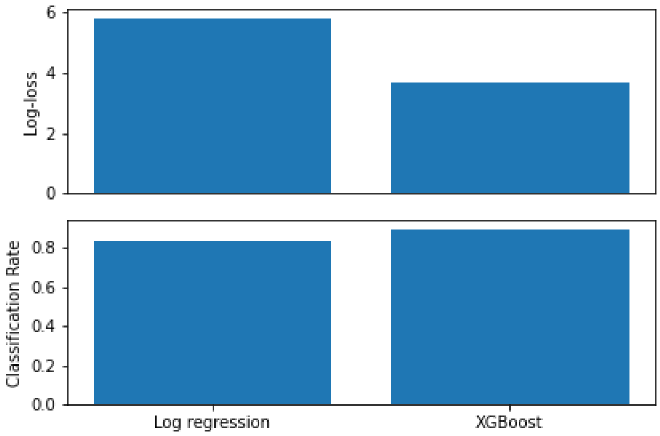


Figure 4.5: Logistic Regression Confusion Matrix

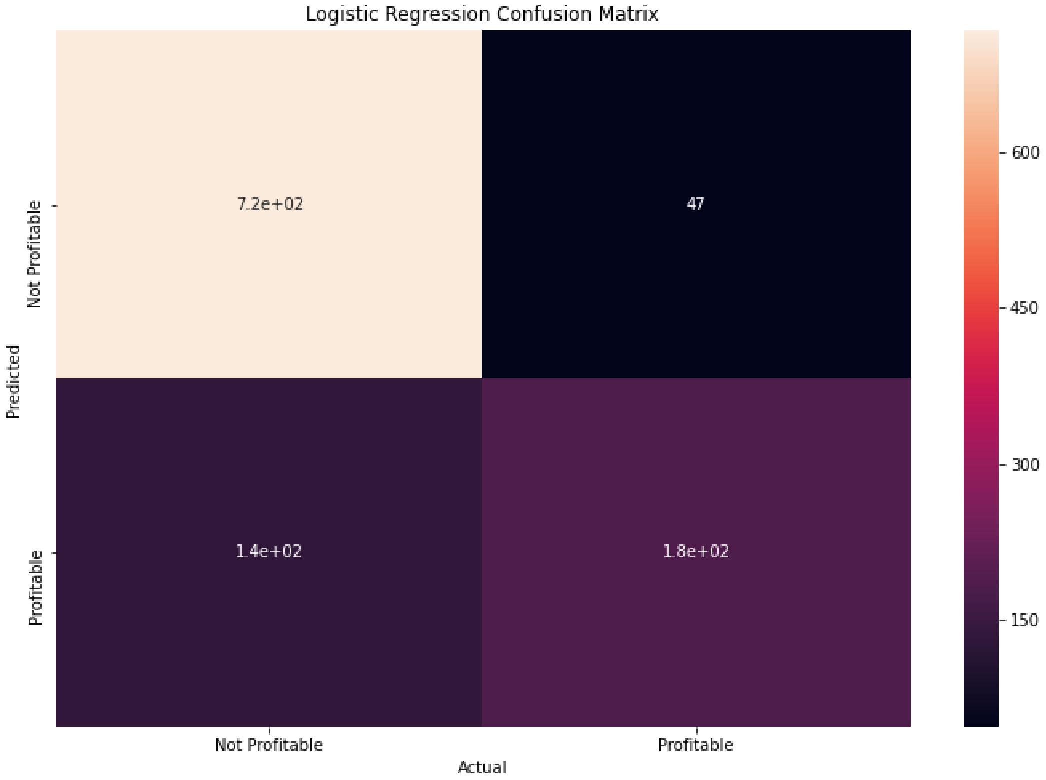
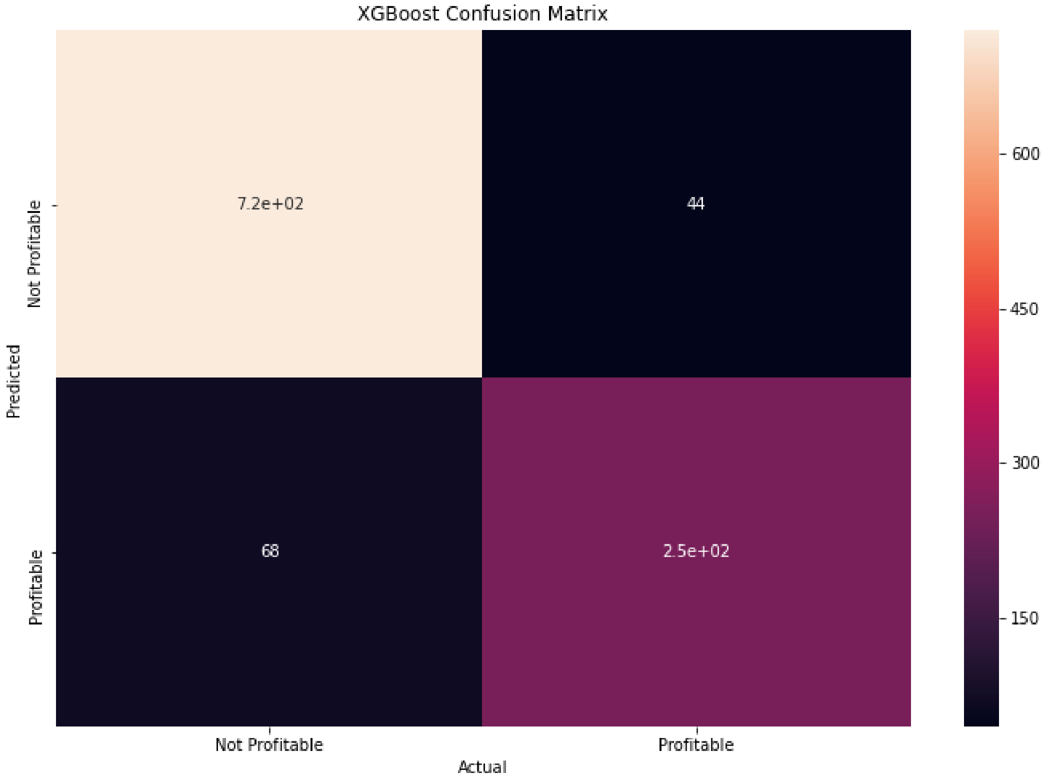
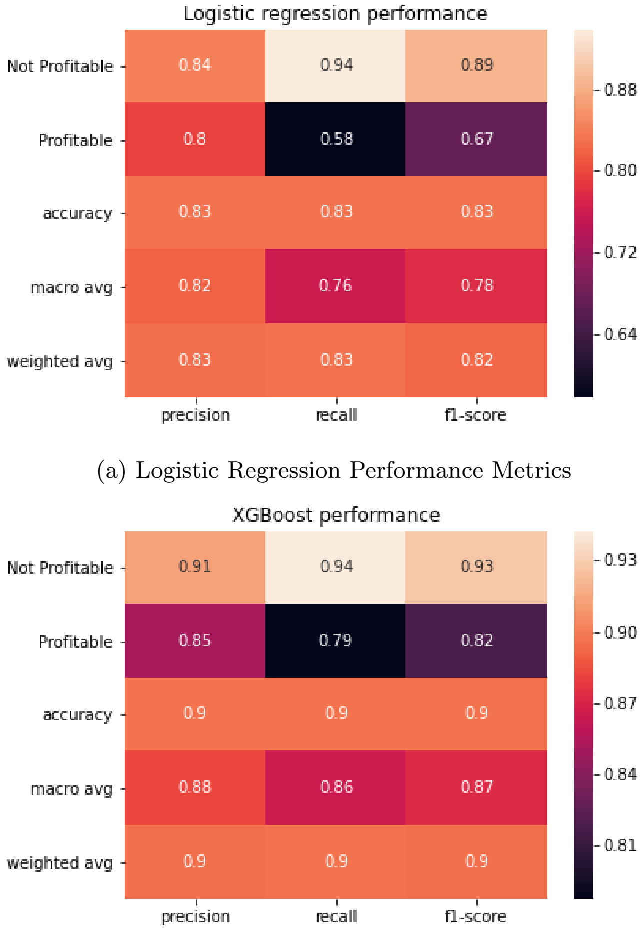


Figure 4.6: *XGBoost* Confusion Matrix





(a) Logistic Regression Performance Metrics

(b) *XGBoost* Performance Metrics

Figure 4.7: Profitability Models Performance Metrics

Comparison with Existing Work

We now consider how our model performs against those in literature; Rhee & Zulkernine (2018) [46] developed a Neural Network and a Support Vector Machine (SVM) to predict film profitability. Much of the data they consider is similar to the features used in this project. In addition they use some data only available post-release such as film ratings from website such as IMDb and Rotten Tomatoes. We summarise the comparison in Table 4.1. Considering we use no post-release information, our models fare well against Rhee & Zulkernine's.

Table 4.1: Our Profitability Models Comparison with Rhee et al.'s Models

Model	Number of Input Features	Classification Rate(%)
Rhee et al. NN	375	0.888
Rhee et al. SVM	375	0.842
Our Logistic Regression	411	0.832
Our XGBoost	411	0.897

4.3 Pre-release Regression

4.3.1 Introduction

We turn our attention to the core problem in pre-release modelling: regression. Models must make granular predictions for key sources of income at strategically significant stages. The results of these predictions can be used in conjunction with the terms of various agreements such as the *Rental Contract* to enable those involved in the filmaking process to calculate their own remuneration.

4.3.2 Target Description

The pre-release models take as input a vector, $\mathbf{x}$, containing the features outlined in Section 3.5.4. These features constitute essential information about the film as well as key factors such as the release date of a film in a particular country. Models must make several predictions, therefore, this is a multiple regression task. We provide an overview of the targets in Table 4.2.

Table 4.2: Outline of Pre-release Regressands

Name	Description
One Week Gross	Revenue generated per country one week after release
One Month Gross	Revenue generated per country after four weeks
Three Month Gross	Revenue generated per country after twelve weeks
Final Gross	Total revenue generated per country
Number of Weeks in Theatres	Number in weeks exhibited for per country
Physical Sales	Total units sold of physical media
Physical Sales Revenue	Total revenue generated from physical media sales

In total, this equates to 237 regressands.

When the expected theatrical run of a film in a country is, for example, more than four weeks and less than twelve weeks, we expect the three month gross and the final gross for that country to be identical.

4.3.3 Dataset Division

The complete pre-release dataset contains approximately 30,000 samples. We split data into training and test sets with 90% of data being allocated for training and the remaining 10% for testing. To ensure no information is leaked from training to test data, feature scalers are fitted on the training data and then used to transform test data.

We do not use multi fold cross validation because for k fold cross validation, each fold must be used as the test set once with the remaining *k-1* folds acting as training data. This results in the model having to be trained from scratch k times. Although cross validation has the advantage of being able to better assess how well a model generalises and offers lower variance for model evaluation, it has a large computational impact. This would be particularly expensive considering the models we use and the volume of data. We consider the number of samples we have to be sufficient to test generalisation performance and hence do not use cross validation.

We also consider another variation of the dataset: a version in which prices are log transformed before scaling. Applying a log transformation reduces the variance of the data and can reduce the skewness of some distributions.

As outlined in Section 3.5.2, all budgets and all box office predictions are in 2000's USD prices.

4.3.4 Overview of the Models

For this section we specify five models of increasing complexity: a Linear Regression Model, two Neural Networks, a Gaussian Process and a Deep Kernel Process.

Linear Regression

Linear regression is an algorithm that attempts to model relationships between independent and dependent variables by fitting an equation to observed data with n features.

$$y = \beta_0 + \beta_1 x_1 ... + \beta_n x_n \tag{4.4}$$

Linear regression is only able to model linear relationships between variables. We use this model for its transparency and to provide a baseline measure of performance. Most importantly, linear regression is able to find the exact solution to an equation which differentiates it from the other model considered.

Neural Network

Neural Networks can model more complex functions between inputs and outputs that may otherwise be ignored in a setting modelling only linear relationships. NN's have had a huge impact in modelling due to their ability to model complex relationships and handle sparse data.

Despite their advantages, NN's have some key disadvantages. NN's have many hyperparameters that need to be optimised such as the network architecture, learning rate and activation function(s). Determining optimal hyperparameters is itself an immense task with techniques such as Random Search and Bayesian Optimisation often used. NN's can also take a long time to train compared to linear regression as they use Gradient Descent or related forms of it to converge on optimal weights and biases for all neurons. This can occasionally lead to the optimisation algorithm becoming stuck in local optima with no guarantee of ever reaching the global optimum.

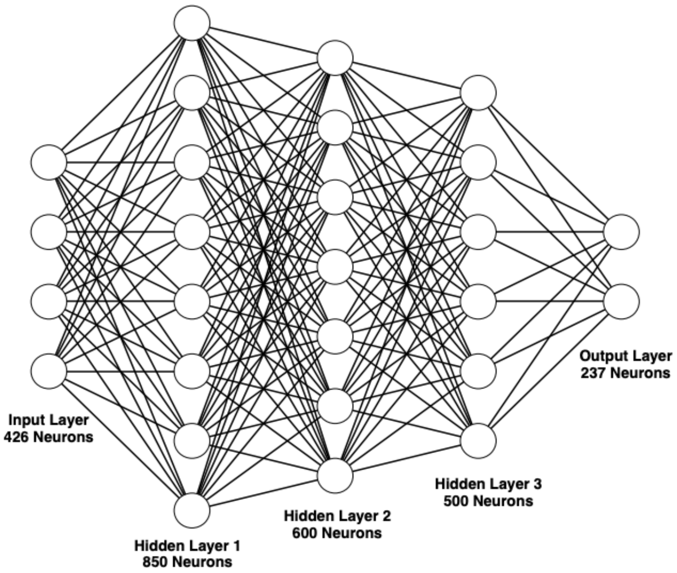
For this task we design two Neural Networks: one trained on the raw scaled data and the other trained on the log transformed, then scaled data. Both models employ identical architectures and hyperparameters.

The activation function we choose is the *Softplus* activation which applies the function

$$\sigma(z) = \ln\,(1 + e^z) \tag{4.5}$$

For both models, the error function used is the Mean Squared Error (MSE), with learning rate 0.0001 and a batch size of 256 samples per batch. Figure 4.8 depicts the architecture of our Neural Network models.

Figure 4.8: Pre-release Neural Network Model



To reduce over-fitting and to enable uncertainty estimation using MC dropout (refer to Section 2.2.2) we set dropout rate to 0.1 for all layers except the output layer for both models. When making predictions we use MC dropout, where dropout is enabled during inference and 200 predictions are obtained for the same data from which the mean and standard deviation are calculated, yielding a Gaussian distributed predictive posterior.

Gaussian Process

Gaussian Processes use a measure of similarity between points in training data, obtained using the kernel function to predict the value of an unseen point. The output takes the form of a predictive posterior Gaussian distribution, or in this use case, a multivariate predictive posterior Gaussian distribution as we perform multiple regression. The most important factor when using a GP is selection of the covariance kernel function, as discussed in Section 2.2.3. For our model we choose a Radial Basis Function (RBF) (also called a Squared Exponential Kernel) which takes two parameters: the lengthscale, l, and output variance, σ^2. The lengthscale determines the length of the perturbations of the function and output variance controls the average distance of the function from the mean, acting as a scale factor. Other kernels were tried such as a polynomial kernel, periodic kernel with limited success. We set the lengthscale parameter, l, to 0.1 and σ^2 to 2.0.

The computational complexity of fitting a Gaussian Process model is $O(N^3$ with N training points. This is infeasible with tens of thousands of training points, so instead we use a Sparse Gaussian Process (see Section 2.2.3) where the training complexity is reduced to $O(N^2M)$ where we choose M, the number of inducing points, to be 400. We maximise the Evidence Lower Bound (ELBO) and use a learning rate of 0.01.

This model is referred to hereafter as the GP model.

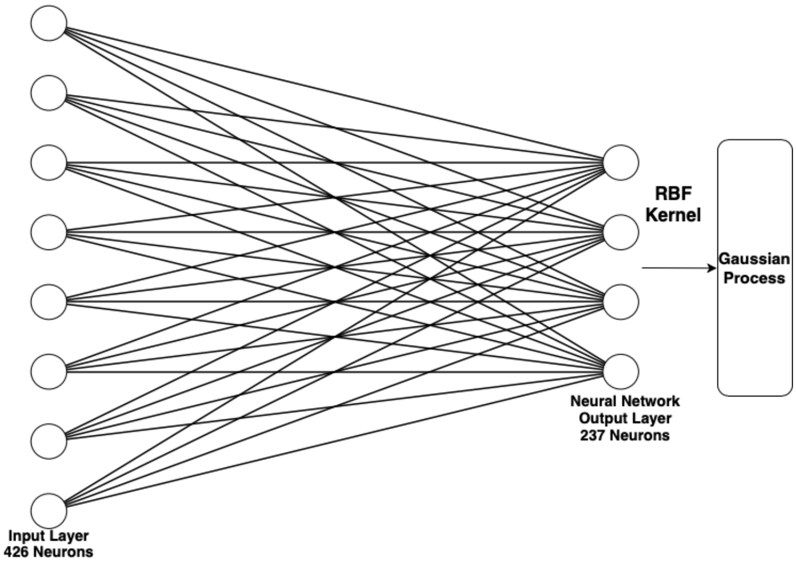
Deep Kernel Process

Deep Kernel Processes (Deep KP) combine the non-parametric capability of a GP and the expressiveness of a NN. The input feature vector $\mathbf{x}$ is fed into a Neural Network which forms and outputs a latent representation of the data. A kernel function is then applied to the representation offered by the NN before being fed into a Gaussian Process which performs final inference. Finding optimal hyperparameters for a Deep KP entails finding optimal parameters for both the NN input as well as the GP output's covariance kernel. For this model a single layer NN was used with a single hidden layer using a *Leaky ReLU* activation function:

$$f(x) = \begin{cases} x & \text{if } x > 0, \\ 0.01x & \text{otherwise.} \end{cases} \tag{4.6}$$

The NN layer is followed by an RBF kernel with lengthscale parameter, l, set to 0.1 and variance, σ^2, set to 1.0. We make use of a Sparse Gaussian Process output layer with inducing size set to 50 and a learning rate of 0.001. The Deep KP architecture is described in Figure 4.9.

Figure 4.9: Pre-release Deep Kernel Process Model



4.3.5 Results

Each model predicts 237 variables as outlined in Section 4.3.2. For clarity, we group individual predictions as described in Table 4.2 and visualise predictions for entire subsets.

Number of Weeks in Theatres

Knowing how long a film will be displayed in theatres beforehand allows distributors and exhibitors to make more informed decisions about the terms of the Rental Contract. It also allows beneficiaries of revenue to more accurately estimate how much money they are likely to earn from the box office gross.

Figure 4.10: Number of weeks *High Crimes* will play in theatres in each country, for a subset of countries. Linear Regression model on raw data.

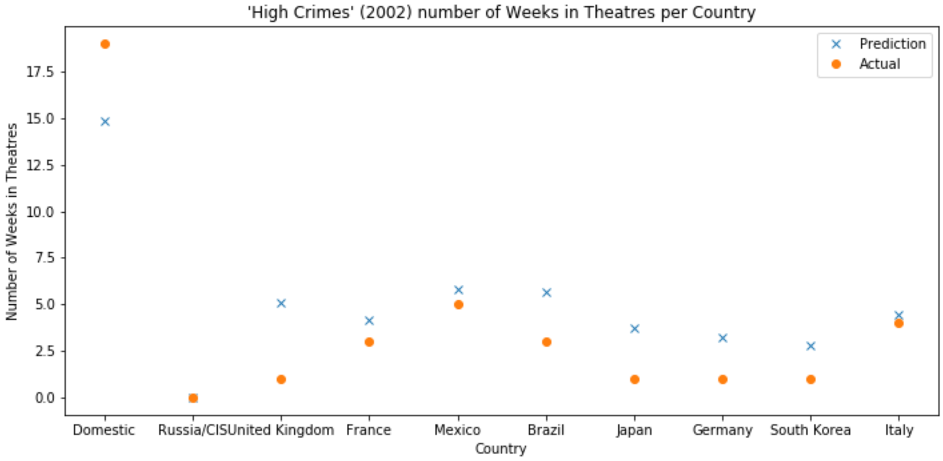


Figure 4.10 shows the linear regression model's prediction for how long the film *High Crimes* will stay in theatres for a subset of countries.

In general films rarely tend to be exhibited in theatres for more than twenty weeks. Continued screening of a film beyond twenty weeks is subject to a combination of factors such as public response to a film, releases of other major films and the value gained from continued showing.

Physical Media Sales

Next, we look at the revenue and units sold from sales of Blu-ray/DVDs in North America. Sales of physical media mark one of the highest margin sources of revenue for films and make up a large proportion of total revenue.

Figure 4.11: Revenue from Sales of Physical Media. GP model.

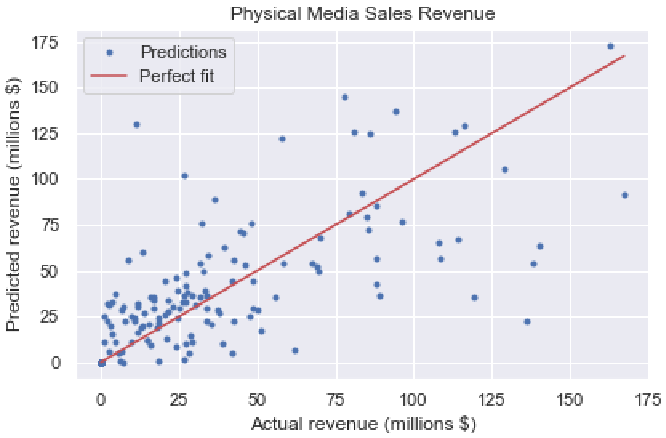
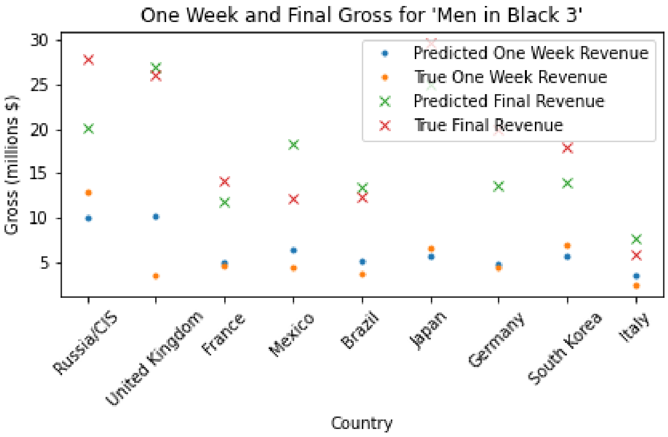


Figure 4.11 shows the true revenue plotted against the predicted revenue on test data with predictions made by the Gaussian Process model. Many predictions exhibit large errors which we put down to the very few training samples available for this regressand. We discuss this in further detail during evaluation.

Box Office Revenue

The models also predict box office gross at the one, four and twelve week marks, as well as the final gross. Figure 4.12 shows the predicted gross after one week and the final gross for the film *Men in Black 3* in a few countries.

Figure 4.12: *Men in Black 3* gross per country. NN model trained on raw data.



Prediction Uncertainty

Finally, we look at how the models can make predictions and characterise uncertainty. We inspect the North American final gross predicted by the NN raw data model on a few unseen films in Figure 4.13. Also shown are the 95% error bounds.

With knowledge of parameters of the distribution, the predictive posterior can also be plotted, as in Figures 4.14, 4.15 and 4.16. Using the parameters of a predictive posterior we can plot a distribution and state the likelihood of achieving certain values.

Figure 4.13: *North America* Final Gross predictions with uncertainty for unseen films. NN model trained on raw data.

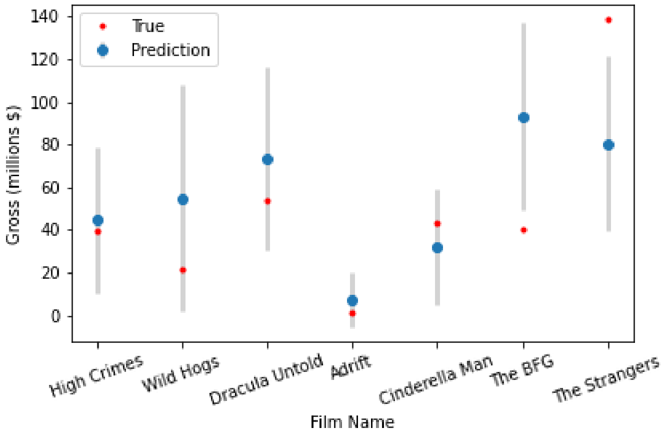


Figure 4.14: *High Crimes* Predictive Posterior

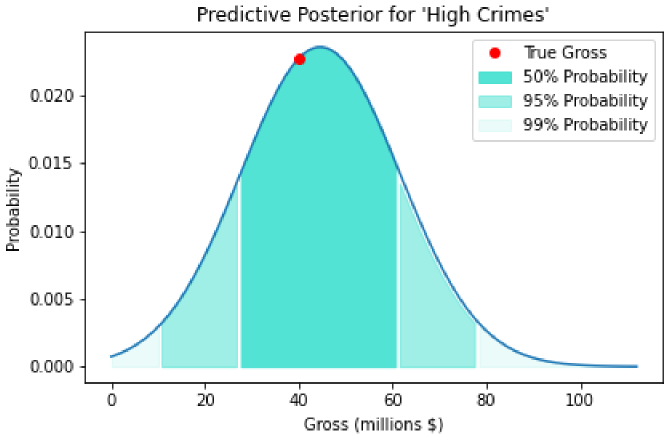


Figure 4.15: *Dracula Untold* Predictive Posterior

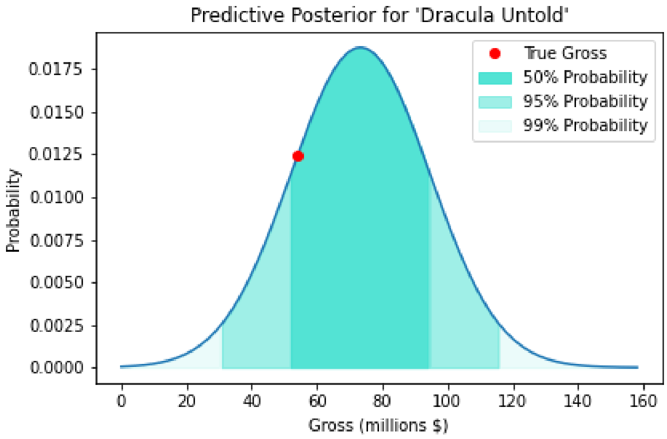
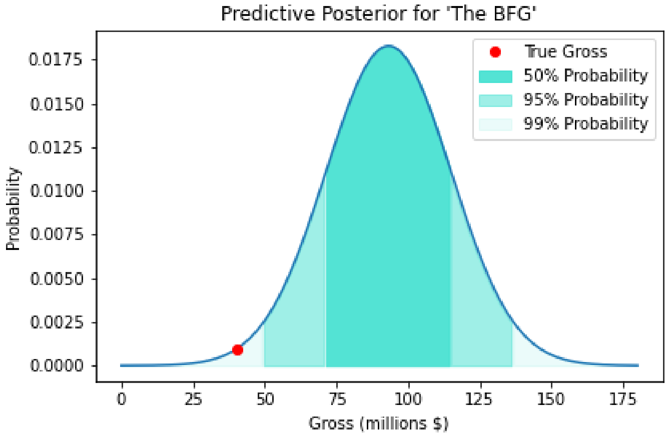


Figure 4.16: *The BFG* Predictive Posterior



4.3.6 Model Interpretation

Most of the models considered for this task are black box models and provide no insight into how they work. In a bid to understand more about the models and discover important relationships between variables, we compute and inspect the Shapley values for each output and identify key and sometimes surprising relationships between features. We display and analyse a few of the most interesting feature importance plots.

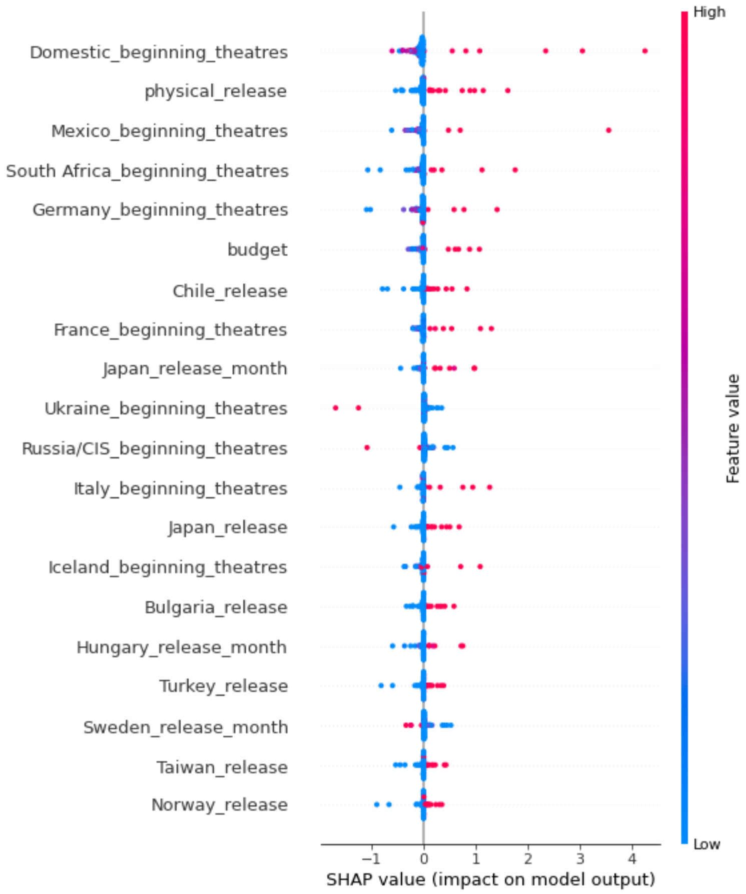
Model interpretation is performed on unseen test data as feature importance is often performed on model estimates of error. This means if importance is evaluated on training data, the model will appear to work better than it does in reality. Furthermore, overfitting on particular features in training data can yield incorrect estimates of true importance.

North American Final Gross

The *North American* market represents the largest box office by revenue, generating 11.4 billion dollars in 2019. For many films released in it, *North America* represents a significant source of revenue and creators make a concerted effort to appeal to and optimise for the *North American* audience.

We consider the feature importance for the final box office gross for *North America* plotted in Figure 4.17. This is obtained from the NN model trained on raw data.

Figure 4.17: *North American* Final Gross Feature Importance. NN model trained on raw data.

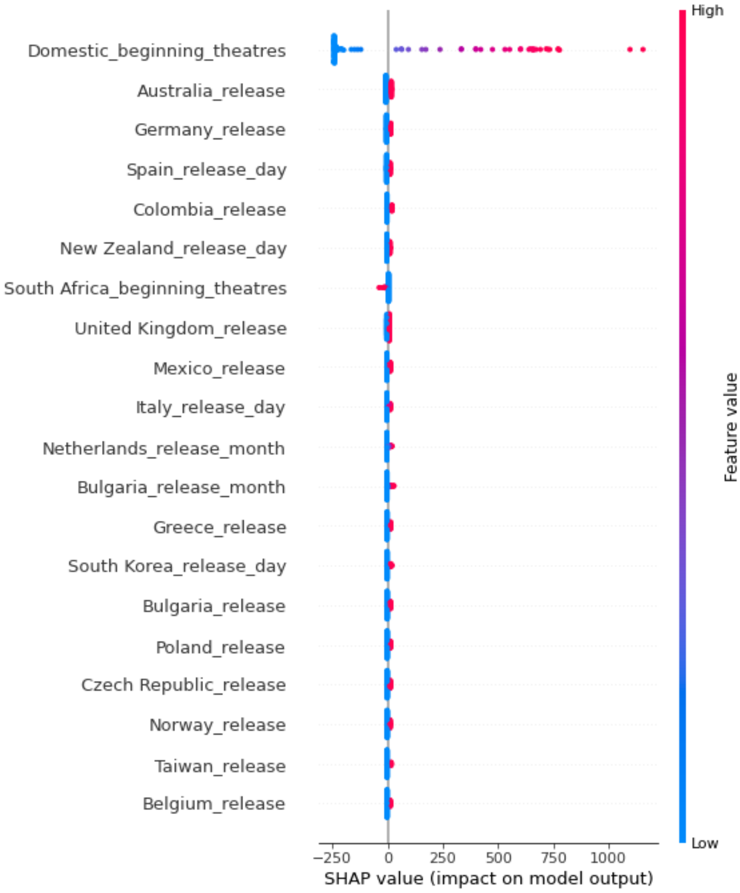


Domestic_ beginning_ theatres – the number of theatres a film is released in initially in *North America* – unsurprisingly, has the largest impact on the final gross. The number of theatres is representative of the nature of the release and by extension, the level of marketing and developed public anticipation for the film. The model also leverages relationships between markets in other countries, taking into account the beginning theatres in large film markets such as *Mexico Germany* and *Japan*.

Surprisingly, the *physical_ release* feature has a positive impact on final box office gross. This relationship may be due to the nature of the dataset – only a small number samples have valid data for a physical media release. These tend to be for films that already have high budgets and consequently are likely to have higher grosses. Acquisition of more data for physical media releases and further modelling would confirm whether these factors are truly related.

Contrasting these importances against the feature importances for the NN model trained on log transformed data in Figure 4.18, we see a significant difference. The *Domestic_ beginning_ theatres* feature is still regarded as the most important, but we see far fewer strong relationships between the final gross and other input features. This means that the log transformed NN model is not exploiting other relationships to the same extent.

Figure 4.18: *North American* Final Gross Feature Importance. NN model trained on log transformed data.



France Final Gross

Inspecting Shapley values for the final gross in *France* in Figure 4.19 yields insights into the relationships between geographically close markets. Similar to what we saw in *North America*, *France_ beginning_ theatres* has a significant impact on final gross, as does whether or not a film is released in French (*language_fr*) and to a lesser extent, English (*language_ en*).

The initial number of theatres in *Germany* has a significant impact on final gross in *France* which is surprising as *Germany* does not share a common language with *France*. This suggests a geographical relationship. But does *France* have a similar impact on final gross in *Germany*? Figure 4.20 shows the Shapley values for final gross in *Germany* where we see that this relationship is in fact reciprocated, as is *Germany*'s relationship with many other European countries.

In general, in both Figures 4.19 and 4.20 we see strong relationships between both geographically related countries (i.e. the European countries) as well as between regions that share similar languages, such as *Chile* (whose official language is Spanish, a European language) in the *Chile_ release* in Figure 4.19. The model also seems to infer relationships when countries are neither related by geography nor language – *Japan_ release* has an impact on *Germany's* final gross in Figure 4.20 which suggests that the film preferences of people and the markets in both countries are related. Identifying key relationships between seemingly unrelated markets can shed insight into how content creators can more effectively optimise for revenue across countries.

Figure 4.19: *France* Final Gross Feature Importance. NN model trained on raw data.

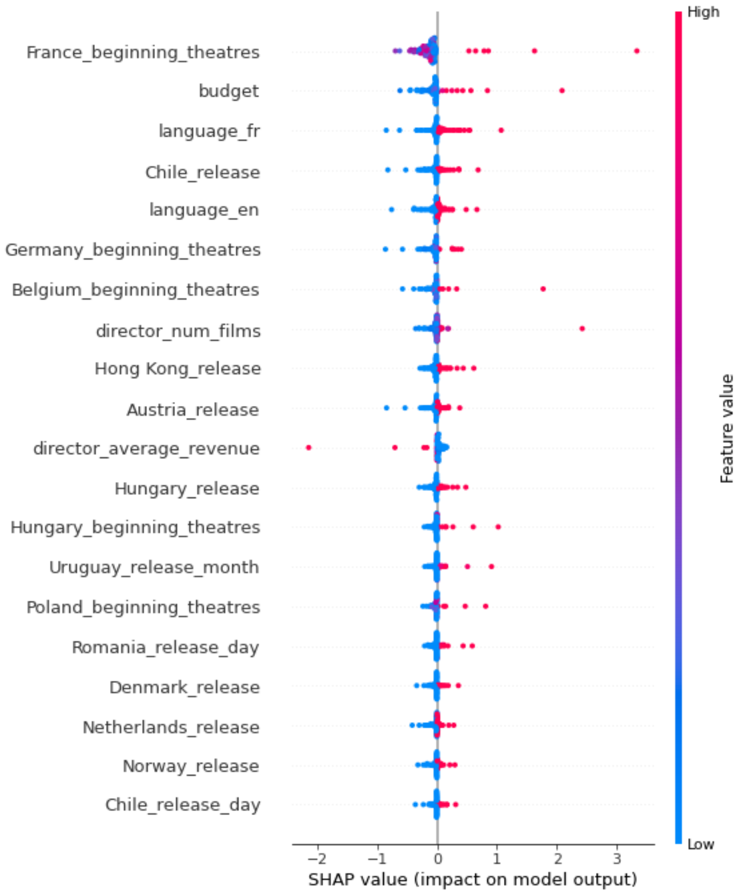
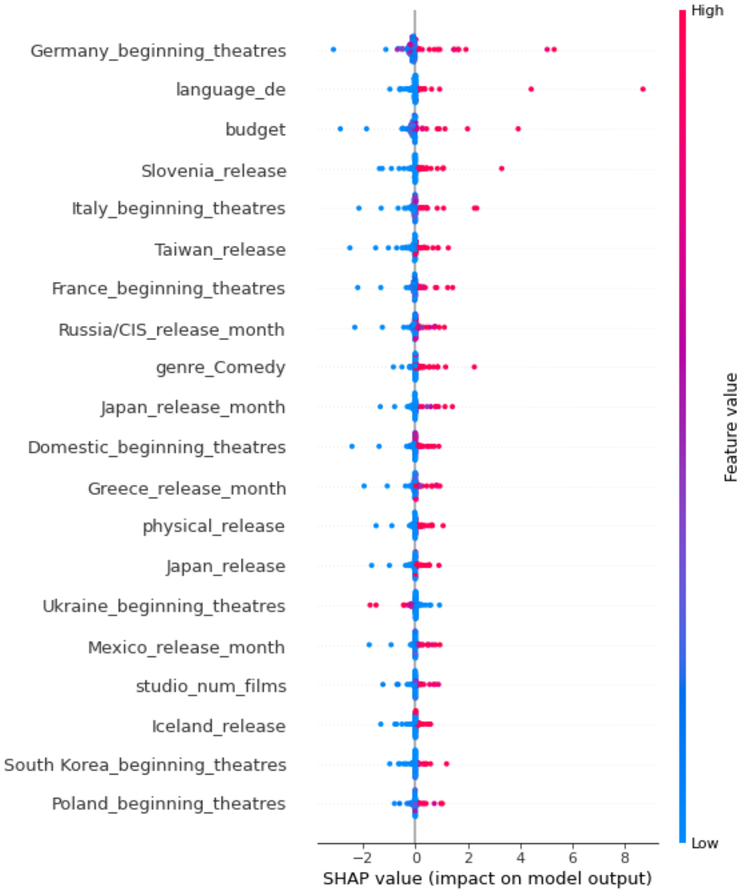


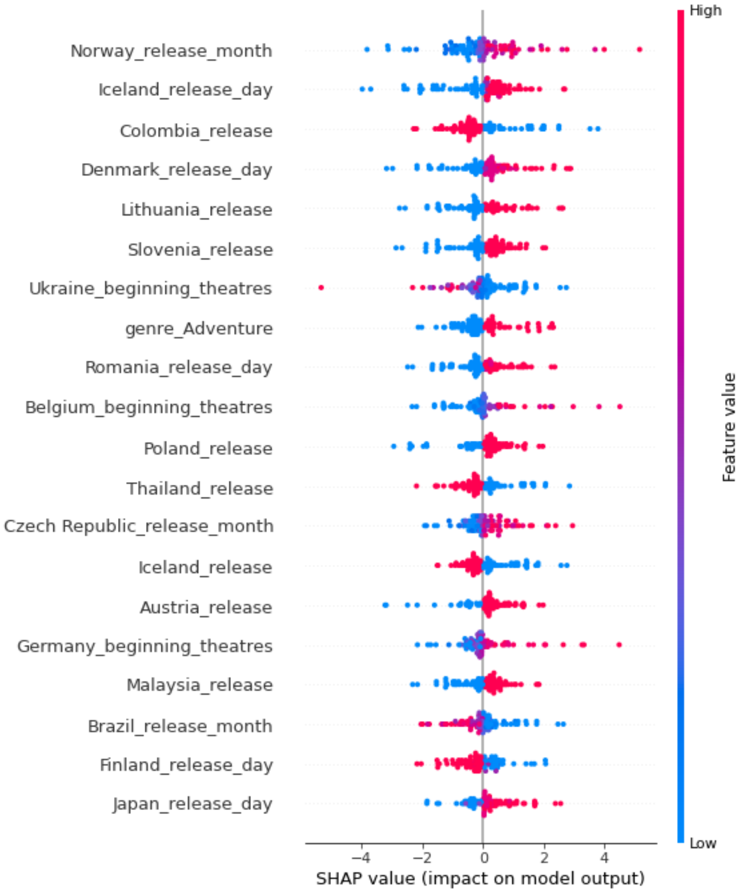
Figure 4.20: *Germany* Final Gross Feature Importance. NN model trained on raw data.



Physical Media Sales

Finally, we inspect feature importance for non-theatrical regressands. Figure 4.21 shows Shapley values for the total unit sales of physical media. There are many features, such as *Norway_release_month*, for which no logical relationship seems to exist with total sales. Perhaps most informative is the *genre_Adventure* feature which suggests that *Adventure* films tend to sell many units. *Domestic_beginning_theatres*, the initial number of theatres a film is released into in *North America* is not considered important by the model at all suggesting that there are too few samples of Blu-ray/DVD unit sales data for the model to identify and build relevant relationships.

Figure 4.21: Physical Media Sales. NN model on raw data.



4.3.7 Evaluation Methodology

Choice of evaluation metric is a key aspect of understanding the performance of a model. Different metrics can lead to different insights. We discuss the choice of performance metrics as well as implications of their use.

Evaluation Metrics

We use two evaluation metrics: the Coefficient of Determination R^2 and the Mean Absolute Error (MAE).

As explained in Section 2.4.1, R^2 provides a measure of how well observations are replicated by the model, based on the proportion of total variation of predictions explained by the model. R^2 treats over-predictions and under-predictions in the same way as both are considered equally as bad. R^2 is related to the Fraction of Variance Unexplained (FRU) which is the fraction of variance of the dependent variable not explained by input features by:

$$FRU = 1 - R^2 \tag{4.7}$$

MAE and RMSE provide measures of the average error per prediction. RMSE is given by:

$$RMSE = \sqrt{\sum_{i=0}^{n} \frac{(y_i - f(x_i))^2}{n}} \tag{4.8}$$

MAE is given by:

$$MAE = \sum_{i=0}^{n} |\frac{(y_i - f(x_i))}{n}| \tag{4.9}$$

MAE measures the average absolute error where all errors are equally weighted whereas with RMSE, larger errors have a higher weighting as errors are squared. Why use MAE over RMSE? RMSE does not increase with the variance of errors, it only increases with the variance of the frequency distribution of error magnitudes. As a result we can put an upper bound on MAE: MAE $\leq$ RMSE and more specifically, if all errors have the same magnitude then RMSE = MAE. This means MAE is a useful metric when we want a true and interpretable idea of the average error.

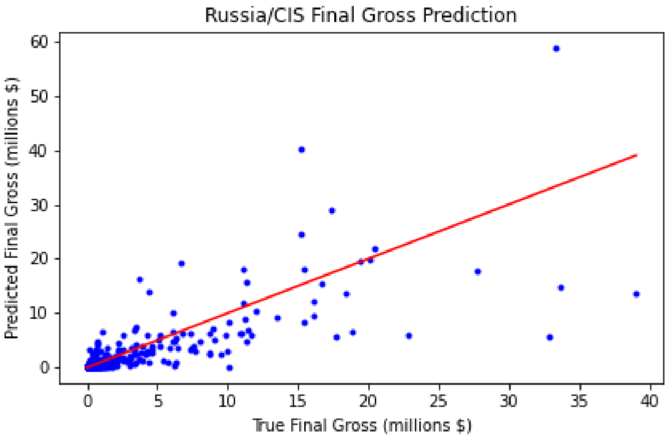
To ensure predictions made by all models are comparable, all predictions are unscaled and inverse transformed, if necessary, to yield raw values which are used to calculate performance metrics. As this is a multiple regression problem there are can be samples for certain outputs are not required. For example, some films may not be released in *Germany* and hence evaluating the error on these points would serve to skew the error metric. To prevent this we calculate the metrics only on samples where the true value is not zero.

A Note on Log Transformed Data Models

Applying a logarithmic transformation to a feature reduces the variance of the feature. This can lead to different interpretation of data as seen in Figures 4.18 and 4.17. After a such a log-predicting model makes a prediction, to obtain raw prices the predictions must be unscaled and exponentiated which can magnify the effect of small log prediction errors.

The NN trained on log transformed prices is especially affected by this as small log term errors can end up amounting to a large absolute error when exponentiated. This manifests as predictions that are very large and far from the expected value. The R^2 metric is especially susceptible to the presence of even a few large errors and this can result in negative R^2 terms for a set of points. An example of the log data trained NN model displaying this behaviour is shown in Figure 4.22.

Figure 4.22: Russia/CIS Final Gross. NN Model trained on log transformed data.



To ensure evaluation graphs remain clear, for some graphs containing an R^2 less than zero a threshold is applied and such values are raised to zero. We state whenever we apply this.

4.3.8 Evaluation

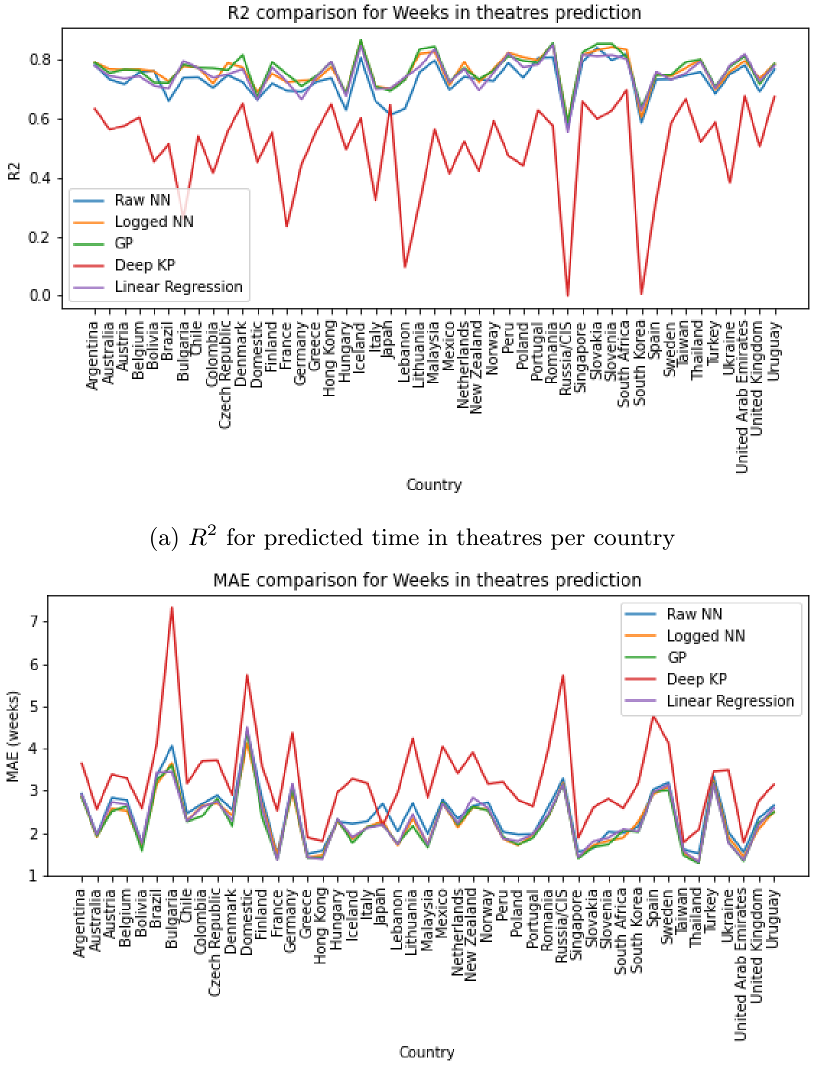
Having trained five different models, four on scaled data with raw prices and one on log transformed and scaled price data, we must find how well the models perform. Each model is trained on the same training samples and evaluated on the same previously unseen samples. Where available, we also compare the performance of our models against those developed in previous work.

Comparing hundreds of regressands individually would be cumbersome so instead we compare groups of similar variables as outlined in Table 4.2. For categories containing a large number of individual regressands, we plot the error metrics and proceed to summarise the comparison using metrics calculated over the whole class.

We extract and use only the specific regressand from our models that are considered by the models from previous works, ensuring we are directly comparing regressand performance.

Number of Weeks in Theatres

We begin by evaluating the predictions for how long a film will be exhibited in theatres per country in Figure 4.23. The Deep KP model exhibits a significantly lower R^2 than the other models suggesting more unexplained variance. In fact, there are countries for which the R^2 of the Deep KP model fall below zero, such as *South Korea*. These have been thresholded at zero for clarity as described previously. An R^2 less than zero suggests an exceptionally poor fit which is reflected in the MAE comparison where the Deep KP model suffers from a higher error than other models for almost all the countries.



(a) R^2 for predicted time in theatres per country

(b) MAE for predicted time in theatres per country

Figure 4.23: Performance metrics per country for number of weeks in theatres

Except for the Deep KP, models show consistent MAE and R^2 performance across most countries. There is a significantly higher MAE for *Domestic*, the *North American* prediction, which can be attributed to both a larger range of values present for *North America* and the fact that the true length of exhibition, especially for non-wide release films is heavily dependent factors such as competition for screens, awards, word of mouth, reviews and perceived value of exhibition by exhibitors. These factors are only available after theatrical release.

We summarise the overall fit metrics for predicting the number of weeks in theatres in Table 4.3 where we confirm the poor performance of the Deep KP model and observe the other models performing very similarly in both metrics. The linear regression model is able to compete and perform well against more complex models.

Physical Media Sales

We next evaluate performance on the regressands with the fewest available training samples: revenue and sales of Blu-ray/DVDs. Having very few training samples can lead to problems in the ability of a model to generalise well to unseen data as it hasn't had enough training to learn the modelling function in the first place.

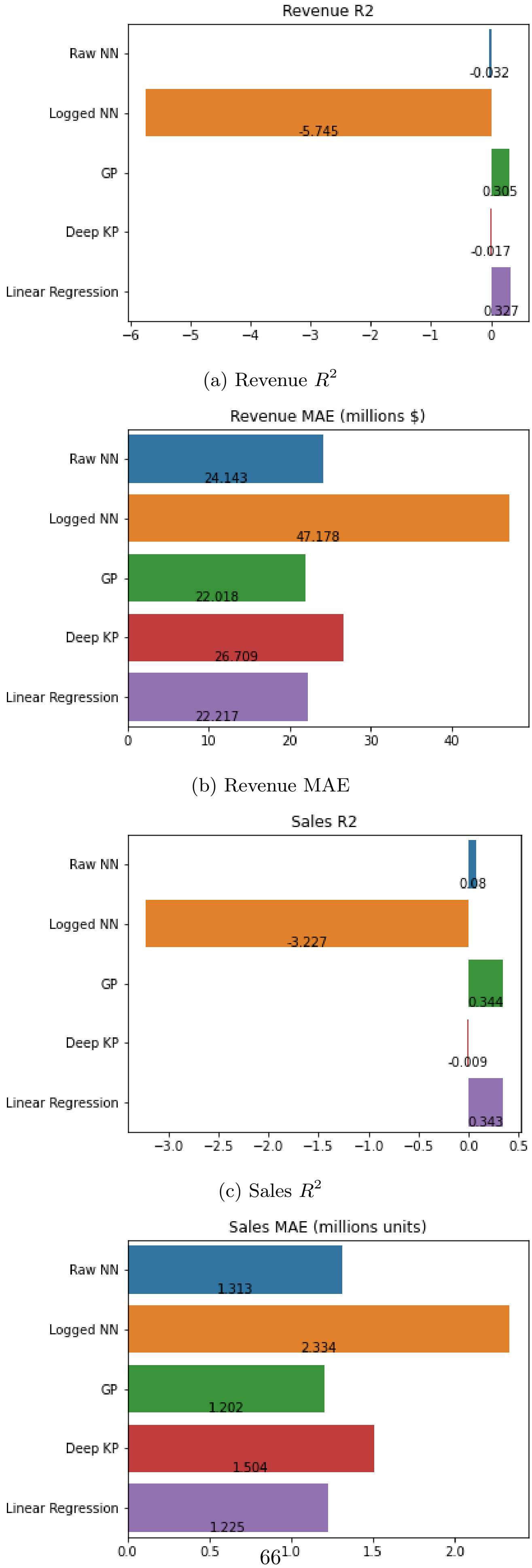
In Figure 4.24 we visualise the performance metrics for both sales and revenue and observe that the NN model trained on log transformed data performs poorly with a high MAE and a negative

Table 4.3: Model Comparison: Number of Weeks in Theatres Summary

Model	R^2	MAE (weeks)
Linear Regression	0.736	0.257
Neural Network (raw data)	0.721	0.268
Gaussian Process	0.748	0.252
Deep Kernel Process	0.524	0.347
Neural Network (log transformed data)	0.750	0.252

R^2. The cause of such a low R^2 is due to a few very large errors which can be seen in Figure 4.25. The Deep KP model too exhibits a small negative R^2. The remaining three models manage to achieve an R^2 of approximately 0.3 for both regressands. The R^2 of even these models suggests a lot of unexplained variance in prediction. The MAE for revenue prediction is unacceptably large with all models achieving an MAE of greater than \$20 million. We consider performance of this calibre to be incredibly weak and attribute this to limited availability of diverse data.

We inspect the violin plots of log-budget and log-gross for films that have had physical media release data in Figure 4.26 where we see that in general, films that are released in Blu-ray/DVD have both higher grosses and higher budgets compared to films that are not. We have also found evidence that the models are not able to identify logical relationships with input features in Figure 4.21, which further suggests that we have too few samples to enable the models to generalise well.



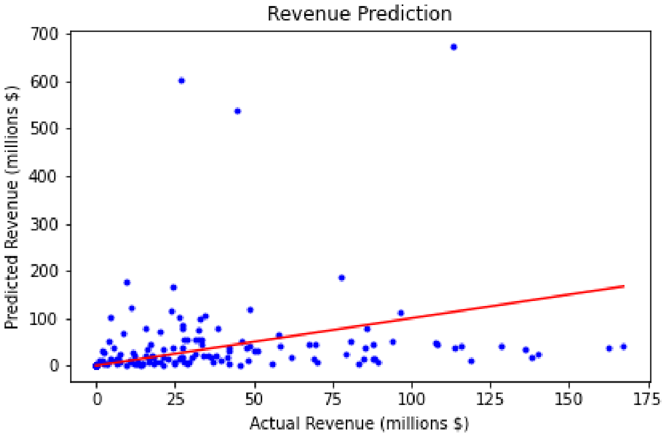
(a) Revenue R^2

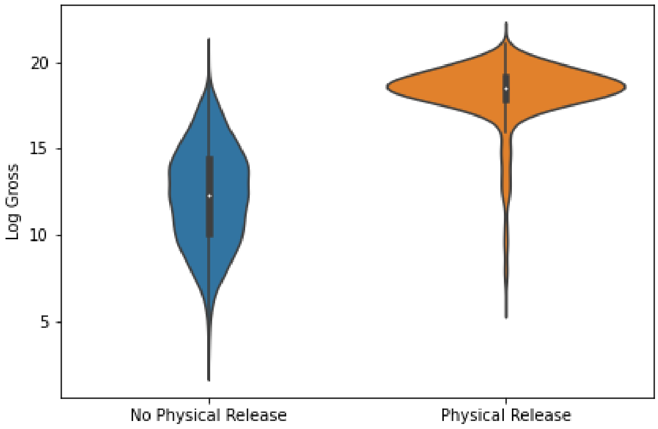
(b) Revenue MAE

(c) Sales R^2

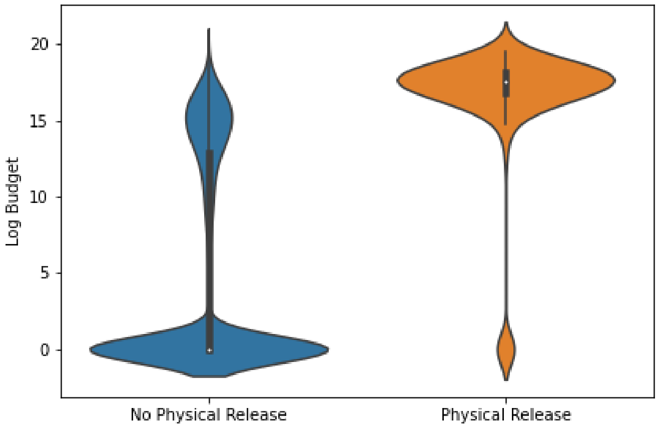
(d) Sales MAE

Figure 4.25: Physical Media Revenue. NN Model trained on log transformed data.





(a) Violin plot of log Gross for Films with and without Physical Media Release data



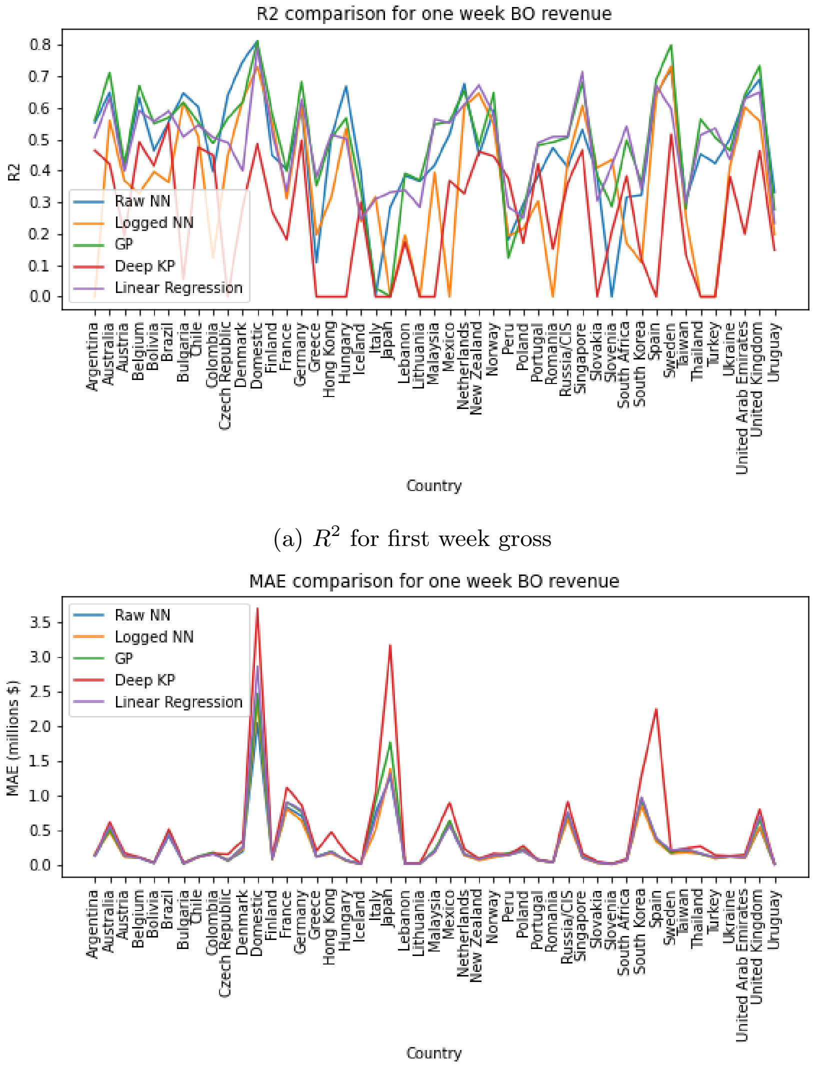
(b) Violin plot of log Budget for Films with and without Physical Media Release data

Figure 4.26: Comparing the log Gross and log Budget of Films with and without a Physical Media Release

Box Office One Week Gross

The opening weekend gross is defined as the revenue collected for the first week of release of a film in a country. The opening weekend often marks the highest revenue generating period of a film's release, making up a third of the box office total. In many cases, a good opening weekend can determine entirely whether or not the film is profitable at all.

Figure 4.27 shows the performance metrics per country for the first week gross. Both the Deep KP and the log transformed data trained NN occasionally achieve R^2 less than zero (which have been thresholded). Inspection of the MAE plot shows the Deep KP model producing large errors. The log NN model performs well, exhibiting a low MAE in line with the others.



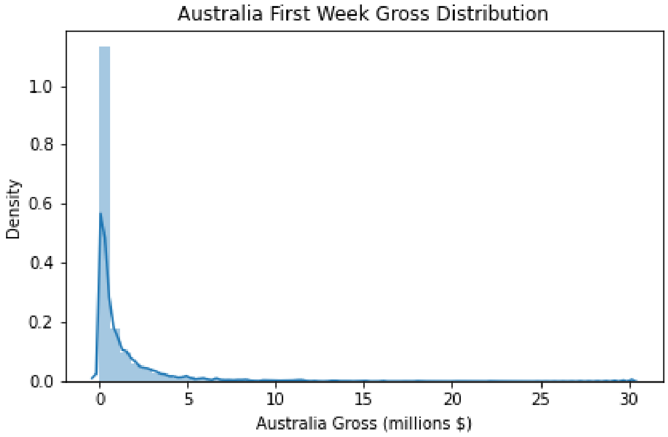
(a) R^2 for first week gross

(b) MAE for first week gross

Figure 4.27: Performance metrics per country for number of weeks in theatres

There are spikes in MAE for a handful of countries such as *Australia*, *Brazil* and the *Domestic* (*North American*) markets which is caused by a combination of these countries having few training samples and several instances of missing data for films released in these countries. For example, only around 2300 samples are available for *Japan* and 4000 for *Australia*. Looking at the distribution of grosses for *Australia* in Figure 4.28, there are fewer than 20 films that grossed more than \$15 million in the first week.

Figure 4.28: *Australia* First Week Gross Distribution



We summarise the model fit metrics of overall class fit in Table 4.4 where both NN models achieve a much lower MAE than other models while maintaining a similar R^2 to the GP and linear regression models.

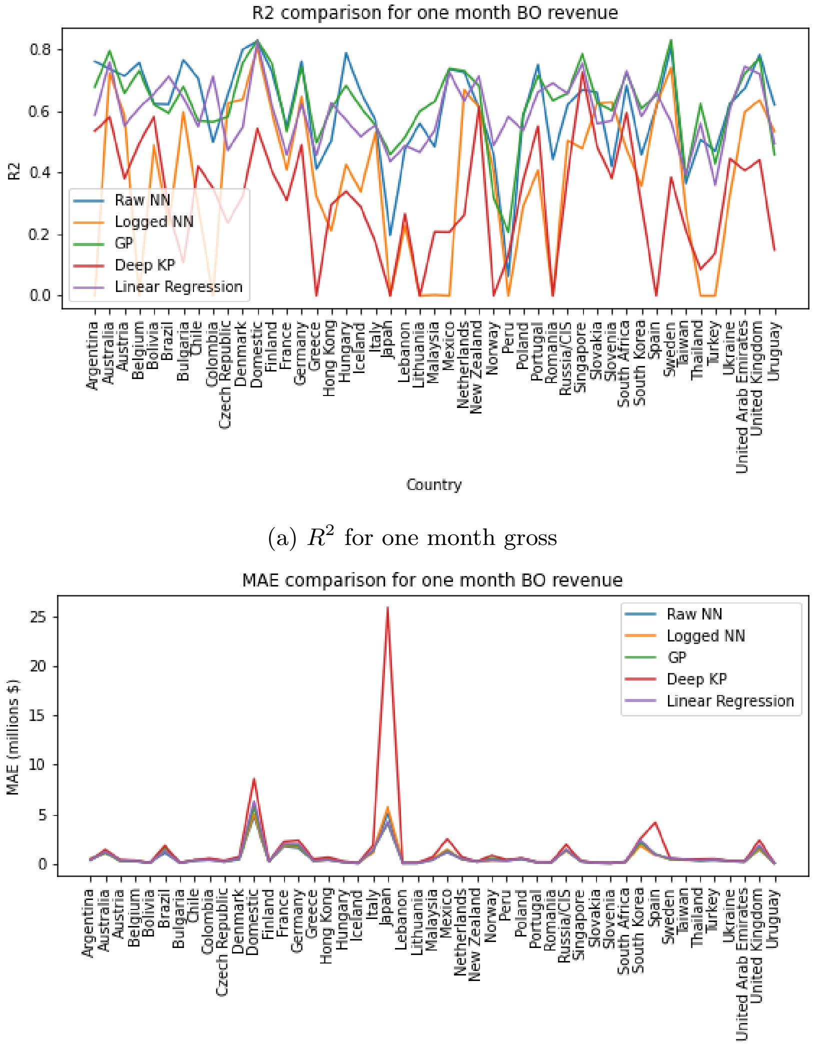
Table 4.4: Model Comparison: First Week Gross Summary

Model	R^2	MAE($)
Linear Regression	0.779	512 013.37
Neural Network (raw data)	0.742	433 846.10
Gaussian Process	0.794	498 252.44
Deep Kernel Process	0.613	749 123.88
Neural Network (log transformed data)	0.710	440 949.18

Box Office One and Three Month Grosses

One month and three month grosses are good indicators of film performance in the long run. We combine the evaluation of these classes as often there is overlap between these predictions, such as when a film is exhibited in theatres for less than four weeks. This most commonly occurs with limited release films.

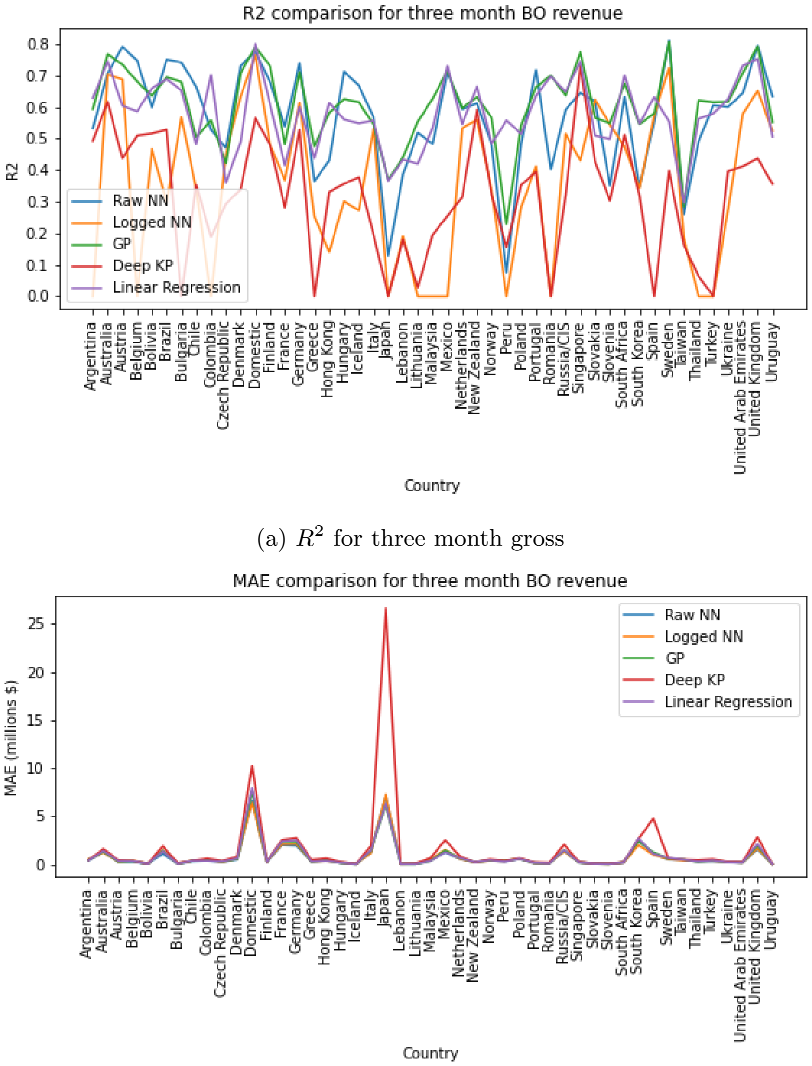
Graphs showing per country performance for both one and three month grosses are in Figures 4.29 and 4.30.



(a) R^2 for one month gross

(b) MAE for one month gross

Figure 4.29: Performance metrics per country gross after one month predictions



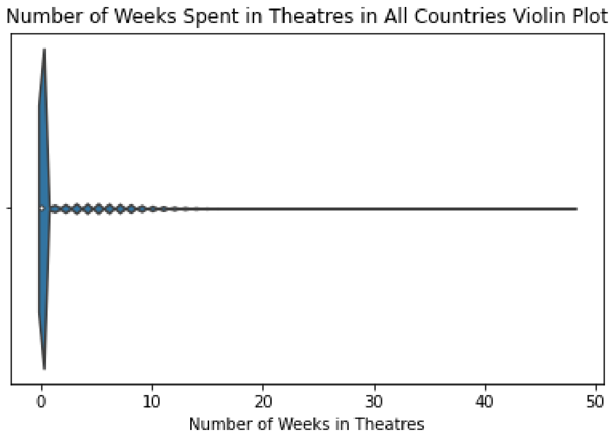
(a) R^2 for three month gross

(b) MAE for three month gross

Figure 4.30: Performance metrics per country gross after three month predictions

Models perform similarly for both classes of predictions. The Deep KP and log transformed NN models' R^2 occasionally fall below zero (thresholded for clarity) and spikes in MAE for both occur in the same countries. In fact the MAE's for both classes of predictions are so similar that it suggests that most films grosses at the four week and twelve week marks are similar and, by extension that most films are in theatres for no more than twelve weeks. We look at the distribution of the number of weeks films spend in theatres in all countries in Figure 4.31 which indicates that the vast majority of films are exhibited for no longer than four weeks.

Figure 4.31: Number of Weeks Spent in Theatres in All Countries Violin Plot



To quantify the similarity between samples of four and twelve grosses, we compute the Bhattacharya Distance (Equation 4.10) – an approximate measure of the overlap between two sets of samples – to be 0.4986. This suggests a reasonable degree of similarity between the two samples which agrees with our observations from the Violin plot of distribution.

$$D_B(p,q) = -\ln\ (BC(p,q)) \tag{4.10}$$

$$BC(p,q) = \sum_{x \in X} \sqrt{p(x)q(x)} \tag{4.11}$$

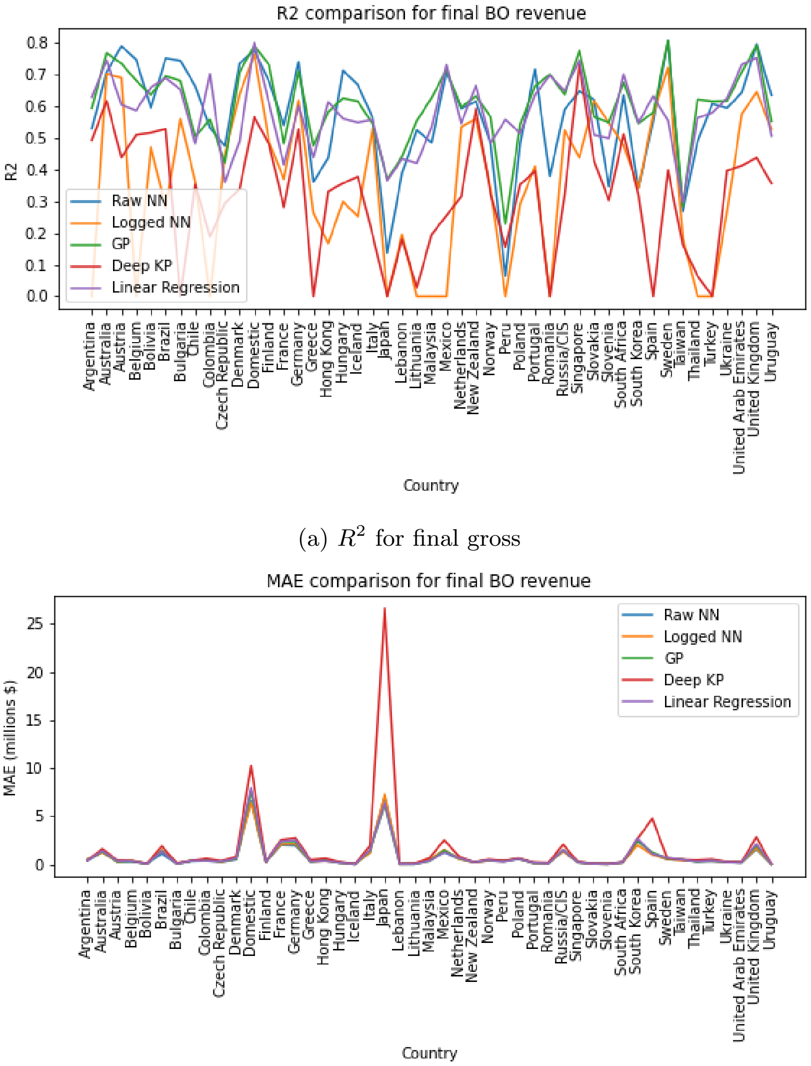
We summarise the class performance metrics in Table 4.5 which predominantly shows a marked drop in R^2 from the one month predictions, to the four month predictions, and a corresponding increase in MAE. Both NN models exhibit similar levels of performance with the GP and linear regression models trailing slightly.

Table 4.5: Model Comparison: One Month and Three Month Gross Summary

Model	One Month R^2	One Month MAE($)	Three Month R^2	Three Month MAE($)
Linear Regression	0.815	1 170 698.15	0.791	1 425 649.80
Neural Network (raw data)	0.816	997 000.01	0.771	1 253 102.59
Gaussian Process	0.823	1 111 591.45	0.788	1 383 767.96
Deep Kernel Process	0.406	1 999 625.09	0.473	2 239 654.66
Neural Network (log transformed data)	0.719	1 044 569.68	0.674	1 253 805.90

Box Office Final Gross

Finally, we evaluation performance for the final gross prediction per country. Final gross is the number that is often quoted as the number than qualifies success of a film. We look at the breakdown of performance per country in Figure 4.32 where there are few spikes in the MAE plot, such sharp increases in MAE being limited to *Domestic* (*North America*) and *Japan*. As in previous sections, any R^2 below zero are thresholded. We see the raw data NN model, the GP and linear regression models performing similarly across all countries, almost always with a high R^2.



(a) R^2 for final gross

(b) MAE for final gross

Figure 4.32: Performance metrics per country for final gross predictions

We summarise class performance in Table 4.6 where, as for the previous section, we see both NN models delivering similar performance followed by the GP and linear regression models. Perhaps most surprising is how similar the MAE's for both NN models are despite their significantly different R^2 metrics. Considering that both NN models exhibit similar performance across most classes of regressands it must be that the log transformed data NN model produces smaller errors across most samples, and occasional large errors whereas the raw data NN model produces more 'consistent' proportional errors. Such divergent and good overall performance leads us to suggest an Ensemble Modelling approach as potential future work in Section 7.3.1.

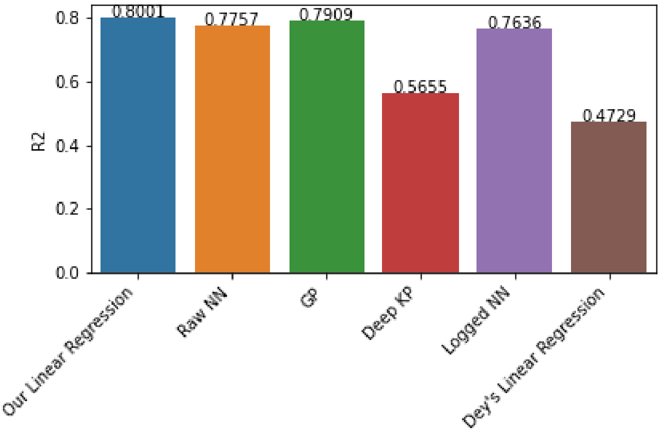
Table 4.6: Model Comparison: Final Gross Summary

Model	R^2	MAE($)
Linear Regression	0.791	1 425 649.80
Neural Network (raw data)	0.771	1 252 471.91
Gaussian Process	0.788	1 383 767.96
Deep Kernel Process	0.473	2 239 654.66
Neural Network (log transformed data)	0.675	1 254 589.83

We now move on to comparing our models against those developed by other researchers. There have been two approaches to modelling final box office: proposing it as a classification problem, often multi-class such as the Sharda-Delen model [28] (see Section 2.2.2) or as a regression problem [26, 48, 49, 50]. We consider three comparable regression models: Dey's [49] model, Pangarker et al.'s [50] model and Simonoff's [48] model.

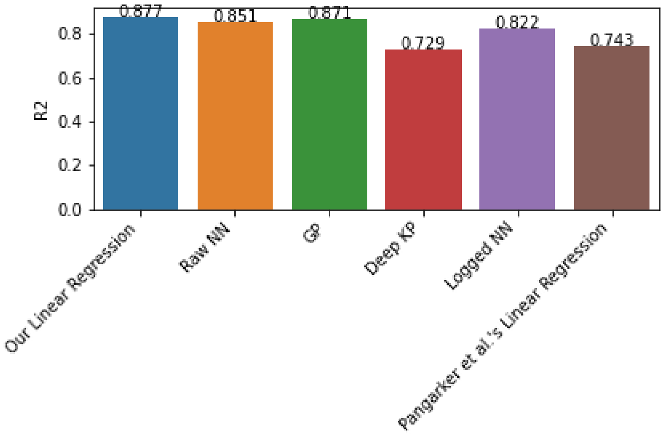
Dey presented a linear regression model to predict gross revenue from the USA box office using features available prior to release such as release date, MPAA rating and genre. Dey's chosen performance metric is the Coefficient of Determination and we directly compare the R^2 of North American gross regressand from our models with this in Figure 4.33 where both NNs as well as the GP and linear regression models achieve a much greater R^2. We attribute this to our models leveraging relationships between countries as well as the method of encoding used to include actors, directors and production companies, whose track record can indicate the potential performance of a film.

Figure 4.33: Developed Models compared to Dey's Regression Model



Pangarker et al. [50] developed a linear regression model to predict global box office revenue using a range of information, some of which is only known post-release such as the number of awards a film was nominated for and ratings from critics. These researchers use the Coefficient of Determination as their chosen performance metric. Our models do not perform inference for the total global gross of a film. To obtain a prediction of a total global gross, we sum the predicted grosses from individual countries and calculate the R^2 for summed predictions and summed true gross. Figure 4.34 shows our models' performance compared to Pangarker et al.'s model. We observe a slight improvement in R^2 for all of our models except the Deep KP. This once again implies that our models are able to capture and use dependencies between countries' markets and actors, directors and production companies effectively even in a pre-release setting.

Figure 4.34: Developed Models compared to Pangarker et al.'s Regression Model



Simonoff developed a linear regression model for predicting the final gross of a film in the USA using post-release information such as the first week gross and Rotten Tomatoes ratings. The model was trained only on wide release films with 125 films in forming the training set and a further 22 used for evaluation. Simonoff does not provide the R^2 or MAE metrics but he does provide the names of the films in the test set, as well as true final gross and the predictions made by his model. We use the reported predictions and true values to calculate the Coefficient of Determination and MAE for Simonoff's model.

Simonoff uses only films that received a wide release and the lowest grossing film in his test made over $7 million in revenue. We, however consider both wide and limited releases and the MAE between our models and Simonoff's wouldn't be comparable.

To ensure comparable results we retrain our models, holding out samples that correspond to the films Simonoff used for testing and compute performance metrics using these only. As a result the evaluation metrics in Table 4.7 for both our models and Simonoff's model are calculated using predictions made on the same films − using samples the models have not previously seen. We observe that that most of our models, with the exception of the Deep KP, exhibit very similar performance. In fact, both NN's and the GP have almost identical MAEs; despite using using only pre-release data these achieve a lower error than Simonoff's model. Our linear regression model too performs well with an R^2 in line with our more complex models and a competitive MAE.

Table 4.7: Model Comparison: Comparing our Models to Simonoff's

Model	R^2	MAE($)
Our Linear Regression	0.596	20 344 236.85
Neural Network (raw data)	0.604	**17 352 162.23**
Gaussian Process	0.669	17 782 225.43
Deep Kernel Process	0.082	27 193 087.37
Neural Network (log transformed data)	**0.692**	17 864 618.21
Simonoff's Linear Regression	0.740	24 991 222.27

4.4 Conclusion

In this chapter we began by asking ourselves: what makes a film profitable? To answer this question, we used a logistic regression model and powerful, tree based *XGBoost* model to analyse the characteristics of profitable films and identify how the magnitudes of features affect profitability.

We then considered five different models to predict pre-release box office performance of films: a linear regression model, two Neural Networks, a Gaussian Process and a Deep Kernel Process. We interpreted the inner workings of some of the black box models to identify relationships between regressands and input features. We have also shown the results of some point predictions from the models, as well as the predictive posteriors of models designed to provide model uncertainty. Although intended as a baseline performance level, the linear regression model was found to be a powerful model during evaluation, often competing well against the more complex models. In fact, we find very little performance difference between all the model, except for the Deep Kernel Process, with the Neural Network models exhibiting slightly lower mean absolute error on most regression tasks.

Chapter 5

Post-release Modelling: Ordinary Differential Equations

> Toto, I've a feeling we're not in Kansas anymore.
>
> *The Wizard of Oz, 1939*

5.1 Motivation

The first step in the process of solving the problem is to explore methods that show promise to this end. Looking at the graphs of how average revenue per theatre and the number of theatres playing the movie change over time (Figure 5.1 shows data for the film *The House at the End of the Street*), we note similarities with the models of an epidemic (Figure 5.2). We begin by taking an approach similar to that of the Edwards Buckmire [9] (EB) model.

This chapter constitutes an exploratory analysis of the application of ODEs to modelling post-release box office performance. We end the chapter (Section 5.5) by highlighting considerations and reasons for not pursuing this modelling approach.

Figure 5.1: Visualisation of theatres and daily revenue over time for *The House at the End of the Street*

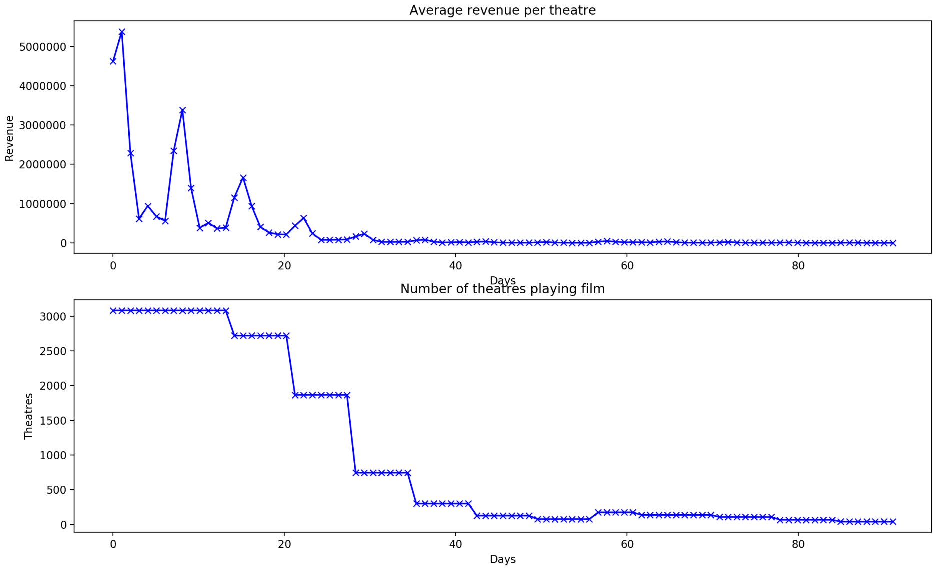
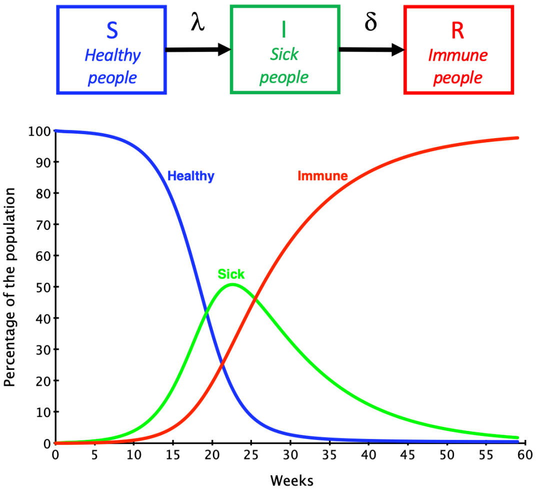


Figure 5.2: Visualisation of the SIR Model. Credit: TU Delft



5.2 Data Format

We begin by looking at daily reported data from The Numbers. The Numbers provides day to day, North American data for films, with records of daily total revenue, revenue per theatre and the number of theatres showing the film.

5.3 Developing the Model

5.3.1 Governing Equations

If the total gross of a film at time t is given by $G(t)$, we want to find $G(\infty)$. We consider modelling the number of theatres playing the film ($T(t)$), as well as the average revenue per theatre ($R(t)$). Thus we define,

$$\frac{dG}{dt} = RT \tag{5.1}$$

$$\frac{dR}{dt} = -\alpha_R R \tag{5.2}$$

$$\frac{dT}{dt} = -\alpha_T T \tag{5.3}$$

where $G(0) = 0$ and $R(0)$ and $T(0)$ are positive real numbers.

$G(\infty)$ can be computed by integrating total daily revenue over all time, therefore

$$G(\infty) = \int_0^\infty RT dt \tag{5.4}$$

5.3.2 Extending the Model

In the real world, movies are released in theatres in one of two ways. In a 'limited release' a movie is released on a small number of screens and positive word of mouth can prompt exhibitors to increase the number of screens it is played in. Case in point, the film *Mission Impossible: Ghost Protocol* actually saw maximum theatre showing a few days after release (Figure 5.3). In a 'wide release' a film is released in a large number of theatres initially which does not increase over time.

The original model is not able to handle this as it assumes all theatres show the film on day of release and the number of theatres decays monotonically from there. This prompts a modification of the model for $T(t)$.

If the number of screens is small but the daily average revenue is high then the number of theatres that play the film must increase up to a value T_{max}. If $T > T_{max}$ then the number of theatres should decrease monotonically. Hence, the equation modelling rate of change of T is modified to:

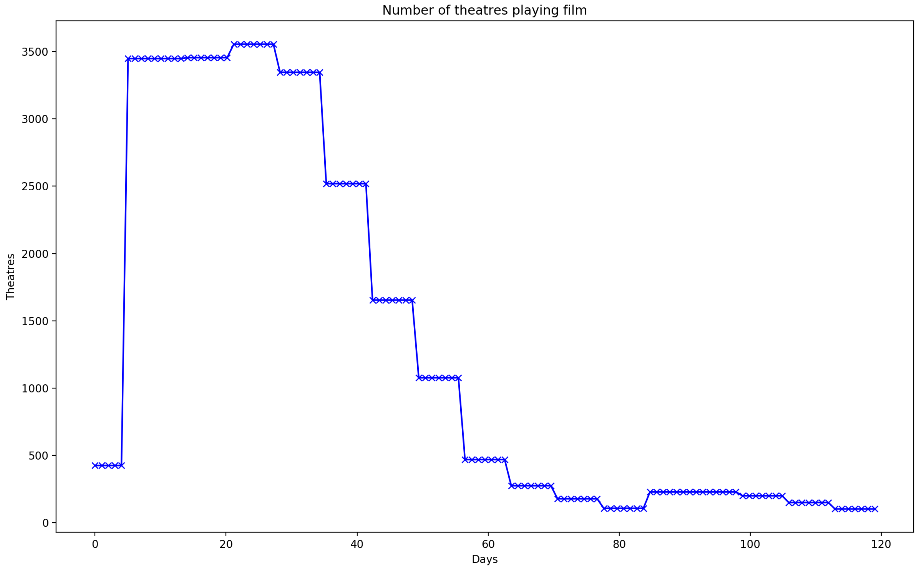
$$\frac{dT}{dt} = \alpha_{T1}\frac{RT}{T_{max}} - \alpha_{T2}T \tag{5.5}$$

In Equation 5.5 we assume that T depends not only on itself, but R as well. This is because for the exhibitor, the revenue from the film is the main factor in deciding whether to continue showing the film. The disadvantage of this is that it ignores the relative stability of screening afforded by the rental contract, where the number of theatres stays stable for several weeks at a a time.

5.3.3 Parameter Estimation

This model requires two types of parameters:

Figure 5.3: Theatres showing *Mission Impossible: Ghost Protocol*



1. Fixed parameters such as the opening day average revenue, opening day number of theatres and maximum number of theatres that show the film. These can be determined a priori using machine learning techniques taking values from an existing database of movies, potentially using clustering algorithms.

2. Free parameters such as the α_i's, the decay rates. For pre-release prediction these are global terms fitted to model a large data set of movies. For post-release, as data is collected the model can then be fitted to the new data for up to date predictions. Sawney and Eliashberg [6] implement just such a technique to a probabilistic model.

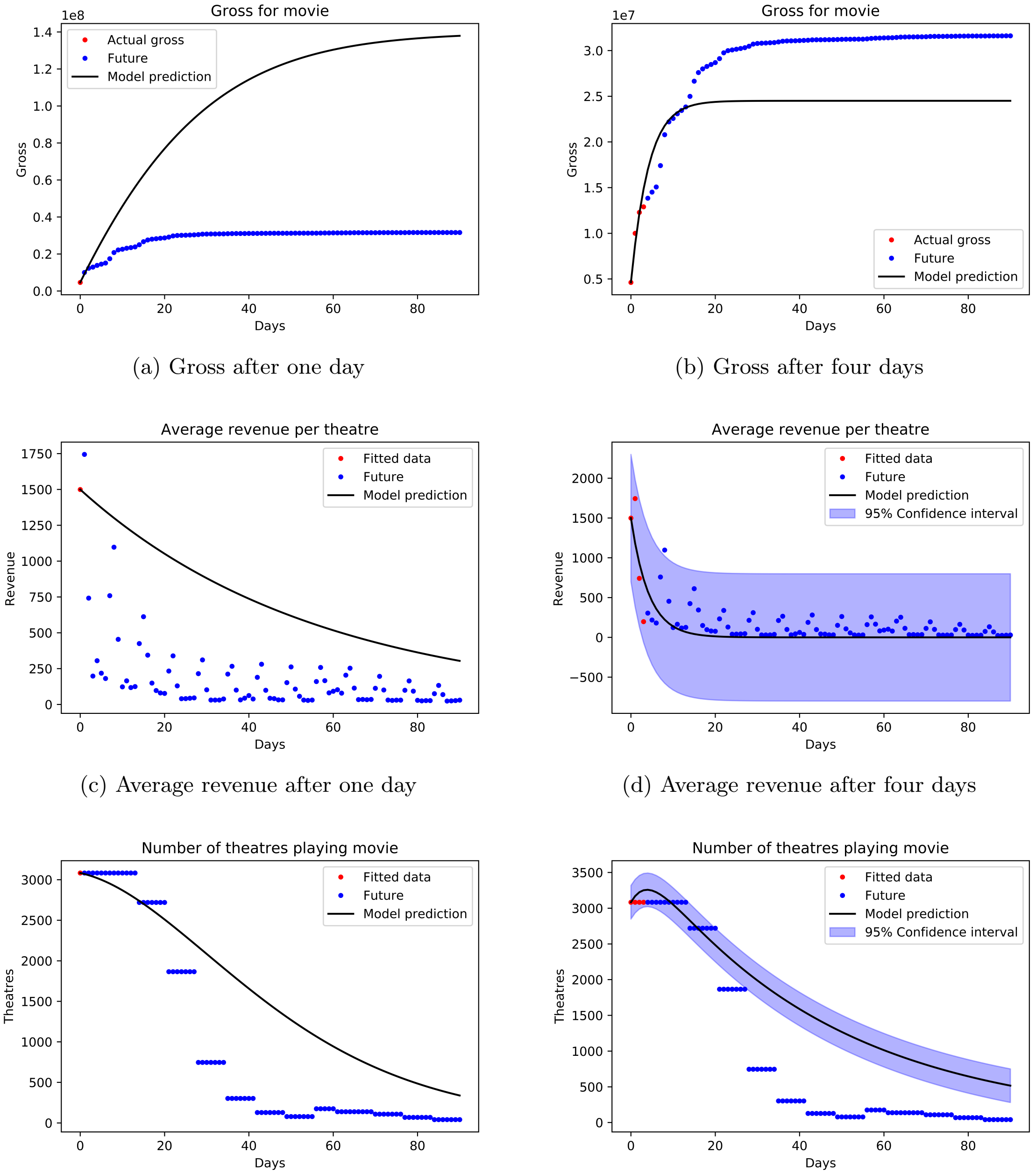
5.4 Modelling Post-release Box Office Performance

Modelling films, unlike modelling viral phenomena [29], have well defined start points – the date of release. With post-release prediction we are primarily concerned with fitting a model using the collected, real world data.

As data is collected on a daily basis, we require measures of the average revenue collected as well as the number of theatres the film is being shown in. The model can then be fitted to the data collected thus far.

As the equations governing the model can't be solved analytically, the *odeint* function in the *scipy.integrate* package can be used to solve the system of equations numerically. Using this requires providing suitable initial conditions such as a time sequence array for which an output is needed, as well as values for the decay parameters. After solving the ODEs, the model must be fitted to provided data for which we use the Nelder-Mead method. The residual function used here is the Least Squares Method. Since we are fitting both $T(t)$ and $R(t)$ in the model, the residual function returns the weighted average of errors for both $T(t)$ and $R(t)$.

Fitting on a daily basis yielded promising results. Graphs generated from post-release models are in Figure 5.4.



(a) Gross after one day

(b) Gross after four days

(c) Average revenue after one day

(d) Average revenue after four days

(e) Number of theatres after one day

(f) Number of theatres after four days

Figure 5.4: Results of model after fitting with one day and four days of data

All predictions were made for up to 90 days. The results of this show that by day four, the model prediction is reasonably accurate − and it continues refining the prediction as more data becomes available and is refitted. There are two significant drawbacks to this method. Firstly, in order to enable fitting, data values are interpolated from the last known data point up to the 90 day mark as linearly decaying points. This accounts for the poor prediction capability with just one day of data. Further work is needed to interpolate with quasi-exponential decay which will trace the curve of the true data much more accurately. The other disadvantage arises due to the discrepancy between average revenues on Monday–Thursday and Friday–Sunday. Generally, the Friday-Sunday revenues are much higher, especially in the first few weeks leading the model to periodically over-predict or under-predict depending on the latest day it is fitted to. This cyclical nature also leads to large confidence intervals, as can be seen in Figure 5.4d. An approach to stabilising the model against this is to model Friday–Sunday and Monday–Thursday revenues separately and interpolating the values in between. This could be achieved by finding peaks in the daily revenue and filtering out the peak with the values to either side (thus separating Friday–Sunday from Monday–Thursday) to obtain two arrays, then interpolating the intermediate values in each.

5.5 Feasibility of this Modelling Technique

Modelling using coupled ODEs thus far has relied on the availability of highly granular, daily revenue and theatre data. In practice we find that this data is only consistently available for films released in North America that have high budgets. This leads to there being very few samples which are also unrepresentative of the box office releases in the country as a whole. Releases in other regions, such the in the UK and India, lack the granularity offered by the North American Box Office and rather follow at best weekly or weekend frequency reporting.

Normally, this wouldn't be an issue alone as the ODE model can be adapted to fit weekly interval data. However, modelling on a weekly basis means that some films in some countries may only have been exhibited for a short period of time − many for less than two weeks.

We find that the lack of available data renders this modelling technique infeasible. Pursuing this approach means that we do not meet our project objectives and as a result, we use other modelling techniques instead.

Chapter 6

Post-release Modelling: Machine Learning

> Just when I thought I was out, they pull me back in.
>
> *The Godfather: Part III, 1990*

6.1 Motivation

Prior to a film's release we are limited to using only the core characteristics of a film − such as budget, genre, actors etc. − and more importantly, we have no information on how audiences will respond to a film. However, post-release, new data about the film comes flooding in. Reviews, ratings and more importantly, figures on revenue and screenings can provide a true indication of how an audience will react to a film which we use to predict the number of theatres that will screen the film in consecutive weeks as well as the future revenue.

This is a task that involves the modelling of several, short time series that are unrelated and non-stationary (see Section 3.4.7). Films can generally stay in theatres for anywhere from one week to an entire year (such as *E.T. the Extra-Terrestrial*).

6.2 Release Pattern Terminology

Films released are usually of two types: wide releases or limited releases each with their own characteristic curves as we have seen before in Section 3.4.7. To clarity and consistency in our modelling, we define three different types of release curve. Firstly, a *Wide* release pattern is so called as the revenue it generates per week and the number of theatres screening it, generally decline over time with the maximum for both occurring at release. Secondly, a *Limited, Then Wide* release pattern is one where the number of theatres and revenue reach their peak after release, following which both decline. Finally, a *Complex* release pattern follows neither of these trends and can in fact have several distinct peaks in both theatres and revenue post release. It is important to note that these definitions are descriptors of the shape release curves and do not relate to the actual values of revenues or theatres that occur.

6.3 Target Description

In post-release modelling we are concerned with predicting how much revenue will a film generate in the next week, and how many theatres will still be screening the film in a week's time.

The models will take in k previous periods' features (see Section 3.5.4) containing the k previous weeks' theatres and revenues as well as constant features such as encoded actors, encoded directors, genres etc.

6.4 Dataset Division

In pre-release modelling each sample corresponds to a film where prices are inflation adjusted and information is not time-varying. With post-release modelling, however, a film's release can generate multiple samples; one for each week of release and with information varying with time. We use an 90-10 split with 90% of data used for training and 10% for testing. We take an Out of Sample approach to testing the performance of these models primarily due to the computational complexity of performing k-fold cross validation which is especially high for the LSTM based models.

6.5 Overview of the Models

This is a distinct time-series prediction problem where we model several different, short time-series and expect the model to learn trends and quickly which release pattern a film will follow: a *Wide Release*, a *Limited, Then Wide Release* or a *Complex Release.*

We propose two approaches to modelling: firstly, using non time-series machine learning models which are provided with only the previous week information; and secondly, more complex models that make use of Long Short-Term Memory and are provided with more previous weeks of information. We use two LSTM models each with different numbers of LSTM units and being designed to use different sequence lengths of data. The LSTM models are also designed to be able to estimate uncertainty (see Section 2.2.2) using MC dropout.

6.5.1 Ridge Regression

We first consider a linear model: Ridge Regression [20] (RR). With RR, we minimise the squared error between predictions and observations, and in addition, penalise the squared magnitude of weights. As a result the optimal coefficients are given by:

$$\beta_{Ridge} = \arg\min_{\beta} ||y - X\beta||_2^2 + \lambda||\beta||_2^2 \tag{6.1}$$

where is a tuning parameter that determines the significance of the penalisation term. We use Ridge Regression to ensure that no single wight is over-fitted upon. The tuning parameter, λ, is set to 1 for this model.

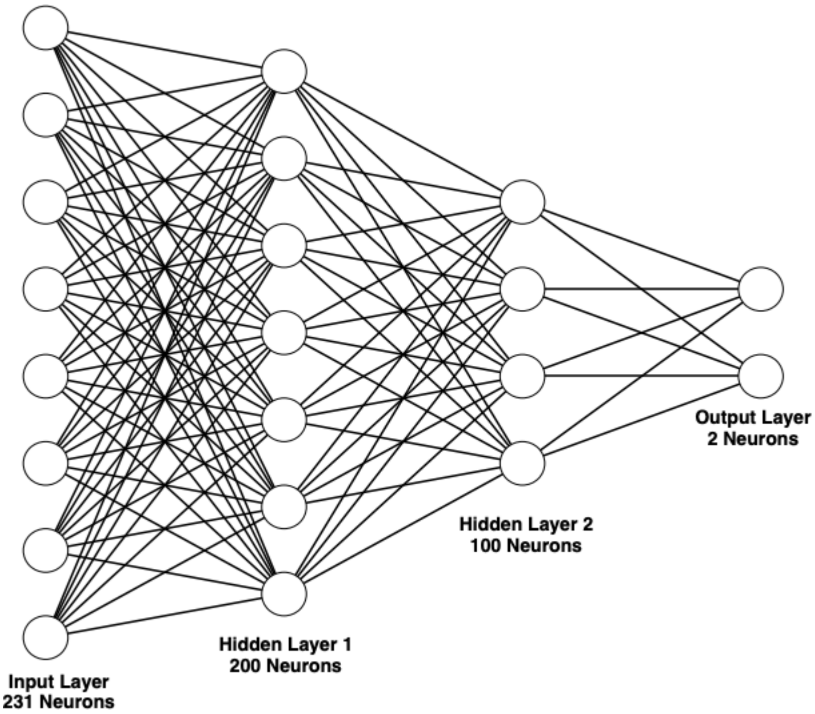
6.5.2 Neural Network

The second model is a Neural Network. NN's can model non-linear relationships between variables and can model far more complex functions than is possible with a linear model like RR. This comes with an associated cost of increased training time, an inability to find exact optimal parameters exactly, a much larger number number of parameters to optimise and potential non-convergence during training. We designed a NN with three hidden layers and a *Softplus* activation function applied to outputs of all layers. The *Softplus* activation function is defined as:

$$\sigma(z) = \ln\ (1 + e^z) \tag{6.2}$$

A description of the NN architecture for this model is in Figure 6.1. We set dropout rate to 0.1 and the model is trained with a batch size of 4096 samples and a learning rate of 0.001.

Figure 6.1: Post-release Neural Network Architecture



6.5.3 LSTM Neural Network 1

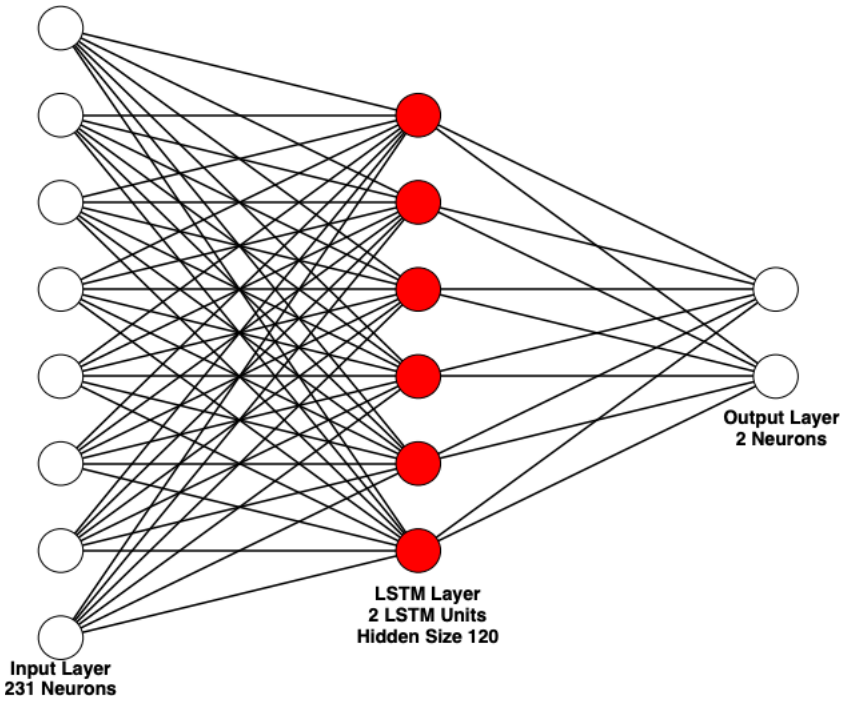
We now design complex models more suited for time-series prediction. The first of these models is a Neural Network which employs the architecture described in Figure 6.2. A *Tanh* activation function is applied to the output of final LSTM layer and the output of this layer is fed to a linear output layer. The *Tanh* function is defined as:

$$\sigma(z) = \frac{2}{1 + e^{-2x}} - 1 \tag{6.3}$$

We provide as input into this model only the previous two weeks of data, with zero padding when a full two weeks of data is not available. To reduce overfitting and enable prediction of uncertainty we set dropout to 0.2. We use a batch size of 512 and a learning rate of 0.001.

We refer to this model as the *LSTM 1* model in future chapters.

Figure 6.2: Post-release LSTM 1 Model Architecture

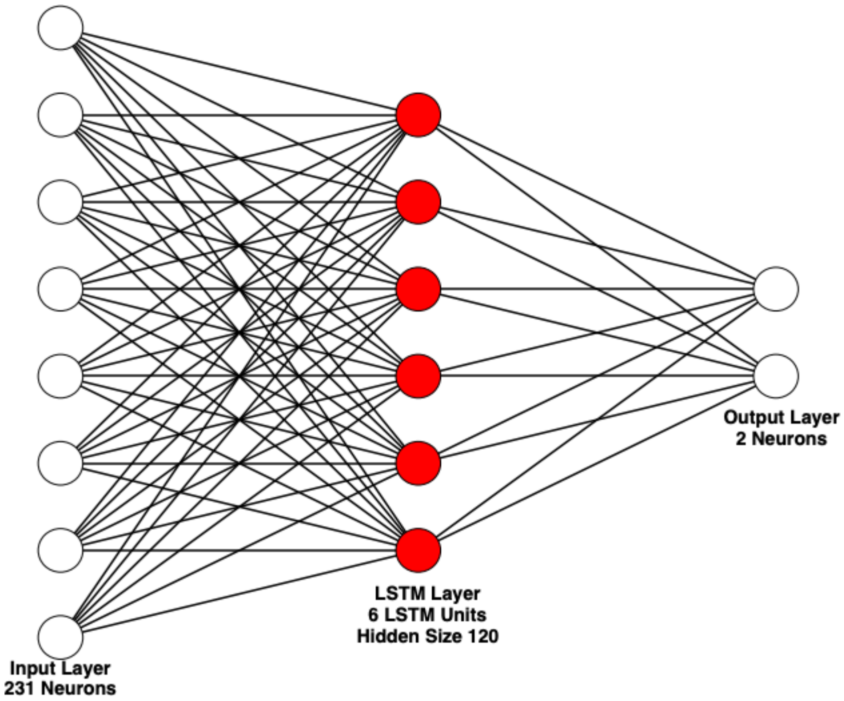


6.5.4 LSTM Neural Network 2

The second LSTM based model is once again based on a two hidden layer NN with the architecture descibed in Figure 6.3 and with a *Tanh* activation applied. This model has the ability to remember longer sequences, hence we also provide as input more past data points - this time data for the past six weeks is used and if necessary, zero padded. As with the previous LSTM model, dropout is set to 0.2 and a batch size of 512 is used with a learning rate of 0.001.

This model will be referred to in subsequent sections as the *LSTM 2* model.

Figure 6.3: Post-release LSTM 2 Model Architecture



6.6 Ridge Regressor Model Interpretation

The primary advantage of using the Ridge regression model is that, as a linear model, its weights are easily interpreted we gain a clear insight into how the model works. We look at the top twenty largest weights in magnitude in Figure 6.4 where we see that the most significant features are the *previous_ week_ theatres* and *previous_ week_ gross*. Surprisingly, the previous week's theatres is more highly weighted than the gross which seems counterintuitive as the number of theatres often remains steady for long periods. However, as we find out in Section 6.7.1, the number of theatres is a very good indicator of future revenue. The Ridge regression model doesn't assign much importance to the other features, with some of the more important ones being whether the film is released in the English language (*language_ en*) as well as the IMDb rating. However, the low importance assigned to the IMDb rating is unexpected − we would think that the rating would prove more informative about performance − so we plot IMDb rating against gross in Figure 6.5 where there are many films with an IMDb rating between 4 and 8, and the gross of films in this range are higher than for other scores. As a result, the IMDb rating is more likely to serve as a filter that likely indicates low performance i.e. a rating less than of less than 4 for greater than 8 could signal a film that generates lower revenue.

Figure 6.4: Ridge Regressor Feature Weights

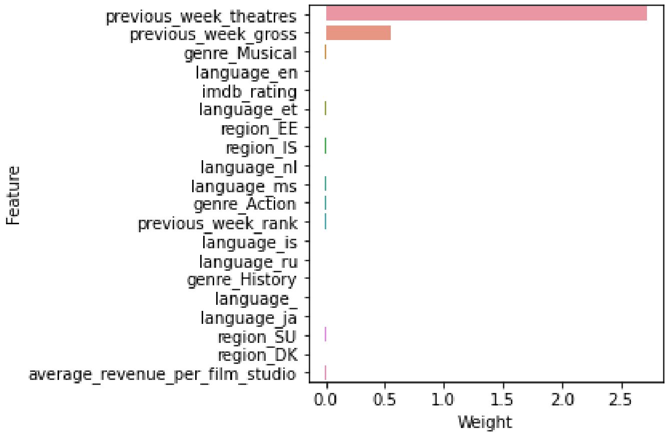
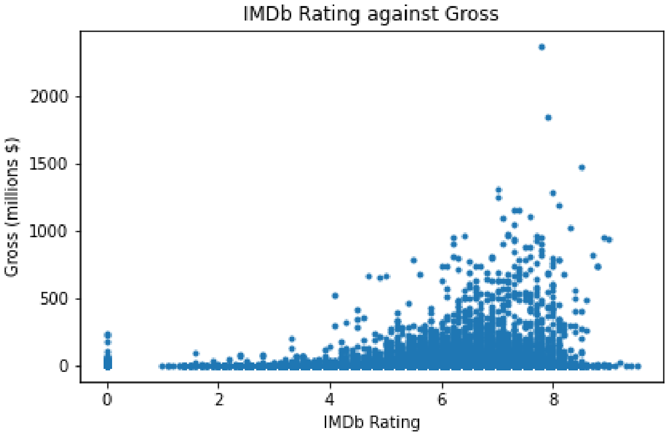


Figure 6.5: IMDb Rating Plotted Against Gross



6.7 Forecasting

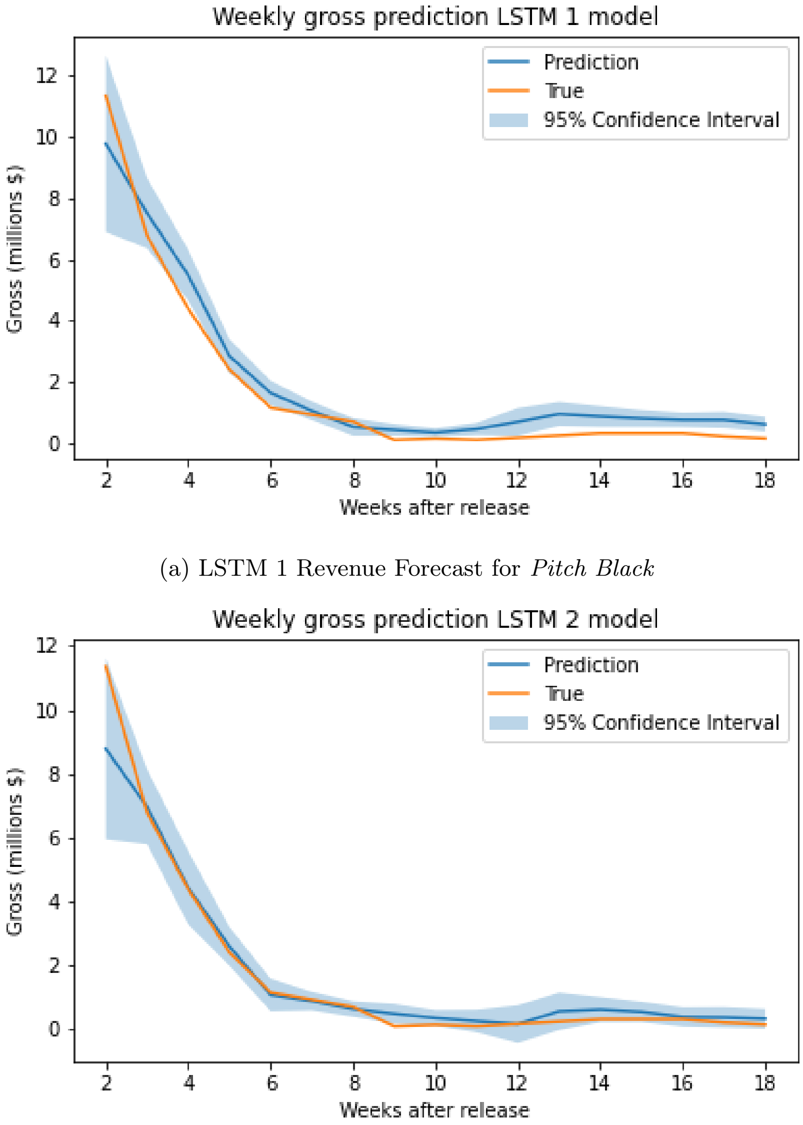
To demonstrate the application of the LSTM 1 and LSTM 2 models to forecasting, we explore how well they apply to both long term and short term prediction. We perform two types of forecasting: firstly showing how the models can be applied to predict for one week ahead at a time; and secondly, how well the models can predict for several weeks ahead.

6.7.1 One Week Ahead Predictions

A Typical *Wide* Release Pattern

Performing one-week ahead predictions provides the most accurate predictions possible as the LSTM models can use real, collected data to base forecasts on. Figures 6.6 and 6.7 show the one week ahead revenue and theatre for each model for the film *Pitch Black*.

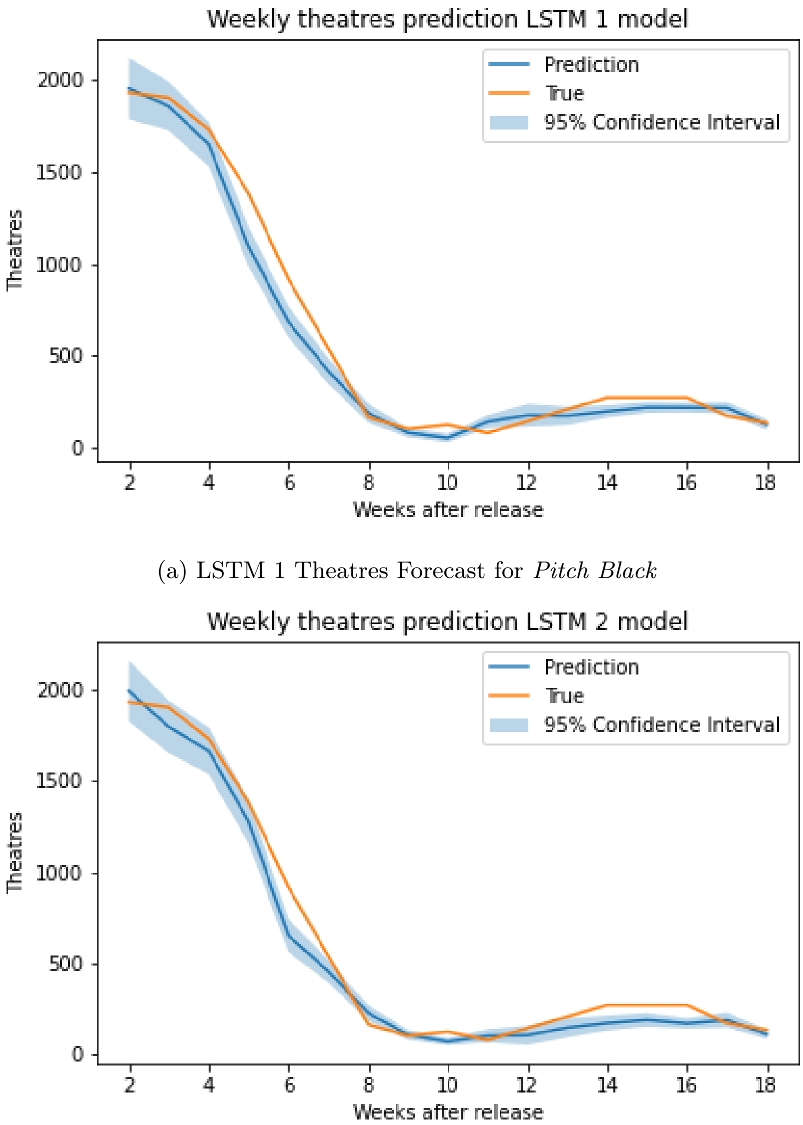
Pitch Black was a film released in 2000 in North America only and follows the revenue pattern of a *wide release* film. We observe good quality forecasting from both models, with the LSTM 1 model under-predicting revenue after the 10 week mark.



(a) LSTM 1 Revenue Forecast for *Pitch Black*

(b) LSTM 2 Revenue Forecast for *Pitch Black*

Figure 6.6: North America Revenue Forecast for *Pitch Black*



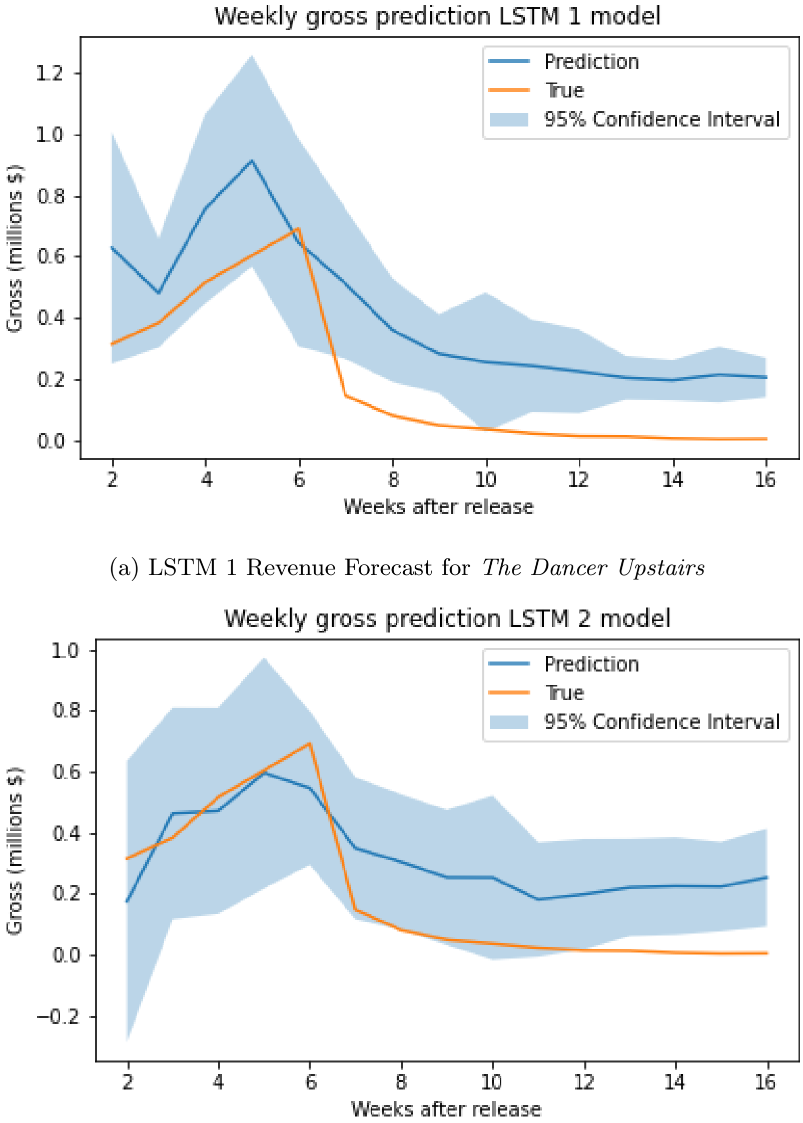
(a) LSTM 1 Theatres Forecast for *Pitch Black*

(b) LSTM 2 Theatres Forecast for *Pitch Black*

Figure 6.7: North America Theatres Forecast for *Pitch Black*

A *Limited, Then Wide* Release Pattern

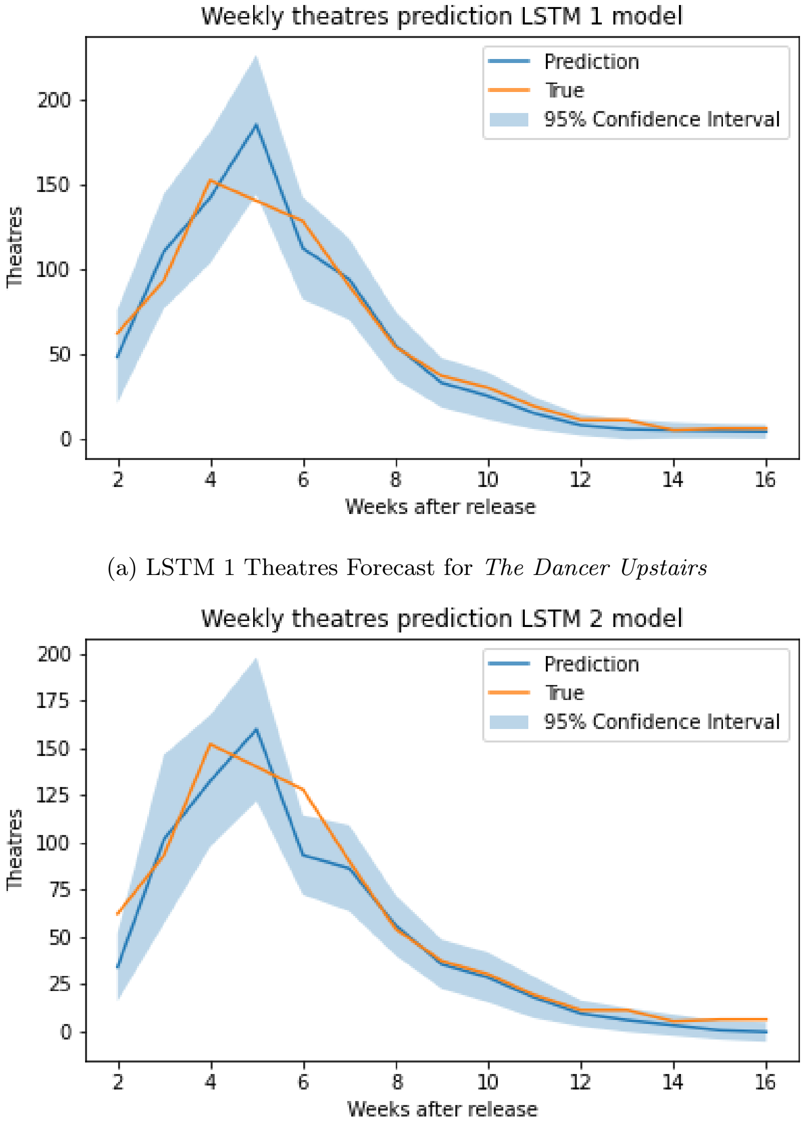
We've seen an example of how the model is able to forecast for a typical wide release pattern – but what about when a film releases in a limited capacity and becomes more widely exhibited post-release? *The Dancer Upstairs* had such a release in North America. Plotting the one week ahead forecast for revenue and gross yields the results in Figures 6.8 and 6.9. Both models are able to model the revenue well, especially as revenue has a more gradual build up to a peak whereas for the theatres plot, both models predict the peak to be shifted in time by one week. The limited ability of the models to predict peaks well is occurs as we consider only a few post-release features: better results could likely be obtained by considering more sources of post-release information that better capture popular appeal such as social media data.



(a) LSTM 1 Revenue Forecast for *The Dancer Upstairs*

(b) LSTM 2 Revenue Forecast for *The Dancer Upstairs*

Figure 6.8: North America Revenue Forecast for *The Dancer Upstairs*



(a) LSTM 1 Theatres Forecast for *The Dancer Upstairs*

(b) LSTM 2 Theatres Forecast for *The Dancer Upstairs*

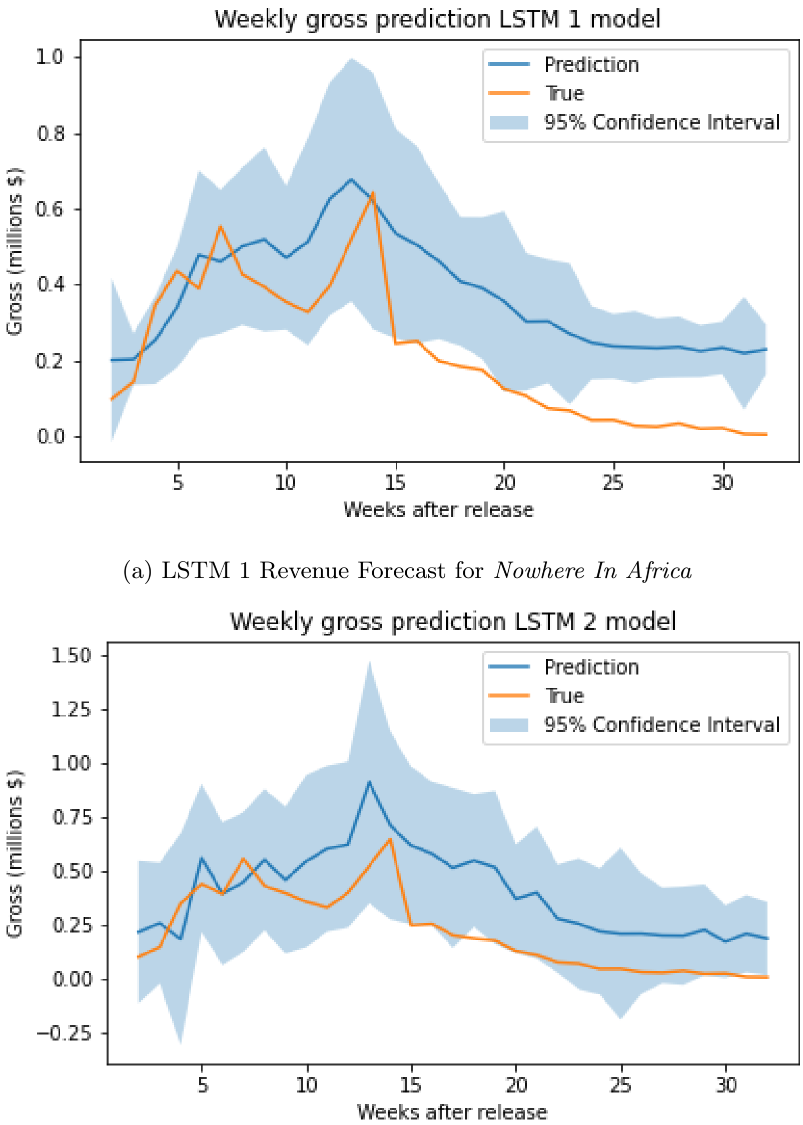
Figure 6.9: North America Theatre Forecast for *The Dancer Upstairs*

A *Complex* Release Pattern

We now look at forecasting for films that cannot be described as neither a *Wide Release* nor a *Limited, Then Wide Release*. *Nowhere In Africa* is a film with especially convoluted revenue and theatre streams. Modelling its' weekly revenue and theatres in North America generates the results in Figures 6.10 and 6.11.

Both models find it difficult to fit the nature of this release as there is no indication of when and why ups and downs in theatres and revenue occur. After week 15, both models consistently over-predict the revenue, however, the theatres predictions are far closer to the true values.

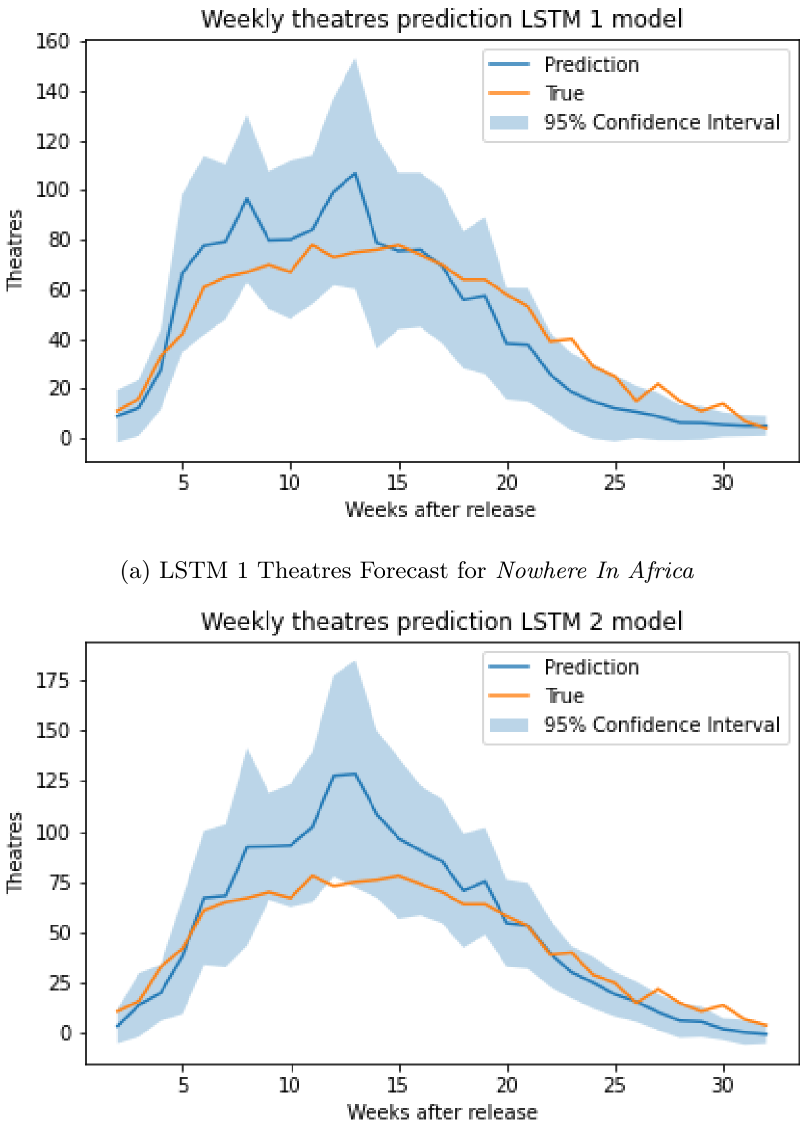
We inspect the cross-correlation for the true revenue and true theatres sequences in Figure 6.12 where we see revenue and theatres being correlated at a slightly negative time lag − in other words, a change in theatres precedes a corresponding change in revenue. This is the likely explanation for why the models are able to predict the peak for revenue so well; the theatres indicate this change beforehand. This is also why we see the peak theatre prediction for a *Limited, Then Wide* release to be delayed by a week whereas the revenue is modelled well.



(a) LSTM 1 Revenue Forecast for *Nowhere In Africa*

(b) LSTM 2 Revenue Forecast for *Nowhere In Africa*

Figure 6.10: North America Revenue Forecast for *Nowhere In Africa*

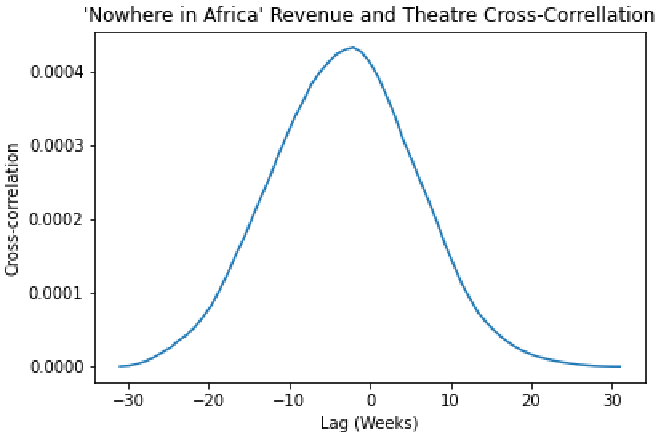


(a) LSTM 1 Theatres Forecast for *Nowhere In Africa*

(b) LSTM 2 Theatres Forecast for *Nowhere In Africa*

Figure 6.11: North America Theatres Forecast for *Nowhere In Africa*

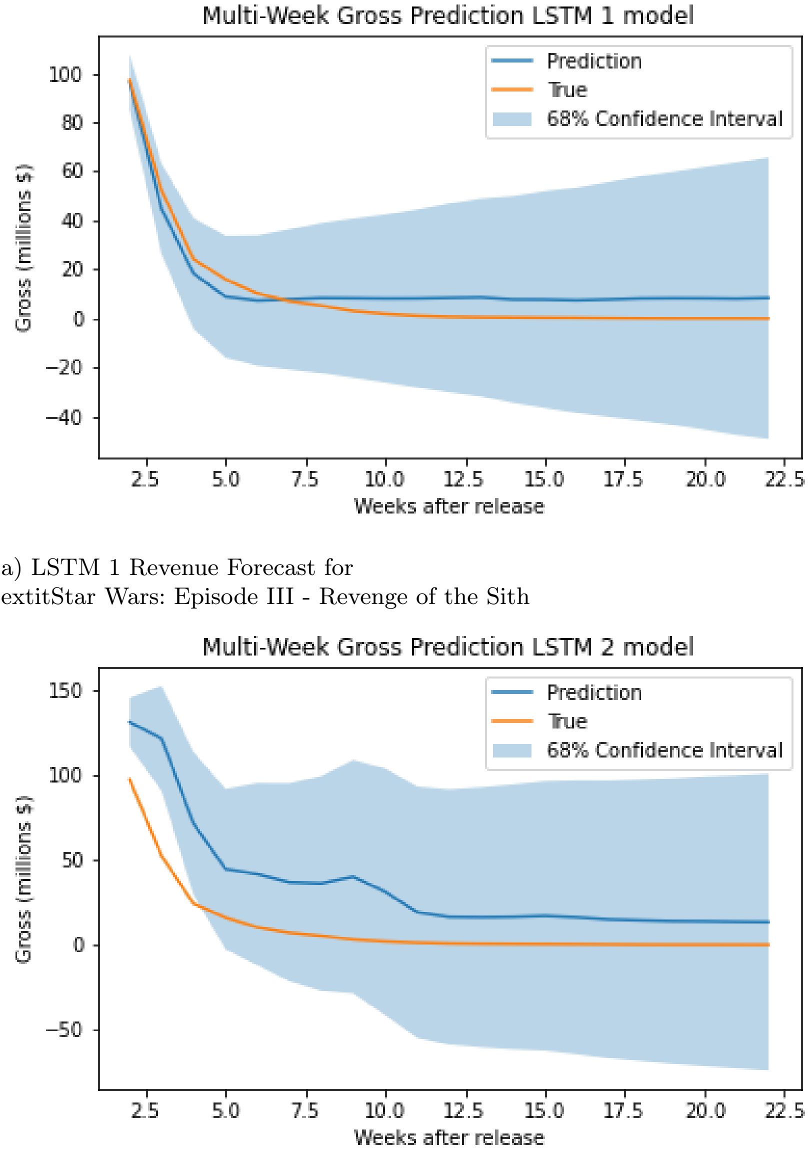
Figure 6.12: *Nowhere In Africa* Revenue and Theatres Cross-Correlation



6.7.2 Multi-Week Ahead Predictions

We next look at how the models can perform multi-week ahead forecasting: a typical use case would be when the first week gross and theatres are known and we wish to predict how these will change over the next several weeks. Hence, any predictions made by a model are fed back into it to be used as the basis for future predictions.

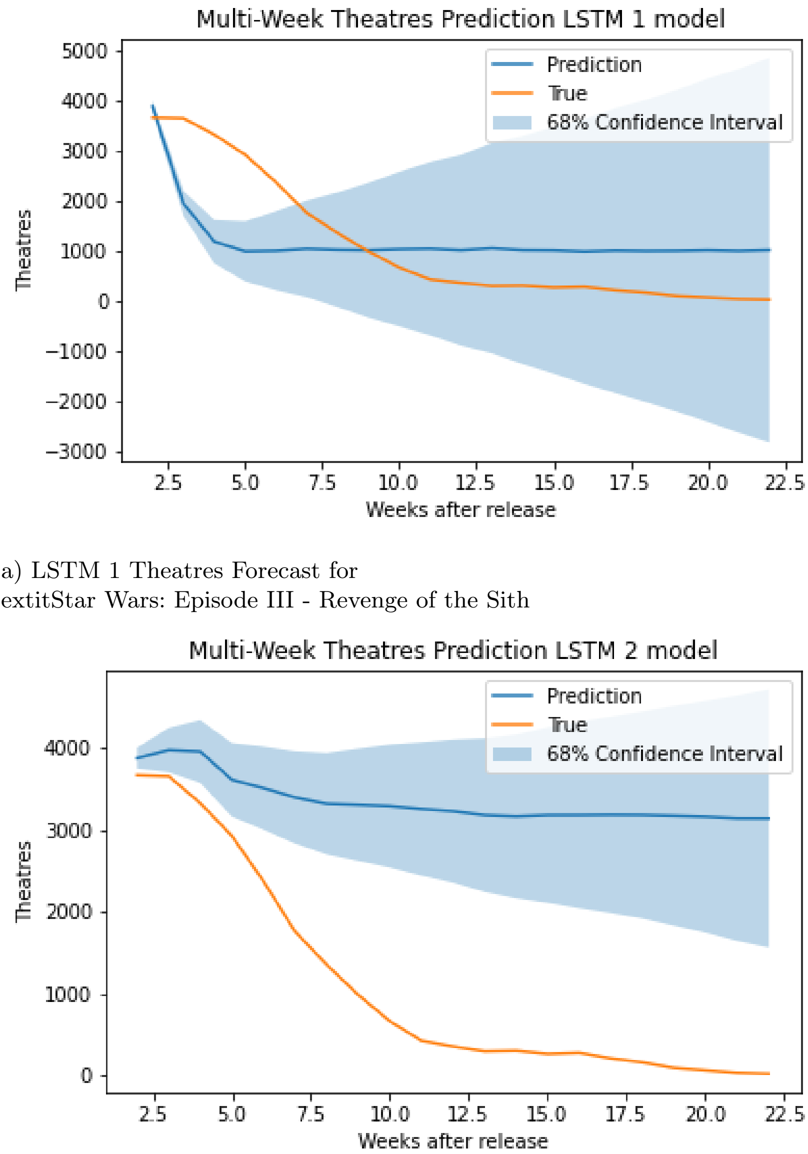
We perform a complete forecast for *Star Wars: Episode III - Revenge of the Sith* beginning from the true opening week revenue and using predictions as future inputs. The results of this forecast are shown in figures 6.13 and 6.14.



(a) LSTM 1 Revenue Forecast for
textitStar Wars: Episode III - Revenge of the Sith

(b) LSTM 2 Revenue Forecast for *Star Wars: Episode III - Revenge of the Sith*

Figure 6.13: North America Revenue Forecast for *Star Wars: Episode III - Revenge of the Sith*

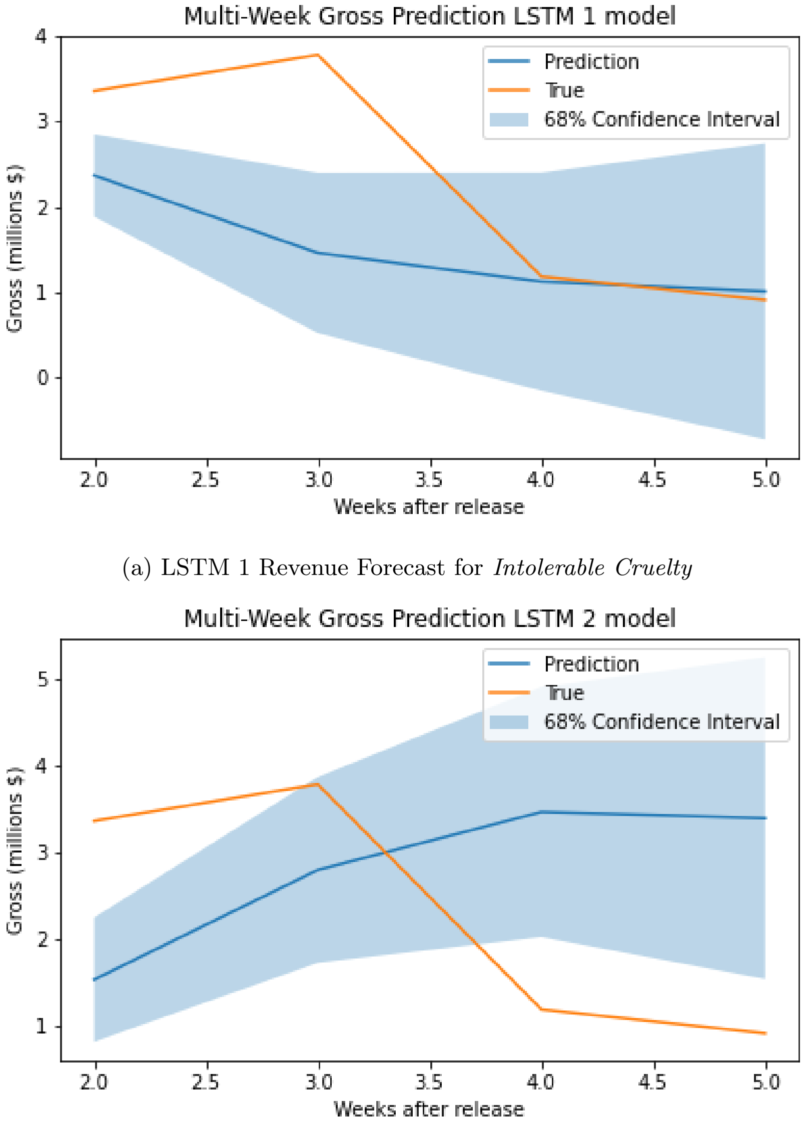


(a) LSTM 1 Theatres Forecast for
textitStar Wars: Episode III - Revenge of the Sith

(b) LSTM 2 Theatres Forecast for *Star Wars: Episode III - Revenge of the Sith*

Figure 6.14: North America Theatres Forecast for *Star Wars: Episode III - Revenge of the Sith*

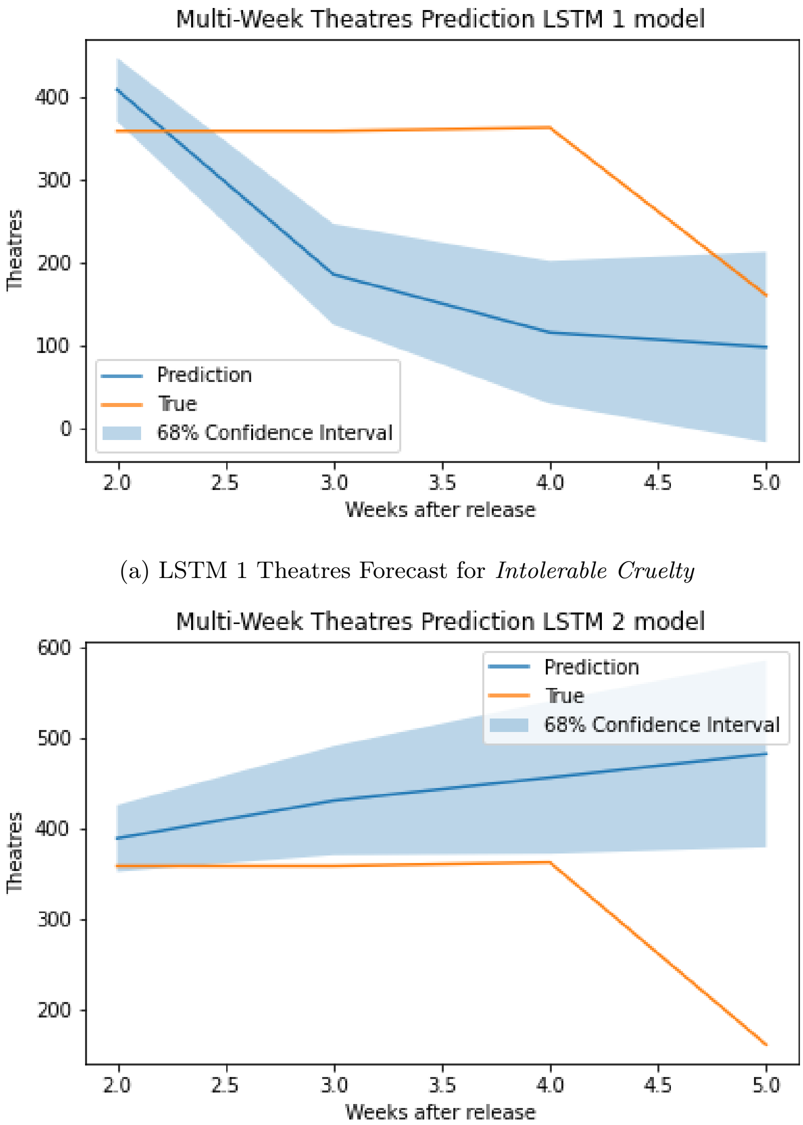
This is an especially good quality forecast with the models closely forecasting over twenty weeks of revenue and theatres with just the first week's information. However, this seems to be the exception rather than the rule; performing a multi-week forecast for the film *Intolerable Cruelty*'s release in Germany (Figures 6.16 and **??**) we find far less encouraging results. Curiously, the LSTM 1 and LSTM 2 model forecast very differing release types with LSTM 1 indicating a decrease in both revenue and theatres suggesting a *Wide* release pattern whereas LSTM 2 predicts an increase in both theatres and revenue akin to a *Limited, Then Wide* release.



(a) LSTM 1 Revenue Forecast for *Intolerable Cruelty*

(b) LSTM 2 Revenue Forecast for *Intolerable Cruelty*

Figure 6.15: Germany Revenue Forecast for *Intolerable Cruelty*



(a) LSTM 1 Theatres Forecast for *Intolerable Cruelty*

(b) LSTM 2 Theatres Forecast for *Intolerable Cruelty*

Figure 6.16: Germany Revenue Forecast for *Intolerable Cruelty*

6.8 Evaluation

6.8.1 Methodology

During pre-release regression in Section (4.3.7), the Coefficient of Determination and the Mean Absolute Error were used as evaluation metrics. The Coefficient of Determination of a measure of the variance that can be explained by the model. This is not a good measure of performance for a time-series prediction model as it does not take into account the type of data being modelled - time-series data tend to be correlated in time and consequently a relationship exists between a sequence and a time-shifted version of itself. A common error of many time-series forecasting models is that they predict values with a small time lag. If the R^2 was calculated for this data it would likely end up indicating well explained variance when in truth this does not reflect the performance of the model.

When evaluation these model we focus on quantifying the size and nature of the error: MAE is a metric that calculates the mean magnitude of error over all predictions, and is an intuitive and interpretable metric.

When time-series models make forecasts, the forecast errors can be positive or negative. The Mean Forecast Error (MFE) (Equation 6.4) is an error metric that aims to quantify both the magnitude and sign of the error, which indicates whether the model tends to over-predict or under-predict.

$$MFE = \sum_{i}^{n} \frac{y_i - f(x_i)}{n} \tag{6.4}$$

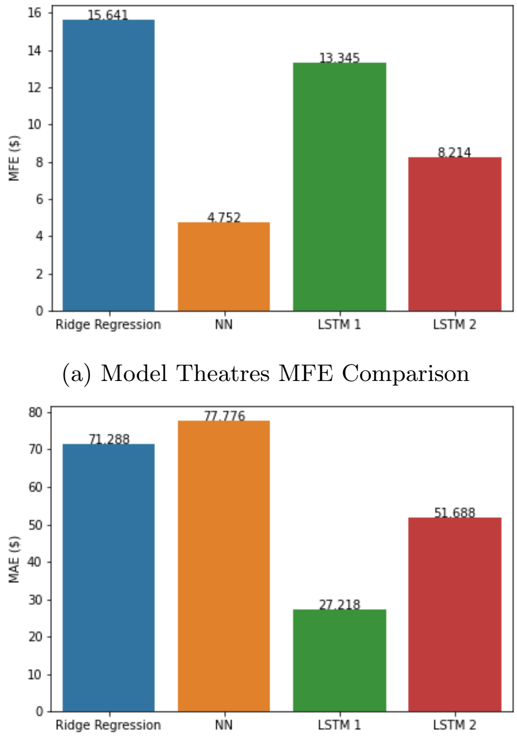
A negative MFE indicates that the model tends to over-predict and any small number close to zero indicates a over/under prediction small in magnitude.

6.8.2 Comparing Model Performance

We now compare the performance of our specified models, having trained them on the same training set and obtained predictions on the same test set. Each model makes two predictions: one prediction about the number of theatres showing the film in the next week and another prediction on the revenue expected in the following week. We evaluate performance on each regressand separately.

Theatre Forecast

Knowledge of the following week's theatres playing a film in every country enables production companies to gauge the long term financial and public impact of a film and optimally schedule release of the film in non-theatrical media. If a film is projected to have a better theatrical run than initially anticipated then better terms can be negotiated for both continued theatrical exhibition as well as for the non-theatrical release.



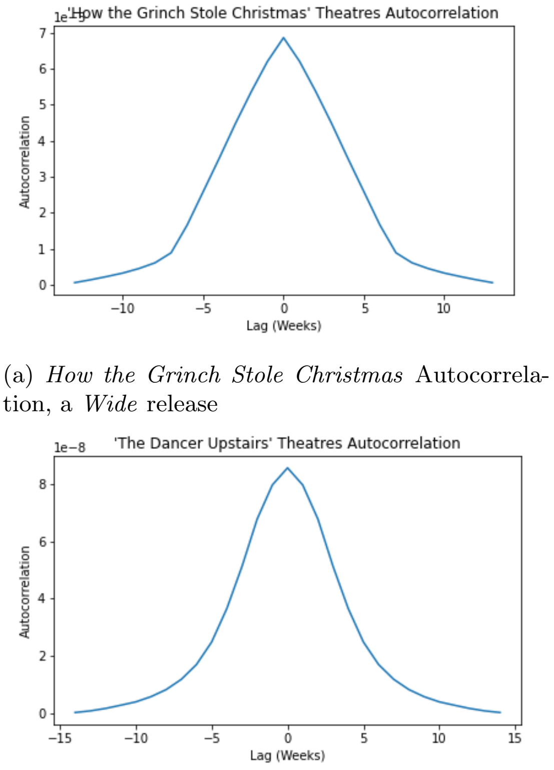
(a) Model Theatres MFE Comparison

(b) Model Theatres MAE Comparison

Figure 6.17: MFE and MAE Comparison of Theatre Forecast

Figure 6.17 shows the MFE and MAE metrics for all the models. The MFE metrics for all the models are positive implying that all the models tend to under-predict; this is preferable to over-predicting as it gives a more 'worst-case' forecast. Most surprisingly, the NN model achieves a lower MFE than any other model - given the nature of the problem one would expect the LSTM models to perform better. However, the nature of MFE metric means that having roughly equal numbers of positive and negative errors equal in magnitude, even if magnitudes are large, can result in a low MFE.

We confirm this by looking to the MAE of the models where the NN exhibits the largest MAE and suggests that the errors of the NN model are in fact larger in magnitude on average. Both the LSTM 1 and LSTM 2 models achieve a far lower MAE than the Ridge regressor and NN. This is expected as the LSTM models are able to use previously seen inputs when making a prediction. What is surprising though is the similarity of their MAEs - provided with more time lagged samples and the ability to remember and use longer sequences we would expect LSTM 2 to perform better than LSTM 1 yet we see almost no difference in performance. This suggests that having knowledge of more past data points makes very little difference. Inspecting theatre autocorrelation plots (Figure 6.18) for the films *How the Grinch Stole Christmas* and *The Dancer Upstairs* which are *Wide* and *Limited, Then Wide* release patterns, respectively, we see that despite the two being of different shapes most of the correlation occurs within a five week period and information beyond five weeks is far more weakly correlated. As the LSTM 1 model already captures and uses the most important information, we conclude that it should perform similarly to the LSTM 2 model.



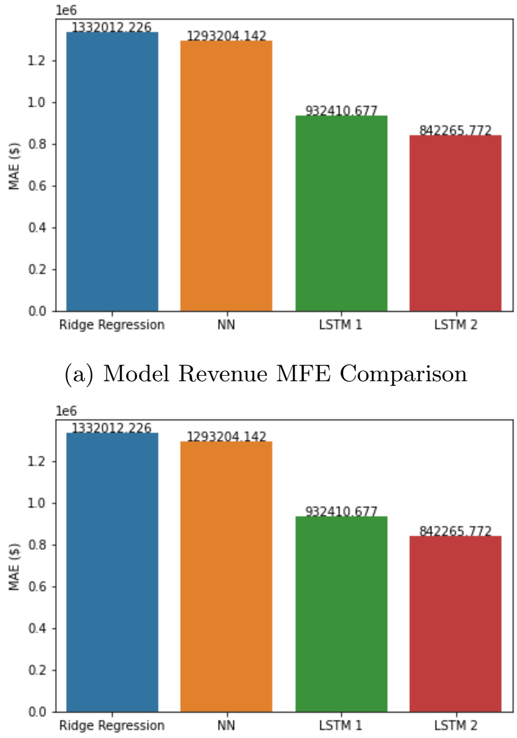
(a) *How the Grinch Stole Christmas* Autocorrelation, a *Wide* release

(b) *The Dancer Upstairs* Autocorrelation, a *Limited, Then Wide* release

Figure 6.18: Theatres Autocorrelation plots for *Wide* and *Limited, Then Wide* release patterns

Revenue Forecast

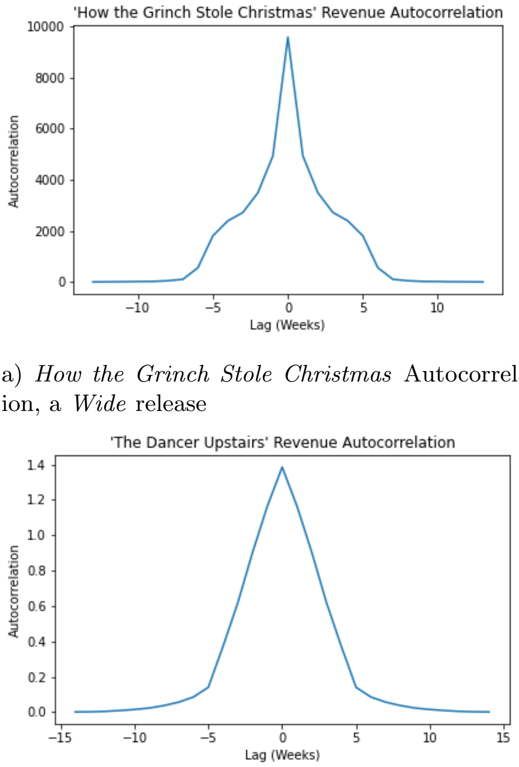
Accurate revenue forecasts enable all involved in the film making process to better estimate how much they are likely earn from the release of a film. Early indication of under-performance can enable studios to cut their losses on films that make less money than anticipated - a poor financial performance could lead to the film's theatrical run being reduced and other means of revenue generation tapped earlier. Figure 6.19 shows the performance metrics for the models for revenue forecasting. The MFEs for most models are positive indicating general under-prediction. The LSTM 1 model, however, exhibits a negative MFE suggesting over-prediction which, although small in magnitude, is nonetheless undesireable. The MAE comparison shows a clear advantage for both LSTM models, with both achieving a significant error reduction over the Ridge regressor and NN models. As with the theatres forecasting, however, both LSTM models achieve very similar performance as the revenue autocorrelation in Figure 6.20 too suggests that even a small window of past data captures most of the important information.



(a) Model Revenue MFE Comparison

(b) Model Revenue MAE Comparison

Figure 6.19: MFE and MAE Comparison of Revenue Forecast



(a) *How the Grinch Stole Christmas* Autocorrelation, a *Wide* release

(b) *The Dancer Upstairs* Autocorrelation, a *Limited, Then Wide* release

Figure 6.20: Revenue Autocorrelation plots for *Wide* and *Limited, Then Wide* release patterns

6.9 Conclusion

In this chapter, we have explored how to model the post-release box office performance of films. After identifying that the short time-series that make up a film's box office performance are non-stationary, we categorised the release patterns of films into three types. We then designed four models to forecast future performance and visualised how forecasts behave for each of these types of release patterns, and identified reasons for their behaviour. We noticed very similar performance between the LSTM 1 and LSTM 2 models, despite LSTM 2 using more LSTM cells that allow it to remember more past data points. We investigated the cause of this and concluded that most of the highly correlated information occurs very shortly before the time of forecast and as a result even models with limited memory capability can perform well, leading to diminishing performance returns as more LSTM cells are added.

Chapter 7

Conclusion

Hasta la vista, baby.

Terminator 2: Judgment Day, 1991

7.1 Summary of Achievements

The aim of this project was to investigate which factors affect the profitability of a film, and develop models to predict film performance in both pre-release and post-release settings. We have explored how to best feature engineer raw data and designed models to answer the *Yes* or *No* question about profitability, as well as make concrete predictions of box office performance before and after release.

We solved a classification problem to predict a film will be profitable using data available prior to release using a Logistic Regression and an *XGBoost* Model. Inspection of the coefficient terms of the logistic regression model showed general insight into which features had a negative or positive impact on profitability and. We performed extended analysis to obtain more informative Shapley values for both models and showed how values of certain features impact profitability. Comparing Shapley feature plots for the logistic regression and *XGBoost* mostly concurred on the importance and effects of features, with the *XGBoost* plot yielding more insight into the effects of magnitudes of values. We identified surprising trends identified during model evaluation and suggested reasons why they may be, recognising that can be other sources of revenue that we do not consider, that contribute to profitability.

In order to make more insightful predictions about film revenues we proposed a regression task to predict revenues per country at key stages in the release, along with length of time the film will spend in theatres per country, and revenues from sales of Blu-rays and DVDs in North America. We designed five different models for this task, with most models also capable of generating a predictive posterior. We calculated Shapley values of input features for each of the regressands and found some very interesting relationships such as the link between geographically close countries despite differences in language, as well as relationships that suggest similarity in the preferences of audiences in even geographically distant countries that do not share a common language. We showed that the box office prediction regressands generally produced high quality estimates and identified a weakness of the models in predicting sales and revenue for Blu-rays and DVDs. We identified the likely cause for this weakness and related this to how well the models were able to utilise relevant features during regression. Finally, we evaluated the performance of these models and compared them with existing models from previous work where we showed that our models perform well despite using only pre-release features.

A technique to model post-release box office using a set of Ordinary Differential Equations was developed and applied to some data. This technique showed promising performance on a few initial data samples. However, we quickly identified limitations of this approach in its' inability to use categorical and quantitative features of films as well as there being insufficient data to pursue this method of modelling.

We developed models that could overcome these limitations to predict post-release performance time-series, two of which employed Long Short-Term Memory units to predict revenue and theatres screening a film a week in the future. On inspecting the feature weights for the Ridge Regression model we found that the number of theatres in the previous week was considered the most informative feature, and that ratings data which was initially considered to be of great value was in fact not as useful. We showed applications of the LSTM based models in predicting for both one week ahead at a time, as well for forecasting for several weeks in the future. We identified interesting features of forecasts and investigated the reasons for these. We evaluated the models and found both LSTM based models outperforming other methods and recognised that the LSTM models exhibited similar performance. The cause of this was further investigated and identified.

7.2 Reflection

7.2.1 Alternative Revenue Streams

Alternative revenue streams include any non-theatrical revenue such as DVD/Blu-Ray, Video on Demand and TV. Major VoD firms often buy the rights to sell/stream films in bulk and do not release detailed breakdowns of sales or views at all. As a result we were not able to obtain any data on sales or revenue from VoD sources. As VoD represents one of the fastest growing revenue streams for films with highly successful films earning huge amounts of money from these sources, we missed out on the opportunity to understand how VoD can affect film revenues. Furthermore, a large proportion of films released see no theatrical release and instead rely on generating revenue from alternative sources in entirety. VoD companies such as Netflix produce films, sometimes with very high budgets and may only offer a limited or no theatrical release. The lack of information on these types of films represents a lost opportunity to gain an insight into this revenue stream.

The one alternative revenue stream we have modelled is the sales of DVDs/Blu-rays for which only limited data was available, primarily being available for films that were already successful at the box office and only for North America. The lack of data and the limited diversity of the data we did have was a hindrance in modelling this aspect of revenue.

7.2.2 Box Office Tracking

Many large markets such as India have no bodies performing box office tracking and as a result the data collected sees many countries underrepresented enough that they offer far too few samples to model. This means that potentially key sources of revenue are omitted entirely. Incomplete tracking of figures greatly affects post-release modelling as having several incomplete release information samples means having to use linear imputation to fill in missing values, which is not representative of real world changes in values.

7.2.3 Model Interpretation

We have primarily made use of Neural Networks and Gaussian Process based models which offer no insight into the reasoning behind the model. Although Shapley feature importance offers some explanation of the inner workings of black box models, they are by no means as informative as inspection of the parameters of a linear regression model would be. As we cannot view the reasoning of black box models we are forced to accept their predictions and evaluate their merits purely on performance in a test scenario.

7.3 Future Work

7.3.1 Combined Models

As we observed in the evaluation of pre-release models, all the models show different levels of performance across each category of regressand. Machine Learning can be used to combine several models in a Stacking approach where predictions from multiple base models can be used as inputs to another, higher level model. The higher level model learns the weaknesses of each base model and uses the base set of predictions to make its own prediction.

7.3.2 Cast/Crew Recommendation

We have tackled the task of predicting performance given some fixed factors, such as the director. But what if we could also recommend actors or directors to maximise performance in something as specific as a single country?

Our method of encoding actors and directors (see Section 3.5.1) can be extended to include more features such as the number of awards received, social media follower counts and product endorsement history. Clustering algorithms can be applied to this data to form groups of similar people. When looking for recommendations, characteristics of a an initially suggested person could be assigned to a cluster and alternatives suggested from both within the cluster, as well as nearby clusters. As a result a selection of people can be suggested for a particular job. Work in this direction is already being explored by the startup, Cinelytic[1].

7.3.3 Using Unstructured Data

A key limitation of our pre-release models is that it does not know anything about the film which it is predicting for. Any metric that quantify the quality of a film prior to release is generated by a human evaluating a film. An interesting extension to our work would be to use a Natural Language Processing model to analyse the script or a Computer Vision model to evaluate the quality of a feature film or trailer and use these in combination with quantitative data for to create a combined model that understands both the structured characteristics of a film as well as the qualitative reasoning about the content to interpret not only how well a film will perform, but also assess how an audience will react to it. Previously, Zhou et al. [58] have used film posters to predict movie box office performance using Convolutional Neural Networks and we believe that building on this approach by analysing trailers, films or scripts could better enable the model to understand what constitutes a successful film and potentially be extended to indicate potential drawbacks or suggest changes to the film.

7.3.4 Search Volumes

An interesting method of gauging the effectiveness of marketing and public interest and anticipation for a film is to look at search query volumes for terms relating to that film. Google published a whitepaper in 2013 (which has since been removed) citing over 90% accuracy in predicting performance. We scraped and attempted to use data from Google Trends[2] for each film. In Google Trends, for search query, the search volumes are normalised to be between 0-100 within a data range. As a result, this doesn't reveal true query numbers. We attempted to reverse the scaling using reference queries that have constant search volumes, but this yielded little success or insight into film performance. However, we believe that using search query volumes can provide great insight into the anticipation for a film that can provide more relevant insight than metrics collected from social media. We maintain that this is an exciting avenue for future research.

[1] https://www.cinelytic.com

[2] https://trends.google.com

Bibliography

[1] MPAA (2018). Theatrical Home Entertainment Report. [online] Available at: https://www.motionpictures.org/research-docs/2018-theatrical-home-entertainment-market-environment-theme-report/ [Accessed 18 Jan. 2020].

[2] European Audiovisual Observatory (2017). VOD distribution and the role of aggregators. [online] Available at: https://rm.coe.int/2017-vod-distribution-and-the-role-of-aggregators-g-fontaine-p-simone-/1680788ff1 [Accessed 18 Jan. 2020].

[3] Stephenfollows.com. (2020). What percentage of independent films are profitable? | Stephen Follows. [online] Available at: https://stephenfollows.com/what-percentage-of-independent-films-are-profitable/ [Accessed 20 Jan. 2020].

[4] Sharon P. Smith and V. Kerry Smith (1986) Successful movies: A preliminary empirical analysis, Applied Economics, 18:5, 501-507, DOI: 10.1080/00036848608537445

[5] Bass, F. (1969). A New Product Growth for Model Consumer Durables. Management Science, 15(5), 215-227. Retrieved January 20, 2020, from www.jstor.org/stable/2628128

[6] Sawhney, M., and Eliashberg, J. (1996). A Parsimonious Model for Forecasting Gross Box-Office Revenues of Motion Pictures. Marketing Science, 15(2), 113-131. Retrieved January 20, 2020, from www.jstor.org/stable/184189

[7] Eliashberg, J., Jonker, J., Sawhney, M., and Wierenga, B. (2000). MOVIEMOD: An Implementable Decision-Support System for Prerelease Market Evaluation of Motion Pictures. Marketing Science, 19(3), 226-243. Retrieved January 20, 2020, from www.jstor.org/stable/193187

[8] John Conlisk (1976) Interactive markov chains, The Journal of Mathematical Sociology, 4:2, 157-185, DOI: 10.1080/0022250X.1976.9989852

[9] D. A. Edwards and R. Buckmire, A differential equation model of North American cinematic box-office dynamics, in IMA Journal of Management Mathematics, vol. 12, no. 1, pp. 41-74, July 2001. DOI: 10.1093/imaman/12.1.41 URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8142457&isnumber=8142453

[10] Kermack, W., and McKendrick, A. (1927). A Contribution to the Mathematical Theory of Epidemics. Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character, 115(772), 700-721. Retrieved January 20, 2020, from www.jstor.org/stable/94815

[11] Clarke E.M., Klieber W., Nováček M., Zuliani P. (2012) Model Checking and the State Explosion Problem. In: Meyer B., Nordio M. (eds) Tools for Practical Software Verification. LASER 2011. Lecture Notes in Computer Science, vol 7682. Springer, Berlin, Heidelberg

[12] Sousi, P. (2009). ODE approximations to some Markov chain models.

[13] Darling, R.W.R.; Norris, J.R. Differential equation approximations for Markov chains. Probab. Surveys 5 (2008), 37–79. DOI:10.1214/07-PS121. https://projecteuclid.org/euclid.ps/1208958281

[14] Pourranjbar A., Hillston J., Bortolussi L. (2013) Don't Just Go with the Flow: Cautionary Tales of Fluid Flow Approximation. In: Tribastone M., Gilmore S. (eds) Computer Performance Engineering. EPEW 2012, UKPEW 2012. Lecture Notes in Computer Science, vol 7587. Springer, Berlin, Heidelberg

[15] J. Hillston, Fluid flow approximation of PEPA models, Second International Conference on the Quantitative Evaluation of Systems (QEST'05), Torino, 2005, pp. 33-42. DOI: 10.1109/QEST.2005.12

[16] L.Bortolussi, J.Hillston, D.Latella, and M.Massink. Continuous Approximation of Collective System Behaviour: A tutorial. Performance Evaluation, 70(5):317 – 349, 2013.

[17] Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological), 58(1), 267-288.

[18] Bozdogan, H. Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. Psychometrika 52, 345–370 (1987) DOI:10.1007/BF02294361

[19] Schwarz, G. (1978). Estimating the Dimension of a Model. The Annals of Statistics, 6(2), 461-464. Retrieved January 20, 2020, from www.jstor.org/stable/2958889

[20] Hoerl, A., and Kennard, R. (2000). Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics, 42(1), 80-86. DOI:10.2307/1271436

[21] En.wikipedia.org. (2020). Least squares. [online] Available at: https://en.wikipedia.org/wiki/Least_squares [Accessed 20 Jan. 2020].

[22] J. A. Nelder, R. Mead, A Simplex Method for Function Minimization, The Computer Journal, Volume 7, Issue 4, January 1965, Pages 308–313.

[23] Mathematics for Machine Learning. (2020). Mathematics for Machine Learning. [online] Available at: https://mml-book.com [Accessed 20 Jan. 2020].

[24] Ruder, S. (2016). An overview of gradient descent optimization algorithms. ArXiv, abs/1609.04747.

[25] Gopinath, S., Chintagunta, P., and Venkataraman, S. (2013). Blogs, Advertising, and Local-Market Movie Box Office Performance. Management Science, 59(12), 2635-2654. Retrieved January 19, 2020, from www.jstor.org/stable/42919500

[26] Kanika Almadi. Quantitative Study of the Movie Industry Based on IMDb Data. Available at: https://dspace.mit.edu/handle/1721.1/113502

[27] En.wikipedia.org. (2020). Artificial neural network. [online] Available at: https://en.wikipedia.org/wiki/Artificia_neural_network [Accessed 20 Jan. 2020].

[28] Sharda, R., and Delen, D. (2006). Predicting box-office success of motion pictures with neural networks. Expert Systems with Applications, 30(2), 243-254.

[29] Nika, M. (2014). Synthedemic Modelling and Prediction of Internet-based Phenomena. Ph.D. Imperial College London.

[30] En.wikipedia.org. (2020). The Irishman. [online] Available at: https://en.wikipedia.org/wiki/The_Irishman [Accessed 20 Jan. 2020].

[31] Rasmussen, C. and Williams, C., 2008. Gaussian Processes For Machine Learning. Cambridge, Mass: MIT Press.

[32] Duvenaud, D., 2014. Automatic Model Construction With Gaussian Processes. Ph.D. University of Cambridge.

[33] Radford M. Neal. 1996. Bayesian Learning for Neural Networks. Springer-Verlag, Berlin, Heidelberg.

[34] Lee, J., Bahri, Y., Novak, R., Schoenholz, S.S., Pennington, J., and Sohl-Dickstein, J. (2018). Deep Neural Networks as Gaussian Processes. ArXiv, abs/1711.00165.

[35] Joaquin Quiñonero-Candela and Carl Edward Rasmussen. 2005. A Unifying View of Sparse Approximate Gaussian Process Regression. J. Mach. Learn. Res. 6 (12/1/2005), 1939–1959.

[36] Edward Snelson and Zoubin Ghahramani. 2005. Sparse Gaussian processes using pseudo-inputs. In Proceedings of the 18th International Conference on Neural Information Processing Systems (NIPS'05). MIT Press, Cambridge, MA, USA, 1257–1264.

[37] Christopher K. I. Williams and Matthias Seeger. 2000. Using the Nyström method to speed up kernel machines. In Proceedings of the 13th International Conference on Neural Information Processing Systems (NIPS'00). MIT Press, Cambridge, MA, USA, 661–667.

[38] Titsias, M.. 2009. Variational Learning of Inducing Variables in Sparse Gaussian Processes. Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics, in PMLR 5:567-574

[39] Molnar, Christoph. Interpretable machine learning. A Guide for Making Black Box Models Explainable, 2019. https://christophm.github.io/interpretable-ml-book/.

[40] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 4768–4777.

[41] Erik Štrumbelj and Igor Kononenko. 2014. Explaining prediction models and individual predictions with feature contributions. Knowl. Inf. Syst. 41, 3 (December 2014), 647–665. DOI:https://DOI.org/10.1007/s10115-013-0679-x

[42] IMDb. IMDb Datasets. Retrieved 30/01/2020 from https://www.imdb.com/interfaces/

[43] Antipov, Evgeny and Pokryshevskaya, Elena. 2010. Accounting for latent classes in movie box office modeling. Journal of Targeting, Measurement and Analysis for Marketing. 19. 10.2139/ssrn.1729631.

[44] Galvão, Marta and Henriques, Roberto. 2018. Forecasting Movie Box Office Profitability. Journal of Information Systems Engineering and Management. 3. 10.20897/jisem/2658.

[45] Quader, Nahid and Gani, Md and Chaki, Dipankar and Ali, Md. 2018. A Machine Learning Approach to Predict Movie Box-Office Success. 10.1109/ICCITECHN.2017.8281839.

[46] T. G. Rhee and F. Zulkernine, Predicting Movie Box Office Profitability: A Neural Network Approach, 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), Anaheim, CA, 2016, pp. 665-670, DOI: 10.1109/ICMLA.2016.0117.

[47] Deloitte. 2017. China's Film Industry – A New Era. [online] Available at: <https://www2.deloitte.com/content/dam/Deloitte/cn/Documents/technology-media-telecommunications/deloitte-cn-tmt-china-film-industry-en-161223.pdf> [Accessed 25 April 2020].

[48] Simonoff, J., 2015. Predicting Total Movie Grosses After One Week. [ebook] Available at: <http://http://people.stern.nyu.edu/jsimonof/classes/2301/pdf/movies.pdf> [Accessed 27 May 2020].

[49] Dey, S. 2018. Predicting Gross Movie Revenue. ArXiv, abs/1804.03565.

[50] Pangarker, N.A. and V.d.M. Smit, E. 2013. The determinants of box office performance in the film industry revisited. South African Journal of Business Management, 44, 47-58 (2013)

[51] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, Ruslan Salakhutdinov. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research 2014; 15(56):1929-1958.

[52] Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48 (ICML'16). JMLR.org, 1050–1059.

[53] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning - Volume 70 (ICML'17). JMLR.org, 1321–1330.

[54] Christopher Olah. Understanding LSTM networks. 2015. URL https://colah.github.io/posts/2015-08-Understanding-LSTMs/[Accessed 19 March 2020].

[55] Zhu, L, Laptev, N. Deep and Confident Prediction for Time Series at Uber. 2017. IEEE International Conference on Data Mining Workshops (ICDMW) 2017.

[56] Markowitz, H. 1952. Portfolio Selection. The Journal of Finance, 7(1), 77-91. DOI:10.2307/2975974

[57] Saylordotorg.github.io. 2020. The History Of Movies. [online] Available at: <https://saylordotorg.github.io/text_understanding-media-and-culture-an-introduction-to-mass-communication/s11-01-the-history-of-movies.html> [Accessed 7 June 2020].

[58] Zhou, Y., Zhang, L. and Yi, Z. 2019. Predicting movie box-office revenues using deep neural networks. Neural Comput and Applic 31, 1855–1865 (2019). https://DOI.org/10.1007/s00521-017-3162-x

[59] Gal, Y. Uncertainty in Deep Learning, Doctoral dissertation, University of Cambridge.

[60] The Connection Of Dropout And Bayesian Statistics. [ebook] Available at: <https://tensorchiefs.github.io/bbs/files/dropouts-brownbag.pdf> [Accessed 12 June 2020].