VISVESVARAYA TECHNOLOGICAL UNIVERSITY

"Jnana Sangama", Belgaum: 590018

A Project report on

"PREDICTIVE TTM OPTIMIZATION"

A Dissertation work submitted in partial fulfillment of the requirement for the award of the degree of

Bachelor of Engineering
in
Computer Science and Engineering

by

Aditya Das	**1AY14CS010**
Ankur Vinekar	**1AY14CS019**
Chirag DK	**1AY14CS033**
M Arjun	**1AY14CS060**

Under the guidance of
Dr. P V Kumar
Professor

ACHARYA

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

ACHARYA INSTITUTE OF TECHNOLOGY

(Affiliated to Visvesvaraya Technological University, Belgaum)

2017-2018

ACHARYA INSTITUTE OF TECHNOLOGY

Acharya Dr. Sarvepalli Radhakrishnan Road, Soladevanahalli, Bangalore – 560107
(Affiliated to Visvesvaraya Technological University, Belgaum)

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

Certificate

Certified that the Project entitled **"Predictive TTM Optimization"** is a bona fide work carried out by **Aditya Das (1AY14CS010), Ankur A Vinekar (1AY14CS019), Chirag DK (1AY14CS033)** and **M Arjun (1AY14CS060)** in partial fulfillment for the award of degree of **Bachelor of Engineering in Computer Science & Engineering of Visvesvaraya Technological University**, Belgaum during the year **2017-2018.** It is certified that all corrections/suggestions indicated for internal assessments have been incorporated in the Report deposited in the departmental library. The project report has been approved as it satisfies the academic requirements in respect of project prescribed for the **Bachelor of Engineering Degree**.

Signature of Guide	**Signature of H.O.D**	**Signature of Principal**
Dr. P V Kumar	Dr. Prashanth C M	Principal
Professor	Head of the Department	

External Viva

Name of the Examiners	**Signature with Date**
1.	
2.	

ACKNOWLEDGEMENT

We express our gratitude to our institution and management for providing us with good infrastructure, laboratory, facilities and inspiring staff, and whose gratitude was of immense help in completion of this report successfully.

We deeply indebted to **Dr. Sharanabasava C Pilli,** Principal, Acharya Institute of Technology, Bangalore, who has been a constant source of enthusiastic inspiration to steer us forward.

We hearty thank **Dr. Prashanth C M**, Head of the Department and Dean R&D, Department of Computer Science and Engineering, Acharya Institute of Technology Bangalore, for his valuable support and for rendering us resources for this project work.

We specially thank **Dr. P V Kumar,** Professor, Department of Computer Science and Engineering who guided us with valuable suggestions in completing this project at every stage.

Also, we wish to express deep sense of gratitude for project coordinator **Prof. B Gayathri Kamath**, Assistant Professor, Department of Computer Science and Engineering, Acharya Institute of Technology for her support and advice during the course of this final year project.

We would like to express our sincere thanks and heartfelt gratitude to our beloved Parents, Respected Professors, Classmates, Friends, Juniors for their indispensable help at all times.

Last but not the least our respectful thanks to the Almighty.

Aditya Das **(1AY14CS010)**

Ankur A Vinekar **(1AY14CS019)**

Chirag DK **(1AY14CS033)**

M Arjun **(1AY14CS060)**

ABSTRACT

Forecasting a movie's opening success is a laborious task, as it does not always depend on its quality. External factors such as competing movies, time of the year and even weather influence the success as these factors impact the Box-office sales for the moving opening.

Nevertheless, predicting a movie's opening success in terms of Box-office ticket sales is essential for a movie studio, to plan its cost and make the work profitable.

Release time being wholly essential to the success of the movie, studios decide and pre-announce the envisaged release date long before the actual reveal of their upcoming movies. This choice of dates and then the following changes are tactical in nature taking into deliberation various factors like release of competing movies, holidays, sports event, natural disaster etc. Machine learning mixed with predictive analytics using previously archived data and their box office performance can help us identify the most apt release date for maximizing the revenue.

Predictive TTM optimization focuses on predicting the optimal release date for a movie using machine learning based on factors such as cast, budget and recent trends in the film industry. This analysis will also empower the producers as to what would be the most apt time to release a movie according to the state of the overall market.

CONTENTS

LIST OF FIGURES

Chapter 1

Introduction

CHAPTER 1

INTRODUCTION

With India being the churner of the largest number of motion pictures and third in the world in terms of revenue, movies are a part and parcel of every Indian's life. The budget of these movies is usually in the order of crores of Rupees, making their box office success utterly imperative for the survival of the industry. Having the knowledge that which movies are likely to make headway and which are likely to fail before the release could support the production houses greatly as it will enable them to focus their advertising campaigns which itself costs crores.

This analysis will also empower the producers as to what would be the most apt time to release a movie according to the state of the overall market.

Release time being wholly essential to the success of the movie, studios decide and pre-announce the envisaged release date long before the actual reveal of their upcoming movies. This choice of dates and then the following changes are tactical in nature taking into consideration various factors like release of competing movies, holidays, sports event, natural disaster etc. Machine learning mixed with predictive analytics using previously archived data and their box office performance can help us identify the most apt release date for maximizing the revenue.

Forecasting a movie's opening success is a laborious task, as it does not always depend on its quality. External factors such as competing movies, time of the year and even weather influence the success as these factors impact the Box-office sales for the moving opening. Nevertheless, predicting a movie's opening success in terms of Box-office ticket sales is essential for a movie studio, to plan its cost and make the work profitable.

Predictive TTM Optimization plans to tap into one of the largest sectors of entertainment available today. Though the box office looks innocuous enough to seem like just an industry that deals with the fun and entertainment, it is far from the truth. The film industry employs lakhs of people in our country itself. They are usually at the forefront of cutting edge technology, funding expensive, next-gen techniques. Techniques that other

sectors cannot invest in due to the unavailability of funds. And because the film industry is cash strapped, they end up indirectly funding a lot of technologies that go onto become next big thing. Things like 3D movies, high pressure cameras are one of few wonders kick-started by the film industry.

With the advent of AI and widespread use of machine learning, we are at the epoch of a new era in computing. The field of machine learning has matured by leaps and bounds. We are able to guess diseases in human beings long before they become fatal, predict and avert natural disasters. We cannot even start contemplating the endless areas machine learning could help humanity to solve their umpteen problems.

Machine learning algorithms are widely used to make predictions such as growth in the stock market, demand for products, nature of tumors, etc. This project presents a brief study on the viability of our project and the types of reinforced Machine Learning models we might use to predict movie success rates and time of release.

Thus, keeping the interests of the film industry is a very crucial thing. We have therefore decided upon a Machine learning project that predicts the success of films, taking into consideration a lot of factors like the track history of the film production house, actors, directors, release time, competing films etc.

Chapter 2

Related Technologies and Concepts

CHAPTER 2

RELATED TECHNOLOGIES AND CONCEPTS

Predictive TTM Optimization is a real-world project which consists of many different concepts, sources and technologies. By bringing machine learning, a highly computer-oriented technology, to the film industry, one of the ever-booming industries, we have combined major aspects of our daily life.

Some of the most predominant matters of importance are explained further in detail to provide a better understanding of the environment of the project.

2.1 The Film Industry

The film industry or motion picture industry comprises the technological and commercial institutions of filmmaking, i.e., film production companies, film studios, cinematography, animation, film production, screenwriting, pre-production, post production, film festivals, distribution; and actors, film directors, music composers, sound effects workers and other film crew personnel.

One of the most important part of the film industry is marketing and release analysis, which is where Predictive TTM Optimization hopes to aid the industry.

Though the expense involved in making films almost immediately led film production to concentrate under the auspices of standing production companies, advances in affordable film making equipment, and expansion of opportunities to acquire investment capital from outside the film industry itself, have allowed independent film production to evolve.

Fig 2.1: The iconic Hollywood Sign

Hollywood is the oldest film industry of the world, and the largest in terms of box office gross revenue. Indian cinema is the largest film industry in terms of the number of films produced and the number of tickets sold domestically, with 1.9 billion tickets sold in 2016.

The film industry has revolutionized technology due to the high demand and the high amount of funding it possesses. It led to the creation of 3D displays, it led to the invention of high pressure underwater cameras and has still been contributing to being the cause of newer technologies being developed to this date.

2.2 IMDb – Internet Movie Database

IMDb, also known as Internet Movie Database, is an online database of information related to world films, television programs, home videos and video games, and internet streams, including cast, production crew, personnel and fictional character biographies, plot summaries, trivia, and fan reviews and ratings. Originally a fan-operated website, the database is owned and operated by IMDb.com Inc., a subsidiary of Amazon.

IMDb does not provide an API for automated queries. However, most of the data can be downloaded as compressed plain text files and the information can be extracted using the command-line interface tools provided. There is also a Java-based graphical user

interface (GUI) application available that is able to process the compressed plain text files, which allows a search and a display of the information. This GUI application supports different languages, but the movie related data are in English, as made available by IMDb.

A Python package called IMDbPY can also be used to process the compressed plain text files into a number of different SQL databases, enabling easier access to the entire dataset for searching or data mining.

IMDb is the place from where this project collects its resources such as the datasets which are used in this project. It provides a full-fledged set of details for every movie ever made (of major languages) which has helped this project become a success.

2.3 Machine Learning

Machine learning is a field of computer science that uses statistical techniques to give computer systems the ability to "learn" (e.g., progressively improve performance on a specific task) with data, without being explicitly programmed.

Machine learning explores the study and construction of algorithms that can learn from and make predictions on data – such algorithms overcome following strictly static program instructions by making data-driven predictions or decisions, through building a model from sample inputs. Machine learning is employed in a range of computing tasks where designing and programming explicit algorithms with good performance is difficult or infeasible; example applications include email filtering, detection of network intruders or malicious insiders working towards a data breach, optical character recognition (OCR), learning to rank, and computer vision.

Machine learning is closely related to (and often overlaps with) computational statistics, which also focuses on prediction-making through the use of computers. It has strong ties to mathematical optimization, which delivers methods, theory and application domains to the field. Machine learning is sometimes conflated with data mining, where the latter subfield focuses more on exploratory data analysis and is known as unsupervised learning. Machine learning can also be unsupervised and be used to learn and establish baseline behavioral profiles for various entities and then used to find meaningful anomalies.

Within the field of data analytics, machine learning is a method used to devise complex models and algorithms that lend themselves to prediction; in commercial use, this

is known as predictive analytics. These analytical models allow researchers, data scientists, engineers, and analysts to "produce reliable, repeatable decisions and results" and uncover "hidden insights" through learning from historical relationships and trends in the data.

2.4 Predictive Analysis

Predictive analytics encompasses a variety of statistical techniques from predictive modelling, machine learning, and data mining that analyze current and historical facts to make predictions about future or otherwise unknown events.

In business, predictive models exploit patterns found in historical and transactional data to identify risks and opportunities. Models capture relationships among many factors to allow assessment of risk or potential associated with a particular set of conditions, guiding decision making for candidate transactions.

Predictive Analysis Process

1. Define Project: Define the project outcomes, deliverable, scope of the effort, business objectives, identify the data sets that are going to be used.
2. Data Collection: Data mining for predictive analytics prepares data from multiple sources for analysis. This provides a complete view of customer interactions.
3. Data Analysis: Data Analysis is the process of inspecting, cleaning and modelling data with the objective of discovering useful information, arriving at conclusion
4. Statistics: Statistical Analysis enables to validate the assumptions, hypothesis and test them using standard statistical models.
5. Modelling: Predictive modelling provides the ability to automatically create accurate predictive models about future. There are also options to choose the best solution with multi-modal evaluation.
6. Deployment: Predictive model deployment provides the option to deploy the analytical results into everyday decision-making process to get results, reports and output by automating the decisions based on the modelling.
7. Model Monitoring: Models are managed and monitored to review the model performance to ensure that it is providing the results expected.

2.5 R Programming Language

R is a programming language and free software environment for statistical computing and graphics that is supported by the R Foundation for Statistical Computing. The R language is widely used among statisticians and data miners for developing statistical software and data analysis. Polls, surveys of data miners, and studies of scholarly literature databases show that R's popularity has increased substantially in recent years.

R is a GNU package. The source code for the R software environment is written primarily in C, Fortran, and R. R is freely available under the GNU General Public License, and pre-compiled binary versions are provided for various operating systems. While R has a command line interface, there are several graphical front-ends and integrated development environments available.

R and its libraries implement a wide variety of statistical and graphical techniques, including linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, and others. R is easily extensible through functions and extensions, and the R community is noted for its active contributions in terms of packages. Many of R's standard functions are written in R itself, which makes it easy for users to follow the algorithmic choices made.

2.5.1 R Studio

RStudio is a free and open-source integrated development environment (IDE) for R, a programming language for statistical computing and graphics. RStudio was founded by JJ Allaire, creator of the programming language ColdFusion. Hadley Wickham is the Chief Scientist at RStudio.

RStudio is available in two editions: RStudio Desktop, where the program is run locally as a regular desktop application; and RStudio Server, which allows accessing RStudio using a web browser while it is running on a remote Linux server. Prepackaged distributions of RStudio Desktop are available for Windows, macOS, and Linux.

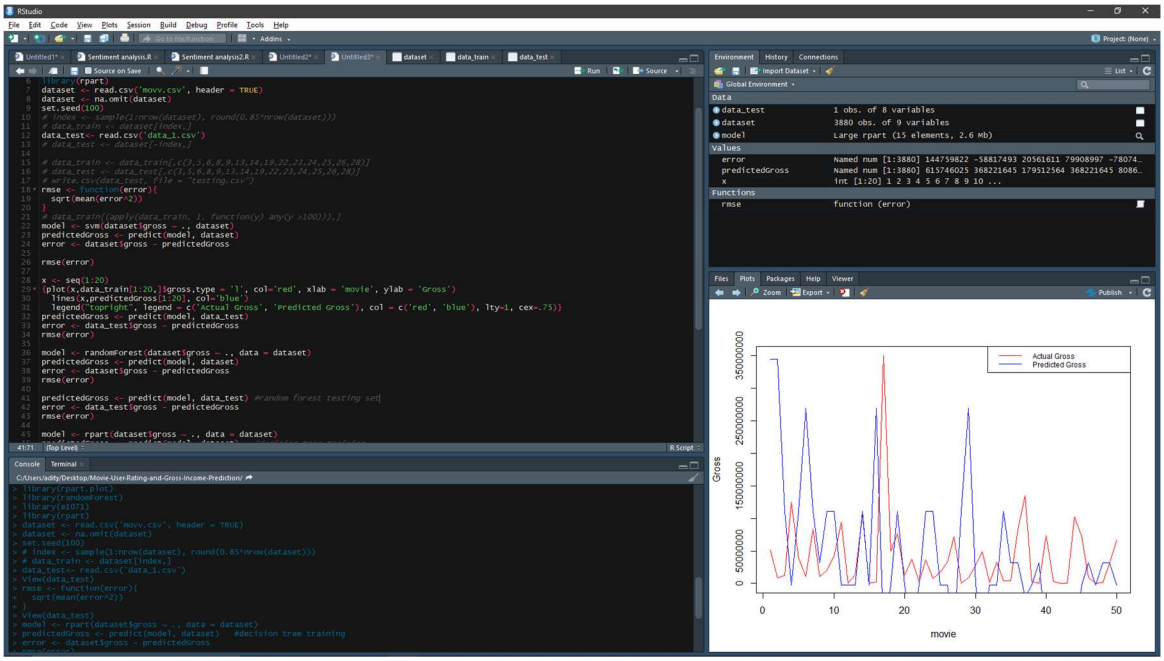


Fig 2.2: R Studio Interface

Chapter 3

Literature Survey

CHAPTER 3

LITERATURE SURVEY

The movie industry is a multi-billion-dollar industry, generating billions in revenue annually. In recent years, movies have generally become divided into two categories: blockbusters and independent movies. Studios have focused on relying on only a handful of extremely expensive movies every year to make sure they remain profitable. It is estimated that 80% of the industry's profits over the last decade is generated from just 6% of the films released. 78% of movies have lost money of the same time-period. These blockbuster movies emphasize the spectacular: casting as much star power as possible and pairing it with high production value. The result of this is a sky-rocketing budget. It is estimated that the average movie now costs $100.3 million after including production and marketing expenses.2 However, "Hollywood is the land of hunch and wild guess," so it's difficult to predict whether these high-budget films will make a profit. As the costs of movies have gone up, it has become paramount that movies are successful to justify such large undertakings. Studios are under great pressure to ensure their movies succeed, trying to find ways to produce movies that are more likely to be successful [1].

Because of this the movie industry has attempted to employ the help of computer scientist to create recommendation and predictive software to tackle this problem. Recommendation software is more common and attempts to make correlations between a consumer's past choices and other products they might like. But there is a great dearth for predicting movies gross and its success based on dependent parameters of a movie with the help of machine learning algorithms and predictive analysis [2].

By analyzing two different models (regression and k-nearest neighbor models), we find models using categorical data from the Internet Movie Database (IMDB). Moreover, we can achieve better performance by using the combination of IMDB data and news data. Further, the improvement is statistically significant [2].

Entities	Duration	Gross (Pre-rel)	Gross (Post-rel)	Budget (Pre-rel)	Budget (Post-rel)
Movie	1 week	0.707	0.781	0.497	0.480
	1 month	0.672	0.779	0.463	0.474
	4 months	0.629	0.749	0.437	0.455
Director	1 week	0.494	0.602	0.311	0.389
	1 month	0.371	0.495	0.218	0.389
	4 months	0.192	0.317	0.117	0.078
Top 3 Actors	1 week	0.640	0.726	0.476	0.528
	1 month	0.569	0.683	0.448	0.477
	4 months	0.493	0.618	0.413	0.424
Top 15 Actors	1 week	0.646	0.725	0.533	0.595
	1 month	0.575	0.686	0.477	0.530
	4 months	0.511	0.618	0.415	0.433

Fig 3.1: Correlation Coefficient of Entities with film affecting parameters

Some of the literature survey on individual domains of this topic are –

- **Data cleaning** - data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data [1]. Data cleansing may be performed interactively with data wrangling tools, or as batch processing through scripting. After cleansing, a data set should be consistent with other similar data sets in the system. The inconsistencies detected or removed may have been originally caused by user entry errors, by corruption in transmission or storage, or by different data dictionary definitions of similar entities in different stores. Data cleaning differs from data validation in that validation almost invariably means data is rejected from the system at entry and is performed at the time of entry, rather than on batches of data [5].
- **R** - R is a programming language and free software environment for statistical computing and graphics that is supported by the R Foundation for Statistical Computing. The R language is widely used among statisticians and data miners for developing statistical software and data analysis. Polls, surveys of data miners, and studies of scholarly literature databases show that R's popularity has increased substantially in recent years. As of May 2018, R ranks 11th in the TIOBE index, a measure of popularity of programming languages [6].

- **Linear regression** - Linear regression is a very simple approach for supervised learning. Though it may seem somewhat dull compared to some of the more modern algorithms, linear regression is still a useful and widely used statistical learning method. Linear regression is used to predict a quantitative response Y from the predictor variable X. Linear Regression is made with an assumption that there's a linear relationship between X and Y [7].
- **Random Forest** - Random Forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because it's simplicity and the fact that it can be used for both classification and regression tasks. One big advantage of random forest is, that it can be used for both classification and regression problems, which form the majority of current machine learning systems.

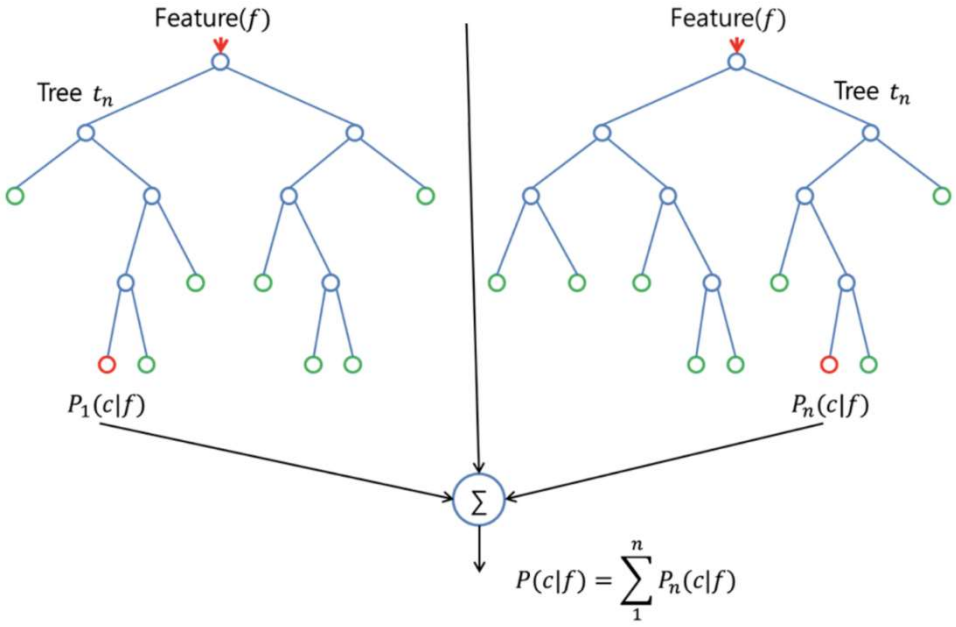


Fig 3.2: Decision Trees

Chapter 4

Implementation

CHAPTER 4

IMPLEMENTATION

Predictive TTM Optimization has multiple steps involved in its implementation. Each step is an important constituent of the process and it is a linear process where the flow goes from one process to the other one after the other. Each of the processes are listed below with details and information about each process in detail.

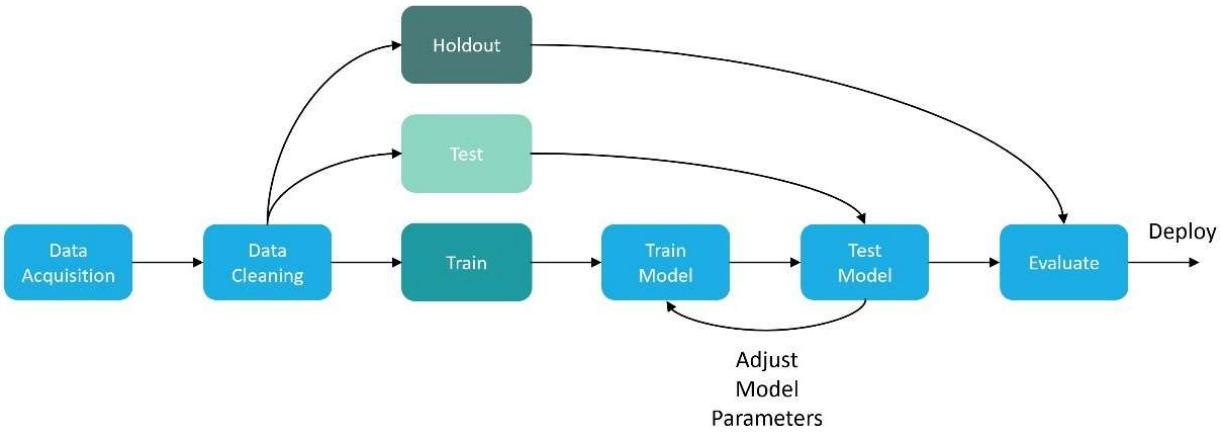


Fig 4.1: ML model and processes

The process has the following steps:

1. Data acquisition
2. Data cleaning
3. Test, train, holdout
4. Train model
5. Test and Evaluate
6. Deploy

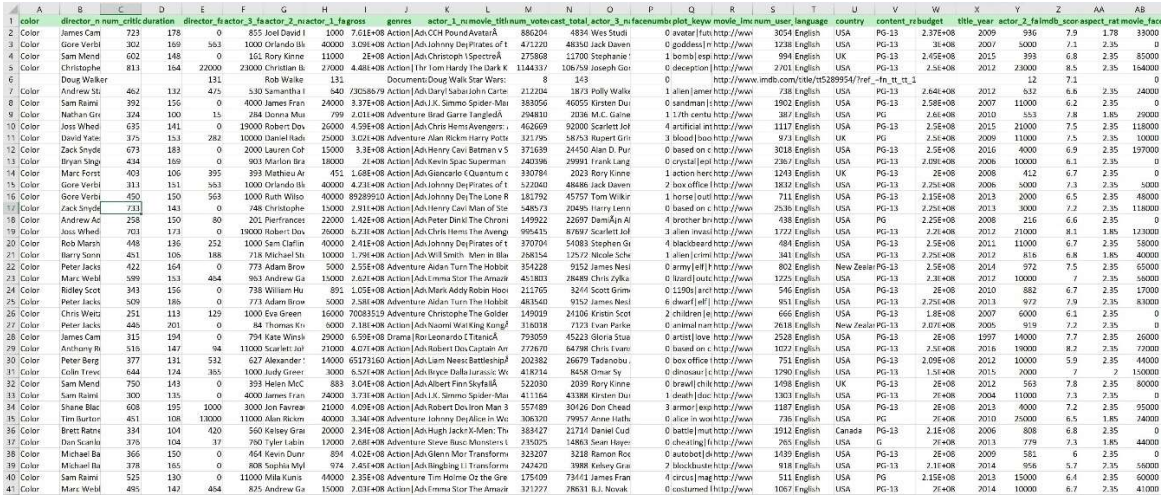
4.1 Data Acquisition

The very first step of the entire process is to acquire data to use for training the machine learning model. This data can be in any form such as images, videos, sounds, textual data, spreadsheets or datasets.

This step is concerned with selecting the subset of all available data that you will be working with. There is always a strong desire for including all data that is available, that the maxim “more is better” will hold. This may or may not be true.

For this project, a dataset suits the necessary requirements and thus the chosen training data is a dataset in the form of a csv file (comma separated values).

Predictive TTM Optimization has a unique requirement of values to be analyzed which were unavailable collectively from a single source. Thus, a dataset had to be manually created by combining various datasets from multiple sources like IMDb and Kaggle.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB
1	color	director_n	num_critic	duration	director_fa	actor_3_fa	actor_2_n	actor_1_fa	gross	genres	actor_1_n	movie_titl	num_voter	cast_total	actor_3_n	facenumbe	plot_keyw	movie_imc	num_user	language	country	content_re	budget	title_year	actor_2_fa	imdb_scor	aspect_rat	movie_face
2	Color	James Cam	723	178	0	855	Joel David I	1000	7.61E+08	Action|Adv	CCH Pound	AvatarÂ	886204	4834	Wes Studi	0	avatar|futu	http://www	3054	English	USA	PG-13	2.37E+08	2009	936	7.9	1.78	33000
3	Color	Gore Verbi	302	169	563	1000	Orlando Bl	40000	3.09E+08	Action|Adv	Johnny De	Pirates of t	471220	48350	Jack Daven	0	goddess|n	http://www	1238	English	USA	PG-13	3E+08	2007	5000	7.1	2.35	0
4	Color	Sam Mend	602	148	0	161	Rory Kinne	11000	2E+08	Action|Adv	Christoph	SpectreÂ	275868	11700	Stephanie	1	bomb|espi	http://www	994	English	UK	PG-13	2.45E+08	2015	393	6.8	2.35	85000
5	Color	Christophe	813	164	22000	23000	Christian Bi	27000	4.48E+08	Action|Thr	Tom Hardy	The Dark K	1144337	106759	Joseph Gor	0	deception|	http://www	2701	English	USA	PG-13	2.55E+08	2012	23000	8.5	2.35	164000
6		Doug Walker		131		Rob Walke		131		Documentar	Doug Walk	Star Wars:		8	143		0	http://www	imdb.com/title/tt5289954/?ref_=fn_tt_tt_1				12		7.1		0
7	Color	Andrew St	462	132	475	530	Samantha I	640	230588679	Action|Adv	Daryl Saba	John Carte	212204	1873	Polly Walke	1	alien|amer	http://www	738	English	USA	PG-13	2.64E+08	2012	632	6.8	2.35	24000
8	Color	Sam Raimi	392	156	0	4000	James Fran	24000	3.37E+08	Action|Adv	J.K. Simmo	Spider-Mar	383056	46055	Kirsten Dur	0	sandman|s	http://www	1902	English	USA	PG-13	2.58E+08	2007	11000	6.2	2.35	0
9	Color	Nathan Gre	324	100	15	284	Donna Mur	799	2.01E+08	Adventure	Brad Garre	TangledÂ	294810	2036	M.C. Gaine	1	17th centu	http://www	387	English	USA	PG	2.6E+08	2010	553	7.8	1.85	29000
10	Color	Joss Whed	635	141	0	19000	Robert Dov	26000	4.59E+08	Action|Adv	Chris Hems	Avengers: /	462669	92000	Scarlett Joh	4	artificial int	http://www	1117	English	USA	PG-13	2.5E+08	2015	21000	7.5	2.35	118000
11	Color	David Yate	375	153	282	10000	Daniel Radc	25000	3.02E+08	Adventure	Alan Rickm	Harry Potte	321795	58753	Rupert Grin	3	blood|boo	http://www	971	English	UK	PG	2.5E+08	2009	11000	7.5	2.35	10000
12	Color	Zack Snyde	673	183	0	2000	Lauren Coh	15000	3.3E+08	Action|Adv	Henry Cavi	Batman v S	371639	24450	Alan D. Pur	0	based on c	http://www	3018	English	USA	PG-13	2.5E+08	2016	4000	6.9	2.35	197000
13	Color	Bryan Singe	434	169	0	903	Marlon Bra	18000	2E+08	Action|Adv	Kevin Spac	Superman	240396	29991	Frank Lang	0	crystal|epi	http://www	2367	English	USA	PG-13	2.09E+08	2006	10000	6.1	2.35	0
14	Color	Marc Forst	403	106	395	393	Mathieu Ar	451	1.68E+08	Action|Adv	Giancarlo E	Quantum c	330784	2023	Rory Kinne	1	action herc	http://www	1243	English	UK	PG-13	2E+08	2008	412	6.7	2.35	0
15	Color	Gore Verbi	313	151	563	1000	Orlando Bl	40000	4.23E+08	Action|Adv	Johnny De	Pirates of t	522040	48486	Jack Daven	2	box office|	http://www	1832	English	USA	PG-13	2.25E+08	2006	5000	7.3	2.35	5000
16	Color	Gore Verbi	450	150	563	1000	Ruth Wilso	40000	89289910	Action|Adv	Johnny De	The Lone R	181792	45757	Tom Wilkir	1	horse|outl	http://www	711	English	USA	PG-13	2.15E+08	2013	2000	6.5	2.35	48000
17	Color	Zack Snyde	733	141	0	748	Christophe	15000	2.91E+08	Action|Adv	Henry Cavi	Man of Ste	548573	20495	Harry Lenn	0	based on c	http://www	2536	English	USA	PG-13	2.25E+08	2013	3000	7.2	2.35	118000
18	Color	Andrew Ad	258	150	80	201	Pierfrances	22000	1.42E+08	Action|Adv	Peter Dinkl	The Chroni	149922	22697	DamÃ¡n Al	4	brother bn	http://www	438	English	USA	PG	2.25E+08	2008	216	6.6	2.35	0
19	Color	Joss Whed	703	173	0	19000	Robert Dov	26000	6.23E+08	Action|Adv	Chris Hems	The Avenge	995415	87697	Scarlett Joh	3	alien invasi	http://www	1722	English	USA	PG-13	2.2E+08	2012	21000	8.1	1.85	123000
20	Color	Rob Marsh	448	136	252	1000	Sam Claflin	40000	2.41E+08	Action|Adv	Johnny De	Pirates of t	370704	54083	Stephen Gr	4	blackbeard	http://www	484	English	USA	PG-13	2.5E+08	2011	11000	6.7	2.35	58000
21	Color	Barry Sonn	451	106	188	718	Michael Sti	10000	1.79E+08	Action|Adv	Will Smith	Men in Blac	268154	12572	Nicole Sche	1	alien|crimi	http://www	341	English	USA	PG-13	2.25E+08	2012	816	6.8	1.85	40000
22	Color	Peter Jacks	422	164	0	773	Adam Brov	5000	2.55E+08	Adventure	Aidan Turn	The Hobbit	354228	9152	James Nesl	0	army|elf|h	http://www	802	English	New Zealar	PG-13	2.5E+08	2014	972	7.5	2.35	65000
23	Color	Marc Webb	599	153	484	963	Andrew Ga	15000	2.62E+08	Action|Adv	Emma Stor	The Amazin	451803	26489	Chris Zylka	0	lizard|outc	http://www	1225	English	USA	PG-13	2.3E+08	2012	10000	7	2.35	58000
24	Color	Ridley Scot	343	156	0	738	William Hu	891	1.03E+08	Action|Adv	Mark Addy	Robin Hood	211765	3244	Scott Grim	0	1190s|arct	http://www	546	English	USA	PG-13	2E+08	2010	882	6.7	2.35	17000
25	Color	Peter Jacks	509	186	0	773	Adam Brov	5000	2.58E+08	Adventure	Aidan Turn	The Hobbit	483540	9152	James Nesl	6	dwarf|elf|	http://www	951	English	USA	PG-13	2.25E+08	2013	972	7.9	2.35	83000
26	Color	Chris Weitz	251	113	129	1000	Eva Green	16000	70083519	Adventure	Christophe	The Golden	149019	24106	Kristin Scot	2	children|e	http://www	666	English	USA	PG-13	1.8E+08	2007	6000	6.1	2.35	0
27	Color	Peter Jacks	446	201	0	84	Thomas Kr	6000	2.18E+08	Action|Adv	Naomi Wat	King Kong	316018	7123	Evan Parke	0	animal nan	http://www	2618	English	New Zealar	PG-13	2.07E+08	2005	919	7.2	2.35	0
28	Color	James Cam	315	194	0	794	Kate Winsl	29000	6.59E+08	Drama|Ror	Leonardo D	TitanicÂ	793059	45223	Gloria Stua	0	artist|love	http://www	2528	English	USA	PG-13	2E+08	1997	14000	7.7	2.35	26000
29	Color	Anthony Ru	516	147	94	11000	Scarlett Joh	21000	4.07E+08	Action|Adv	Robert Dov	Captain Am	772670	84798	Chris Evans	0	based on c	http://www	3022	English	USA	PG-13	2.5E+08	2016	19000	8.2	2.35	72000
30	Color	Peter Berg	377	131	532	627	Alexander I	14000	65173160	Action|Adv	Liam Neeso	Battleship	202382	26679	Tadanobu	0	box office|	http://www	751	English	USA	PG-13	2.09E+08	2012	10000	5.9	2.35	44000
31	Color	Colin Treve	644	124	365	1000	Judy Greer	3000	6.52E+08	Action|Adv	Bryce Dalla	Jurassic Wo	418714	8458	Omar Sy	0	dinosaur|c	http://www	1290	English	USA	PG-13	1.5E+08	2015	2000	7	2	150000
32	Color	Sam Mend	750	143	0	393	Helen McC	883	3.04E+08	Action|Adv	Albert Finn	SkyfallÂ	522030	2039	Rory Kinne	0	brawl|chil	http://www	1498	English	UK	PG-13	2E+08	2012	563	7.8	2.35	80000
33	Color	Sam Raimi	300	135	0	4000	James Fran	24000	3.73E+08	Action|Adv	J.K. Simmo	Spider-Mar	411164	43388	Kirsten Dur	1	death|doc	http://www	1303	English	USA	PG-13	2E+08	2004	11000	7.3	2.35	0
34	Color	Shane Blac	608	195	1000	3000	Jon Favreau	21000	4.09E+08	Action|Adv	Robert Dov	Iron Man 3	557489	30426	Don Cheadl	3	armor|exp	http://www	1187	English	USA	PG-13	2E+08	2013	4000	7.2	2.35	95000
35	Color	Tim Burton	451	108	13000	11000	Alan Rickm	40000	3.34E+08	Adventure	Johnny De	Alice in Wo	306320	79957	Anne Hathe	0	alien in wo	http://www	736	English	USA	PG	2E+08	2010	25000	6.5	1.85	24000
36	Color	Brett Ratne	334	104	420	560	Kelsey Gra	20000	2.3E+08	Action|Adv	Ash Hugh Ja	X-Men: Thi	388427	21716	Daniel Cud	0	battle|mut	http://www	1912	English	Canada	PG-13	2.1E+08	2006	808	6.8	2.35	0
37	Color	Dan Scanlo	376	104	37	780	Tyler Labin	12000	2.68E+08	Adventure	Steve Busc	Monsters U	235025	14863	Sean Hayes	0	cheating|fi	http://www	765	English	USA	G	2E+08	2013	779	7.3	1.85	44000
38	Color	Michael Ba	366	150	0	464	Kevin Dunr	894	4.02E+08	Action|Adv	Glenn Mor	Transforme	323207	3218	Ramon Roc	0	autobot|di	http://www	1439	English	USA	PG-13	2E+08	2009	581	6	2.35	0
39	Color	Michael Ba	378	165	0	808	Sophia Myl	974	2.45E+08	Action|Adv	Bingbing Li	Transforme	242470	3988	Kelsey Gra	2	blockbuste	http://www	918	English	USA	PG-13	2.1E+08	2014	956	5.7	2.35	56000
40	Color	Sam Raimi	525	130	0	11000	Mila Kunis	44000	2.35E+08	Adventure	Tim Holme	Oz the Gre	175409	73441	James Fran	4	circus|mag	http://www	511	English	USA	PG	2.15E+08	2013	15000	6.4	2.35	60000
41	Color	Marc Webb	495	142	484	825	Andrew Ga	15000	2.03E+08	Action|Adv	Emma Stor	The Amazin	321277	28631	B.J. Novak	0	costumed|	http://www	1067	English	USA	PG-13	2E+08	2014	10000	6.7	2.35	41000

Fig 4.2: A screenshot of the final dataset

This proved to be a very difficult task. Matching individual dataset’s values with the others and making sure no redundancy exists and that all values are in their right place and order was found to be a laborious task.

Once the dataset has been created, the project progresses into its next phase.

4.2 Data Cleaning

Data cleansing or data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data.

After cleansing, a data set should be consistent with other similar data sets in the system. The inconsistencies detected or removed may have been originally caused by user entry errors, by corruption in transmission or storage, or by different data dictionary definitions of similar entities in different stores.

Missing data is one of the most error causing scenario as the algorithm needs values to be present in all points at all times.

Color	Gore Verbi	302	169	563	1000
Color	Sam Mend	602	148	0	161
Color	Christophe	813	164	22000	23000
				131	
Color	Andrew Sta	462	132	475	530
Color	Sam Raimi	392	156	0	4000
Color	Nathan Gre	324	100	15	284
Color		635	141	0	19000
Color	David Yate:	375	153	282	10000
Color	Zack Snyde	673	183	0	2000
Color	Bryan Singe	434	169		903
Color	Marc Forst	403	106	395	393
Color	Gore Verbi	313		563	1000
Color	Gore Verbi	450	150	563	1000
Color	Zack Snyde	733	143	0	748

Fig 4.3: Missing data in the dataset

There are 3 common ways of cleaning the data, which are as follows.

1. Mean Value
2. Removal
3. Default Value

4.2.1 Mean Value

This is a technique where the values for any missing data blocks are filled with the mean for that column or that attribute in the dataset.

The mean is the average of the numbers. It is easy to calculate: add up all the numbers, then divide by how many numbers there are. In other words, it is the sum divided by the count.

This will fill in every missing block with a single value and will cause a slight median in the dataset, thus it must not be relied upon too often.

4.2.2 Removal

The second type of data cleaning is to remove those entries in the dataset which lack some values entirely. Although this will reduce the size of the dataset, it will eliminate any error causing entries and will guarantee an error free process.

Large datasets can easily handle removal as one entry will not make a difference as compared to a million. But smaller datasets who have around 100 to 1000 values will be heavily affected by such changes.

There are also chances that any vital value or an exceptional entry would get removed, thus caution should be exercised when practicing the removal method.

4.2.3 Default Value

The last form of data cleaning is called default value, where a single preset default value will be used to fill in every missing or empty values in the entire dataset.

This provides a very quick and easy way of cleaning the dataset, but this makes it more robotic and it tends to reduce the dataset's variations but putting in similar values for all cases. Thus, this is the least preferred method for cleaning of all the three methods mentioned above.

For the purpose of this project, we will be using only the mean value method to clean our datasets.

4.3 Test, Train and Holdout Data

Training: In order to find patterns in a dataset from which it can make predictions, an algorithm must first learn from a historical example – typically from a historical dataset that contains the output variable we want to predict. Basically, it is the data we use to build the model.

Testing: Once an algorithm is trained on one subset of historical data, we need to make sure the patterns it "learns" are relevant. This second set is what we call the validation set. We use validation to rank the accuracy of algorithms to find the most accurate one and for making decisions about which algorithms are useful.

Holdout: Sometimes referred to as "testing" data, the holdout is also used to determine the algorithm's predictive prowess and provides a final estimate of model performance after it has been trained and validated. Basically, it's the section of the original data that we keep separate from the training and validation data to provide a final check to make sure that the algorithm that we chose works on data that was not used in the choice of that algorithm, which can happen when you evaluate multiple models on validation data alone. Holdouts should never be used to make decisions about which algorithms to use or for improving or tuning algorithms.

4.4 Training the Model

To train an ML model, you need to specify the following:

- Input training data source
- Name of the data attribute that contains the target to be predicted
- Required data transformation instructions

The process of training an ML model involves providing an ML algorithm (that is, the learning algorithm) with training data to learn from. The term ML model refers to the model artefact that is created by the training process.

The training data must contain the correct answer, which is known as a target or target attribute. The learning algorithm finds patterns in the training data that map the input data attributes to the target (the answer that you want to predict), and it outputs an ML model that captures these patterns.

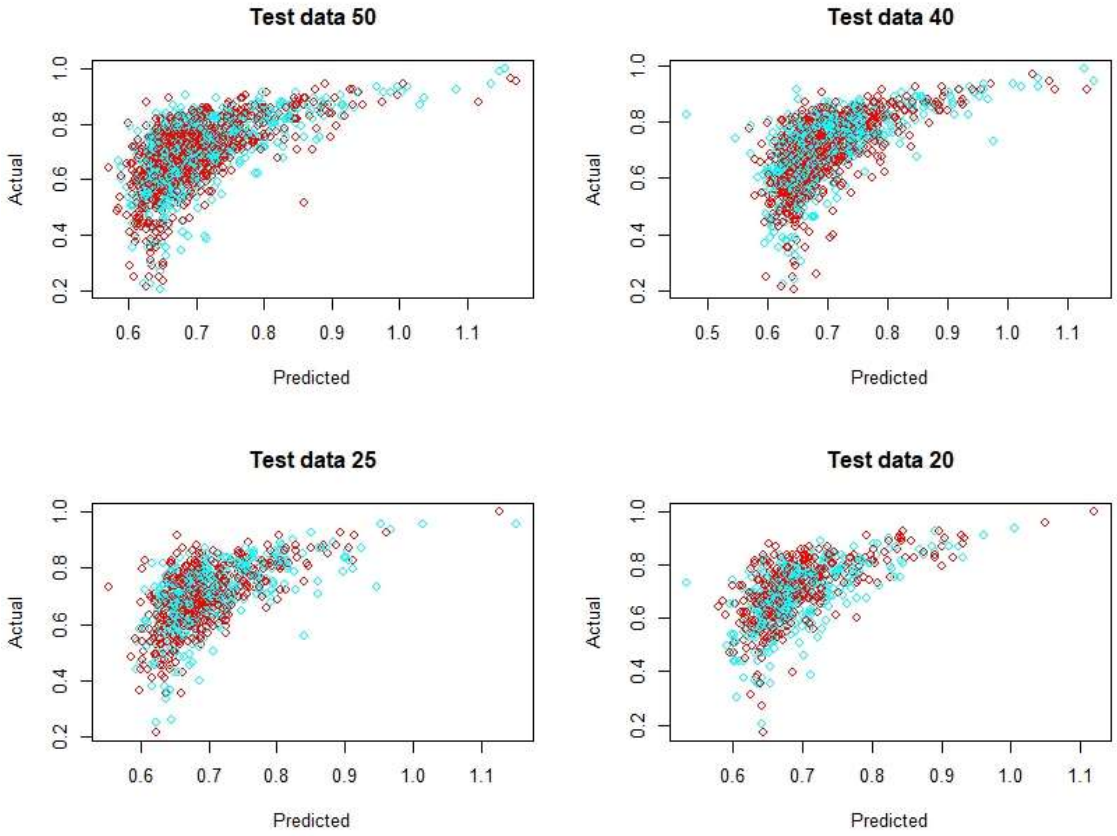


Fig 4.4: Different test and training cases

We can use the ML model to get predictions on new data for which you do not know the target. For example, let's say that you want to train an ML model to predict if an email is spam or not spam. We would provide ML with training data that contains emails for which you know the target (that is, a label that tells whether an email is spam or not spam).

ML would train an ML model by using this data, resulting in a model that attempts to predict whether new email will be spam or not spam.

4.5 Testing and Evaluating the Model

One should always evaluate a model to determine if it will do a good job of predicting the target on new and future data. Because future instances have unknown target values, you need to check the accuracy metric of the ML model on data for which you already know the target answer and use this assessment as a proxy for predictive accuracy on future data.

To properly evaluate a model, you hold out a sample of data that has been labeled with the target (ground truth) from the training data source. Evaluating the predictive accuracy of an ML model with the same data that was used for training is not useful, because it rewards models that can "remember" the training data, as opposed to generalizing from it. Once you have finished training the ML model, you send the model the held-out observations for which you know the target values. You then compare the predictions returned by the ML model against the known target value. Finally, you compute a summary metric that tells you how well the predicted and true values match.

4.6 Deploy

As all the development steps of predictive analysis is completed at this point, the software shall be open for client usage. This is can be done in various ways. Machine learning models and its technical stack is versatile enough to be hosted and be accessible through any platform. These models can be hosted on multiple platforms like Cloud Infrastructure, PaaS, SaaS, IaaS, Server based website/app platforms and many more.

Chapter 5

Working Methodology

CHAPTER 5

WORKING METHODOLOGY

Predictive TTM Optimization proposes a linear model with no recursions or loops. It follows a single path in its predictions as shown below.

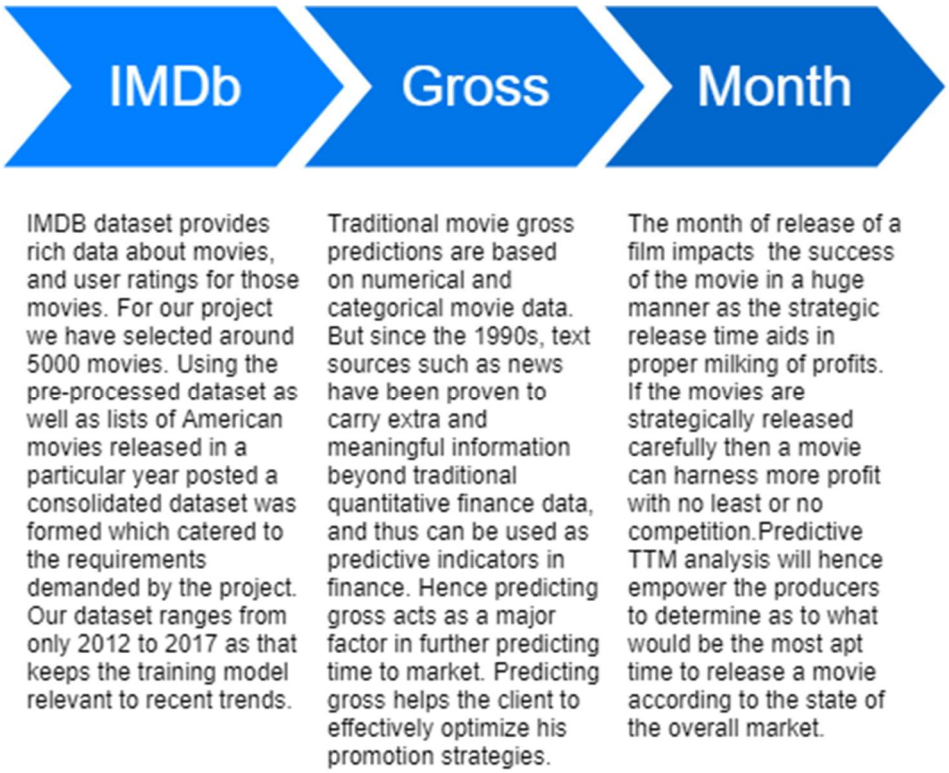


Fig 5.1: Working Methodology

5.1 Predicting IMDb Score

Our first step is to predict the IMDb Score, which is like a universally accepted standard rating system for all sorts of movies provided by IMDb Inc. This is our basis for the entire prediction and one of the most important values.

To predict the IMDb Score, we have used Linear Regression algorithm, which gave us the best-case results after a lot of trial and error and testing with different algorithms.

We use the dataset which includes values such as Duration and Budget, and we use Linear Regression to train the model.

5.2 Predicting Gross of the Movie

The newly predicted IMDb score will be entered into the dataset again, and this time the dependent variable will be the gross of the movie.

The dataset used for this has values such as Budget, Duration, the newly predicted IMDb Score.

The machine learning algorithm used for predicting the gross is random forest algorithm.

5.3 Predicting the month for optimal gross

The algorithm used to predict the release date is linear regression. We use the predicted IMDb Score and the predicted Gross along with the other factors in the dataset to predict the month of launch which would result.

This month of launch is based on the hit movies and those movies who are believed to have made optimum gross and their launch dates.

Chapter 6

Algorithms Tested and Used

CHAPTER 6

ALGORITHMS TESTED AND USED

Predictive TTM Optimization requires a high level of accuracy to be able to work in the real-world scenario, we cannot simply assume and use any one algorithm. We must make sure that we pick the right algorithm that is perfectly suitable for our scenarios.

To ensure that the right algorithm is picked, we have tried and tested various algorithms.

6.1 Algorithms Tested

Before deciding on which algorithm to use, we implemented our models with the following algorithms and decided to not use them for the reasons mentioned below.

1. Polynomial Regression Algorithm
2. Decision Tree
3. Artificial Neural Networks

6.1.1 Polynomial Regression Algorithm

Polynomial regression is a form of regression analysis in which the relationship between the independent variable x and the dependent variable y is modelled as a nth degree polynomial in x. Polynomial regression fits a nonlinear relationship between the value of x and the corresponding conditional mean of y, denoted E(y|x), and has been used to describe nonlinear phenomena such as the growth rate of tissues. Polynomial regression extends the linear model by adding extra predictors, obtained by raising each of the original predictors to a power.

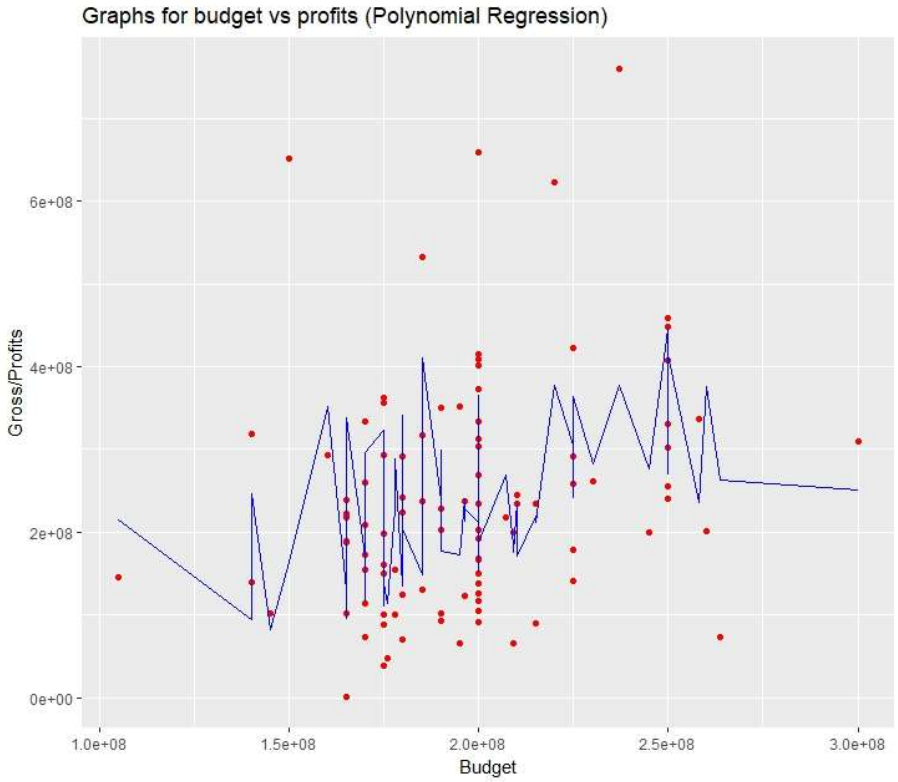


Fig 6.1: Polynomial Regression

Polynomial regression was one of the first algorithms tested apart from Linear Regression, and it was purely rejected for the fact that for the conditions and scenarios of our dataset, Linear Regression simply provided a higher accuracy range as compared to Polynomial Regression. Linear regression gave an accuracy range of, whereas Polynomial Regression was lacking by about 7% to 8% as compared to Linear Regression.

6.1.2 Decision Tree

A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

It is a type of supervised learning algorithm (with a predefined target variable) that is mostly used in classification problems and works for both categorical and continuous input and output variables. It is one of the most widely used and practical methods for

inductive inference. (Inductive inference is the process of reaching a general conclusion from specific examples.)

Although these are highly effective in cases where a lot of binary values are present in the dataset, such as a Boolean true/false, a binary 1/0 etc., this doesn't make much sense for a dataset with widely varying values such as the IMDb scores ranging from 0 to 10, budget and gross moving from 5,00,00,000 to 10,00,00,000.

A decision tree cannot handle such varying values and ended up getting very low accuracy rates in the range of 50% to 55%.

6.1.3 Artificial Neural Networks

Artificial neural networks (ANNs) or connectionist systems are computing systems vaguely inspired by the biological neural networks that constitute animal brains. Such systems "learn" (i.e. progressively improve performance on) tasks by considering examples, generally without task-specific programming.

We were successfully able to predict the IMDb scores for a given dataset and a given movie from the test set with a very high accuracy value.

Unfortunately, the hardware and resources available from our end were not enough to process through ANN effectively. It took us about 3 hours to get one value, and we realized that this is not efficient despite the accuracy.

ANN requires very high specifications and high-quality servers and is not viable to perform this on the current systems available to us. Thus, we had to give up on ANN and look for other alternative solutions.

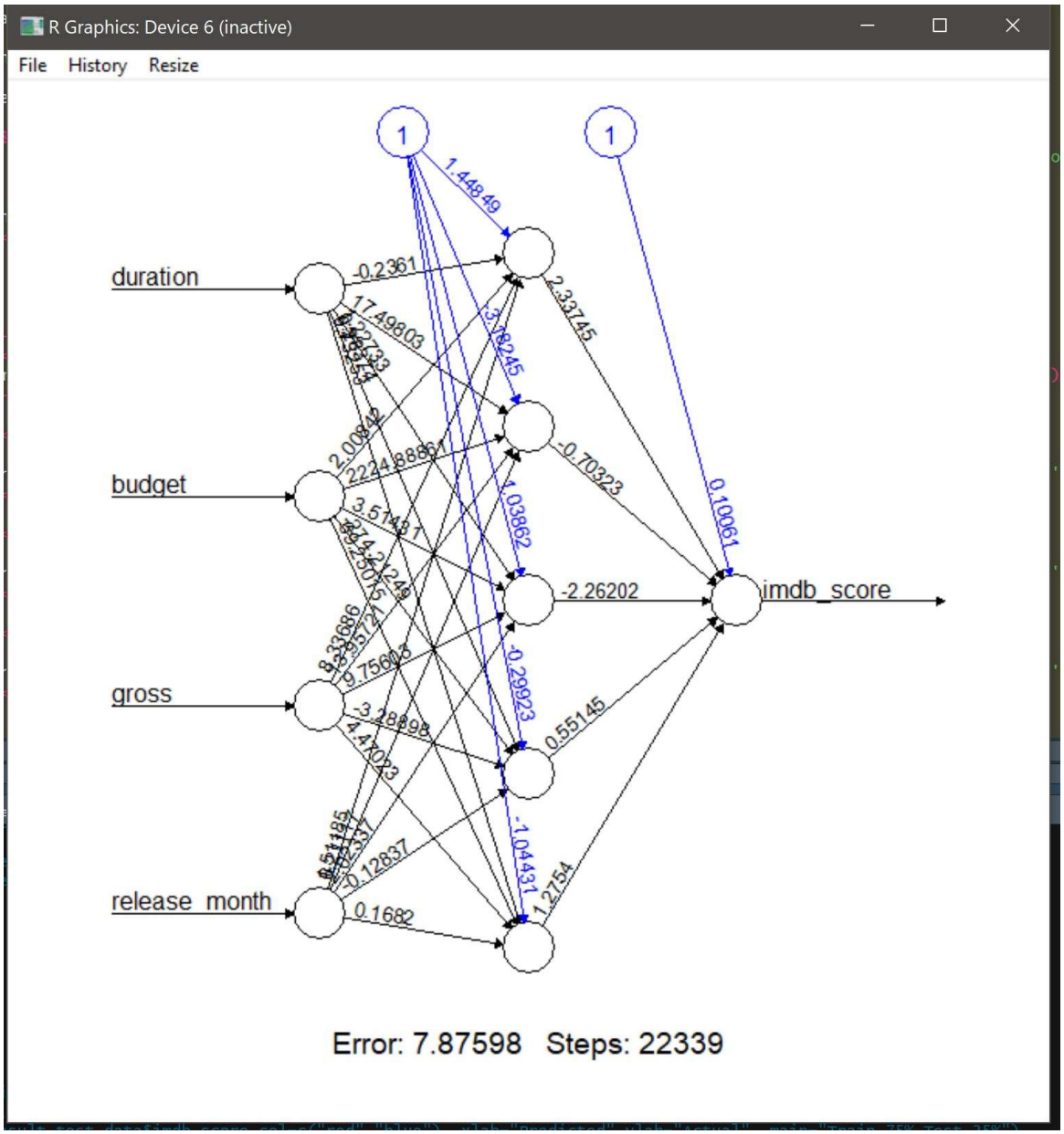


Fig 6.2: Artificial Neural Network for prediction IMDb Score

6.2 Algorithms used

We have used Linear Regression for predicting IMDb scores and the month of launch which will result in optimal gross and have used Random Forest Algorithm to find the gross.

6.2.1 Linear Regression

Linear regression is a very simple approach for supervised learning. Though it may seem somewhat dull compared to some of the more modern algorithms, linear regression is still a useful and widely used statistical learning method. Linear regression is used to predict a quantitative response Y from the predictor variable X. Linear Regression is made with an assumption that there's a linear relationship between X and Y.

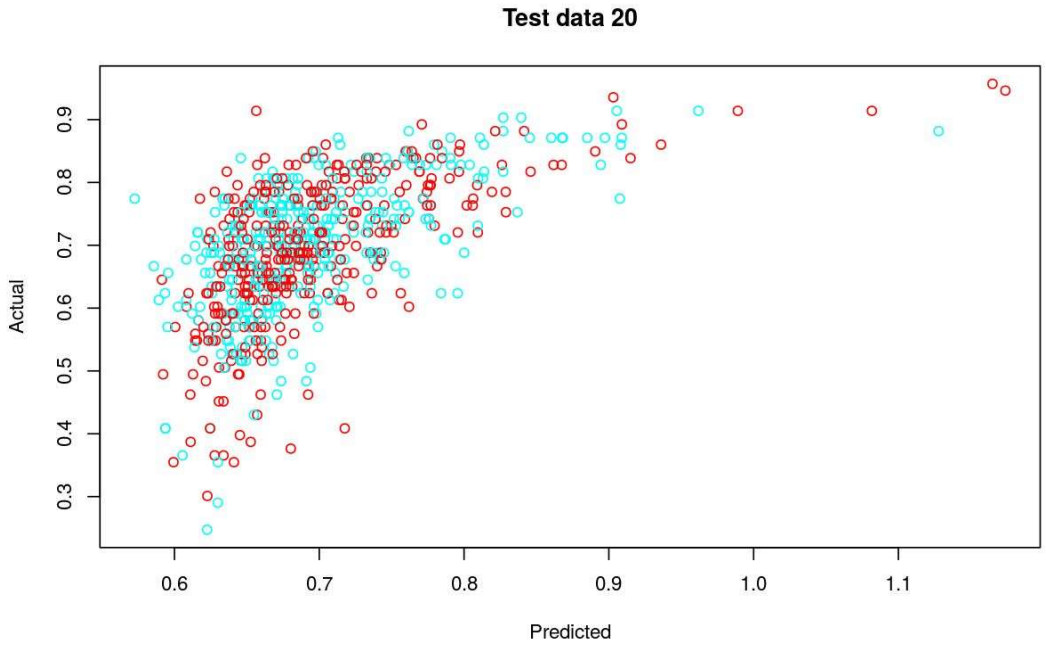


Fig 6.3: Linear Regression Training Set

This algorithm has proven to give results in a range of 72.04% to 98.4%.

6.2.2 Random Forest Algorithm

Random forest algorithm was used to predict the gross. The random forest algorithms have proven to give a positive response with an accuracy of 74% to 92%.

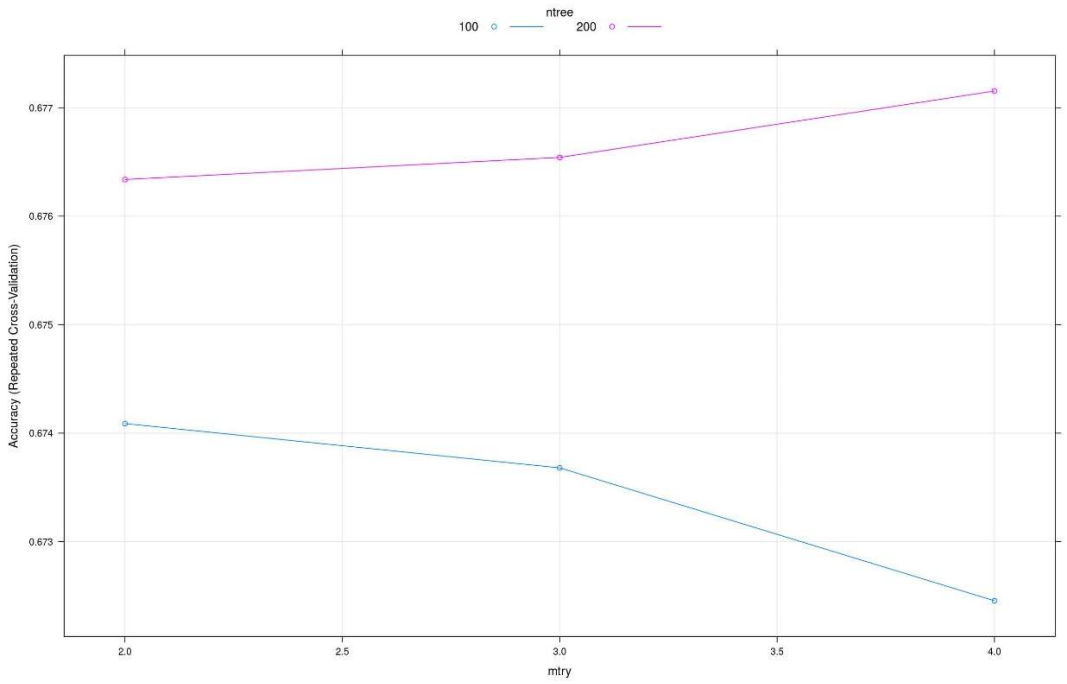


Fig 6.4: Random Forest Visualization

Random Forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because it's simplicity and the fact that it can be used for both classification and regression tasks. One big advantage of random forest is, that it can be used for both classification and regression problems, which form the majority of current machine learning systems.

Chapter 7

Visualization

CHAPTER 7

VISUALIZATION

Once the machine learning models are trained and tested, the necessity to package them into a working application is the next natural task. Just the code to be run line by line without a user interface is neither satisfactory nor the way to present an application. Hence a cloud-based platform comes into the foray to make this R code integrate with the front-end user interface. The platform used for the same is called shinnyapps.io and as this is a platform provided by the RStudio itself, it becomes astonishingly easy to deploy the R code on it and to monitor its traffic usage and also resource allocation of the same.

7.1 Shinnyapps.io

Shinyapps.io is a self-service platform that makes it easy for you to share your shiny applications on the web in just a few minutes. Many customers use shinyapps.io to prove out some concepts, build out a prototype, or just run it for a short period of time for their own purposes, while others are using it as a core component of their analytical offerings within a larger online property.

The service runs in the cloud on shared servers that are operated by R Studio. Each application is self-contained and operates on either data that is uploaded with the application, or data that the code pulls from third-party data stores, such as databases or web services.

7.2 Output Obtained

With the help of the aforementioned shiny, we have been able to successfully generate all sorts of demographics, which contain a lot of insights about the movies to help the producer to have a better idea about the competition and the market trends.

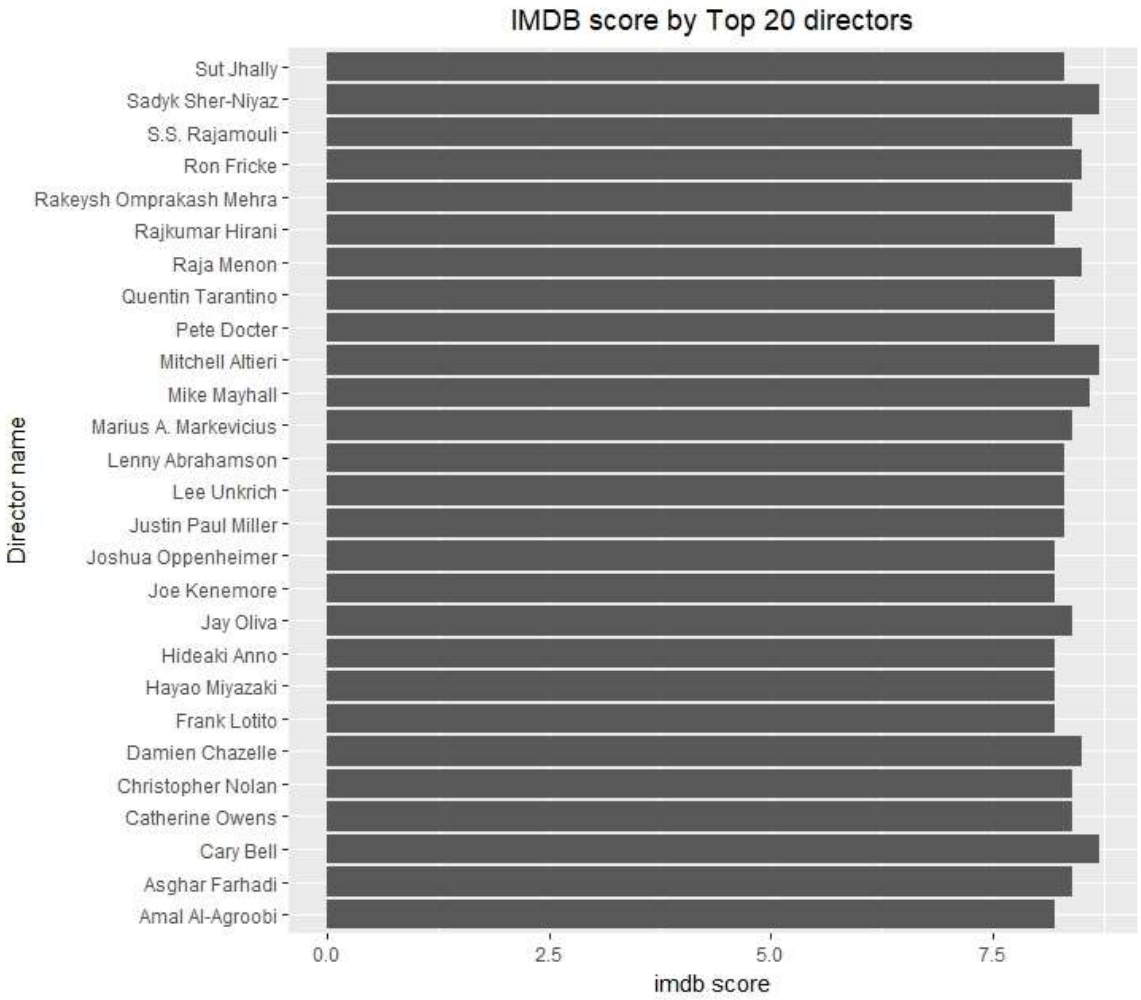


Fig 7.1: Directors vs their average IMDb ratings

This graph helps visualize the trends between the directors and the IMDb scores that their movies receive on an average. This helps us factor in the director factor when deciding on a release month.

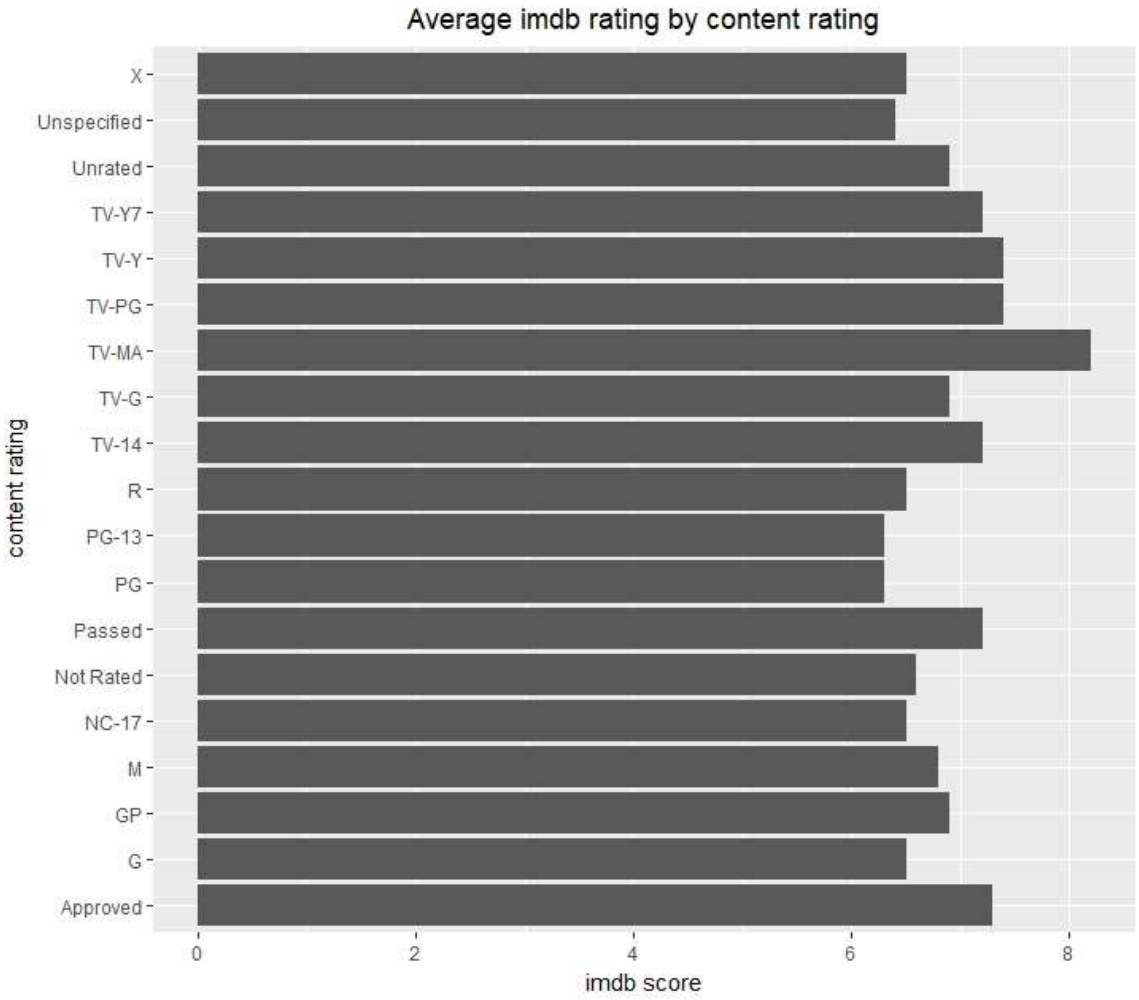


Fig 7.2: Average IMDb rating by content rating

This is one more factor in deciding the status of the movie. The content rating decides the kind of audience that a particular movie might receive and decide how it is welcomed in the society.

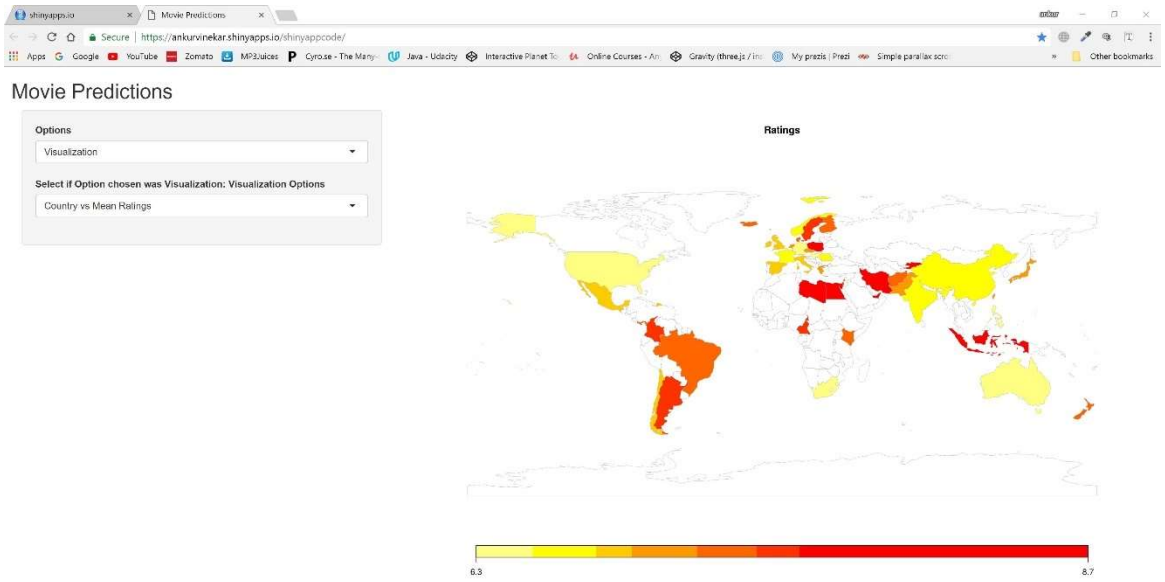


Fig 7.3: Country vs IMDb Score

The release area of a movie heavily depends on where it is launched. Launching a Hindi movie in the United States of America might not be as well received as an English movie over there, as many people might not watch it due to the language barriers.

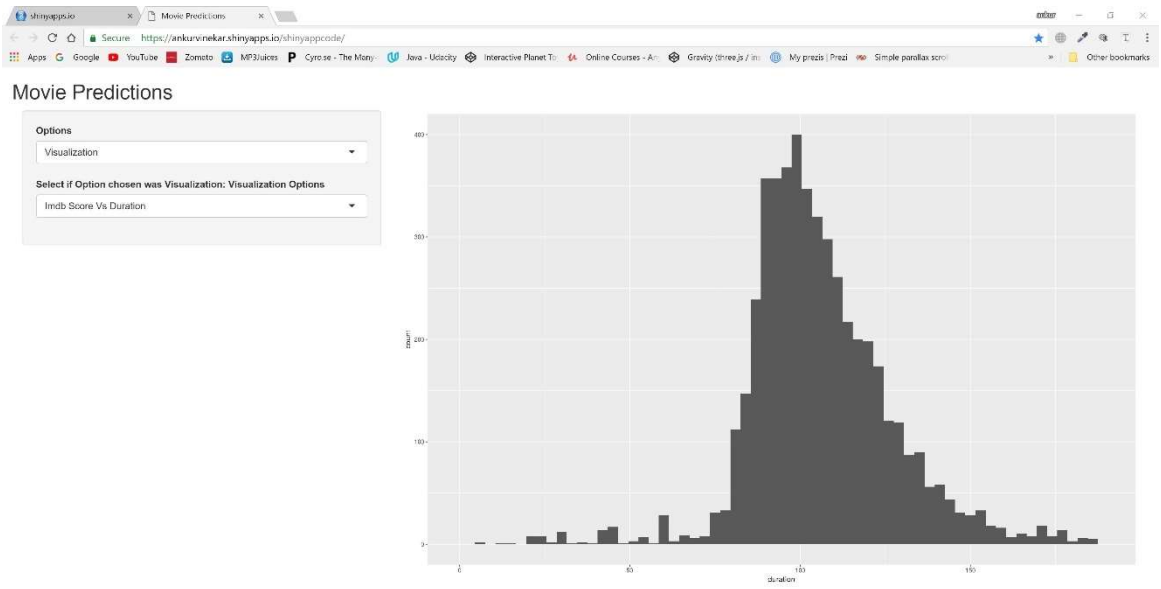


Fig 7.4: IMDb Score vs Movie Duration

The duration of the movie is a highly important aspect as it not only tests the availability of the fans, it also tests their patience levels. This might be a major factor in the success of a movie as a 4-hour movie is highly unlikely to be a hit.

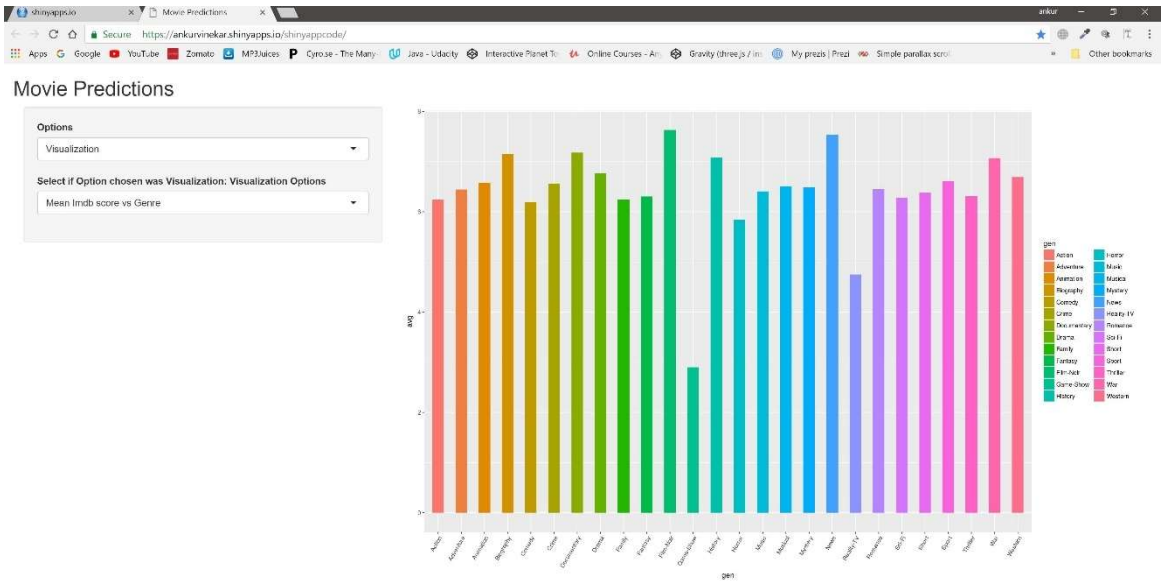


Fig 7.5: IMDb Score vs Genre

A genre is what defines the taste of a person as they watch a movie based on their likes and dislikes. This will highly influence your target audience and your specific targets will improve if you're trying to get a particular attraction but might be a risky move to launch a specific genre movie if you are expecting a general turnout.

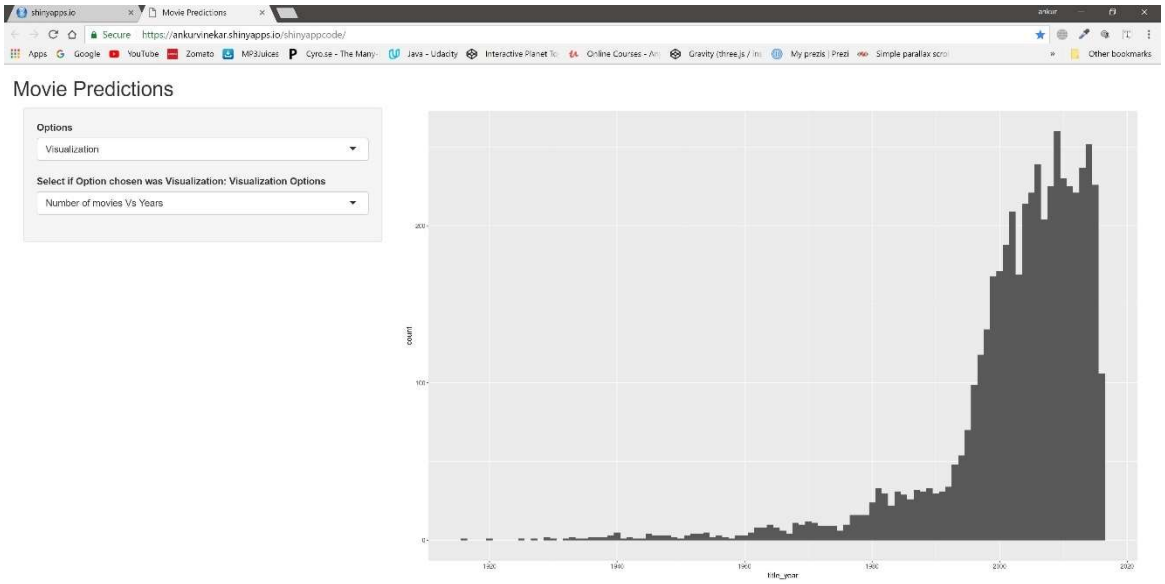


Fig 7.6: No. of movies vs Years

The number of movies launched per year has slowly been increasing and has seen a tremendous increase recently, signifying the boom in the film industry. And when we compare this information with the IMDb score information, we find that despite the rise in movies per year, the popularity only keeps increasing and the quality of the movies aren't degrading either.

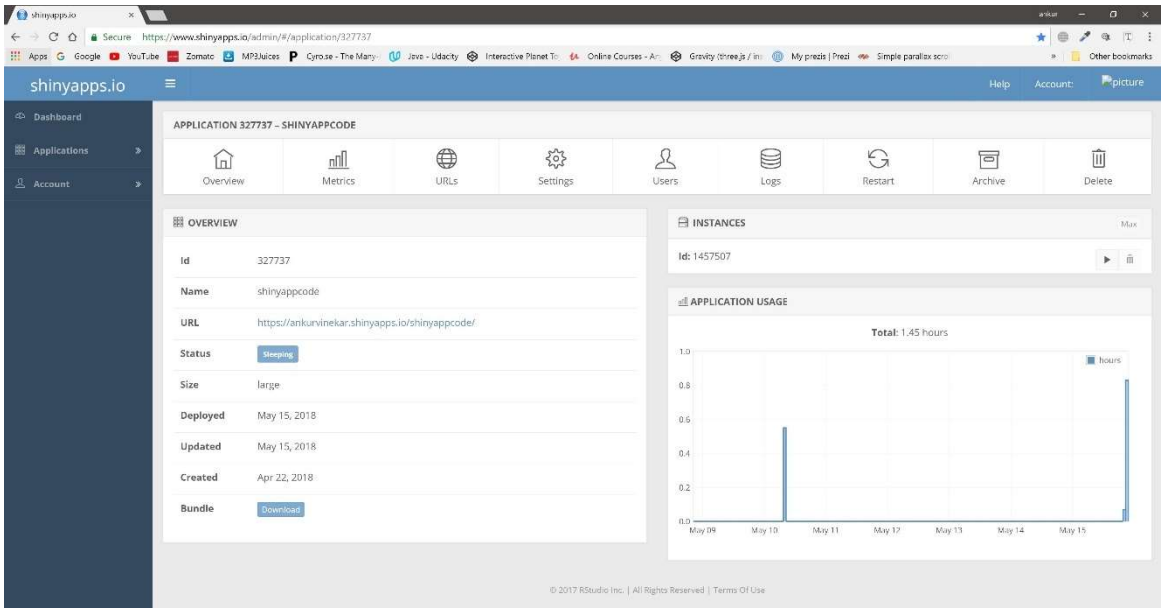


Fig 7.7: The Shiny app developer's console

This is the interface that we see when the developers are trying to visualize their data. The interface is provided by R studio's very own Shiny.io service. This service is a great addition to all the new and experienced developers alike, and this is a great way of displaying the results too.

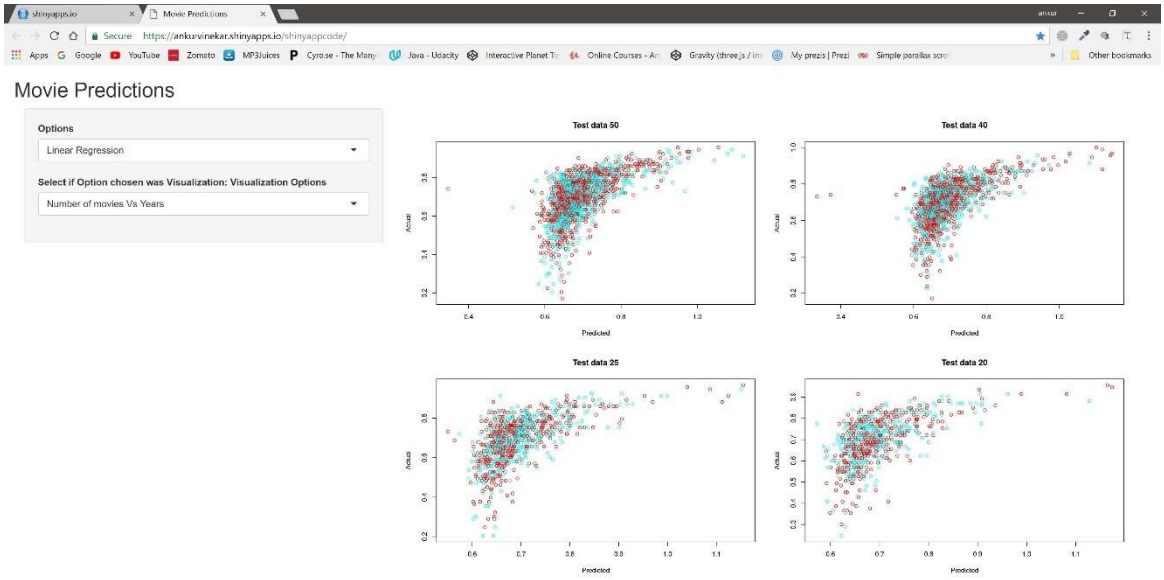


Fig 7.8: Test Scenarios for the Linear Regression as seen on shiny

This is a representation of the shiny interface and how it helps in displaying the graphs and the details provided by the dataset.

Chapter 8

Future Scope

CHAPTER 8

FUTURE SCOPE

This project has a lot of potential that can be delved into further if given the time and dedication. A few of the things that can be done in the future with this project as its base are

1. **Including textual factors:** Currently, only numerical factors are being considered as text cannot be directly comprehended by the ML Algorithms, ad they need special steps to be performed to get them working.
2. **Including more factors:** There is always room for more, and it is never too much. More factors mean more views and more room for consideration. This will lead to a higher accuracy in the future.
3. **Testing out advanced technologies:** This will allow us to use higher and more complex algorithms. Better hardware would mean that we could finally complete the ANN algorithm that we could not because of hardware limitations.
4. **Extended to other fields:** This project can not only apply to movies but can extend to other fields also such as art, music etc.

Chapter 9

Conclusion

CHAPTER 9

CONCLUSION

We have successfully been able to predict IMDb Scores and the Gross of the movie, eventually leading up to the prediction of the optimal launch period which should result in the highest gross based on the trained data from the past trends and scenarios.

Predictive TTM analysis will go a long way in helping the film industry and has a great potential of becoming a heavily researched field as well as a highly demanded career option as it bridges two of the most advanced fields in todays world: Computer science and the film industry.

By using this model, and movies make the most out of their money, the higher profit and the cash flowing through the film industry will only lead to the invention of newer technologies as demand grows.

References

REFERENCES

[1]. Darin Im, Minh Thao, Dang Nguyen, Predicting Box Office of movies in the U.S Market, CS 229, Fall 2011

[2]. W. Zhang and S. Skiena, Improving movie gross prediction through news analysis, IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Milan, 2009

[3]. Aju Thalappillil Scaria, Rose Marie Philip, Sagar V Mehta, Predicting Star Rating of Movie Review Comments, 21st Pacific Asis Conference on Language, 2007

[4]. Deniz Demir, Olga Kapralova, Hongze Lai, Predicting IMDb Movie Ratings Using Google Trends, Dept.Elect.Eng,Stanford Univ., California, December, 2012.

[5]. Wikipedia - https://en.wikipedia.org/wiki/R_(programming_language)

[6]. Wikipedia - https://en.wikipedia.org/wiki/Data_cleansing

[7]. Medium - https://medium.com/simple-ai/linear-regression-intro-to-machine-learning-6-6e320dbdaf06