CINEMA TICKET SALES FORECASTING

Team 1

Aakash A Aundhkar(G01371754)

Sanika Suhas Dalvi(G01398375),

Keerthana Cheruvu(G01160202)

George Mason University

Dr. Duoduo Liao

AIT-614-001

October 24, 2023

Aakash A Aundhkar, Sanika Suhas Dalvi, Keerthana Cheruvu

INDEX

1. INTRODUCTION:

The entertainment industry, and specifically the cinema sector, faces the task of streamlining its operations, increasing revenue, and improving the whole cinematic experience in an era of quickly growing technology and ever-changing consumer demands. Strong reliance on data-driven decision-making, intense competitiveness, and high operating costs are all characteristics of the film industry. For cinema chains, accurate sales forecasting is an essential component of strategic planning since it enables them to spend resources wisely, optimize their pricing schemes, and offer outstanding experiences to customers. With this initiative, we hope to accurately estimate ticket sales for theatre operators by leveraging data analytics, machine learning, and predictive modelling.

To be able to create predictive models, our research will use information on ticket sales. These models will not only predict ticket sales but also offer useful information for enhancing marketing initiatives and managerial choices, ultimately resulting in more earnings and happier customers.

The project will help in reducing risk by using a data-driven strategy. Additionally, the project will serve as a first step towards realizing the potential of cutting-edge technologies like artificial intelligence and machine learning to completely transform the movie industry.

2. RELATED WORK:

In 2017, Javaria and Ahmad predicted Movie Success by using various data mining. They did an analysis to find which are the top factors which affect the movie's success and found that genres, movie actors and ratings of the movie contribute the most towards the sales of the

tickets. [1] Our Project is going to forecast the ticket sales for a given period. Hence by taking this analysis in consideration, we will be able to select which attributes are important.

In 2015, Michael and Kang introduced a decision support system designed for guiding investment decisions in the early stages of movie production. The primary objective of this system is to forecast a movie's potential success based on its profitability, utilizing historical data from diverse sources. Employing social network analysis and text mining techniques, the system autonomously extracts multiple feature sets, encompassing information on the cast, the movie's plot, its release date, and unique hybrid features that combine cast with plot and release date with plot. The paper's experiments, conducted over an 11-year period with various films, demonstrate the system's noteworthy capability in predicting a movie's success. Furthermore, the results of these experiments also indicate that different feature sets, including newly proposed ones, all contribute significantly to the prediction. [2] Our project is different from this project as this project takes data consideration of 11 years whereas our project will use recent data which will make the model more accurate.

In 2020, Pawet and Karol introduced the paper which focuses on short-term forecasting of cinema attendance, a topic with limited prior research compared to aggregate movie performance modeling. The research employs data at the individual show level, encompassing 179,103 shows, as opposed to aggregate box office sales. Multiple linear regression models are applied to generate one-week ahead attendance forecasts, and the models are ranked based on out-of-sample fit. The findings indicate that the most effective models incorporate cinema- and region-specific variables in addition to movie parameters and title popularity. In conclusion, the study suggests that regression models using a comprehensive set of variables

can successfully predict attendance at individual cinema shows. [3] Our project will use modern and more efficient forecasting algorithms such as Facebook Prophet and ARIMA.

3. OBJECTIVES:

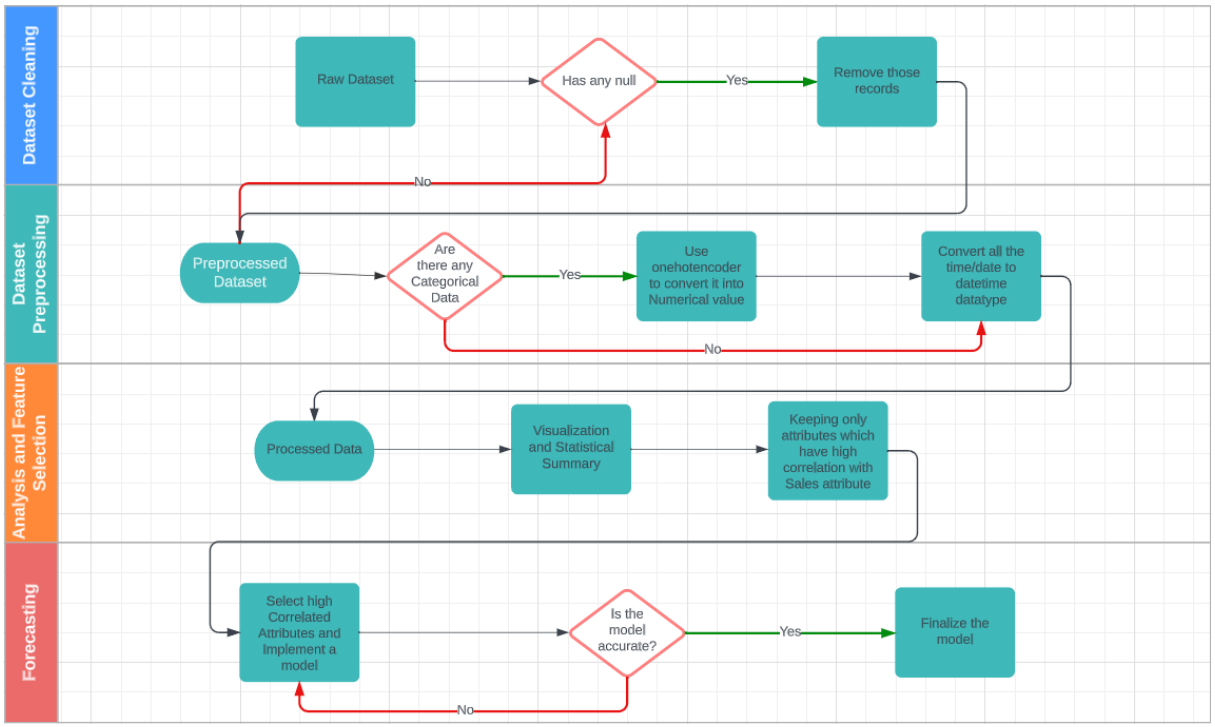
The project's main goal is to create and execute a cutting-edge data-driven solution that precisely forecast movie ticket sales. Below are few more objectives of our project.

- Evaluating potential issues and uncertainties in ticket sales to enable proactive decision-making and market-dynamics adaption.
- Leveraging data analytics and machine learning to beat competitors and satisfying the changing needs of the entertainment industry to stay ahead in a highly competitive sector.
- Determining probable difficulties and uncertainties in ticket sales to enable proactive decision-making and market-dynamics adaption.

4. SUMMARY:

In summary, we seek to improve resource allocation and enable theatre owners to manage personnel, concessions, and screening schedules. This ultimately results in lower costs and more effective operations. The project aims to improve revenue generation by introducing dynamic pricing methods and customized promotional campaigns based on sales estimates. The project also addresses the need to reduce industry-specific risks and uncertainties, enabling proactive decision-making in the face of shifting market dynamics. Additionally, by offering data-driven insights for investment and expansion decisions, it aids in strategic planning. In the end, our project makes sure theatre operators can compete in a dynamic entertainment environment.

5. METHODOLOGY:



6. PROPOSED SELECTED DATASET:

We selected a dataset from Kaggle which consists of 14 columns and 142524 records. By using this dataset, we are going to forecast the ticket sales which indeed help the cinema places to assign screening frequency for maximizing the profits.

Column Name	Data Type	Description
film_code	int64	A numerical code representing a film.
cinema_code	int64	A numerical code representing a cinema.
total_sales	int64	The total sales revenue generated.
tickets_sold	int64	The total number of tickets sold.
tickets_out	int64	The total number of tickets distributed.

show_time	int64	The duration of the show in minutes.
occu_perc	float64	The occupancy percentage for the show.
ticket_price	float64	The price of a single ticket.
ticket_use	int64	The number of tickets used by customers.
capacity	float64	The maximum seating capacity of the cinema.
date	object	The date of the cinema show.
month	int64	The month when the show occurred.
quarter	int64	The quarter in which the show took place.
day	int64	The day of the month for the show date.

According to our initial analysis, we have decided to move forward with ticket_price, occu_perc, show_time, tickets_sold, ticket_use and capacity columns as we think they are the most important attributes that will contribute towards the model accuracy.

7. PROPOSED DEVELOPMENT PLATFORMS:

7.1 Databricks dbfs:

The primary component of the Databricks Unified Analytics Platform, Databricks DBFS (Databricks File System), is a distributed and scalable file system created to simplify data management and access for big data and machine learning applications. Data engineers and data scientists working within the Databricks environment will have a seamless and effective experience because of DBFS's unification of structured and unstructured data. It

offers close connection with Databricks Workspace, distributed storage, high availability, data replication across clusters and availability zones, mounting of external data sources, security features, and versioning capabilities. As a result, DBFS is a key element for streamlining data operations, facilitating group data analysis, and guaranteeing the dependability and security of data assets.

7.2 PySpark:

An open-source, quick, and flexible framework for distributed data processing, PySpark is a component of the Apache Spark ecosystem. Built on top of the Spark core, PySpark enables Python developers and data scientists to analyze massive volumes of data and carry out a variety of data manipulation and analysis activities. PySpark is a crucial tool for big data processing and analytics owing to its user-friendly API and integration with well-known Python libraries, which give data professionals a seamless interface for working with structured and unstructured data, performing complex data transformations, running machine learning and data analytics tasks, and achieving scalability and performance improvements on distributed clusters.

7.3 Spark MLlib:

Apache Spark ecosystem includes the robust machine learning package MLlib, which provides a variety of tools and methods to make scaled and distributed machine learning operations easier. A popular option for big data analytics, MLlib enables users to carry out a variety of tasks at scale, including classification, regression, clustering, recommendation, and more. Data scientists and engineers can efficiently build and deploy machine learning models with the help of MLlib's user-friendly API and support for multiple programming languages. They can also take advantage of Spark's distributed computing capabilities to

process and analyse large datasets with ease. It is a flexible and essential part of machine learning in contexts with distributed computing.

7.4 Databricks:

A cloud-based unified analytics platform called Databricks enables businesses to fully utilize their data. Built on the well-known Apache Spark framework, Databricks gives data engineers, data scientists, and analysts a collaborative and integrated environment for processing, analysing, and visualizing huge and complicated information. It combines procedures for data engineering and data science into a single platform, allowing teams to work more productively and gain insightful knowledge from their data. Databricks provides features including Databricks Delta for data management, Databricks Runtime for optimal data processing, and Databricks Workspace for collaborative development. Databricks has emerged as a crucial tool for businesses looking to accelerate their adoption of the cloud owing to its scalability, usability, and extensive interaction with well-known cloud providers like AWS, Azure, and Google Cloud.

8. TIME PLAN OF PROJECT:

Project Timeline	Date to be Completed	Team Member
Project Proposal	24-Oct-23	Aakash, Keerthana, Sanika
Start Coding	28-Oct-23	Aakash
Determine Analytical Method	30-Oct-23	Sanika
Code Review	10-Nov-23	Keerthana
Implement Analytical Method	11-Nov-23	Sanika
Perform Analysis	13-Nov-23	Keerthana
Review Result	15-Nov-23	Aakash

Slide Creation	20-Nov-23	Aakash
FInal Report Submission	25-Nov-23	Aakash, Keerthana, Sanika
Final Project Submission	25-Nov-23	Aakash, Keerthana, Sanika

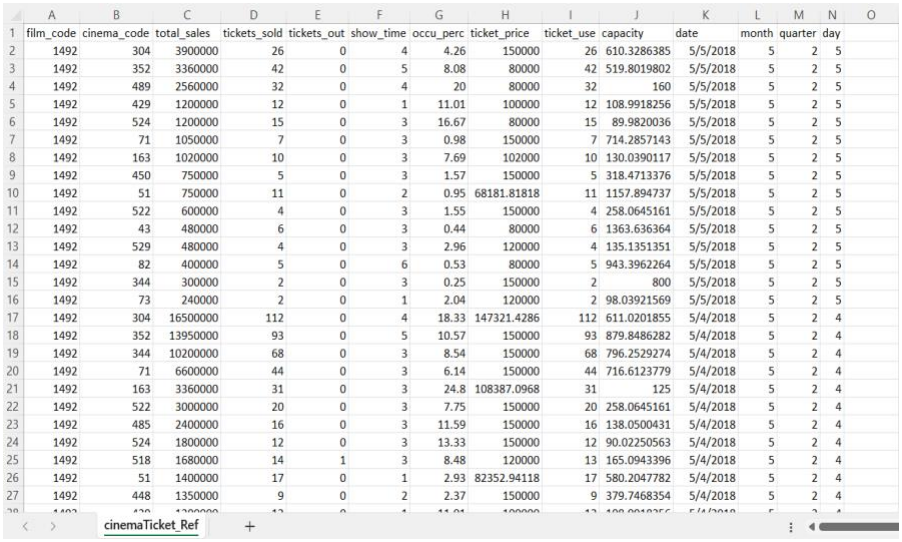
References :

[1] J. Ahmad, P. Duraisamy, A. Yousef and B. Buckles, "Movie success prediction using data mining," 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Delhi, India, 2017, pp. 1-4, doi: 10.1109/ICCCNT.2017.8204173.

[2] Lash, Michael & Zhao, Kang. (2015). Early Predictions of Movie Success: The Who, What, and When of Profitability. Journal of Management Information Systems. 33. 10.1080/07421222.2016.1243969.

[3] Baranowski, P., Korczak, K., & Zając, J. (n.d.). Forecasting cinema attendance at the movie show level: Evidence from Poland. Business Systems Research : International journal of the Society for Advancing Innovation and Research in Economy. https://hrcak.srce.hr/ojs/index.php/bsr/article/view/12668

Appendix:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	film_code	cinema_code	total_sales	tickets_sold	tickets_out	show_time	occu_perc	ticket_price	ticket_use	capacity	date	month	quarter	day	
2	1492	304	3900000	26	0	4	4.26	150000	26	610.3286385	5/5/2018	5	2	5	
3	1492	352	3360000	42	0	5	8.08	80000	42	519.8019802	5/5/2018	5	2	5	
4	1492	489	2560000	32	0	4	20	80000	32	160	5/5/2018	5	2	5	
5	1492	429	1200000	12	0	1	11.01	100000	12	108.9918256	5/5/2018	5	2	5	
6	1492	524	1200000	15	0	3	16.67	80000	15	89.9820036	5/5/2018	5	2	5	
7	1492	71	1050000	7	0	3	0.98	150000	7	714.2857143	5/5/2018	5	2	5	
8	1492	163	1020000	10	0	3	7.69	102000	10	130.0390117	5/5/2018	5	2	5	
9	1492	450	750000	5	0	3	1.57	150000	5	318.4713376	5/5/2018	5	2	5	
10	1492	51	750000	11	0	2	0.95	68181.81818	11	1157.894737	5/5/2018	5	2	5	
11	1492	522	600000	4	0	3	1.55	150000	4	258.0645161	5/5/2018	5	2	5	
12	1492	43	480000	6	0	3	0.44	80000	6	1363.636364	5/5/2018	5	2	5	
13	1492	529	480000	4	0	3	2.96	120000	4	135.1351351	5/5/2018	5	2	5	
14	1492	82	400000	5	0	6	0.53	80000	5	943.3962264	5/5/2018	5	2	5	
15	1492	344	300000	2	0	3	0.25	150000	2	800	5/5/2018	5	2	5	
16	1492	73	240000	2	0	1	2.04	120000	2	98.03921569	5/5/2018	5	2	5	
17	1492	304	16500000	112	0	4	18.33	147321.4286	112	611.0201855	5/4/2018	5	2	4	
18	1492	352	13950000	93	0	5	10.57	150000	93	879.8486282	5/4/2018	5	2	4	
19	1492	344	10200000	68	0	3	8.54	150000	68	796.2529274	5/4/2018	5	2	4	
20	1492	71	6600000	44	0	3	6.14	150000	44	716.6123779	5/4/2018	5	2	4	
21	1492	163	3360000	31	0	3	24.8	108387.0968	31	125	5/4/2018	5	2	4	
22	1492	522	3000000	20	0	3	7.75	150000	20	258.0645161	5/4/2018	5	2	4	
23	1492	485	2400000	16	0	3	11.59	150000	16	138.0500431	5/4/2018	5	2	4	
24	1492	524	1800000	12	0	3	13.33	150000	12	90.02250563	5/4/2018	5	2	4	
25	1492	518	1680000	14	1	3	8.48	120000	13	165.0943396	5/4/2018	5	2	4	
26	1492	51	1400000	17	0	1	2.93	82352.94118	17	580.2047782	5/4/2018	5	2	4	
27	1492	448	1350000	9	0	2	2.37	150000	9	379.7468354	5/4/2018	5	2	4	

cinemaTicket_Ref