Cinema Audience Forecasting Challenge

Project Report

1. Introduction

Accurate forecasting of cinema audience attendance is critical for theater operators, movie distributors, and marketing teams. Reliable predictions enable better resource allocation, optimized show scheduling, staffing decisions, and targeted promotions. Traditional forecasting approaches often fail to capture complex temporal patterns and dynamic consumer behavior.

This project addresses the problem of **predicting daily cinema audience counts** using historical booking and visit data collected from multiple theater management systems. The task is framed as a **time-series regression problem**, where the goal is to forecast future attendance based on past trends, calendar effects, and booking activity.

2. Problem Statement

Given historical data consisting of:

- Daily audience counts
- Booking transactions
- Theater metadata
- Calendar information

the objective is to build a **robust regression model** that can:

- Accurately predict future audience counts
- Generalize well to unseen dates
- Handle seasonality, trends, and data sparsity
- Generate reliable predictions for a test period without future ground truth

3. Dataset Description

The project integrates data from two independent booking systems, providing a comprehensive view of theater operations.

3.1 Dataset Overview

File Name	Description
`booknow_visits.csv`	Contains daily audience counts (target variable)
`booknow_booking.csv`	Booking transactions from BookNow platform
`cinePOS_booking.csv`	Booking transactions from CinePOS platform
`booknow_theaters.csv`	Theater metadata (type, location, coordinates)
`cinePOS_theaters.csv`	Theater metadata from CinePOS
`date_info.csv`	Calendar mapping (date → weekday)
`movie_theater_id_relation.csv`	Mapping between theater IDs across systems

3.2 Target Variable

- **`audience_count`**: Total number of people attending a specific theater on a given date.

4. Exploratory Data Analysis (EDA)

EDA was conducted to understand the structure, distribution, and temporal behavior of the data.

4.1 Temporal Patterns

- Strong **weekly seasonality** observed
  - Peak attendance on **Saturdays and Sundays**
- Clear **upward demand trend** starting around **July 2023**

4.2 Distribution Analysis

- Audience count is **right-skewed**
- Majority of shows attract **0–100 viewers**
- Few high-attendance days create long-tail behavior

4.3 Correlation Analysis

- Moderate positive correlation (**~0.45**) between:
  - Total tickets booked
  - Final audience count

- Confirms booking data is a meaningful predictive signal

4.4 Data Challenges Identified

- Sparse attendance for some theaters
- Missing dates for certain theater histories
- Non-stationary trends across time

5. Data Preprocessing

The following preprocessing steps were applied:

- Date parsing and chronological sorting
- Alignment of theater IDs across booking systems
- Aggregation of booking data at the **theater-date level**
- Handling missing values using forward-fill where appropriate
- Removal of data leakage from future dates

6. Feature Engineering

Feature engineering was the most critical component of the project.

6.1 Temporal Features

- Day of week
- Month
- Quarter
- Binary indicator for Indian public holidays

These features capture weekly and seasonal effects.

6.2 Lag Features

To model temporal dependency:

- Audience count lagged by **1, 3, 7, 14, and 21 days**

Lag features allow the model to learn from historical attendance behavior.

6.3 Rolling Window Statistics

To capture short-term volatility:

- Rolling mean (7, 14, 21 days)

- Rolling standard deviation (7, 14, 21 days)

6.4 Exponentially Weighted Means (EWM)

- Applied decay-based averaging
- Gives higher importance to recent audience values
- Helps capture trend shifts faster than simple rolling averages

7. Model Development

Multiple regression models were implemented and evaluated.

7.1 Baseline Models

- Ridge Regression
- Random Forest Regressor

These models served as performance baselines.

7.2 Gradient Boosting Models

- XGBoost (tuned)
- LightGBM (tuned)

Gradient boosting models performed significantly better due to:

- Ability to model non-linear relationships
- Robust handling of sparse and skewed data
- Built-in regularization

8. Model Evaluation and Comparison

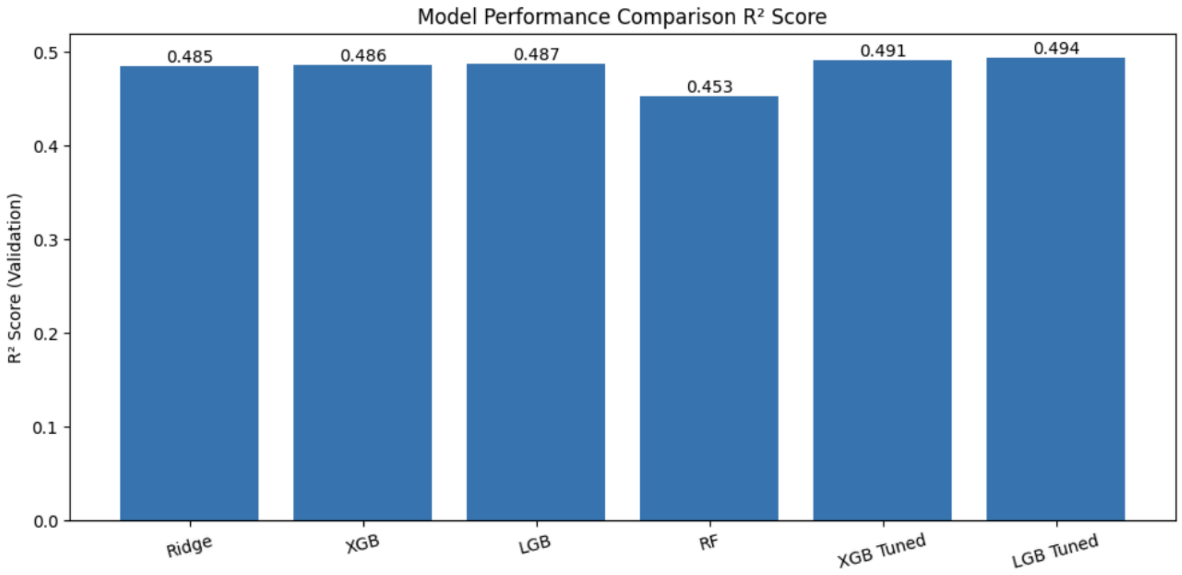
8.1 Evaluation Metric

- Validation score (R^2-based metric)

8.2 Performance Comparison

Model	Validation R^2 Score
LightGBM (Tuned)	**0.494**
XGBoost (Tuned)	**0.491**

LightGBM (Base)	0.487
XGBoost (Base)	0.486
Ridge Regression	0.485
Random Forest	0.453



8.3 Model Selection

LightGBM was selected as the final model due to:

- Highest validation performance
- Faster training time
- Better generalization on time-series data

9. Iterative Forecasting Strategy

9.1 Motivation

For future dates:

- Lag features depend on values that are not yet known
- Direct prediction would cause data leakage

9.2 Iterative Forecasting Process

1. Predict audience count for day *t+1*
2. Append prediction to historical data
3. Recompute lag and rolling features
4. Repeat for each subsequent day in test period

This approach simulates real-world forecasting conditions.

10. Key Learnings and Insights

- Feature engineering has a larger impact than model complexity
- Lag and rolling features are essential for time-series regression
- Gradient boosting models outperform traditional ensembles
- Iterative forecasting is mandatory for real-world deployment
- Handling skewed targets improves model stability

11. Future Work

- Incorporate movie metadata and release schedules
- Add deep learning models (LSTM / Temporal CNN)
- Perform per-theater error analysis
- Introduce probabilistic forecasting for uncertainty estimation

14. Conclusion

This project demonstrates a complete, production-style approach to time-series forecasting in the cinema domain. Through careful preprocessing, extensive feature engineering, and iterative prediction strategies, the final LightGBM model achieved strong performance while maintaining realism and robustness.

The methodology and insights from this project are directly applicable to other demand forecasting problems in retail, transportation, and entertainment industries.