# Real-Time Anomaly Detection in the Movie Industry Using Time Series and Ensemble Models

Chaitanya Sura
*Computer Science*
*Georgia State University*
Atlanta, USA
csura1@student.gsu.edu

***Abstract*—Sales forecasting is the key to business intelligence that allows organizations to plan resources, manage finances, and optimize operations. It serves as the foundation for production planning, financial forecasting, and inventory management in many industries. In the movie industry, ticket sales are highly influenced by film releases, seasonal trends, and customer behavior, making accurate forecasting essential for planning show schedules, pricing, and promotions. In addition, detecting anomalies in ticket sales can help identify operational problems, unexpected demand spikes, or data quality problems in real time. Forecasting also supports long-term strategic planning by helping stakeholders anticipate revenue patterns and align business growth initiatives. This project introduces a hybrid framework that combines classical time series models with machine learning algorithms to forecast daily cinema ticket sales and detect anomalies. The data set is sourced from Kaggle, includes historical sales data from multiple cinema locations along with temporal and operational attributes. After cleaning, transforming the data and applying log transformations—various models were trained and tested. These included Simple and Weighted Moving Average, ARIMA, Random Forest, XGBoost, and Linear Regression, using lag-based feature engineering. Evaluation based on MSE, RMSE, MAE, and $R^2$ metrics showed that ARIMA was the best-performing statistical model, while XGBoost provided the highest accuracy among machine learning approaches. This forecasting system supports data-driven decision-making for efficient scheduling, anomaly detection, and resource management. Future work will focus on deep learning models such as LSTM Autoencoders for more accurate real-time anomaly detection and leveraging external factors like holidays, promotional events, and weather data to enhance the forecasting capability and contextual understanding of sales patterns.**

***Index Terms*—Sales forecasting, Time Series Analysis, Anomaly detection, ARIMA, SARIMAX, Random Forest, XGBoost, LSTM**

## I. Introduction

In today's data-driven world, the ability to predict future trends and respond to unexpected changes is essential for effective business operations. One of the most important tools in this process is sales forecasting, which helps organizations plan ahead, allocate resources efficiently, manage financial risks, and improve decision making at both operational and strategic levels. In the cinema industry, accurate ticket sales forecasting enables operators to optimize show schedules, staffing, and concession planning, while also adjusting ticket pricing and marketing efforts based on expected demand. It also supports long-term growth by helping stakeholders anticipate revenue streams, evaluate investment opportunities, and align content acquisition with audience trends.

Despite the advantages of forecasting, real-world conditions are rarely predictable. Sudden events such as technical glitches, extreme weather, or public transportation failures can cause sharp and unexplained spikes in ticket sales. These irregular behaviors, known as anomalies, are not captured by standard forecasting models and can go unnoticed until they have already impacted revenue or customer satisfaction. For example, a surge in sales might indicate fraudulent activity or data input errors, while a drop may point to disruptions in the ticketing system or last-minute event cancellations.

To address such challenges, organizations are increasingly adopting real-time anomaly detection systems. These systems are designed to monitor the data as they flow, identify unexpected deviations from normal patterns, and trigger alerts for immediate investigation. In the movie domain, real-time anomaly detection can play a crucial role in maintaining service quality, minimizing losses, and responding quickly to unexpected circumstances.

This report introduces a hybrid approach that combines classical time series models with ensemble-based machine learning techniques to provide accurate sales forecasting and effective real-time anomaly detection. By taking advantage of historical patterns through models like ARIMA and combining them with supervised learning approaches such as Random Forest and XGBoost, the system aims to offer a practical solution for data-driven decision-making. The framework is designed not only to predict what is likely to happen, but also to detect what should not be happening, ensuring cinema operators can plan smarter and react faster in a highly dynamic business environment.

## II. Related Work

Sales forecasting and anomaly detection have long been central to operations research and data-driven business planning. Traditional time series models such as the AutoRegressive Integrated Moving Average (ARIMA) have been widely used for forecasting temporal data due to their simplicity and interpretability. Box and Jenkins formalized the ARIMA model framework in their foundational work on time series analysis [2]. While ARIMA remains a popular choice for univariate

forecasting, its limitations become evident when handling non-linear or irregular patterns in highly dynamic environments like the cinema industry.

To address these limitations, researchers have explored more advanced forecasting techniques such as seasonal ARIMA (SARIMA), exponential smoothing, and hybrid models that integrate statistical and machine learning approaches. In [6], Hyndman et al. compared exponential smoothing methods and ARIMA for forecasting business time series and showed that no single model outperforms others across all scenarios. More recently, ensemble methods like Random Forest and gradient boosting (e.g., XGBoost) have been adopted to model complex non-linear relationships in sales and demand forecasting [4], [5]. These methods benefit from their ability to learn from lag-based features and capture intricate patterns beyond linear assumptions.

Anomaly detection, on the other hand, has historically relied on statistical thresholds such as Z-scores and interquartile range (IQR) filters. These methods are effective for identifying outliers in normally distributed data but often underperform when faced with non-stationary or time-dependent signals. Chandola et al. provided a comprehensive survey on anomaly detection techniques, highlighting their applications in time series data and noting the challenges of real-time detection [3]. With the increasing availability of real-time data streams, researchers have begun incorporating machine learning and deep learning approaches, such as LSTM-based Autoencoders and Isolation Forests, for more robust anomaly detection [7],[9].

In the context of the cinema or entertainment industry, literature specifically addressing real-time anomaly detection is limited. However, similar frameworks have been applied in e-commerce, retail, and transportation sectors where ticketing or transaction volumes exhibit seasonal and event-driven fluctuations. For example, Laptev et al. demonstrated the use of time series forecasting and outlier detection techniques at scale within large-scale industrial monitoring systems [11], while Ren et al. proposed an interpretable multivariate anomaly detection model using deep attention-based networks, which could generalize well to domains like cinema where multiple contextual variables are in play [10].

Further, research from the streaming and media services sector has shown how forecasting user activity and detecting anomalies in content consumption can improve operational efficiency and user satisfaction [8]. These insights are particularly relevant to the cinema domain, where audience engagement, show scheduling, and operational resources all hinge on demand prediction and timely issue identification.

This project builds upon these prior works by developing a hybrid system that integrates classical time series forecasting with ensemble-based machine learning models for both sales prediction and real-time anomaly detection in the cinema industry. By leveraging historical patterns and machine-learned features, the system aims to support more accurate forecasting and proactive anomaly response.

## III. Technical Overview

### A. Time Series Forecasting

In many industries, time series data plays a critical role in operational forecasting and monitoring. Time series forecasting refers to the task of using past observations of a variable, arranged chronologically, to predict future values. In the cinema industry, ticket sales represent a classic time series problem: sales data is collected daily and is influenced by temporal patterns such as weekends, public holidays, release schedules, and seasonal audience behavior.

An effective forecasting system must not only capture the expected patterns in sales but also be sensitive enough to detect unexpected changes—known as anomalies. These anomalies may be caused by technical issues, sudden demand shifts, or external events such as weather disruptions, special film promotions, or social media virality. A robust forecasting system, therefore, must be capable of modeling regular patterns and flagging deviations in near real time.

### B. Components of Time Series Data

Time series data typically includes several components:

- Trend: The long-term movement in the data, which may be upward or downward.
- Seasonality: Regular patterns that repeat over a fixed period (e.g., weekends, holidays).
- Cyclic behavior: Irregular fluctuations influenced by non-seasonal factors.
- Noise: Random variation or measurement error.

For accurate modeling, the series often needs to be transformed into a stationary format—where the mean and variance remain constant over time. This is achieved through techniques such as differencing, detrending, and log transformation.

### C. Time Series Forecasting Models

Traditional time series forecasting methods rely on the assumption that future values can be modeled as a function of past observations. These models are widely used for their interpretability and effectiveness in capturing linear trends and seasonality. The following classical statistical models are commonly applied in business analytics and have been evaluated in this study.

1) Simple Moving Average (SMA): SMA calculates the mean of the most recent n observations to forecast the next value. It assumes that all selected past data points are equally relevant. While easy to implement, SMA is sensitive to lag and does not account for trends or seasonality.

2) Weighted Moving Average (WMA): WMA improves upon SMA by assigning higher weights to more recent data points. This helps the model respond more effectively to recent changes in the time series, which is particularly beneficial in environments like cinema ticket sales where short-term fluctuations are common.

3) Exponential Smoothing (ETS / Holt-Winters): Exponential smoothing assigns exponentially decreasing weights to older

observations. Variants like Holt's linear method and the Holt-Winters method (Triple Exponential Smoothing) extend this approach to model both trend and seasonality.

- Holt's Linear Trend Method: Captures level and trend.
- Holt-Winters (Additive or Multiplicative): Captures level, trend, and seasonality explicitly.

These models are computationally efficient and effective for real-world sales data with strong seasonal patterns, such as weekend or holiday peaks in cinema attendance.

4) AutoRegressive Integrated Moving Average (ARIMA): ARIMA is a statistical model that captures the autocorrelation structure in a stationary time series using three components:

- AR (Auto-regression): Forecasting based on lagged values.
- I (Integration): Differencing the series to achieve stationarity.
- MA (Moving Average): Modeling the forecast error using past residuals.

The model is represented as ARIMA(p,d,q), where p is the number of autoregressive terms, d is the degree of differencing, and q is the order of the moving average.

5) Seasonal ARIMA (SARIMA): SARIMA extends ARIMA by incorporating seasonal components: (p,d,q,m), where m is the number of observations per seasonal cycle. This makes it suitable for data with repeating patterns, such as weekly cycles in cinema sales.

These classical models provide a strong foundation for understanding temporal dynamics and are often used as baselines in time series forecasting tasks. They are favored for their transparency and minimal data preprocessing requirements but may require additional tuning when faced with noisy or highly non-linear data.

### *D. Machine Learning Models for Forecasting*

While classical models are effective for linear and stationary data, they may fall short in capturing complex, non-linear relationships. Machine learning (ML) models provide greater flexibility and can incorporate additional contextual features (e.g., holidays, promotions, cinema location) for improved accuracy.

To apply ML models to time series forecasting, the problem is reframed as a supervised regression task. Historical observations are converted into lag-based features, such as values from the previous 5 days, which serve as inputs to predict future sales.
The following ML models were implemented:

1) Linear Regression (LR): A simple model that assumes a linear relationship between input features and the target variable. While not suited for capturing complex trends, it serves as a baseline for comparison.

2) Random Forest Regressor (RFR): An ensemble of decision trees that reduces variance by averaging the results of multiple trees trained on different subsets of the data. RFR handles non-linear relationships well and is robust to outliers and noise.

3) XGBoost Regressor: XGBoost is a gradient boosting framework that builds trees sequentially, each one correcting the errors of its predecessor. It includes regularization techniques to reduce overfitting and is optimized for speed and scalability. XGBoost is particularly effective in learning intricate patterns in cinema sales data influenced by various operational and temporal variables.

These models provide the advantage of learning from both past observations and engineered features, offering a dynamic alternative to purely statistical approaches.

### *E. Anomaly Detection Techniques*

In time series forecasting, anomalies refer to observations that deviate significantly from expected patterns. Detecting these anomalies is essential in operational settings, where a sudden drop in ticket sales might indicate a system malfunction, external event, or reporting issue.

1) Statistical Techniques

*Z-score Method:* Calculates the number of standard deviations an observation is from the mean. Points with $|z| > 3$ are typically flagged as anomalies.

*Interquartile Range (IQR):* Defines anomalies as values below $Q_3 - 1.5 \times IQR$ or above $Q_3 + 1.5 \times IQR$, where $Q_1$ and $Q_3$ are the 25th and 75th percentiles.

*Residual-Based Detection:* After forecasting, the residual (difference between actual and predicted value) is analyzed. Large residuals exceeding a dynamic threshold ($\pm 3\sigma$) suggest an anomaly.

2) Machine Learning Techniques

*Isolation Forest:* An ensemble-based anomaly detection algorithm that isolates outliers by randomly selecting features and split values. Anomalies require fewer splits to isolate, making them stand out.

*Autoencoders (LSTM-based):* Neural networks trained to reconstruct input sequences. When presented with an anomalous sequence, the reconstruction error becomes large—this deviation is used as an anomaly signal.

Dynamic Thresholding: Adapts threshold values based on rolling statistics, seasonality, or contextual inputs, making detection more responsive to changing patterns in cinema sales.

In this study, a residual-based detection framework was used due to its interpretability and real-time applicability. Anomalies were flagged by comparing residuals to thresholds derived from historical error distributions.

## IV. Dataset and Preprocessing

### *A. Dataset Overview*

The dataset used in this study was sourced from a publicly available cinema sales dataset on Kaggle. It contains detailed records of daily cinema ticket sales across multiple cinema locations. Each row in the dataset represents a specific film screened at a particular cinema on a given date, making

the dataset suitable for granular time series analysis and forecasting.

The dataset includes attributes such as: `film_code`, `cinema_code`, `date` – for identifying each unique screening, `total_sales`, `tickets_sold`, `capacity` – for understanding cinema performance, and temporal attributes like `month`, `quarter`, `day`, and `show_time`.

These features provide both temporal and operational context, enabling richer forecasting and anomaly detection.

### B. Data Cleaning

To prepare the dataset for analysis, several cleaning steps were performed:

- Date Parsing and Formatting: The `date` column was converted to a proper datetime object using `pd.to_datetime()`. This enabled sorting, resampling, and temporal grouping by day, week, or month.
- Null Value Removal: A small number of records with missing values were identified and removed using `dropna()`, ensuring no missing data interfered with time-dependent transformations.
- Duplicate Record Handling: Duplicate records were detected using `.duplicated()` and dropped. To ensure uniqueness, a new `id` column was constructed by combining `film_code`, `cinema_code`, and `date`.
- Reordering Columns: The newly created `id` column was moved to the front of the dataset to simplify navigation during analysis and debugging.

### C. Exploratory Analysis and Feature Engineering

To understand the behavior and structure of cinema ticket sales, a comprehensive exploratory data analysis was conducted. The central attribute under study, `total_sales`, was first visualized using a histogram, which revealed a heavily right-skewed distribution. This indicated that while most film screenings generated relatively low revenue, a few high-performing releases contributed to unusually large spikes in sales. This imbalance can cause modeling challenges due to the disproportionate influence of extreme values.

To quantify the skewness, the sales values were grouped into intervals of 100 million using the `pd.cut()` method. The resulting binned distribution showed that over 98% of the sales records fell into the lowest sales range, reaffirming the need for data normalization.

To handle extreme values in the dataset, the Interquartile Range (IQR) method was used for outlier detection. The IQR is calculated as the difference between the 75th percentile ($Q_3$) and 25th percentile ($Q_1$). Any value outside the range: Lower Bound = $Q_3 - 1.5 \times IQR$, Upper Bound = $Q_3 + 1.5 \times IQR$ was classified as an outlier. These outliers were subsequently removed to reduce their influence on model training. Figure 1 shows the resulting distribution of `total_sales` after outlier removal.

*1) Log Transformation for Skewness Correction:* Given the significant skewness of the original `total_sales` distribution, a logarithmic transformation was applied to stabilize variance

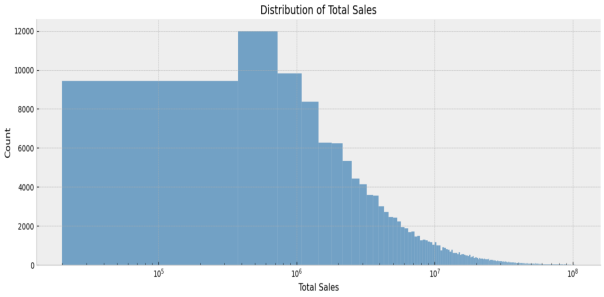


Fig. 1. Histogram of total ticket sales after outlier removal using the IQR method

and make the data more suitable for time series modeling. The transformation compresses the range of large values while stretching out smaller ones, thus normalizing the distribution.

`total_sales_log` = log(`total_sales`)

This step was particularly important for classical models like ARIMA and SARIMAX, which assume that the time series is homoscedastic (i.e., has constant variance over time).

*2) Feature Engineering for Time-Aware Modeling:* To prepare the dataset for machine learning models, which require supervised learning formats, several engineered features were added:

- Lag Features: Sales values from the previous five days were created as separate columns (`lag_1` to `lag_5`). These lag features help capture temporal dependencies and allow models like Random Forest and XGBoost to learn from past behavior.
- Temporal Features: From the date column, additional time-based features such as weekday, month, and quarter were extracted. These variables enable models to understand cyclical sales trends — for instance, increased traffic on weekends or during certain months.
- Aggregated Trends: Grouped sales by time period, such as day of the week or month, were analyzed and visualized to uncover recurring seasonal or weekly trends. These patterns also play a key role in anomaly detection, where deviations from expected behavior can be flagged.

At the conclusion of preprocessing, two final datasets were constructed:

- A log-transformed, stationary version of the dataset for use with time series forecasting models like ARIMA and SARIMAX.
- A supervised learning format with lag-based and temporal features suitable for training machine learning models such as Random Forest, XGBoost, and Linear Regression.

This structured preprocessing pipeline ensured that the dataset was both clean and analytically rich, enabling accurate forecasting and effective anomaly detection in downstream modeling.

## V. Methodology

This section outlines the modeling workflow adopted to forecast cinema ticket sales and detect anomalies. A dual-

modeling strategy was used — combining classical time series forecasting with machine learning regression — to leverage both statistical rigor and data-driven flexibility.

### A. Time Series Modeling Approach

Every meaningful time series analysis begins with an important question — is the data stationary? Stationarity means the statistical properties (mean, variance, autocorrelation) of the series do not change over time. To evaluate it, we applied the Augmented Dickey-Fuller (ADF) test, a statistical test where:

- Null hypothesis ($H_0$): The series is not stationary
- Alternative hypothesis ($H_1$): The series is stationary

This ADF test was applied to the log-transformed `total_sales` column. The test confirmed that the transformed series was stationary, allowing us to proceed with modeling. First-order differencing was applied to eliminate any remaining linear trends and make the data suitable for autoregressive modeling.

As an initial step, we implemented two simple time series forecasting methods: Simple Moving Average (which calculates the unweighted average of the previous n observations) and Weighted Moving Average (which assigns higher weights to more recent observations in the calculation window).

Both SMA and WMA were used to predict daily ticket sales based on short historical windows and were compared against actual values. While they offered a basic understanding of trend behavior, their performance in high-variance regions was limited, making them more suitable as benchmarking tools rather than production-ready models.

To determine the optimal parameters for ARIMA models, we analyzed Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots. These plots revealed the time lags with the highest correlation to current values, which informed the choice of autoregressive (AR) and moving average (MA) terms.

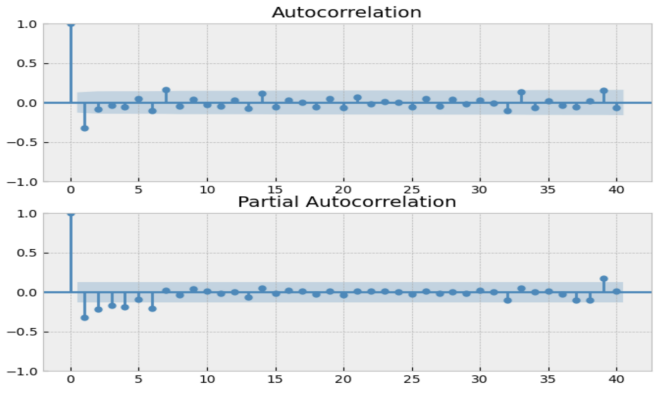


Fig. 2. Auto Correlation and Partial Auto Correlation Plot

The ARIMA model was configured using the `auto_arima()` function, which automatically selected the (p, d, q) values by minimizing the Akaike Information Criterion (AIC). To incorporate seasonality—such as monthly peaks in cinema attendance—we extended the ARIMA model into a SARIMAX model, specifying a seasonal component with a period of 12 (months). These models were trained on the log-transformed, differenced dataset and produced one-step-ahead forecasts.

Additionally, residuals from ARIMA and SARIMAX models were retained for subsequent anomaly detection, where significant deviations between actual and predicted values signal unusual behavior.

### B. Machine Learning Feature Engineering

Machine learning models require the dataset to be in a structured, supervised format. To achieve this, we engineered a range of input features that capture both temporal dependencies and cyclical patterns in ticket sales:

Lag Features: The `total_sales` values from the previous 1 to 5 days (`lag_1` to `lag_5`) were added as independent variables. These features allow ML models to recognize short-term trends and recent momentum in sales.

Calendar-based Features: From the date column, we extracted day of the week, month, and quarter, which capture weekly and seasonal audience behaviors. For example, sales typically spike during weekends or holidays.

These engineered features transformed the original univariate time series into a multivariate regression problem, enabling the use of supervised learning models like Random Forest and XGBoost.

### C. Model Training Strategy

We implemented two parallel modeling pipelines:

1) Time Series Modeling Pipeline: The ARIMA and SARIMAX models were trained using the transformed time series. One-step-ahead predictions were generated, and rolling forecasts were used to predict the next 30 days. Residuals were monitored and plotted for potential anomaly detection.

2) Machine Learning Modeling Pipeline: The feature-engineered dataset was split into training and test sets, with the last 30 days of records reserved for validation. Three models were trained:

- Random Forest Regressor: an ensemble model based on multiple decision trees.
- XGBoost Regressor: a boosting algorithm that builds trees sequentially to minimize errors.
- Linear Regression: used as a baseline to evaluate the effectiveness of complex models.

To improve robustness and prevent overfitting, 5-fold cross-validation was applied to the training data for each ML model. Each model was trained to predict daily ticket sales using the lag and calendar-based features.

### D. Evaluation Strategy

All models were evaluated on their ability to forecast the next 30 days of cinema ticket sales. The following performance metrics were used:

- Mean Squared Error (MSE): Measures the average squared difference between actual and predicted values.

- Root Mean Squared Error (RMSE): Provides the standard deviation of the prediction errors.
- Mean Absolute Error (MAE): Measures the average magnitude of errors in the predictions.
- $R^2$ Score: Indicates how well the model explains the variance in the target variable.

In addition to these numerical metrics, forecast vs. actual plots were generated for visual inspection of each model's performance. For ARIMA and SARIMAX, residual plots were used to assess whether the errors were randomly distributed and to flag potential anomalies.

A final performance comparison was drawn between all models to identify the best approach in terms of accuracy, stability, and responsiveness to sales fluctuations.

## VI. Results

This section presents the evaluation of both time series approaches and machine learning regressors on a 30-day test set. Four key metrics were used for model comparison: RMSE, MAE, MSE and $R^2$ Score (Coefficient of Determination). These metrics collectively measure prediction accuracy, variance error, and the model's explanatory power.

The time series models were trained on log-transformed and differenced data to ensure stationarity. Baseline methods such as SMA and WMA failed to capture the volatility in cinema ticket sales, yielding $R^2$ scores of -0.387 and -0.770, respectively. In contrast, the ARIMA model performed significantly better, achieving the lowest RMSE ($2.67 \times 10^9$) and the highest $R^2$ score (0.553), effectively capturing short-term fluctuations in the data.

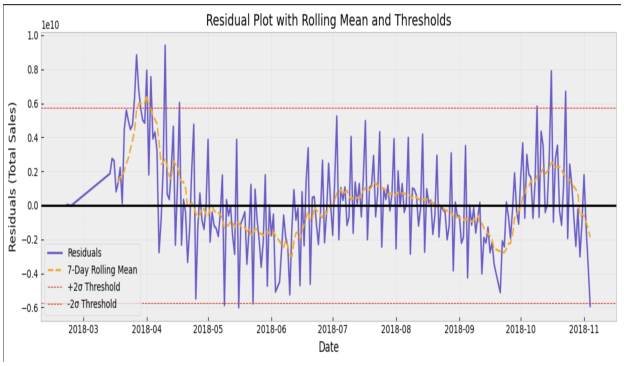


Fig. 3. Residual Plot

To further validate model performance, residual analysis was conducted. Residuals from the ARIMA model were tightly centered around zero with no visible autocorrelation, indicating that the model adequately captured the underlying structure of the time series. This was supported by a Ljung-Box test, which returned high p-values—confirming the residuals resembled white noise. A visual residual plot (Fig. 3) further illustrated this behavior, where residuals appeared randomly distributed with no distinct pattern.

In contrast, SARIMAX residuals exhibited larger variability and deviations from zero, suggesting the model struggled to account for irregular fluctuations in the data. This aligns with SARIMAX's higher RMSE ($6.11 \times 10^9$) and negative $R^2$ score (-1.34). Spikes observed in the residual plot may correspond to anomalies such as promotional events, viral popularity, or abrupt drops in audience attendance, making residuals a useful tool for anomaly detection.

Machine learning models were trained on a restructured dataset featuring lag variables (`lag_1` to `lag_5`) and calendar-based attributes (weekday, month, quarter). These features allowed the models to learn temporal dependencies in a supervised setting.

| | Method | Algorithm | MSE | RMSE | MAE | R2 |
|---|---|---|---|---|---|---|
| 0 | Time Series | Simple Moving Average | 1.340669e+19 | 3.661515e+09 | 3.179639e+09 | -0.387093 |
| 1 | Time Series | Weighted Moving Average | 1.711698e+19 | 4.137267e+09 | 3.139920e+09 | -0.770970 |
| 2 | Time Series | ARIMA | 7.127937e+18 | 2.669820e+09 | 2.131719e+09 | 0.553647 |
| 3 | Time Series | SARIMAX | 1.934418e+19 | 4.398201e+09 | 3.374759e+09 | -0.211337 |
| 4 | Machine Learning | RFR(n_estimators=100) | 7.716221e+18 | 2.777809e+09 | 2.402536e+09 | 0.201658 |
| 5 | Machine Learning | RFR(n_estimators=200) | 7.470949e+18 | 2.733304e+09 | 2.339553e+09 | 0.227035 |
| 6 | Machine Learning | RFR(n_estimators=300) | 7.442448e+18 | 2.728085e+09 | 2.335873e+09 | 0.229984 |
| 7 | Machine Learning | XGBRegressor | 7.186090e+18 | 2.680688e+09 | 2.126586e+09 | 0.256507 |
| 8 | Machine Learning | Linear Regression | 9.677848e+18 | 3.110924e+09 | 2.918928e+09 | -0.001297 |

Fig. 4. Model Performance Summary

The XGBoost Regressor outperformed other ML models with an RMSE of $2.68 \times 10^9$, MAE of $2.13 \times 10^9$, and an $R^2$ score of 0.256. This model effectively learned from complex, non-linear patterns and showed robustness across both peak and low sales periods.

The Random Forest Regressor, configured with 300 estimators, performed comparably with an RMSE of $2.73 \times 10^9$ and $R^2$ of 0.229. Its ensemble nature helped smooth out predictions, but it slightly lagged behind XGBoost in capturing sharper sales changes.

Linear Regression, due to its simplicity, underperformed with an $R^2$ score near zero (-0.001) and high residual errors, showing its inability to model non-linear temporal patterns effectively.

A consolidated comparison of all models is provided in Fig. 4. ARIMA remains the top-performing model overall, followed closely by XGBoost in the machine learning category.

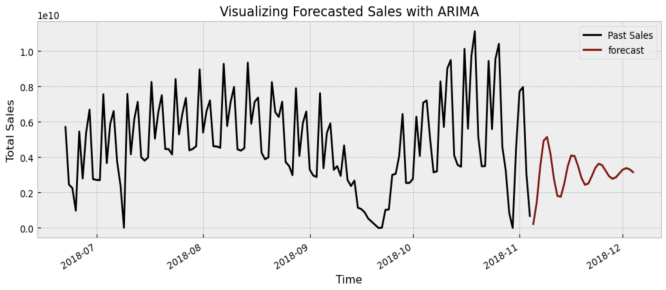


Fig. 5. Forecasted Sales with ARIMA

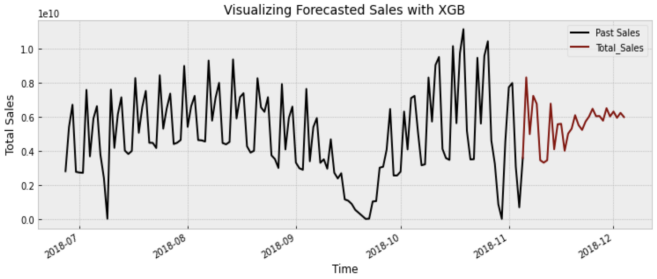


Fig. 6. Forecasted Sales with XGB

Both ARIMA and XGBoost were used to forecast the next 30 days of ticket sales. Forecast plots (Figures 3 and 4) revealed that ARIMA generated smoother trends, making it ideal for anomaly detection and stable trend projection. In contrast, XGBoost produced more responsive predictions, suitable for scenarios requiring dynamic adjustments.

## VII. Conclusion

This project focused on building a reliable system to forecast cinema ticket sales and detect unusual patterns using both time series and machine learning models. We used historical ticket sales data and applied a mix of classical statistical methods (like ARIMA and SARIMAX) and modern machine learning models (Random Forest, XGBoost, and Linear Regression) to compare their forecasting accuracy.

Among all the models tested, ARIMA gave the most accurate results. It had the lowest prediction errors and the highest $R^2$ score, which means it was best at capturing the underlying patterns in the sales data. Its residuals also showed a random distribution, making it a strong choice for detecting anomalies in real-time. XGBoost was the best performer among the machine learning models, especially when we included lag-based features and time-related information. It handled sudden changes in ticket sales better than other models.

The SARIMAX model, which is designed to model seasonality, did not perform well, likely because the data did not have strong seasonal trends. Basic models like SMA and WMA also struggled to handle the high variability in ticket sales.

This project highlights the importance of thorough data preparation—like handling missing values, removing outliers, checking for stationarity, and applying log transformation—before applying forecasting models. The results show that combining both statistical and machine learning methods can lead to better sales predictions and help cinema operators make more informed, data-driven decisions.

## VIII. Future Scope

While the current system provides a strong foundation for forecasting cinema ticket sales and identifying anomalies, there are several ways it can be improved and expanded in the future.

One major area for enhancement is the use of deep learning models, such as LSTM (Long Short-Term Memory) networks and LSTM Autoencoders. These models are specifically designed to handle sequential data and can capture long-term dependencies better than traditional time series models. LSTMs could improve forecast accuracy in scenarios where user behavior changes gradually or where patterns are more complex and less obvious. LSTM Autoencoders, in particular, offer promising results for real-time anomaly detection, as they can learn normal behavior and identify any unusual patterns by reconstruction error.

Another opportunity is to integrate external data sources. Currently, the model is trained solely on internal ticket sales data. However, adding variables such as holiday calendars, promotional campaigns, weather conditions, or social media sentiment could help the model better understand what drives sudden spikes or drops in ticket sales. This could also improve the accuracy of predictions during irregular events.

In a real-world setting, it would also be useful to implement the system as a real-time dashboard that provides continuous monitoring of ticket sales. Such a system could automatically flag anomalies as they happen and generate short-term forecasts to support decisions about show scheduling, marketing, and staffing.

Lastly, the same framework could be extended to other entertainment sectors like concerts, sports events, or streaming platforms, where demand forecasting and anomaly detection are equally important. By adapting the methodology to different datasets, the system could serve as a general solution for time-based business intelligence.

## References

[1] Kaggle. *Cinema Ticket Dataset*. Available at: https://www.kaggle.com/datasets/arashnic/cinema-ticket/data. Accessed: March 27, 2025.
[2] G. E. P. Box and G. M. Jenkins, Time Series Analysis: Forecasting and Control, Holden-Day, 1976.
[3] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly Detection: A Survey," ACM Computing Surveys, vol. 41, no. 3, pp. 1–58, 2009.
[4] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
[5] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp. 785–794, 2016.
[6] R. J. Hyndman and G. Athanasopoulos, Forecasting: Principles and Practice, 2nd ed., OTexts, 2018.
[7] F. Hundman, V. Constantinou, C. Laporte, I. Colwell, and T. Soderstrom, "Detecting Spacecraft Anomalies Using LSTMs and Nonparametric Dynamic Thresholding," Proc. 24th ACM SIGKDD, 2018.
[8] S. Munagala, V. Suryanarayana, and S. Gollapalli, "Anomaly Detection for a Video Streaming Service," arXiv preprint arXiv:2104.10736, 2021.
[9] S. Blázquez-García, A. Conde, U. Mori, and J. A. Lozano, "A Review on Outlier/Anomaly Detection in Time Series Data," ACM Computing Surveys, vol. 54, no. 3, pp. 1–33, 2021.
[10] J. Ren, X. Zhang, X. Wang, M. Zeng, L. Sun, and Q. He, "Time-Series Anomaly Detection Service at Microsoft," in Proc. 25th ACM SIGKDD, pp. 3009–3017, 2019.
[11] N. Laptev, S. Amizadeh, and I. Flint, "Generic and Scalable Framework for Automated Time-Series Anomaly Detection," in Proc. 21st ACM SIGKDD, pp. 1939–1947, 2015.
[12] H. Chen, H. Dai, X. Wang, and G. Xu, "Online Sales Forecasting Using Deep Learning Models: A Case Study," IEEE Access, vol. 7, pp. 173173–173183, 2019.
[13] A. Lavin and S. Ahmad, "Evaluating Real-Time Anomaly Detection Algorithms – The Numenta Anomaly Benchmark," in Proc. 14th IEEE ICMLA, pp. 38–44, 2015.