Optimal Flight Ticket Price Discovery Using Time Series Analysis SARIMAX Model

Avinash Reddy Kovvuri, P. Shyamala Madhuri(✉), D. Shankar, Mallela Santhi Priya, Mohammad Sajidh Ali, and Surya RamTeja Managam

Vishnu Institute of Technology, Bhimavaram, India
shyamalamadhuri.y@vishnu.edu.in

Abstract. Flight fare analysis predicts daily lowest prices for specific routes using SARIMAX, a sophisticated time series model. Multiple SARIMAX models are employed for each airline on a route, considering seasonality, trends, cyclicality, and demand fluctuations. This ensures complexity and efficiency, aiding informed decisions for airlines and travelers. By integrating daily fluctuations, the model offers a comprehensive understanding of price dynamics, enhancing forecast accuracy. SARIMAX's adaptability allows seamless integration with diverse data sources, facilitating versatile applications. Meticulous analysis of historical and real-time data empowers stakeholders to optimize travel plans and financial strategies. This approach benefits airlines by enabling effective revenue management and pricing strategies, while also aiding travelers in budgeting and informed decision-making. Continued research will enhance SARIMAX's accuracy, contributing to aviation industry efficiency and transparency.

Keywords: Flight Fare Analysis · Machine Learning · Time Series Analysis · SARIMAX · Optimal Ticket Prices · Dynamic Routes · Airline Data · Forecasting · Seasonality · Trend Analysis · Cyclicality · Demand Fluctuations

1 Introduction

Flight fare analysis is to generate an analysis with respect to the prediction values of each model in complex architecture and the prediction values enable the people, especially middle class and below middle-class people to have budget-friendly flight journeys. And this analysis is based on machine learning time series analysis [1]. In contemporary times, machine learning techniques wield substantial influence in both society and the market. Additionally, the SARIMAX model, a time series model reliant on only one variable column data that fluctuates over time, holds considerable significance. The word Time Series Analysis refer by analyzing a data collected over an interval of time in a sequence. The SARIMAX model, which stands for Seasonal AutoRegressive Integrated Moving Average with eXogenous factors, is an extension of the ARIMA (AutoRegressive Integrated Moving Average) model. SARIMAX incorporates additional features to account for both seasonality and exogenous variables. The incident laid the foundation for the

P. Pareek et al. (Eds.): IC4S 2024, LNICST 599, pp. 79–99, 2025.
https://doi.org/10.1007/978-3-031-77081-4_8

idea- Due to the huge value change in the prices of the flight, it was difficult for the people to book tickets and have a journey and the application was completely based on the normal regression techniques which doesn't involve trend, seasonality, cyclicality, etc. Finally, the purpose of the flight fare prediction, based on the described approach using SARIMAX models, is to provide valuable insights and assistance in making informed decisions related to air travel such as

Cost Optimization. By forecasting the lowest ticket prices, travelers can plan their journeys to take advantage of more affordable periods, helping them optimize their travel costs.

Decision Support for Booking. Travelers can use the predictions to empower individuals to make well-informed choices regarding the timing of their flight bookings. to secure the most cost-effective options.

Planning and Budgeting. Travel agencies, and passengers can benefit from accurate fare predictions to plan budgets, marketing strategies, and operational activities more effectively.

Flight fare analysis using SARIMAX models intersects with the themes of the Cyber-Physical Systems (CPS) by employing sophisticated time series modeling techniques, contributing to the advancement of CPS methodologies, particularly in the domain of transportation. The analysis of flight fare data not only involves processing large volumes of information but also integrates diverse data sources, reflecting the conference's focus on the integration of CPS with various technologies. Additionally, the paper emphasis on optimizing decision-making processes for airlines and travelers resonates with the conference's goal of improving system efficiency and functionality through data-driven insights. Overall, the chat project aligns with the broader objectives of the CPS conference, aiming to enhance the efficiency, reliability, and adaptability of cyber-physical systems across different domains [2].

2 Related Work

In the realm of flight fare analysis, traditional regression [3] techniques have historically been the go-to method for modeling the relationship between various factors and ticket prices. These methods, while effective to a certain extent, often exhibit limitations when it comes to capturing the intricate temporal dynamics inherent in the flight fare data. While regression models can account for certain influential factors such as flight duration, time of booking, and route popularity, they may struggle to capture the nuanced patterns of seasonality, trends, and irregularities that characterize the airline industry. Moreover, flight fare data is inherently time-dependent, with prices fluctuating over different time intervals due to various factors like demand, competition, and external events. Traditional regression models, which assume independence among observations, may fail to capture the autocorrelation and temporal the dependencies existing within the dataset, leading to suboptimal forecasts. Consequently, there has been an increasing acknowledgment within the field of aviation economics and revenue management of the need for more sophisticated modeling approaches to address these challenges. Time series analysis, particularly utilizing models like SARIMAX (Seasonal Auto-Regressive Integrated

Moving Average with Exogenous Variables), has gained prominence as a powerful tool for forecasting flight ticket prices. SARIMAX models are specifically designed to handle time-dependent data and can capture complex patterns of seasonality, trends, and autocorrelation, thereby providing more accurate forecasts compared to traditional regression techniques. By embracing time series analysis, researchers and practitioners in the field of flight fare analysis can overcome the limitations of traditional regression methods and unlock deeper insights into the temporal dynamics of flight ticket prices. This shift towards more sophisticated modeling approaches not only enhances the accuracy of fare forecasts but also empowers airlines to make data-driven decisions to optimize pricing strategies, improve revenue management, and enhance the overall customer experience.

3 Literature Survey

Flight fare analysis and forecasting using time series models like SARIMAX entails a comprehensive examination of existing academic and industry literature on the subject. It encompasses various aspects, including the application of traditional regression techniques and the advantages of time series analysis in capturing temporal dynamics and trends. Traditional regression methods, while commonly used in flight fare analysis, often struggle to capture the complexities of time-dependent patterns and seasonality inherent in the flight fare data. This limitation has led to a growing recognition of the need for more sophisticated modeling approaches that can better handle these temporal dynamics [4]. Time series analysis, particularly utilizing models like SARIMAX, has emerged as a powerful tool for forecasting flight ticket prices. SARIMAX models excel in capturing seasonality, trend, and other temporal patterns, thereby providing more accurate forecasts compared to traditional regression techniques. Through a literature survey, researchers can review studies that have applied SARIMAX models for flight fare forecasting, examining the methodologies, techniques, and practical implementations. Examples and instances from real-world scenarios provide valuable insights into how airlines have used forecasting models to optimize pricing strategies and improve revenue management. The literature survey also involves comparing SARIMAX models with other time series models, such as ARIMA and smoothing exponentially, to assess their performance in flight fare forecasting. By synthesizing existing research, researchers can identify best practices, challenges, and prospective avenues for research in the realm of flight fare analysis. This includes addressing challenges such as data availability, model interpretability, and scalability, and exploring opportunities for innovation and improvement. Ultimately, a literature survey provides a solid foundation for informing the development of research questions, methodologies, and practical implications for airlines and researchers in the field of flight fare analysis and forecasting.

4 Architecture

4.1 Existing Architecture

In the realm of flight fare analysis, traditional methodologies frequently turn to tried-and-tested regression techniques to forecast ticket prices as shown in Fig. 1. These techniques, such as multiple linear regression and logistic regression, are adept at establishing relationships between historical data and various influencing factors. These factors span

a wide range, encompassing variables like booking timeframes, flight durations, route popularity, seasonal trends, and the competitive landscape among airlines [5].

By delving into these correlations and estimating coefficients for each predictor, regression models offer valuable insights into the expected fare for a specific flight. This analytical prowess empowers airlines to navigate the intricacies of pricing strategies and revenue management with greater precision and foresight. Despite their straightforward nature, these regression-based approaches provide interpretable and actionable insights into the complex interplay of factors driving ticket prices.

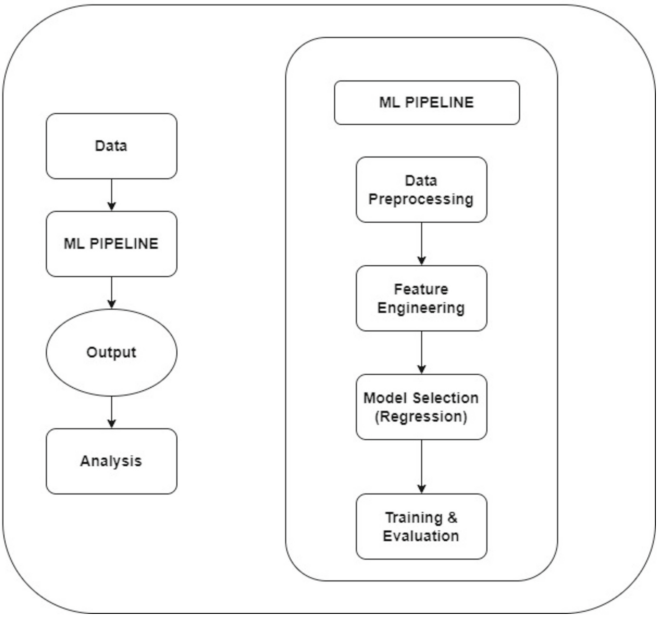

Fig. 1. Existing Architecture

In an industry as dynamic and competitive as aviation, the ability to anticipate and respond to market fluctuations is paramount. Hence, regression-based flight fare analysis serves as a cornerstone for airlines striving to optimize their pricing strategies and maximize profitability. Furthermore, these methodologies enable airlines to adapt to all threats, thereby ensuring resilience and agility in an ever-evolving landscape. Moreover, the robustness and versatility of regression techniques make them invaluable tools for airlines of all sizes and market positions. Whether it's a legacy carrier or a budget airline, the insights gleaned from regression analysis can inform strategic decision-making across various facets of the business, from pricing and revenue management to marketing and route planning. Overall, while regression-based approaches may seem conventional, their enduring relevance and effectiveness in flight fare analysis underscore their importance in an industry where precision and profitability go hand in hand. As airlines continue to navigate the complexities of an increasingly competitive market, leveraging

regression techniques remains a cornerstone for staying ahead of the curve and driving sustained success in the dynamic world of aviation.

4.2 Proposed Architecture

In contrast to contemporary architectures lacking considerations for seasonality, trend, and cyclicality, the proposed methodology segregates data based on airlines operating on a route, subsequently inputting this segmented data into distinct SARIMAX Time Series machine learning models. The resulting outputs from each model undergo comprehensive analysis to generate actionable insights. For a clearer understanding of Fig. 2, let's consider the route from Bangalore to Delhi. In this scenario, the data is partitioned according to the airlines servicing this route, such as Air India, Air Asia, IndiGo, Vistara, etc. Each subset of data is then fed into individual SARIMAX models, yielding outputs that are subsequently analyzed. The results provide insights into the lowest prices along the route, incorporating indications of the respective airlines associated with these prices. By dividing the data according to airlines, the model accounts for variations in pricing strategies, customer demand, and service offerings. Enhancing the accuracy of predictions through this granular approach empowers stakeholders to make more informed decisions regarding ticket purchases and route planning. Furthermore, the SARIMAX models are equipped to capture and analyze the intricate patterns inherent in airline ticket pricing, including seasonal fluctuations, long-term trends, and cyclical variations. This sophisticated analysis empowers airlines to optimize their revenue management strategies while providing travelers with valuable information to plan their journeys more effectively. The generated advice based on the analysis offers actionable recommendations for both airlines and travelers. Airlines can adjust pricing strategies,

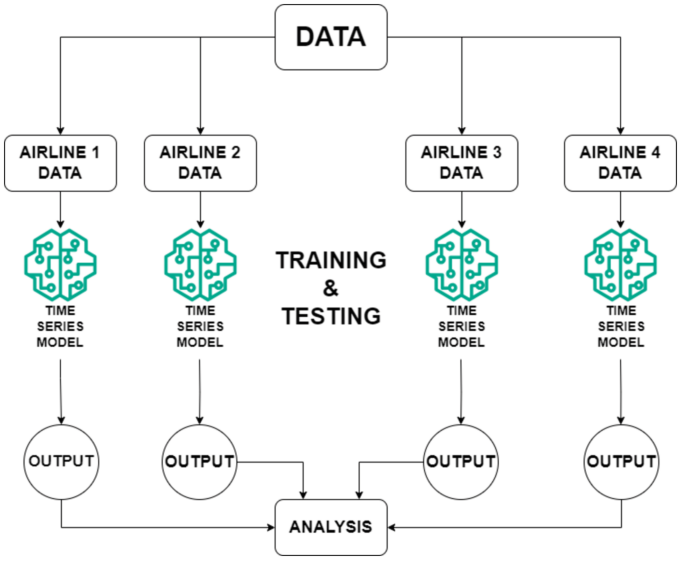

Fig. 2. Complex Architecture

promotional offers, and capacity planning based on the insights gleaned from the model outputs. Meanwhile, travelers can benefit from timely information on optimal ticket prices, enabling them to secure the best deals for their trips. Overall, the proposed architecture revolutionizes flight fare analysis by incorporating machine learning techniques tailored to the complexities of airline pricing dynamics. By leveraging SARIMAX models and a segmented approach based on airlines, the methodology provides a robust framework for uncovering optimal ticket prices on dynamic routes, ultimately enhancing the efficiency and profitability of the aviation industry. To explain Fig. 2 clearly let's take an example of the route Bangalore to Delhi in which the data is divided with respect to airlines were Air India, Air Asia, Indigo, Vistara, etc. Divided data set is fed to an individual SARIMAX model and output is generated and results with an analysis which showcases the least price in route with an indication of the airline also into it.

5 Time Series Analysis

Time series analysis stands as a potent method used to analyze sequential data recorded or stored at specific time intervals, arranged chronologically. Its application extends across various domains, including economics, commerce, and finance, where it plays a pivotal role in forecasting and decision-making processes. Particularly in the face of natural disasters or emergencies, time series analysis assumes even greater significance, aiding in understanding trends, patterns, and potential impacts on economic and societal landscapes. Its ability to uncover insights from historical data makes it an indispensable tool for businesses and policymakers alike, enabling them to anticipate and mitigate risks, optimize strategies, and make informed decisions in dynamic environments. Moreover, time series analysis serves as a crucial component in predictive modeling and risk management, offering valuable insights into future trends and behaviors based on past data patterns. By leveraging techniques such as analysts can extract meaningful information from time series data using ARIMA and SARIMA allowing for more accurate forecasts and scenario planning. In addition to its application in traditional economic and financial contexts, time series analysis is increasingly being utilized in fields such as healthcare, meteorology, and environmental science. For instance, in healthcare, it aids in forecasting patient admissions, resource allocation, and disease outbreak predictions. Similarly, in meteorology, time series analysis helps in weather forecasting, climate modeling, and studying long-term [6] climate trends. Overall, the versatility and efficacy of time series analysis make it an indispensable tool for understanding and navigating complex systems characterized by sequential data. Utilizing its knack for distilling valuable insights from historical data, it empowers decision-makers to optimize resource allocation and make informed choices, proactively managing risks in an ever-evolving landscape. Features of time-series are: trend, seasonality, cyclicality, white noise.

In time-series forecasting, data stationarity is essential for accurate predictions and meaningful insights. Stationarity implies that the statistical characteristics of the data remain consistent over time, including parameters such as the mean, variance, and autocorrelation. Precisely, stationary data exhibits a consistent mean, variance, and autocorrelation structure throughout the time series. Achieving stationarity frequently involves preprocessing steps to stabilize the data, where techniques like differencing or transformation are employed to eliminate any existing trends or seasonality. By ensuring

data stationarity, analysts can enhance the reliability and precision of their time-series forecasts. Stationary data provides a stable foundation for forecasting models, allowing for more accurate predictions of future trends and patterns. This, in turn, enables more informed decision-making across various domains, from economics and finance to healthcare. Furthermore, stationarity simplifies the modeling process and improves the interpretability of results. Models built on stationary data are easier to understand and validate, as they capture the underlying patterns and relationships more effectively. This facilitates the identification of meaningful insights and the development of actionable strategies based on the forecasted outcomes. In summary, data stationarity is a top prerequisite for successful time series analysis. By stabilizing the statistical properties of the data, analysts can unlock the full potential of forecasting models and generate valuable insights to support decision-making processes in diverse fields [7].

Within a time series, the trend can either ascend, descend, or remain steady. If the time series is stationary, the trend remains constant. Put simply, variance represents the average deviation of the data points from the zero line on a graph depicting the varying quantity over a specified time interval. In time series analysis, the trend within a time series can either be upward, downward, or constant. In the case of a stationary time series, however, the trend must remain constant. Essentially, variance denotes the average deviation of data points from the zero line on a graph, considering a specific time interval. It signifies how spread out the data points are around their mean value as they change over time. In essence, variance quantifies the degree of dispersion or variability within the time series data, providing insights into its stability and predictability. When non-stationary data is used as input for a forecasting model, the resulting predictions can be highly inaccurate. To tackle this problem, non-stationary data is frequently converted in to a stationary format through methods like differencing or applying logarithmic transformations to the series. Differencing is the most commonly employed method for achieving stationarity in time series analysis. This process involves subtracting each data point from its preceding point, effectively removing the trend and stabilizing the variance over time. By transforming non-stationary data into a stationary form, analysts can ensure the reliability and accuracy of their forecasting models, leading to more robust predictions and informed decision-making [8].

6 SARIMAX

The SARIMAX extends the capabilities of the ARIMA model by integrating supplementary components. Unlike ARIMA, SARIMAX can accommodate both seasonal patterns and exogenous variables. This renders it a versatile instrument for both time series analysis and forecasting purposes. By incorporating exogenous factors, SARIMAX allows analysts to account for external influences that may impact the time series data. These factors could include economic indicators, weather variables, or any other external variables that are known to affect the time series under consideration. By including these additional variables in the modeling process, SARIMAX enables more accurate and robust forecasts, as it captures the complex interactions between the time series and its external drivers. Moreover, SARIMAX retains the ability of the [9] ARIMA model to capture the autoregressive and moving average components of the time series, along

with the integration of differencing to achieve stationarity. This comprehensive approach makes SARIMAX well-suited for modeling time series data with both seasonal patterns and exogenous influences, offering analysts a powerful framework for understanding and forecasting complex temporal dynamics. In summary, the SARIMAX model represents a significant advancement in time series analysis, providing analysts with a flexible and comprehensive tool for modeling and forecasting. By incorporating both seasonal patterns and exogenous variables, SARIMAX enables more accurate and insightful predictions in a wide range of domains, facilitating better decision making.

6.1 ARIMA Model

ARIMA, short for AutoRegressive Integrated Moving Averages stands as a cornerstone in time series analysis. This statistical model utilizes time series data to provide insights or forecast future values. ARIMA amalgamates (AR) and (MA) elements, employing differencing to transform non-stationary data into a stationary form. By encompassing these elements, ARIMA offers a comprehensive framework for analyzing and predicting time series data, making it a versatile tool for various forecasting tasks across different domains. The (AR) aspect within ARIMA captures the correlation between a current observation and its lagged counterpart in the series. By incorporating past values, it enables the model to grasp trends and patterns within the data. Conversely, the (MA) element models the connection between an observation and the residual error derived from a moving average model applied to lagged observations. This component aids in attenuating random fluctuations or noise within the dataset. Furthermore, ARIMA incorporates differencing to transform non-stationary data into a stationary form. Stationarity is crucial for time series analysis as it ensures that the statistical properties of the data remain constant over time. By differencing the data, ARIMA removes trends and seasonality, allowing for more accurate modeling and forecasting. Overall, [10] ARIMA's ability to capture both the AR and MA components, along with its capability to handle non-stationary data through differencing, makes it a powerful and versatile tool for time series analysis and forecasting. Its widespread application across various domains underscores its importance in understanding and predicting temporal data patterns.

6.2 Auto Regression (AR)

In essence, ARIMA is a model that relies on its own past values, using regression to predict future outcomes. Equation 1 represents the Autoregressive (AR) model, where (y(t)) denotes the current value, (c) represents a constant term. White noise (ϵ_t) is a sequence of uncorrelated measurements, where each value is independent of previous values and ϕ_t represents weight [11]

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots\cdots\cdots + \phi_p y_{t-p} + \epsilon_t(1) \tag{1}$$

6.3 Integrated (I)

It denotes certain methods wherein the original data undergoes differencing to achieve stationarity, ultimately replacing the old non-stationary data with the new stationary one.

This process involves taking the difference between consecutive observations in the time series, effectively removing trends and seasonality [12]. By transforming the data into a stationary form, ARIMA can better capture the underlying patterns and relationships present in the time series, leading to more accurate and reliable forecasts. Differencing plays a crucial role in preparing the data for analysis and is a key step in the ARIMA modeling process.

6.4 Moving Averages (MA)

It signifies an equation that regresses the values of a time series against previous shock values of the same time series.

$$y_t = c + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots\cdots\cdots + \theta_q \epsilon_{t-q} \tag{2}$$

Equation 2 represents the MA model, where (y(t)) denotes the current value, (c) represents a constant term. White noise (ϵ_t) is a sequence of uncorrelated measurements, where each value is independent of previous values and θ_t represents weight. Equations 1 and 2. depict the equations that illustrate the relationship between specified data and its previous values over time, with an order of q representing the moving average component. The three integral components of ARIMA serve distinct functions [13].

p: An integer value determining the number of lagged values to be regressed in the AR model.

- d: An integer value indicating the number of times the differencing technique is applied to the data to achieve stationarity.
- q: An integer value determining the number of shock terms over time used in the MA model.

$$\hat{y}_t = \mu + \sum_{i=1}^{p} a_i y_{t-i} + \epsilon_{t-j} + \epsilon_t \tag{3}$$

$$\epsilon_t = \sqrt{\sigma_t z_t}, \sigma^2 = \omega + \sum\nolimits_{i=1}^{p} \alpha_i \epsilon_{t-i.}^2 + \sum\nolimits_{i=1}^{q} \beta_i \sigma_{i-1}^2. \tag{4}$$

Equations 3 and 4 outline the equations of the ARMA model, a time-series model integrating regression on past values and shock terms. Although ARIMA fore-casting necessitates stationary values, instances arise where non-stationary data is essential based on input. To convert stationary data back to a non-stationary state, differencing is commonly employed, often utilizing the cumulative sum technique. The number of times differencing is applied to achieve stationarity before feeding the data to the model determines how many times the cumulative sum must be applied to restore non-stationarity in the outcome. In another approach, the logarithmic method can be applied, utilizing an exponential function. The frequency of employing the logarithmic method to achieve stationarity before feeding the data into the model will dictate how many times the exponential function must be utilized to revert to non-stationarity in the outcome [14].

6.5 Seasonal (P, D, Q, S)

Seasonal AutoRegressive (SAR) component (P). Similar to the AR component but applied to the seasonal part of the time series.

Seasonal Integrated (D). Analogous to the 'd' parameter, but implemented on the seasonal component of the time series [15].

Seasonal Moving Average (SMA) component (Q). Similar to the MA component but applied to the seasonal part of the time series.

Seasonal Period (s). This represents the number of observations per season. For example, if the data has a yearly seasonality, 's' would be 12 for monthly data.

6.6 Exogenous (X)

Exogenous variables, in the context of time series analysis, play a crucial role in enriching the predictive capabilities of models. These variables represent external factors that are not directly part of the time series being analyzed but may still have a significant influence on its behavior. By incorporating exogenous variables into the modeling process, analysts can capture additional sources of variability and improve the accuracy of their forecasts. The inclusion of exogenous variables allows the model to account for factors such as economic indicators, weather conditions, marketing campaigns, or policy changes that could impact the observed time series. For example, in the case of forecasting sales data, exogenous variables could include advertising expenditure, competitor pricing, or consumer sentiment indices. One of the key advantages of incorporating exogenous variables is the ability to better capture and explain the underlying drivers of the time series. This can lead to more robust and interpretable models, as well as more accurate forecasts. Additionally, by considering external factors that may influence the time series, analysts can gain deeper insights into the dynamics of the system and make more informed decisions. However, it's important to note that the selection and inclusion of exogenous variables require careful consideration. Analysts must assess the relevance and impact of each variable on the time series, as well as any potential relationships or correlations with other variables. Additionally, the availability and quality of exogenous data may vary, which can pose challenges in model development and validation [16, 17].

In summary, exogenous variables offer a powerful means of enhancing time series analysis and forecasting by capturing external influences on the data. By carefully incorporating and modeling these variables, analysts can improve the accuracy, interpretability, and usefulness of their models, ultimately leading to better decision-making and outcomes. The SARIMAX model is especially useful for time series data demonstrating both a trend and seasonality, and when external factors (exogenous variables) need to be considered. It is widely used in forecasting applications where understanding and capturing seasonality, trends, and external influences are essential for accurate predictions.

7 Data Preparation

The data, collected from diverse sources, is time stamped and spans from 2019 to 2020, with a gap due to flight cancellations during the COVID-19 pandemic, followed by data from 2022 to 2023. This interruption in data collection reflects the impact of external events on the aviation industry, underscoring the need to account for such disruptions in analysis and forecasting. Organized according to specific flight routes and airline combinations, the dataset offers a comprehensive view of the dynamic landscape of air travel. For example, flights from Bangalore to Delhi are operated by IndiGo and Jet Airways, while the Delhi to Kochi route includes services from Jet Airways and multiple carriers. Similarly, Air India and Jet Airways service the Kolkata to Bangalore route. Table 1 and Table 2 provides a sample dataset illustrating the data for both Jet Airways and multiple carriers on the Delhi to Cochin route. Table 3 and Table 4 showcases data related to Air India and Jet Airways for the Kolkata to Bangalore route, while Table 5 and Table 6 presents data concerning IndiGo and Jet Airways for the Bangalore to New Delhi route. This meticulous categorization of data facilitates a granular analysis of flight fares and trends specific to each route and airline combination. Stakeholders can leverage this organized dataset to discern patterns, seasonal variations, and competitive dynamics, thereby making informed decisions regarding pricing strategies, revenue management, and operational planning. By incorporating historical data from pre- and post-pandemic periods, analysts gain insights into the resilience and adaptability of the aviation industry in response to unprecedented challenges. This comprehensive dataset serves as a valuable resource for understanding the evolving dynamics of air travel and navigating future uncertainties with greater agility and foresight. The dataset used in this research has been sourced from two primary channels: Kaggle and web scraping. Kaggle, a renowned platform for data science competitions and datasets, provides curated and readily available data for analysis. Complementing this, web scraping techniques have been employed to extract data from websites, broadening the dataset's scope and accessing information not available through traditional means. By combining data from both sources, researchers can construct a comprehensive dataset offering diverse insights. However, ethical considerations such as proper attribution of sources must be maintained, especially when utilizing web scraping methods to gather data from external websites.

Table 1. Delhi to Cochin - JetAirways

Date	Airlines	From	To	Fare
2019-01-01	Jet Airways	Delhi	Cochin	17024
2019-01-02	Jet Airways	Delhi	Cochin	1879
2019-01-03	Jet Airways	Delhi	Cochin	19828
2019-01-04	Jet Airways	Delhi	Cochin	17234

Table 2. Delhi to Cochin - Multiple carriers

Date	Airlines	From	To	Fare
2019-01-01	Multiple carriers	Delhi	Cochin	36983
2019-01-02	Multiple carriers	Delhi	Cochin	23533
2019-01-03	Multiple carriers	Delhi	Cochin	29528
2019-01-04	Multiple carriers	Delhi	Cochin	23170

Table 3. Kolkata to Bangalore - Air India

Date	Airlines	From	To	Fare
2019-01-01	Air India	Kolkata	Banglore	6535
2019-01-02	Air India	Kolkata	Banglore	13885
2019-01-03	Air India	Kolkata	Banglore	4435
2019-01-04	Air India	Kolkata	Banglore	4960

Table 4. Kolkata to Bangalore - Jet Airways

Date	Airlines	From	To	Fare
2019-01-01	Jet Airways	Kolkata	Banglore	14231
2019-01-02	Jet Airways	Kolkata	Banglore	14321
2019-01-03	Jet Airways	Kolkata	Banglore	10031
2019-01-04	Jet Airways	Kolkata	Banglore	13759

Table 5. Bangalore to Delhi - Indigo

Date	Airlines	From	To	Fare
2019-01-01	Indigo	Banglore	New Delhi	13302
2019-01-02	Indigo	Banglore	New Delhi	8153
2019-01-03	Indigo	Banglore	New Delhi	9694
2019-01-04	Indigo	Banglore	New Delhi	14306

7.1 Ad-Fuller Test

The Augmented Dickey-Fuller test stands as a cornerstone in time series analysis, widely utilized to determine whether the data under study is stationary or not. It serves as the most commonly employed test in procedures involving time series analysis, offering

Table 6. Bangalore to Delhi - Jet Airways

Date	Airlines	From	To	Fare
2019-01-01	JetAirways	Banglore	New Delhi	22270
2019-01-02	JetAirways	Banglore	New Delhi	26890
2019-01-03	JetAirways	Banglore	New Delhi	26890
2019-01-04	JetAirways	Banglore	New Delhi	25735

specific values that aid in decision-making. When conducting the Augmented Dickey-Fuller test, if the obtained value falls below a certain threshold, it suggests that the data is stationary. This pivotal test can be executed conveniently by importing the function from the "statsmodels.tsa.stattools" module, streamlining the analysis process [18].

However, if the Augmented Dickey-Fuller test indicates that the data is non stationary, further steps are necessary to transform it into a stationary state. This is typically achieved through techniques such as differencing or taking the logarithm of the data. These transformations are iteratively applied until the Augmented Dickey-Fuller test verifies that the data has achieved stationarity. The number of times these transformations are applied is counted, representing the parameter "d" in the SARIMAX model.

Once the data is deemed stationary, the next crucial step involves training the SARIMAX model using the transformed data. This enables accurate forecasting and analysis, empowering analysts to make informed decisions based on reliable insights derived from the time series data. The SARIMAX model leverages the stationary data to capture intricate patterns, seasonal variations, and dependencies within the time series, thereby enhancing the precision and reliability of forecasts. By following this systematic approach, analysts can extract valuable insights and uncover actionable trends, contributing to more effective decision-making processes across various domains.

8 Training SARIMAX Model

The SARIMAX model, short for Seasonal Autoregressive Integrated Moving Average with eXogenous regressors, is a robust tool for forecasting time series data. It integrates autoregressive, integrated, and moving average components, with the added capability of accommodating exogenous variables. This model is especially effective for analyzing time series data characterized by seasonality, trend, and external influences on observed patterns [19].

Figure 3 shows the graph plot of the data and SARIMAX proves effective for analyzing time series data, so it's crucial to organize your dataset as a chronological sequence. Identify any inherent seasonality, trends, or patterns that the model should capture. To meet SARIMAX's assumption of stationarity, consider applying differencing if your time series is non-stationary. Determine the appropriate orders (p, d, q) by analyzing autocorrelation and partial autocorrelation plots.

If seasonality is present, ascertain the seasonal orders (P, D, Q, and m). SARIMAX allows the incorporation of exogenous variables, which can influence the time series.

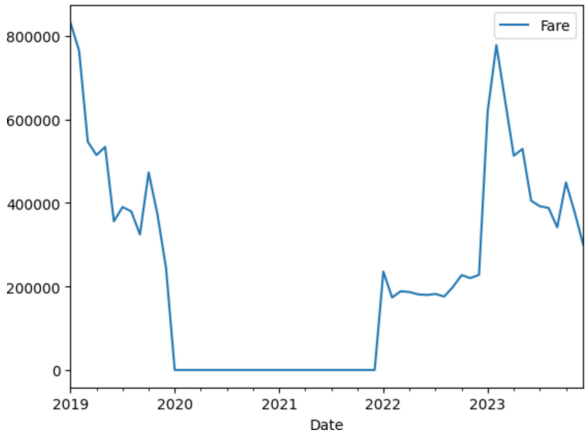

Fig. 3. Graph Plot

Identify and prepare these variables for integration into the model. Next, partition your dataset into training and validation sets. Utilize the training set to train the SARIMAX model, refining parameters and incorporating exogenous variables as needed. Assess the model's performance on the validation set using relevant metrics such as Mean Squared Error (MSE) or Mean Absolute Error (MAE). After the model is trained and validated, employ it to predict future periods by supplying exogenous variables. In cases where the model's performance is unsatisfactory, refine parameters or consider additional features to enhance accuracy. Validate the final model on a separate test set to ensure its generalization to new, unseen data. Finally, interpret the model coefficients and diagnostics to gain insights into the relationships and dynamics captured by the SARIMAX model [20].

8.1 Experimentation

During the experimentation phase, rigorous adjustments were made to the hyperparameters by systematically varying their values. This iterative process aimed to identify the optimal parameters that would yield the best scores for both the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), with a preference for minimal values for both metrics. Achieving optimal scores for AIC and BIC is crucial as they serve as indicators of the model's goodness of fit and parsimony. The determination of the parameters (p), (q), (P), (Q), (d), and (D) was carried out through a thorough analysis of Partial Autocorrelation (PACF) and Autocorrelation (ACF) plots. These plots provided valuable insights into the temporal dependencies and seasonal patterns present in the time series data, guiding the selection of appropriate values for the SARIMAX model parameters.

For the SARIMAX model, both the normal hyper parameters and seasonal hyper parameters underwent fine-tuning. This comprehensive approach ensured that the model could effectively capture both the non-seasonal and seasonal variations in the data. Once the optimal parameters were determined, the data was fed into the SARIMAX model,

incorporating the tuned hyper parameters, including both normal and seasonal ones. To assess the performance of the SARIMAX model, a portion of consecutive data was reserved for independent testing. This held-out dataset allowed for the evaluation of the model's predictive accuracy and generalization capabilities on unseen data. By comparing the model's forecasts with actual observations from the test dataset, analysts could validate the model's effectiveness and identify any potential areas for improvement. Overall, this systematic approach to hyper parameter tuning and model evaluation ensures that the SARIMAX model is robust and reliable for forecasting time series data. By leveraging advanced statistical techniques and rigorous experimentation, analysts can develop accurate and actionable forecasts to support decision-making processes in various domains.

9 Auto ARIMA Function

In the context of SARIMAX modeling, the 'auto_arima' function from the 'pmdarima' library serves as a powerful tool for automating the hyperparameter tuning process. SARIMAX with eXogenous regressors, extends the capabilities of the ARIMA model by enabling the incorporation of exogenous variables. This feature is particularly useful when additional factors beyond the time series data itself need to be considered in the forecasting process. In this code snippet, the 'auto_arima' function performs a grid search over various combinations of orders and seasonalities, selecting the combination that minimizes a specified criterion, such as AIC or BIC. This automated search process saves time and effort compared to manual hyperparameter tuning. Additionally, the resulting 'sarimax_model' object contains the best-fitting SARIMAX model based on the provided dataset, ready for further analysis and forecasting. However, it's important to note that the choice of evaluation criteria, such as AIC or BIC, may impact the model selection. Therefore, it's advisable to review the model's performance and adjust parameters accordingly based on specific needs and domain knowledge. Additionally, thorough validation and testing of the model against independent datasets are essential to ensure its robustness and reliability in real-world applications. By leveraging the 'auto_arima' function and carefully evaluating the results, analysts can develop accurate and effective SARIMAX models for forecasting and decision-making purposes [21].

10 Evaluation

In the final stages of model evaluation, Figs. 4, 5, 6, 7, 8, 9 depict plots showing the Test, Predicted Data, and Training Data. These visualizations offer a comprehensive insight into the alignment between the model's predictions and the actual data across time, aiding in an intuitive comprehension of the model's performance. For quantitative assessment of the model's performance, an evaluation metric like the r2 score is utilized. These metric gauges the proportion of variance in the dependent variable that can be foreseen from the independent variables. A score nearing 1 indicates a superior fit of the model to the data, whereas a score of 0 implies that the model doesn't account for any variability in the dependent variable.

In our evaluation, a portion of the entire dataset, typically ranging from 20 to 30%, is designated as test data. This test data is isolated from the training dataset and used to evaluate the model's predictions. By comparing the predicted data from the model with the actual test data, we can calculate the r2 score as a measure of the model's predictive accuracy. For the SARIMAX model under consideration, the obtained r2 score is 0.7284. This score, greater than zero and closer to 1, indicates that our model fits well and effectively predict future data based on past outcomes. However, it's important to interpret this score in the context of the specific domain and the requirements of the forecasting task. Further analysis and validation may be necessary to ensure the robustness and reliability of the model in real-world scenarios.

Overall, Figs. 4, 5, 6, 7, 8, 9 shows the graphical representation of the trained and tested values with respect to the routes and airlines. And calculated r2 score, provide valuable insights into the performance of the SARIMAX model, enabling stakeholders to make informed decisions and adjustments as needed. Through this rigorous evaluation process, analysts can ensure the accuracy and effectiveness of the forecasting model, ultimately supporting better decision-making and planning efforts in various domains. Lastly, an evaluation metric is required to assess the model's performance. Therefore, we designate 20 or 30% of the entire dataset as test data and compare it with the predicted data from our model to obtain a score for the model. In this process, the last consecutive data points are isolated from the training dataset and employed to test the model.

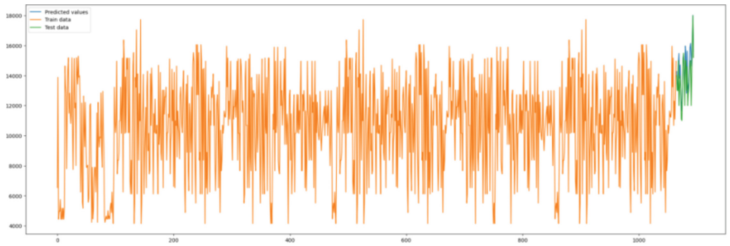

Fig. 4. JetAirways from Kolkata to Bangalore

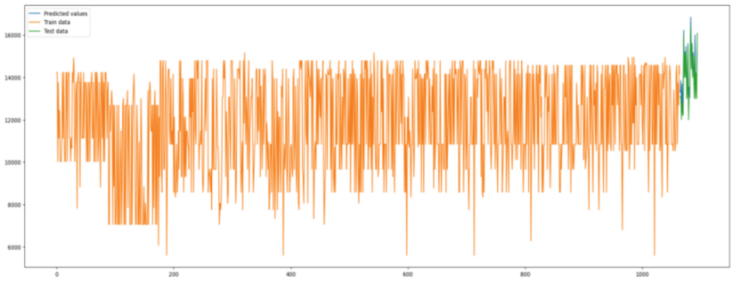

Fig. 5. MultiCarriers from Kolkata to Bangalore

The provided model accuracy scores, ranging from 0.51 to 0.85 based on the R2 Score, showcase the predictive performance of various airline routes analyzed in the

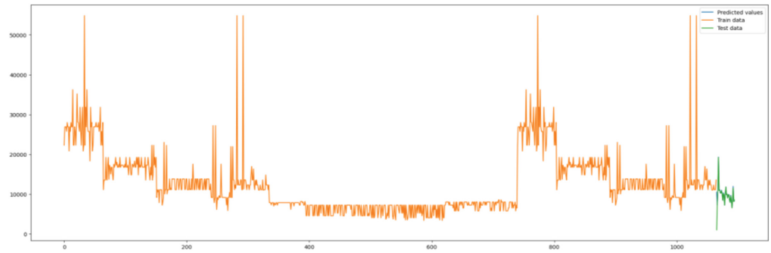

Fig. 6. Indigo from Bangalore to Delhi

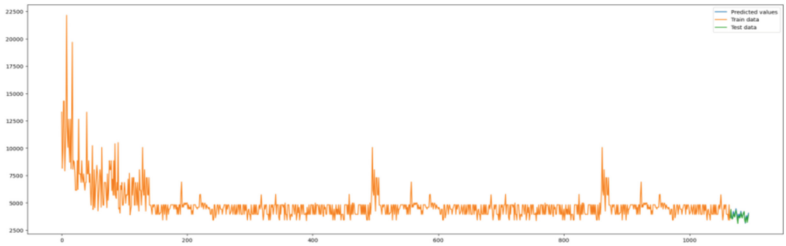

Fig.7. JetAirways from Bangalore to Delhi

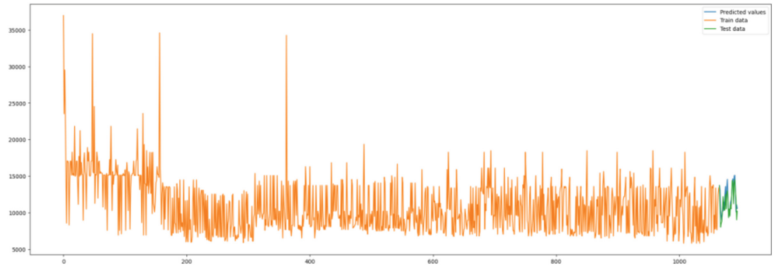

Fig. 8. Multiple Carriers from Delhi to Kochi

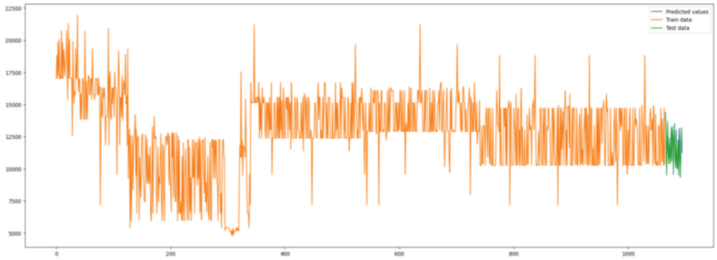

Fig. 9. JetAirways from Delhi to Kochi

study. Table 7. Shows the R2 Score for the Kolkata to Bangalore route indicates a moderate predictive capability for Air India, while Jet Airways demonstrates a significantly higher accuracy score, reaching 0.81. Similarly, for the Bangalore to Delhi route, Indigo exhibits a commendable R2 Score of 0.729, while Jet Airways scores slightly lower at 0.57. Transitioning to the Delhi to Kochi route, Multi carriers achieve an R2 Score of 0.63, with Jet Airways demonstrating the highest predictive accuracy at 0.85. The graphical representations provided in Figs. 4, 5, 6,7, 8, 9 offer visual insights into the comparison between actual and predicted values for each route, aiding in the interpretation and assessment of model performance. These findings contribute to a comprehensive understanding of the predictive capabilities of the analyzed models across different airline routes, facilitating informed decision-making for both airlines and travelers in the aviation industry.

Table 7. Models Accuracy

Model	Accuracy (R2 Score)
Kolkata to Bangalore - Air India	0.51
Kolkata to Bangalore - Jet Airways	0.81
Bangalore to Delhi - Indigo	0.729
Bangalore to Delhi - Jet Airways	0.57
Delhi to Kochi - Multi carriers	0.63
Delhi to Kochi - Jet Airways	0.85

10.1 Workflow

Figure 10 shows the workflow diagram begins with the acquisition of raw data encompassing flight fare information. This raw data is subsequently split into subsets based on specific routes, including Kolkata to Bangalore, Bangalore to Delhi, Delhi to Kochi, and further divided based on the operating airlines. These subsets are delineated into flights operated by Air India, Jet Airways, and IndiGo, as well as combinations such as Jet Airways and multi-carriers. Following the segmentation of data into these distinct subsets, each subset undergoes preprocessing procedures tailored to its respective characteristics. Data preprocessing involves steps such as handling missing values, stationarity check is conducted to ensure that the data is appropriately prepared for modeling.

Once the preprocessing steps are completed for each subset, the datasets are then fed into separate machine learning pipelines. These pipelines incorporate various machine learning algorithms and techniques to train and optimize predictive models. Within each pipeline, the datasets undergo feature selection, model training, hyperparameter tuning, and cross-validation to ensure robust performance. In the last phase of the workflow, the [21] SARIMAX (Seasonal AutoRegressive Integrated Moving Average with Exogenous Variables) model is utilized. Each preprocessed dataset is input into a customized SARIMAX model designed for the particular route and airline combination. Leveraging

time series analysis techniques, the SARIMAX model predicts future flight fares by analyzing historical data and incorporating exogenous variables.

Following the training of SARIMAX models on the preprocessed datasets, evaluation metrics are applied to gauge the performance of each model. Moreover, visualizations such as time series plots and forecast plots are generated to offer insights into the anticipated fare trends. Subsequently, a thorough analysis is conducted on the outputs derived from the SARIMAX models. This analysis entails comparing forecasted fares with actual fares, identifying data trends and patterns, and evaluating the influence of external factors on fare fluctuations. Insights gleaned from this analysis inform decision-making processes related to pricing strategies, revenue management, and operational planning within the aviation industry.

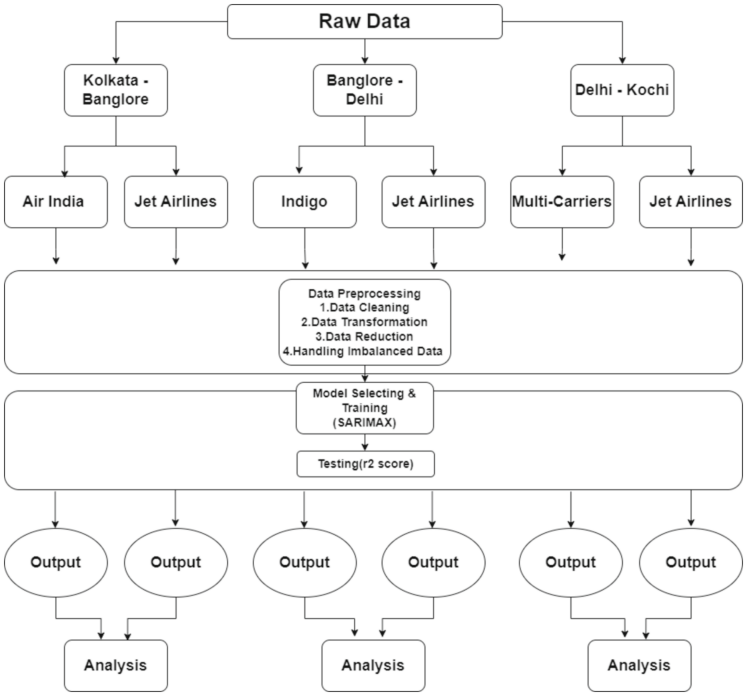

Fig. 10. Complete workflow diagram

11 Conclusion and Future Work

In conclusion, the utilization of SARIMAX models has proven highly effective for airfare analysis, successfully capturing the intricate temporal patterns inherent in airfare data. The automatic hyperparameter tuning facilitated by tools like the 'auto_arima' function has significantly enhanced the accuracy of forecasting, while the incorporation

of exogenous variables allows for a more comprehensive understanding of external influences on airfares. The model's ability to adapt to seasonality and dynamically changing conditions makes it a valuable asset for stakeholders in the aviation industry seeking precise and reliable predictions for airfare trends. For future work, there is ample room for refinement and expansion in the realm of airfare analysis using SARIMAX models. Exploring the integration of additional features and machine learning techniques could further boost the model's predictive power, capturing nuanced relationships and nonlinear patterns within the data. Continuous adaptation through dynamic model updating, ensemble modeling, and scenario analysis will ensure the SARIMAX approach remains resilient and responsive to the evolving dynamics of the air travel landscape, offering valuable insights for strategic decision-making within the aviation industry.

References

1. Kovvuri, A.R., Uppalapati, P.J., Bonthu, S., Kandula, N.R.: Water level forecasting in reservoirs using time series analysis–auto ARIMA model. In: International Conference on Cognitive Computing and Cyber Physical Systems, pp. 192–200 (2022)
2. Gokasar, I., Karakurt, A., Kuvvetli, Y., Deveci, M., Delen, D., Pamucar, D.: Sustainable regional rail system pricing using a machine learning-based optimization approach. Ann. Oper. Res., 1–28 (2023)
3. Nadeem, R., Sivakumar, T.: Flight fare forecasting: a machine learning approach to predict ticket prices. In: Chaki, N., Roy, N.D., Debnath, P., Saeed, K. (eds.) ICDAI 2023. LNNS, vol. 727, pp. 703–713. Springer, Singapore (2023). https://doi.org/10.1007/978-981-99-3878-0_60
4. Kaur, J., Parmar, K.S., Singh, S.: Autoregressive models in environmental forecasting time series: a theoretical and application review. In: Environmental Science and Pollution Research International, pp. 19617–19641 (2023)
5. Degife, W.A., Lin, B.-S.: Deep-learning-powered GRU model for flight ticket fare forecasting. Appl. Sci. **13**, 6032 (2023)
6. He, H., Chen, L., Wang, S.: Flight short-term booking demand forecasting based on a long short-term memory network. Comput. Ind. Eng. **186**, 109707 (2023)
7. Jiang, Y., Tran, T.H., Williams, L.: Machine learning and mixed reality for smart aviation: applications and challenges. J. Air Transp. Manag. **111**, 102437 (2023)
8. Naidu, J., Varma, M.A., Madhuri, P.S., Shankar, D., Matta, D.S., Ramya, S.: Enhancing heart disease prediction through a heterogeneous ensemble DL models, pp. 58–73 (2024)
9. Corselli, G.: Collusion and dynamic pricing - enablers and origins of airlines' cartels - webthesis. Polito.it (2023)
10. Dadoun, A., Defoin-Platel, M., Fiig, T., Landra, C., Troncy, R.: How recommender systems can transform airline offer construction and retailing. J. Revenue Pricing Manag. **20**, 301–315 (2021)
11. Guo, Y., Lai, X., Gan, M.: Cyanobacterial biomass prediction in a shallow lake using the time series SARIMAX models. Ecol. Inform. **78**, 102292 (2023)
12. Yassine, S., Stanulov, A.: A comparative analysis of machine learning algorithms for the purpose of predicting Norwegian air passenger traffic. Int. J. Math. Stat. Comput. Sci. **2**, 28–43 (2023)
13. Elshabrawy, M., Eid, M.M., Abdelhamid, A.A., El-Kenawy, E.-S.M., Ibrahim, A.: Forecasting of Monkeypox cases using optimized SARIMAX based model. In: 2023 3rd International Conference on Electronic Engineering (ICEEM), Menouf, Egypt, pp. 1–6 (2023). https://doi.org/10.1109/ICEEM58740.2023.10319521

14. Simaiya, S., et al.: A hybrid cloud load balancing and host utilization prediction method using deep learning and optimization techniques. Sci. Rep. **14**. https://doi.org/10.1038/s41598-024-51466-0
15. Wei, Y., Jang-Jaccard, J., Xu, W., Sabrina, F., Camtepe, S., Boulic, M.: LSTM-autoencoder-based anomaly detection for indoor air quality time-series data. IEEE Sens. J. **23**, 3787–3800 (2023)
16. Espinosa, R., Jiménez, F., Palma, J.: Multi-surrogate assisted multi-objective evolutionary algorithms for feature selection in regression and classification problems with time series data. Inf. Sci. **622**, 1064–1091 (2023)
17. Uppalapati, P.J., Gontla, B.K., Gundu, P., Hussain, S.M., Narasimharo, K.: A machine learning approach to identifying phishing websites: a comparative study of classification models and ensemble learning techniques. EAI Endorsed Trans. Scalable Inf. Syst. **10** (2023)
18. Khan, W.A., Chung, S.-H., Eltoukhy, A.E.E., Khurshid, F.: A novel parallel series data-driven model for IATA-coded flight delays prediction and features analysis. J. Air Transp. Manag. **114**, 102488 (2024)
19. Sohrabi, P., Shokri, B.J., Dehghani, H.: Predicting coal price using time series methods and combination of radial basis function (RBF) neural network with time series. Miner. Econ. **36**, 207–216 (2021)
20. Mokhtarimousavi, S.M., Mehrabi, A.: Flight delay causality: machine learning technique in conjunction with random parameter statistical analysis. Int. J. Transp. Sci. Technol. **12**, 230–244 (2023)
21. Maheswara Rao, V.V.R., Nrusimhadri, S., Gadiraju, M., Reddy, S., Bonthu, S., Kurada, R.R: A plausible RNN-LSTM based profession recommendation system by predicting human personality types on social media forums, pp. 850–855. https://doi.org/10.1109/ICCMC56507.2023.10083557