Data Visualization and Prediction Model Analysis

Ziyi Hua

University of Mount Saint Vincent, New York City, 10471, The United States of America

Abstract. Data visualization plays an important role in the process of data analysis. It can transform complex data into intuitive charts, which helps us understand data more effectively. The purpose of this paper is to study the application of data visualization and compare different prediction models. In this work, I present three different prediction models and apply them on the US airline industry datasets by using Python as the programming language. Then these data sets are made into a bar graph and analy1zed through these three prediction models. Finally, this paper provides some suggestions on how to choose appropriate visualization tools based on data characteristics. This study not only enriches the application examples of data visualization, but also provides valuable reference for decision-making and predicting in data analysis.

1 Introduction

1.1 Previous research on data visualization and data prediction

Some of the data visualization and graphing work are heavy tasks and always have high computational requirements [1]. Data visualization can represent complex and difficult data through graphics, making the information more intuitive for people to understand and compare. It can help observers quickly identify trends, distribution states, and outliers in data, thereby promoting better decision-making. Data visualization is widely used in research for exploring quantitative data, validating hypotheses, and communicating results, primarily using statistical charts such as bar charts, line charts, or scatter plots [2]. And data visualization get more important in nowadays society. In the context of business intelligence, DV is essential for exploring, analyzing, and understanding vast amounts of information, which in turn helps make informed decisions to improve business efficiency. As a result, it becomes more important in organizations over time [3]. There are many recent developments in data visualization. For example, the integration of artificial intelligence and data visualization can intelligently select visualization methods. Real-time data visualization can meet the market's demand for dynamic data. Virtual reality and augmented reality technologies will also rely on data visualization. In terms of decision-making, data visualization has also become an important means in business intelligence and data analysis, which can help companies make decisions. In the fields of medical care and health, it also

Corresponding author email: zhua.student@umsv.edu

helps patients and doctors better coordinate. Data prediction is very important in helping people identify how to exploit them in the context of enhancing business opportunities. They can perform business analysis, such as how to use it in various aspects of the organization to solve problems or improve its results. Data predicting is grounded on the instance of a predictive model and employing it to anticipate future outcomes [4].

1.2 The Role of predicting model and Data Visualization

Predictive models and data visualization play an important role in transforming raw data into actionable data. Predictive models are able to predict future trends based on historical data using statistical algorithms and machine learning techniques. By identifying patterns in the data, these models can effectively provide predictions for reference. For example, predictive models can predict and detect fraudulent activities or predict customer behavior, thereby avoiding some malicious events. Building these models requires the use of techniques such as regression analysis, classification, and time series forecasting. Each predictive model is generally designed specifically for a specific type of data and forecasting task. Data visualization generally displays data in charts to reveal insights, trends, and patterns that raw numbers cannot present. Data visualization tools present complex datasets in intuitive graphical formats, making it easier for users to spot trends and anomalies [5]. Visual visualization models can transform complex data into intuitive visual representations, making it easier for researchers to interpret information and take action. Such visualizations not only improve the clarity of data, but also help to communicate research results to a wider audience, thereby assisting decision-making. Predictive models analyze historical data to identify patterns and trends, enabling accurate forecasts and supporting decision-making [6].In business decision-making, predictive models provide data-driven forecasts, while data visualization transforms these predictions into easily understandable graphical representations [7]. Integrating predictive models with data visualization clearly displays potential risks and opportunities, improving strategic planning efficiency [8]. The predictions generated by predictive models can be visualized through dynamic charts and graphics, allowing users to grasp the meaning of the predictions. This interplay enhances the ability to make data-driven decisions, optimize operations, and identify new opportunities, ultimately driving success in fields ranging from finance to healthcare to marketing.

1.3 Disadvantages of data visualization

Data visualization has some disadvantages. First, if it is not designed properly, data visualization may mislead the observer, which may lead to wrong decisions and thus have a bad impact on the company. Second, the visual presentation of charts and other visuals may be too concise, which makes people ignore key details, thus affecting the integrity and accuracy of the data. Additionally, data visualization often demands sophisticated technology, with specific equipment and technical requirements for organizations. The creation of interactive visualizations frequently necessitates substantial technical expertise, which can hinder their effective implementation [9]. Finally, enterprises must keep the visualization updated at all times, which has high requirements for data storage and computing ability.

1.4 Principles of Data Visualization

The aim of visualization is to provide insights through interactive graphics into different elements of a process or phenomenon of interest, such as a scientific simulation or a real-world situation [10]. Here is the principles of data visualization. First, the model will collect

the raw data, and then remove some erroneous data through preprocessing operations such as data cleaning, conversion and integration. This will make the data available. Then, after selecting the appropriate visualization form, such as bar charts, line charts, pie charts, etc. The computer will use mathematical calculations to determine the coordinates, size, scale and other parameters of the graphics, so that the relationship between the data can be accurately reflected. Sometimes, people will pursue more visual effects, so graphics are needed to render technology to add a variety of different effects such as color, shadow, texture, etc. to these visualization elements to enhance the visual appeal. Under the combined effect of these technologies, data visualization can express massive and abstract data in a clear and easy-to-understand way, thereby helping users quickly obtain key information, grasp the laws of data and predict future trends.

1.5 The goal of this research

The purpose of this research is to use three prediction models that are used to predict data in a specific field, and the results are presented through data visualization. Finally, the results are compared to obtain a suitable model for predicting the data, thus filling the gap in previous researchers' research on the differences between different prediction models. This study selected relevant data values of American Airlines, selected ARIMA, Neural Network, and Linear forecasting models to visualize these forecast results, evaluated the benefits of different models, selected the best forecasting model and gave relevant suggestions.

1.6 The Basic methodology of this research

The core of the methodology of this study is to obtain appropriate data through reasonable channels, pre-process the data by using scientific strategies, convert the data format reasonably, and select a suitable visualization platform based on function and efficiency. And select a suitable model to analyze and present the data, so as to provide strong support for the research conclusions.

2. Prediction Model

The prediction model of data visualization can use algorithms to judge future trends by deeply exploring the potential laws contained in the data that has been presented. In economic activities, these models can help predict market dynamics and provide a scientific basis for corporate strategic planning and future decision-making. In public service areas such as traffic management and resource allocation, demand changes can be estimated to make resource allocation more reasonable. Common prediction models include linear regression models, ARIMA models and neural network models.

2.1 Linear Regression Model

The linear regression model is an analytical tool that is generally used in fields such as statistics. This model aims to reveal the linear relationship between the independent variable and the dependent variable. Usually this model will calculate a concise mathematical expression, and then fit the data to calculate the appropriate regression coefficient, so that a prediction equation can be constructed. For example, when it comes to the relationship between the number of aircraft and time, the linear regression model can help evaluate the impact of changes in the number of aircraft over a period of time. This prediction model is

simple and easy to understand, and because there is often only one formula, it is very intuitive. It can provide an intuitive basis for future decisions. But it also has disadvantages. It requires the processed data to conform to linear data as much as possible, and it is susceptible to interference from outliers. In general, linear regression models play an important role in fields such as economic forecasting.

2.2 ARIMA Model

ARIMA models, which stands for Autoregressive Integrated Moving Average models, are used for forecasting and analyzing time series data [11]. The ARIMA model is an advanced statistical model for time series analysis. It combines the concepts of autoregression, differencing, and moving average. This model can capture the interdependence of time series data on the time axis, which allows the model to perceive the historical impact of the data. The ARIMA model also has a difference link, which can effectively deal with non-stationary factors in the data. Through transformation, the data has the ability to be more stable, which adds more accuracy and reliability to the model. The ARIMA model can dig deep into the complex structure of time series data and capture the hidden dynamic laws and patterns. In this way, relatively accurate predictions of future development trends can be made.

2.3 Neural Network Model

The main function of a neural network lies in creating an output pattern in response to an input pattern [12]. For instance, pattern classification, which is one of the significant operations a neural network can execute, will be thoroughly analyzed. A neural network serves to produce an output pattern as a reaction to an input pattern. Among the various capabilities of a neural network, pattern classification stands out and will be described in great detail [12]. Neural network models learn complex data and discover patterns through connections and interactions between a large number of neurons. In prediction tasks, neural networks can identify and learn features in input data through multi-layer structures. Once the neural network model receives the data, it can produce prediction results after complex calculations and transformations in the hidden layer. The learning process of neural networks relies on the back propagation algorithm, which minimizes the error between the predicted value and the actual value by continuously adjusting the connection weights between neurons. In addition, deep neural networks capture richer and deeper information by increasing the number of layers. However, there are also some shortcomings in the theory of neural networks. For example, neural network models may fall into local optimal solutions during the training process, resulting in poor performance of the model. At the same time, this model also has the problem of overfitting, which has the impact of poor generalization ability of the model on new data and high computational cost. However, neural network models still play an important role in the field of prediction. Its most prominent advantage lies in its learning ability. It can handle complex nonlinear relationships and can fit and learn well even in the face of complex data. It can automatically extract features from raw data, has strong adaptability, and can also adjust the prediction results according to various data types and prediction tasks. It should be noted that this model has a complex and opaque internal mechanism, making it difficult to clearly explain the prediction generation process. This model has high data requirements. Once the data volume is insufficient, the prediction results will be very inaccurate. At the same time, its training process is computationally intensive, which will consume a lot of computing resources and time.

3. Research Design

3.1 Prepare Us Airline Industry Dataset through Kaggle

This dataset provides comprehensive details on U.S. airline flight routes, fares, and passenger volumes from 1993 to 2024. It includes key metrics such as origin and destination cities, distances between airports, passenger numbers, and fare information by airline carrier. This rich dataset is ideal for analyzing trends in air travel, pricing, and airline competition over three decades.

3.1.1 Data Cleaning

When processing data, missing values must be processed first. After checking the entire data set, the rows and columns containing missing values are directly deleted. Since the number of missing values is small, this operation has little impact on the overall data. In addition, there are some outliers. With the help of statistical methods, indicators such as the mean of the data can be calculated to determine the range of outliers and delete unreasonable data to avoid interference with subsequent analysis.

3.1.2 Data format conversion

Google Chart has certain requirements for the data set to be processed. So convert the data into a format supported by Google Chart. The format used in this study is CSV.

3.1.3 Data visualization platform

This study used the Googlechart platform for data visualization. Google Charts is a very useful free tool that allows you to create a variety of charts. Users can use it to create line charts, bar charts, pie charts, scatter charts, and many other charts. These operations can be achieved through simple code. It is particularly convenient because you can directly extract data from services such as Google Sheets without manually entering it. The charts generated by this platform are very sophisticated and also allow for interactivity, such as tooltips and zooming, which makes the data presentation more vivid. For the need for a free and easy-to-use data visualization tool, Google Charts is a good choice.

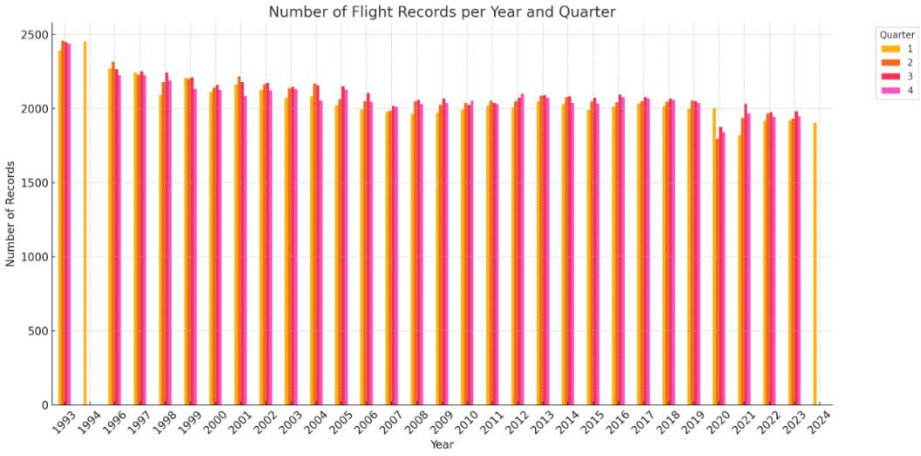

Fig. 1 The number of flight records

3.2 Analysis of visualization results

This chart, shown as Figure 1, shows the number of flight records per year and quarter. It provides a visualization of the fluctuations in flight records over time. From the perspective of time trend, the number of flight records fluctuates between 1993 and 2024. For example, in some years and quarters, there are more flight records, which can be inferred that they may be in certain economic booms or tourist seasons. In other periods, the number of these flight records is relatively small, which may be affected by negative factors such as the epidemic. From the specific perspective of each quarter, it can be inferred that these flight records have certain seasonal patterns. For example, the number of flight records in the quarter with a tourist peak season such as summer may be higher than in other quarters. However, from this chart alone, it is still difficult for us to accurately judge the specific growth or decline trend. To conduct a further in-depth interpretation, it is need to use some forecasting models.

4. Results

4.1 ARIMA process for analyzing flight record data

First, the given data is tested for stationarity. This is a very critical step. In this experiment, the ADF (Augmented Dickey-Fuller) test statistical method is used to determine whether the data is stationary. This ensures that the data is stable and the prediction effect is good. After completing the stationarity processing, start observing the autocorrelation function (ACF) and partial autocorrelation function (PACF) of the data. These two functions can provide characteristics about the internal correlation of the data, and it is useful to infer the order (p, d, q) of the ARIMA model. Here p represents the autoregressive order, d represents the difference order, and q represents the moving average order. After that, multiple ARIMA models are fitted based on the preliminarily determined order range. In this process, the AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) values of the models under different order combinations are compared. AIC and BIC are indicators used to measure model complexity and goodness of fit. The smaller the value, the better the model. After completing the model fitting, the obtained model is thoroughly checked and judged.

Focus on checking whether the residual of the model meets the white noise assumption. If it meets the white noise assumption, it means that the model has fully captured the effective information in the data. Otherwise, it means that the model is not perfect and needs to be adjusted. Finally, the selected ARIMA model is used to predict the number of future flight records. At the same time, the accuracy and reliability of the prediction results should be fully evaluated. The degree of deviation between the predicted value and the actual value can be measured by calculating, for example, the mean square error (MSE), so as to judge whether the prediction performance of the model meets the requirements.

4.2 Prediction results

This forecast chart predicts the trend of the US aviation industry in the next year. In terms of the number of passengers, there is a certain seasonal fluctuation overall, and the number of aircraft reaches a peak in the summer, which is consistent with the characteristics that summer is the peak season for tourism and the right time for business trips. Although there may be some missing values in the data, it does not affect the judgment of the overall trend. From the forecast results, it is logical as a whole. Overall, this forecast chart shows us the development trend of the US aviation industry and has a certain reference value for the planning and decision-making of the future aviation industry. At the same time, this chart shows that the highest data reached 8,000, and the lowest data was concentrated around 4,000, which is also in line with the actual situation and has a certain reference value.

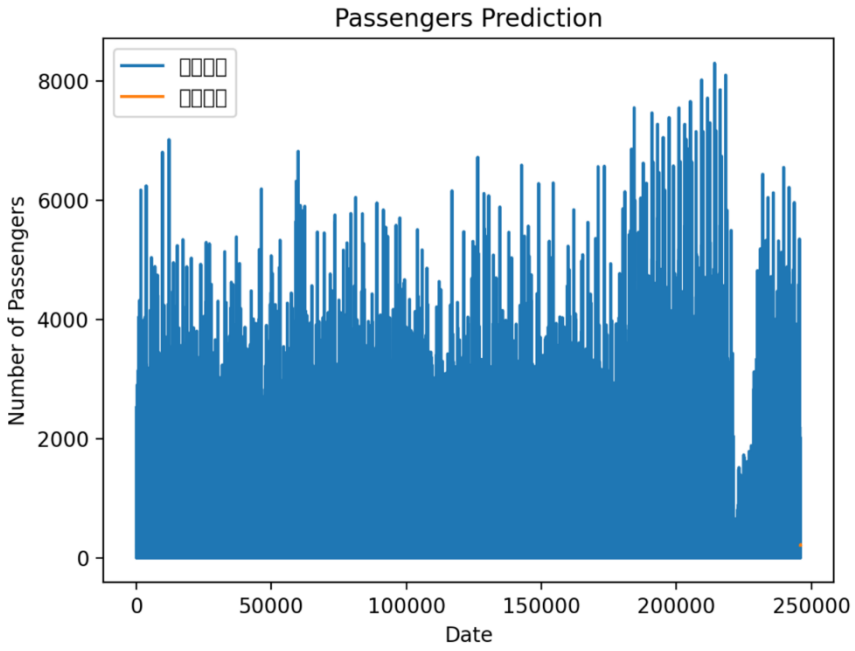

Fig. 2. The prediction result

4.3 The significance of prediction results

Three models are introduced above. For the time series data of the number of flight records per year and per quarter in this experiment, the ARIMA model may be a more appropriate choice. The reason is that this is data with obvious time sequence characteristics, and the ARIMA model can handle the characteristics of trend, seasonality and periodicity in time series data. Therefore, ARIMA can extract the rules in the data through the analysis of historical data and make predictions based on it. The practical significance of the forecast results of the ARIMA model has a very positive effect in various aspects. For example, for airlines, they can set and adjust the departure time and ticket prices of flights based on the forecast results, so as to optimize the company's resource allocation and improve operational efficiency. For the airport itself, these forecast results have a very positive impact on the deployment of personnel and the use of infrastructure. Of course, these results also have a considerable impact on decision-making. Managers responsible for decision-making can decide the distribution of capital investment based on the growth trend of flight records. For the government, there is a reasonable basis for taking measures and regulating the airport and its surroundings to ensure safety and stability. And the hidden market opportunities brought by these forecast results are huge. For example, industries related to tourism and the hotel industry can prepare for demand in advance. New entrants to the flight market can seize opportunities based on the forecast results, etc.

5. Model comparison

The ARIMA model focuses mainly on the linear characteristics and seasonality of time series, so when it processes data with relatively stable trends or data with obvious differences due to different seasons, it will produce relatively stable prediction results. When faced with sudden changes or unstable nonlinear patterns, it cannot handle problems sensitively, resulting in some deviations in the prediction results. Linear models often give simple prediction results. If the data itself conforms to the linear characteristics, this model can make reasonable predictions. However, when the relationship between the data is nonlinear, the prediction results of this model are not accurate because the linear model has limited ability to deal with such problems. Neural Model has strong learning and fitting capabilities, so it still has good prediction performance when dealing with nonlinear data relationships or data where the results are affected by multiple factors. Neural Model can dig out some deep-level features and relationships hidden in the data, so it can effectively deal with large-scale, high-dimensional and very complex internal structure data. But sometimes it also leads to overfitting problems due to excessive learning of noise in the data, which also makes the prediction effect poor.

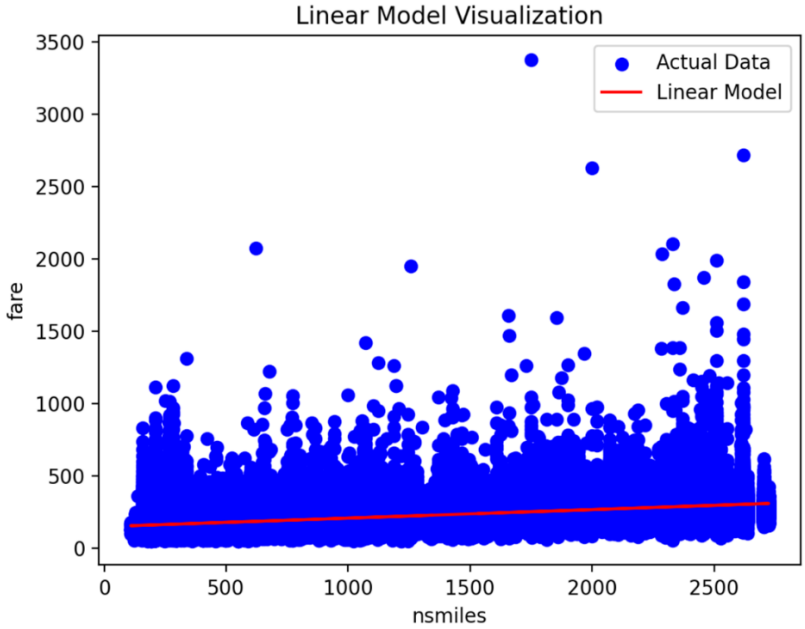

Fig. 3. Linear Model Prediction

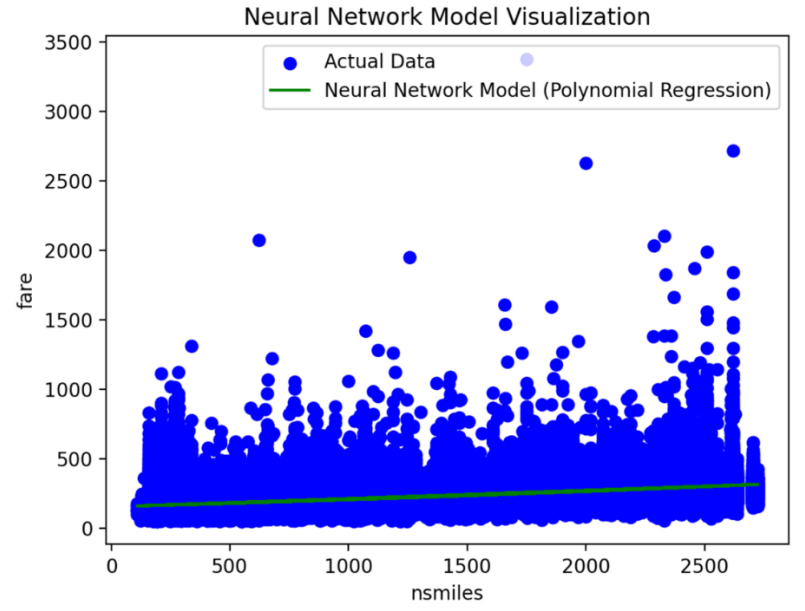

Fig. 4. Neural Network Model Prediction

Linear model predictions are shown in Figure 3, which roughly illustrates the linear predictions. As shown in Figure 4, the neural network model also gives intuitive prediction results. From the prediction results, Linear Model and Neural Network Model are similar. However, the prediction results are too low and do not match the previous data. The prediction results of these two models cannot match most of the historical data values. In contrast, as with the previous predictions, the ARIMA model has a more reasonable ability to process these aviation data. ARIMA does not give a simple linear prediction relationship, but provides a detailed, complex and non-linear prediction relationship based on the seasonality and regularity of previous data, which is more convincing and reliable. Shown as figure 2, these data themselves contain certain patterns, and the ARIMA model captures these patterns well and makes predictions. The performance of these models varies greatly in different situations. In a stable economic environment, the forecast results of linear models may be more reasonable. However, in the event of unexpected events such as global epidemics, the ability of ARIMA models to handle seasonality and irregularities is more important than linear models. When aviation data is affected by many factors, neural network models may play a prominent role, such as resource prices, political factors, etc., and it can handle the complexity between these factors. However, if the data is scarce or noisy, it may overfit, so that the results given do not meet the researcher's expectations. For short-term forecasts, if the trend is linear, linear models will have more advantages. But for long-term forecasts under the condition of seasonal changes, ARIMA models may be more appropriate. In general, understanding the specific context is crucial to making accurate forecasts in the aviation field.

6. Conclusion

This study focuses on the application of data visualization in the analysis of US aviation industry data sets and the analysis and comparison of different forecasting models. Data visualization can transform complex data into intuitive charts and other forms, which can promote better understanding of observers and enterprises and make reasonable decisions for future operations. Various visualization methods, mainly bar charts, have their own unique characteristics and application scenarios. This study chooses the Googlechart platform to generate visualization charts. This platform enables users to efficiently create a variety of charts. In this study, the analysis of visualization results reveals the fluctuations and patterns reflected in the number of flight records over time. For a deeper explanation and trend prediction, this study discusses forecasting models. The three common forecasting models are linear regression, ARIMA, and neural networks. Each model has its advantages and disadvantages. Although the linear regression model is simple and intuitive to observe, it has obvious limitations in dealing with nonlinear data and outliers. The neural network model shows strong learning ability, but there are still problems such as local optimal solution, overfitting, and high computational cost. The ARIMA model has been proven to be a more suitable choice for flight record time series data because it can handle trend, seasonality, and periodicity characteristics. Therefore, this study finally focused on the prediction of future flight data by ARIMA and obtained corresponding results. Overall, this study not only enhances the understanding of data visualization and prediction models, but also provides references for data analysis and decision-making in related fields in the future. However, this study also has certain limitations. In terms of data, the data set used only covers a specific period and a specific region of the US aviation industry, which cannot fully represent the situation of the entire aviation industry. At the same time, there may be some deviations or errors in these data, which will affect the accuracy of the analysis. In terms of models, although the ARIMA model performed well in this study, it still has some limitations. For example, it assumes that there is a linear relationship in the data, and its ability to handle data

with non-linear relationships may not be good. At the same time, the prediction accuracy of the model also needs to be further improved. For future research, researchers have some optimization directions. First, the data range should be expanded to obtain data information on the aviation industry in various periods. Second, the prediction model can be improved, such as improving the prediction accuracy. In addition, it can be considered to apply the research results and methods to other related fields, such as transportation or logistics, to help managers in these fields make more effective decisions.

References

1. Tang, D., Chen, M., Huang, X., Zhang, G., Zeng, L., Zhang, G., ... & Wang, Y. (2023). SRplot: A free online platform for data visualization and graphing. PLoS One, 18(11), e0294236.
2. Aurich, D., & Horaniet Ibañez, A. (2023). How can data visualization support interdisciplinary research? LuxTIME: studying historical exposomics in Belval. Frontiers in big Data, 6, 1164885.
3. Inastrilla, C. R. A. (2023, September). Data visualization in the information society. In Seminars in Medical Writing and Education (Vol. 2, pp. 25-25).
4. Bajić, F., Job, J., & Nenadić, K. (2020). Data visualization classification using simple convolutional neural network model. International journal of electrical and computer engineering systems, 11(1), 43-51.
5. Liu, A., Mahapatra, R. P., & Mayuri, A. V. R. (2023). Hybrid design for sports data visualization using AI and big data analytics. Complex & Intelligent Systems, 9(3), 2969-2980.
6. Ramaswami, G., Susnjak, T., Mathrani, A., & Umer, R. (2023). Use of predictive analytics within learning analytics dashboards: A review of case studies. Technology, Knowledge and Learning, 28(3), 959-980. [7] Wang, G., Zhao, B., Wu, B., Zhang, C., & Liu, W. (2023). Intelligent prediction of slope stability based on visual exploratory data analysis of 77 in situ cases. International Journal of Mining Science and Technology, 33(1), 47-59.
7. Khodadadi, E., & Towfek, S. K. (2023). Internet of Things Enabled Disease Outbreak Detection: A Predictive Modeling System. Journal of Intelligent Systems & Internet of Things, 10(1).
8. Kraak, M. J., & Ormeling, F. (2020). Cartography: visualization of geospatial data. CRC Press.
9. Narechania, A., Srinivasan, A., & Stasko, J. (2020). NL4DV: A toolkit for generating analytic specifications for data visualization from natural language queries. IEEE Transactions on Visualization and Computer Graphics, 27(2), 369-379.
10. Ospina, R., Gondim, J. A., Leiva, V., & Castro, C. (2023). An overview of forecast analysis with ARIMA models during the COVID-19 pandemic: Methodology and case study in Brazil. Mathematics, 11(14), 3069.
11. Chen, Y., Tong, Z., Zheng, Y., Samuelson, H., & Norford, L. (2020). Transfer learning with deep neural networks for model predictive control of HVAC and natural ventilation in smart buildings. Journal of Cleaner Production, 254, 119866.
12. Montesinos López, O. A., Montesinos López, A., & Crossa, J. (2022). Fundamentals of artificial neural networks and deep learning. In Multivariate statistical machine learning methods for genomic prediction (pp. 379-425). Cham: Springer International Publishing.