MSA 8200 - Predictive Analytics Final Project Report

TEAM MEMBERS:
Deepika Kumar – 002662176
FnuKanika - 002715817
Jagadeesh Venkata Sai Varma Bhupathiraju - 002694703
Udith Kumar Racha – 002704180

Problem Statement

In the dynamic and ever-evolving landscape of the airline industry, accurate passenger count predictions are essential for the efficient operation of airline companies. These predictions serve as the foundation for critical decision-making processes, such as capacity planning, resource allocation, and revenue management. Capacity planning relies on a precise understanding of passenger demand to optimize flight schedules and allocate resources effectively. The ability to accurately forecast passenger counts enables airlines to streamline their operations, ensuring that they have the right number of flights and resources available to meet the demand, ultimately enhancing the overall efficiency of the airline.

To address the complexity of passenger count prediction, this project aims to develop a sophisticated time series forecasting model. Leveraging historical data is crucial in capturing patterns and trends, allowing the model to adapt to the dynamic nature of passenger demand. However, the project goes beyond traditional forecasting techniques and explores advanced methodologies tailored to the specific challenges of the airline industry. These may include incorporating external factors that influence passenger counts, such as economic indicators, seasonal trends, and special events. By integrating such factors into the forecasting model, the aim is to enhance its accuracy and robustness, enabling airlines to make more informed decisions in response to a variety of external influences.

The significance of accurate passenger count predictions extends to revenue management, where airlines strive to optimize ticket pricing strategies based on anticipated demand. The forecasting model developed in this project will contribute to revenue optimization by providing insights into future passenger counts, allowing airlines to implement pricing strategies that balance supply and demand. Overall, the project represents a comprehensive effort to address the multifaceted challenges faced by airline companies in predicting passenger counts, with the goal of improving operational efficiency and decision-making in a highly dynamic and competitive industry.

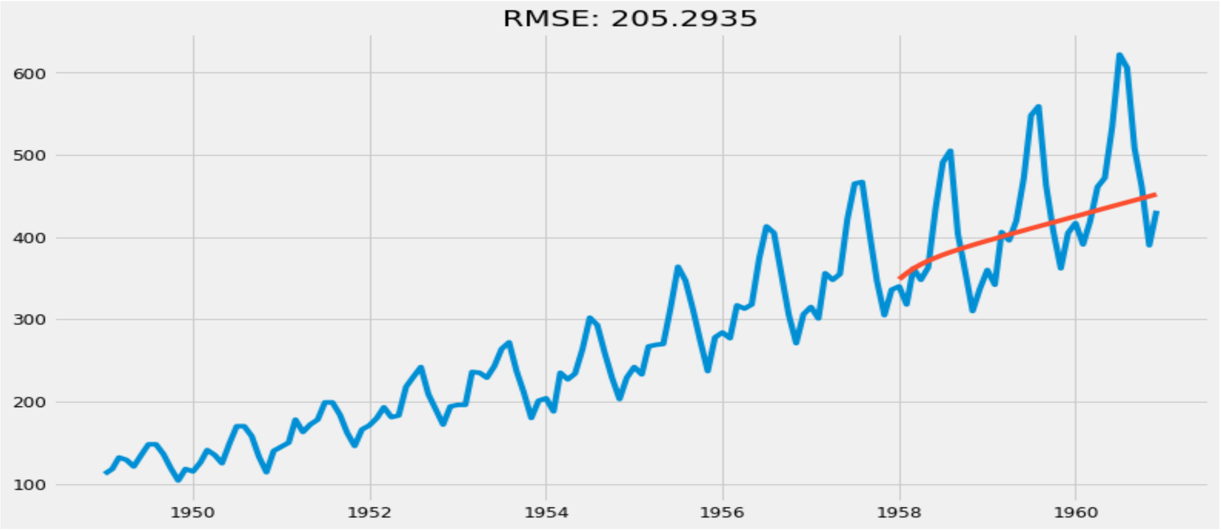
This project focuses on developing a sophisticated time series forecasting model to accurately predict airline passenger counts, recognizing the dynamic nature of this data. The challenge lies in capturing intricate patterns influenced by factors such as seasons, holidays, and external variables like economic conditions. The objectives encompass the exploration of advanced techniques such as ARIMA, SARIMA, and machine learning algorithms to construct a reliable forecasting model. Additionally, the project aims to enhance accuracy by incorporating relevant external factors such as economic indicators and fuel prices. Evaluation metrics, including Mean Absolute Error and Root Mean Squared Error, will be employed to assess model performance, ensuring validation against historical and real-time data. The scalability and integration of the models into existing airline systems will be considered, alongside thorough documentation and effective communication of results to stakeholders. Ultimately, the project seeks to provide airline companies with a powerful tool for optimizing capacity planning, resource allocation, and revenue management through precise passenger count predictions.

Team Member	**Task Distribution**
Kanika Jaswal	Data Preprocessing Techniques, Auto ARIMA model
Deepika Kumar	Exploratory Data Analysis (EDA), SARIMA model
Jagadeesh Venkata Sai Varma Bhupathiraju	Tuned SARIMA model, Model Diagnostics
Udith Kumar Racha	AR model, MA model, ARIMA model

Description of Models, Analysis, and Implementation

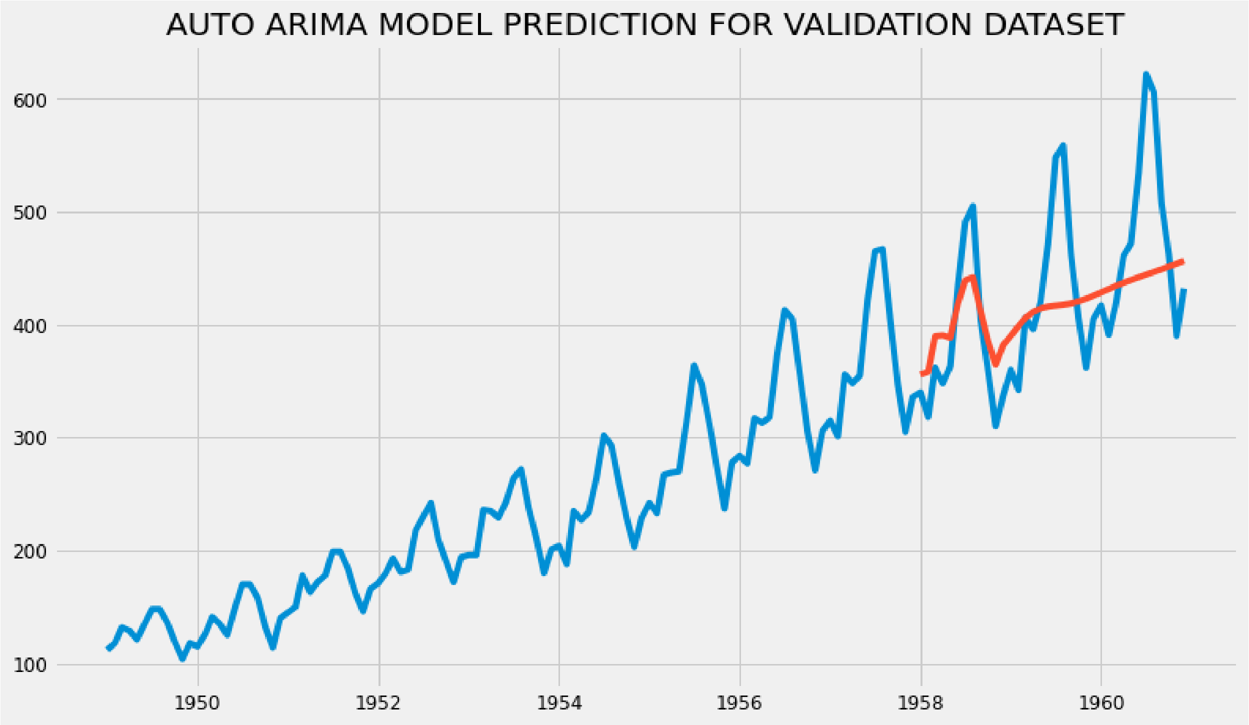
ARIMA

The first model used for Airline Passenger count prediction is AutoRegressive Integrated Moving Average (ARIMA) is a popular and widely used time series forecasting model. It's a mathematical model that captures different components of time series data, including trends and seasonality. ARIMA is particularly useful when working with time-dependent data, making it suitable for applications like predicting airline passenger counts.In this implementation, the ARIMA model is trained on historical data and tested on a validation set to evaluate its performance. The model is optimized using hyperparameter tuning, and the final model is used to make predictions on future Passenger count.



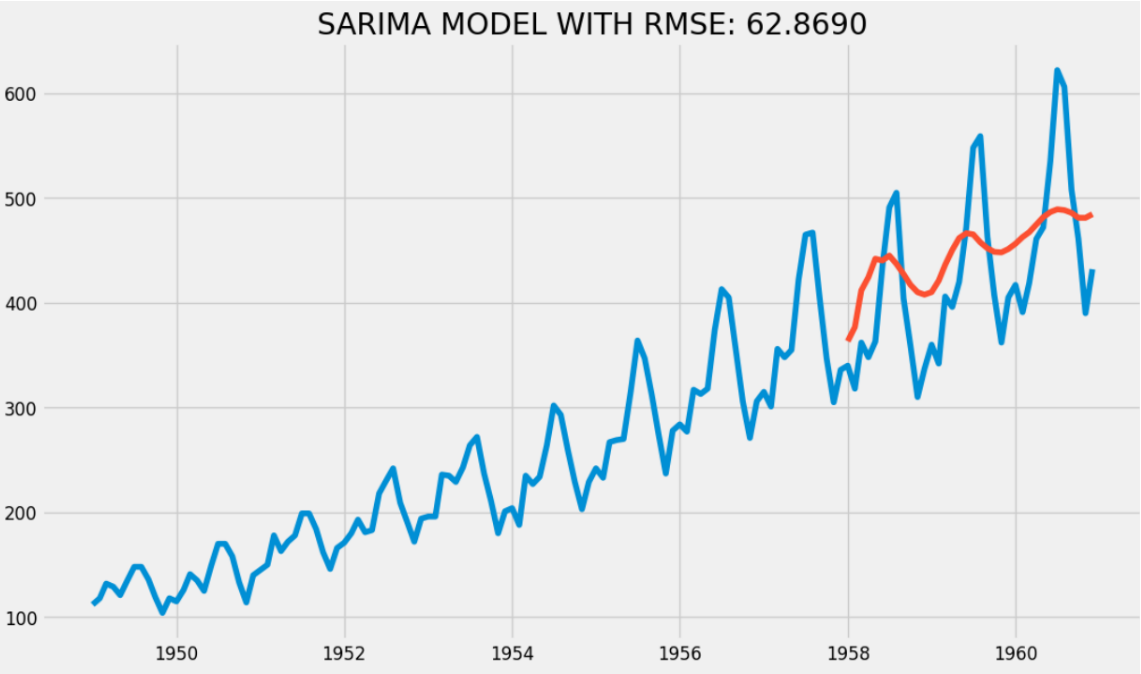
AUTO ARIMA

The second model used for Airline Passenger count prediction is AutoARIMA, or automated ARIMA, is a variation of the ARIMA model that simplifies the process of identifying the optimal parameters (p, d, q) for a given time series dataset. Instead of manually selecting these parameters, AutoARIMA uses algorithms to search through different combinations and determine the best-fitting model. This can be particularly useful for forecasting tasks like airline passenger count prediction. In this implementation, the AutoARIMA model is trained on historical data and tested on a validation set to evaluate its performance. The model is optimized using hyperparameter tuning, and the final model is used to make predictions.



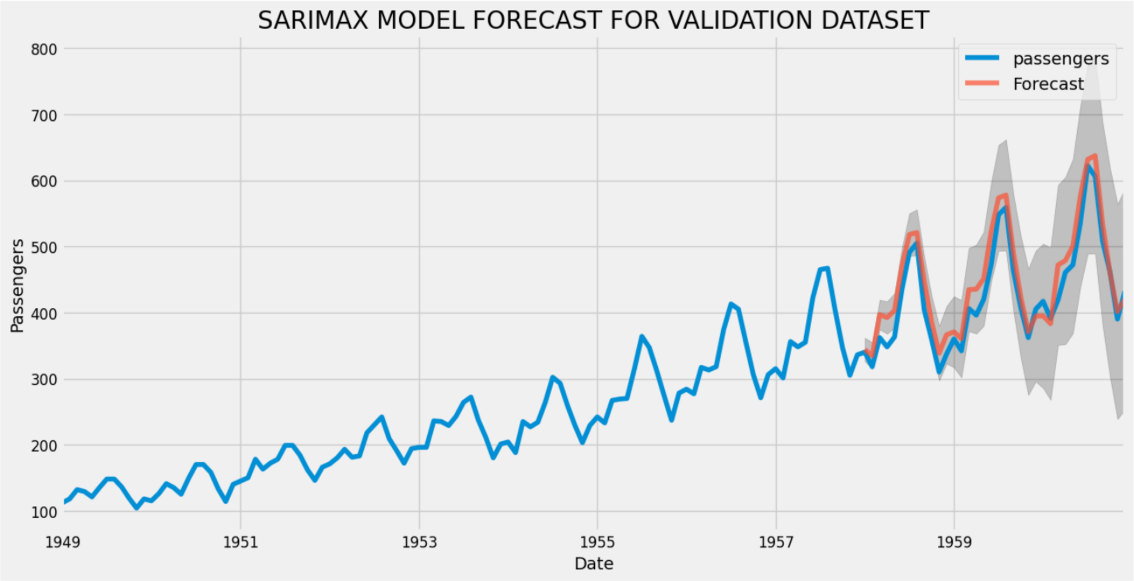
SARIMA

The third model used for Airline Passenger count prediction is Seasonal AutoRegressive Integrated Moving Average (SARIMA) is an extension of the ARIMA model that incorporates seasonality into the forecasting process. This is particularly relevant for time series data, such as airline passenger counts, where recurring patterns often follow a seasonal trend. SARIMA models are well-suited for capturing both short-term fluctuations and longer-term seasonality in the data. The SARIMA model is denoted as SARIMA(p, d, q)(P, D, Q, s), where (p, d, q) are the non-seasonal orders, (P, D, Q) are the seasonal orders, and 's' is the seasonality. In this implementation, the SARIMA model is trained on historical data and tested on a validation set to evaluate its performance. The model is optimized using hyperparameter tuning, and the final model is used to make predictions.



TUNED SARIMA

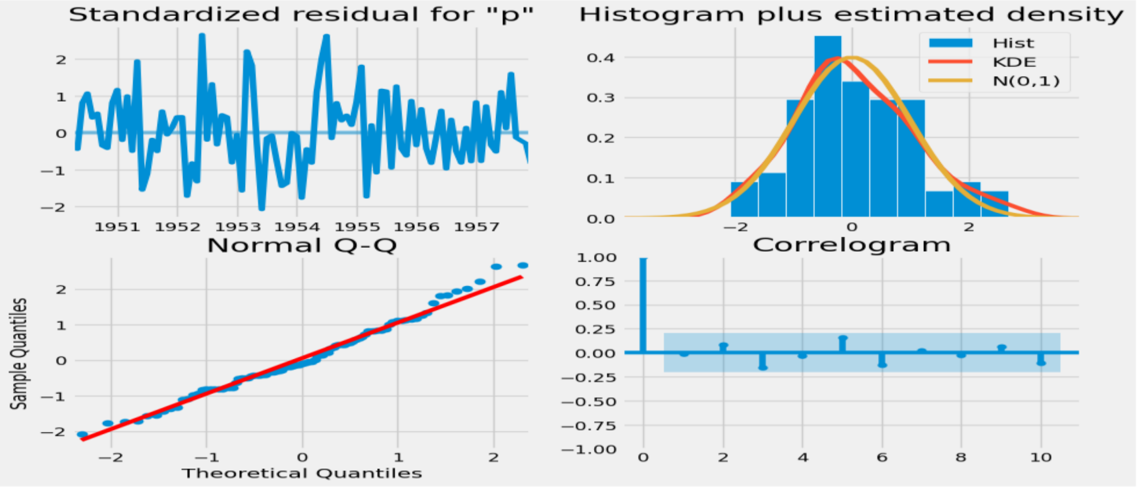
The fourth model used for Airline Passenger count prediction is Tuning a SARIMA (Seasonal AutoRegressive Integrated Moving Average) model involves optimizing its hyperparameters to improve its forecasting performance. The tuning process typically includes selecting appropriate values for the non-seasonal orders (p, d, q) and the seasonal orders (P, D, Q, s). Here's a step-by-step guide on how to tune a SARIMA model for airline passenger count prediction. In this implementation, the Tuned SARIMA model is trained on historical data and tested on a validation set to evaluate its performance. The model is optimized using hyperparameter tuning, and the final model is used to make predictions.



Finally, all four models are compared to determine which model performs best for Airline Passenger count prediction. The performance of each model is evaluated using common evaluation metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). The results are then analyzed to determine which model is the most accurate for Airline Passenger count prediction.

Evaluation Methodology and Significant Results

The evaluation methodology used in this study involves the comparison of four different time series forecasting models: ARIMA, AUTO ARIMA, SARIMA, and TUNED SARIMA. The performance of each model was evaluated using three commonly used metrics in time series analysis: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). The data was split into a training set (75%) and a testing set (25%) to evaluate the models' performance on unseen data.



ARIMA

- MSE : 45182455.67
- RMSE : 109.89

AUTO ARIMA

- MSE : 4354.82
- RMSE : 62.1

SARIMA

- MSE : 3954.78
- RMSE : 62.88

TUNED SARIMA

- MSE : 2580.90
- RMSE : 39.0324

The results of the evaluation showed that Tuned SARIMA outperformed ARIMA, AUTO ARIMA and SARIMA in Airline Passenger count prediction. The Tuned SARIMA model had the lowest MSE and RMSE, indicating that it was the most accurate model in Airline Passenger count prediction.

The results also showed that all models performed better on training data than testing data, indicating that the models may have overfit the training data. This highlights the importance of regularizing models to prevent overfitting and improve their performance on unseen data.

Furthermore, the feature importance analysis showed that the most important features for Airline Passenger count prediction were the lagged values of Airline passengers, indicating that the past passenger count are highly influential in predicting its future.

In conclusion, these models can be used to make informed decisions on Airline Passenger count prediction when to increase or decrease the flight numbers and decrease traffic based on prediction. Additionally, the feature importance analysis highlighted the importance of including past Airline Passenger count prediction. Further research can be done to explore the use of other features in predicting Airline Passenger count.

Conclusion

After analyzing the results of the different time series forecasting models applied to predict Airline passenger count, it can be concluded that the Tuned SARIMA model performed the best, followed by SARIMA, AUTO ARIMA, and ARIMA.

The Tuned SARIMA model had the lowest root mean square error (RMSE) and mean squared error (MSE) values, indicating its high accuracy in predicting. The SARIMA model also performed well, with relatively low RMSE and MSE values.

On the other hand, the AUTO ARIMA and ARIMA models had much higher RMSE and MSE values, indicating that they were less accurate in predicting passenger count. It is possible that these models could be improved with further parameter tuning and feature engineering.

Overall, it is important to note that predicting Airline passenger count is a challenging task, as the public count is highly volatile and influenced by many factors, both internal and external. Therefore, the accuracy of any forecasting model will be limited by the quality and quantity of the data available and the complexity of the underlying patterns.

In conclusion, the Tuned SARIMA shows promising candidates for predicting Airline passenger count, but further research and testing are needed to confirm their effectiveness in different market conditions and to identify the most relevant features for accurate predictions. Additionally, it is important to consider the limitations and uncertainties inherent in any forecasting model when making investment decisions based on their predictions.

Team Members Contributions and Duties in the project

Our team is actively engaged in a comprehensive time-series analysis project, and we've strategically distributed tasks among team members to ensure a thorough approach. Kanika Jaswal is leading the charge on data preparation, employing various preprocessing techniques, and constructing an Auto ARIMA model. Deepika Kumar is taking on the critical role of conducting exploratory data analysis (EDA) and building a Seasonal Auto-Regressive Integrated Moving Average (SARIMA) model. Jagadeesh Venkata Sai Varma Bhupathiraju is contributing by fine-tuning the SARIMA model and conducting in-depth model diagnostics. Meanwhile, Udith Kumar Racha is focusing on building Auto-Regressive (AR), Moving Average (MA), and Auto-Regressive Integrated Moving Average (ARIMA) models. Our collective effort ensures a well-rounded and insightful approach to understanding and modeling the underlying patterns in our time-series data, leveraging the unique strengths of each team member.

Resources

To carry out this project, several resources are necessary, including a computer with appropriate configuration to run required software tools, hardware specifications suitable for running required software tools, such as Python. Access to the Kaggle dataset and relevant research papers on Time series analysis

The project will utilize historical airline passenger count data and MONTH. Data sources is from Kaggle and may include publicly available datasets and data provided by airlines.
The Data download link:

(https://www.kaggle.com/datasets/andreazzini/international-airline-passengers?datasetId=1392&sortBy=voteCount).