Project Report: Time Series Analysis and Forecasting of Airline Passenger Traffic

1. Introduction

1.1 Background

Airline passenger traffic data is a critical metric for understanding the trends, seasonality, and overall growth in the airline industry. Accurate forecasting of passenger numbers can aid in resource planning, budgeting, and strategic decision-making. This project focuses on time series analysis and forecasting of airline passenger traffic using various statistical and machine learning techniques.

1.2 Objective

The primary objective of this project is to analyze historical airline passenger traffic data, understand its underlying patterns, and develop predictive models to forecast future passenger traffic. The models will be evaluated using metrics such as Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE).

2. Data Overview

2.1 Dataset Description

The dataset used in this analysis is a time series representing monthly airline passenger traffic. The data consists of two columns:

- **Month:** A timestamp representing the month of observation.
- **Passengers:** The number of passengers in that particular month.

2.2 Data Preprocessing

The data was first loaded and converted into a time series format. Missing values were handled using mean imputation and linear interpolation. Additionally, boxplots and histograms were used to analyze the distribution of the data and identify any potential outliers.

3. Exploratory Data Analysis (EDA)

3.1 Time Series Decomposition

Time series decomposition was performed to identify the trend, seasonality, and residual components. Both additive and multiplicative models were applied to understand the nature of the seasonality and trend in the data.

3.2 Visualization

- **Line Plot:** The raw time series data was visualized using a line plot to observe overall trends and seasonality.
- **Boxplots:** Boxplots with different whisker settings (1.5 and 1.0) were used to detect outliers.
- **Histograms:** A histogram of passenger numbers was plotted to understand the distribution of the data.

4. Model Development and Forecasting

4.1 Naive Method

The naive method assumes that the next observation will be equal to the last observation from the training set. The RMSE and MAPE for the naive method were calculated to serve as a benchmark.

4.2 Simple Average Method

The simple average method uses the average of all observations in the training set to forecast future values. This method was evaluated using RMSE and MAPE.

4.3 Simple Moving Average (SMA)

The SMA method was applied with different window sizes (12, 6, and 3 months). The forecast was then compared to the actual test data, and the RMSE and MAPE were calculated.

4.4 Simple Exponential Smoothing (SES)

The SES method was applied with a smoothing factor ($\alpha = 0.2$). The model was fitted to the training data, and forecasts were generated for the test period.

4.5 Holt's Linear Trend Model

Holt's method was used to capture both level and trend components in the data. An additive trend model was fitted, and forecasts were generated for the test period.

4.6 Holt-Winters Seasonal Method

Both additive and multiplicative seasonal methods were applied using Holt-Winters exponential smoothing. The model parameters (α, β, γ) were optimized, and forecasts were generated for the test period.

4.7 ARIMA Model

ARIMA models were implemented, including AR(1), MA(1), and ARMA(1,1) configurations. Box-Cox transformation and differencing were applied to make the data stationary before fitting the models.

5. Model Evaluation

5.1 Evaluation Metrics

The performance of each forecasting method was evaluated using the following metrics:

- **Root Mean Squared Error (RMSE):** Measures the average magnitude of the error.
- **Mean Absolute Percentage Error (MAPE):** Provides an error percentage relative to the observed data.

5.2 Results Comparison

The following table summarizes the RMSE and MAPE values for each method:

The following table summarizes the RMSE and MAPE values for each method:

Method	RMSE	MAPE
Naive Method	137.51	23.62
Simple Average Method	219.69	44.28
Simple Moving Average Method (Window=12)	103.33	15.54
Simple Moving Average Method (Window=6)	103.33	16.49
Simple Moving Average Method (Window=3)	107.65	11.11
Simple Exponential Smoothing	71.94	6.53
Holt's Linear Trend Method	35.10	6.82
Holt-Winters Additive Method	34.46	13.77
Holt-Winters Multiplicative Method	88.88	13.39
AR(1) Method	88.88	12.89
MA(1) Method	37.32	7.32
ARMA(1,1) Method	26.63	4.65

6. Stationarity Tests

6.1 Augmented Dickey-Fuller (ADF) Test

The ADF test was performed to check for stationarity. The test results indicated whether the time series was stationary or required further transformation.

6.2 KPSS Test

The KPSS test was also conducted to validate the results of the ADF test, providing additional insights into the stationarity of the series.

7. Data Transformation

7.1 Box-Cox Transformation

A Box-Cox transformation was applied to stabilize the variance and make the series more stationary. Differencing was also performed to remove trend and seasonality.

7.2 ACF and PACF Plots

ACF and PACF plots were generated to identify the order of AR and MA components for the ARIMA models.

8. Conclusion

8.1 Summary of Findings

- The time series data exhibited strong seasonality and a linear trend.
- Holt-Winters multiplicative method provided the best forecast accuracy with the lowest RMSE and MAPE.
- Stationarity was achieved using Box-Cox transformation and differencing.

8.2 Future Work

Future work could involve exploring more advanced models such as SARIMA, LSTM neural networks, or ensemble methods to further improve forecasting accuracy.