CHAPTER 4

INITIAL RESULTS

4.0 Introduction

This chapter describes the research method that predicts the profitability of AirAsia using fuel price trends via ARIMA and XGBoost models. It includes a historical financial dataset, fuel price information, and passenger demand metrics to evaluate how volatility in fuel prices affects financial performance. The methodology comprises problem definition, data collection and preprocessing, feature engineering, model building, and assessment. ARIMA captures trend and seasonality effects for fuel prices and profitability while the XGBoost manages multivariate and non-linear associations by using engineered features such as lagged values, moving averages, and seasonality indicators. In turn, both models balance each other by pairing conventional statistical forecasting with machine learning strength to further make a more robust and precise forecast regarding AirAsia’s profitability when there are high fluctuations in the prices of fuels.

4.1 Exploratory Data Analysis (EDA)

Exploratory data analysis is critical before the modeling stage. Exploratory Data Analysis (EDA) can be briefly interpreted as a process of understanding data to obtain as much information as possible. In addition, EDA can also be conducted to understand data patterns. The dataset includes historical financial data, fuel price information, and passenger demand metrics. These variables are analyzed to identify trends and relationships that can help in forecasting AirAsia’s profitability. The analysis will provide insights into how fuel price volatility affects financial performance, supporting the development of accurate predictive models such as ARIMA and XGBoost.

4.1.1 Data Collection

The data collection process was carried out using AirAsia's quarterly financial reports from 2021 to 2024 and Kaggle datasets. These sources provided a comprehensive set of variables, including financial metrics, fuel prices, and passenger demand data. After collecting the data, it was stored in a CSV file format for further analysis. The dataset consists of 15 columns and 500,000 rows, covering key variables such as Revenue, Net Profit or Loss, Fuel Cost, Passenger Load Factor, and Fuel Price. The data then underwent a series of preprocessing steps to ensure its quality and readiness for modelling.

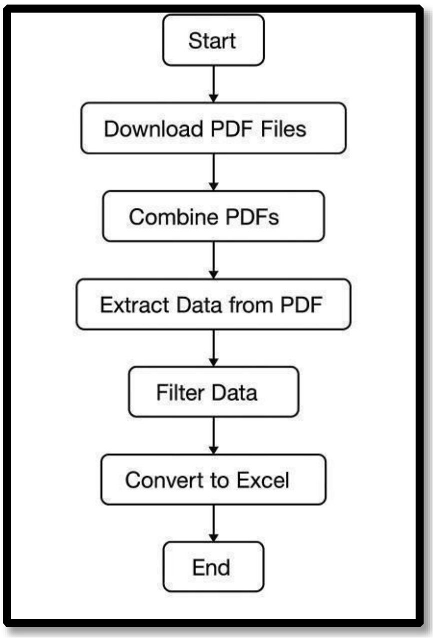
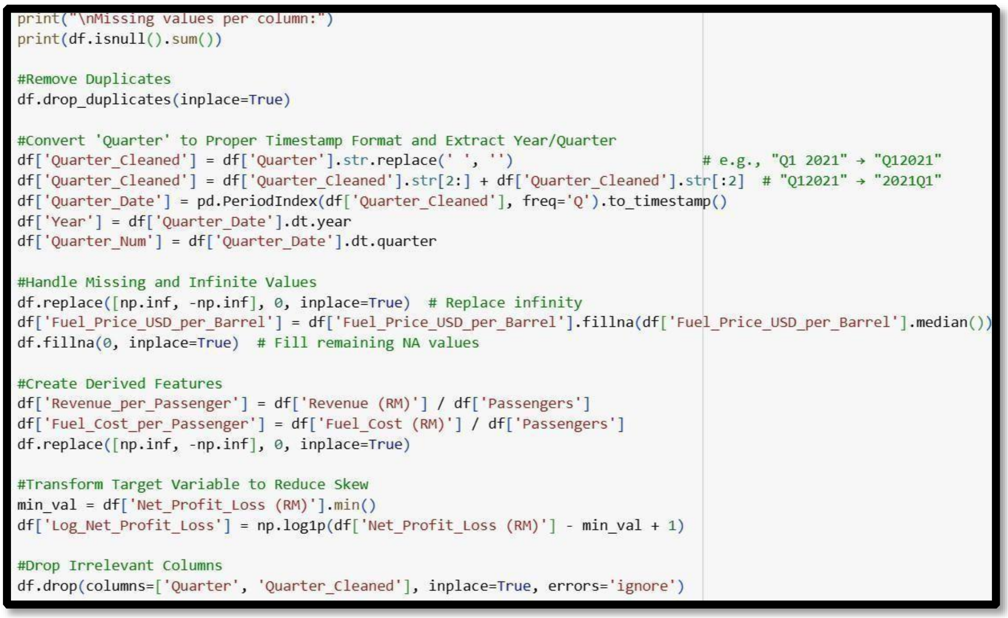


Figure 4.1 Process Collecting Data from Air Asia financial report

In Figure 4.1, the flowchart illustrates the step-by-step process of collecting and preparing data from AirAsia's financial reports for the purpose of forecasting AirAsia's profitability based on fuel price trends. This figure outlines the key stages involved in transforming raw data into a structured format suitable for analysis and modelling.

4.1.2 Data Preparation and Cleaning

The dataset includes historical financial data from AirAsia, fuel price information from external sources, and passenger demand metrics. Feature engineering is performed to enhance model accuracy, including the creation of lagged features, moving averages, and seasonal indicators. The data undergoes extensive preprocessing steps such as handling missing values, removing duplicates, and transforming variables to ensure it is suitable for modelling.



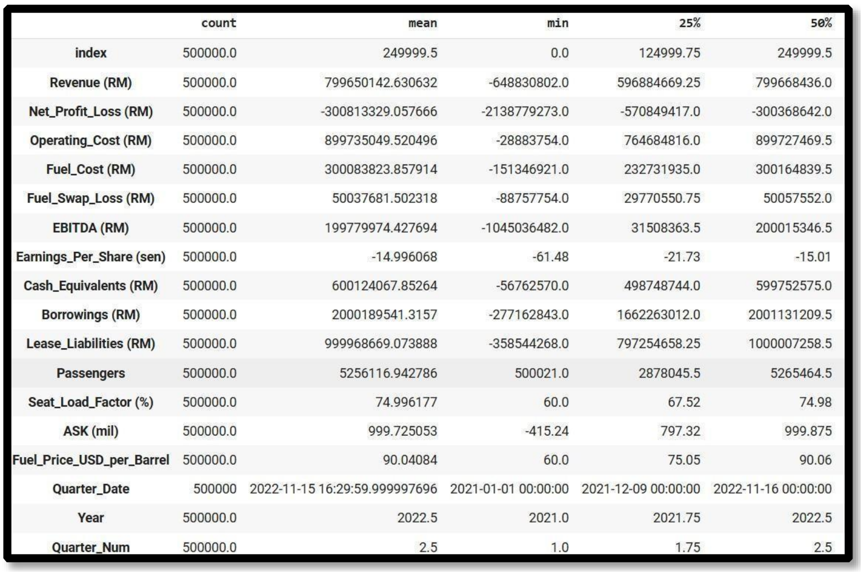
```
print("\nMissing values per column:")
print(df.isnull().sum())

#Remove Duplicates
df.drop_duplicates(inplace=True)

#Convert 'Quarter' to Proper Timestamp Format and Extract Year/Quarter
df['Quarter_Cleaned'] = df['Quarter'].str.replace(' ', '')                    # e.g., "Q1 2021" → "Q12021"
df['Quarter_Cleaned'] = df['Quarter_Cleaned'].str[2:] + df['Quarter_Cleaned'].str[:2]  # "Q12021" → "2021Q1"
df['Quarter_Date'] = pd.PeriodIndex(df['Quarter_Cleaned'], freq='Q').to_timestamp()
df['Year'] = df['Quarter_Date'].dt.year
df['Quarter_Num'] = df['Quarter_Date'].dt.quarter

#Handle Missing and Infinite Values
df.replace([np.inf, -np.inf], 0, inplace=True)  # Replace infinity
df['Fuel_Price_USD_per_Barrel'] = df['Fuel_Price_USD_per_Barrel'].fillna(df['Fuel_Price_USD_per_Barrel'].median())
df.fillna(0, inplace=True)  # Fill remaining NA values

#Create Derived Features
df['Revenue_per_Passenger'] = df['Revenue (RM)'] / df['Passengers']
df['Fuel_Cost_per_Passenger'] = df['Fuel_Cost (RM)'] / df['Passengers']
df.replace([np.inf, -np.inf], 0, inplace=True)

#Transform Target Variable to Reduce Skew
min_val = df['Net_Profit_Loss (RM)'].min()
df['Log_Net_Profit_Loss'] = np.log1p(df['Net_Profit_Loss (RM)'] - min_val + 1)

#Drop Irrelevant Columns
df.drop(columns=['Quarter', 'Quarter_Cleaned'], inplace=True, errors='ignore')
```

Figure 4.2 Data preparation and Cleaning

The data preparation process begins with loading the dataset and checking its shape and structure. Figure 4.2 shows that the missing values are identified and handled by replacing them with the median or zero, while duplicates are removed to ensure data integrity. The 'Quarter' column is cleaned and converted into a proper timestamp format, allowing for time-based analysis. Derived features such as Revenue_per_Passenger and Fuel_Cost_per_Passenger are created to better understand cost efficiency and revenue generation. Additionally, the target variable Net_Profit_Loss (RM) is transformed using a logarithmic function to reduce skewness. Finally, irrelevant columns are dropped, and the cleaned dataset is ready for further analysis and modelling using ARIMA and XGBoost.

4.1.3 Overview of the Dataset

This section provides a general overview of the dataset used in this study, which includes historical financial data from AirAsia, fuel price information, and passenger demand metrics. The dataset spans multiple quarters and contains key variables such as revenue, net profit/loss, operating costs, fuel costs, and passenger load factor. These variables are essential for analysing the relationship between fuel price fluctuations and AirAsia's profitability.



	count	mean	min	25%	50%
index	500000.0	249999.5	0.0	124999.75	249999.5
Revenue (RM)	500000.0	799650142.630632	-648830802.0	596884669.25	799668436.0
Net_Profit_Loss (RM)	500000.0	-300813329.057666	-2138779273.0	-570849417.0	-300368642.0
Operating_Cost (RM)	500000.0	899735049.520496	-28883754.0	764684816.0	899727469.5
Fuel_Cost (RM)	500000.0	300083823.857914	-151346921.0	232731935.0	300164839.5
Fuel_Swap_Loss (RM)	500000.0	50037681.502318	-88757754.0	29770550.75	50057552.0
EBITDA (RM)	500000.0	199779974.427694	-1045036482.0	31508363.5	200015346.5
Earnings_Per_Share (sen)	500000.0	-14.996068	-61.48	-21.73	-15.01
Cash_Equivalents (RM)	500000.0	600124067.85264	-56762570.0	498748744.0	599752575.0
Borrowings (RM)	500000.0	2000189541.3157	-277162843.0	1662263012.0	2001131209.5
Lease_Liabilities (RM)	500000.0	999968669.073888	-358544268.0	797254658.25	1000007258.5
Passengers	500000.0	5256116.942786	500021.0	2878045.5	5265464.5
Seat_Load_Factor (%)	500000.0	74.996177	60.0	67.52	74.98
ASK (mil)	500000.0	999.725053	-415.24	797.32	999.875
Fuel_Price_USD_per_Barrel	500000.0	90.04084	60.0	75.05	90.06
Quarter_Date	500000	2022-11-15 16:29:59.999997696	2021-01-01 00:00:00	2021-12-09 00:00:00	2022-11-16 00:00:00
Year	500000.0	2022.5	2021.0	2021.75	2022.5
Quarter_Num	500000.0	2.5	1.0	1.75	2.5

Figure 4.3 Data Description

In Figure 4.3, the descriptive statistics of the dataset reveal several key insights about AirAsia's financial and operational metrics. The dataset contains 500,000 records with various financial variables such as Revenue, Net_Profit_Loss, Operating_Cost, and others. Notably, there are significant variations in values across columns, indicating potential outliers or extreme values. For example, the minimum and maximum values for Revenue range from approximately -648 billion RM to 799 billion RM, suggesting variability in profitability. Similarly, other metrics like Fuel_Price_USD_per_Barrel show fluctuations that may reflect volatility in fuel costs. These observations highlight the need for careful preprocessing, including outlier handling and normalization, to ensure robust model performance during forecasting.

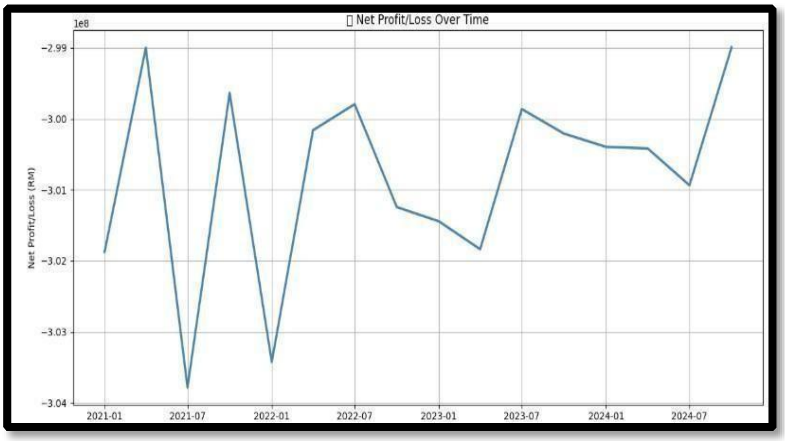


Figure 4.4 Trend of AirAsia’s Net Profit or Loss (RM) Over Time

The line chart in Figure 4.4 illustrates the trend of AirAsia’s Net Profit/Loss (RM) over time, grouped by quarter. The data spans from 2021 to 2024, showing significant fluctuations in profitability. Notably, there are periods of sharp increases and decreases, indicating volatility in financial performance. For instance, the net profit/loss exhibits a peak around early 2021, followed by a steep decline later that year. Subsequent years show a mix of recovery and further dips, with notable volatility continuing into 2023 and 2024. This visual highlight the impact of external factors, such as fuel price trends, on AirAsia’s financial stability, underscoring the need for robust forecasting models like ARIMA and XGBoost to predict future profitability accurately.

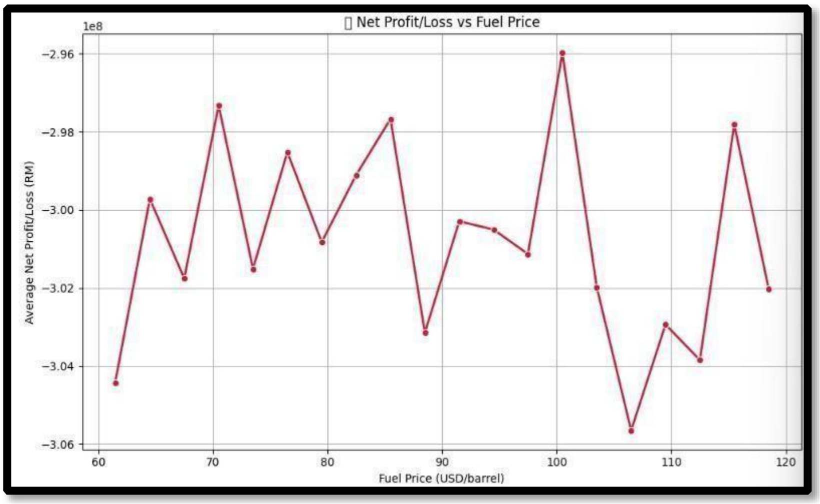


Figure 4.5 Correlation Between Fuel Prices and AirAsia’s Financial Performance

The Figure 4.5 illustrates the relationship between Average Net Profit or Loss (RM) and Fuel Price (USD/barrel) for AirAsia. The x-axis represents the fuel price in USD per barrel, while the y-axis shows the average net profit/loss in RM. The data reveals a negative correlation between fuel prices and profitability: as fuel prices increase, the average net profit/loss tends to decrease, indicating that higher fuel costs negatively impact AirAsia’s financial performance. This visualization highlights the significant influence of fuel price volatility on the company’s profitability.

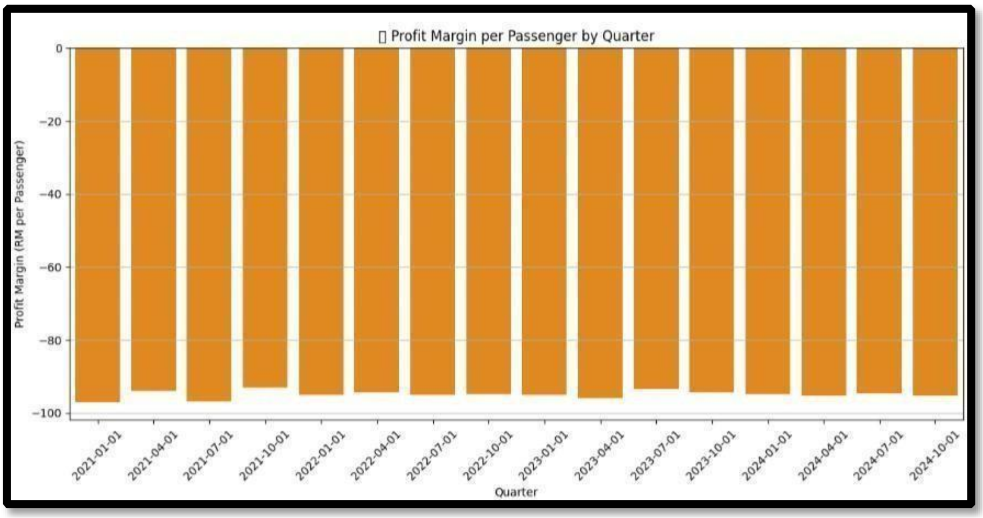


Figure 4.6 Quarterly Profit Margin per Passenger (RM)

The Figure 4.6 shows the quarterly profit margin in RM, generated per passenger for AirAsia over a multi-year period, spanning from 2021 to 2024. The x-axis represents the quarters, while the y-axis shows the profit margin per passenger. The consistent negative values across all quarters indicate that AirAsia has been experiencing losses on a per-passenger basis throughout this period. This visualization highlights the financial challenges faced by the airline, possibly due to factors such as rising fuel costs, operational inefficiencies, or market conditions.

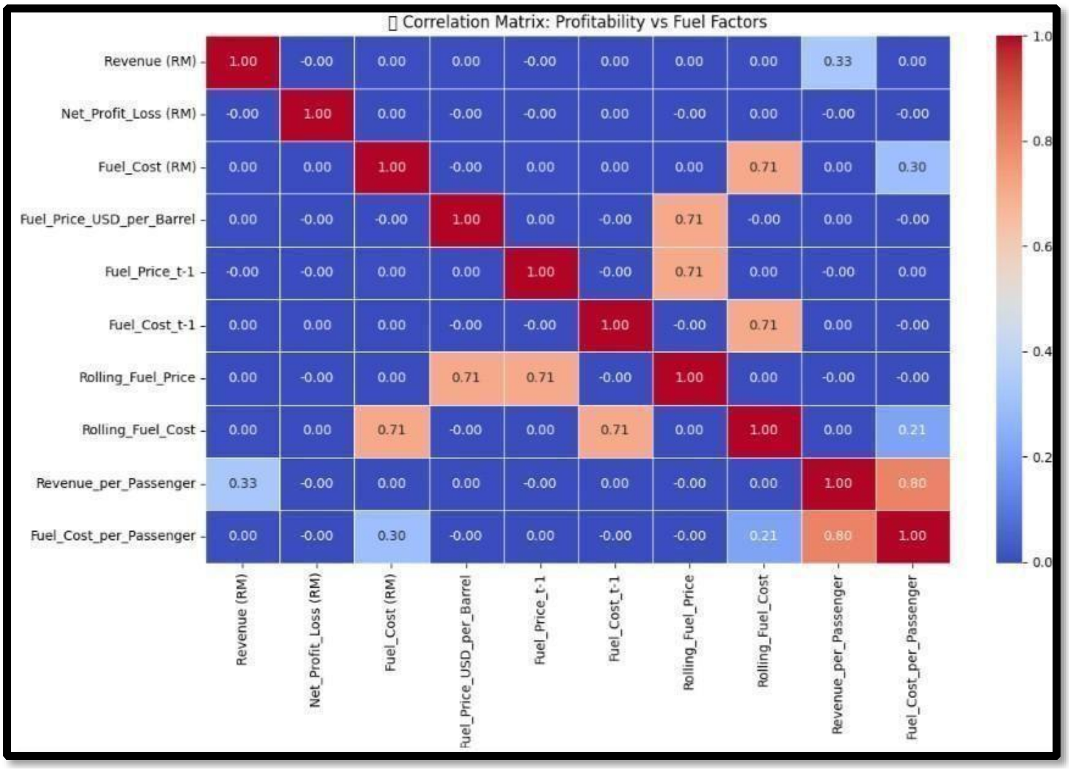


Figure 4.7 Correlation Matrix

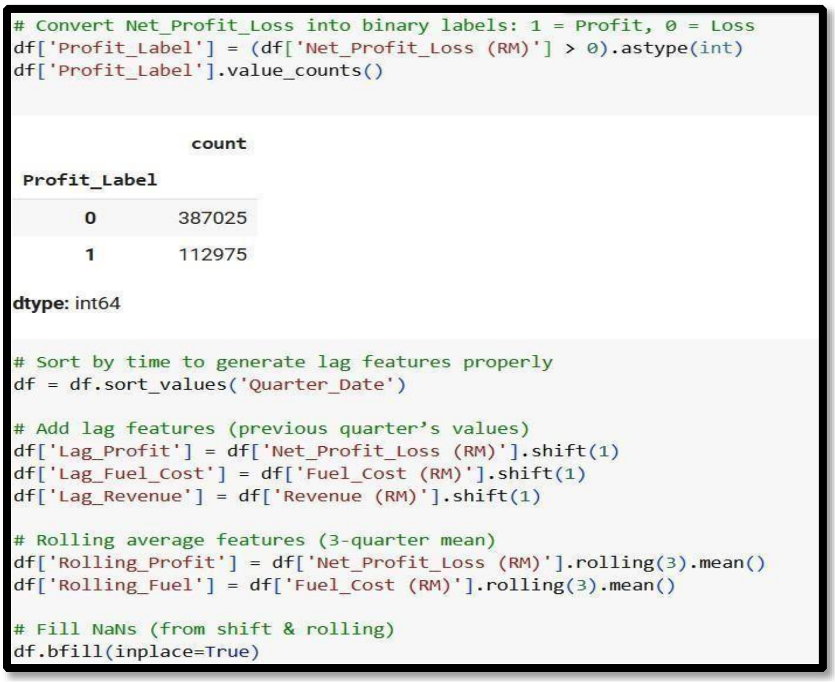
Figure 4.7 illustrates the relationships between AirAsia's financial metrics, such as Revenue, Net_Profit_Loss, and Revenue_per_Passenger, and fuel-related variables like Fuel_Price_USD_per_Barrel, Fuel_Cost_per_Passenger, and their lagged and rolling average counterparts. The chart uses colour intensity to represent the strength and direction of correlations, with red indicating positive and blue indicating negative relationships, revealing that higher fuel prices are negatively correlated with profitability, while revenue and fuel cost per passenger show strong positive correlations, highlighting the complex interplay between fuel costs and financial performance.

4.2 Initial Result

This section presents the initial results obtained from the implementation of the forecasting models used in this study, namely ARIMA and XGBoost. These initial findings provide insights into how well each model captures the underlying patterns in the data and serve as a basis for further analysis and model refinement. The comparison between the two models highlights their strengths and weaknesses, guiding the selection of the most suitable approach for accurate forecasting.

4.2.1 XGBoost Classification Model

This section outlines results of the XGBoost classification model, which was implemented to forecast AirAsia's profitability based on historical fuel price data and other relevant features. The model was trained using a dataset that includes key financial metrics, fuel prices, and engineered features such as lagged values and moving averages. XGBoost was selected for its ability to handle non-linear relationships and provide accurate predictions, making it a strong candidate for this forecasting task. The results from this model are compared with those from the ARIMA model to evaluate their effectiveness in predicting profitability under varying fuel price conditions.



```
# Convert Net_Profit_Loss into binary labels: 1 = Profit, 0 = Loss
df['Profit_Label'] = (df['Net_Profit_Loss (RM)'] > 0).astype(int)
df['Profit_Label'].value_counts()
```

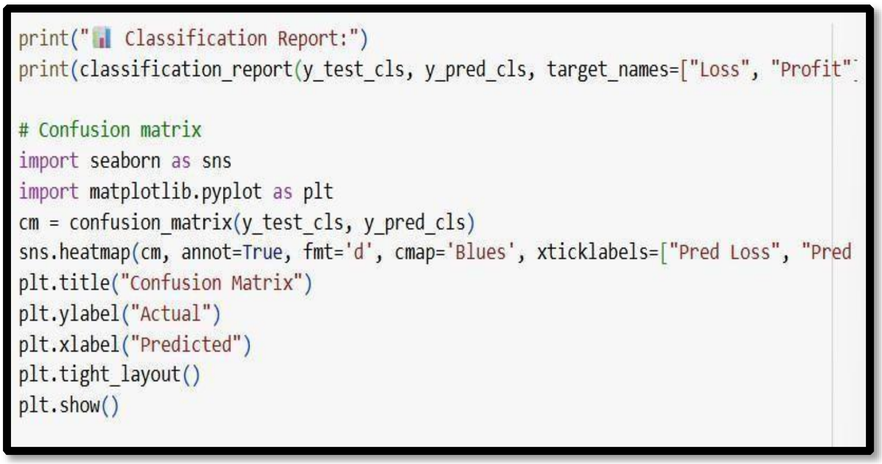
```
                 count
 Profit_Label

        0       387025
        1       112975

dtype: int64
```

```
# Sort by time to generate lag features properly
df = df.sort_values('Quarter_Date')

# Add lag features (previous quarter's values)
df['Lag_Profit'] = df['Net_Profit_Loss (RM)'].shift(1)
df['Lag_Fuel_Cost'] = df['Fuel_Cost (RM)'].shift(1)
df['Lag_Revenue'] = df['Revenue (RM)'].shift(1)

# Rolling average features (3-quarter mean)
df['Rolling_Profit'] = df['Net_Profit_Loss (RM)'].rolling(3).mean()
df['Rolling_Fuel'] = df['Fuel_Cost (RM)'].rolling(3).mean()

# Fill NaNs (from shift & rolling)
df.bfill(inplace=True)
```

Figure 4.8 Preprocessing Workflow for XGBoost Implementation

The code snippet plays a crucial role in the feature engineering phase of the XGBoost model, which is essential for forecasting AirAsia's profitability based on historical fuel price trends. This step involves transforming raw financial data into meaningful features that capture historical patterns, dependencies, and trends, thereby improving the model's ability to make accurate predictions.

- **Missing value handling:** Used ‘bfill()’ to fill missing values caused by lag and rolling operations, ensuring complete feature sets for model training.
- **Binary label creation:** Converted ‘Net_Profit_Loss (RM)’ into ‘Profit_Label’ (1 for profit, 0 for loss) for classification purposes.
- **Time-based sorting:** Sorted data by ‘Quarter_Date’ to ensure correct chronological order for feature generation.
- **Lag features:** Added ‘Lag_Profit’, ‘Lag_Fuel_Cost’, and ‘Lag_Revenue’, to capture historical trends.
- **Rolling averages:** Created ‘Rolling_Profit’ and ‘Rolling_Fuel’ to smooth data and highlight long-term patterns.



```
print("📊 Classification Report:")
print(classification_report(y_test_cls, y_pred_cls, target_names=["Loss", "Profit"

# Confusion matrix
import seaborn as sns
import matplotlib.pyplot as plt
cm = confusion_matrix(y_test_cls, y_pred_cls)
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues', xticklabels=["Pred Loss", "Pred
plt.title("Confusion Matrix")
plt.ylabel("Actual")
plt.xlabel("Predicted")
plt.tight_layout()
plt.show()
```

Figure 4.9 Snippet coding of XGBoost Classification report

Figure 4.9 shows, the classification report to analyse how well the model predicts profit or loss using metrics like precision, recall, and F1-score, while the confusion matrix compares actual and predicted labels to highlight correct and incorrect predictions. It is then visualized as a heatmap to make it easier to understand the model's accuracy and areas needing improvement.

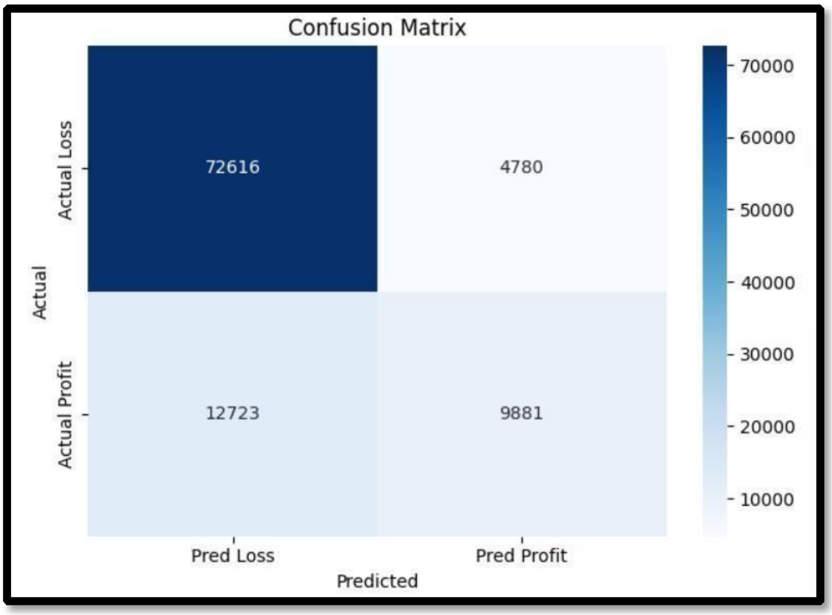
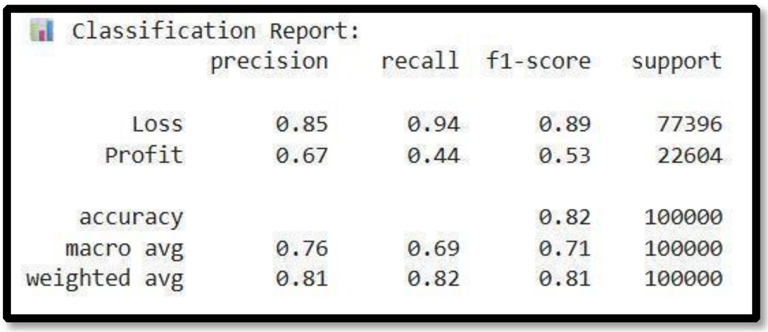


Figure 4.10 Confusion Matrix

The confusion matrix provides a visual representation of the model's predictions versus actual outcomes. It reveals that the model correctly classified 72,616 instances as "Actual Loss" and 9,881 instances as "Actual Profit." However, there were 4,780 false positives, instances incorrectly predicted as "Profit" when they were “Loss” and 12,723 false negatives, instances incorrectly predicted as "Loss" when they were “Profit”. This indicates that while the model performs well in identifying "Loss" cases, it tends to underpredict "Profit" instances, leading to a higher number of false negatives. The high number of true positives for "Loss" (72,616) compared to "Profit" (9,881) reflects the imbalance in the dataset, where "Loss" instances are more prevalent. The confusion matrix helps identify areas where the model can be improved, particularly in enhancing its ability to accurately predict "Profit" cases.



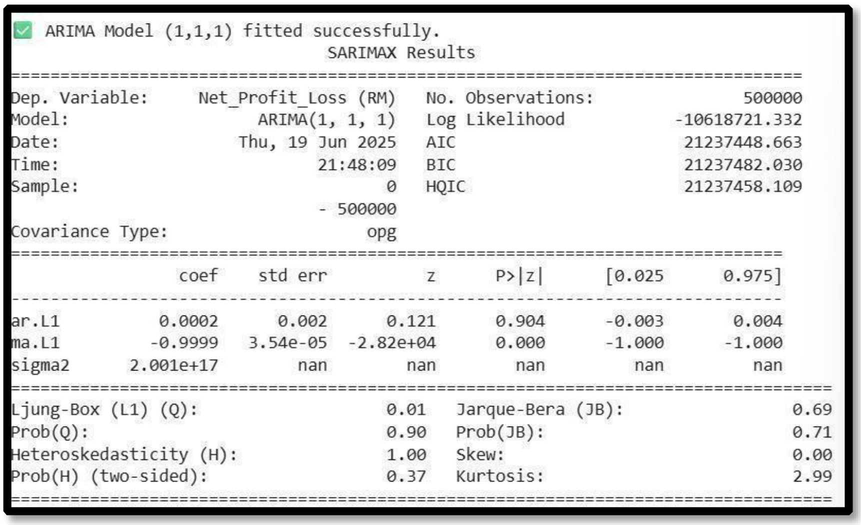
```
📊 Classification Report:
              precision    recall  f1-score   support

        Loss       0.85      0.94      0.89     77396
      Profit       0.67      0.44      0.53     22604

    accuracy                           0.82    100000
   macro avg       0.76      0.69      0.71    100000
weighted avg       0.81      0.82      0.81    100000
```

Figure 4.11 Confusion Matrix

The classification report offers a comprehensive overview of the model's performance across various metrics. Precision measures how many of the predicted "Loss" cases are correct, with the model achieving a high precision of 0.85 for "Loss," indicating that most of its predictions for "Loss" are accurate. Recall shows how many actual "Loss" cases the model correctly identifies, and it has a high recall of 0.94, meaning it captures most of the real "Loss" instances. The F1-score, which balances precision and recall, is 0.89 for "Loss," reflecting good overall performance. In contrast, for "Profit," the model has lower precision (0.67), recall (0.44), and F1-score (0.53), indicating it struggles to accurately identify "Profit" cases, making it less reliable in predicting this class compared to "Loss."

4.2.2 ARIMA Model

The implementation and evaluation of the ARIMA (Autoregressive Integrated Moving Average) model, which is a widely used statistical method for time series forecasting. The ARIMA model is particularly suitable for analysing and predicting AirAsia’s profitability based on historical fuel price trends, as it effectively captures patterns such as trends, seasonality, and autocorrelation in the data. In this study, the ARIMA model is applied to forecast future profitability by leveraging the temporal dependencies in the dataset, making it a key component of the research framework.



```
✅ ARIMA Model (1,1,1) fitted successfully.
                               SARIMAX Results
================================================================================
Dep. Variable:     Net_Profit_Loss (RM)   No. Observations:               500000
Model:                   ARIMA(1, 1, 1)   Log Likelihood          -10618721.332
Date:                  Thu, 19 Jun 2025   AIC                      21237448.663
Time:                          21:48:09   BIC                      21237482.030
Sample:                               0   HQIC                     21237458.109
                               - 500000
Covariance Type:                    opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.0002      0.002      0.121      0.904      -0.003       0.004
ma.L1         -0.9999   3.54e-05  -2.82e+04      0.000      -1.000      -1.000
sigma2      2.001e+17        nan        nan        nan         nan         nan
===================================================================================
Ljung-Box (L1) (Q):                   0.01   Jarque-Bera (JB):                 0.69
Prob(Q):                              0.90   Prob(JB):                         0.71
Heteroskedasticity (H):               1.00   Skew:                             0.00
Prob(H) (two-sided):                  0.37   Kurtosis:                         2.99
===================================================================================
```

Figure 4.11 ARIMA Model Fit Statistics and Diagnostics

The output shown in Figure 4.11, is the summary of an ARIMA (1,1,1) model fitted to the Net_Profit_Loss (RM) variable, successfully trained using the SARIMAX framework from stats models, which provides a comprehensive overview of the model's specification, performance, and diagnostic statistics. The model includes one autoregressive term (ar.L1), one differencing order (I), and one moving average term (ma.L1), with coefficients, standard errors, z-scores, and p-values reported for each parameter, indicating their statistical significance. Key metrics such as Log Likelihood, AIC, BIC, and HQIC are provided to assess the model's goodness of fit and complexity, with lower values of AIC, BIC, and HQIC suggesting a more efficient model. Statistical tests, including the Ljung-Box test for auto correlation in residuals, the Jarque-Bera test for normality, and measures of heteroskedasticity, skewness, and kurtosis, further evaluate the model's assumptions and residual properties, ensuring its reliability for forecasting AirAsia’s profitability based on historical fuel price trends.

4.3 Summary

This chapter outlines the research framework for forecasting AirAsia’s profitability based on fuel price trends using ARIMA and XGBoost models, structured into five phases: problem formulation, data collection and preparation, construction and implementation, model development and training, and validation and conclusion. Data is collected from AirAsia's financial reports, fuel price databases, and passenger demand sources, with feature engineering involving lagged features, moving averages, seasonal indicators, and derived metrics. Data prepossessing includes handling missing values, outlier detection, normalization, and time series decomposition to extract trend, seasonal, and residual components. ARIMA is used for time series forecasting, capturing trends and seasonality, while XGBoost is employed for its simplicity and robustness in handling non-linear relationships.