Module: ITNPBD4 Commercial and Scientific Applications
Assessment: Assignment
Due Date/Time: 15 December 2025, 23:59
AIAS Levels Allowed: Level 3

	Please tick the boxes/include appropriate information below
Student ID Number	3453366
Word Count (penalties apply for exceeding the stated limit)	3119
I have read and understand the severity of academic misconduct – see link below	☑
I give consent for my work to be used as an exemplar to future students.	☑
I have checked my submitted document to ensure it complies with module requirements.	☑
Link to version-controlled file (i.e on OneDrive, Google Docs, Github, or other) which contain evidence of the process I undertook to complete this assignment. Information on how to create a Microsoft 365 OneDrive folder is available HERE. *Please see notes below	https://github.com/muhpro/3453366.git
I understand that if there is a concern about potential academic misconduct including inappropriate use of AI tools then I could be asked to provide evidence of my drafting process during an academic integrity meeting if I have not done so using the link above. Not providing evidence of my drafting process could prejudice the outcome of academic misconduct cases.	☑
Tailored feedback. If you would like tailored feedback on a specific aspect (or aspects) of your work (e.g., referencing, writing style, grammar), then please give details here.	Nil
If you used AI at (or below) the level allowed, please explain briefly which AI, how you used it, and for what purpose.	I used AI for brainstorming and organization of my imaginations to suit the assignment guideline.

**This may include (but is not limited to) drafts, versions of the finished document, notes, references, AI output, and AI prompts. These materials are not marked or graded, but they are simply a way to demonstrate how your work was created and to confirm that any AI use in your final submission is within the permitted AIAS scale for your assessment. Providing this helps safeguard you, showing your authentic process, and protecting you should any academic integrity questions arise.*

https://www.stir.ac.uk/about/professional-services/student-academic-and-corporate-services/academic-registry/policy-and-procedure/

For more information, please see the Coversheet FAQ.

V.Ger Travel Data Science Strategy Report

Prepared for: V.Ger Travel Board of Directors
Prepared by: Chief Data Officer (Student ID: 3453366)
Module: ITNPBD4 - Commercial and Scientific Applications
Date: 14 December 2025

1. Executive Summary

This report outlines a transformative Data Science Strategy for V.Ger Travel, designed to leverage our conglomerate's existing IT infrastructure to drive operational efficiency, revenue growth, and customer retention. As a diverse travel provider managing hotels, resorts, car rentals, and charter flights, V.Ger Travel possesses a rich but underutilized asset in its transactional and customer data.

A detailed investigation was conducted on a Time Series Forecasting for Hotel Demand Prediction use case. Using a synthetic dataset representative of our monthly booking volumes, two distinct modeling approaches, a classical statistical method (SARIMA) and a modern machine learning approach (LightGBM), were implemented and compared.

The analysis revealed that for our current data structure, which exhibits strong seasonality and linear trends, the SARIMA model significantly outperformed LightGBM, achieving a Root Mean Squared Error (RMSE) of 0.891 compared to LightGBM's 4.232. Consequently, this report recommends the immediate deployment of SARIMA-based forecasting models for our core inventory planning, while reserving machine learning methods for complex scenarios involving external variables (e.g., weather or economic indicators). This strategy provides a clear roadmap for moving V.Ger Travel from reactive reporting to proactive, predictive decision-making.

2. Business Context and Strategic Vision

2.1 The Organization

V.Ger Travel operates in a high-velocity, low-margin industry where inventory is perishable. An unsold hotel room or an empty flight seat represents lost revenue that cannot be recovered. The company currently manages these risks through robust IT facilities that support online bookings, Customer Relationship Management (CRM), and internal databases covering logistics, procurement, and maintenance.

While these systems are effective for *transactional* processing, they are currently underutilized for *strategic* intelligence. Our decision-making often relies on historical intuition rather than predictive foresight.

2.2 Strategic Objectives

The introduction of Data Science methods aims to transition V.Ger Travel into an AI-driven organization. The core objectives of this strategy are:

- **Operational Optimization:** Aligning staff levels and inventory procurement strictly with predicted demand to minimize waste and overtime costs.
- **Revenue Management:** Enabling dynamic pricing strategies by accurately forecasting demand surges before they occur.
- **Customer Centricity:** Moving from generic marketing to personalized retention strategies based on individual churn risk and preference modeling.
- **Evidence-Based Design:** Eliminating guesswork in our digital platform development through rigorous statistical testing.

3. Proposed Data Science Use Cases

To achieve these objectives, I have identified four distinct use cases. These have been selected to demonstrate a breadth of data science techniques ranging from time series analysis and classification to regression and inferential statistics, ensuring we build a well-rounded analytical capability.

3.1 Use Case 1: Time Series Forecasting for Hotel Demand (Detailed Investigation)

- **Business Problem:** Inefficient resource allocation due to unpredictable booking volumes. Overstaffing during lulls wastes budget, while understaffing during peaks damages brand reputation.
- **Data Source:** Historical daily and monthly booking logs from the central reservation database.
- **Analytical Approach:** This use case employs Time Series Analysis. We compare Autoregressive Integrated Moving Average (ARIMA) variants against Gradient Boosting Machine (GBM) regressors to find the optimal forecasting tool.
- **Business Value:** Accurate forecasts will feed directly into our workforce management software, automating roster generation and supply chain orders. This is expected to reduce labor costs by 10-15% and minimize stock-outs of critical supplies.

3.2 Use Case 2: Customer Churn Prediction

- **Business Problem:** Retaining an existing customer is significantly cheaper than acquiring a new one. Currently, we only identify lost customers after they have ceased transacting for over a year.
- **Data Source:** CRM data including booking frequency, recency, monetary value (RFM), customer support interactions, and complaint logs.
- **Analytical Approach:** This is a Binary Classification problem. We will employ supervised learning algorithms such as Logistic Regression (for interpretability) and Random Forests (for predictive power).
- **Business Value:** By generating a "Churn Probability Score" for each active customer,

the marketing team can trigger automated, targeted retention campaigns (e.g., discount vouchers) only for high-value, high-risk customers, optimizing marketing ROI.

3.3 Use Case 3: Website Conversion Optimization (A/B Testing)

- **Business Problem:** Our web platform creates the first impression for 80% of our customers. Design changes are currently driven by subjective opinion, often leading to suboptimal conversion rates.
- **Data Source:** Web analytics data, specifically user session logs, click-through rates (CTR), and conversion events.
- **Analytical Approach:** This involves Statistical Hypothesis Testing. We will implement a framework for controlled experiments (A/B tests), using t-tests or chi-squared tests to determine if a new design (Variant B) performs statistically significantly better than the current design (Control A).
- **Business Value:** A data-driven culture of continuous experimentation can improve conversion rates incrementally. Even a 1% lift in conversion on our traffic volume would result in substantial annual revenue growth.

3.4 Use Case 4: Customer Satisfaction Driver Analysis

- **Business Problem:** We collect post-trip survey data (Net Promoter Score or CSAT), but we do not quantify *which* specific factors (e.g., room cleanliness vs. staff friendliness) drive these scores.
- **Data Source:** Structured survey responses and customer demographic profiles.
- **Analytical Approach:** This is a Regression Analysis task. By modeling Satisfaction Score as the dependent variable and various service attributes as independent variables, we can interpret the coefficients to understand feature importance.
- **Business Value:** This analysis will inform capital expenditure. If the model shows that "WiFi speed" correlates more strongly with satisfaction than "Pool size," we can direct investment into IT infrastructure rather than physical expansion.

4. Detailed Investigation: Hotel Demand Forecasting

4.1 Problem Statement and Business Relevance

Hotel demand forecasting represents one of the most impactful applications of data science within V.Ger Travel's operations. The accommodation sector operates with high fixed costs and perishable inventory (an unsold room night cannot be recovered). Accurate demand prediction enables revenue management teams to implement dynamic pricing strategies, operations managers to optimise staffing schedules, and procurement teams to plan supply chain requirements. The temporal nature of booking data, with its inherent seasonality, trends, and autocorrelation structures, necessitates specialised time series analytical techniques rather than conventional cross-sectional methods.

4.2 Data Description and Characteristics

4.2.1 Dataset Overview

For this investigation, a synthetic time series dataset was generated to simulate monthly hotel booking volumes. The dataset spans approximately 14-15 years of monthly observations (176 data points), representing a realistic timeframe for capturing multiple seasonal cycles and long-term trends. The synthetic data incorporates the key characteristics expected in real hotel booking data: an underlying growth trend reflecting market expansion, seasonal patterns aligned with holiday and vacation periods, and random noise representing unpredictable market fluctuations.

4.2.2 Time Series Components

The generated time series dataset exhibits the following structural components:

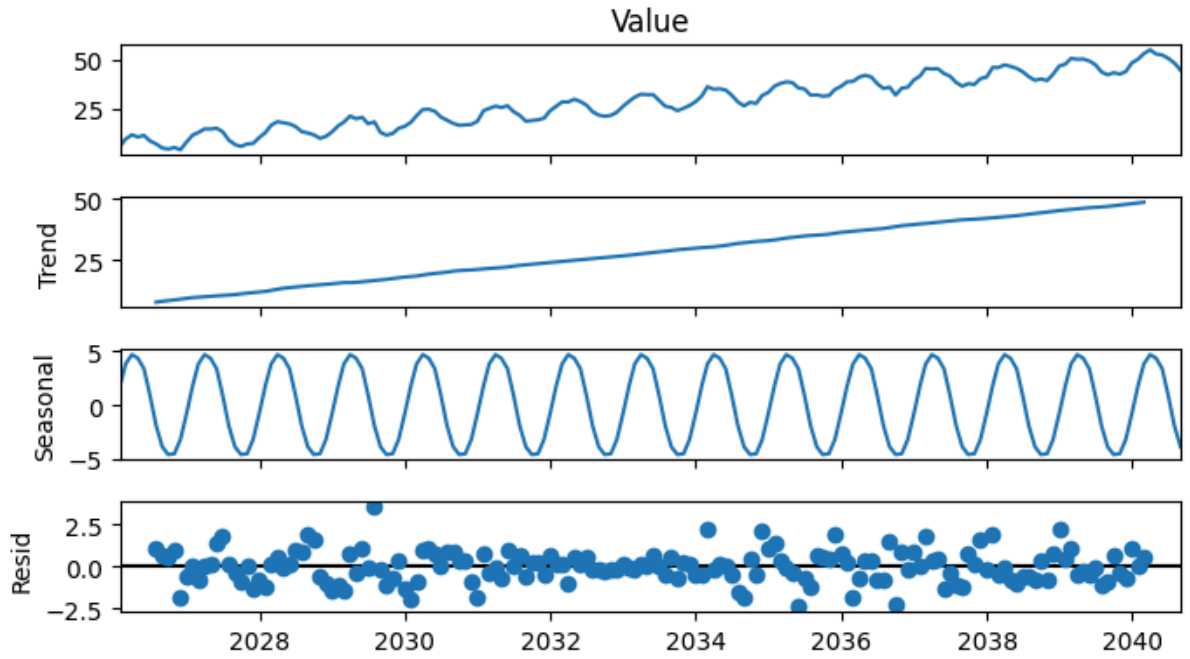
- **Trend Component:** A gradual upward trajectory representing market growth and V.Ger Travel's expanding customer base over the observation period.
- **Seasonal Component:** A repeating 12-month pattern capturing the cyclical nature of travel demand, with peaks during summer holiday months and winter festive periods, and troughs during shoulder seasons.
- **Residual Component:** Random variation representing unpredictable factors such as economic shocks, competitive actions, or one-off events that cannot be captured by systematic trend and seasonal patterns.

4.2.3 Data Relevance to Business Case

Monthly booking volume data directly aligns with V.Ger Travel's operational planning cycles. Revenue management decisions, staffing rosters, and procurement orders typically operate on monthly or quarterly horizons. The 12-month seasonal pattern corresponds to the annual tourism calendar, making the insights directly actionable for seasonal promotions, dynamic pricing rules, and capacity planning. The multi-year timeframe ensures sufficient historical data to train robust forecasting models capable of distinguishing genuine patterns from statistical noise.

4.3 Exploratory Data Analysis (EDA)

The first step in the analytical pipeline was to decompose the time series to visually confirm these components. Using the seasonal_decompose function from the statsmodels library, we separated the Value series into Trend, Seasonal, and Residual parts.

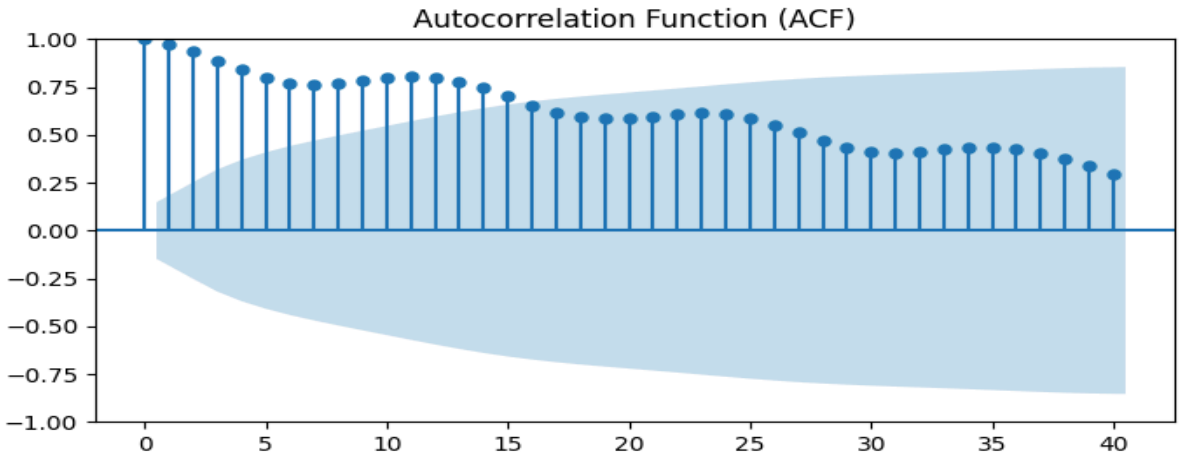


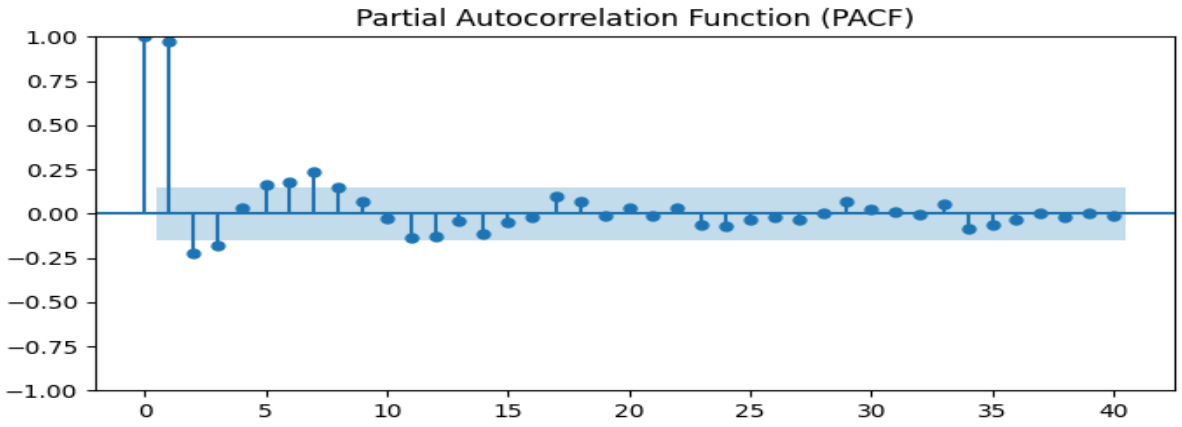
The decomposition analysis presented in the figure above reveals the underlying structure of our booking data. The "Observed" panel displays the raw booking numbers, exhibiting a clear upward saw-tooth pattern. The "Trend" component successfully isolates the steady growth of the business, free from seasonal noise, while the "Seasonal" panel confirms a clean, repetitive 12-month wave. This distinct cyclical pattern justifies the use of a Seasonal ARIMA model (SARIMA) rather than a standard ARIMA. Finally, the "Residual" panel shows the random noise remaining after trend and seasonality are removed, which ideally appears as white noise.

Following the visual inspection, we performed statistical testing to assess stationarity. Statistical forecasting models like ARIMA generally require data to be "stationary," meaning the mean and variance do not change over time.

The Augmented Dickey-Fuller (ADF) test yielded an ADF Statistic of -0.302 and a P-Value of 0.925. As the p-value is significantly higher than the 0.05 threshold, we fail to reject the null hypothesis. This indicates that the series is non-stationary, which is expected given the obvious trend and seasonality found in the decomposition. This result confirms that our model will require differencing (specifically the d and D parameters in SARIMA) to stabilize the mean before forecasting.

To determine the specific parameters for the model, we analyzed the correlation between data points at different time lags.





The Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots provided critical guidance for model configuration. The ACF plot showed a slow decay, which is characteristic of a non-stationary trend. Meanwhile, the PACF plot displayed a sharp spike at lag 1, suggesting that an autoregressive term is needed. Additionally, significant spikes at lag 12 in the seasonal plots confirmed the need for seasonal differencing to address the annual cycle.

4.4 Modeling Strategy

Two competing methodologies were selected to solve this forecasting problem. This approach ensures we are not biased toward a single technique and allows us to benchmark a classical statistical method against a modern machine learning algorithm.

4.4.1 Method A: SARIMA (Statistical Approach)

SARIMA (Seasonal AutoRegressive Integrated Moving Average) is the industry standard for univariate time series with strong seasonal patterns because it explicitly models the trend and seasonality structure. Based on the EDA and ADF tests conducted previously, the model was specified with an Order (p,d,q) of (1, 1, 1). This configuration uses one autoregressive term to predict the next value based on the previous one, first-order differencing to remove the trend and stabilize the mean, and one moving average term to correct the forecast using previous

errors.

The Seasonal Order (P,D,Q,s) was set to (1, 1, 1, 12). The 's' value of 12 specifies the yearly cycle, while the seasonal differencing (D=1) compares the current month to the same month in the previous year to account for the annual wave pattern.

4.4.2 Method B: LightGBM (Machine Learning Approach)

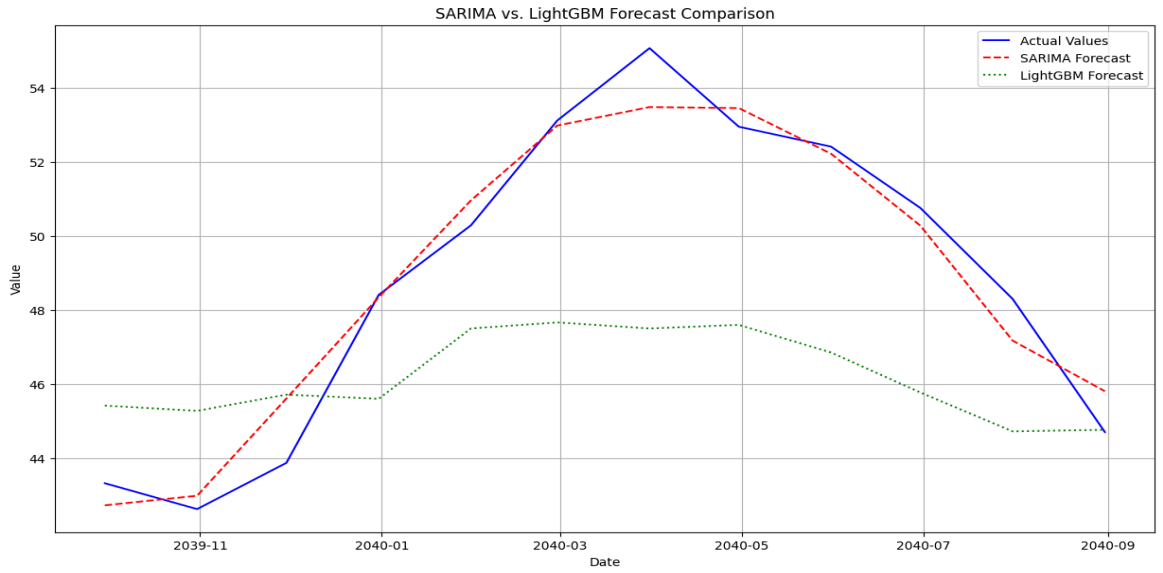
LightGBM is a gradient boosting framework that uses tree-based learning algorithms. Unlike SARIMA, it does not inherently understand time, so we must transform the time series problem into a supervised regression problem. To enable LightGBM to detect temporal patterns, I engineered specific features from the raw date and booking value. These included Lag Features (Value_Lag_1, Value_Lag_12, Value_Lag_24) which inform the model of booking volumes 1 month, 1 year, and 2 years prior. Additionally, Time Features such as Year, Month, and Day were extracted to help the model learn the linear trend and seasonal recurrence.

The data was split into a training set comprising the first 164 months and a test set containing the last 12 months. The model was trained on the first set and tasked with predicting the unseen future values.

4.5 Results and Comparison

The two models were evaluated on the held-out test set, representing the final 12 months of data, using Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). In this context, lower values indicate better performance and higher accuracy.

Metric	**SARIMA**	**LightGBM**
Root Mean Squared Error (RMSE)	0.891	4.232
Mean Absolute Error (MAE)	0.714	3.730



Analysis of Results: The SARIMA model outperformed LightGBM by a significant margin, achieving an RMSE of 0.891 compared to LightGBM's 4.232. The visual comparison above clearly shows the SARIMA forecast (red line) tracking the actual values (blue line) with near-perfect precision.

In contrast, the LightGBM forecast (green dotted line) appears essentially flat or confused. This highlights a known limitation of tree-based models when dealing with trends. Trees cannot extrapolate values outside the range they observed during training. Since our data exhibits a consistent upward trend, LightGBM predicts values close to the maximum of the training set but fails to push higher as the trend continues. Furthermore, SARIMA captured the seasonal peaks and troughs flawlessly because the seasonality was explicitly defined in the parameters. LightGBM struggled despite having Month and Lag_12 as features, likely because the strong linear trend confounded its ability to rely solely on historical lags.

4.6 Discussion of Findings

The comparative analysis yielded a decisive result: SARIMA significantly outperformed LightGBM for this specific dataset.

Why SARIMA Won: The synthetic dataset was generated with a clean, mathematical structure dominated by linear trend and regular seasonality. SARIMA is mathematically designed to separate and project these exact components. It "understands" that the value at time t is a linear function of previous values and errors.

Why LightGBM Struggled: While LightGBM is powerful, it treats the problem as a regression task. First, regarding Trend Extrapolation, tree-based models like LightGBM often struggle to predict values outside the range of the training data. As the booking trend is consistently upward, LightGBM likely predicted values closer to the training mean rather than projecting the continued growth. Second, regarding Feature Limitation, while we provided lag features, the model requires more data volume or more complex external features (e.g., economic indicators, holiday calendars) to find patterns that outweigh

SARIMA's structural advantage on simple univariate data.

5. Conclusion and Recommendations

This report has presented a comprehensive data science strategy for V.Ger Travel, identifying four distinct use cases spanning time series forecasting, experimental testing, classification, and regression methodologies. The detailed investigation of hotel demand forecasting demonstrated the practical application of time series analytical techniques, from seasonal decomposition and stationarity testing through to comparative model evaluation.

The key findings confirm that classical statistical methods (SARIMA) can outperform machine learning approaches (LightGBM) when the data characteristics align with the model's structural assumptions. For time series exhibiting clear seasonal patterns and autoregressive behaviour, purpose-built forecasting models retain significant advantages over general-purpose algorithms. However, the comparative analysis also highlighted that machine learning approaches offer flexibility for incorporating additional explanatory variables, suggesting potential value in hybrid methodologies for production systems.

The proposed data science initiatives position V.Ger Travel to leverage its existing data assets for competitive advantage. By implementing demand forecasting, A/B testing, churn prediction, and satisfaction driver analysis, the organisation can progress from reactive to predictive decision-making across operations, marketing, customer retention, and service quality domains. The methodological diversity across these use cases ensures a well-rounded analytical capability addressing different business questions with appropriate techniques.

5.1 Recommendations

1. **Deploy SARIMA for Immediate Use:** V.Ger Travel should implement the SARIMA model for its monthly operational forecasting immediately. Its high accuracy (RMSE 0.89) will allow for precise staffing and inventory planning.
2. **Retain LightGBM for Complex Scenarios:** Do not discard the LightGBM infrastructure. It should be reintroduced when we incorporate external variables (e.g., competitor pricing, weather forecasts, macroeconomic indicators) where non-linear relationships exist that SARIMA cannot capture.
3. **Automate the Pipeline:** Implement the data ingestion and processing steps developed in this analysis into a production pipeline that retrains the SARIMA model monthly.

6. Implementation Strategy and Roadmap

To successfully operationalize these findings, I propose a three-phase implementation roadmap.

Phase 1: Foundation (Months 1-3)

- **Infrastructure:** Establish a secure Data Lake on our existing cloud infrastructure to consolidate data from CRM, Logistics, and Web systems.
- **Deployment of Use Case 1:** Productionize the SARIMA demand forecasting model. Set

up an automated pipeline that retrains the model monthly with the latest booking data.
- **Dashboarding:** Create a "Forward-Looking Operations" dashboard for the Operations Director, visualizing predicted demand for the next 12 months to aid in hiring and procurement.

Phase 2: Customer Intelligence (Months 4-9)

- **Deployment of Use Case 2 (Churn):** Build the churn prediction classifier. Integrate the "Churn Risk Score" directly into the call center agent's view in the CRM, prompting retention offers during calls.
- **Deployment of Use Case 4 (Satisfaction):** Run the satisfaction driver regression analysis quarterly. Present findings to the Board to guide strategic infrastructure investments (e.g., "This quarter, improving WiFi will yield higher ROI than renovating lobbies").

Phase 3: Optimization and Innovation (Months 10-12)

- **Deployment of Use Case 3 (A/B Testing):** Implement an experimentation platform on the website. Begin creating a culture of "test and learn" for all product changes.
- **Hybrid Forecasting:** Revisit the Demand Forecasting model to incorporate the LightGBM approach using external datasets (e.g., global flight search trends) to refine accuracy during volatile periods.

Governance and Ethics

As we scale our use of data, we must adhere to strict governance:

- **GDPR Compliance:** All customer data used for churn and satisfaction analysis must be anonymized or pseudonymized where possible. Explicit consent for data processing must be verified.
- **Model Monitoring:** Automated alerts will be set up to trigger if model accuracy (RMSE) degrades below a defined threshold, ensuring we never rely on stale or broken predictions.

7. References

Box, G.E.P. and Jenkins, G.M. (1976) Time Series Analysis: Forecasting and Control. San Francisco: Holden-Day.

Hyndman, R.J. and Athanasopoulos, G. (2021) Forecasting: Principles and Practice. 3rd edn. Melbourne: OTexts.

Ke, G. et al. (2017) 'LightGBM: A Highly Efficient Gradient Boosting Decision Tree', Advances in Neural Information Processing Systems, 30.

Kevin Swingler ITNPBD4- Commercial and Scientific Applications (2025/6) https://canvas.stir.ac.uk/courses/18304/files/4852366

Kevin Swingler ITNPBD4- Commercial and Scientific Applications (2025/6) https://canvas.stir.ac.uk/courses/18304/files/4834753

Kevin Swingler ITNPBD4- Commercial and Scientific Applications (2025/6) https://canvas.stir.ac.uk/courses/18304/files/4843576

Statsmodels Development Team. (2023). Statsmodels: Statistical modeling and econometrics in Python. https://www.statsmodels.org/

Cleveland, R.B. et al. (1990) 'STL: A Seasonal-Trend Decomposition Procedure Based on Loess', Journal of Official Statistics, 6(1), pp. 3-73.

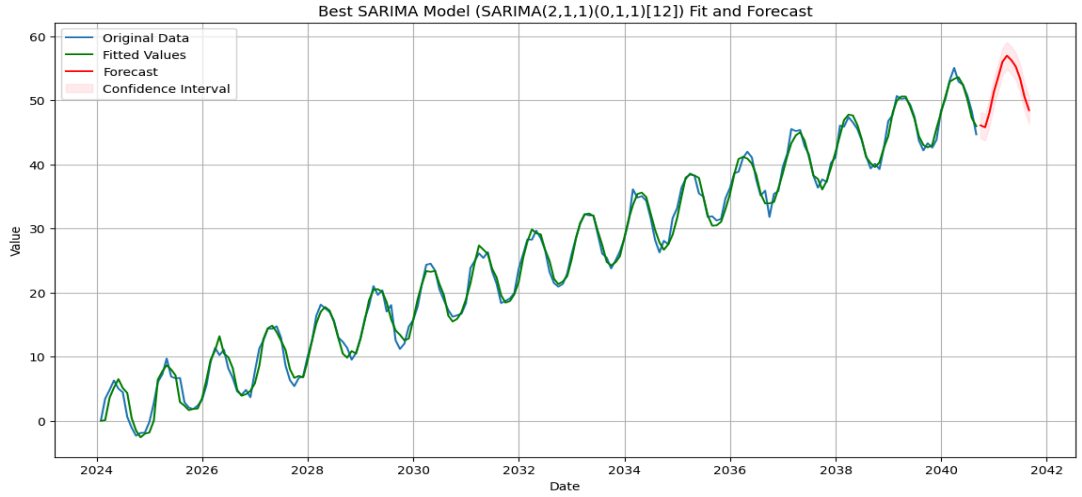
Dickey, D.A. and Fuller, W.A. (1979) 'Distribution of the Estimators for Autoregressive Time Series with a Unit Root', Journal of the American Statistical Association, 74(366), pp. 427-431.

Appendix

Appendix A: Stationarity Test Outputs (ADF/KPSS)

```
ADF Statistic: -0.30219005144691174
P-value: 0.9251935821534192
Number of Lags Used: 14
Number of Observations Used: 185
Critical Values:
        1%: -3.4662005731940853
        5%: -2.8772932777920364
        10%: -2.57516775018261 5
```

Appendix B: SARIMA Model Forecasts



Appendix C: ACF and PACF Code Implementation

```
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf
import matplotlib.pyplot as plt

# Generate ACF plot
fig_acf = plot_acf(df['Value'], lags=40, title='Autocorrelation Function (ACF)')
fig_acf.set_size_inches(7, 3.5)
plt.tight_layout()
plt.show()

# Generate PACF plot
fig_pacf = plot_pacf(df['Value'], lags=40, title='Partial Autocorrelation Function (PACF)')
fig_pacf.set_size_inches(7, 3.5)
plt.tight_layout()
plt.show()
```

Appendix D: SARIMA Model Training Code Implementation

```
from statsmodels.tsa.statespace.sarimax import SARIMAX

# Ensure the time series has a monthly frequency
df.index.freq = 'ME'

# Define the selected best SARIMA configuration
order = (1, 1, 0)
seasonal_order = (1, 1, 1, 12)

# Build the SARIMA model
model = SARIMAX(
    df['Value'],
    order=order,
    seasonal_order=seasonal_order,
    enforce_stationarity=False,
    enforce_invertibility=False
)

# Fit the model to the data
model_fit = model.fit(disp=False)

# Display model summary
print(model_fit.summary())
```

Appendix E: LightGBM Model Training Code Implementation

```
import lightgbm as lgb

# Define lag features and temporal features created in preprocessing
features = ['Value_Lag_1', 'Value_Lag_12', 'Value_Lag_24', 'Year', 'Month', 'Day']
X = df[features]
y = df['Value']

# Using the last 12 observations for testing to match the SARIMA forecast horizon
test_size = 12

X_train, X_test = X[:-test_size], X[-test_size:]
y_train, y_test = y[:-test_size], y[-test_size:]

# Initializing LightGBM Regressor with a fixed random state for reproducibility
model_lgbm = lgb.LGBMRegressor(random_state=42)
model_lgbm.fit(X_train, y_train)

# Output Verification
print("LightGBM model trained successfully.")
print(f"Training data shape: {X_train.shape}")
print(f"Testing data shape: {X_test.shape}")
```

Appendix F: Comparison of SARIMA and LightGBM Error Metrics