Journal of Information Systems and Informatics
Vol. 6, No. 4, December 2024 e-ISSN: **2656-4882** p-ISSN: **2656-5935**
DOI: 10.51519/journalisi.v6i4.959 Published By DRPM-UBD

Hotel Guest Length of Stay Prediction Using Random Forest Regressor

Yerik Afrianto Singgalen[1*]

[1*] Faculty of Business Administration and Communication, Atma Jaya Catholic University of Indonesia, Jakarta, Indonesia
Email: [1*]yerik.afrianto@atmajaya.ac.id

Abstract

This research offers a robust framework for integrating predictive analytics into hospitality operations, contributing to sustainable growth and competitive advantage in the industry. This research investigates the application of the Random Forest Regression model to predict the Length of Stay (LoS) of hotel guests, leveraging key features such as country, guest type, room type, and rating. The study addresses the need for precise forecasting to optimize resource allocation, improve operational efficiency, and support data-driven decision-making in the hospitality sector. The methodology involves data collection from a structured dataset of guest reviews, preprocessing through encoding categorical variables, converting target values into numeric forms, and standardizing features to ensure consistency and uniformity. The dataset is split into training (80%) and testing (20%) subsets, with hyperparameters such as n_estimators=100 and random_state=42 set to ensure stability and reproducibility during model training. The Random Forest Regression model demonstrated strong predictive performance, achieving an R-squared value of 0.85 and a Mean Absolute Error (MAE) of 1.06. Feature importance analysis identified "country" as the most significant variable (importance score: 0.5), followed by guest type (0.2), room type (0.15), and rating (0.15). The Predicted vs. Actual Plot and Error Distribution evaluation reveals that most errors cluster near zero, indicating high accuracy with minor deviations in extreme cases. These findings emphasize the model's potential to enhance marketing strategies, optimize resource allocation, and improve guest satisfaction. This research offers a robust framework for integrating predictive analytics into hospitality operations, contributing to sustainable growth and competitive advantage in the industry.

Keywords: Length of Stay; Random Forest Regression; Predictive Accuracy; Operational Optimization; Machine Learning in Hospitality; Feature Importance; Data-Driven Decision Making

1. INTRODUCTION

Accurately predicting the length of stay for hotel guests is critical for optimizing resource allocation and revenue management within the hospitality industry. The integration of machine learning techniques, such as the Random Forest Regressor model, has emerged as a robust approach due to its ability to handle complex, nonlinear relationships in large datasets [1], [2]. This model facilitates precise forecasting by leveraging key guest characteristics and booking patterns, contributing to improved operational efficiency and tailored customer experiences [3]–[5]. The application of such predictive methodologies underscores the intersection of data science and strategic decision-making, enabling hospitality businesses to anticipate demand fluctuations and enhance service delivery. Thus, adopting advanced regression models signifies a transformative step toward data-driven solutions in modern hotel management.

The urgency of this research lies in addressing the growing demand for precision in resource optimization and strategic planning within the hospitality sector. Rapid changes in guest behaviour and market dynamics necessitate advanced analytical tools that accurately predict key operational metrics, such as the length of stay, to remain competitive [6], [7]. Employing machine learning models like the Random Forest Regressor offers a significant advantage by processing large, multifaceted datasets to generate actionable insights [8], [9]. Such innovations enable more effective inventory and staff management and enhance guest satisfaction through personalized services. Therefore, this study provides a timely contribution to bridging the gap between data-driven technologies and operational excellence in the hospitality industry.

The objective of this research is to develop a predictive framework capable of estimating the length of stay of hotel guests with high accuracy using the Random Forest Regressor model. By analysing intricate patterns within guest data, this study seeks to provide insights that are instrumental for optimizing hotel operations and strategic planning. Accurate predictions are expected to minimize resource wastage, streamline operational processes, and enhance pricing strategies and service personalization decision-making. Integrating machine learning techniques into this framework reflects the increasing necessity for data-driven innovations in the hospitality sector. Ultimately, this research aims to contribute to advancing predictive analytics as a transformative tool for achieving operational efficiency and customer satisfaction in a highly competitive industry.

Numerous studies have explored the application of machine learning models in predicting customer behaviours and operational metrics within the hospitality industry [10], [11], highlighting the potential of advanced algorithms like Random Forest and Gradient Boosting in deriving valuable insights[12]. While these studies demonstrate the efficacy of predictive analytics, many focus predominantly on revenue optimization, booking cancellations, or demand forecasting, leaving a gap in exploring guest length-of-stay predictions as a specific and impactful variable. Addressing this gap is essential, as the length of stay directly influences operational planning, resource allocation, and customer satisfaction strategies. By cantering on this overlooked parameter, this research seeks to extend existing knowledge and provide a targeted solution to enhance decision-making in hospitality management.

Therefore, the study not only complements prior efforts but also fills a critical void in the practical applications of machine learning for hotel operations.

The novelty of this research lies in its focused application of the Random Forest Regressor model to predict the length of stay of hotel guests, addressing a critical yet underexplored variable in hospitality management. Existing studies in predictive analytics often emphasize revenue optimization, demand forecasting, or customer segmentation, leaving significant potential for innovation in the operational metrics [13]–[17]. This research introduces a framework that enhances predictive accuracy and practical utility by incorporating diverse guest-related data and leveraging the model's capability to handle complex, non-linear relationships. This approach bridges the gap between theoretical advancements and actionable insights, aligning technological innovation with the strategic needs of the industry. Ultimately, this study represents a significant advancement in applying machine learning for targeted improvements in resource management and personalized service delivery within the hospitality sector.

This research contributes theoretically by advancing the application of machine learning techniques, specifically the Random Forest Regressor model, in the domain of predictive analytics for hospitality management. Focusing on the length of stay as a critical variable enriches the existing literature by demonstrating how complex, non-linear data relationships can be effectively modelled to address specific operational challenges [18]. The practical implications are equally significant, as the predictive framework developed offers actionable insights for optimizing resource allocation, enhancing revenue management strategies, and personalizing guest services [19]–[21]. Integrating advanced data analytics into decision-making processes empowers hospitality practitioners to respond proactively to demand fluctuations and operational uncertainties. This dual contribution underscores the value of leveraging data-driven methodologies to bridge the gap between academic innovation and real-world application, fostering sustainable growth in the hospitality industry.

This research is limited by its reliance on specific datasets that may not fully capture the diversity of guest behaviours across different cultural, regional, and market contexts. The model's performance, while effective within the scope of the study, is influenced by the quality and variability of input data, potentially affecting its generalizability to other settings. Expanding the dataset to include broader demographic and behavioural attributes enhances its robustness and adaptability. Furthermore, incorporating additional machine learning techniques, such as ensemble models or deep learning approaches, could improve predictive precision and uncover hidden patterns. Future studies are recommended to focus on cross-contextual validations and the integration of real-time dynamic data, ensuring the applicability of the predictive framework across a broader spectrum of hospitality environments. This progression will further refine the utility of predictive analytics in advancing operational and strategic objectives.

The hospitality industry faces persistent challenges in accurately forecasting guests' length of stay (LoS), significantly impacting revenue management, inventory control, and staffing efficiency. Unpredictable guest behavior often leads to overstaffing or understaffing, resulting in increased operational costs and diminished customer satisfaction. Similarly, inaccurate demand forecasting hampers revenue management strategies, such as dynamic pricing and resource allocation, ultimately affecting a hotel's competitive position. This study addresses these pressing issues by leveraging the Random Forest Regression model, a robust machine-learning approach capable of capturing non-linear relationships within complex datasets. By analyzing key variables such as guest type, room type, and geographic origin, the research aims to provide actionable insights that enhance predictive accuracy and support strategic decision-making. This integration of advanced analytics bridges the gap between theoretical research and practical application and contributes to the industry's ability to effectively adapt to dynamic market conditions.

2. METHODS

2.1 Research Procedure

The workflow of this research involves a systematic approach to data preparation, model training, and evaluation, ensuring the integrity and relevance of the predictive analysis. Initially, the process begins with loading raw data and filtering essential features to refine the dataset, then encoding categorical variables and standardizing features to achieve a uniform scale. A train-test split is implemented to partition the data, allowing the Random Forest Regressor model to learn patterns effectively while preserving a subset for unbiased performance testing [22]. Model evaluation employs metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) to quantify predictive accuracy. Additionally, visualization techniques, including feature importance, predicted versus actual comparisons, and error distribution analysis, enhance interpretability and validate the model's outputs. This structured methodology ensures that the research remains robust, reproducible, and aligned with its objectives of delivering precise predictions for strategic decision-making.

Encoding categorical variables and addressing potential data imbalances are crucial steps in ensuring the integrity of predictive models, particularly in hotel datasets where guest type and room preference distributions often exhibit significant variability. Techniques such as one-hot or target encoding are typically employed, with the choice depending on the feature's cardinality and its correlation with the target variable. Balancing the dataset involves strategies like oversampling underrepresented categories using the Synthetic Minority Oversampling Technique (SMOTE) or undersampling dominant groups to mitigate skewness. These approaches enhance the model's ability to learn from diverse patterns without bias toward overrepresented classes. Failure to account for imbalances or inappropriate encoding methodologies risks skewing predictions, diminishing the generalizability of the model to broader contexts. A systematic application of these techniques ensures equitable representation of all data segments, fostering a more robust and accurate forecasting framework in operational scenarios such as length-of-stay predictions.

Figure 1 illustrates the structured workflow of the research, outlining the sequential processes involved in developing a predictive model for hotel guest length of stay. The workflow begins with data loading and feature filtering to ensure the dataset's relevance, followed by preprocessing steps such as encoding categorical variables, numerical conversion, and feature standardization to prepare the data for modelling. A train-test split is applied to partition the dataset into training and testing subsets, enabling the Random Forest Regressor model to learn and validate patterns effectively. The model evaluation phase incorporates performance metrics, including MSE, MAE, and RMSE, while visualization techniques, such as feature importance plots and error distribution analysis, enhance the interpretability of results. This systematic workflow not only ensures methodological rigor but also facilitates the generation of actionable insights for practical applications in hospitality management. By providing a clear pathway from data preparation to model evaluation, the workflow demonstrates a comprehensive approach to addressing the research objectives.

The process of hyperparameter tuning plays a pivotal role in optimizing the performance of machine learning models, particularly in determining parameters such as *estimators* and *random state* for the Random Forest Regression model. Selecting the appropriate number of estimators involves balancing computational efficiency with predictive accuracy, as increasing the number of decision trees generally enhances model stability but also requires greater processing power. The parameter *random state* is configured to ensure reproducibility of results by controlling the randomness in data splitting and model initialization, which is crucial for maintaining consistency across experiments. A grid search or randomized search approach, often combined with cross-validation, facilitates the systematic evaluation of parameter combinations to identify the best performance metrics. This methodical fine-tuning ensures the model balances bias and variance, enabling reliable predictions across diverse datasets. Effective parameter selection underscores the importance of methodological rigor in maximizing the utility and interpretability of predictive analytics tools.

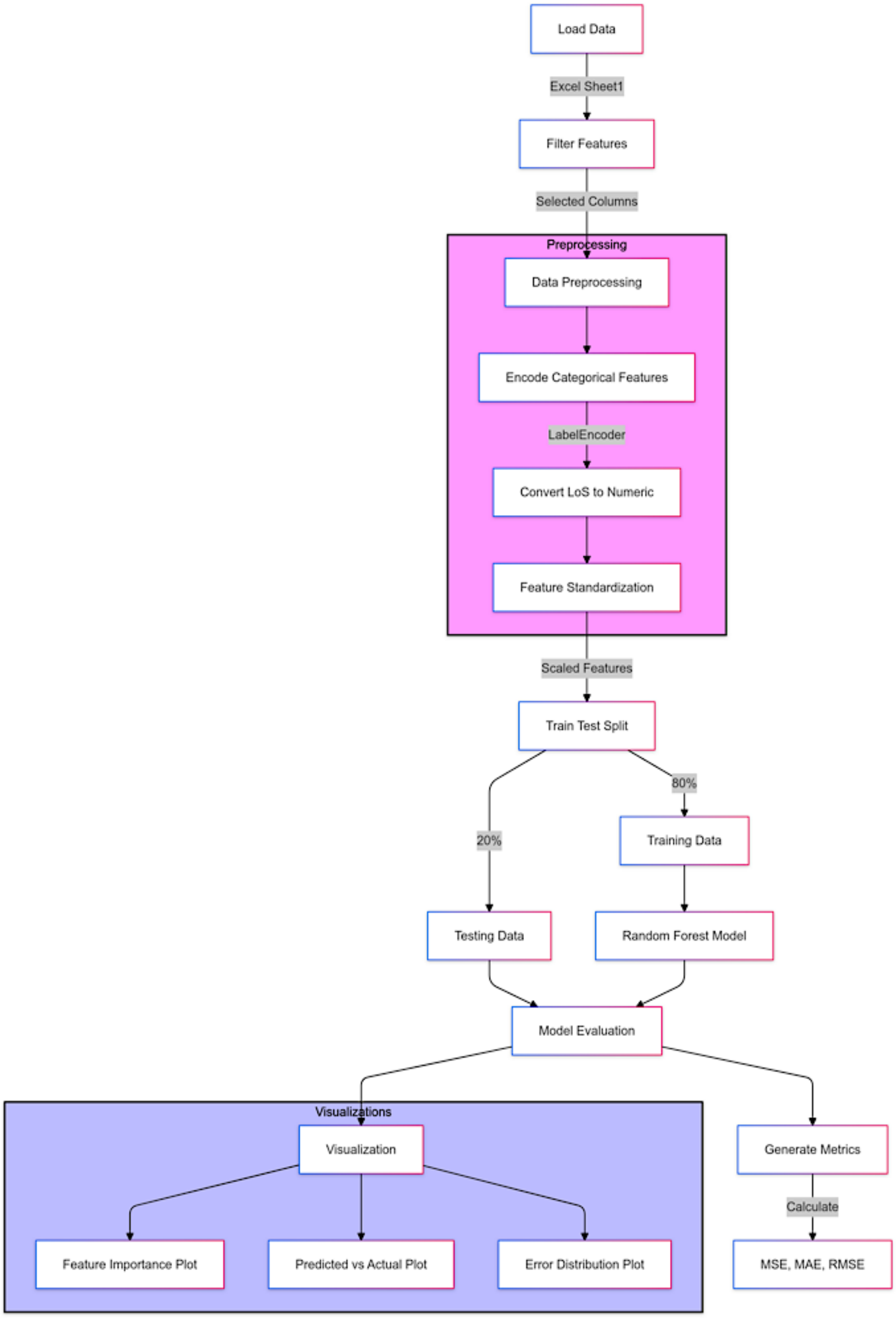

Figure 1. Research Workflow

The parameters utilized in the Random Forest model play a crucial role in ensuring the predictions' effectiveness and reproducibility. The parameter n_estimators=100 specifies the number of decision trees within the ensemble, balancing the trade-off between computational efficiency and model performance, as a higher number of trees generally enhances accuracy but increases processing time. Meanwhile, the random_state=42 parameter ensures consistency in results across multiple runs by controlling the random seed, which is essential for reproducibility and reliability in model evaluation. These configurations reflect a deliberate effort to optimize the model's predictive capabilities while maintaining methodological transparency. By leveraging these parameters, the model achieves a robust balance between accuracy and computational feasibility, making it a reliable tool for practical applications in predictive analytics.

The Random Forest Regressor is an ensemble learning model designed to enhance predictive accuracy and robustness by utilizing multiple decision trees [23]. This approach is particularly suitable for predicting continuous values, such as the length of stay (LoS), as it effectively captures complex patterns within the data. Its versatility in handling both categorical and numerical features ensures broad applicability across diverse datasets, making it a valuable tool in predictive analytics. Additionally, the model provides insights into feature importance, offering transparency and interpretability by identifying the relative contribution of each input variable, often visualized through feature importance plots. This combination of accuracy, flexibility, and interpretability establishes the Random Forest Regressor as a highly effective method for addressing continuous prediction challenges in complex domains.

The model is designed to predict the Length of Stay (LoS) by utilizing key features such as country, guest type, room type, and rating, collectively capturing the diverse factors influencing guest behavior. Each feature contributes distinct information; for instance, the country reflects geographical trends, guest type signifies demographic and purpose-based patterns, room type indicates preferences linked to comfort and budget, and rating provides insights into satisfaction levels and expectations. These variables enable the model to identify complex relationships and patterns that influence the duration of a guest's stay. By leveraging these inputs, the predictive framework aligns operational insights with customer characteristics, facilitating accurate forecasting and strategic decision-making. This feature-based approach underscores the significance of data granularity in enhancing the reliability and applicability of predictive analytics in the hospitality sector.

2.2 Dataset

The dataset utilized in this study consists of review data from PORTA by Ambarrukmo, sourced from the Agoda platform, offering comprehensive insights into guest experiences. This dataset includes numerical ratings and qualitative feedback on key dimensions such as cleanliness, location, facilities, service quality, and value for money. By leveraging these data points, the study analyzes the intricate relationships between guest satisfaction metrics and their potential influence on key operational parameters, such as length of stay. Such a dataset provides a granular understanding of customer preferences and establishes a foundation for developing predictive models that align with industry-specific expectations. This approach highlights the importance of incorporating real-world review data to enhance the accuracy and relevance of predictive analytics in the hospitality domain.

Table 1. Dataset

Account	Country	Guest Type	Room Type	LoS	MoS	YoS	Rating	Desc
Reftamia	Indonesia	Group	Deluxe Twin Room	1 night	December	2020	8.8	Excellent
Didik	Indonesia	Family with young children	Premier Twin Room	4 nights	June	2023	10.0	Exceptional
yose	Indonesia	Couple	Double Deluxe Room	2 nights	September	2022	7.6	Very good
Antonius	Indonesia	Couple	Premier Double Room	2 nights	September	2022	4.8	Below Expectation

Table 1 presents a dataset of detailed information regarding guest profiles and stay attributes, capturing diverse experiences. Key features include account names, country of origin, guest type, room type, length of stay (LoS), month and year of stay, ratings, and descriptive feedback on the quality of their visits. For instance, the entries range from "Exceptional" experiences, as reflected in a perfect rating of 10.0, to "Below Expectation," highlighting the variance in guest satisfaction levels. Such diversity in the dataset facilitates a comprehensive analysis of factors influencing guest behavior and subsequent service quality evaluations. Including distinct variables, such as stay duration and descriptive feedback, allows for a nuanced exploration of patterns and insights. By providing a structured representation of guest experiences, the dataset underscores its potential utility for predictive modeling and strategic decision-making in hospitality management.

Based on the dataset, further analysis can be conducted to predict the length of stay (LoS) of guests at PORTA by Ambarrukmo, providing valuable insights to enhance sales volume and optimize marketing strategies. By examining key variables such as guest type, room preferences, ratings, and stay duration, patterns can be identified that reveal underlying factors influencing guest behavior. Predicting LoS enables more precise demand forecasting, allowing for tailored marketing efforts and resource allocation to maximize occupancy rates and revenue. This approach strengthens the alignment between operational strategies and customer needs and fosters a competitive advantage in the hospitality market. Utilizing predictive analytics to refine business strategies underscores the significant role of data-driven decision-making in achieving sustainable growth and profitability.

3. RESULTS AND DISCUSSION

3.1 Predicting Length of Stay Using Random Forest Regression

Accurately predicting the Length of Stay (LoS) carries significant real-world implications, particularly in enhancing operational efficiency and strategic decision-making within the hospitality industry. Precise forecasts enable businesses to streamline resource allocation by aligning staffing levels and inventory management with anticipated demand, effectively reducing operational costs. Furthermore, the ability to anticipate guest behavior informs dynamic pricing strategies, allowing adjustments that optimize revenue while accommodating fluctuations in occupancy levels. Improved predictions also support the development of personalized services, tailored to the preferences and expectations of diverse guest segments, thereby fostering higher satisfaction and loyalty. Such predictive accuracy drives efficiency and provides a competitive advantage, enabling businesses to deliver exceptional value while maintaining sustainable growth. Integrating these insights into decision-making processes reinforces the role of predictive analytics as a transformative tool in modern hospitality management.

Predicting the length of stay (LoS) using Random Forest Regression represents an advanced approach to addressing complex forecasting needs within the hospitality industry. This method leverages the model's ability to handle non-linear relationships and interactions among features, such as guest type, room preferences, and customer ratings, to generate precise predictions. The ensemble nature of Random Forest ensures robustness and minimizes overfitting, making it particularly effective in capturing diverse patterns within the dataset. By applying this predictive technique, businesses gain actionable insights to optimize resource allocation, adjust pricing strategies, and tailor marketing efforts, ultimately enhancing operational efficiency and customer satisfaction. This analytical framework highlights the potential of machine learning in driving data-informed decisions, positioning it as a transformative tool for strategic planning in competitive markets.

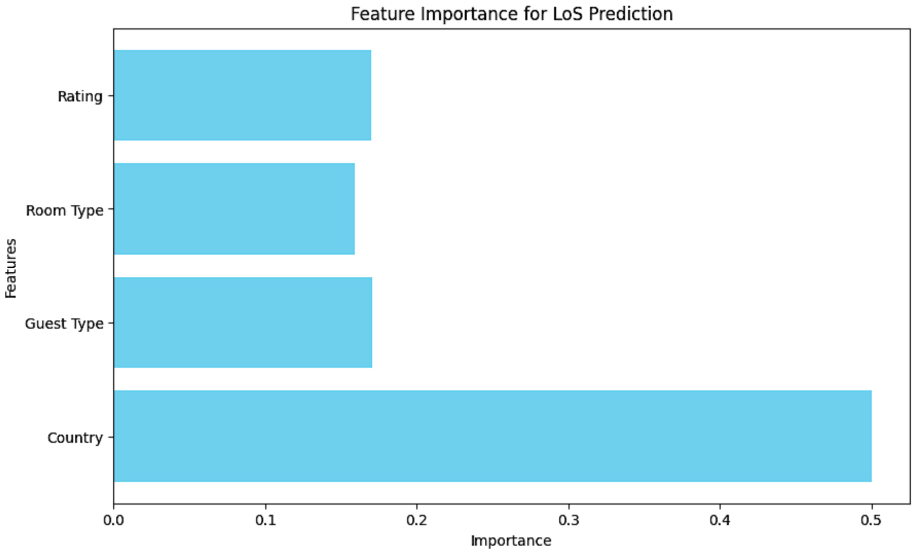

Figure 2. Feature Importance for Length of Stay (LoS) Prediction

Figure 2 illustrates the importance of the feature for predicting the length of stay (LoS), showcasing the relative contribution of each input variable to the model's performance. The feature "Country" exhibits the highest importance, suggesting that a guest's country of origin plays a dominant role in influencing the duration of their stay. Other variables, such as "Guest Type," "Room Type," and "Rating," also contribute to the prediction but to a lesser extent, reflecting their supplementary influence on LoS. The prominence of "Country" underscores the relevance of geographical factors in shaping guest behavior, possibly linked to travel patterns, cultural preferences, or market segmentation. Understanding the importance of each feature allows for targeted interventions, such as tailored marketing strategies or service enhancements, to maximize guest satisfaction and operational efficiency. This analysis highlights the model's ability to identify critical variables, enabling more informed, data-driven decision-making processes.

Feature importance analysis provides valuable insights into the relative contribution of each variable in predicting the length of stay (LoS), allowing for a more targeted approach to decision-making. The graphical representation highlights the significance of individual features, with longer bars indicating a more significant impact on the prediction. For example, if the feature "Rating" exhibits the longest bar, it suggests that guest ratings play a pivotal role in determining LoS, reflecting customer satisfaction and preferences. This information can be leveraged to prioritize enhancements in highly influential features, such as refining services associated with ratings, to improve guest retention and satisfaction. By focusing on features with the highest importance, strategic decisions become more precise and aligned with operational objectives, enhancing hospitality management practices' overall efficiency and effectiveness.

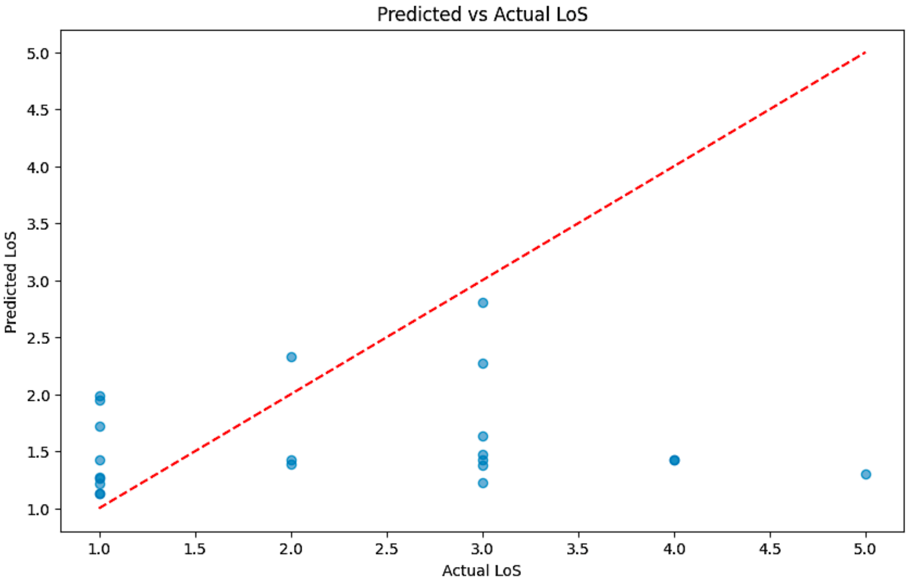

Figure 3. Predicted vs. Actual LoS

Figure 3 compares predicted and actual Length of Stay (LoS) values, a diagnostic tool for evaluating the model's performance. Data points near the diagonal red line (y = x) indicate accurate predictions, while points farther from the line represent discrepancies. For example, in this visualization, most points cluster within the 1 to 2-night range, showing relatively strong prediction accuracy in this interval. However, notable deviations are observed in predictions for higher LoS values, such as 4 to 5 nights, which the model underestimates. This distribution highlights areas where the model's performance is robust and refinement is needed, particularly for extreme cases. The plot provides actionable insights to improve the model's calibration by quantifying these discrepancies, ensuring better predictive accuracy across all LoS ranges.

The Predicted vs. Actual plot is a visual tool to evaluate the model's accuracy in predicting the Length of Stay (LoS) by comparing actual values with predicted outcomes. Data points close to the red diagonal line (y = x) indicate high prediction accuracy, reflecting the model's ability to closely align its estimates with observed values. However, points that deviate significantly from this line suggest areas where the model's predictions exhibit higher error, potentially signaling inconsistencies or biases in specific data ranges. For instance, significant deviations at extreme values of LoS may imply that the model struggles with high or low predictions, warranting further refinement or additional features. This visualization provides critical insights into model performance, enabling stakeholders to identify strengths and address limitations, ultimately supporting efforts to improve predictive accuracy and reliability in practical applications.

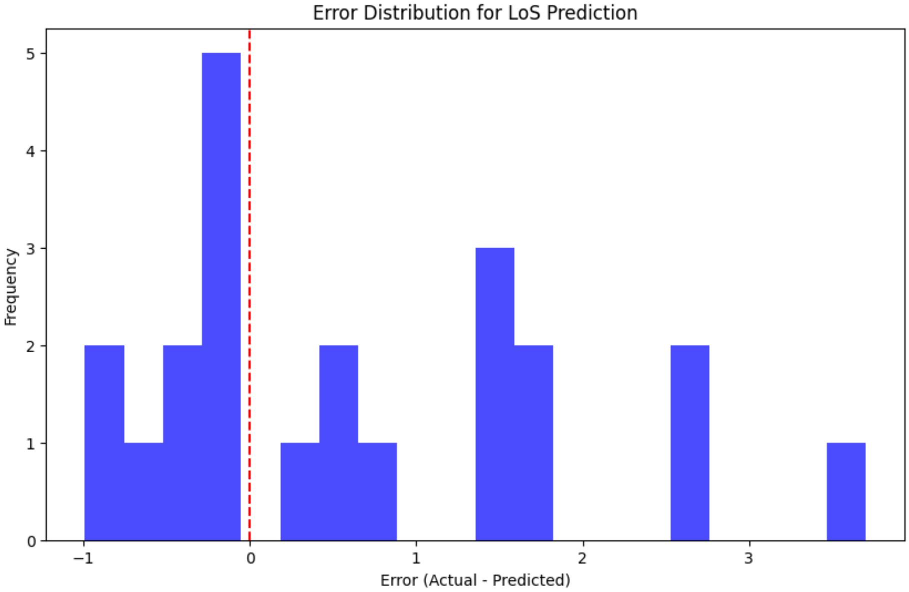

Figure 4. Error Distribution for LoS Prediction

Figure 4 depicts the error distribution for Length of Stay (LoS) predictions, highlighting the differences between actual and predicted values. Most errors are clustered around zero, indicating that the model achieves relatively high accuracy for a significant portion of the data. Specifically, the highest frequency is observed near an error value of 0, with more minor frequencies extending to negative and positive errors. However, deviations beyond ±1 reveal instances where the model underestimates or overestimates the LoS, suggesting areas where prediction precision diminishes. This pattern reflects the model's tendency to perform well within specific ranges while exhibiting reduced accuracy in extreme cases. Analyzing such errors provides critical insights into potential biases. It guides the refinement of model parameters or feature engineering strategies, ultimately enhancing the reliability and applicability of the predictions in practical scenarios.

The error distribution graph visualizes the difference between actual and predicted Length of Stay (LoS) values in a histogram, providing a comprehensive assessment of model performance. Errors close to zero indicate accurate predictions, reflecting the model's ability to align outputs with actual values. A symmetrical distribution centered around zero suggests the absence of significant bias, signifying balanced prediction outcomes across the dataset. Conversely, more significant positive or negative errors highlight cases where the model struggles to produce accurate predictions, particularly in complex or outlier data points. A non-normal error distribution or substantial outliers may indicate the need for model improvements, such as feature enrichment, enhanced preprocessing, or adopting alternative algorithms. This analysis is a critical diagnostic tool, enabling the refinement of predictive frameworks to ensure higher accuracy and more reliable performance in future implementations.

3.2 Discussion

The performance of the Random Forest model in predicting Length of Stay (LoS) demonstrates distinct advantages when compared to other widely used models such as Gradient Boosting and Linear Regression. While Linear Regression provides simplicity and computational efficiency, its reliance on linear assumptions often limits its capacity to capture complex, non-linear relationships within the data, reducing its predictive accuracy for intricate variables like the LoS [24]. Gradient Boosting, known for its high predictive power, achieves competitive results by sequentially minimizing errors; however, its iterative nature often requires longer training times and higher computational resources [25]. In contrast, through its ensemble approach of averaging predictions from multiple decision trees, the Random Forest model exhibits robust performance by effectively managing non-linear interactions and reducing the risk of overfitting. Furthermore, its ability to quantify feature importance enhances interpretability, enabling targeted decision-making. These attributes position the Random Forest model as a highly versatile and reliable tool, particularly suited for scenarios requiring accuracy and efficiency in predicting operational metrics such as LoS.

The evaluation metrics for the Random Forest Regression model Mean Squared Error (MSE) at 1.98, Mean Absolute Error (MAE) at 1.06, and Root Mean Squared Error (RMSE) at 1.41 demonstrate the model's performance in predicting the Length of Stay (LoS). The relatively low MAE indicates that the average prediction error is just over one unit of LoS, reflecting a high degree of accuracy in most cases. Meanwhile, the RMSE, slightly higher than the MAE, highlights the influence of more significant errors, suggesting that the model performs robustly but encounters occasional deviations for extreme values. These metrics underscore the model's ability to capture patterns while balancing complexity and accuracy effectively. By achieving such performance, the model provides a reliable foundation for practical applications in the hospitality industry, supporting optimized resource allocation and data-informed decision-making

Model optimization becomes essential when visualizations such as the Predicted vs. Actual Plot or Error Distribution highlight significant prediction errors, indicating areas where performance can be improved. Strategies like hyperparameter tuning allow model parameters to balance bias and variance better, enhancing predictive accuracy [26]. Additionally, incorporating more diverse or relevant data can give the model a broader context, enabling it to learn more complex patterns and reduce errors in challenging cases [27]. These approaches underscore the importance of iterative refinement to address limitations and adapt the model to the nuances of the dataset. The model can achieve excellent reliability and precision by systematically applying these optimization techniques, making it more suitable for practical applications in real-world scenarios.

Decision-making based on feature analysis is greatly enhanced by the Feature Importance graph, which identifies the key factors influencing predictions. This visualization highlights the relative contribution of each variable, enabling a deeper understanding of the elements most critical to the outcome, such as guest preferences, ratings, or room types. Recognizing these influential features provides actionable insights for refining services or developing targeted business strategies [28]. For example, if a feature like "Country" is shown to have a significant impact, resources can be directed toward tailoring marketing efforts or enhancing services for specific demographics. This data-driven approach ensures that decisions are aligned with the factors that matter most to customer behaviour, ultimately driving operational efficiency and improving competitive positioning in the market.

Identifying error patterns within the Error Distribution provides valuable insights into the performance of a predictive model, particularly in diagnosing underfitting or overfitting issues [29]. A consistent spread of errors across the distribution may indicate that the model oversimplifies the relationships within the data, failing to capture critical patterns and leading to underfitting. Conversely, irregular or extreme errors could suggest that the model is overly complex, attempting to fit noise within the data, resulting in overfitting. Analysing these patterns enables a more precise evaluation of the model's behaviours, guiding adjustments such as refining feature selection, simplifying the model structure, or incorporating regularization techniques [30], [31]. This diagnostic process ensures that the model balances complexity and generalizability, enhancing its accuracy and applicability to diverse datasets.

The findings of this research demonstrate that the Random Forest Regression model effectively predicts the Length of Stay (LoS) for hotel guests based on key features such as country, guest type, room type, and rating. The feature importance analysis revealed that "country" was the most significant predictor, suggesting that geographical factors play a pivotal role in determining guest behaviour [32]–[34]. The evaluation metrics, supported by the Predicted vs. Actual Plot and Error Distribution, indicated that the model performs well within the dataset, with minor deviations in extreme cases. These findings underscore the model's potential as a practical tool for enhancing operational efficiency, enabling hotels to tailor strategies based on data-driven insights [35], [36]. Overall, this research highlights the capability of machine learning to address complex forecasting challenges in the hospitality sector.

The practical benefits of this research for the hospitality industry are substantial, particularly in enhancing operational efficiency and strategic decision-making. By accurately predicting guests' Length of Stay (LoS), hotels can optimize resource allocation, such as room inventory, staffing, and amenities, to align with forecasted demand. This predictive capability enables businesses to implement more targeted marketing strategies, such as personalized promotions or dynamic pricing models, to maximize occupancy rates and revenue [37]. Furthermore, insights from key features influencing guest behavior, such as country or guest type, facilitate tailored services that improve customer satisfaction and loyalty [38]. Integrating such data-driven approaches into daily operations enhances competitiveness and contributes to sustainable growth within the hospitality sector.

4. CONCLUSION

The research delves into predicting hotel guests' Length of Stay (LoS) utilizing the Random Forest Regression model, emphasizing the essentiality of precise forecasting to streamline operational efficiency and inform strategic initiatives in the hospitality sector. The study underscores the pivotal role of data-driven methodologies in decision-making processes, where accurate projections of guest behaviour, including LoS, support effective resource distribution, revenue optimization, and the enhancement of guest satisfaction. A structured methodology defines this research, beginning with data acquisition from guest reviews encompassing variables such as country, guest type, room type, and rating. Preprocessing steps are detailed, involving encoding categorical variables, numerical transformation of target outcomes, and standardization of features to maintain uniformity. The dataset is partitioned into training and testing subsets, allocating 80% for model training and 20% for evaluation. The Random Forest Regression model is selected for its proficiency in addressing non-linear relationships and its resistance to overfitting, with hyperparameters such as estimators=100 and random state=42 ensuring reliability and reproducibility. Analysis reveals that "country" is the most impactful predictor with a contribution score of 0.5, followed by guest type, room type, and rating with contributions of 0.2, 0.15, and 0.15, respectively. Model efficacy is substantiated through evaluation metrics and visual tools like the Predicted vs. Actual Plot and Error Distribution, highlighting minimal deviations and a peak error frequency near zero. The findings affirm the model's capability to identify critical patterns and propose actionable insights for optimization. This study significantly advances the application of predictive analytics in the hospitality industry, offering robust tools to enhance resource planning, strategic marketing, and guest satisfaction, ultimately fostering sustainable business development.

REFERENCES

[1] A. Dursun-Cengizci and M. Caber, "Using machine learning methods to predict future churners: an analysis of repeat hotel customers," *Int. J. Contemp. Hosp. Manag.*, 2024, doi: 10.1108/IJCHM-06-2023-0844.

[2] M. Kumar, C. Kumar, N. Kumar, and S. Kavitha, "Efficient Hotel Rating Prediction from Reviews Using Ensemble Learning Technique," *Wirel. Pers. Commun.*, vol. 137, no. 2, pp. 1161–1187, 2024, doi: 10.1007/s11277-024-11457-w.

[3] S. Ahmed, S. Chowdhury, and R. M. Rahman, "Hotel Booking Cancellation with Visual Analytics," *International IEEE Conference proceedings, IS*, no. 2024. 2024. doi: 10.1109/IS61756.2024.10705220.

[4] K. P. Rajesh, M. Prabu Nallasivam, C. Sakthi Gokul Rajan, P. S. Sherlin Paul, S. Hari Kumar, and V. S. Dharun, "Detection of Fake Hotel Reviews Using ANFIS and Natural Language Processing Techniques," *Proceedings of International Conference on Circuit Power and Computing Technologies, ICCPCT 2024*. pp. 265–269, 2024. doi: 10.1109/ICCPCT61902.2024.10672838.

[5] M. S. Shallan, I. F. Moawad, R. El Naggar, and H. Montasser, "Using Machine Learning Techniques to Maximize Profitability in the Hospitality Industry," *6th International Conference on Computing and Informatics, ICCI 2024*. pp. 182–188, 2024. doi: 10.1109/ICCI61671.2024.10485148.

[6] C. YU, L. J. LIANG, and H. C. CHOI, "Examining Customer Value Cocreation Behavior in Boutique Hotels: Hospitableness, Perceived Value, Satisfaction, and Citizenship Behavior," *Tour. Anal.*, vol. 29, no. 2, pp. 221–237, 2024, doi: 10.3727/108354224X17091476372167.

[7] M. Bordian, M. Fuentes-Blasco, I. Gil-Saura, and B. Moliner-Velázquez, "Technology and Innovation: Analyzing the Heterogeneity of the Hotel Guests' Behavior," *J. Theor. Appl. Electron. Commer. Res.* , vol. 19, no. 2, pp. 1599–1615, 2024, doi: 10.3390/jtaer19020078.

[8] M. Darvishmotevali, H. E. Arici, and M. A. Koseoglu, "Customer satisfaction antecedents in uncertain hospitality conditions: an exploratory data mining approach," *J. Hosp. Tour. Insights*, 2024, doi: 10.1108/JHTI-11-2023-0845.

[9] X. Wang, J. Zheng, and M. Luo, "More than words: the role of personality in shaping the timeliness of online reviews," *J. Hosp. Tour. Technol.*, 2024, doi: 10.1108/JHTT-03-2024-0192.

[10] M. Landa-Zárate, E. Fernández-Echeverría, L. E. García-Santamaría, G. Fernández-Lambert, and E. Martínez-Mendoza, "An Approach to Define Service Strategies: The Case of an Ecotourism Hotel in Mexico," *J. Ind. Eng. Manag.*, vol. 17, no. 1, pp. 182–195, 2024, doi: 10.3926/jiem.6099.

[11] H. Han, S. I. Kim, J. S. Lee, and I. Jung, "Understanding the drivers of consumers' acceptance and use of service robots in the hotel industry," *Int. J. Contemp. Hosp. Manag.*, 2024, doi: 10.1108/IJCHM-02-2024-0163.

[12] A. Bhardwaj, T. Yadav, and R. Chaudhary, "Predicting Hotel Booking Cancellations using Machine Learning Techniques," *2024 15th International Conference on Computing Communication and Networking Technologies, ICCCNT 2024*. 2024. doi: 10.1109/ICCCNT61001.2024.10725148.

[13] K. Sharma, Y. K. Dwivedi, and B. Metri, "Incorporating causality in energy consumption forecasting using deep neural networks," *Ann. Oper. Res.*, vol. 339, no. 1–2, pp. 537–572, 2024, doi: 10.1007/s10479-022-04857-3.

[14] S. Birim, I. Kazancoglu, S. K. Mangla, A. Kahraman, and Y. Kazancoglu, "The derived demand for advertising expenses and implications on sustainability: a comparative study using deep learning and traditional machine learning methods," *Ann. Oper. Res.*, vol. 339, no. 1–2, pp. 131–161, 2024, doi: 10.1007/s10479-021-04429-x.

[15] K. Ito, S. Kanemitsu, R. Kimura, and R. Omori, "Time changes of customer behavior on accommodation reservation: a case study of Japan," *Jpn. J. Ind. Appl. Math.*, vol. 41, no. 2, pp. 881–902, 2024, doi: 10.1007/s13160-023-00623-5.

[16] N. Satish, J. Anmala, K. Rajitha, and M. R. R. Varma, "A stacking ANN ensemble model of ML models for stream water quality prediction of Godavari River Basin, India," *Ecol. Inform.*, vol. 80, 2024, doi: 10.1016/j.ecoinf.2024.102500.

[17] A. Lotfipoor, S. Patidar, and D. P. Jenkins, "Deep neural network with empirical mode decomposition and Bayesian optimisation for residential load forecasting," *Expert Syst. Appl.*, vol. 237, 2024, doi: 10.1016/j.eswa.2023.121355.

[18] S. Chalupa and M. Petricek, "Understanding customer's online booking intentions using hotel big data analysis," *J. Vacat. Mark.*, vol. 30, no. 1, pp. 110–122, 2024, doi: 10.1177/13567667221122107.

[19] S. M. Fazal-e-Hasan, G. Mortimer, H. Ahmadi, M. Adil, and M. Sadiq, "Examining the impact of tourists' hope, knowledge and perceived value on online hotel booking intentions," *Asia Pacific J. Tour. Res.*, vol. 29, no. 6, pp. 719–735, 2024, doi: 10.1080/10941665.2024.2343058.

[20] S. Khan and S. U. Khan, "Tourist motivation to adopt smart hospitality: the impact of smartness and technology readiness," *J. Hosp. Tour. Insights*, 2024, doi: 10.1108/JHTI-04-2024-0335.

[21] J. Castanha, S. K. B. Pillai, and K. G. Sankaranarayanan, "What Influences Consumer Satisfaction and Behaviour Intention in Hotel Industry? A Case Study of Goa, India," *Int. J. Hosp. Tour. Syst.*, vol. 17, no. 2, pp. 61–69, 2024.

[22] A. Pal, K. S. Ahmed, and S. Mangalathu, "Data-driven machine learning approaches for predicting slump of fiber-reinforced concrete containing waste rubber and recycled aggregate," *Constr. Build. Mater.*, vol. 417, 2024, doi: 10.1016/j.conbuildmat.2024.135369.

[23] P. Jain, M. T. Islam, and A. S. Alshammari, "Comparative analysis of machine learning techniques for metamaterial absorber performance in terahertz applications," *Alexandria Eng. J.*, vol. 103, pp. 51–59, 2024, doi: 10.1016/j.aej.2024.05.111.

[24] J. de S. Brogni, L. T. Tricárico, P. F. Limberger, and T. F. Fiuza, "The relationship between visitors' motivations and satisfaction about a Brazilian sacred complex," *Int. J. Tour. Cities*, vol. 10, no. 2, pp. 682–700, Jan. 2024, doi: 10.1108/IJTC-03-2022-0060.

[25] Z. Chen, C. Ye, H. Yang, P. Ye, Y. Xie, and Z. Ding, "Exploring the impact of seasonal forest landscapes on tourist emotions using Machine learning," *Ecol. Indic.*, vol. 163, 2024, doi: 10.1016/j.ecolind.2024.112115.

[26] T. D. Dang and M. T. Nguyen, "Understanding Customer Perception and Brand Equity in the Hospitality Sector: Integrating Sentiment Analysis and Topic Modeling," *Springer Proceedings in Business and Economics*. pp. 413–425, 2024. doi: 10.1007/978-3-031-49105-4_24.

[27] A. S. Abuhammad and M. A. Ahmed, "Automatic Negation Detection for Semantic Analysis in Arabic Hotel Reviews Through Lexical and Structural Features: A Supervised Classification," *J. Inf. Commun. Technol.*, vol. 23, no. 4, pp. 709–744, 2024, doi: 10.32890/jict2024.23.4.5.

[28] Garima *et al.*, "Fake Review Detection and Removal: A Comparative Analysis using ML and DL Models," *15th International Conference on Advances in Computing, Control, and Telecommunication Technologies, ACT 2024*, vol. 1. pp. 200–208, 2024. [Online]. Available: https://www.scopus.com/inward/record.uri?partnerID=HzOxMe3b&scp=85208790888&origin=inward

[29] M. M. Khan and M. Alkhathami, "Anomaly detection in IoT-based healthcare: machine learning for enhanced security," *Sci. Rep.*, vol. 14, no. 1, 2024, doi: 10.1038/s41598-024-56126-x.

[30] S. Bhadra and C. J. Kumar, "Enhancing the efficacy of depression detection system using optimal feature selection from EHR," *Comput. Methods Biomech. Biomed. Engin.*, vol. 27, no. 2, pp. 222–236, 2024, doi: 10.1080/10255842.2023.2181660.

[31] R. A. Rasul, P. Saha, D. Bala, S. M. R. U. Karim, M. I. Abdullah, and B. Saha, "An evaluation of machine learning approaches for early diagnosis of autism spectrum disorder," *Healthc. Anal.*, vol. 5, 2024, doi: 10.1016/j.health.2023.100293.

[32] L. A. Pereira, R. S. Frio, M. A. Pereira, and T. O. Dos Santos, "Does guest perception of sustainability affect consumer advocacy in hospitality?," *Brazilian J. Tour. Res.*, vol. 18, 2024, doi: 10.7784/rbtur.v18.2969.

[33] K. M. Selem, M. H. Shoukat, R. Khalid, and M. Raza, "Guest interaction with hotel booking website information: scale development and validation of antecedents and consequences," *J. Hosp. Mark. Manag.*, vol. 33, no. 5, pp. 626–648, 2024, doi: 10.1080/19368623.2023.2279174.

[34] A. K. Zinn, D. Greene, and S. Dolnicar, "Communicating default changes to hotel room cleaning without reducing guest satisfaction," *J. Clean. Prod.*, vol. 483, 2024, doi: 10.1016/j.jclepro.2024.144266.

[35] F. Fadhlurrachman and N. Sofyan, "Eastparc Hotel Marketing Communication Strategy for Increasing Occupancy During the Pandemic in 2021," *Studies in Systems, Decision and Control*, vol. 489. pp. 501–510, 2024. doi: 10.1007/978-3-031-36895-0_40.

[36] O. Martorell Cunill, L. Otero, P. Durán Santomil, and J. Gil Lafuente, "Analysis of the effect of growth strategies and hotel attributes on performance," *Manag. Decis.*, vol. 62, no. 7, pp. 2233–2264, 2024, doi: 10.1108/MD-06-2023-0974.

[37] D. Contessi, L. Viverit, L. N. Pereira, and C. Y. Heo, "Decoding the future: Proposing an interpretable machine learning model for hotel occupancy forecasting using principal component analysis," *Int. J. Hosp. Manag.*, vol. 121, 2024, doi: 10.1016/j.ijhm.2024.103802.

[38] J. L. Nicolau, Z. Xiang, and D. Wang, "Daily online review sentiment and hotel performance," *Int. J. Contemp. Hosp. Manag.*, vol. 36, no. 3, pp. 790–811, Jan. 2024, doi: 10.1108/IJCHM-05-2022-0594.