NOVA SCHOOL OF BUSINESS & ECONOMICS

REPORT

HOTEL INDUSTRY

MACHINE LEARNING

SPRING 2024

Abstract

This research investigates machine learning techniques for predicting hotel booking trends to minimize cancellations and optimize occupancy management. Analyzing a dataset of over 100,000 hotel booking records from two Portuguese hotels, the study evaluates models like Logistic Regression, Random Forest, XGBoost, and Stacking. Stacking emerges as the best-performing model, offering high accuracy and discrimination. Recommendations include ongoing improvement, model validation, interpretability, and adaptability to market changes. This study underscores the potential of predictive modeling to enhance operational efficiency and guest satisfaction in the hospitality sector.

FRANCISCO GOMES | 39350
GUILHERME RAIMUNDO | 58053
LUÍS OLIVEIRA | 43123
MARTA DINIS | 43027
MARTIM COSTA | 39358

Introduction

The capability to accurately predict hotel booking trends represents a critical advantage in the competitive hospitality industry. This research report focuses on the application of machine learning techniques to forecast dynamics such as booking cancellations, crucial for strategic decision-making in hotel management. The study is particularly centred around two distinct types of accommodations: a Resort Hotel (H1) and a City Hotel (H2), both located in Portugal. These establishments provide a rich context for examining varied booking behaviours and the impact of multiple factors on hotel occupancy and guest satisfaction.

This study intersects the domains of predictive analytics and hospitality management, emphasizing the growing significance of data-driven approaches in the industry. Through a comprehensive analysis of a substantial dataset detailing the operational characteristics of the chosen hotels, this research strives to uncover underlying patterns and insights. The goal is to improve booking management strategies to minimize cancellation rates, thereby enabling the development of more advanced and efficient operational practices in the hospitality sector.

The core challenge addressed by this research involves developing predictive models that can foresee hotel booking cancellations and occupancy rates with a high degree of accuracy. Preliminary analysis of the data has revealed several patterns of interest, such as the significant presence of domestic guests, the influence of seasonal trends on booking volumes, and the relationship between cancellation rates and the likelihood of repeat bookings. These findings highlight the complexity of the hotel bookings landscape and serve as a basis for our methodological approach.

Data

This dataset provides a deep analysis of over 100,000 hotel booking records from two different hotels in Portugal: A City Hotel and a Resort Hotel. It covers a wide range of variables, including guest demographics, booking details, stay characteristics, and financial transactions. With its comprehensive nature, this dataset serves as an excellent resource for investigating the complexities of hotel booking dynamics and building predictive models to understand and predict booking trends.

Our decision to use this dataset for research was motivated by its extensive and real-world data, offering a unique opportunity to explore the operational intricacies of hotel management through data analysis. The availability of detailed and comprehensive data allows for a thorough exploration and examination of patterns and insights relevant to the hospitality industry, using techniques such as exploratory data analysis and machine learning.

Exploratory Data Analysis (EDA)

The initial phase of our analysis involved a comprehensive exploratory data analysis (EDA) to understand the dataset's characteristics, the distribution of key variables, and identify any underlying patterns or anomalies. This critical step set the stage for the subsequent application of machine learning techniques to predict booking trends and cancellations.

Our EDA revealed several insightful trends and anomalies:

- Booking Insights: The dataset reveals distinct booking behaviors between the Resort Hotel (H1) and the City Hotel (H2), indicating the diverse demand dynamics and guest profiles unique to each hotel category. Recognizing these differences is essential for developing predictive models that effectively mirror the operational characteristics of each type of accommodation.
- Cancellation Patterns: Initial analysis brought to light key observations regarding cancellation rates, emphasizing the influence of variables like the advance booking period, guest origin, and seasonal shifts. Such variables play a crucial role in understanding booking fluctuations and developing effective approaches to reduce cancellations.
- Guest Demographics and Preferences: The exploration of guest demographics and stay preferences highlighted the diverse guest segments served by the two hotels. Notably, a considerable proportion of bookings originate from domestic guests, indicating the importance of local market dynamics in overall booking trends.

An in-depth examination of the dataset verified its quality and extensive inclusion of variables that are very important to our study. The dataset covers a broad spectrum of information, ranging from fundamental booking details to more intricate data such as special requests and interactions with guests, facilitating a detailed and multifaceted analysis. To maintain the integrity of the data, steps were taken to rectify missing values, amend inaccuracies, and confirm the dataset’s dependability through comparative analyses and statistical overviews. The exploratory data analysis laid a solid foundation for our research, uncovering crucial insights that informed our methodological approach and model development. The identification of key variables and their interactions will guide our predictive modelling efforts, aiming to enhance booking management and operational efficiency.

Model Selection

In predictive analytics, choosing the appropriate model is essential. It ensures accurate predictions and better decision-making with available resources. Selecting the right model enhances operational efficiency, ultimately driving improvements in guest satisfaction and revenue generation within the competitive hospitality industry.

Given that the goal is to forecast whether a guest will cancel a booking or not, this task falls under the domain of classification problems in machine learning. In classification, the aim is to categorize input data into distinct labels, making it suitable for predicting binary outcomes like booking cancellations. Therefore, we have chosen models specifically designed for binary classification tasks, ensuring they align with the nature of our problem. Evaluating multiple models provides a broader perspective, increasing our assurance in selecting the optimal one.

We selected Logistic Regression as our starting point due to its simplicity and interpretability, serving as a foundational model. However, it's important to note its disadvantages, such as the assumption of linear relationships between features and the target variable, which may limit its capacity to capture complex relationships in the data. For that reason, we have proceeded to explore more advanced models that can overcome these limitations and capture the intricate relationships present in the data more effectively.

Following this, we explored Random Forest and its optimized variant, both of which employ a bagging approach within ensemble learning. The Optimized Random Forest incorporates fine-tuning techniques to further enhance its predictive capabilities through parameter optimization and feature selection. Random Forest is robust against overfitting and is effective for handling high-dimensional data. Yet, it can be computationally expensive and slow to train on large datasets, and it may struggle to perform well when faced with highly imbalanced datasets. In the context of hotel booking prediction, imbalanced datasets can naturally occur where the non-cancelled bookings vastly outnumber the cancelled bookings, reflecting the common scenario where most hotel reservations are not cancelled.

Then, we delved into XGBoost, a boosting approach that combines several weak models iteratively to produce a powerful one with superior predictive performance. XGBoost offers superior predictive performance, handling complex relationships well and providing regularization techniques to prevent overfitting. However, it requires careful parameter tuning and may be computationally expensive and slow.

Lastly, we experimented with Stacking, a meta-learning technique that combines predictions from multiple diverse models to produce a final prediction. Stacking has the advantage of potentially improving predictive accuracy and robustness by leveraging the strengths of multiple models, but it requires careful implementation to avoid overfitting.

Results

The performance of the models was evaluated based on various metrics, including accuracy, ROC-AUC score, and F1 score. Accuracy measures the overall correctness of the model's predictions, representing the proportion of correctly predicted instances (both cancellations and non-cancellations) out of the total instances in the dataset. The ROC-AUC score evaluates the model's

ability to rank cancellations higher than non-cancellations across different threshold settings, indicating its discrimination between cancelled and non-cancelled bookings. A higher ROC-AUC score signifies better discrimination ability. The F1 score is the harmonic mean of precision and recall, and it balances the trade-off between these two metrics. Precision measures the proportion of true positive predictions (cancellations correctly identified) out of all positive predictions, while recall measures the proportion of true positive predictions out of all actual cancellations. In our context, a higher F1 score indicates a more accurate prediction of cancellations by achieving a better balance between precision and recall. The results can be seen in the following table:

Model	Accuracy	ROC-AUC	F1-Score
Logistic Regression	0.785236	0.734096	0.649248
Random Forest	0.867126	0.844815	0.809281
Optimized Random Forest	0.867294	0.845163	0.809690
XGBoost	0.846409	0.818141	0.774207
Stacking	0.869081	0.849199	0.814319

Before delving into the analysis of each model's results, it is important to note that models that excel in one metric tend to perform well across all metrics. This highlights the interconnected nature of evaluation in machine learning, implying that improvements in one aspect often lead to enhancements in others.

Regarding Logistic Regression, it consistently shows the lowest scores across all metrics, suggesting its limitations in accurately predicting hotel booking cancellations compared to other models. As mentioned before, although this is the simplest model selected and therefore easier to interpret, its inability to capture complex relationships may have led to its lower performance.

While XGBoost exhibits better performance than Logistic Regression, it falls behind other ensemble techniques in terms of performance scores. Despite this difference being relatively small, it still indicates that further exploration of different ensemble techniques may be beneficial for improving predictive accuracy.

Furthermore, the comparable performance between the optimized variant of Random Forest and the standard Random Forest suggests that the fine-tuning techniques applied may not have significantly enhanced predictive accuracy in this context. Despite efforts to optimize parameters and select relevant features, the improvement in performance appears to be marginal. This observation could indicate that the default settings of Random Forest already capture the underlying patterns in the data effectively, leaving limited room for further enhancement through fine-tuning.

Recommendations

Among the evaluated models, Stacking emerges as the standout performer, consistently exhibiting the highest scores across all key metrics. With an accuracy score of 0.869081, it correctly predicts hotel booking cancellations approximately 86.91% of the time. The ROC-AUC score of 0.849199 indicates a robust ability to distinguish between cancelled and non-cancelled bookings, suggesting reliable discrimination. Additionally, an F1-Score of 0.814319, demonstrates a balance between precision and recall in identifying cancelled bookings. Given this, we recommend using this model, but hotel managers should remain mindful of several key considerations when implementing it.

Firstly, while the model has shown promising performance, there is room for improvement. Therefore, hotels should prioritize continuous improvement efforts, such as refining model parameters and incorporating additional relevant features, to enhance its predictive accuracy over time. Additionally, it is crucial to regularly validate the model to ensure its performance remains consistent and reliable. This involves monitoring the model's predictions against actual booking cancellations and adjusting strategies accordingly. Furthermore, hotels should be mindful of the trade-off between model complexity and interpretability. While the Stacking model may offer superior predictive power, its interpretability may be compromised. Thus, hotels should carefully consider this trade-off and ensure they can understand and justify the model's predictions to stakeholders. Lastly, given the dynamic nature of the hospitality industry, hotels must remain adaptable to changing market conditions and customer preferences. This adaptability ensures that strategies informed by the model can effectively respond to evolving trends.

Conclusion

With a well-developed and tuned model, hotels can take great advantage of it. For instance, by accurately anticipating booking cancellations, hotels can optimize their room inventory management, ensuring optimal occupancy levels while minimizing the risk of overbooking. Moreover, the model enables hotels to implement personalized customer engagement strategies. By identifying guests at risk of cancelling their bookings, hotels can proactively reach out with tailored incentives or alternative arrangements to retain their business. This personalized approach not only improves guest satisfaction and loyalty but also strengthens the hotel's reputation for exceptional customer service. In addition, the model's predictive insights can help design pricing strategies. Hotels have the flexibility to adapt their room rates in response to anticipated cancellation probabilities, which enables them to boost revenue while also ensuring they stay competitive within the market. Overall, by leveraging the predictive power of the model, hotels can optimize their operations, maximize revenue potential, and maintain high service standards.