



**DBA4714 Deep Learning & Generative AI in Business
Final Project Report
Team SAGA**

| Name | Matriculation Number |
|----------------------|----------------------|
| Heng Chin Hwee, Dave | A0217633H |
| Keynes Yeo Jun Yan | A0217770A |
| Mevin Singh | A0217526E |

Table of Contents

| | |
|---|-----------|
| Table of Contents..... | 2 |
| 1. Introduction..... | 3 |
| 2. Data Cleaning and Preprocessing..... | 3 |
| 2.1 Removal of Unnecessary Features..... | 3 |
| 2.2 Null Values..... | 3 |
| 2.3 Scaling & Splitting..... | 3 |
| 3. Customer Segmentation..... | 4 |
| 4. Preliminary Analysis (Exploratory Data Analysis)..... | 4 |
| 4.1 Correlation Analysis..... | 4 |
| 4.2 Association Analysis..... | 5 |
| 4.3 Analysis of Consumer Booking Patterns Over the Years..... | 5 |
| 4.4 Analysis of Month-on-Month Cancellation Rates Changes..... | 6 |
| 4.5 Analysis of Cancellation Rates Across HG88's Hotel Types..... | 6 |
| 4.6 Identifying Factors Driving Cancellation Rates..... | 6 |
| 4.6.1 Booking Lead Time Hypothesis..... | 6 |
| 4.6.2 Deposit Type Hypothesis..... | 7 |
| 4.6.3 Seasonal Trends Hypothesis..... | 8 |
| 5. Models..... | 9 |
| 5.1 Neural Network..... | 9 |
| 5.2 Random Forest..... | 9 |
| 5.3 XGBoost..... | 9 |
| 6. Model Evaluation and Comparison..... | 9 |
| 7. Recommendations..... | 10 |
| 7.1 Dynamic Pricing Strategy..... | 10 |
| 7.2 Tiered Incentives based on Consumer Profile..... | 11 |
| 7.3 Strategic Overbooking..... | 11 |
| 8. Conclusion..... | 11 |
| 9. References..... | 12 |
| 10. Appendix..... | 13 |
| Appendix A - Table containing columns in a dataset..... | 13 |
| Appendix B - Centroids of K-prototype Clustering..... | 15 |
| Appendix C - Clustering: Summary of Responses by LLM..... | 16 |
| Appendix D - ChatGPT and Gemini's response to Clustered Features and Segment Name..... | 17 |
| Appendix E - ChatGPT & Gemini's Response to High Correlation: Weekday vs. Weekend Stays..... | 21 |
| Appendix F - ChatGPT code to calculate Cramér's V values..... | 22 |
| Appendix G - Drivers for High Cancellation Rates in City Hotels..... | 23 |
| Appendix H - Contradicting Conclusions from ChatGPT for Similar Issues..... | 24 |
| Appendix I - Outlier Situations Where Non-refundable Deposits Generate Higher Cancellation Rates..... | 25 |
| Appendix J - Estimate of Type I and Type II errors..... | 26 |
| Appendix K - XGB Hyperparameter Tuning..... | 27 |
| Appendix L - GridSearchCV Scoring..... | 28 |
| Appendix M - Alternative Hyperparameter Tuning Method..... | 28 |
| Appendix N - Metric vs Threshold Curve + Feature importance for all 3 models..... | 29 |
| Appendix O - Possible Reason for Increase in False Negatives for XGB After Threshold Tuning..... | 30 |
| Appendix P - Code for Neural Network Model..... | 31 |
| Appendix Q - Consumer Preferences Based On The Identified Consumer Segments..... | 33 |

1. Introduction

In the dynamic world of the hospitality industry, managing cancellation rates stand out as a significant challenge for hoteliers worldwide. These cancellations disrupt revenue, affect operational efficiency, and impact overall guest satisfaction. Likewise, prominent hotelier Hotel Group 88 (HG88) has faced a considerable cancellation rate issue, peaking at around 37% across their two properties. To address this issue, they have partnered with SAGA Consulting Group (SAGA) to enhance their operational efficiency and reduce cancellation rates.

SAGA will leverage data analytics to delve into the complexities of cancellations, aiming to extract consumer insights and identify high-risk customers prone to cancelling their bookings. The approach will involve several steps:

- Step 1: Segment consumers to understand cancellation rates within each group.
- Step 2: Develop prediction models to forecast booking cancellations.
- Step 3: Select the most effective model for the task.
- Step 4: Identify the primary drivers behind cancellation rates.
- Step 5: Formulate actionable recommendations to address these factors.

This comprehensive process will empower HG88 to understand the nuances of cancellation management. Armed with this knowledge, HG88 can target high-risk consumers with personalised marketing campaigns to reduce cancellation rates. Additionally, insights into the drivers of cancellations will enable HG88 to refine its business practices and mitigate high cancellation rates effectively.

2. Data Cleaning and Preprocessing

We conducted data cleaning and preprocessing to prepare the dataset for analysis to ensure its quality and usability. Considering the large size of our dataset, we decided to drop some columns that were ambiguous or contained null values, reducing the noise in our dataset.

2.1 Removal of Unnecessary Features

“Agent”, “company”, “reservation_status_date” were removed from the dataset as the data source did not provide detailed explanations for these columns. Based on our domain knowledge from secondary research, we concluded that these variables are unlikely to affect model predictions.

The “reservation_status” column is also dropped as it perfectly correlates with our target variable “is_cancelled”. The reservation status of customers will be updated accordingly in the event of a cancellation and our target variable can be derived from the “reservation_status” column.

2.2 Null Values

The team made the decision to drop all null values since they made up an insignificant amount of our dataset (0.4%). Only the “children” and “country” columns are affected, with each containing 4 and 488 rows of Null values respectively. The final dataset has 28 columns (17 numerical and 7 categorical variables) with 118,898 rows (see Appendix A).

2.3 Scaling & Splitting

The categorical variables identified were converted to binary variables using one-hot encoding. The dataset was then scaled to improve model performance and to normalise the data. The ‘MinMaxScaler’ was used instead of the usual ‘StandardScaler’ as our data contains non-negative variables (eg. lead time, daily rate). The training and test data was then derived from the main dataset using the 80/20 split, which was used to train and evaluate the models.

3. Customer Segmentation

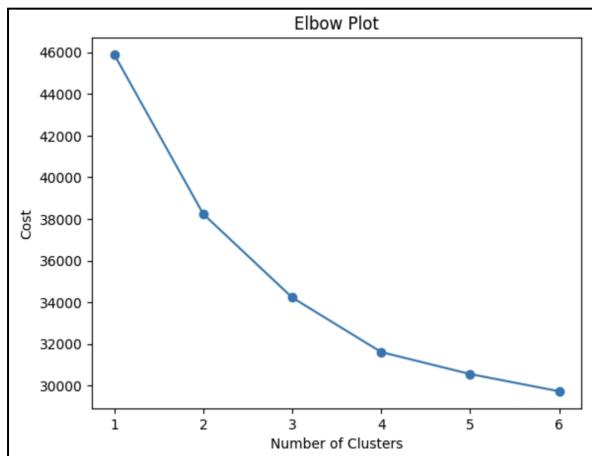


Figure 1: Elbow Plot of K-means Clustering

In order to better understand the hotel industry, we decided to identify customer segments using K-prototype clustering, which is an extension of k-means clustering that is able to process both numerical and categorical variables. Our goal from clustering is to further breakdown the data and find potential insights within clusters and to engineer new features which may help with model prediction. Looking at the elbow plot, we concluded that there are 4 customer segments within HG88. Once the clusters were identified, we engaged the help of LLMs to name the clusters and identify unique properties of each cluster.

| LLM | Cluster 1 (26.9%) | Cluster 2 (24.1%) | Cluster 3 (26.5%) | Cluster 4 (22.5%) |
|------------|---|---|---|--|
| Perplexity | Long-Stay Guests | Short-Stay Leisure Travellers | Families | Business Travellers |
| | High average lead time (124.73 days) | Moderate lead time (93.49 days) | High lead time (114.51 days) | Lowest average lead time (79.41 days) |
| | Highest average number of stays in weekend nights (0.97) and week nights (2.62) | Lowest average number of previous cancellations (0.04) and booking changes (0.22) | Highest Average number of children (0.12) | Highest average number of previous bookings not cancelled (0.20) |
| | High ADR (106.23) | Moderate ADR (97.03) | High ADR (107.47) | Moderate ADR (95.81) |

Table 1: LLM Generated Cluster Names

Table 1 shows the cluster names and properties proposed by the chosen LLM: Perplexity (refer to Appendix C for the comparison of different LLM responses)

We decided to use the responses generated by Perplexity and have identified 4 main customer segments of HG88: (1) Long-Stay Guests, (2) Leisure Travellers, (3) Families and (4) Business Travellers. Perplexity provided the best justifications for segmenting the customers into the 4 groups which made business sense and is consistent with our interpretation of the data. With this information, we aim to enable HG88 to understand their client base better and adopt specialised marketing tools for each group.

4. Preliminary Analysis (Exploratory Data Analysis)

4.1 Correlation Analysis

A correlation analysis was conducted among the numerical variables in the data to uncover any relationships among variables. Figure 2 shows a correlation heatmap of the numerical variables in the dataset.

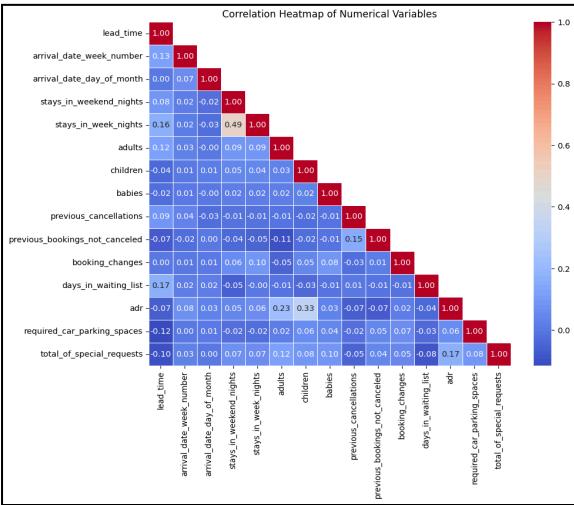


Figure 2: Correlation Heatmap

4.2 Association Analysis

We employed the Chi-Squared Test of Independence to conduct an association analysis between categorical variables. Recognizing our limited expertise in this area, we turned to ChatGPT for assistance in generating the necessary code. Additionally, we evaluated the association among categorical variables using Cramér's V, a metric ranging from 0 (no association) to 1 (perfect association), akin to the Pearson correlation coefficient. A threshold of 0.3 was set by the LLM and subsequently validated as appropriate. Our analysis, detailed in Appendix F, identified the highest association (Cramér's V = 0.778) between 'reserved_room_type' and 'assigned_room_type'. This finding is intuitive, as customers typically receive the rooms they reserved during booking, barring unforeseen circumstances.

4.3 Analysis of Consumer Booking Patterns Over the Years

To gain an understanding of the consumer booking pattern over the years, we proceeded to visualise the number of bookings over time using a line graph (refer to Figure 3).

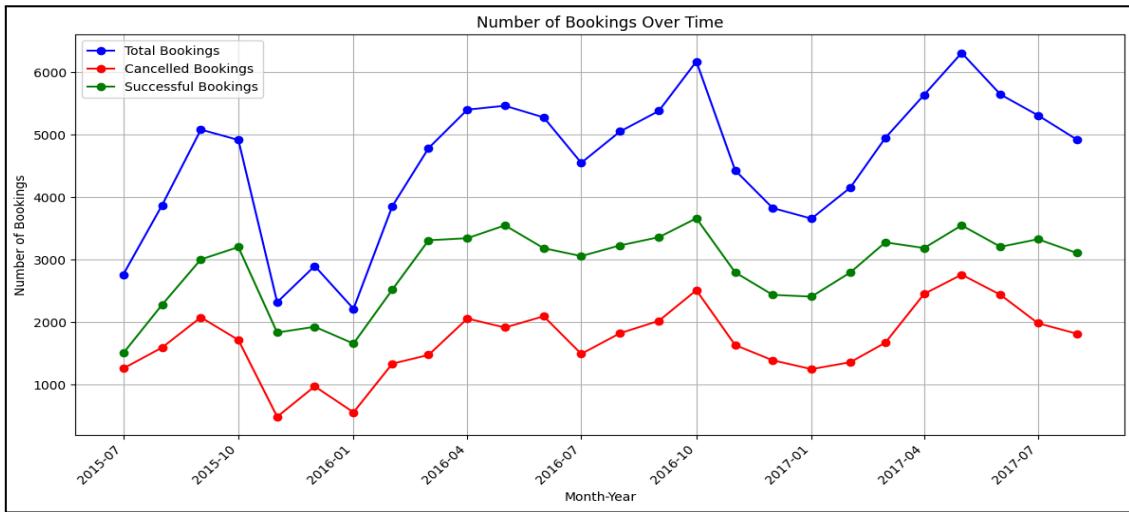


Figure 3: Number of Bookings Over Time

HG88 saw an increase in customer engagement and a steady rise in total bookings throughout the year. However, this surge has also brought about a corresponding increase in cancellations (Chua, 2020). Although the growth rate of total bookings exceeds that of cancellations, leading to a higher number of successful bookings, addressing this ongoing concern is imperative despite a decline in cancellation rates over time.

4.4 Analysis of Month-on-Month Cancellation Rates Changes

Examining the cancellation changes month-on-month is vital for understanding booking cancellation trends. We aim to gain insights to inform strategic decisions and understand customer behaviour dynamics.

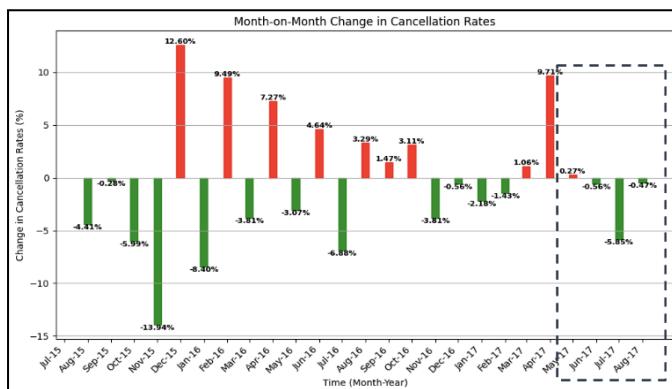


Figure 4: Month-on-Month Cancellation Rates Changes

Our analysis shows a consistent decline in Month-on-Month (MoM) cancellation rates over the years. Recent months have seen more stable fluctuations, indicating positive progress. This reduced volatility enhances our planning and resource allocation accuracy. Effective management of cancellation rates improves demand forecasting and model reliability. However, this should not breed complacency, and more proactive efforts are still needed to further reduce cancellations.

4.5 Analysis of Cancellation Rates Across HG88's Hotel Types

Since the data is collected from 2 different hotels, City Hotel and Resort Hotel, the cancellation rates might not be consistent between the 2. Hence, we investigated the cancellation rates by hotel types, so as to possibly identify underlying reasons driving cancellation rates.

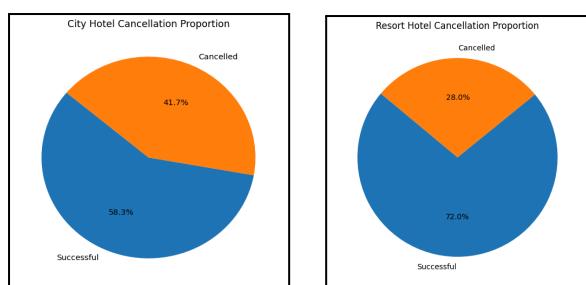


Figure 5 & 6: Cancellation rates in City Hotel and Resort Hotel respectively

The charts show that City Hotel contributed significantly to the cancellation rate and we leveraged on LLMs to uncover the potential reasons for this trend. Our findings revealed two main reasons: (1) leisure travellers tend to book Resort Hotels with firm plans (2) City Hotels attract last-minute planners, leading to more cancellations due to shorter booking lead times. (see Appendix G for LLM's full response).

However, during our investigative process, the LLM provided conflicting conclusions based on query nuances. For instance, when prompting the LLM about the drivers of cancellations in City Hotel specifically, low booking lead time was highlighted as a major driver. On the other hand, general inquiries about cancellation rates cited high booking lead time instead. This discrepancy, despite similar queries, generated conflicting results (see Appendix H). To reconcile this, further secondary research was performed to identify the true relationship between lead time and cancellations. We relied on sources with significant domain knowledge like HotelMinder (Verot, 2024) and HotelManagement.net (Hertzfeld, 2019), which are hospitality CRM solution providers, and they both support the theory that high lead time is a key driver of cancellation rates.

4.6 Identifying Factors Driving Cancellation Rates

Consumer behaviours are complex, influenced by various factors and to varying degrees. To streamline the analysis for SAGA, we employed LLMs, supplemented by secondary research and domain expertise, to pinpoint potential drivers of cancellation rates. Three key factors emerged as primary hypotheses, providing a foundation for SAGA's strategy formulation process.

4.6.1 Booking Lead Time Hypothesis

The first hypothesis seeks to prove if the time gap between booking and arrival significantly impacts cancellation rates. Advanced bookings tend to have higher cancellation rates due to increased chances of plan changes or unforeseen circumstances. Conversely, last-minute bookings may see lower cancellation

rates because of limited alternative options. Thus, retaining customers with longer booking lead times could be a viable strategy to reduce cancellations.

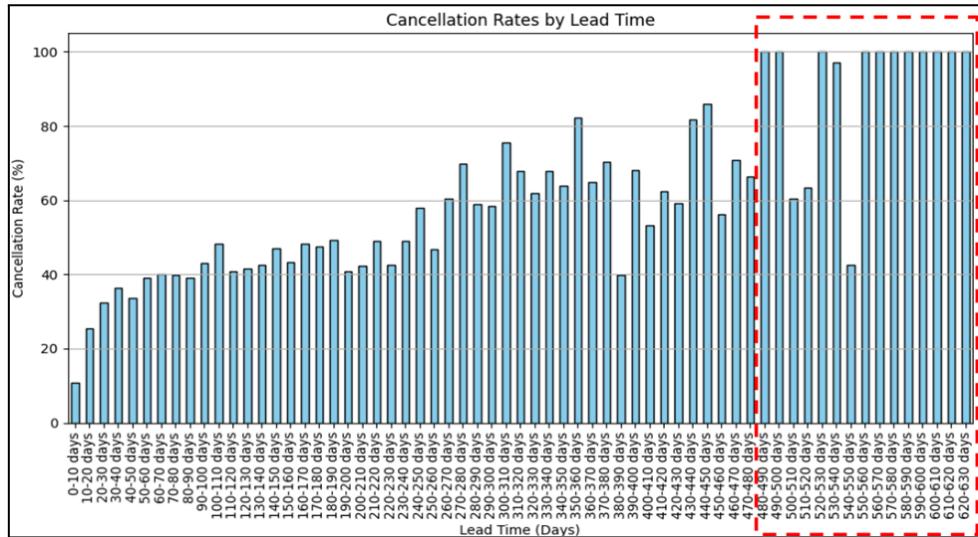


Figure 7: Cancellation Rates by Lead Time

Figure 7 appears to validate the hypothesis, clearly depicting that cancellation rates rise as booking lead time increases. Moreover, most bookings that were made over 480 days in advance were cancelled. Based only on this chart, it appears that these bookings are vulnerable to cancellations and HG88 should prioritise minimising cancellation rates among reservations with high lead times since there is more room for improvement. However, while it is important to consider cancellation rates, the business impact should also be looked at.

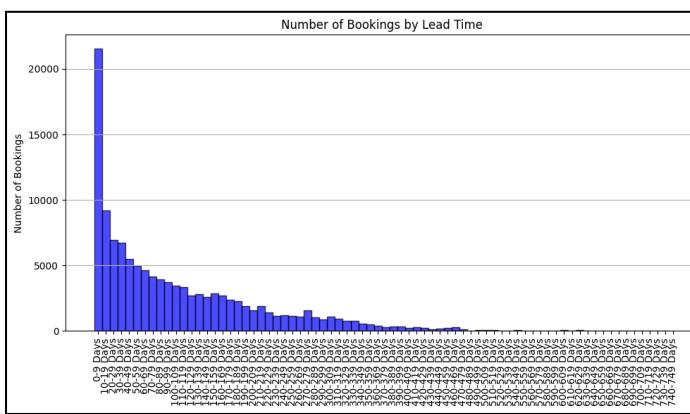


Figure 8: Number of Bookings by Booking Lead Time

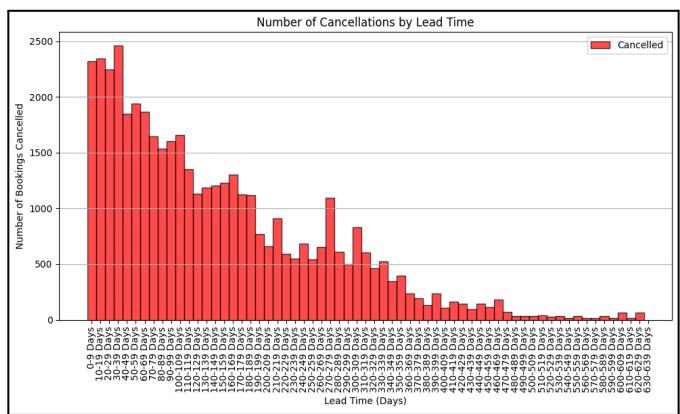


Figure 9: Number of Cancellations by Booking Lead Time

As seen from Figures 8 and 9, the number of bookings and cancellations are clustered around shorter booking lead times. There were only an insignificant number of bookings that were made more than 480 days in advance, and of which an even lower number of them were cancelled. Hence, focusing on reservations with long booking lead times is misguided, and it will be more beneficial for the hotel group to reduce cancellation rates for bookings with shorter lead times since it will cover more reservations.

4.6.2 Deposit Type Hypothesis

The hypothesis expresses that the type of deposit made by the customer will significantly impact the cancellation of bookings. More specifically, the “refundability” of booking deposits can significantly impact

cancellations. Thus, with the arrangements made with a "no deposit" or "refundable" requirement, customers are freer to cancel their bookings since they are less financially obligated to stay with an existing reservation. Conversely, travellers who have already made a non-refundable deposit are financially incentivized to not cancel their reservations.

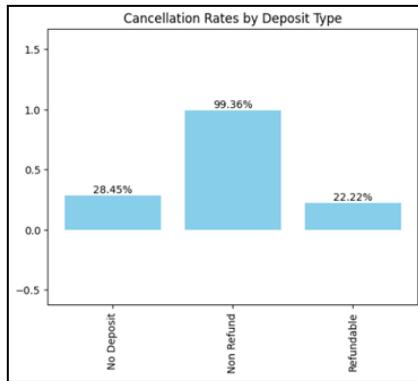


Figure 10: Cancellation Rates by Deposit Type

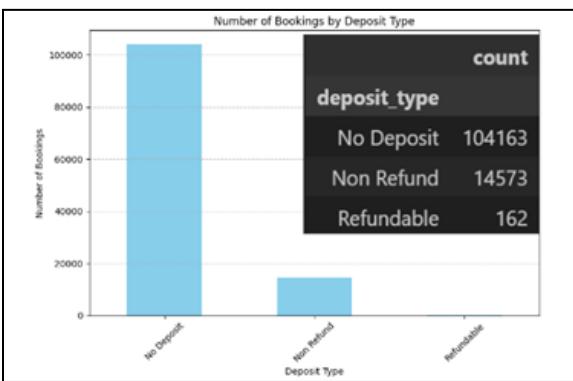


Figure 11: Number of Bookings by Deposit Type

Surprisingly, our data indicate the opposite relationship between deposit type and cancellation rates. Not only do non-refundable bookings have the highest cancellation rates, but it is abnormally high, sitting at 99.36% (Figure 10).

To understand this anomaly, LLM was used to identify possible reasons for this discrepancy. As seen from Appendix I, the reasons listed were all very situational and generally referred to outlier scenarios that are harder to generalise to a particular trend or group behaviour.

Upon closer inspection of the data set, it is observed from Figure 11 that only about 0.1% of total bookings were refundable, which supports the theory that the phenomenon is an outlier situation. Hence, more data is required before the hypothesis can be accepted or rejected conclusively.

4.6.3 Seasonal Trends Hypothesis

The final hypothesis is that different seasonal travelling periods can affect cancellation rates differently. For instance, peak tourist seasons draw in more holiday-goers. However, it also invites fierce competition from rival chains who offer huge discounts to entice customers, drawing them away from their existing reservations, and raising cancellation rates. Another possible seasonal reason is because of changes in seasons and weather conditions. Certain travel periods are more prone to adverse weather conditions which dissuades customers from committing to their existing plans. In extreme circumstances, extended periods of typhoons or blizzards can force customers to cancel their bookings due to safety concerns.

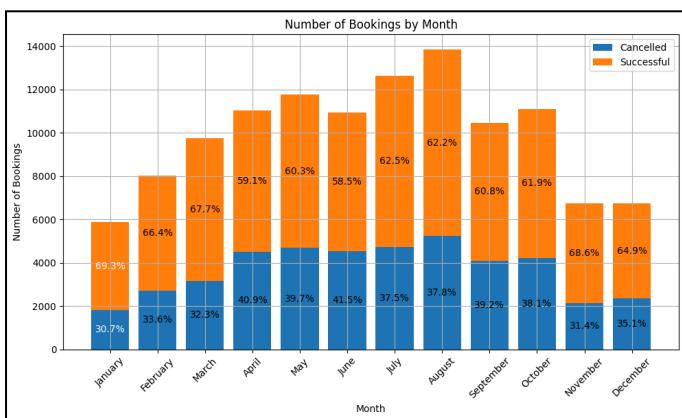


Figure 12: Number of Bookings by Month

Figure 12 demonstrates the fact that cancellation rates can be cyclical. As seen, cancellation rates peaked between April – June, a popular travelling period indicated by the higher number of bookings. Hence, countermeasures, such as dynamic pricing or issuing promo codes, can be rolled out during these periods to encourage consumers to stay with their bookings.

5. Models

Considering the inherent nature of hotel cancellations, our dataset exhibits an imbalance, with 37% of the binary target variable, 'is_canceled', being 1 (cancelled). Consequently, we employed ensemble learning techniques such as Random Forest and XGBoost (Wu, 2022), along with a Neural Network (Huang et al., 2022) as these techniques are particularly effective for handling imbalanced data distributions.

5.1 Neural Network

As shown in Appendix P, we leveraged on the prowess of LLMs to obtain the code for the Neural Network model. The final model that was created generated very positive results, achieving the highest AUC score of all models at 0.845, while the meals under the "BB" category (Breakfast & Bed) were found to be the most significant driver of cancellation.

While the results were desirable, the code that the LLM first provided required several rounds of fine-tuning. For instance, the code suggested utilising pre-built functions like LabelEncoder to encode features (Appendix P) when it should only be used for target variables. In addition, when fixing this issue, the LLM suggested static code that is inflexible and harder to fit into the rest of our code, which reduces efficiency and diminishes the value that is at the core of LLMs.

5.2 Random Forest

In order to improve model performance, we tuned the hyperparameters of the RF model using the GridSearchCV. These are the hyperparameters of the model after tuning: *criterion* = 'entropy', *max_depth* = 15, *n_estimators* = 300. The model has a moderately high AUC of 0.841 and identified the account of special requests as the most important variable in determining cancellation rates.

However, when evaluating the LLM's responses, we realise ChatGPT's default scoring metric for the "GridSearchCV" function in the SKLearn package is 'accuracy' (Appendix L). The same scoring metric of using 'accuracy' is generated by ChatGPT multiple times across different models. We felt that a more robust metric is needed to capture the model's performance relative to misclassifications (eg. False Positive/Negative cases), which can be extremely costly to HG88. More of this cost to HG88 will be elaborated in [Section 6](#). Hence, 'roc_auc' will be a more suitable metric instead of 'accuracy'.

5.3 XGBoost

The initial hyperparameter tuning code provided by ChatGPT included a comprehensive set of parameters and correctly utilised the scoring metric specified in the prompt. However, the use of GridSearchCV resulted in lengthy execution times, prompting us to seek an alternative method like RandomisedSearchCV. After tuning the model, we obtained the following hyperparameters from RandomSearchCV: *learning_rate* = 0.082, *max_depth* = 8, *n_estimators* = 483, with an AUC of 0.843. From Appendix N, we can see the classifier's Metric vs Threshold curve and the top 10 feature importances.

6. Model Evaluation and Comparison

When predicting hotel cancellations, considering Type I and Type II errors is crucial. Type I errors refers to inaccurately predicting cancellations, while Type II errors mistakenly anticipate non-cancellations. We assigned \$50 to false positives and \$100 to false negatives, given the greater risk of Type II errors (Appendix J), since the loss of revenue from a cancelled booking is far greater than the cost of following up on false positive cancellations. Thus, we aim to optimise Recall and AUC to identify cancellations accurately.

The optimal threshold is determined by the point of intersection between the AUC and Recall graph. Tuning the threshold is important as it affects model predictions, business cost, and F1-score. After threshold

adjustments, we successfully reduced false positives across three models, except XGBoost. A possible reason is because of complex hyperparameter tuning by RandomSearchCV, leading to vulnerability to noise or outliers and potential data leakage (refer to Appendix O).

The table below summarises the false negatives pre and post-tuning, as well as the optimal probability threshold in determining classification that maximises the AUC and Recall for the models.

| Model | False Negatives Before Tuning | False Negatives After Tuning | Optimal Threshold |
|----------------|-------------------------------|------------------------------|-------------------|
| Neural Network | 2171 | 1377 | 0.361 |
| XGBoost | 1148 | 1349 | 0.551 |
| Random Forest | 2970 | 1354 | 0.369 |

The table below summarises all metrics used to evaluate the models:

| Model | AUC | Recall | Type-2 Error | F1 | Business Cost |
|----------------|-------|--------|--------------|-------|---------------|
| Neural Network | 0.845 | 0.846 | 0.154 | 0.804 | \$250,150 |
| XGBoost | 0.843 | 0.844 | 0.156 | 0.796 | \$254,800 |
| Random Forest | 0.841 | 0.844 | 0.156 | 0.794 | \$258,000 |

Models are evaluated based on the matrices above, with the final model selected for the highest AUC and Recall and lowest Business Cost. Thus, our chosen model would be the Neural Network.

7. Recommendations

With the help of LLMs, we gathered the findings from our preliminary analysis, customer segmentation and model results to propose the following recommendations to HG88. Our objective is to help HG88 achieve operational efficiency and maximise profits through our recommendations.

7.1 Dynamic Pricing Strategy

Dynamic pricing allows businesses to stay above the competition and is now standard practice in the hospitality industry (Paddle, 2023) by accounting for factors such as seasonality, competitor pricing and booking patterns. This strategy optimises revenue without sacrificing occupancy rates. Considering the importance of demand forecasting for effective resource management, dynamic pricing can encourage earlier bookings while reducing cancellation rates by introducing price differentiation based on lead time. Additionally, given the seasonal fluctuations of HG88 identified in [Section 4.6.3](#). HG88 can maximise revenue by setting higher prices during peak demand periods of April to June and offering attractive deals during low-demand periods to stimulate bookings.

HG88 can also implement tiered cancellation penalty structures to encourage customers to commit to bookings. Last-minute bookings are extremely costly to hotels as not only does that constitute a loss of revenue, it is much more challenging to secure a new booking in such a short time to make up for the cancellation. Hence we propose a penalty structure whereby the cancellation fee is proportionate to how late the cancellation is. Used in conjunction with dynamic pricing, HG88 can expect a more competitive pricing strategy while reducing cancellation rates.

7.2 Tiered Incentives based on Consumer Profile

One of the ways to increase hotel occupancy is to tie in packaged deals and services, increasing the value for customers, and deterring them from cancelling bookings (Cvent, 2023). Once our model identifies a potential booking cancellation, HG88 can consider incentivising guests to encourage them to stay. However, the hotel group handles a myriad of consumer profiles with different tastes and preferences, hence it is unwise to offer all consumers the same type of incentives to entice them. Additionally, consumers value personalisation and research has found that 71% of consumers expect personalization from the businesses they choose (McKinsey, 2021). Hence, customising the incentives based on the consumer profile identified in [Section 3](#) can reduce cancellation rates more effectively.

To identify the preferences of each consumer profile, we leveraged on LLMs and some of the proposed incentives to prevent cancellations are as follows (Appendix Q):

| Long-Stay Guests | Short-Stay Leisure Travellers | Families | Business Travellers |
|--|---|---|---|
| Laundry Services, Complimentary Parking | Meals, Room Upgrades, Late Check-outs | Kids Meal, Extra Beds, Child-Friendly Amenities | Ironing Services, Transportation, Express Check-ins |

7.3 Strategic Overbooking

Based on the dataset, HG88 has a cancellation rate of about 37% which is in line with the global average of 40% (HT, 2019). Given the competitive landscape of the hospitality industry, many hotels are adopting flexible cancellation policies to promote more bookings which results in a high discrepancy between bookings and actual stays. We drew parallels from the aviation industry and sought to maximise room occupancy by allowing overbooking. The typical “no-show” rate for airlines is about 15% and overbooking can generate an additional 10% of passenger revenue for airlines. With this strategy, American Airlines saw a \$500 million increase in annual revenue (Suzuki, 2006). This shows the success of overbooking in the aviation industry which can be replicated by hotels to maximise room usage. Given the even higher “no-show” rates of HG88, we can expect to generate more than 10% of additional sales by adopting the overbooking strategy.

While overbooking can enhance overall room utilisation, it does entail inherent risks. In cases of overbooked scenarios, a robust compensation strategy similar to that of airlines is needed to ensure customer satisfaction (USA Today, 2023). HG88 can look to offer substantial compensations or alternative accommodations. An important success factor is to balance the predictive capabilities of our model and maintain customer satisfaction. We are confident the reward of adopting the overbooking strategy exceeds the risks and will benefit HG88 in the long run.

8. Conclusion

SAGA's analysis has provided actionable insights to mitigate cancellations and increase revenue. Through EDA, clustering, and predictive modelling, we identified key factors influencing cancellations and developed accurate forecasting models. Our recommendations aim to reduce cancellation rates, increase revenue, and improve customer satisfaction. With our strategies, we hope to mitigate current gaps and differentiate HG88 from its competitors through our data-driven approach

9. References

- Benjamin Verot. (2024, January 21). Everything you need to know about hotel cancellations. Hotel Services, Software & Systems Comparison | HotelMinder. <https://www.hotelminder.com/everything-you-need-to-know-about-hotel-cancellations>
- Chua, J. (2020, October 28). Hotel Cancellation Rates. LinkedIn. <https://www.linkedin.com/pulse/hotel-cancellation-rates-joeys-chua/>
- Cvent. (2023, November 1). Increase hotel occupancy: 9 proven strategies | Cvent blog. Cvent | Event Platform for In-person, Virtual, and Hybrid Events & Webinars. <https://www.cvent.com/en/blog/hospitality/increase-hotel-occupancy>
- Esther Hertzfeld. (2019, April 23). Study: Cancellation rate at 40% as OTAs push free change policy. Hotel Management. <https://www.hotelmanagement.net/tech/study-cancellation-rate-at-40-as-otas-push-free-change-policy>
- Hospitality Technology. (2019, April 24). Global cancellation rate of hotel reservations reaches 40% on average. <https://hospitalitytech.com/global-cancellation-rate-hotel-reservations-reaches-40-average>
- Mckinsey. (2021, November 12). *The value of getting personalization right—or wrong—is multiplying.* McKinsey & Company. <https://www.mckinsey.com/capabilities/growth-marketing-and-sales/our-insights/the-value-of-getting-personalization-right-or-wrong-is-multiplying>
- Paddle. (2022). Dynamic pricing strategy: Definition, types, benefits & examples. Paddle – The complete payments, tax, and subscriptions solution for SaaS. <https://www.paddle.com/resources/dynamic-pricing-model>
- Ye Wu, & Rick Radewagen. (2022, August 24). 7 Techniques to Handle Imbalanced Data. KD Nuggets. <https://www.kdnuggets.com/2017/06/7-techniques-handle-imbalanced-data.html>
- Yoshinori Suzuki. (n.d.). The net benefit of airline overbooking. ScienceDirect. <https://www.sciencedirect.com/science/article/abs/pii/S1366554504000845>
- Zach Wichter. (2023, May 10). Your browser is not supported. Your browser is not supported | usatoday.com. <https://www.usatoday.com/story/travel/columnist/2023/05/10/airline-overbooking-bumped-flight-cruising-altitude/70199667007/>
- Zhan ao Huang, Yongsheng Sang, Yanan Sun, & Jiancheng Lv. (2022, September 8). A neural network learning algorithm for highly imbalanced data classification. <https://www.sciencedirect.com/science/article/abs/pii/S0020025522009847>

10. Appendix

Appendix A - Table containing columns in a dataset

| Column | Description | Data Type |
|--------------------------------|--|-------------|
| hotel | Indicates whether the booking is for a city hotel or resort hotel | Categorical |
| is_canceled | Indicates if the booking is canceled (1) or not (0) Target Variable | Binary |
| lead_time | Number of days between booking date and the arrival date | Integer |
| arrival_date_year | Year of arrival date | Integer |
| arrival_date_month | Month of arrival date | Categorical |
| arrival_date_week_number | Week number of arrival date | Integer |
| arrival_date_day_of_month | Day of the month of the arrival date | Integer |
| stays_in_weekend_nights | Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay | Integer |
| stays_in_week_ends | Number of week nights (Monday to Friday) the guest stayed or booked to stay | Integer |
| adults | Number of adults | Integer |
| children | Number of children | Integer |
| babies | Number of babies | Integer |
| meal | Type of meal booked | Categorical |
| country | Country of origin | Categorical |
| market_segment | Market segment designation | Categorical |
| distribution_channel | Booking distribution channel | Categorical |
| is_repeated_guest | Indicates if the booking customer is a repeated guest (1) or not (0) | Categorical |
| previous_cancellations | Number of previous bookings that were canceled by the customer | Integer |
| previous_bookings_not_canceled | Number of previous non-canceled bookings by the customer | Integer |
| reserved_room_type | Code of the reserved room type | Categorical |
| assigned_room_type | Code of the assigned room type | Categorical |
| booking_changes | Number of changes/amendments made to the | Integer |

| | | |
|-----------------------------|--|-------------|
| | booking | |
| deposit_type | Indication of the type of deposit | Categorical |
| days_in_waiting | Number of days the booking was in the waiting list | Integer |
| customer_type | Type of customer | Categorical |
| adr | Average Daily Rate for the booking | Integer |
| required_car_parking_spaces | Number of car parking spaces required by the customer | Integer |
| total_special_requests | Number of special requests made by the customer (e.g., high floor, view from the room, etc.) | Integer |

Appendix B - Centroids of K-prototype Clustering

| | | | | | | | | |
|---|------------------------------------|------------|----------|----------|----------|------------|-----------|----------|
| [31]: | sorted_centroid_df[numerical_cols] | | | | | | | |
| [31]: lead_time arrival_date_week_number arrival_date_day_of_month stays_in_weekend_nights stays_in_week_nights adults children babies adr previous_cancellations previous_bookings_not_canceled booking_changes days_in_waiting_list | | | | | | | | |
| 0 | 124.234244 | 36.763821 | 8.348933 | 0.960782 | 2.616061 | 1.877668 | 0.110196 | 0.009609 |
| 1 | 91.892061 | 15.12798 | 23.15605 | 0.872166 | 2.342033 | 1.814017 | 0.092535 | 0.006059 |
| 2 | 81.14827 | 14.317675 | 7.997924 | 0.89752 | 2.480442 | 1.839568 | 0.088688 | 0.006303 |
| 3 | 114.360591 | 38.51915 | 23.43045 | 0.971336 | 2.542796 | 1.892426 | 0.121028 | 0.009267 |
| babies previous_cancellations previous_bookings_not_canceled booking_changes days_in_waiting_list adr | | | | | | | | |
| 0.009609 | 0.173843 | | 0.082366 | 0.213965 | 2.116561 | 106.715672 | | |
| 0.006059 | 0.042124 | | 0.157474 | 0.222449 | 2.634094 | 95.175977 | | |
| 0.006303 | 0.039783 | | 0.194802 | 0.215602 | 1.794705 | 95.848875 | | |
| 0.009267 | 0.078879 | | 0.106004 | 0.231866 | 2.731404 | 108.221826 | | |
| booking_changes days_in_waiting_list adr required_car_parking_spaces total_of_special_requests | | | | | | | | |
| 0.213965 | 2.116561 | 106.715672 | | 0.061681 | | 0.580182 | | |
| 0.222449 | 2.634094 | 95.175977 | | 0.060668 | | 0.509655 | | |
| 0.215602 | 1.794705 | 95.848875 | | 0.059583 | | 0.569167 | | |
| 0.231866 | 2.731404 | 108.221826 | | 0.065025 | | 0.617703 | | |
| hotel arrival_date_year arrival_date_month meal country market_segment distribution_channel is_repeated_guest | | | | | | | | |
| 0 | City Hotel | 2016 | August | BB | PRT | Online TA | TA/TO | No |
| 1 | City Hotel | 2017 | April | BB | PRT | Online TA | TA/TO | No |
| 2 | City Hotel | 2017 | May | BB | PRT | Online TA | TA/TO | No |
| 3 | City Hotel | 2016 | July | BB | PRT | Online TA | TA/TO | No |
| distribution_channel is_repeated_guest reserved_room_type assigned_room_type deposit_type customer_type | | | | | | | | |
| TA/TO | No | | A | | A | No Deposit | Transient | |
| TA/TO | No | | A | | A | No Deposit | Transient | |
| TA/TO | No | | A | | A | No Deposit | Transient | |
| TA/TO | No | | A | | A | No Deposit | Transient | |

Appendix C - Clustering: Summary of Responses by LLM

| | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 |
|------------|---|---|---|--|
| ChatGPT | Long-term Planners | Extended Stay Guests | Standard Stay Guests | Seasonal Visitors |
| | High lead time (124.23) | High lead time (91.89) | Moderate lead time (81.15) | High lead time (114.36) |
| | Moderate previous cancellations (0.17) | High arrival date week number (15.13) | Moderate arrival date week number (14.32) | High arrival date week number (38.52) |
| | Moderate ADR (106.72) | High stays in week nights (2.34) | Moderate ADR (95.85) | High stays in week nights (2.54) |
| Gemini | Planned Weekend Travelers | Last Minute Business Travelers | Spontaneous Weekend Travelers | Planned Business Travelers |
| | High Lead Time (124.23 days): | Low Lead Time (91.89 days) | Very Short Lead Time (81.15 days) | High Lead Time (114.36 days) |
| | High Weekend Stays (0.96 nights) | High Weekday Arrivals (23.16 days of month) | Moderate Weekend Stays (0.89 nights) | High Weekday Arrivals (23.43 days of month): |
| | Moderate Booking Changes (0.21). | Moderate Booking Changes (0.22) | Low Booking Changes (0.22) | Low Booking Changes (0.23) |
| Perplexity | Long-Stay Guests | Short-Stay Leisure Travellers | Families | Business Travellers |
| | high average lead time (124.73 days) | Moderate lead time (93.49 days) | High lead time (114.51 days) | lowest average lead time (79.41 days) |
| | highest average number of stays in weekend nights (0.97) and week nights (2.62) | lowest average number of previous cancellations (0.04) and booking changes (0.22) | High Average number of children (0.12) | highest average number of previous bookings not cancelled (0.20) |
| | High ADR: (106.23) | Moderate ADR: (97.03) | High ADR: (107.47) | Moderate ADR: (95.81) |

Appendix D - ChatGPT and Gemini's response to Clustered Features and Segment Name

K You

I have a dataset regarding hotel cancellation rates. I have clustered the data using k-prototype and decided there are 4 clusters. I need you to do the following:

1. Identify the top 3 traits of each cluster based on the centroids below
2. Name the 4 clusters based on their properties

Here is the value of the centroids



ChatGPT

To identify the top 3 traits of each cluster, we'll examine the centroids provided for each cluster.

Here are the top 3 traits for each cluster based on the highest centroid values:

Cluster 0:

1. High lead time (124.23)
2. Moderate previous cancellations (0.17)
3. Moderate ADR (average daily rate) (106.72)

Cluster 1:

1. High lead time (91.89)
2. High arrival date week number (15.13)
3. High stays in week nights (2.34)

Cluster 3:

1. High lead time (114.36)
2. High arrival date week number (38.52)
3. High stays in week nights (2.54)

Based on these properties, we can name the clusters as follows:

1. Cluster 0: **Long-term Planners**
2. Cluster 1: **Extended Stay Guests**
3. Cluster 2: **Standard Stay Guests**
4. Cluster 3: **Seasonal Visitors**

These names reflect the dominant characteristics of each cluster, making it easier to interpret and understand their differences.





Analyzing Hotel Cancellation Rates by Cluster

Based on the centroids you provided, here's an analysis of the top 3 traits and suggested names for each of the 4 clusters:

Cluster 1: Long Lead Time, Weekend Stays

- Top Traits:
 - **High Lead Time (124.23 days):** Guests book far in advance, potentially indicating planned vacations.
 - **High Weekend Stays (0.96 nights):** Guests are more likely to book for weekend getaways.
 - **Moderate Booking Changes (0.21):** Guests might make some adjustments to their reservations.
- Name: **Planned Weekend Travelers**

Cluster 2: Short Lead Time, Weekday Arrivals

- Top Traits:
 - **Low Lead Time (91.89 days):** Guests book closer to arrival, suggesting spontaneous trips.
 - **High Weekday Arrivals (23.16 days of month):** Bookings are concentrated on weekdays.
 - **Moderate Booking Changes (0.22):** Similar to Cluster 1, some adjustments might be made.
- Name: **Last Minute Business Travelers**

Cluster 3: Short Lead Time, Weekend Stays

- Top Traits:
 - **Very Short Lead Time (81.15 days):** Bookings happen very close to arrival date.
 - **Moderate Weekend Stays (0.89 nights):** Weekend presence is still significant.
 - **Low Booking Changes (0.22):** Guests seem confident about their reservations.
- Name: **Spontaneous Weekend Travelers**

Cluster 4: Long Lead Time, Weekday Arrivals

- Top Traits:
 - **High Lead Time (114.36 days):** Similar to Cluster 1, bookings are made well in advance.
 - **High Weekday Arrivals (23.43 days of month):** Focus on weekday stays, potentially for business trips.
 - **Low Booking Changes (0.23):** Guests are likely confident in their plans.
- Name: **Planned Business Travelers**

Perplexity

Answer

Based on the cluster centroids provided, I would suggest the following cluster names and rationale:

Cluster 0: Long-Stay Guests

- This cluster has the highest average lead time (124.73 days) and the highest average number of stays in weekend nights (0.97) and week nights (2.62), indicating these guests tend to book well in advance and stay for longer durations.
- The average daily rate (ADR) for this cluster is also the highest at 106.23, suggesting these are higher-value customers.
- This cluster also has the highest average number of previous cancellations (0.17) and booking changes (0.21), implying these guests may be more flexible with their travel plans.

Cluster 1: Short-Stay Leisure Travelers

- This cluster has the lowest average lead time (93.49 days) and the lowest average number of stays in weekend nights (0.88) and week nights (2.36), indicating these guests tend to book closer to their travel dates and stay for shorter durations.
- The average daily rate (ADR) for this cluster is the lowest at 97.03, suggesting these are more price-sensitive customers.
- This cluster also has the lowest average number of previous cancellations (0.04) and booking changes (0.22), implying these guests may be more rigid with their travel plans.

Cluster 2: Families

- This cluster has the second highest average number of children (0.12) and the highest average number of babies (0.01), indicating these guests are likely traveling with families.
- The average lead time (114.51 days) and number of stays in weekend nights (0.96) and week nights (2.54) are also higher than Cluster 1, suggesting these guests plan their trips further in advance.
- The average daily rate (ADR) for this cluster is 107.47, which is higher than Cluster 1 but lower than Cluster 0, indicating these are mid-range customers.

Cluster 3: Business Travelers

- This cluster has the lowest average lead time (79.41 days) and the lowest average number of stays in weekend nights (0.90) and week nights (2.47), indicating these guests tend to book closer to their travel dates and stay for shorter durations.
- The average daily rate (ADR) for this cluster is 95.81, which is lower than Clusters 0 and 2, suggesting these are more price-sensitive customers.
- This cluster also has the highest average number of previous bookings not canceled (0.20), implying these guests may be more reliable and less likely to cancel their reservations.

ChatGPT Link: <https://chat.openai.com/share/3e99929c-4588-4c39-8346-8491d972e7bf>

Gemini Link: <https://gemini.google.com/app/5344480aa2a11239>

Perplexity Link: https://www.perplexity.ai/search/after-ive-applied-EZhc_mQeTa.tMnqjQGip1w#0

Appendix E - ChatGPT & Gemini's Response to High Correlation: Weekday vs. Weekend Stays

Prompt:

M You
for the dataset (<https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand>), the highest correlation seems to be between stay_in_week_nights and stay_in_weekend_nights with a correlation coefficient of 0.49. Give me 3 possible explanations for it.

I think that it could be due to package deals or promotions, length of stay preferences and availability

Response:

ChatGPT
Certainly! The correlation between "stay_in_week_nights" and "stay_in_weekend_nights" could be influenced by several factors. Here are three possible explanations:

- Package Deals or Promotions:** Hotels often offer package deals or promotions that encourage guests to extend their stays over both weekdays and weekends. For example, a hotel might offer a discounted rate or additional amenities for guests who stay for a certain number of nights, regardless of whether those nights fall on weekdays or weekends. This incentive could lead to a positive correlation between the number of nights stayed during the week and on weekends.
- Length of Stay Preferences:** Some guests may have preferences for longer stays that span both weekdays and weekends. This could be due to various reasons such as leisure travel, business trips, or attending events that occur over multiple days. As a result, individuals who prefer longer stays are likely to book accommodations for both weekdays and weekends, leading to a positive correlation between the durations of weekday and weekend stays.
- Availability and Booking Patterns:** The correlation could also be influenced by the availability of accommodations and booking patterns. Hotels may experience higher demand for rooms during weekends compared to weekdays, leading to a higher likelihood of guests extending their stays into the weekends if they initially booked for weekdays. Similarly, guests may prefer to book longer stays that cover both weekdays and weekends when there is limited availability or when they want to secure a continuous period of accommodation. This pattern of booking behavior could contribute to the observed correlation between weekday and weekend stays.

These explanations highlight how factors such as promotions, guest preferences, and booking dynamics can contribute to the correlation between stay durations on weekdays and weekends in hotel bookings.

These are just a few possibilities, and the real-world reasons could be a combination of these factors or something else entirely. The data itself might not reveal the definitive cause, but it highlights an interesting trend for further exploration.

Share icon: Copy icon: Gemini icon: Google icon: More options icon:

ChatGPT link: <https://chat.openai.com/share/fb5e6599-903c-42cb-8d97-b8be52b19ee0>

Gemini link: <https://g.co/gemini/share/33bc64df9233>

Appendix F - ChatGPT code to calculate Cramér's V values

Prompt:

 You
give me the code to calculate the association between categorical variables in the dataset and then calculate the cramer's V values
Import

Response:

```
python Copy code  
  
import pandas as pd  
import scipy.stats as ss  
  
def cramers_v(x, y):  
    confusion_matrix = pd.crosstab(x, y)  
    chi2 = ss.chi2_contingency(confusion_matrix)[0]  
    n = confusion_matrix.sum().sum()  
    phi2 = chi2 / n  
    r, k = confusion_matrix.shape  
    phi2corr = max(0, phi2 - ((k-1)*(r-1)) / (n-1))  
    rcorr = r - ((r-1)**2) / (n-1)  
    kcorr = k - ((k-1)**2) / (n-1)  
    return np.sqrt(phi2corr / min((kcorr-1), (rcorr-1)))  
  
# Example usage  
# Assuming you have a DataFrame named 'data' with categorical variables  
# For example:  
# data = pd.DataFrame({'A': ['Yes', 'Yes', 'No', 'No'], 'B': ['High', 'Low', 'High',  
# Calculate Cramer's V for columns 'A' and 'B'  
cramers_v_value = cramers_v(data['A'], data['B'])  
print("Cramer's V value:", cramers_v_value)
```

```
python Copy code  
  
# Assuming you have calculated Cramer's V values for pairs of categorical variables at  
threshold = 0.3  
highly_correlated_pairs = [(pair, value) for pair, value in cramers_v_values.items()  
  
print("Pairs with Cramer's V > 0.3:")  
for pair, value in highly_correlated_pairs:  
    print(pair, "-", value)
```

Cramer's V results:

| Variable 1 | Variable 2 | Cramer's V |
|----------------------|----------------------|------------|
| reserved_room_type | assigned_room_type | 0.778 |
| market_segment | distribution_channel | 0.614 |
| deposit_type | is_canceled | 0.481 |
| arrival_date_year | arrival_date_month | 0.428 |
| hotel | assigned_room_type | 0.391 |
| market_segment | deposit_type | 0.374 |
| country | is_canceled | 0.358 |
| market_segment | is_repeated_guest | 0.353 |
| hotel | reserved_room_type | 0.325 |
| hotel | meal | 0.318 |
| country | deposit_type | 0.312 |
| distribution_channel | is_repeated_guest | 0.302 |
| hotel | country | 0.301 |

Link: <https://chat.openai.com/share/7167c04e-1e5e-423b-b5fe-b0b64a1a935f>

Appendix G - Drivers for High Cancellation Rates in City Hotels

Prompt:

 You
looking at the dataset <df>, what is another 1 reason why city hotels have higher cancellation rates than resort hotels


Responses:

Purpose of Travel

 ChatGPT
Another reason for the higher cancellation rates in city hotels compared to resort hotels, as indicated by the dataset, could be the presence of a larger proportion of business travelers.

City hotels often attract a significant number of business travelers who may need to adjust their travel plans due to changes in business meetings, conferences, or other work-related obligations. Business trips are often subject to last-minute alterations, such as cancellations or rescheduling, which can lead to a higher incidence of canceled bookings in city hotels.

In contrast, resort hotels primarily cater to leisure travelers who are typically on vacation or holiday trips. Leisure travelers tend to have more stable vacation plans and are less likely to cancel their bookings unless faced with unavoidable circumstances.

Therefore, the prevalence of business travelers in city hotels, as reflected in the dataset, may contribute to the higher cancellation rates observed in comparison to resort hotels.

Low Booking Lead Time:

 ChatGPT
Another reason for the higher cancellation rates in city hotels compared to resort hotels, as evident from the dataset, could be the difference in the average lead time for bookings.

Lead time refers to the number of days that elapsed between the entering date of the booking into the hotel's system and the arrival date. In the dataset, it might be observed that city hotels have a shorter average lead time compared to resort hotels.

Shorter lead times in city hotels could indicate a higher proportion of last-minute bookings or bookings made closer to the travel date. Guests who book accommodations with shorter lead times may be more likely to cancel their reservations if they encounter changes in their plans or find alternative options available.

In contrast, resort hotels may have a longer average lead time, with guests typically booking their accommodations well in advance of their travel dates. Guests who book accommodations with longer lead times may have firmer travel plans and are less likely to cancel their reservations unless faced with significant unforeseen circumstances.

Therefore, the difference in average lead time for bookings between city hotels and resort hotels, as observed in the dataset, may contribute to the higher cancellation rates experienced by city hotels.

Link: <https://chat.openai.com/share/9fd77b2d-4380-46b2-a0e5-143435b0669f>

Appendix H - Contradicting Conclusions from ChatGPT for Similar Issues

Prompt:

 You
looking at the dataset <df>, what is another 1 reason why city hotels have higher cancellation rates than resort hotels



Response:

 ChatGPT
Another reason for the higher cancellation rates in city hotels compared to resort hotels, as indicated by the dataset, could be the presence of a larger proportion of business travelers.

City hotels often attract a significant number of business travelers who may need to adjust their travel plans due to changes in business meetings, conferences, or other work-related obligations. Business trips are often subject to last-minute alterations, such as cancellations or rescheduling, which can lead to a higher incidence of canceled bookings in city hotels.

In contrast, resort hotels primarily cater to leisure travelers who are typically on vacation or holiday trips. Leisure travelers tend to have more stable vacation plans and are less likely to cancel their bookings unless faced with unavoidable circumstances.

Therefore, the prevalence of business travelers in city hotels, as reflected in the dataset, may contribute to the higher cancellation rates observed in comparison to resort hotels.

Prompt:

 You
<is_canceled> is the target variable, while the other columns are all features in the dataset that may influence <is_canceled>.

Come up with 2 different hypothesis as to which factors will influence the target variables the most

Response:

 ChatGPT
Certainly! Here are two different hypotheses regarding factors that may influence the target variable `<is_canceled>` the most:

1. **Lead Time Hypothesis:** This hypothesis suggests that the lead time between the booking date and the arrival date will strongly influence the cancellation of bookings. The rationale behind this hypothesis is that customers who book well in advance might be more likely to cancel their reservations due to changes in plans or unforeseen circumstances. Conversely, last-minute bookings might have lower cancellation rates because customers are less likely to cancel when the arrival date is imminent.
2. **Deposit Type Hypothesis:** This hypothesis posits that the type of deposit made by the customer will significantly impact the cancellation of bookings. Specifically, bookings made with a "No Deposit" requirement may have higher cancellation rates compared to bookings where a non-refundable deposit was made. The reasoning behind this hypothesis is that customers who

Link (Top): <https://chat.openai.com/share/9fd77b2d-4380-46b2-a0e5-143435b0669f>

Link (Bottom): <https://chat.openai.com/share/5fc2935-2371-4fa7-82cc-77bb7075fd17>

Appendix I - Outlier Situations Where Non-refundable Deposits Generate Higher Cancellation Rates

Prompt:

 You
In what situation will non-refundable deposits generate higher cancellation rates

Response:

 ChatGPT
Non-refundable deposits might generate higher cancellation rates in certain situations where the perceived value or circumstances change for the customer after making the deposit. Here are a few scenarios where non-refundable deposits could lead to higher cancellation rates:

1. **Unforeseen Circumstances:** If the customer experiences unexpected events such as illness, family emergencies, or travel restrictions, they may need to cancel their booking despite the non-refundable deposit. In such cases, the customer might prioritize addressing the urgent situation over the financial loss incurred from the deposit.
2. **Change in Plans:** Customers might change their travel plans due to evolving preferences, changes in schedules, or alternative opportunities. Even though the deposit is non-refundable, customers may still decide to cancel their booking if they perceive greater value in alternative options or destinations.
3. **Financial Constraints:** Economic factors such as job loss, financial instability, or unforeseen expenses might force customers to reconsider their travel expenses, leading to cancellations despite the non-refundable deposit.
4. **Unsatisfactory Conditions:** If customers find out new information about the accommodation or destination that they find unsatisfactory (e.g., negative reviews, safety concerns, or unfavorable weather forecasts), they may choose to cancel their booking, accepting the loss of the deposit as a lesser evil compared to experiencing a disappointing trip.
5. **Booking Errors or Misunderstandings:** Customers might make reservations with incorrect dates, room types, or other details, leading to the need for cancellation. In such cases, customers might be less willing to proceed with the booking, even with a non-refundable deposit, if it doesn't meet their expectations or requirements.

Overall, while non-refundable deposits are intended to discourage cancellations and secure commitment from customers, various factors can still influence customers to cancel their bookings despite the financial loss. Therefore, it's essential for businesses to consider these factors and provide flexibility or alternatives to mitigate cancellation rates and maintain customer satisfaction.

Appendix J - Estimate of Type I and Type II errors

Prompt:

 You

given that you have expert domain knowledge about hotel booking cancellations, what would be the cost of false positives and false negatives in monetary value? Give me an estimated figure and provide reasons to justify that amount

Response:

 ChatGPT

As an AI language model, I don't possess real-time data or access to proprietary information on specific industries or businesses, including the hotel booking industry. However, I can provide you with a general framework for understanding the costs associated with false positives and false negatives in hotel booking cancellations, along with some hypothetical figures to illustrate potential impacts.

1. False Positives:

- **Definition:** In the context of hotel booking cancellations, a false positive occurs when a booking is incorrectly flagged as likely to be canceled, leading to unnecessary actions such as offering discounts or allocating resources to reconfirm the booking.
- **Costs:**
 - **Loss of Revenue:** Offering discounts or incentives to retain customers who were not actually planning to cancel their bookings can lead to a direct loss of revenue.
 - **Operational Costs:** Allocating staff time and resources to follow up on false positives incurs operational expenses.
 - **Customer Dissatisfaction:** In some cases, offering discounts or additional services to prevent cancellations can set a precedent for future bookings, potentially leading to lower revenue per booking.
- **Estimated Cost:** Depending on the scale of false positives and the actions taken to mitigate them, the cost per false positive could range from \$50 to \$200 or more.

2. False Negatives:

- **Definition:** A false negative occurs when a booking that is likely to be canceled is not identified as such, leading to missed opportunities to mitigate the cancellation.
- **Costs:**
 - **Lost Revenue:** When a booking that could have been canceled is not identified in advance, the hotel may lose out on potential revenue if the cancellation occurs without opportunity for intervention.
 - **Opportunity Cost:** In addition to direct revenue loss, there may be missed opportunities to offer alternative arrangements or incentives to retain the booking.
 - **Reputational Damage:** If a customer's booking is canceled unexpectedly and no proactive measures were taken by the hotel, it could result in negative reviews or damage to the hotel's reputation.
- **Estimated Cost:** The cost per false negative can vary significantly depending on factors such as the average booking value, occupancy rates, and the effectiveness of mitigation strategies. A conservative estimate could range from \$100 to \$500 per false negative.

It's important to note that these estimated costs are hypothetical and may vary widely based on factors specific to individual hotels, such as location, brand reputation, customer demographics, and pricing strategies. Actual costs would need to be determined through careful analysis of historical data and ongoing monitoring of cancellation patterns.

Link: <https://chat.openai.com/share/f17e1cd3-9884-4f48-9c43-c9b21d6b6812>

Appendix K - XGB Hyperparameter Tuning

Prompt:

M You
I would tune my XGBclassifier model for improved performance. Provide me the code to conduct hyper parameter tuning to maximise my AUC.

```
import
```

Response:

ChatGPT
Certainly! Below is an example code for conducting hyperparameter tuning for an XGBoost classifier using GridSearchCV to maximize the Area Under the ROC Curve (AUC):

```
python
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split, GridSearchCV
from xgboost import XGBClassifier
from sklearn.metrics import roc_auc_score

# Assuming you have your data loaded into X and y variables
# X should be your features and y should be your target variable

# Splitting the data into train and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Defining the XGBoost classifier
xgb = XGBClassifier()

# Defining a parameter grid for hyperparameter tuning
param_grid = {
    'n_estimators': [100, 200, 300],
    'learning_rate': [0.01, 0.1, 0.2],
    'max_depth': [3, 5, 7],
    'min_child_weight': [1, 3, 5],
    'gamma': [0, 0.1, 0.2],
    'subsample': [0.6, 0.8, 1.0],
    'colsample_bytree': [0.6, 0.8, 1.0],
}
```

Link: <https://chat.openai.com/share/360b2639-cfad-4837-b02e-eb1a3393c5bd>

Appendix L - GridSearchCV Scoring

Original response by ChatGPT (Scoring = “accuracy”)

GPT Providing wrong scoring matrix

```
grid_search = GridSearchCV(estimator=rf_classifier, param_grid=param_grid, cv=5, scoring='accuracy', n_jobs=-1)
```

Corrected GridSearchCV (Scoring = “roc_auc”)

Adjusting Scoring to "roc_auc"

```
grid_search = GridSearchCV(estimator=rf_classifier, param_grid=param_grid, cv=5, scoring='roc_auc', n_jobs=-1)
```

Appendix M - Alternative Hyperparameter Tuning Method

Prompt:



You

this gridsearch is taking too long to run. What other ways can i tune my model?

Response:



ChatGPT

If GridSearchCV is taking too long to run due to the large number of combinations, you can consider using RandomizedSearchCV instead. RandomizedSearchCV randomly selects a subset of the hyperparameter combinations specified rather than exhaustively searching through all possible combinations like GridSearchCV. This can significantly reduce the time taken for hyperparameter tuning while still providing good results.

Here's how you can modify the code to use RandomizedSearchCV:

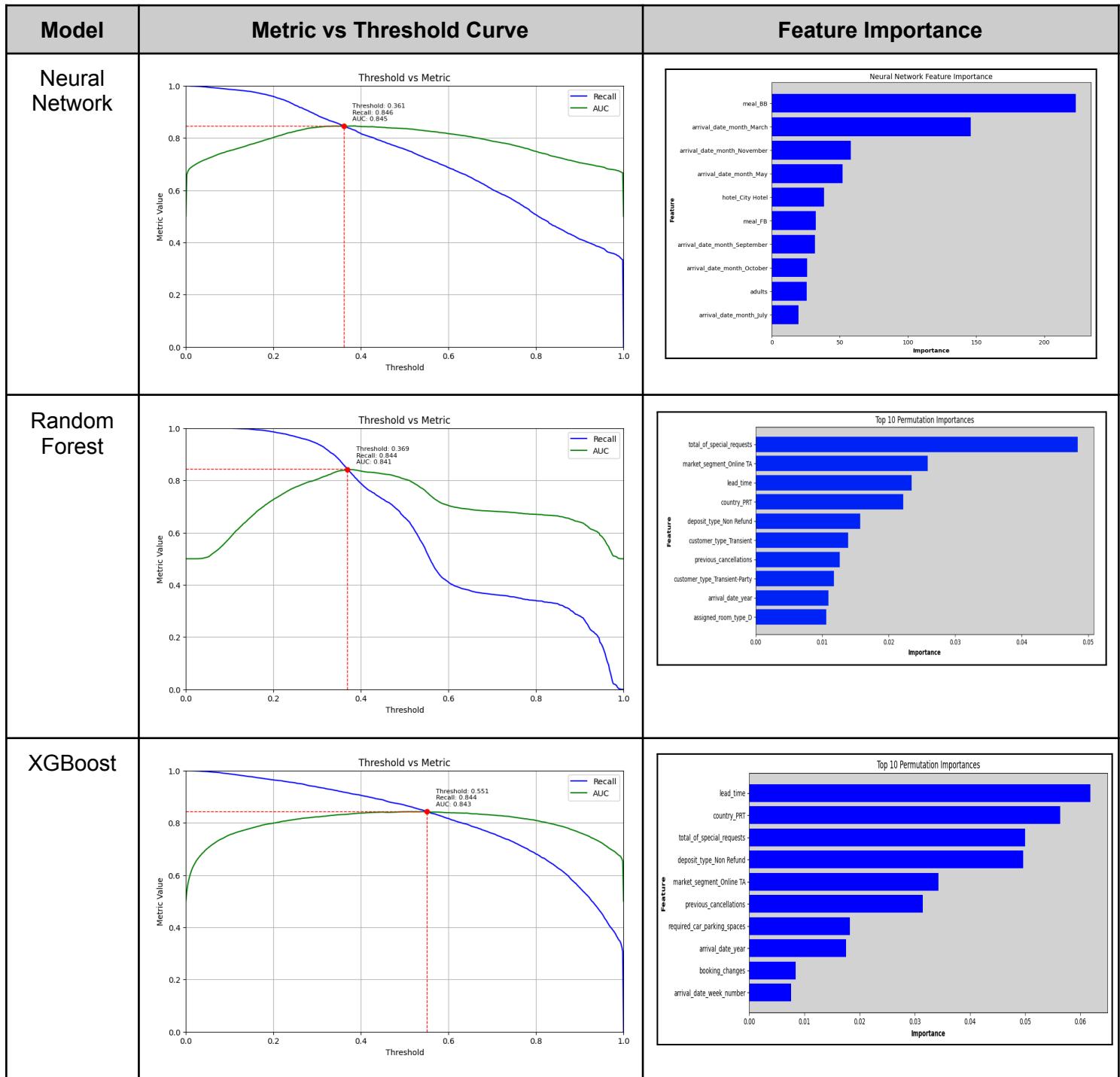
```
python Copy code
from sklearn.model_selection import RandomizedSearchCV

# Defining a parameter grid for hyperparameter tuning
param_grid = {
    'n_estimators': [100, 200, 300, 400, 500],
    'learning_rate': [0.01, 0.05, 0.1, 0.2],
    'max_depth': [3, 4, 5, 6, 7, 8],
    'min_child_weight': [1, 2, 3, 4, 5],
    'gamma': [0, 0.1, 0.2, 0.3, 0.4],
    'subsample': [0.6, 0.7, 0.8, 0.9, 1.0],
    'colsample_bytree': [0.6, 0.7, 0.8, 0.9, 1.0],
}

# Creating RandomizedSearchCV object with specified parameters
random_search = RandomizedSearchCV(estimator=xgb, param_distributions=param_grid, n_i
```

Link: <https://chat.openai.com/share/360b2639-cfad-4837-b02e-eb1a3393c5bd>

Appendix N - Metric vs Threshold Curve + Feature importance for all 3 models



Appendix O - Possible Reason for Increase in False Negatives for XGB After Threshold Tuning

Prompt:

 You
in a hotel cancellation prediction project where the positive class is 'cancelled' and negative class is 'not cancelled' and the goal is to maximise both recall and auc by varying the threshold and finding the intersection between the 2 graphs. The table below summarises the false negatives pre and post tuning, as well as the optimal threshold that maximises the AUC and Recall for the models.

Model

| Model | False Negatives Before | False Negatives After | Threshold |
|----------------|------------------------|-----------------------|-----------|
| Neural Network | 2171 | 1377 | 0.361 |
| XGBoost | 1148 | 1349 | 0.551 |
| Random Forest | 2970 | 1354 | 0.369 |

Given that you're a senior data scientist with expert domain knowledge in the hotel industry, give me possible reasons why the false negatives before tuning is lower than that after tuning when it should be the other way around for xgb

Response:

 ChatGPT
There could be several reasons why the false negatives before tuning are lower than those after tuning for XGBoost. Here are some possible explanations:

- Feature Selection:** Before tuning, the model might have been trained on a different set of features or a subset of the available features. After tuning, feature selection might have been refined, leading to better discriminatory power and higher sensitivity to cancellations.
- Hyperparameter Tuning:** During the tuning process, hyperparameters such as the learning rate, maximum depth of trees, and regularization parameters might have been adjusted. While these adjustments aim to improve overall model performance, they could inadvertently increase the number of false negatives if not carefully tuned.
- Data Preprocessing:** Before tuning, the data might have undergone less preprocessing compared to after tuning. More thorough preprocessing steps such as feature scaling, handling missing values, or outlier detection could have resulted in a better understanding of the data and hence a more accurate model.
- Class Imbalance:** The distribution of cancelled and not cancelled instances might have been skewed differently before and after tuning. If the tuning process inadvertently skewed the model towards the majority class (not cancelled), it could lead to an increase in false negatives for the minority class (cancelled).
- Overfitting:** Tuning hyperparameters might have unintentionally led to overfitting on the training data, causing the model to generalize poorly on unseen data and thus increasing false negatives.
- Data Leakage:** During the tuning process, there might have been inadvertent data leakage, where information from the validation or test set influenced the model's training process, leading to overly optimistic performance metrics.
- Model Complexity:** After tuning, the model might have become more complex, potentially leading to higher false negatives due to increased susceptibility to noise or outliers.

Considering these possibilities, it's important to carefully analyze the tuning process, evaluate its impact on model performance, and potentially iterate to find the best balance between maximizing AUC and recall while minimizing false negatives. Additionally, domain expertise in the hotel industry can provide valuable insights into the specific characteristics and nuances of the data that may influence model performance.

Link: <https://chat.openai.com/share/6fb80d07-4863-418f-ac49-d97e74088695>

Appendix P - Code for Neural Network Model

Prompt:

DH You

Using the dataset <df>, I want to create a neural network model that can predict if a guest will cancel their bookings. Using the <is_canceled> column as the target column, what is the best model for this task? Evaluate the model using roc_auc. Include methods that can reduce overfitting.

```
import
```

Response:

python

 Copy code

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler, LabelEncoder
from sklearn.metrics import roc_auc_score
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense, Dropout
from tensorflow.keras.callbacks import EarlyStopping

# Load the dataset
# Assume <df> is already loaded here

# Encode categorical variables
encoder = LabelEncoder()
df_encoded = df.apply(encoder.fit_transform)

# Split the data into features and target
X = df_encoded.drop('is_canceled', axis=1)
y = df_encoded['is_canceled']

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Scale numerical features
scaler = StandardScaler()
```



```

# Scale numerical features
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# Define the neural network architecture
model = Sequential([
    Dense(64, activation='relu', input_shape=(X_train_scaled.shape[1],)),
    Dropout(0.5), # Dropout layer to reduce overfitting
    Dense(32, activation='relu'),
    Dropout(0.5), # Dropout layer to reduce overfitting
    Dense(1, activation='sigmoid')
])

# Compile the model
model.compile(optimizer='adam', loss='binary_crossentropy', metrics=['accuracy'])

# Implement early stopping to prevent overfitting
early_stopping = EarlyStopping(monitor='val_loss', patience=3, restore_best_weights=True)

# Train the model
history = model.fit(X_train_scaled, y_train, epochs=50, batch_size=64, validation_split=0.2)

# Evaluate the model using ROC AUC on the testing data
y_pred_proba = model.predict(X_test_scaled).ravel()
roc_auc = roc_auc_score(y_test, y_pred_proba)
print("ROC AUC:", roc_auc)

```

Link: <https://chat.openai.com/c/6dab2e24-fb57-4536-b5a4-06f47776ad98>

Appendix Q - Consumer Preferences Based On The Identified Consumer Segments

Prompt:

 You

For a hotelier, there are multiple guest profile types that must be considered. Considering the following guest profiles, what are the things that each profile would value such that they will not cancel their bookings:

<Long-Stay Guests>: Guests who has the highest number of stays, high average booking lead time, high revenue per room booked

<Short-Stay Leisure Travellers>: moderate booking lead time, moderate revenue per room booked

<Families>: High booking lead time, usually has children, high revenue per room booked

Business travellers: Lowest average booking lead time, highest number of previous bookings not cancelled, moderate revenue per room booked

Responses:



ChatGPT

Understanding what each guest profile values can help tailor offerings and services to meet their needs, reducing the likelihood of cancellations:

1. Long-Stay Guests:

- Value Comfort and Convenience: They seek a home-away-from-home experience, so amenities like kitchenettes, laundry facilities, and spacious rooms are important.
- Personalized Services: Offering personalized services such as grocery shopping assistance, housekeeping preferences, and local area recommendations can enhance their experience.
- Loyalty Programs: Implementing a robust loyalty program with perks like room upgrades, extended stay discounts, and exclusive events can incentivize them to continue booking.
- Special Long-Stay Packages: Offering discounted rates for extended stays or bundling services like meals or spa treatments can encourage them to commit to their bookings.

2. Short-Stay Leisure Travelers:

- Experience and Activities: They value experiences and activities, so offering packages that include local tours, dining vouchers, or spa treatments can entice them to keep their bookings.
- Flexibility: Providing flexible booking options such as free cancellation up to a certain date or no prepayment requirements can alleviate any hesitations they may have about booking in advance.



- **Social Media Presence:** Having a strong presence on social media platforms with engaging content about local attractions, events, and experiences can capture their interest and encourage them to keep their bookings.
- **Seamless Booking Process:** Ensuring a user-friendly and intuitive booking process on both desktop and mobile platforms can enhance their booking experience and reduce the likelihood of cancellations due to frustration.

3. Families:

- **Child-Friendly Amenities:** Offering amenities such as cribs, high chairs, children's menus, and family-friendly entertainment options can cater to their needs and make their stay more enjoyable.
- **Spacious Accommodations:** Providing larger room options or interconnected rooms can accommodate families of different sizes and ensure everyone has enough space.
- **Safety and Security:** Highlighting safety measures such as childproofing features, secure premises, and trained staff can reassure families and make them more comfortable keeping their bookings.
- **Family-Focused Activities:** Organizing family-friendly activities or events such as movie nights, craft workshops, or outdoor adventures can create memorable experiences for families and encourage them to keep their bookings.

4. Business Travelers:

- **Efficiency and Convenience:** They prioritize efficiency and convenience, so offering services such as express check-in/out, business centers, and complimentary Wi-Fi can make their stay more productive.
- **Location Proximity:** Being located near business districts, conference centers, or transportation hubs can save them time and effort, making them more likely to keep their bookings.
- **Flexible Booking Policies:** Providing flexible booking policies such as last-minute cancellation options or corporate rates can accommodate their unpredictable schedules and encourage them to book with confidence.
- **Business-Friendly Amenities:** Offering amenities like meeting rooms, in-room workspaces, and business concierge services can cater to their professional needs and make their stay more comfortable.



Thanks prof, we enjoyed this module. Happy RUNning