Introduction

The hospitality industry faces ongoing challenges in managing reservation uncertainty, particularly due to high rates of booking cancellations, which disrupt operational planning, reduce revenue efficiency, and create difficulties in inventory management. In large-scale hotel operations, even small inaccuracies in demand forecasting can significantly impact Revenue Per Available Room (RevPAR), staffing decisions, and the overall customer experience. As a result, understanding and predicting cancellation behavior has become an important problem that can be addressed with data-driven solutions.

In this project, we apply data analytics and data mining techniques to analyze hotel booking data and uncover patterns associated with reservation cancellations. Using a real-world dataset containing over 119,000 booking records, we aim to move beyond simple descriptive analysis and instead extract meaningful insights through structured analytical methods. As part of this process, we first prepare the dataset through data preprocessing steps such as handling missing values, encoding categorical variables, and scaling features to ensure the data is suitable for further analysis.

The primary goal of this project is to develop a deeper understanding of booking behavior and to support predictive modeling of cancellation risk. By transforming raw booking data into a structured, analyzable format, this work lays the foundation for applying more advanced techniques, such as classification, clustering, and association rule mining. Ultimately, the insights generated from this analysis can help inform more effective decision-making strategies in the hospitality industry, including improved cancellation policies, better resource allocation, and more accurate demand forecasting

Brief Description of the Data Set

The dataset used for this project contains information on hotel reservations, including cancelled and existing bookings. Each row in this dataset represents a single booking, and it records comprehensive features describing the booking, which supports the analysis of booking cancellations.

The wide range of features focuses on three major aspects regarding the reservation. Customer-related features describe the behavioural and historical characteristics, such as customer type, previous cancellations and demographics. The second type is time-related features, including lead time and length of stay. Third, operational features like deposit types, distribution channels and room assignments describe

The target in this analysis is the indicator variable that shows whether the reservation was eventually cancelled. When it is cancelled, it shows 1 and 0 when it is not. This variable serves as the foundation for conducting predictive data analysis.

For the purposes of three data analysis techniques used to explore patterns related to booking cancellation, which are Association Rule Mining, Clustering, and Classification, this dataset is cleaned, and several subsets are derived to align with the requirements of each of these three techniques.

Problem definition and motivation

The core problem addressed in this project is the high degree of uncertainty surrounding hotel reservation cancellations, which directly leads to suboptimal inventory allocation and significant revenue leakage. While cancellations are a standard aspect of the hospitality industry, the inability to accurately anticipate which specific bookings will be canceled prevents hotels from optimizing their pricing and availability strategies.

For example, if a hotel anticipates a fully booked weekend and turns away new potential customers, but ultimately experiences a 20% cancellation rate at the last minute, those vacant rooms represent irrecoverable lost revenue (lowering the Revenue Per Available Room, or RevPAR). Conversely, if management attempts to mitigate this by blindly overbooking the property and fewer guests cancel than expected, the hotel will be forced to "walk" guests to other properties, resulting in severe customer dissatisfaction, penalty costs, and long-term reputational damage.

The motivation for this study is to transition hotel management from a reactive operational model to a proactive, data-driven strategy. By identifying the underlying patterns and risk factors associated with cancellations, hoteliers can implement dynamic overbooking limits, establish targeted deposit policies (e.g., requiring stricter non-refundable deposits for high-risk profiles), and optimize staffing and resource planning. Solving this problem provides immediate, quantifiable value to hospitality operations.

Data Analysis Method/Task

To comprehensively address the problem of reservation uncertainty, this study employs a structured, multi-faceted data analytics pipeline. Moving beyond basic descriptive statistics, we apply three advanced data mining techniques: Classification, Clustering, and Association Rule Evaluation. Each method was selected for its distinct analytical strengths, requiring specific data preparation steps and evaluation metrics to ensure robust, actionable results for the hospitality domain.

1. Classification: Decision Tree using Information Gain

The primary predictive task is to classify whether a booking will be canceled (is_canceled = 1) or honored (is_canceled = 0). A Decision Tree classifier was selected for its high interpretability: unlike "black-box" models, it generates explicit decision rules and inherently performs feature selection via Information Gain (Entropy), identifying the most critical variables such as deposit type, lead time, and special requests.

The dataset was split using proportionally stratified sampling (80% training, 20% testing) to preserve the natural class distribution. Accuracy and ROC-AUC were selected as primary evaluation measures—ROC-AUC being critical for assessing class separability across all probability thresholds. The confusion matrix was additionally used to analyze the precision-recall trade-off, as false negatives (missed cancellations) carry a distinct financial penalty compared to false positives.

2. Clustering: K-Means for Customer Segmentation

K-Means clustering was applied to segment bookings into distinct behavioral personas based on continuous features: lead_time, adr, and total_of_special_requests. Since K-Means relies on Euclidean distance, standardizing all numerical variables was a mandatory preprocessing step to prevent large-range features from dominating. The Elbow Method and Silhouette Analysis were used to select the optimal K, with business interpretability retained as the final arbitration criterion.

3. Association Rule Evaluation: Probabilistic Dependency Analysis

Rather than exhaustive frequent itemset mining (e.g., Apriori), a targeted probabilistic approach was used to directly compute conditional probabilities for pre-identified business-relevant rules. Continuous variables were discretized into binary indicators (e.g., lead_time > 90 days → lead_time_long). Three metrics were computed: Support (rule frequency), Confidence (P(cancellation | antecedent)), and Lift (confidence relative to baseline cancellation rate). A Lift > 1.0 confirms a positive statistical association, allowing management to isolate genuine risk factors without redundant rule generation.

Experimental Results

In this section, we present the quantitative findings and empirical evaluations of the three data analytics techniques applied to the structured hotel booking dataset. All experiments were conducted using Python, with data preprocessed to handle the 16,344 missing values and scaled to ensure algorithmic stability.

1. Predictive Modeling and Classification Metrics

A Decision Tree (entropy criterion) was trained on an 80% training set (95,512 records) and evaluated on a 20% hold-out set (23,878 records). Hyperparameter tuning on max_depth (3 to None) revealed a clear overfitting onset at higher depths. As shown in Figure 1, max_depth=8 was selected as optimal, yielding training/test accuracies of 0.828/0.827 and ROC-AUC of 0.909.

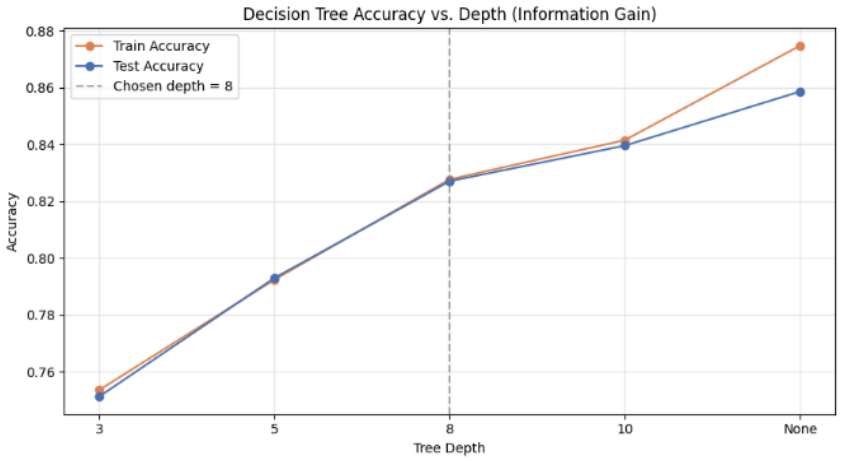

Figure 1

As shown in Figure 2, the confusion matrix reveals a cancellation recall of 66%, meaning roughly 3,000 genuine cancellations were misclassified as confirmed stays—false negatives that directly translate to unrecovered room revenue. Feature importance identified deposit_type_NonRefund as dominant (0.415), but this was diagnosed as a recording artefact (99.4% canceled). The strongest genuine predictors were market_segment_Online TA (0.137), lead_time (0.077), total_of_special_requests (0.068), and required_car_parking_spaces (0.066).

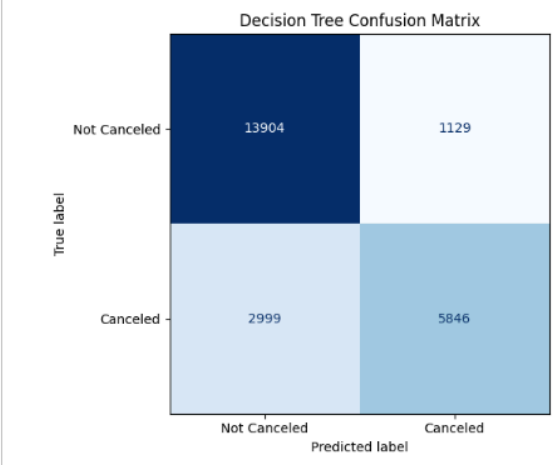

Figure 2

2. Customer Segmentation via Clustering

Determining the optimal K required a dual-validation approach, as the Elbow curve showed a smooth decrease from K=2 through K=10 without a clear inflection point. Silhouette analysis peaked at K=2 (0.9627), but this was diagnosed as a data artefact driven by highly correlated binary features. A stabilization plateau at K=4 (Silhouette Score: 0.1720) was selected for its superior business interpretability.

As shown in Figure 3, K=4 partitioned the data into four operationally distinct archetypes: **Cluster 0** (Early Placeholder Bookers, N=1,345) showed the highest cancellation risk (55%) with extreme lead times (250 days) but near-zero engagement; Cluster 1 (Habitual Cancellers, N=72,568), the largest segment, had moderate lead times (102 days) and the highest prior cancellation rate (0.13), representing ~30,479 lost room-nights; Cluster 2 (Committed Self-Drive Guests, N=7,409) achieved a 0% cancellation rate, defined entirely by parking requirements (avg. 1.006); Cluster 3 (High-Value Planners, N=38,068) featured the highest ADR (123.9) and most special requests (1.08) with a moderate 33% cancellation rate.

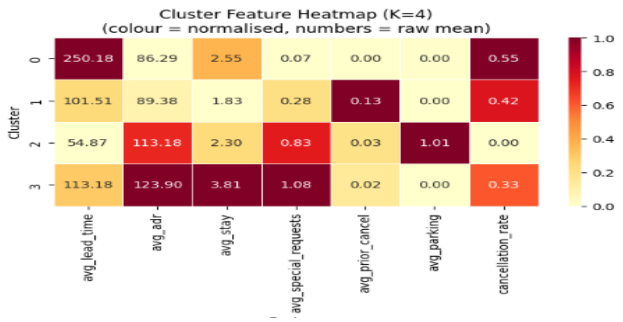

Figure 3

3. Probabilistic Dependency Analysis (Association Rules)

Finally, computing custom association rules yielded highly specific, quantifiable metrics mapping booking traits to cancellation probabilities. The strongest behavioral indicator was prior history: the rule {previous_cancel_yes} → {canceled} yielded a support of 0.049, a confidence of 91.64%, and a lift of 2.47, indicating these guests are nearly 2.5 times more likely to cancel than the baseline average. The data artefact observed in classification was empirically confirmed here, with {deposit_type_Non Refund} → {canceled} showing 99.36% confidence and a lift of 2.68. Conversely, {deposit_type_Refundable} → {canceled} showed only 22.2% confidence and a lift of 0.60.

Analyzing temporal factors, the rule {lead_time_long} → {canceled} (representing lead times exceeding 90 days) produced a confidence of 50.65% and a lift of 1.36. Customer type also proved highly discriminative: the rule {customer_type_Transient} → {canceled} resulted in a 40.74% confidence (Lift = 1.10), whereas {customer_type_Transient-Party} → {canceled} dropped to a 25.42% confidence (Lift = 0.68). This stark numerical contrast empirically proves that group or party bookings are substantially more stable than individual transient reservations.

Discussion

The experimental results validate that hotel reservation cancellations are not random occurrences but predictable events driven by specific behavioral and temporal patterns. However, translating these data-driven insights into operational business strategies requires a critical evaluation of model performance, data quality, and the financial cost of classification errors.

1. Model Evaluation and the Cost of Misclassification

The Decision Tree achieved 82.7% accuracy and ROC-AUC of 0.909, but the 66% cancellation recall highlights a key operational gap: one-third of true cancellations are missed (false negatives), directly causing unrecovered room revenue. Conversely, false positives force hotels to walk arriving guests, incurring penalty costs and reputational damage. Since the default threshold (0.5) optimizes overall accuracy rather than minimizing false negatives, management should lower this threshold or apply asymmetric class weights depending on their overbooking aggressiveness.

2. Data Quality and the "Non-Refundable" Artefact

deposit_type_NonRefund was identified as the strongest predictor (0.415), yet 99.4% of these bookings were recorded as canceled—the opposite of its business logic. This strongly suggests a system recording artefact where property management systems automatically log unfulfilled non-refundable bookings as "canceled" to close the ledger. This underscores that algorithmic feature importance must always be validated with domain expertise; the genuine predictive signals lie in lead_time, special_requests, and market_segment.

3. Actionable Segmentation and Dependency

The Elbow Method produced no clear inflection point across K=2 to K=10, making it inconclusive as a standalone criterion. Silhouette analysis peaked at K=2 (0.9627), but this near-perfect score was diagnosed as an artefact of highly correlated binary features producing trivially broad segments with no operational utility. K=4 was ultimately selected because it represents the minimum granularity at which meaningfully differentiated customer profiles emerge—a decision grounded in business interpretability rather than metric optimization alone.

The resulting segmentation delivers direct operational value: Cluster 2's 0% cancellation rate, driven entirely by parking requirements, offers an immediately actionable lever—prioritizing parking-inclusive packages during peak seasons locks in guaranteed occupancy. Cluster 1's 72,568 habitual cancellers represent the largest aggregate revenue risk (~30,479 lost room-nights), justifying targeted deposit policies for this segment without penalizing lower-risk guests.

Conclusions

This project successfully applied Classification, Clustering, and Probabilistic Dependency Analysis to a large-scale hotel booking dataset to systematically decode reservation uncertainty. Moving beyond simple descriptive statistics, the analysis constructed a robust predictive framework for identifying cancellation risks.

Ultimately, this study demonstrates that data analytics can transition hotel inventory management from a reactive guessing game to a proactive, precision-driven strategy. By integrating these predictive models and segment-specific rules into their reservation systems, hospitality operators can confidently optimize overbooking limits, dynamically adjust deposit policies, and minimize revenue leakage caused by last-minute cancellations.

Appendix: AI Tool Usage Documentation

Tool used: Claude Sonnet 4.6

Section assisted: Coding — Select final k

Prompt provided:

> "I obtained different conclusions using the elbow method and the profile coefficient method. The elbow method did not have a clear inflection point, and the profile coefficient was highest at K=2 at 0.9627, but only 0.1720 at K=4. Which K should I choose?"

AI response summary: Claude explained that when the two methods conflict, neither result should be accepted blindly. The near-perfect Silhouette score at K=2 is likely an artefact caused by highly correlated binary features that create artificially separable clusters rather than meaningful customer segments. The Elbow Method's lack of inflection suggests no single mathematically optimal K exists. In such cases, business interpretability should serve as the final arbitration criterion—K=4 is preferred because it produces the minimum number of segments needed for differentiated operational responses, whereas K=2 yields only trivially broad groupings with no actionable distinction.

How it was used: This explanation informed the justification for selecting K=4 in both the Experimental Results and Discussion sections. The framing of K=2 as an artefact and K=4 as the business-interpretable choice is our own analytical conclusion, supported by Claude's methodological guidance.

Tool used: Claude Sonnet 4.6

Section assisted: Coding — visualization heatmap

Prompt provided:

> “I've decided to use K=4. Could you please use Python to draw a heatmap of the cluster, displaying the mean characteristics of each cluster, including lead time, adr, special requests, prior cancellations, parking, and cancellation rate?”

How it was used: The suggested code was adapted and integrated into our analysis pipeline. The feature list and cluster labels were modified to match our dataset's column names. The final heatmap (Figure 3) was generated by running this code on our processed dataset, and the visual output was incorporated into the Experimental Results section to support the cluster profile descriptions.

Tool used: Claude Sonnet 4.6

Section assisted: Coding — print evaluation metric

Prompt provided:

> print out the final model's performance metrics, including accuracy, precision, recall, F1 score, ROC-AUC, and a complete classification report. The target names should be 'Not Canceled' and 'Canceled'.

AI response summary: Claude provided the following code snippet:

python

```
# Print final model performance metrics (Information Gain, depth=8)

print(f'\nFinal Model Performance (Information Gain, depth=8):')

print(f'Accuracy : {accuracy_score(y_test, y_pred):.4f}')

print(f'Precision: {precision_score(y_test, y_pred):.4f}')

print(f'Recall:    {recall_score(y_test, y_pred):.4f}')

print(f'F1:        {f1_score(y_test, y_pred):.4f}')

print(f'ROC-AUC:   {roc_auc_score(y_test, y_prob):.4f}')

print(classification_report(y_test, y_pred, target_names=["Not Canceled", "Canceled"]))
```

How it was used: The code was directly incorporated into the classification evaluation section of our analysis pipeline. Variable names (y_test, y_pred, y_prob) were already consistent with our existing code structure and required no modification.