For ABC Hotels, the central business problem is the high rate of booking cancellations, which creates uncertainty in occupancy planning, increases the chance of unsold rooms, and ultimately reduces revenue. The hotel has supplied a dataset of 36,238 historical bookings, each annotated with whether the reservation was cancelled. This creates a natural supervised classification problem where the objective is to develop a predictive model that estimates the probability of cancellation for each booking. With such a model, hotel management can identify high-risk bookings in advance and apply targeted interventions to reduce cancellation rates, improving both revenue and efficiency.

The dependent variable for this supervised classification problem is the booking_status column, which records whether a booking was "canceled" or "not_canceled." In the dataset, approximately 32.8% of bookings were cancelled, while 67.2% were not cancelled. This proportion is significant: nearly one in three bookings results in lost revenue unless mitigated. For modeling purposes, the target will be encoded as a binary outcome, with 1 representing a cancellation and 0 representing a non-cancellation. The fact that the data is somewhat imbalanced with two-thirds of cases are not cancelled means that evaluation metrics such as ROC-AUC, precision, recall, and F1-score will be more informative than accuracy alone, since a model that simply predicts "not cancelled" for every case would still appear to perform reasonably under accuracy but would fail to meet the business need.

Before modeling can proceed, careful preprocessing is required. The variable Booking_ID is a unique identifier and will be excluded from the analysis, as it carries no predictive value. The categorical variables type_of_meal_plan, room_type_reserved, and market_segment_type will be transformed into one-hot encoded binary variables so that algorithms can interpret them effectively. The arrival_date column will be converted from string to datetime format and engineered into new features such as month of arrival and season, since cancellations often vary with seasonal demand cycles. Numerical features such as lead_time and avg_price_per_room will be standardized where appropriate, though tree-based models will be robust to scaling. The binary variable repeated_guest and integer counts such as no_of_special_requests, no_of_previous_cancellations, and no_of_previous_bookings_not_canceled will be retained directly. This ensures that nearly all variables contribute to the initial model, with feature selection or importance analysis used later if necessary.

Exploratory data analysis provides insight into which features are most promising for predicting cancellations. The most striking difference is in lead time. For cancelled bookings, the mean lead time is approximately 139 days, with a median of 122. By contrast, for non-cancelled bookings the mean is only 59 days, with a median of 39. This suggests that customers who book far in advance are more likely to cancel, likely because their plans are less certain over long horizons. Another notable feature is average price per room. Cancelled bookings had a higher mean room price ($110.60) compared to non-cancelled bookings ($99.94). This may reflect the fact that

higher rates are often charged during high-demand periods, when customers may make speculative reservations they later cancel. The number of special requests also reveals a strong relationship. Customers who eventually cancelled made an average of only 0.33 special requests, whereas those who honored their bookings averaged 0.76 requests. This suggests that engaged customers who customize their stays are less likely to cancel. Similarly, repeat guests are dramatically less likely to cancel: only 0.1% of cancelled bookings were repeat guests, compared to nearly 4% among non-cancelled bookings. Finally, prior history is meaningful: customers with previous cancellations are naturally more likely to cancel again, while those with prior non-cancelled bookings are more reliable. Together, these exploratory findings provide a strong basis for including all variables in the initial model, as many of them show meaningful separation between cancelled and non-cancelled classes.

The analytic outcomes expected from this project are twofold: predictive and informational. On the predictive side, the model will provide a probability between 0 and 1 that any given booking will be cancelled. This allows hotel management to stratify bookings by risk level and design tailored interventions. On the informational side, the model will highlight which factors contribute most to the prediction. For example, if lead time, lack of special requests, and higher average prices are found to be the strongest predictors, then hotel management can design policies around these insights. Feature importance methods such as SHAP values or permutation importance will be used to quantify and explain the role of each variable, making the results transparent and actionable.

In practice, the model would be integrated into the hotel's booking system so that when a new reservation is recorded, the cancellation probability is immediately generated. High-risk bookings might be flagged for follow-up actions, such as sending reminder emails, offering discounts for non-refundable confirmations, or requesting partial prepayment. Medium-risk bookings could receive softer nudges, while low-risk bookings could be left unaltered, conserving resources. Over time, aggregate predictions could inform strategic decisions, such as how much to overbook to offset expected cancellations. This would allow the hotel to optimize occupancy rates without overshooting and creating customer dissatisfaction. The model could be retrained periodically with new data to adapt to changes in customer behavior, ensuring continued relevance.

The implications of this project extend beyond predictive accuracy. By transforming historical booking records into forward-looking risk assessments, ABC Hotels can align operational practices with data-driven insights. This has tangible financial benefits in reducing lost revenue from cancellations and intangible benefits in strengthening customer engagement by offering proactive, personalized communication. Moreover, the exploratory findings already provide valuable managerial insight: customers who book far in advance and make no special requests are disproportionately likely to cancel, while repeat guests and engaged customers are far more

reliable. These insights alone suggest avenues for loyalty-building initiatives and targeted deposit policies. When formalized into a predictive model, these patterns will be quantifiable and operationally usable, giving ABC Hotels a clear competitive advantage.