Introduction

Hotel booking cancellations are a common but costly issue in the hospitality industry. While each cancellation may be small, it translates to huge financial loss and operational inefficiency when they are added up. To compensate for this uncertainty, most hotels engage in overbooking in anticipation that some of the guests will fail to turn up. This is also done in many other industries such as the airlines industry. Yet this approach can generate its own set of problems, such as guest complaints and logistical strain when too many bookings are taken. These risks suggest the appeal of more predictive, data-driven approaches to forecasting cancellations.

Not only can proper forecasting allow hotel managers to make better-informed decisions but also allow proactive intervention in terms of improving guest experience and reducing monetary losses. Knowing which bookings are most likely to cancel, hotels can dynamically yield prices, send reminders to the target, or offer incentives to encourage the subject to ensure honoring the booking. In addition, an understanding of the trends behind cancellations improves overall operational planning, including staff scheduling, room scheduling, and promotional campaigns. Not only is the desire to predict cancellations, but to understand the dynamics at play in how likely a cancelation may be.

This project has the objective of developing a predictive model to estimate the likelihood of hotel booking cancellations and identifying the key drivers of cancellations. Both model interpretability and accuracy are important. Accuracy will ensure the predictions are close to reality, while interpretability will allow those who run the business to understand the reasoning of each prediction and use the information to improve efficiency. To that end, both transparency of the model and predictive performance are emphasized

Dataset

The data employed in this study is based on a Lisbon hotel, Portugal, and comprises 36,275 booking records collected between 2017 and 2018. The dataset contains 16 variables that comprise both numerical attributes (e.g., lead time, number of special requests, average room price) and categorical attributes (e.g., room type, market segment, meal plan). The variables provide a well-rounded view of the guest booking process, allowing for complete analysis. Such variables as lead time and market segment will tend to describe different guest behaviors, while other attributes like special requests or repeat booking history might indicate guest commitment or purpose.

Objective

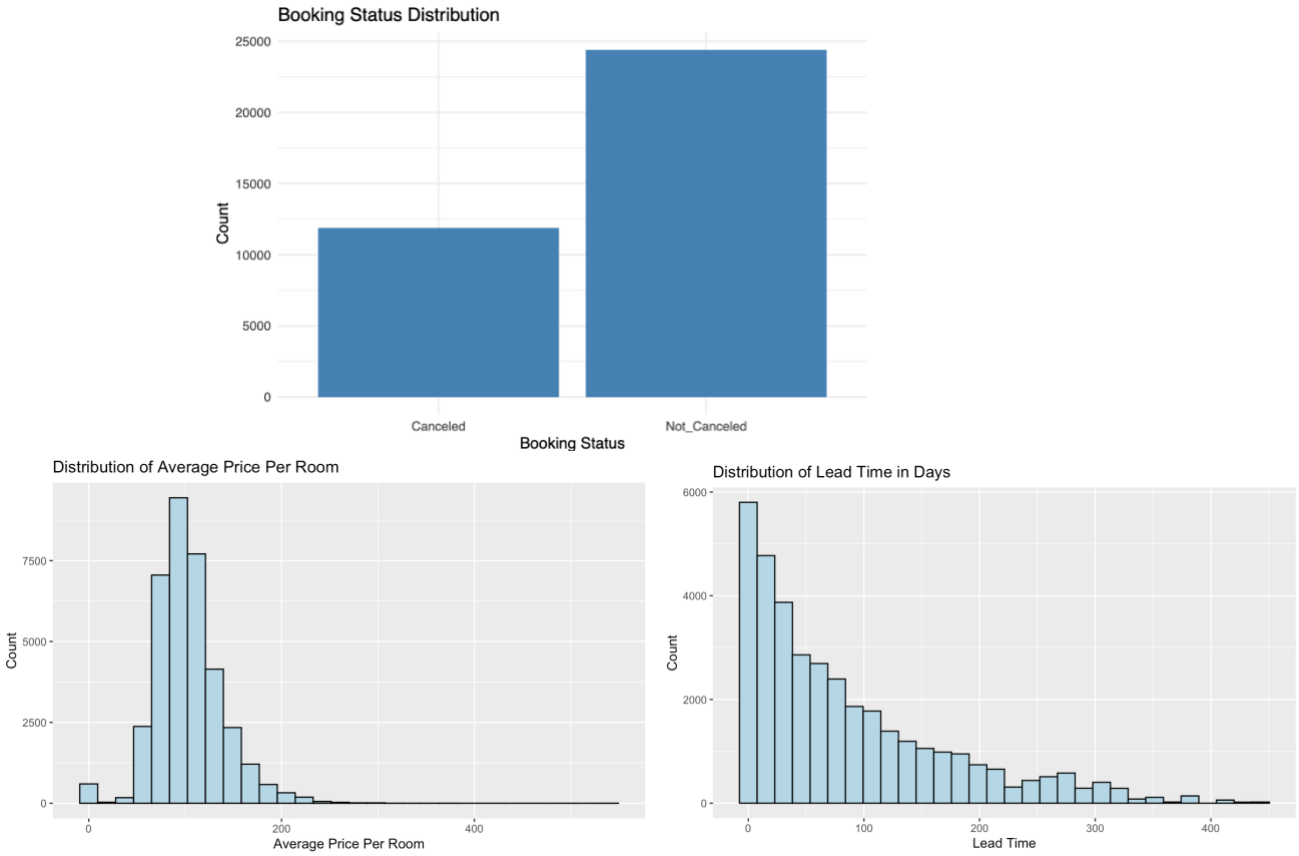
The key questions this report answers are: (1) Can we predict with reliability whether a hotel booking will be canceled? and (2) What are the most informative predictors for estimating cancellation risk? These are the questions that guide model choice and interpretation of results.

Methods

Exploratory Analysis

The data was then checked for missing variables and there were none. Thereafter, exploratory data analysis was done to look at the distributions of the variables. I then did

exploratory data analysis to view the distribution of the variables. For numeric variables, standard summary statistics were computed.

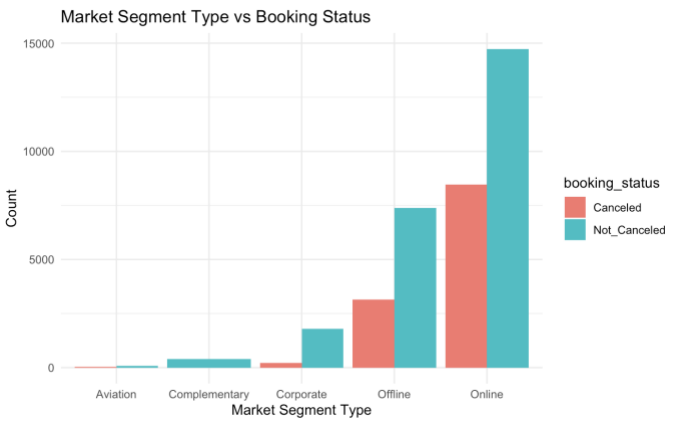


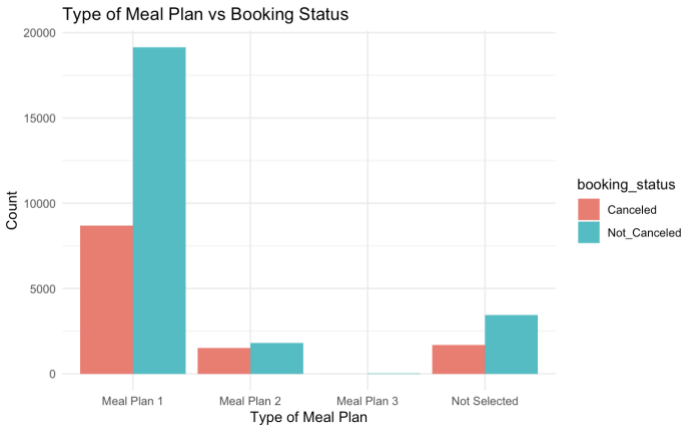
Some of the important exploratory analysis is included. Figure 1 shows a large class imbalance between bookings that were cancelled and not cancelled with about 1/3 of bookings be cancelled. From the histograms from both average price per room and lead time there appears to be skewness and outliers.

I then created some visualizations for some of the possible important predictors and their respective distributions against the response. From the box and whisker plot in figure # there appeared to be a strong pattern indicating that longer lead times are associated with higher rates of cancelation. For market segment type we can see that most of the bookings are made online, then offline, then corporate. We can also see that the proportion of bookings for each segment

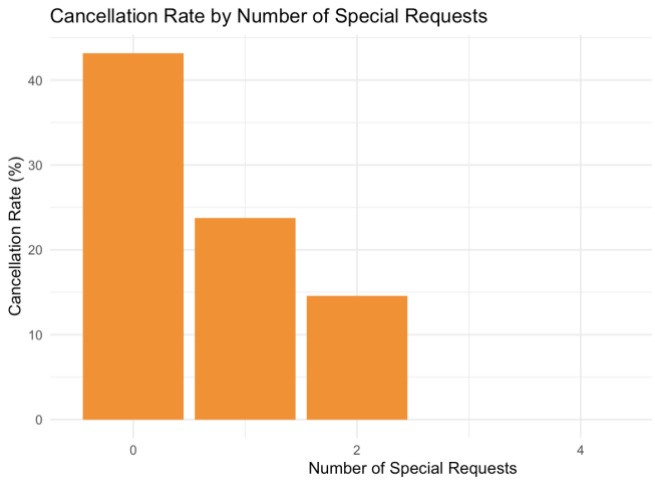
differs with online bookings having the highest percentage of bookings cancelled. We can see from figure # that the distribution differs between meal plans as well with meal plan 1 having the largest difference in proportions between canceled and not cancelled.

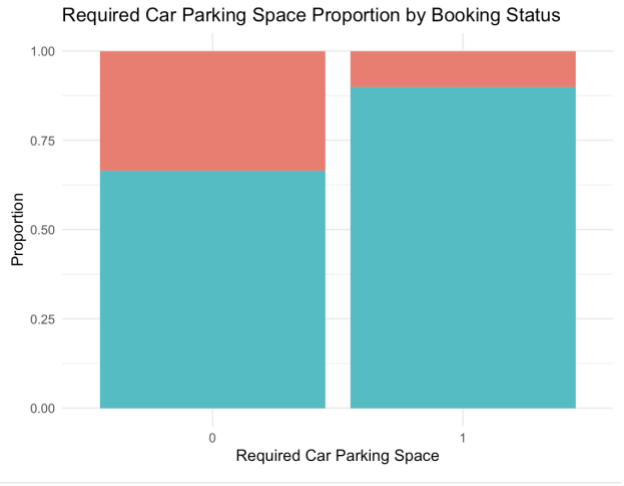
Overall, these in-depth summaries provided a solid foundation for the subsequent analysis by clearly establishing both the numeric distributions and the categorical separations within the data. I then created visualizations of the various variables to gain a better understanding of their distributions. In particular, lead time and average price per room were





both heavily skewed. For this reason we will consider a log transformation. The data exploration also revealed that there were large class differences for many of the categorical variables. This will be something to keep in mind when building our model.





Plotting the cancelation rates against the number of special requests indicated 0 special requests was associated with a much higher cancelation rate than 1 or 2. While for car parking spots required there appeared to be an association that a car parking spot is associated with a lower proportion of cancelations. These patterns in the data were useful for gauging the inference of the models.

The dataset was already fairly clean to begin with so only some minor data cleaning was required. Categorical variables such as market type and room type were converted to factors. The date variable was turned into a month variable with 12 factors. This was done because there did not appear to be a pattern between the years or the days of the months. The response variable, booking status, was turned into a binary variable with a value of 1 corresponding to a cancellation and a value of 0 corresponding to a booking not being canceled. In addition, the data was standardized to increase inference of the regression coefficients.

Models

When selecting a model to build, robust and inferable models were first preferred. This was to meet the first objective question. Regression was chosen for its inference ability and predictive power. There were many predictor variables that almost all seemed they could contribute to the model from the exploratory analysis. For this reason, LASSO regression was chosen for its built in feature selection. To help with inference of the variables, the predictors were standardized. This will allow the size of the coefficient to be a good indicator of how important a variable is as a predictor in the model. A LASSO model was fit using 10 fold cross validation to choose the best lambda value. However, the initial lambda value did not shrink any of the coefficients to 0. For this reason lamda.1se was chosen and compared to lambda. Lambda.1se is a value that is one standard errir higher than the lambda value from cross validation. This new value of lambda increases the penalty for the size of the coefficients in the LASSO model. This creates a model that is usually simpler. This did infact happen as many coefficients were shrunk to 0 and were thus not included in the model. It is important to note that the way LASSO regression handles factor variables is that each factor is assigned its own coefficient. The model used a default threshold of 0.5.

LASSO Model Comparison

Metric	LASSO (λmin)	LASSO (λmin.1se)
Lambda	4.55×10^{-5}	1.30×10^{-3}
Accuracy	80.51%	80.51%
Sensitivity	88.03%	88.44%
Specificity	65.08%	64.24%
Positive Predictive Value	83.80%	83.54%
Negative Predictive Value	72.60%	73.03%
Balanced Accuracy	76.55%	76.34%

The Lasso model provided predictive power, but not enough. I suspected that the Lasso model could not pickup some of the non linear patterns in the dataset. To address this, I selected a Random Forest model for comparison. Random Forest is an ensemble learning method that combines multiple decision trees and is well-suited to handling non-linearity and interaction effects between variables. It also provides a built-in measure of feature importance using the Mean Decrease in Gini, which reflects how much each variable contributes to improving node purity across all trees in the model. This offered an alternative way to rank features, independent of coefficient values.

The model was tuned to identify the ideal value of mtry, the number of features to randomly select at each node. Various values were attempted, and mtry = 4 was the ideal trade-off between accuracy and F1 measure. The model was tested on the test split of data set consisting of 20%. As with the LASSO model, the default classification threshold for cancellation prediction was 0.5, and therefore any booking with a probability of cancellation of more than 50% was tagged as canceled. The threshold was then adjusted to optimize performance, most notably in getting a good balance between false positives and false negatives.

Results

Accuracy

Following training for both LASSO and Random Forest on an 80–20 train–test split, I evaluated them using various measures: accuracy, F1 score, sensitivity, specificity, and balanced accuracy. The Lasso model is lambda1.se = 1.3×10^{-3} gave a test accuracy of 79.74% and balanced accuracy of 75.40%. The sensitivity was 87.99% and the specificity was 63.81%.

Confusion Matrix Metrics: LASSO

	Metric	Value
Accuracy	Accuracy	0.7983
Sensitivity	Sensitivity	0.8805
Specificity	Specificity	0.6298
Pos Pred Value	Pos Pred Value	0.8300
Neg Pred Value	Neg Pred Value	0.7197
Balanced Accuracy	Balanced Accuracy	0.7551

Random Forest Performance Metrics

	Metric	Value
Accuracy	Accuracy	0.9002
Sensitivity	Sensitivity	0.9469
Specificity	Specificity	0.8044
Pos Pred Value	Pos Pred Value	0.9085
Neg Pred Value	Neg Pred Value	0.8807
Balanced Accuracy	Balanced Accuracy	0.8756

The Random Forest model had much more predictive power. The model had an accuracy of 90.32% and balanced accuracy of 87.91%. The sensitivity and specificity were 94.87% and 80.94%. Both models had large imbalances in the sensitivity and specificity likely due to the large imbalance in cancelled vs non cancelled bookings

Threshold Tuning

I also experimented with changing classification thresholds to optimize performance instead of using the default value of 0.5. Lowering the threshold enabled the models to spot more potential cancellations (fewer false negatives) at the cost of more false positives. Where revenue loss due to vacant rooms is a major concern, this trade-off might be a good suggestion. By contrast, low-resource hotels that are unable to find every "likely to cancel" reservation might prefer to employ a higher threshold, avoiding unnecessary interventions at the expense of missing some genuine cancellations. Both the Random Forest and Lasso model's thresholds were tuned using both the F1 score and Youden's index. Youden's index is just a measure of balancing the sensitivity and specificity.

LASSO (λmin.1se) - Performance Metrics

Metric	Value
Accuracy	0.7847
Sensitivity	0.7749
Specificity	0.7895
Pos Pred Value	0.6420
Neg Pred Value	0.8780
F1 Score	0.7022
Balanced Accuracy	0.7822

Random Forest - Performance Metrics

Metric	Value
Accuracy	0.9023
Sensitivity	0.8271
Specificity	0.9389
Pos Pred Value	0.8684
Neg Pred Value	0.9177
F1 Score	0.8472
Balanced Accuracy	0.8830

For the F1 tuning, the best cutoff for the LASSO model was found to be at 0.38. The overall accuracy of 78.47% and balanced accuracy of 78.22% were similar to the base threshold cutoff. However, the sensitivity (0.7749) and specificity (0.7895) became much closer to each other. For the Random Forest model was 0.42. The random forest model with the new threshold had an accuracy of 90.23% with sensitivity 82.71% and specificity 93.89%. For both models the overall accuracy stayed roughly ethe same. For the LASSO model the sensitivity and specificity balanced out but this was not the case for the random forest model.

Random Forest - Performance Metrics

Metric	Value
Accuracy	0.8944
Sensitivity	0.8616
Specificity	0.9104
Pos Pred Value	0.8241
Neg Pred Value	0.9310
F1 Score	0.8425
Balanced Accuracy	0.8860

LASSO (λmin.1se) - Performance Metrics

Metric	Value
Accuracy	0.7847
Sensitivity	0.7749
Specificity	0.7895
Pos Pred Value	0.6420
Neg Pred Value	0.8780
F1 Score	0.7022
Balanced Accuracy	0.7822

Thus they were also tuned using Youden's index. Using Youden's index to tune, the best threshold for lasso was 0.37 and for Random Forest it was 0.38. The Lasso model had accuracy 78.47% with sensitivity of 77.49% and specificity of 78.95%. For the Random forest model with cutoff threshold of 0.38 the overall accuracy was 89.44% with sensitivity 86.16% and specificity of 91.04%. The threshold tuning was able to balance out the false negatives and false positives while keeping most of the overall predictive power.

Overall, all these performance metrics indicate the robustness of Random Forest in effectively classifying bookings, minimizing cancellations missed, and avoiding unnecessary follow-ups. LASSO, despite poorer accuracy and balanced accuracy, even so, is useful in identifying which predictors have the most significant impacts on cancellation risk, giving insight into the relationships between lead time, room price, and client behavior and outcomes.

This project compared the performance of two models: LASSO logistic regression and Random Forest. Both were trained on 80% of the dataset and tested on the remaining 20%. While Random Forest had stronger predictive performance, the LASSO model offered clearer insights into which features most affected cancellation risk.

Lasso Feature Importance

LASSO (λmin.1se) Coefficients

Variable	Coefficient	Variable	Coefficient
log_avg_price_per_room	1.7052	room_type_reservedRoom_Type 6	-0.1600
market_segment_typeOffline	-1.5857	arrival_month10	0.1122
no_of_special_requests	-1.3949	arrival_month6	0.0989
required_car_parking_space	-1.2854	no_of_children	0.0693
arrival_month12	-1.0269	no_of_weekend_nights	0.0679
log_lead_time	1.0004	arrival_month5	-0.0670
market_segment_typeAviation	0.9548	no_of_previous_cancellations	0.0635
repeated_guest	-0.8569	arrival_month8	0.0533
market_segment_typeCorporate	-0.8373	no_of_adults	0.0455
arrival_month2	0.7035	no_of_week_nights	0.0111
arrival_month11	0.3432	type_of_meal_planMeal Plan 3	0.0000
room_type_reservedRoom_Type 7	-0.3409	room_type_reservedRoom_Type 3	0.0000
type_of_meal_planNot Selected	0.2502	arrival_month4	0.0000
room_type_reservedRoom_Type 4	-0.2399	arrival_month7	0.0000
type_of_meal_planMeal Plan 2	0.2264	arrival_month9	0.0000
arrival_month3	0.2146	market_segment_typeComplementary	0.0000
room_type_reservedRoom_Type 5	-0.2145	no_of_previous_bookings_not_canceled	0.0000
room_type_reservedRoom_Type 2	-0.1768	NA	NA

The data was standardized for the LASSO model. This means that we can interpret the magnitude of the coefficients as how important they are to the model. A positive coefficient indicates that that predictor increases the chance of a cancellation while a negative predictor decreases the chance of a cancellation. The LASSO coefficients can be seen in the table above. It is important to note that the LASSO model assigns each level of each categorical variable its own coefficient allowing us to see how it affects the model.

Random Forest Feature Importance

Variable	MeanDecreaseGini
log_lead_time	3910.362015
log_avg_price_per_room	1875.055989
no_of_special_requests	1195.862982
arrival_month	1154.430567
market_segment_type	739.233646
no_of_week_nights	557.446215
no_of_weekend_nights	422.980713
no_of_adults	281.926857
type_of_meal_plan	239.726613
room_type_reserved	205.433089
no_of_children	90.181161
required_car_parking_space	79.744477
repeated_guest	18.862150
no_of_previous_bookings_not_canceled	17.484149
no_of_previous_cancellations	3.743464

The random forest importance table can be seen above. The importance comes from the average reduction in impurity (Gini impurity or variance reduction) achieved by splits using each feature across all trees in the forest. Unlike LASSO, which transforms every factor level into a separate coefficient, Random Forest tends to consider an entire categorical variable (such as market_segment_type or type_of_meal_plan) holistically, splitting the data in various ways across all trees. This can capture nonlinear interactions or subtle differences among multiple factor levels simultaneously, rather than assigning one coefficient to each level independently.

Comparison of LASSO vs. Random Forest Feature Importance

Overall, both models agree that lead time, market segment type, average price, and special requests are top drivers of cancellations. However, their approaches differ in two key ways. First, LASSO provides coefficients, each level of a categorical variable shows up as a

positive or negative effect on cancellation odds, while Random Forest assigns a single score to the entire factor, measuring how much it reduces impurity overall. Second, LASSO's interpretability stems from explicit positive or negative coefficients, whereas Random Forest excels at finding interactions and nonlinearities, but provides fewer insights to how the features affect the model. Despite these differences, the consistency in identifying log lead time and log average price as leading factors, in addition to market segment type reinforces their significance across modeling techniques.

Discussion

Model Accuracy

Through both the random forest model and the LASSO model I was able to answer both of my objective questions. The random forest model was capable of making accurate predictions of whether a booking would be cancelled or not. The model had an overall accuracy of 90.31% with the base threshold of 0.5 and an accuracy of 89.44% with the tuned threshed of 0.38. Raising or lowering the threshold allows the sensitivity and specificity of the model to change. Lowering the threshold increases the sensitivity and decreases the specificity. Raising the threshold does the opposite. By tuning using Youden's index I was able to find a threshold that balanced the sensitivity to 86.16% and the sensitivity to 91.04%.

In a business setting, this threshold could be adjusted to meet business needs. For example, if a hotel was using the model and they wanted to minimize false negatives (bookings that are predicted to be not cancelled but are cancelled) then lowering the threshold to increase sensitivity would be appropriate. This would make sense in a setting where a hotel would want to minimize the amount of vacant rooms to increase revenue. However, if a hotel wanted to optimize for specificity then raising the threshold would be appropriate. This may be for many

reasons, such as the hotel not wanting to overbook rooms to avoid upsetting customers. In the end the best threshold will come to business needs of minimizing false positives or false negatives.

The LASSO model was not as accurate at making predictions with an accuracy of BLANK. However, the model was able to fulfill our second objective question: what factors are important in predicting whether a booking will be cancelled. We can get this answer from the coefficients of the model. LASSO adds an additional penalty in regression that minimizes the size of the coefficients. Because of this and the fact that the data has been standardized, we can interpret the magnitude of the coefficient as how important that predictor is to the model. A positive coefficient means that variable increases the chance of cancellation while a negative variable means it decreases the chance of cancellation.

Interpretation of Variables

From table BLANK we can see that the variable with the largest coefficient is the average price per room (1.7052). This shows that the more expensive a room is, the higher the likelihood of a cancellation is. The next highest positive coefficient was lead time (1.0004). This confirms the pattern from the exploratory data analysis that longer lead times are associated with higher rates of cancellations. For the month of the booking, the baseline is month 1 or January. The table shows that all the other months except May (-0.0670) and December (-1.0269) had positive coefficients, which can be interpreted as those months have a higher likelihood of being cancelled compared to bookings in January. The LASSO model breaks down the categorical variables into their different factors. This is an added benefit of the LASSO model. Other variables had positive coefficients that were smaller in size, such as number of children (0.0693),

number of previous cancellations (0.0635), number of week nights (0.0111), and number of adults (0.0455).

Predictors with negative coefficients decrease the chance of a cancellation. The numeric predictor with the largest negative coefficient was number of special requests (-1.3949). This was more surprising than some of the other results but it does intuitively make sense. The more special requests that are made by a guest, the lower the likelihood of cancellation is. Many factor or binary predictors were important in lowering the chance of cancellation. The way the booking was made (market segment type) seems to be most important in predicting a cancellation if the booking was made offline (-1.5857). The baseline for this variable is online bookings. Corporate bookings (-0.8373) were also important in reducing the chance of cancellation when compared to online bookings. Aviation (0.9548) and complementary (0.0000) bookings were shrunk to small or zero values in the model and thus are not significant predictors. This is likely due to how few of them were in the dataset. Arrival months May (-0.0670) and December (-1.0269) were both associated with lower chances of cancellations when compared to January. This could be because May is typically the start of warm weather, so people are less likely to cancel their trips. Similarly, December has many holidays and this could mean people are less likely to cancel their trips. Interestingly, if a guest booked a car parking spot (-1.2854), they also seemed less likely to cancel their bookings. Lastly, repeated guests (-0.8569) seemed to play a moderate role in decreasing the chance of a booking.

The categorical predictors room type and meal plan had minor roles in the model. Unfortunately, however, I could not find further information corresponding to the different meal plans and room types. They are still included in the model because we can still identify patterns in the data without that information.

Takeaways

Hotels could use this model or train one on their own data in order to minimize revenue loss due to cancellations. By forecasting cancellation risk, hotels can actively manage these risks through a number of targeted interventions. Specifically, hotels can dynamically adjust room prices based on the forecasted risk, offer personalized incentives to guests who are predicted to cancel, and send timely reminders or messages to reassure guests and encourage booking fulfillment. They could also strategically overbook their room to decrease lost revenue if they have an accurate prediction of how many bookings will be cancelled.

The average price per room, lead time, number of special requests, whether the guest reserved a parking space, and the method of booking (market segment type) were the most significant predictors of cancellation, according to the model. Longer lead times and higher room rates were linked to higher cancellation rates, indicating that travelers might cancel more costly or far-in-advance reservations. However, those who reserved a parking spot or made special requests were less likely to cancel, which may have indicated a stronger commitment to the stay. Furthermore, reservations made offline or via corporate channels had a lower cancellation rate than reservations made online. In addition to increasing the precision of their predictive models, these insights can assist hotels in better understanding the behavior of their guests and customizing interventions (e.g., providing exclusive benefits to lower-risk guests while targeting higher-risk guests with reminders, flexible options, or retention offers).

The threshold-adjusting flexibility of predictive modeling allows hotels to balance the cost of false positives (wasted incentives or over-discounts) against the cost of false negatives (revenue lost on unscheduled cancellations). It has the ability to make the model more sensitive,

picking up more potential cancellations but with decreased specificity, such that threshold adjustment becomes a revenue management strategy imperative.

Limitations and Future Work

Even though the predictive model has promising outcomes, several limitations must be considered. First, the analysis performed was conducted based on information for a single hotel alone, and hence these findings might not hold for other hotels or geographical regions. Future research should utilize data from more than one property or chain to make the findings stronger and more applicable. Additionally, the present study does not account for temporal dynamics or exogenous shocks, such as impacts of the COVID-19 pandemic on cancellation tendencies. Adding temporality to future modeling efforts could greatly increase predictive utility and relevance to evolving market conditions.

Lastly, while the LASSO model used in this study was effective in identifying the key predictors, exploring more advanced techniques such as ensemble methods or gradient boosting models such as XGBoost could potentially increase predictive accuracy further. Utilizing these models could reveal more insights and added precision, allowing hotels to manage cancellations more effectively and maximize revenue.