NOVA IMS
Information Management School

Data-Driven Decision Making

Lorem Ipsum Hotel Group: Hotel Cancellation Prediction

Gabriel Calero - 20240357
Tingting Mo - 20241230
Dylan Fourie - 20211651
Gonçalo Duarte - 20241313

EXECUTIVE SUMMARY

To help reduce losses from booking cancellations, we developed a prediction model based on 20,000 past reservations. This model identifies which bookings are most likely to be canceled, allowing the hotel to take early action.

After testing different methods, the best-performing model reached an accuracy of 86%, meaning it can reliably flag high-risk bookings. We also calculated when it makes the most financial sense to act. The analysis showed that when a booking has a 60% or higher chance of being canceled, it's best to treat it as high risk.

By using this model, the hotel can:

- Spot likely cancellations before they happen;
- Send reminders or special offers to reduce no-shows;
- Adjust overbooking strategies to avoid empty rooms;
- Improve revenue and planning.

This tool can be added to the hotel's reservation system and updated regularly to stay effective. All customer data will be handled securely and responsibly.

BUSINESS AND DATA UNDERSTANDING & DATA ENGINEERING

During the modeling phase, several different approaches to feature selection were explored to assess their impact on the performance of the hotel booking cancellation prediction model. Specifically, three main strategies for selecting variables were tested, each with its own motivations and process.

- **Filter method:** We used statistics to pick variables most related to cancellations. For example, we used the Spearman correlation (which measures how strongly two things are related) to identify useful numerical features like Lead Time, Previous Cancellations, Booking Changes, ADR (average daily rate), and Special Requests. For categorical data (like Country or Customer Type), we selected those with enough variety. We excluded variables where one category made up more than 80% of the data to ensure the model learns useful patterns.

- **All Variables:** As a baseline, we also tested a model using all available variables without filtering. This helped us compare whether more data always leads to better performance.

- **Wrapper Method:** In this approach, the genetic algorithm tries different combinations of variables, and the Decision Tree evaluates how well each combination performs. This more advanced method automatically tests many combinations of variables to find the best mix. While it takes longer to run, it can discover hidden relationships between variables that may not be obvious.

Before training the models, we performed data processing:

- Missing values: Filled using the most common value for categories or the median for numerical values.
- Outliers: Adjusted extremely high or low values using winsorization to reduce their impact.
- Normalization: Scaled all numerical values to a standard range (Min-Max) to ensure balanced input for the model.

This process helped us prepare a clean, balanced dataset that improves the model's ability to make reliable predictions.

MODEL ENGINEERING AND EVALUATION

During the model development process, our main goal was to find out which algorithm could best predict whether a hotel booking would be canceled. We also compared how different ways of choosing input variables (features) would impact the results.

We tested three feature selection strategies: Filter Method, All Variables and Wrapper Method.

For each feature set, we tested four commonly used models: Logistic Regression, Decision Tree, Random Forest, and Gradient Boosted Learner. These models range from easy-to-understand to more advanced ones that can capture complex patterns.

We used 10-fold cross-validation—a process that splits the data into 10 parts and tests the model multiple times—to ensure the performance of model was consistent and not due to chance.

To evaluate the models, we used two key metrics:

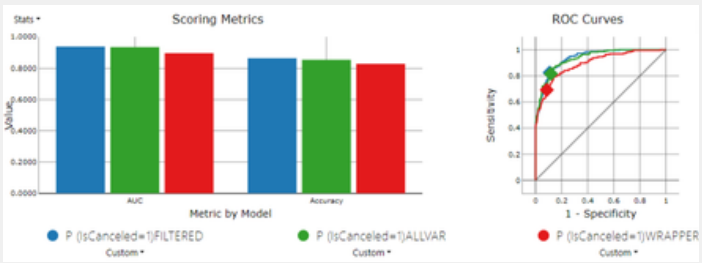
- Accuracy: The percentage of correct predictions.
- AUC (Area Under the ROC Curve): A score between 0 and 1 that shows how well the model distinguishes between cancellations and non-cancellations. A higher AUC means better predictive power. This is especially important in real business situations, where different types of mistakes (like predicting a cancellation that doesn't happen) have different costs.

After testing, the Gradient Boosted Learner gave the best results for both Accuracy and AUC in all three scenarios, meaning the model was both accurate and effective at identifying at-risk bookings. This made it the best choice overall. The results for Gradient Boosted Learner were:

- **Filter method**: Accuracy = 0.8470, AUC = 0.92
- **All variables**: Accuracy = 0.8510, AUC = 0.9315
- **Wrapper method:** Accuracy = 0.82, AUC = 0.8944

To make sure the model wasn't overfitting (performing well only on known data), we also tested it on a separate validation set the model hadn't seen before. Results were consistent, confirming that the model can generalize well to new data.

In the end, the combination of the Filter Method and Gradient Boosted Learner was selected as the final solution. This model helps the hotel identify potential cancellations in advance, allowing the team to take proactive steps—such as overbooking adjustments or offering incentives to reduce last-minute losses and improve revenue management.



DEPLOYMENT

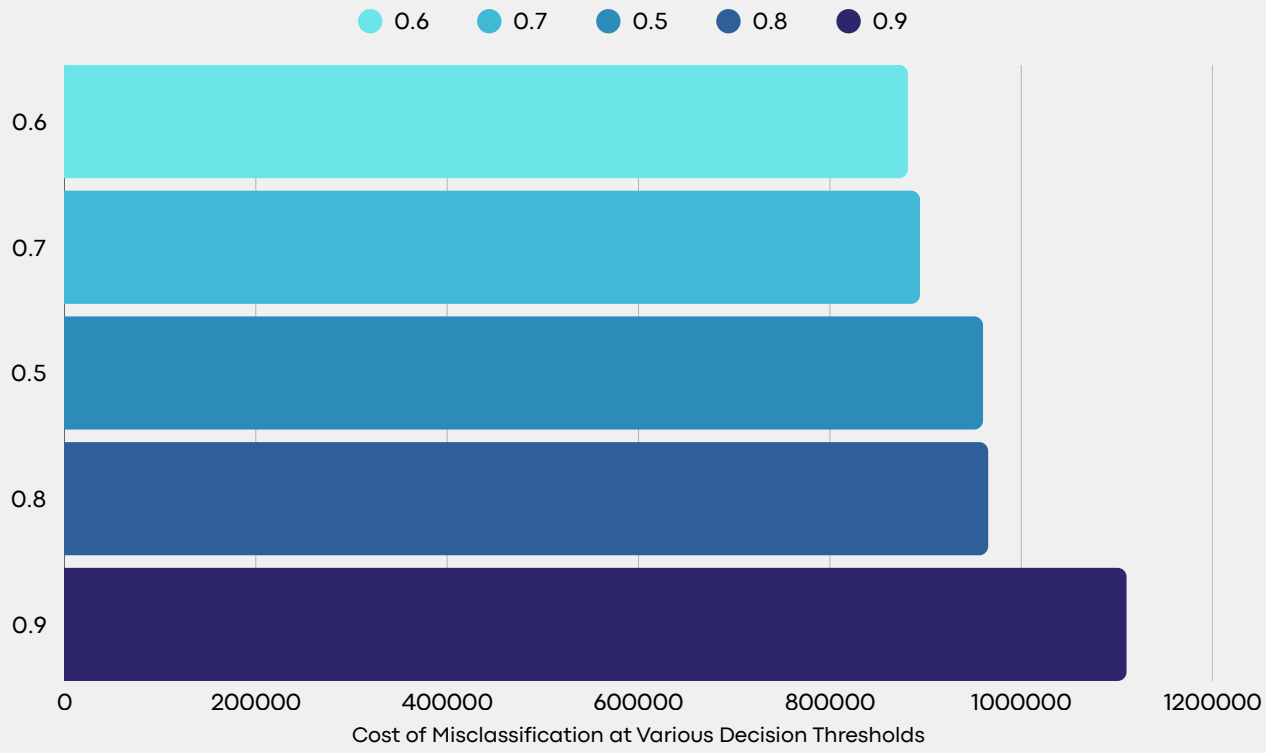
To reduce the impact of cancellations, we recommend integrating the prediction model into the hotel's reservation system. This would allow the system to automatically estimate the cancellation risk for each booking.

For bookings with a high risk of cancellation, hotel staff can take early actions—such as sending reminders or special offers to encourage guests to keep their bookings—or adjust overbooking strategies to reduce empty rooms.

A key part of this strategy is understanding the cost of prediction errors. Based on industry research:

- Predicting a cancellation that doesn't happen (false positive) costs about $633.27
- Missing a real cancellation (false negative) costs $316.63

We tested different thresholds for the model (the probability at which a booking is flagged as "likely to cancel") and found that a threshold of 0.6 leads to the lowest overall cost ($881,520.19). So, this is our recommended setting.



If our hotel decides to deploy the model, it can expect a reduction in losses due to last-minute cancellations and an increase in occupancy rates, resulting in more stable and predictable income. However, it is important to regularly review the model's performance and update it as customer behavior evolves. Actions based on predictions, such as sending offers or reminders, should be carried out thoughtfully to avoid overwhelming guests. Additionally, the company must always protect customer data and comply with privacy regulations.

In summary, deploying the prediction model with a cost-based threshold selection can bring significant financial and operational benefits, as long as the model's performance is monitored, the approach is updated when necessary, and guest experience remains a priority.

MONITORING AND MAINTENANCE

To keep the cancellation prediction model accurate and useful, it's important to monitor and update it regularly. Customer behavior and booking trends can change over time, so the model must stay up to date.

We recommend checking the model's performance every few months by comparing predictions with actual cancellations. If the accuracy or AUC (a measure of how well the model separates likely cancellations from confirmed bookings) drops, the model should be retrained using the latest data.

The current model uses basic booking data, such as how far in advance the booking was made and the number of special requests. However, adding new types of data could improve its predictions. For example, knowing if a guest has canceled before, how they respond to reminders, or if they booked during a special event might help better estimate risk.

While our current model (Gradient Boosted Learner) performs well, trying other models in the future—like ones that look at booking trends over time—might bring new insights and improve results further.

In short, we suggest:

- Monitoring the model's accuracy regularly
- Updating it with new data as needed
- Collecting additional useful information
- Exploring other types of models if needed

This ensures the hotel keeps making informed, smart decisions to reduce cancellations and improve guest satisfaction.

CONCLUSION

To keep the cancellation prediction model accurate and useful over time, it's important to monitor its performance regularly. Customer behavior and booking patterns can change due to seasons, promotions, or market shifts. We recommend reviewing the model every 3–6 months by comparing its predictions to actual cancellations. If accuracy or AUC (a measure of prediction quality) drops, the model should be retrained using recent data.

Now, the model uses booking details such as lead time, special requests, and customer type. However, the hotel could improve the model by collecting new data. Useful additions might include a guest's past booking or cancellation history, how they respond to emails, or even local event information that could affect travel plans.

Although our current model (Gradient Boosted Learner) performs well, testing other types of models in the future may provide even better results, especially as the business environment changes. For example, time-series models could help capture trends across seasons or years.

In short, we suggest a long-term strategy of monitoring model accuracy, updating it with fresh data, exploring new types of customer information, and testing other models when needed. This will help the hotel make smarter decisions, reduce cancellations, and improve revenue stability.