Understanding and Reducing Hotel Booking Cancellations

A Data-Driven Approach

Capstone Project

Model Governance

Business Analytics and Insights

Maria Alejandra Barrios Meneses

Student Number: 301314502

August, 2024

Validation, Monitoring and Governance

1.0. Introduction

As the hospitality industry continues to evolve, the importance of managing booking cancellations has become increasingly significant. Predictive models that forecast booking cancellations are crucial for maintaining profitability and operational efficiency. Over time, these models may exhibit different performance levels due to shifts in booking behaviors, changes in customer preferences, or variations in market conditions. To ensure continued accuracy and effectiveness, robust validation and governance practices are essential. This section outlines the steps taken to monitor model performance, manage potential drift, and validate the stability of the model in production.

2.0. Variable Level Monitoring

In predictive modeling, especially within the hospitality industry, variable level monitoring is a critical component of ensuring that the models remain accurate over time. This section details the methodologies applied to monitor variables in the hotel booking cancellation model, including descriptive statistics, acceptable ranges, handling outliers, and strategies for managing missing values. Additionally, variable drift will be explored to monitor and manage the model's performance over time.

2.1. Variable Level Statistics

During the model development phase, it is essential baseline the statistics for each feature. These statistics include the mean, median, min, max, and standard deviation, which provide a clear understanding of the data distribution.

For numerical features, descriptive statistics such as mean, median, and standard deviation will be compared between the training dataset and real-time data. The next table showcase the descriptive statistics for numerical variables for the training dataset:

Feature	Mean	Standard Deviation	Min	25%	50%	75%	Max
lead_time	101.41	93.5	0	23	78	159	373
stays_in_weekeend_nights	0.91	0.94	1	1	1	2	5
stays_in_week_nights	2.30	1.38	1	1	2	3	6
adults	1.87	0.59	1	2	2	2	55
children	0.10	0.39	0	0	0	0	10
babies	0.007	0.09	0	0	0	0	9
previous_cancellations	0.11	0.96	0	0	0	0	26
days_in_waiting_list	2.57	18.38	0	0	0	0	391
average_daily_rate	97.74	39.4	0	70	95	120	211
required_car_parking_spaces	0.049	0.22	0	0	0	0	8
total_of_special_requests	0.52	0.77	0	0	0	1	5

Any significant deviations may indicate a change in the underlying data distribution; values fall outside the expected range (min and max) will be identified as outliers. Outliers can significantly skew model predictions if not appropriately managed. For further deployments, outliers' values will be replaced with the median value of each numerical feature. By implementing this technique, the model will be better equipped to handle real-world variability maintaining accuracy.

Missing data is a common challenge in predictive modeling. In the context of hotel booking prediction, different strategies can be employed depending on the nature of the missing data. For numerical variables, missing values will be imputed using the median of each feature. For more complex variables, particularly ***average_daily_rate***, missing values will be imputed using a K-Nearest Neighbors (KNN) approach, leveraging the similarity between records to estimate the most likely value.

In the case where categorical data was missing, missing values will be imputed depending on the proportion out of the total of records. When missing values are less than 10% of the population, they will be imputed using a K-Nearest Neighbors. Otherwise, an additional binary flag will be created to indicate missing values and validate if the missing values are relevant for the model.

2.2. Variable Drift Monitoring Tolerance

The variables in this project are selected based on their predictive power for forecasting hotel booking cancellations. However, factors such as seasonality, market shifts, or changes in customer behavior can lead to drift in these variables over time. Tuycheiev (2024) defines data drift when there is a significant shift in the distribution between the training and the data in production (para. 11). Monitoring for variable drift is therefore essential to maintain model accuracy.

Different variables have varying levels of importance in predicting cancellations. For the more important variables are assigned lower drift tolerance levels of 5% of the mean, while less critical variables the tolerance levels are settled to 10% of the mean.

3.0. Model Monitoring, Health & Stability

To ensure the long-term success of the predictive model, continuous monitoring of its health and stability is required. This involves tracking key performance metrics and establishing thresholds for action.

3.1. Initial Model Fit Statistics

Upon deployment, the model's initial statistics are recorded. These metrics provide a baseline for further comparisons. The initial model got an accuracy of 83.20% and a ROC AUC

score of 0.9189. These statistics indicate the model's ability to correctly predict cancellation and distinguish between canceled and non-canceled bookings, respectively.

Metric	Initial Model	Minimum Threshold
Accuracy	0.8320	0.7488
ROC AUC	0.9189	0.8270

If accuracy falls below 74.88%, a review of recent booking data and model parameters needs to be initiated. Monitoring accuracy ensures that the model continues to deliver reliable predictions. A drop below 0.8270 for ROC AUC will require a deeper analysis of variable importance and potential drift. By tracking the ROC AUC score, the model's discriminatory power is maintained.

4.0. Risk Tiering

Risk tiering is a systematic approach to managing model drift and degradation. By categorizing the severity of drift, appropriate actions can be taken to maintain the model's performance.

4.1. Low Risk

In the case of drift less than 2%, the model is likely still performing with acceptable limits. Continued monitoring is recommended to ensure that the drift does not increase, but no immediate changes are necessary.

4.2. Moderate Risk

If drift reaches 2-5%, it suggests that some variables may be shifting in a way that could degrade model accuracy. Reporting the drift to stakeholders and considering minor adjustments or recalibration is advised. It suggests using techniques such as hyperparameter tuning, principal component analysis, or even data feature.

4.3. High Risk

A drift greater than 5% is considered high risk and typically necessitates refitting or rebuilding the model. This level of drift suggests that the underlying data has changed sufficiently that the current model can no longer accurately predict outcomes.

5.0. References

Tuychiev, B. February, 2024. *An End-to-End ML Model Monitoring Workflow with NannyML in Python.* Datacamp. https://www.datacamp.com/tutorial/model-monitoring-with-nannyml-in-python