

Introduction

Hotels face an ongoing challenge in mitigating cancellations while attempting to operate at maximum capacity. Booking cancellations influence the demand-management decisions that determine overall revenue and customer satisfaction. If cancellations are carefully considered, hotels could serve more individuals and make a higher revenue. The problem lies in the uncertainty of cancellations, however, access to additional information that could aid in determining the validity of a reservation can develop a model. With a formula capable of predicting which reservations are most likely to be canceled and how many reservations will remain, the hotel could run most efficiently. The project would influence the hospitality industry, especially during times of mass reservations such as holidays or school breaks.

Similar studies have been conducted with the same intention of improving the hospitality industry. Realistically, no model will perfectly predict the cancellations, but each similar project provides a new perspective and unique model in predicting the cancellations. In our search for similar projects, we found use of probability, machine learning classification, and Bayesian models. Each study had a similar goal accomplished through different methods.

The first study developed a hotel booking cancellation prediction using a Kaggle dataset like ours (Jishan, 2024). The intention was to improve hotel resource management revenue and satisfaction using a predictive model. To perform the model, the study tested applied Bayesian Models and a Beta-Binomial. Findings suggested that the Bayesian approach was an effective tool in prediction and key factors included the number of adults, children, stay duration, lead time, car parking space, room type and special requests.

The second study focused on an integrating model based on personal name records (Chen,2023). The goal was to establish operational strategies for hotel management. The authors propose a combination of two methods to determine relationships between correlated variables and identify the most important predictors. This was accomplished through a combination of Bayesian Network and Lasso Regression models to develop a more accurate prediction rather than using one model alone.

The third similar study used data from a hotel's property management system to train a daily classification model (Antonio, 2017). It was tested in two hotels alongside a measurement of the impact of action taken with bookings "likely to cancel." Results on the impact of action taken indicated bookings contacted were less likely to be canceled. This further reaffirms the value of a predictive model considering that action can be taken alongside it.

Overall, previous studies validate the importance of the predictive algorithm while providing insight into how to improve the model. Additionally, the last study suggests that the algorithm could be used alongside preventative measures to decrease cancellations. In summary, the objective to maximize operational capacity is attainable and the information from our study can add to improving the hospitality industry.

Data and Methods

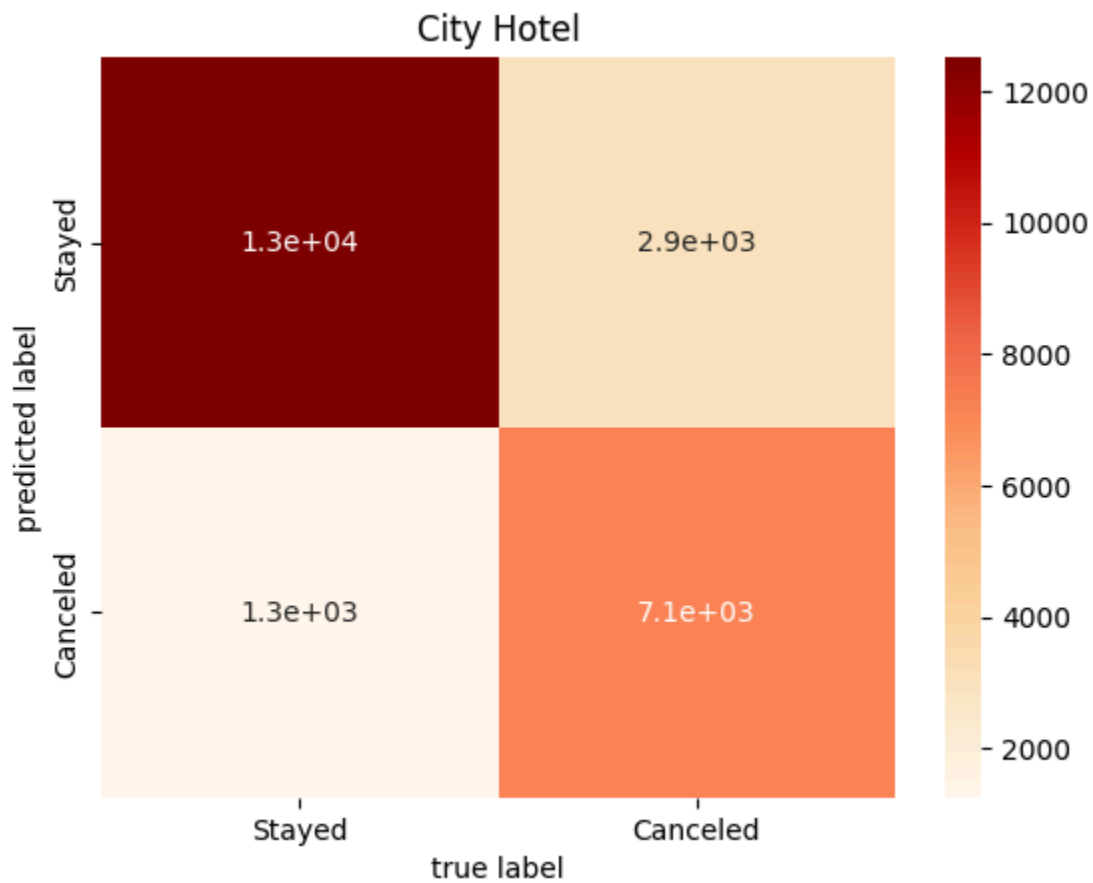
For our specific analysis we used a dataset from Kaggle to develop a predictive model. We were able to interpret which variables are most important in developing a model and could be used as predictor variables while analyzing our data on Google Collab. We then ran multiple models to determine which would perform as a more effective predictor for the specific situation. When running a logistic regression and random forest classification, we found that the RFC performed better. The measure of accuracy and success is based on the results of both the confusion matrices and classification reports of each model. In interpreting the report and matrices we find that the RFC had a higher accuracy report and fewer type 2 (False negative errors). In other words, the model was more likely to infer a type 1 error (False positive) that no cancellation would be made. This allows leniency and prevents overbooking based on the predictive analysis.

Results

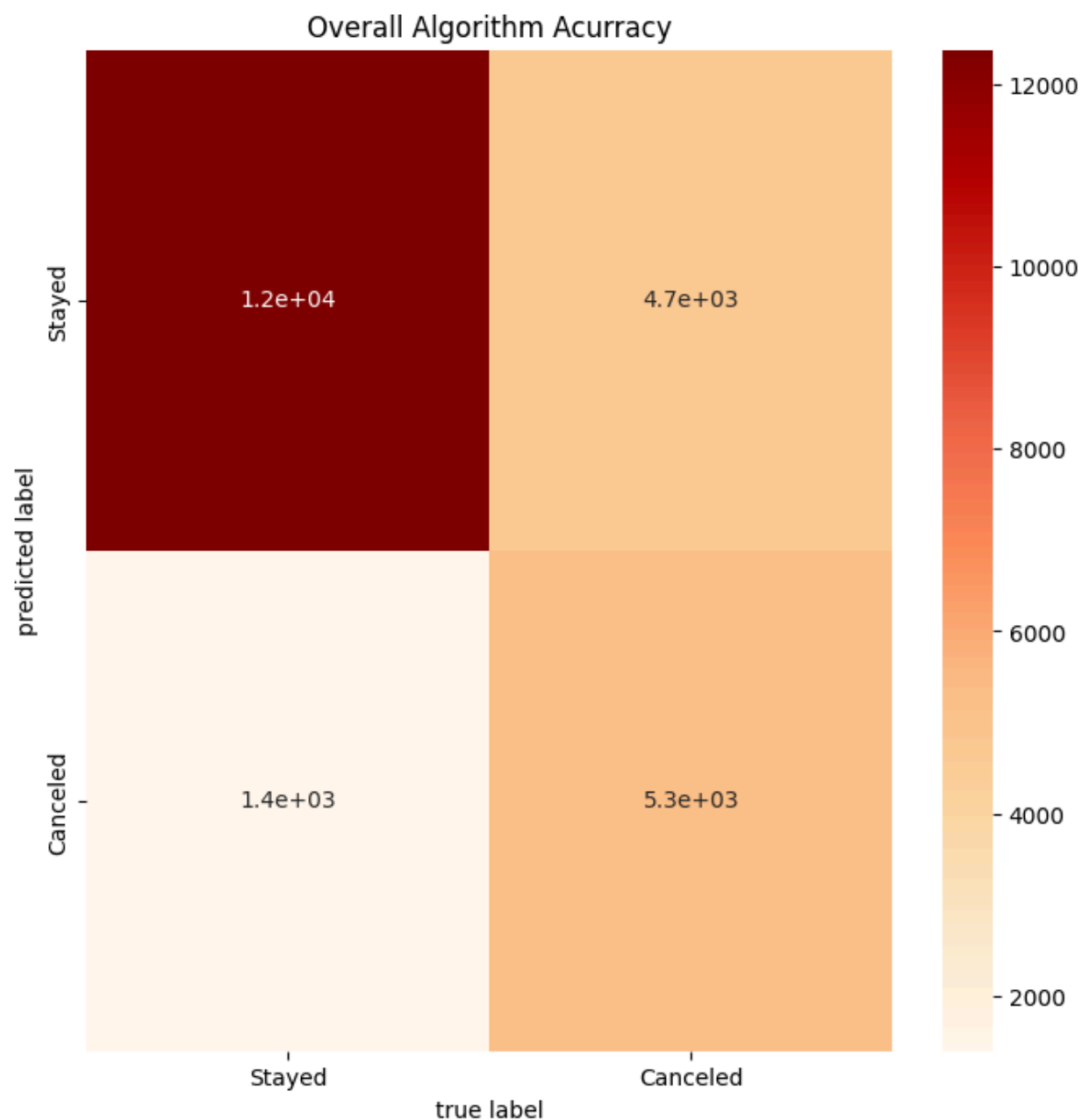
After testing our two models, Logistic regression(LR) and random forest classification(RFC), RFC performed significantly better at predicting booking cancellations. With 82-83% accuracy across city and resort hotels, this was a large improvement over the minimum expected baseline of simply predicting 66% of bookings as non-cancelled based on historical trends. Below are various confusion matrices of our models:



Accuracy report(Resort): 0.8290896987851556



Accuracy report(City): 0.8246491301790067



Accuracy report(Resort): 0.8248592318682242

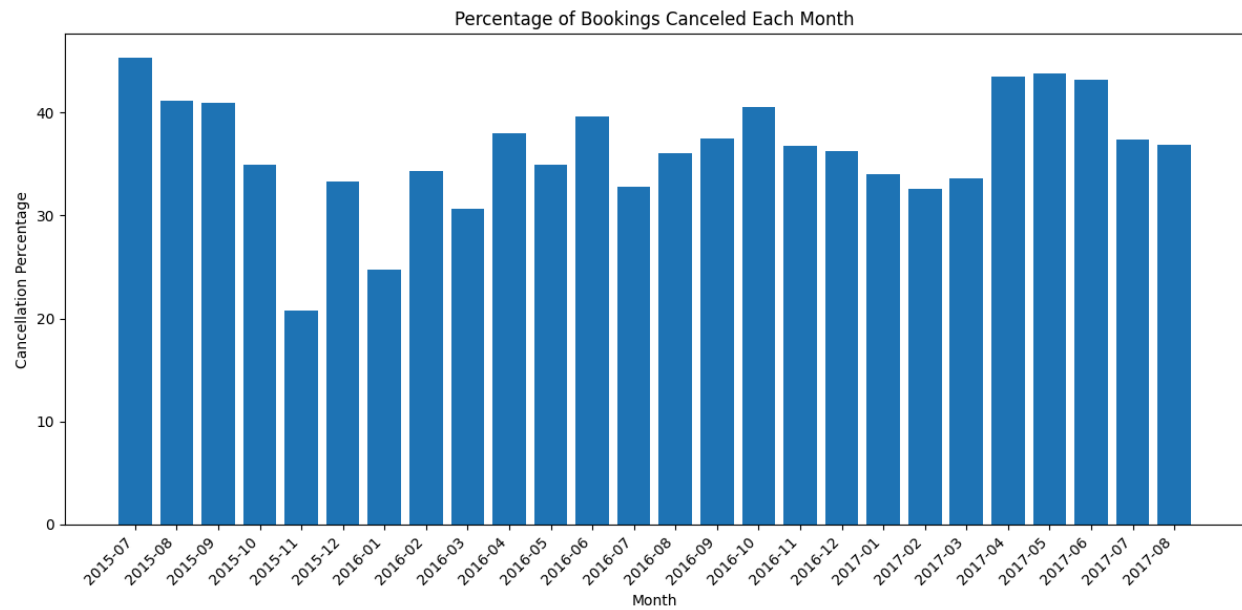
Our initial exploratory data analysis examined the correlation coefficient of various features to cancellation. As you can see, the features with the highest correlation were deposit type, lead time, booking changes, hotel type, and customer type.

```
deposit_type    0.468634
lead_time       0.293123
booking_changes 0.144381
hotel           0.136531
customer_type   0.121862
Name: is_canceled, dtype: float64
```

In contrast, during the predictive modeling, the features with the greatest importance were deposit type, lead time, and average daily rate(adr), as well as arrival date. This indicates that the price of the rooms as well as what time of year the booking is placed for both play a more significant role in cancellation likelihood than previously expected.

	Feature	Importance
14	deposit_type	0.222838
0	lead_time	0.193751
17	adr	0.144373
4	arrival_date_day_of_month	0.091790
3	arrival_date_week_number	0.068400

This matches some of our previous findings, where there was a large variance in bookings over the course of the two years that our data spans.



Work Cited

- Antonio, Nuno, et al. "Predicting Hotel Bookings cancellation with a machine learning classification model." *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Dec. 2017, pp. 1049–1054, <https://doi.org/10.1109/icmla.2017.00-11>.
- Chen, Shuixia, et al. "Prediction of hotel booking cancellations: Integration of machine learning and probability model based on interpretable feature interaction." *Decision Support Systems*, vol. 170, July 2023, p. 113959, <https://doi.org/10.1016/j.dss.2023.113959>.
- Jishan, Md Asifuzzaman, et al. "Hotel Booking Cancellation Prediction Using Applied Bayesian Models." *arXiv.Org*, 23 Oct. 2024, arxiv.org/abs/2410.16406. Accessed 13 Nov. 2024.