Hotel Reservation Cancellation Flexibility Risk

Anthony L Wilson

College of Science and Technology, Bellevue University

DSC 680: Applied Data Science

Professor Catie Williams

5/26/2023

Hotel Reservation Cancellation Flexibility Risk

Background and History

The hospitality industry has evolved in the last decade with the rise of online bookings. Customers can conveniently search and book travel reservations online. With the rise of online booking, free cancellations have also become popular. This feature has revolutionized the industry and changed the way hotels do business. Hotels need to be able to forecast what actual reservations will look like for several reasons. First, hotels must be able to plan for the number of guests for the bottom line and staffing, bedding, towels, food, and accommodations. Second, hotels must also deal with under-booking due to last-minute cancellations. Third, hotels need to account for overbooking. It is a common practice for hotels to overbook to account for last-minute cancellations. (Schwartz, 2021)

To better understand the risks of reservation cancellations, two data sets will be used from Kaggle. The first is "Reservation Cancellation Prediction," by Gaurav Dutta, updated January 2023. This data set identifies customers' behavior with booking, including when, length of stay, number of guests and children, type of room, repeated guest, lead time to stay, price of a room, etc. (Dutta, 2023). The second data set was created based on a deep learning model where they took the first data set and made it similar with different outputs so that more testing and validation could be done on the models. (Chow & Reade, 2023) This data set is called "Binary Classification with Tabular Reservation Cancellation Dataset," by Ashley Chow and Walter Reade in 2023.

Business Problem

Online hotel reservation cancellations can pose an issue for hospitality companies as many offers free cancellations 24 hours in advance. The convenience of booking and canceling online puts the industry at risk of losing profits due to flexibility and competition. "By exposing cancellation drivers, models help hoteliers to understand booking cancellation patterns better and enable the adjustment of

a hotel's cancellation policies and overbooking tactics according to the characteristics of its bookings" (Antonio, de Almedia & Lunes 2019). This project will try to determine a cancellation model on historical reservations to help the industry better mitigate risk.

Data Explanation

Before beginning to build a model to forecast cancellations, the data from both sets needed to be prepped. Overall, the data was clean and did not need any significant changes. Categorical variables were all set with numerical values. A couple of categorical variables must be transposed and converted to binary values. The features chosen to do this with were the type of meal plan, room type, market segment, and day of the week. Arrival year, month, and day were concatenated into a date and then converted to a day number in the year. New features included length of stay, total guests, single parents with kids, and bookings without adults. A Data Dictionary depicting new features is included in Appendix A.

Methods

To begin creating the model features first needed to be selected. The open-source Python library Featurewiz was used for feature selection. This was used because it effectively identifies significant variables relative to the target variable (Sharma, 2020). Based on Featurewiz output, 17 features were chosen as most important for booking cancellations. We used the package random forest classifier from the Scikit Learn library to fit and train the model. The training data sets from Kaggle were used to train and test the model. "Reservation Cancellation Prediction" (Dutta, 2023) was used to train the model, and "Binary Classification with Tabular Reservation Cancellation Dataset" (Chow & Reade, 2023) training data set was used to test the model. Three different models were created, two with Random Forest Classifiers. One used all the features, and the other only used feature selection. The third model used Random Forest Classifiers with Hyperparameter tuning. We use hyperparameter tuning to allow multiple

models to be created using Random Forest Classifiers, with different random numbers of trees and tree lengths to help under and overfit the model. We tested the accuracy and balanced accuracy of each model.

Analysis

The three different model accuracy scores are in Figure 1. Two different types of accuracy scores were used for each model. The average accuracy score and the balance were used to identify which model provided more balance. When both accuracy scores are close, the data is well-balanced. (Olugbenga, 2023) The hyperparameter tuning ended up choosing a better tree where the data was more balanced.

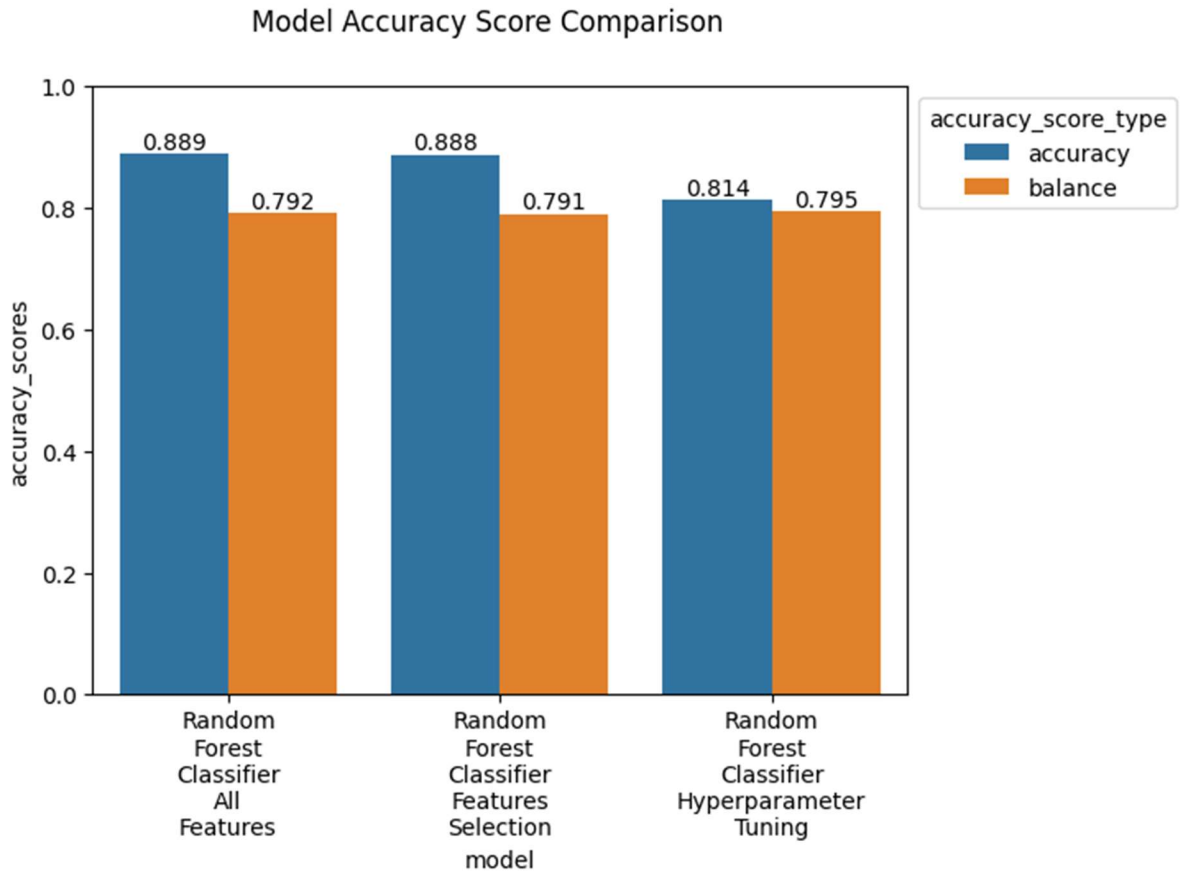

Figure 1: Model Accuracy Score Comparison (Wilson, 2023)

The most influential variable in the features was lead time, the days between reserving the hotel, and the actual stay. It seemed skewed when comparing the training data to the test see Figure 2.

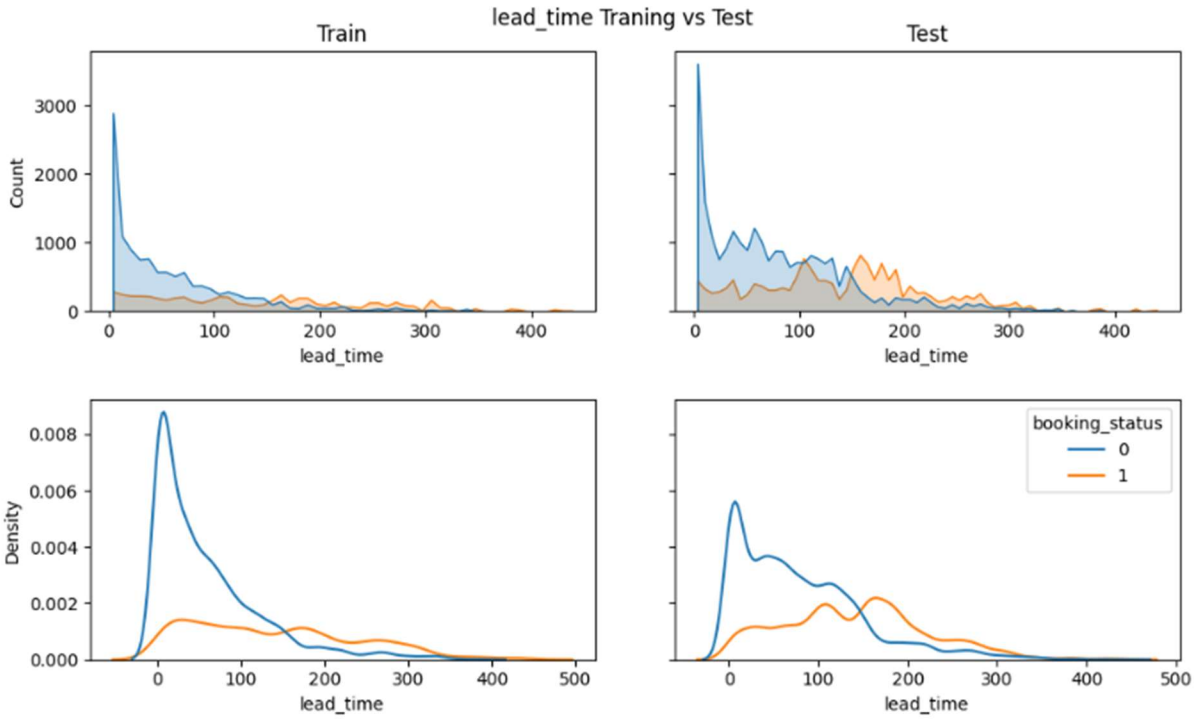

Figure 2: Lead Time vs. Booking Status (Wilson,2023)

In the top two polygon histograms from Figure 2, the graphs look similar between test and train. The test dataset has more cancelations around the 200 mark. Looking at the bottom two probability density graphs in Figure 2, we see differences, with the non-booking status more skewed to the right. The test data has data with more cancellations in the center.

The arrival day number was a newly created feature representing the day number in the year between 0 and 365. This was used so dates could be removed from the calculation. It would give numerical values for the model to ingest. In Figure 3, the test environment cancellations are above the non-canceled bookings. Looking at Figure 4, the percentage of completions for the test environment is up by 5% of the total sample.

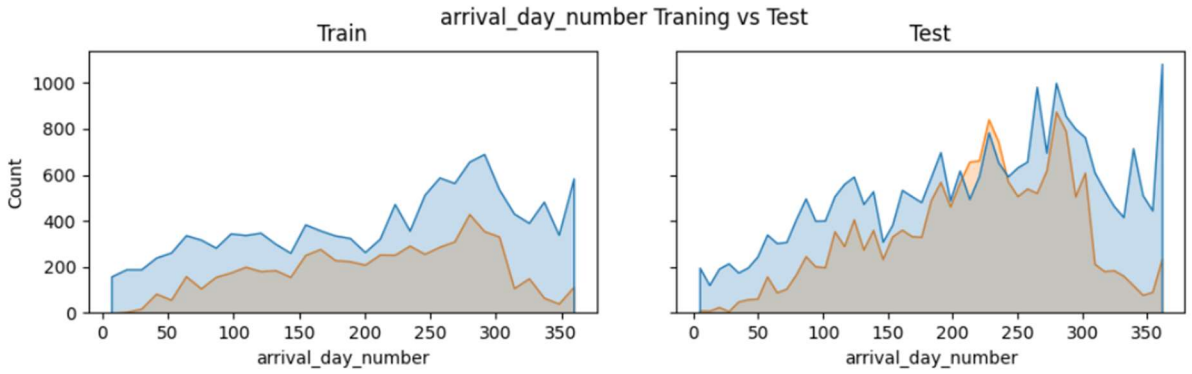

Figure 3: Arrival Day Number Vs. Booking Status (Wilson 2023)

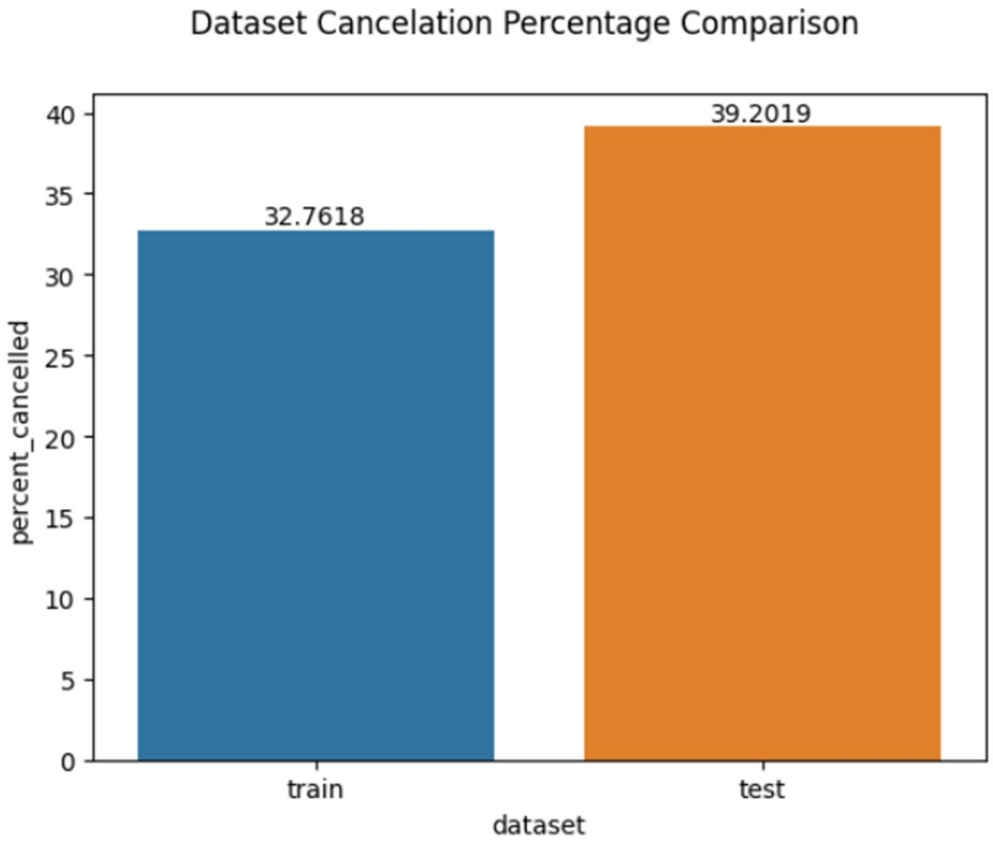

Figure 4: Dataset Cancelation Percentage Comparison

Assumptions

There were several assumptions made throughout this project. First, it is assumed that all data comes from one reliable source that accurately represents the hotel industry. If data had come from multiple sources, it could have given a better representation of the population. Second, it is assumed that the data was gathered ethically without bias. Third, although the data is five years old, it is assumed that the data and findings are still relevant to current industry trends. Finally, an assumption that the data had been cleaned accurately before our access.

Limitations

Some limitations need to be considered in the scope of the study. First, there was limited information about the customers' financial and demographic status. This information could have helped identify additional trends in the data. Additionally, the data did not indicate the type of hotel or amenities. For example, it is unknown the hotel quality rating, local attractions, pool, laundry services, etc. These are features that could impact industry trends. A final limitation is the age of the data. This data was all gathered pre-COVID. Industry standards, availability, etc., have likely evolved post-COVID.

Challenges

There were some challenges with this data in creating a successful model. The data was already clean, which reduced a lot of the prep work. However, because it has already been cleaned, we lose an understanding of the source of the data. There is not sufficient information to know if it accurately represents the population. Another challenge was that there was not a high correlation between many variables, potentially making forecasting difficult.

Future Uses/Additional Applications

There is the potential to use this model in future decisions. First, this could be used with other data sets to research and better understand the hospitality market. This could help identify growth opportunities, inventory and staffing needs, and day-to-day operational logistics such as maintenance needs, future budgets, and plans. Beyond logistical applications, this model could be applied to the short-term vacation rental industry by aiding in research and development for future opportunities.

Recommendations

Based on the data, it will let us know how accurate the model is in predicting whether or not a customer will cancel or retain their reservation. The first data set model predicts with a 79% accuracy. This is based on one data set, where the second data set was generated off of the first one. The model has the potential to help the industry, but due to a lack of information about the source, it is recommended to continue research and development.

Implementation Plan

If this model were used within the business setting, it would need to be tied to a reservation system or software program to help hotels determine whether to allow for more overbooking to ensure they can keep the hotel filled and account for cancellations. This would indicate how much overbooking reduces the risk of under booking due to cancellations. This would allow the data to keep retraining and fine-tuning the model as more and more data comes in from the system or software. Over time the model will become more and more accurate.

Ethical Assessment

There is not a lot of information as far as where the data is coming from. This could threaten the validity of the model and interpretation of the results. If used in the industry, it could negatively affect business decisions if invalid.

Another ethical consideration is the lack of information on the sampling. This data could come from only one source; if it came from several different sources, the data could better represent the market. This will need to be considered as a recommendation is developed.

Conclusion

The hospitality industry is changing due to online booking opportunities. Hotels need to mitigate risk with overbooking and under booking strategies. This model could potentially help businesses forecast member cancellations and reduce the risk of lost revenue due to its 79% accuracy in predicting cancellations or retainment of bookings. Successfully implementing this model utilizing software into business operations would allow additional data to be gathered, refinement of the model, and ultimately the opportunity to improve accuracy over time.

References

Antonio, N., de Almeida, A., & Nunes, L. (2019). *Big Data in Hotel Revenue Management: Exploring Cancellation Drivers to Gain Insights Into Booking Cancellation Behavior*. Cornell Hospitality Quarterly, 60(4), 298–319. htt ps://doi.org/10.1177/1938965519851466

Chow, A., & Reade, W. (2023). *Binary classification with a tabular reservation cancellation dataset*. Kaggle. https://www.kaggle.com/competitions/playground-series-s3e7/overview

Dutta, G. (2023, January 6). *Reservation cancellation prediction*. Kaggle. https://www.kaggle.com/datasets/gauravduttakiit/reservation-cancellation-prediction

Hollander, J. (2023, February 16). *What is overbooking in hotels? - hotel tech report*. Hotel Tech Report. https://hoteltechreport.com/news/what-is-overbooking-in-hotels

Hotel cancellations pose a great challenge. (2020, January 21). https://www.rateboard.io/en/blog/hotel-cancellations-pose-a-great-challenge

Olugbenga, M. (2023, May 9). *Balanced accuracy: When should you use it?*. neptune.ai. https://neptune.ai/blog/balanced-accuracy

Schwartz, Z. (2021, June 29). *Consumers vs. revenue managers? the case of cancellations and no shows*. Boston Hospitality Review Consumers vs Revenue Managers The Case of Cancellations and No Shows Comments. https://www.bu.edu/bhr/2021/06/29/consumers-vs-revenue-managers-the-case-of-cancelations-and-no-shows/

Appendix A: Data Dictionary

Number of Children	Description
no_of_adults	Number of adults
no_of_children	Number of Children
no_of_weekend_nights	Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
no_of_week_nights	Number of weeknights (Monday to Friday) the guest stayed or booked to stay at the hotel
required_car_parking_space	Does the customer require a car parking space? (0 - No, 1- Yes)
lead_time	Number of days between the date of booking and the arrival date
arrival_year	Year of arrival date
arrival_month	The month of arrival date
arrival_date	Date of the month
repeated_guest	Is the customer a repeated guest? (0 - No, 1- Yes)
no_of_previous_cancellations	Number of previous bookings that were canceled by the customer prior to the current booking
no_of_previous_bookings_not_canceled	Number of previous bookings not canceled by the customer prior to the current booking
avg_price_per_room	Average price per day of the reservation; prices of the rooms are dynamic. (in euros)
no_of_special_requests	Total number of special requests made by the customer (e.g. high floor, view from the room, etc)
booking_status	Flag indicating if the booking was canceled or not.
type_of_meal_plan_[0-3]	Type of meal plan booked by the customer plan id
room_type_reserved_[0-6]	Type of room reserved by the customer. The values are ciphered (encoded) by INN Hotels.
market_segment_type_[0-4]	Market segment designation.
length_of_stay	Number of weekend nights plus Number of weeknights
arrival_day_number	Arrival Day Number in year
day_of_week_-1	Null Day of Week
day_of_week_[1-7]	Day of the week based on values Monday -Sunday.
total_guests	Number of adults + Number of Children
single_parent_with_children	Reservation with no adults and children
bad_booking	Reservations where there are children but no adult

Appendix B: Questions

1. **Could this model be applied to Ma and Pa hotels and motels?**

   Yes, the timing of when they could start using the model would be the better question to ask. Implementing the model would require available data to be used. If they don't have data available, the company must gather or start a process to store the data.

2. **Could this model help in the apartment rental industry?**

   Yes, but a different model would be required. Hotels are not generally used as a resident. The features needed for apartments would be additional. An apartment company would have more demographic data to be used as the features that could help them understand their residents. For example, if you know one of your attendants works as a cashier making $25,000 a year and the company wants to raise the rent by $300, there may be a chance they don't renew.

3. **How adaptable is this model to market disruptions?**

   I would say it captures some based on how many rooms are going for, assuming the model would continue collecting data and tuning. Adding items such as the stock market trends may offer features that could help it become more accurate.

4. **How much more data needs to be gathered to make this model usable?**

   We would need to gather at least one year's worth of current historical data, but if we could get up to three years, that would be ideal. We are looking into the day number for each year, and understanding each day may help us understand daily trends.

5. **Did you consider running any other models?**

   Yes, I would want to take more time and look at other potential models to see if something is better. For example, the XGBoost would most likely do well for this dataset.

6. **Are there other hospitality data sets available to gain a better understanding?**

   Yes, there are other datasets available on a couple of different websites.

7. **What are the benefits of raw data that hasn't already been cleaned?**

   When cleaning data, one can have specific things to do and remove information valuable for other projects and analysis.

8. **How is overbooking handled in the industry today?**

   Hotels use booking software to help manage the booking. Similar to this model, they use historical trends. When hotels do not have enough rooms, they book another room in another hotel. (Hollander, 2023)

9. **Can the charts, graphs, and analysis be converted into a readable, non-technical business white paper?**

   Yes, this would make sense as it would help promote the business. It will also help educate businesses about what the model can and can't do.

10. **For businesses to implement a model like this into their software, how much historical data would they need for the model to capture their business and make predictions effectively?**

    A minimum of six months and one to two years would be ideal.