Predicting Rebooking with Multiple Logistic Regression

# Executive Summary

When attempting to forecast demand within the hospitality industry, analyses are often inflated due to canceled and rebooked reservations. This occurs because hotel reservations get canceled and rebooked when the same package becomes available at a lower rate (Clay, 2023). Because it was not always possible to know which reservations were rebooked, this analysis aimed to develop a Multiple Logistic Regression model of hotel reservations to predict whether or not a reservation was canceled and rebooked. The null and alternative hypotheses were:

**H0:** The booking lead time and average price per room will not statistically significantly predict that a reservation will be rebooked.

**Ha:** The booking lead time and average price per room will statistically significantly predict that a reservation will be rebooked.

# Data Analysis

The Hotel Reservations Dataset (Raza, 2023), obtained from Kaggle, contained 19 columns, shown in Table 1, and 36,275 rows, each about a single hotel reservation.

**Table 1**

*Variables Included in the Hotel Reservations Dataset*

|  |  |  |
| --- | --- | --- |
| Field | Type | Dependency |
| no\_of\_adults | Continuous | Independent |
| no\_of\_children | Continuous | Independent |
| no\_of\_weekend\_nights | Continuous | Independent |
| no\_of\_week\_nights | Continuous | Independent |
| type\_of\_meal\_plan | Categorical | Independent |
| required\_car\_parking\_space | Categorical | Independent |
| room\_type\_reserved | Categorical | Independent |
| lead\_time | Continuous | Independent |
| arrival\_year | Continuous | Independent |
| arrival\_month | Continuous | Independent |
| arrival\_date | Continuous | Independent |
| market\_segment\_type | Categorical | Independent |
| repeated\_guest | Categorical | Independent |
| no\_of\_previous\_cancellations | Continuous | Independent |
| no\_of\_previous\_bookings\_not\_canceled | Continuous | Independent |
| avg\_price\_per\_room | Continuous | Independent |
| no\_of\_special\_requests | Continuous | Independent |
| booking status | Categorical | Independent |

## Data Preparation

Because rebooking for a lower rate generally only occurs online, only reservations booked online were included. The remaining data was relatively clean. There were no missing values or duplicate entries. To prepare the data for analysis the following steps were taken:

1. Removal of numeric outliers, defined as having a z-score greater than 3 or less than -3.
2. Deterministic matching to link canceled reservations to rebooked reservations.
3. Univariate and bivariate exploratory analysis to identify data structures and relationships.
4. Continuous variables were scaled to prevent model skew.
5. Categorical variables were one-hot encoded. One level per variable was dropped to prevent multicollinearity.
6. Identified multicollinearity using variance influence factor. Highly correlated variables were removed.
7. The data was split so that 20% was reserved for testing.

## Model Development

The dataset was first tested to ensure it met the assumptions of Multiple Logistic Regression analysis. The initial model was built and fitted to all remaining variables using the training data. The model was then reduced using backward stepwise feature selection to include only the statistically significant variables. The reduced model summary can be seen in Figure 1.

**Figure 1**

*Screenshot: Reduced Model Summary*

Table

Description automatically generated

## Post-Hoc Model Development

The data was re-split into training, validation, and testing sets. The validation and testing set each contained 20% of the dataset with the remaining 60% used for training. A Feed-Forward Neural Network was built to include an input layer, two dense hidden layers with 500 and 250 nodes respectively, and an output layer with one node. Both hidden layers were activated by the Rectified Linear Unit (reLU) function, and the output layer was activated by the sigmoid function. The loss was measured using Balanced Cross-Entropy to account for class imbalance in the dataset. The metrics used were Accuracy and recall, and the Adam optimizer was used. During fitting, an early stopping monitor was used on the validation loss with patience of 5 epochs.

# Findings

The final Multiple Logistic Regression model had a test Accuracy of 0.95 and a test recall of 0. Because both lead time and average price per room were statistically significant, the null hypothesis can be rejected in favor of the alternate hypothesis. However, the model was unable to identify reservations that will be rebooked. The Feed-Forward Neural Network model had a test Accuracy of 0.70 and a test recall of 0.73. The model was able to identify reservations that will be rebooked 73% of the time but at the cost of overall model Accuracy.

# Limitations

According to the author of the dataset used for this analysis, the data was extracted from a hospitality training environment and included no identifying information. To account for this, deterministic matching was used to link cancellations to rebooked reservations. This means both models resulting from this analysis can only be considered proof of concepts. The Multiple Logistic Regression model was most heavily limited by the class imbalance present in the data. The standard optimization function was unable to detect rebooked reservations when they occurred. The Feed-Forward Neural Network was most heavily limited by the low volume of data. Neural networks are known to perform better when trained on larger volumes of data. As always, correlation does not imply causation. Yes, statistically significant relationships were found. However, only a controlled experiment can prove one or more of the predictors caused an outcome.

# Actions

Based on the findings of this analysis, it is recommended that the next step be to extract hotel reservation data that contains confirmed links between the canceled and rebooked portions of the rebooked reservations. The new data should then be used to fit the neural network and determine whether any side effects were introduced by using deterministic matching. In addition to this follow-up, the following two analyses are proposed:

1. Exclude the initially canceled portion of the deterministically matched rebooked reservations and use the data to produce a time-series model of hotel demand. Compare how the results differ when the time-series model is applied to the same data but with the predicted rebooks removed instead of the deterministically matched rebooks.
2. Analyze the seasonality of rebooked reservations and average price per room to determine if any pattern exists between the time of year, how the price fluctuates, and whether a reservation will be rebooked.

# Benefits

During the course of analysis, it was discovered that rebooked reservations accounted for approximately 5% (998 out of 20,223) of online hotel reservations. When forecasting hotel reservations, it is important to identify and exclude these records to prevent future demand from being falsely inflated. The dataset contains reservations that were booked between 7/1/2017 and 12/31/2018, a period of 17 months. When averaging the number of rebooks that occurred by month, this equals an approximate inflation of 60 reservations. When running a hotel, 60 reservations a month can make a big difference in areas such as staffing, stocking of supplies, and marketing to name a few.

# References

Clay, B. (2023). *Why Cancelling & Re-Booking is Killing Your Revenue*. Retrieved from IDeaS: <https://ideas.com/cancelling-re-booking-killing-revenue/>

Raza, A. (2023). *Hotel Reservations Dataset.* Retrieved from Kaggle: <https://www.kaggle.com/datasets/ahsan81/hotel-reservations-classification-dataset>