Karan Modi, Folarin Omotoriogun, Mark Tegeler, Shih-Yuan Wang, Yujie Wen

# Predicting Hotel Booking Cancellations

## A. Introduction

Our team used the *Hotel Booking Demand* dataset found on Kaggle[1]. This is a robust dataset containing 119,390 data points. The dataset contains daily booking data from 2015 to 2017 for both city and resort hotels located around the world. From Appendix Exhibit 1, you can see there are 32 variables contained in the data. The data contains information such as when the booking was made, reserved room type, arrival day, length of stay (week and weekend nights), if the customer is a repeat guest, lead time (number of days between booking date and arrival date), and if a cancellation occurred (previous and current bookings). Appendix Exhibit 1 also lists the variables further broken down into subcategories. A few examples of data further categorized include customer type, market segment, and reservation status.

## B. Business Objectives

There are many questions one could use this data to answer. During our initial exploration of the data using Tableau we looked at various possible relationships. Our team decided to focus on reservation cancellations, as this has a substantial impact on hotel revenue and profitability. We used the confusion matrix to better understand the logic and implications behind possible cancellation scenarios. Table 1 represents the relationships between predicted and actual cancellations. Our models focused on maximizing sensitivity (true positive rate) and specificity (true negative rate) for predicting cancellations, but emphasized specificity over sensitivity.

**Sensitivity (True Positive Rate): TP/(TP+FN)**

- Probability of ***correctly predicting a cancellation*** while the booking ***was actually canceled***

**Specificity (True Negative Rate): TN/(TN+FP)**

- Probability of ***correctly predicting non-cancellation*** while the booking ***was actually not cancelled***

The false negative scenario has the biggest potential impact to hotel owners. With a false negative, hotels would predict that a reservation is not cancelled, but a customer is a "no-show" at the time of check in. This has serious implications to hotels, as it creates underutilization of capacity, and with enough cancellations, creates a bullwhip effect. Therefore, we believe that hotel owners must proactively counter against false negatives. The recommendations section discusses ways to address this false positive scenario, but as a teaser, our main recommendation is thwarting such scenarios by overbooking.

With a false positive, a hotel would have predicted a customer-initiated cancellation, which did not evolve, with the customer actually showing up for her or his stay. A false positive can also be concerning for hotels. If a hotel doesn't reserve a room for a customer, believing she or he would be a “no-show,” and the customer actually shows up and doesn’t have a room, the customer would be upset. In this scenario, the hotel would likely receive negative customer feedback and PR, potentially impacting future bookings. Traditionally, this is not how hotels operate though, as they don’t use predictive models and instead plan for all bookings as if customers will truly “show,” but as has been shown from our analysis below, a large percentage of customers actually cancel, leaving hotels stuck with unfilled rooms.

**Table 1 – Reservation Cancellation Confusion Matrix**

| | | Actual Class | |
|---|---|---|---|
| **Predicted Class** | | **Not Canceled (0/-)** | **Canceled (1/+)** |
| | **Not Canceled (0/-)** | True Negative (TN) ↑ | False Negative (FN) ↓ |
| | **Canceled (1/+)** | False Positive (FP) ↓ | True Positive (TP) ↑ |

Table 2 below represents the reservation cancellation rate found in the dataset. To ensure data splitting was not biased, we checked whether the cancellation rate (proportion of cancellation) was similar across initial, training, and test datasets. As you can see, the cancellation rate was consistent.

**Table 2 – Reservation Cancellation Rate**

| | Not Canceled (0/-) | Canceled (1/+) |
|---|---|---|
| **Whole dataset** | 63% | 37% |
| **Training (75%)** | 63% | 37% |
| **Test (25%)** | 63% | 37% |

The dataset showed an actual cancellation rate of 37%, which is quite alarming. Such a high cancellation rate poses serious problems for hotels. Thus, our goal is to analyze what factors are most correlated with cancellations in order to predict which customers are most likely to cancel their reservation. Hotels could utilize such predictive models to better manage bookings and overall capacity levels.

## C. Data Exploration & Tableau Observations

We made a number of observations while exploring the data. Please see Appendix Exhibit 1 for the variable names used in our dataset.

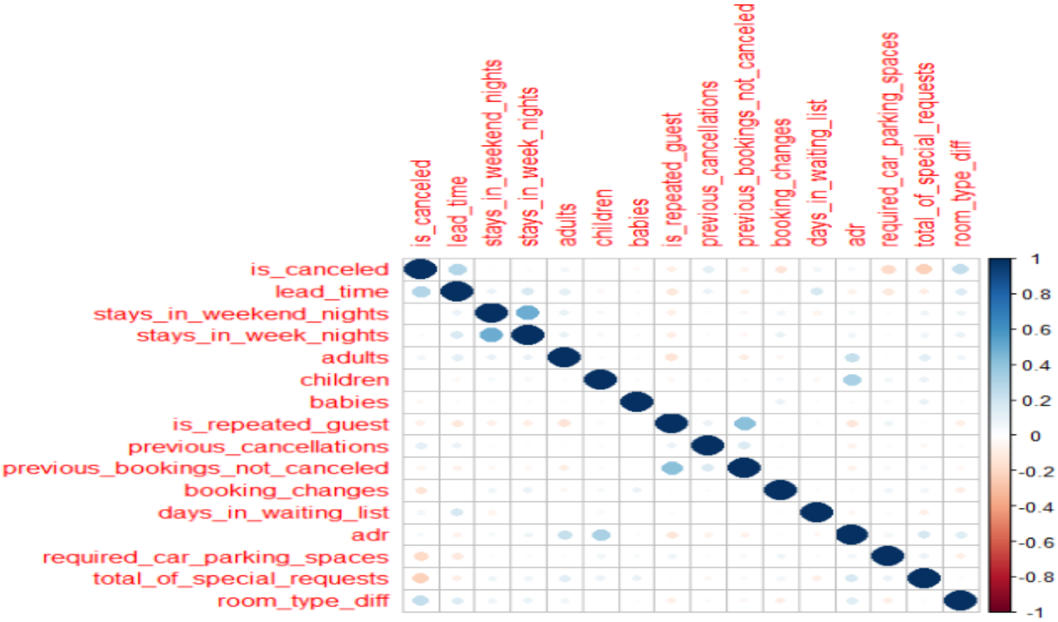
- *Lead Time Showed an Upward Trend in Cancellations* – From our dataset, we noted that there was somewhat an upward trend in average cancellations as lead time increased. See Appendix Exhibit 2.a for visualization.
- *Non-Refundable Deposits represented the highest cancellation rate at 14,494. However, we traced over 97% of these to booking from Puerto-Rico (PRT). Although PRT has the highest number of cancellations, its cancellation rate is not the highest at 56%.* See Appendix Exhibit 2.b for visualization.

- *More generally we see that customers who have made more than 10 previous cancellations are more likely to make future cancellations, however, we observed that over 5,000 bookings were cancelled with 1 cancellation history.* See Appendix Exhibit 2.c for visualization.
- *Customer Type* – From our dataset, we noted that Transient customer types are more likely to result in cancellations See Appendix Exhibit 2.d for visualization.

## D. Data Manipulation

To prepare the data for predictive models, we need to do data cleaning and partitions. First, we converted all NULLs into NA and calculate the number of NA in all variables. The "country," "agent," and "company" variables include many NA values, and have too many categories that require further dimensionality reduction. Moreover, some models cannot handle categorical predictors with more than a specified number of categories (e.g., random forest cannot handle categorical predictors with more than 53 categories, and classification tree model can only take factor variables with less than 32 levels) Thus, we decided to drop these variables for modeling. For further improvement, we could use Principal Component Analysis (PCA) or other dimensionality-reduction methods to reduce the number of features. The variable "children" has 4 NAs, so we decided to replace them with corresponding values in the "babies" column (=0). According to the variable description on Kaggle, both "Undefined" and "SC" meal categories represent no meal package, so we replaced "Undefined" values with "SC." We also converted character variables and date variables into factor variables. For all models we used, we added a new column to illustrate whether reserved room type is the same as assigned room type (1 if equal, 0 otherwise), and dropped "reserved_room_type" and "assigned_room_type."

After cleaning up the dataset and converting all variables into a reasonable format, we examined the correlation between numeric variables to avoid potential multicollinearity.



Based on the correlation graph above, numeric variables do not have strong correlation with each other. Last but not Least, we separated the dataset into training dataset (75%) and testing dataset (25%) with random seed (123).

## E. Predictive Models

Models Used: Association Rules, Logistic Regression, Classification Trees, K-nearest Neighbors, Random Forests

### 1. Association Rules

To consolidate our exploration of trends within our dataset to identify key independent variables, we used association rules:

| Support | Confidence | Minimum Length | Maximum Length |
|---|---|---|---|
| **0.1** | 0.1 | 2 | 5 |

This generated over 351,000 rules. However, we narrowed down our rules by exploring the variable most likely to lead to cancellations:

```
book_sub_rules<-subset(bookrules, subset = lift > 2 & support > 0.1)
summary(book_sub_rules)

# Rules for cancellations sort in descending order so that lastest results show first
cancellations <- subset(book_sub_rules, subset = rhs %in% "is_cancelled=1")
summary(cancellations)
cancellations<-sort(cancellations, decreasing=FALSE, by="lift")
inspect(cancellations)
```

**Results:** The top rules identified as likely to result in cancellations are: babies=0, previous_bookings_not_canceled=0, deposit_type=Non Refund, customer_type=Transient. This is also in alignment with our Tableau visualizations.

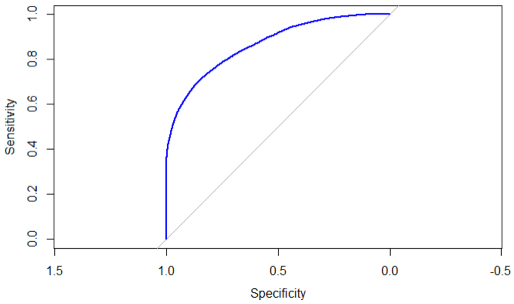
### 2. Logistics Regression

- **Variables used in the model (auto selection):** All variables except “country", "agent", "company", "reservation_status", “days_in_waiting_list”, “reserved_room_type”, “assigned_room_type”, and "reservation_status_date” in the initial dataset, and additional variable “room_type_diff”.
- **AIC:** 74359
- **Cutoff value used**: 0.5
- **The analysis and performance of the model:**

The variable “lead_time”, “room_type_diff”, “total_of_special_requests”, “adr”, “deposit_typeNon Refund”, “booking_changes”, “previous_cancellations”, “arrival_date_month”, “arrival_date_week_number” have the most significant impact in predicting cancellation.

- **Overall Accuracy:** 80.8%
- **Sensitivity (TPR):** 61.3%
- **Specificity (TNR):** 92.3%
- **AUC:** 0.8653

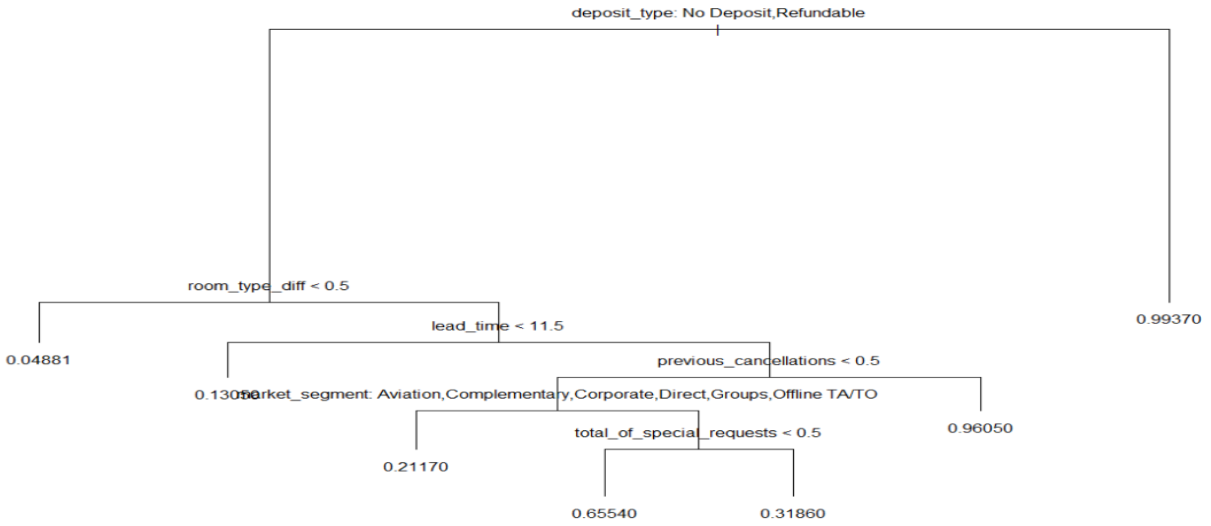
| | | Actual Class | |
|---|---|---|---|
| | | Not Canceled (0/-) | Canceled (1/+) |
| Predicted Class | Not Canceled (0/-) | 17,380 | 4,267 |
| | Canceled (1/+) | 1,453 | 6,747 |



### 3. Classification Tree

- **Variables used in the model (auto selection):** All variables except “country", "agent", "company", "reservation_status", “arrival_date_week_number”, “reserved_room_type”, “assigned_room_type”, and "reservation_status_date” in the initial dataset, and additional variable “room_type_diff”.
- **Cutoff value used**: 0.5
- **The analysis and performance of the model:**
  - **Overall Accuracy:** 80.5%
  - **Sensitivity (TPR):** 60.3%
  - **Specificity (TNR):** 92.2%
  - **AUC:** 0.7629

| | | Actual Class | |
|---|---|---|---|
| | | Not Canceled (0/-) | Canceled (1/+) |
| Predicted Class | Not Canceled (0/-) | 17,371 | 4,367 |
| | Canceled (1/+) | 1,462 | 6,647 |


deposit_type: No Deposit,Refundable
room_type_diff < 0.5
0.99370
0.04881
lead_time < 11.5
previous_cancellations < 0.5
0.13050
market_segment: Aviation,Complementary,Corporate,Direct,Groups,Offline TA/TO
0.96050
total_of_special_requests < 0.5
0.21170
0.65540
0.31860

We noticed that the classification tree classifies observations and predicts the cancellation based on deposit type, room type difference, lead time, previous cancellations, market segment, and total special requests.

### 4. K-Nearest Neighbors (KNN)

- **Variables used in the model:** All variables except "country", "agent", "company", "reservation_status", "reservation_status_date”, “reserved_room_type”, and “assigned_room_type” variables in the initial dataset, and additional variable “room_type_diff”.
- **The value of k used in the model:** K=13
- **The analysis and performance of the model:**
  Performance of the model using different K values:

| K | Overall Accuracy | Sensitivity (TPR) | Specificity (TNR) |
|---|---|---|---|
| **7** | 0.7851 | 0.6583 | 0.8601 |
| **9** | 0.7808 | 0.6403 | 0.8640 |
| **11** | 0.7782 | 0.6292 | 0.8664 |
| **13** | 0.7750 | 0.6169 | 0.8685 |
| **15** | 0.7719 | 0.6041 | 0.8713 |

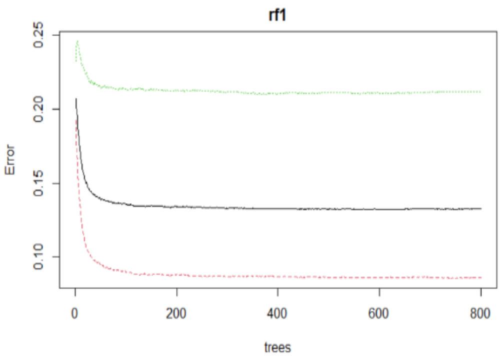
We built multiple models to see how K values influence the prediction accuracy and tried to find a tradeoff between decrease in sensitivity (TPR) and increase in specificity (TNR). Since the proportion of cancellations is much lower than the proportion of non-cancellations, a larger K would lead to higher specificity but lower sensitivity. As mentioned in the business objective, we prefer a higher specificity (i.e., lower false positive rate) over higher sensitivity while not sacrificing too much overall accuracy and sensitivity. Additionally, we keep K as an odd number to avoid ties in the voting. As a minor decrease in accuracy rate would cause a large number of false predictions, we decided to choose K value of 13.

**Performance results:**
Overall Accuracy: 77.5%
Sensitivity (TPR): 61.7%
Specificity (TNR): 86.9%

| | | **Actual Class** | |
|---|---|---|---|
| | | **Not Canceled (0/-)** | **Canceled (1/+)** |
| **Predicted Class** | **Not Canceled (0/-)** | 16,284 | 4,252 |
| | **Canceled (1/+)** | 2,465 | 6,847 |

### 5. Random Forest (Best Model)

- **Variables used in the model:** All variables except "country", "agent", "company", "reservation_status", "reservation_status_date”, “reserved_room_type”, and “assigned_room_type” variables in the initial dataset, and additional variable “room_type_diff”.
- **The value of parameters used in the model:**
  I. ntree=300: Number of trees to grow.
  II. mtry=5: Number of variables randomly sampled as candidates at each split.
- **The analysis and performance of the random forest model:**
  I. Tuning parameter ntree (=300)



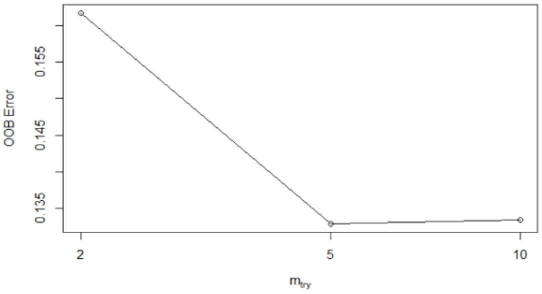
First, we used the default value of mtry (the square root of the number of predictor variables - 5) and the ntree value of 800 to train the model. In the left plot, the black curve represents overall Out-of-Bag error rate, and the other two lines are error rates for each class. We can see that the error rate does not decrease a lot after the number of trees is greater than 200, so we tried to use smaller ntree to test our model performance.

| ntree | Overall Accuracy | Sensitivity (TPR) | Specificity (TNR) | AUC* |
|---|---|---|---|---|
| **800** | 0.8668 | 0.7810 | 0.9176 | 0.9313 |
| **500** | 0.8667 | 0.7809 | 0.9174 | 0.9312 |
| **300** | 0.8673 | 0.7815 | 0.9181 | 0.9309 |

* AUC (Area under the ROC Curve): provides an aggregate measure of performance across all possible classification thresholds. The higher AUC is, the better the performance is.

The model performance for different values of ntree is quite similar with the cutoff 0.5 in different measures. We chose the one with higher specificity, that is, ntree is 300.

II. Tuning parameter mtry (=5) using the tuneRF() function in the randomForest package



Next, we used the ntree value of 300 and the tuneRF() function to find the optimal value of mtry. The left picture shows that when mtry value is 5, the Out-of-Bag error rate is the smallest. It seems that the initial value of mtry (=5) is the optimal value.

III. Check which variables are most important in driving the predictions
The arrival date, deposit type, lead time (Top 3), average daily rate (adr), previous cancellations, total of special requests, room type difference, and market segment are the most impactful variables for predicting cancellations in the model. (See Appendix Exhibit 3)

IV. The model performance with tuning parameters (ntree = 300, mtry = 5)
The performance based on different cutoff values:

| Cutoff | Overall Accuracy | Sensitivity (TPR) | Specificity (TNR) |
|---|---|---|---|
| **0.4** | 0.8558 | 0.8368 | 0.8671 |
| **0.5** | 0.8673 | 0.7815 | 0.9181 |
| **0.6** | 0.8641 | 0.7200 | 0.9494 |
| **0.7** | 0.8527 | 0.6529 | 0.9710 |

The specificity (TNR) gets higher while the cutoff value increases. Again, we prefer higher specificity (i.e., lower false positive rate) over higher sensitivity while not sacrificing too much overall accuracy and sensitivity. We chose cutoff value 0.6 for classifying cancellations in this model.

**Performance results:**
Overall Accuracy: 86.4%
Sensitivity (TPR): 72%
Specificity (TNR): 94.9%

| | | Actual Class | |
|---|---|---|---|
| | | **Not Canceled (0/-)** | **Canceled (1/+)** |
| **Predicted Class** | **Not Canceled (0/-)** | 17,800 | 3,108 |
| | **Canceled (1/+)** | 949 | 7,991 |

## F. Evaluation of Models

**The performance of different models:**

| Model | Overall Accuracy | Sensitivity (TPR) | Specificity (TNR) | AUC |
|---|---|---|---|---|
| **Logistic Regression** | 80.8% | 61.3% | 92.3% | 0.8653 |
| **Classification Trees** | 80.5% | 60.3% | 92.2% | 0.7629 |
| **K-Nearest Neighbors** | 77.5% | 61.7% | 86.9% | NA |
| **Random Forests** | 86.4% | 72.0% | 94.9% | 0.9309 |

The Random Forest model has the best performance, and the performance of the Logistic Regression and Classification Trees is quite similar in all measures. The KNN model is the least accurate and useful as it did not explain which variables were important and all variables that were included had equal influence. We also found that using the association rules to discover the key predictors for cancellations/non-cancellations and assist in building predictive models was a good method. This way we could see which variables might have higher impact on cancellations using association rules and decide whether to include them in the models.

**Similarities among different models:**

- In all the Logistic Regression, Classification Trees, and Random Forest models, the variable deposit_type, room_type_diff, lead_time, previous_cancellations, and total special requests are the most impactful factors in predicting cancellations. In the Association Rules, lead_time, previous_bookings_not_canceled, and deposit_type are also among the top variables that predict cancellations.
- In the Random Forests and Classification Trees, market segment is among the most impactful variables.
- In the Random Forest and the Logistic Regression, arrival date and average daily rate (adr) also have the most significant impact in predicting cancellation.
- In the Logistic regression and Association Rules, booking_changes is among the top variables that predict cancellations.

**Differences among different models:**

- In the Association Rules, customer type and number of babies are among the top variables for predicting cancellations, but we did not find that these variables are that important in other predictive models.

## G. Recommendations

Customer-initiated cancellations have a substantial impact on hotel revenue and profitability. By utilizing our predictive models, we feel that hotels could employ proactive tactics to better manage capacity. For instance, we believe that hotels should overbook customers meeting the risk factors that we have identified. Our team deems this tactic in the best interest of hotels, not customers, as there will be instances when customers are notified that their room is no longer available. Although, one could argue that by employing our strategies, hotels would increase margins, and as a result, could offer better overall pricing to customers.

Appendix Exhibit 4 shows a visualization of our overbooking decision process. We suggest that hotels double book high risk customers up to a defined percentage threshold, knowing that many of these customers won't show up. For example, a hotel could set a policy of double-booking high-risk customers by say 25%. The hotel would send autogenerated emails to customers 48 hours prior to check in requesting confirmation of their intended stay. If double-booked high-risk customers have not responded or cancelled by 24 hours prior to check in, they would be "bumped" by the overbooked guest with a higher "value score" (described in the paragraph below) and offered compensation (restitution), also based on their value score. We see such restitution on a spectrum, with options varying from meal, entertainment, or discounted stay vouchers. Overall, we see restitution being quite low for most displaced customers, as most are likely low value. We also suggest that hotels provide information on other local hotel openings to help ease tensions.

Although considered outside the scope of this analysis, which requires creating an additional model, we propose a bumping methodology to determine which customers get to stay and which get displaced based on a value score. Conceptually, we envision using criteria such as a customer's risk of cancellation (a cumulative score based on weighted cancellation risk factors), value of the stay, and booking history as primary factors for deciding which customers get to stay and which ones don't. Those customers with the higher value score would bump less valued customers. To employ our overbooking strategy, hotels must ensure they have language incorporated in their terms and conditions allowing hotel-initiated cancellations. Also, note that the overbooking percentage threshold, value score criteria, and bump notice period should be considered a baseline which can be catered based on varying business philosophies on risk and reward, as well as hotel-specific historical data.

A variation of the solution described above is currently used by airlines. An overbooked flight occurs when an airline sells more tickets on a plane than there are seats. This is a way to avoid empty seats from "no-show" passengers or missed connections. Airlines understand that a certain percentage of passengers will no-show on every flight. As a result, they sell more tickets than there are seats to ensure planes are at capacity in order to maximize revenue and profit. To determine how many seats to oversell, airlines use models that predict how many people might miss the flight based on factors associated with past no-shows. Of course, if everyone does show up and the flight is oversold, some customers will be bumped. Each airline has its own factors for determining which customers to bump. For instance, Delta Air Lines uses check-in order, loyalty status, and first class/business class fare vs economy class fare as determining factors. Meaning, passengers who check in later, are not in the loyalty program, and hold economy tickets are most likely to be bumped first. However, those with disabilities, unaccompanied minors, and members of the military are all protected from such bump practices.

Our team also believes that it would be prudent to exclude customers requesting disability accommodations as well as government bookings from hotel-initiated cancellations. Unaccompanied minors are not allowed to stay at hotels for liability reasons, thus this customer type is not a concern for excluding. Like our solution, airlines vary the restitution offered to customers. If the wait time for a comparable flight is less than one hour, no restitution is given. The maximum restitution is $775 (or 200% of the ticket price) for a two-hour delay and $1,550 (400% of the ticket price) for longer delays. Flight, hotel, and meal vouchers are used a means of restitution, depending on the length of the delay. As discussed earlier, our team envisions a similar restitution scale for any hotel-initiated cancellations, but with much lower restitution levels since there are typically many other lodging options available to customers.

There are some additional tactics that hotels could consider. For instance, hotels could hold rooms up to a certain date for bookings by high value consumers only. Premium pricing could be charged as well. We envision this practice being especially relevant during peak seasons as well as when there are nearby major events such as sports games, concerts, and conferences scheduled. To play on a similar concept, hotels also could set high deposits and non-refund policies during such high peak dates, as well as for last-minute bookings. And a final option would be for hotels to provide vouchers of a defined dollar amount to previous customers with high value scores as a way to entice additional stays.

Overall, our team believes all the possible strategies discussed thus far would be more profitable for hotels rather than continuing on with current practices and losing revenue due to ongoing cancellations. There are numerous positive supply chain implications of employing our proposed strategies. Having a more stable room utilization rate would allow hotels to fully utilize their support staff such as cleaning and restaurant employees. Additionally, hotels could better forecast demand for raw materials and finished goods, leading to more consistent purchase orders, or establishing long-term contracts for better pricing and terms. By having a predictable forecast, hotels could better buy in bulk for non-perishable items as well as ensure less spoilage/waste of perishable items. The bullwhip effect experienced by large swings in demand from cancellations would be minimized with more stable and smoothed demand.

## H. Conclusion

Our team applied many of the statistical methodologies covered in class to a real world supply chain issue. Through our analysis, we found that hotels face a significant problem with customer-initiated cancellations. A cancellation rate of 37% equates to a substantial under-utilization of capacity, which limits revenue and profit. Additionally, haphazard cancellations create a lumpy demand pattern, which ripples through the supply chain with detrimental effects. Our predictive models offer a solution to hotels. By anticipating which customers will likely cancel, hotels can employ proactive strategies to ensure demand. We have provided a few such strategies that hotels can utilize to reduce customer-initiated cancellations in order to create a higher, as well as more consistent, room utilization rate.

# Appendix

## Exhibit 1 – Hotel Dataset

### *A. Variables*

| | | |
|---|---|---|
| • Is Cancelled (Dependent) | • Meal | • Days in Waiting List |
| • Arrival Date Day of Month | • Reservation Status | • Is Repeated Guest |
| • Arrival Date Month | • Reservations Status Date | • Lead Time |
| • Arrival Date Week Number | • Reserved Room Type | • Previous Bookings Not Cancelled |
| • Arrival Date Year | • Average Daily Rate (adr) | • Previous Cancellations |
| • Assigned Room Type | • Adults | • Required Car Parking Spaces |
| • Country | • Agent | • Stays in Week Nights |
| • Customer Type | • Babies | • Stays in Weekend Nights |
| • Deposit Type | • Booking Changes | • Total of Special Requests |
| • Distribution Channel | • Children | • Hotel Type |
| • Market Segment | • Company | |

### *B. Variables with Subcategories*

#### Hotel Type

| | |
|---|---|
| • City | • Resort |

#### Assigned Room Type / Reserved Room & Type: Code of Room Type Reserved / Assigned

| | | | | | |
|---|---|---|---|---|---|
| • A | • B | • C | • D | • E | • F |
| • G | • H | • I | • K | • L | • P |

#### Country

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ABW | BEL | CHL | DMA | GEO | IND | KNA | MDG | NCL | PRT | STP | UGA |
| AGO | BEN | CHN | DNK | GGY | IRL | KOR | MDV | NGA | PRY | SUR | UKR |
| AIA | BFA | CIV | DOM | GHA | IRN | KWT | MEX | NIC | PYF | SVK | UMI |
| ALB | BGD | CMR | DZA | GIB | IRQ | LAO | MKD | NLD | QAT | SVN | URY |
| AND | BGR | CN | ECU | GLP | ISL | LBN | MLI | NOR | ROU | SWE | USA |
| ARE | BHR | COL | EGY | GNB | ISR | LBY | MLT | NPL | RUS | SYC | UZB |
| ARG | BHS | COM | ESP | GRC | ITA | LCA | MMR | NZL | RWA | SYR | VEN |
| ARM | BIH | CPV | EST | GTM | JAM | LIE | MNE | OMN | SAU | TGO | VGB |
| ASM | BLR | CRI | ETH | GUY | JEY | LKA | MOZ | PAK | SDN | THA | VNM |
| ATA | BOL | CUB | FIN | HKG | JOR | LTU | MRT | PAN | SEN | TJK | ZAF |
| ATF | BRA | CYM | FJI | HND | JPN | LUX | MUS | PER | SGP | TMP | ZMB |
| AUS | BRB | CYP | FRA | HRV | KAZ | LVA | MWI | PHL | SLE | TUN | ZWE |
| AUT | BWA | CZE | FRO | HUN | KEN | MAC | MYS | PLW | SLV | TUR | |
| AZE | CAF | DEU | GAB | IDN | KHM | MAR | MYT | POL | SMR | TWN | |
| BDI | CHE | DJI | GBR | IMN | KIR | MCO | NAM | PRI | SRB | TZA | |

#### Customer Type

| | | | |
|---|---|---|---|
| • Contract | • Group | • Transient | • Transient-Party |

#### Deposit Type

| | | |
|---|---|---|
| • No Deposit | • Non Refund | • Refundable |

Distribution Channel

| | | | | |
|---|---|---|---|---|
| • Corporate | • Direct | • GDS | • TA/TO (Travel Agents/ Tour Operators) | • Undefined |

Market Segment

| | | | |
|---|---|---|---|
| • Aviation | • Complementary | • Corporate | • Direct |
| • Groups | • Offline TA/TO | • Online TA | • Undefined |

Meal

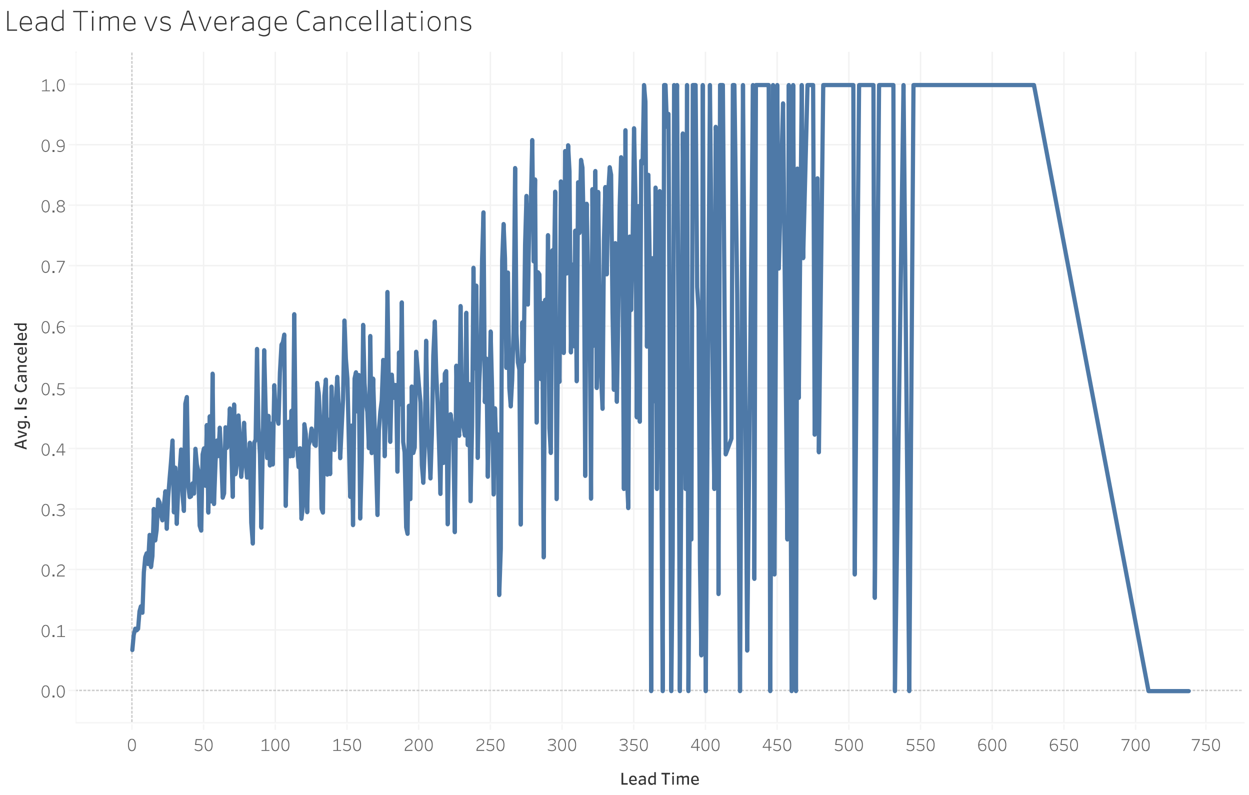
| | | | | |
|---|---|---|---|---|
| • BB (Bed & Breakfast) | • FB (Full board) | • HB (Half board) | • SC (no meal package) | • Undefined (no meal package) |

Reservation Status: Reservation last status

| | | |
|---|---|---|
| • Cancelled | • Check-Out | • No-Show |

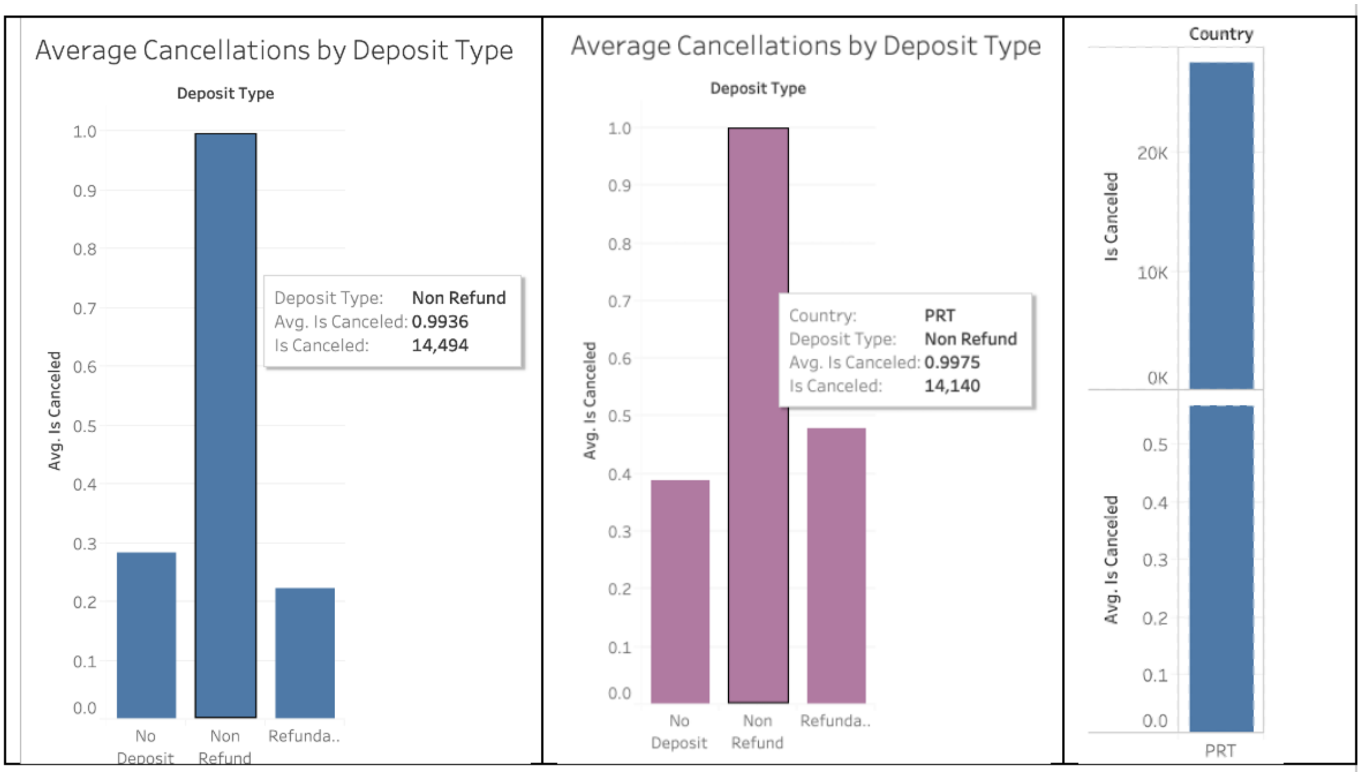
## Exhibit 2 – Tableau Observations

### (a)

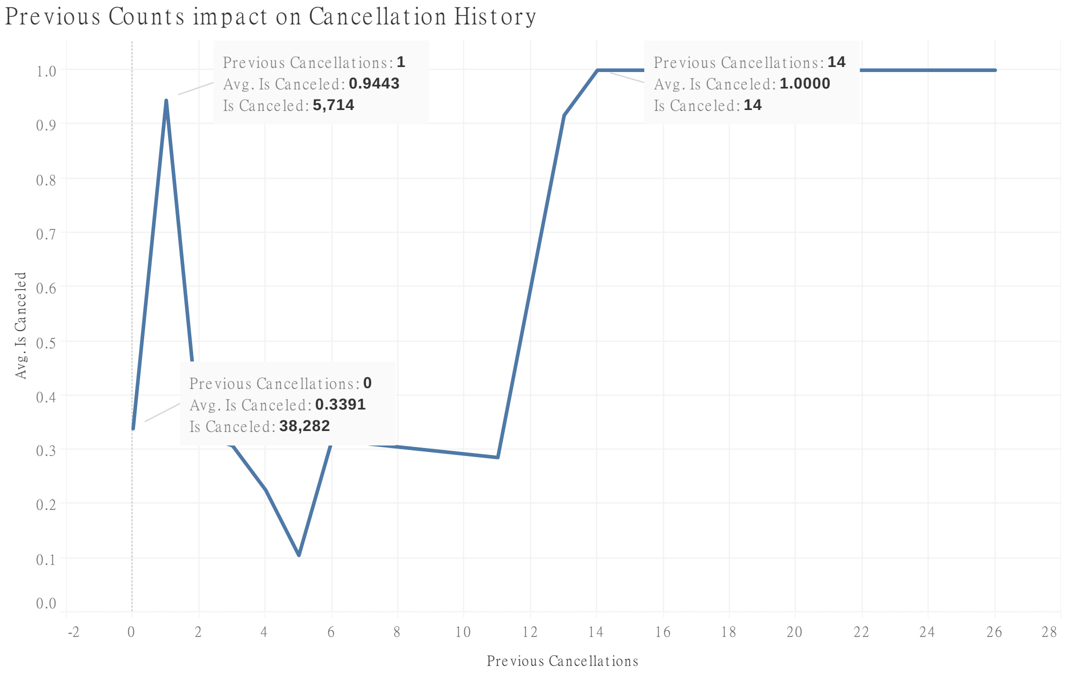


The trend of average of Is Canceled for Lead Time.

(b)



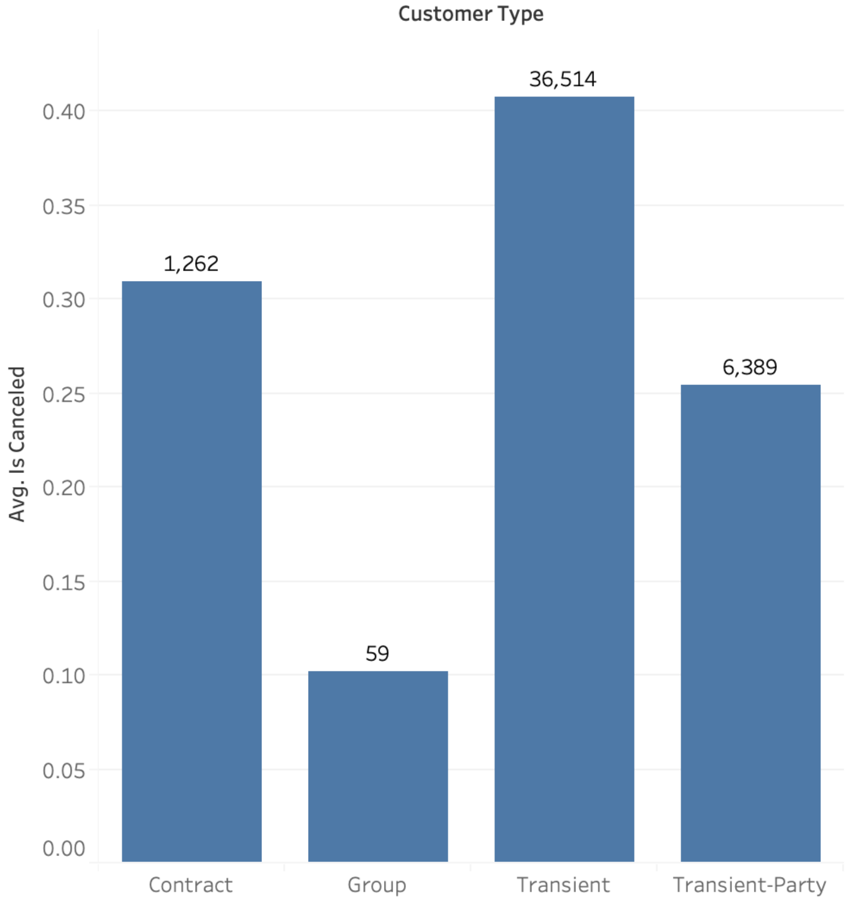
(c)



The trend of average of Is Canceled for Previous Cancellations.

**(d)**

Cancellations by Customer Type



## Exhibit 3 – Random Forest

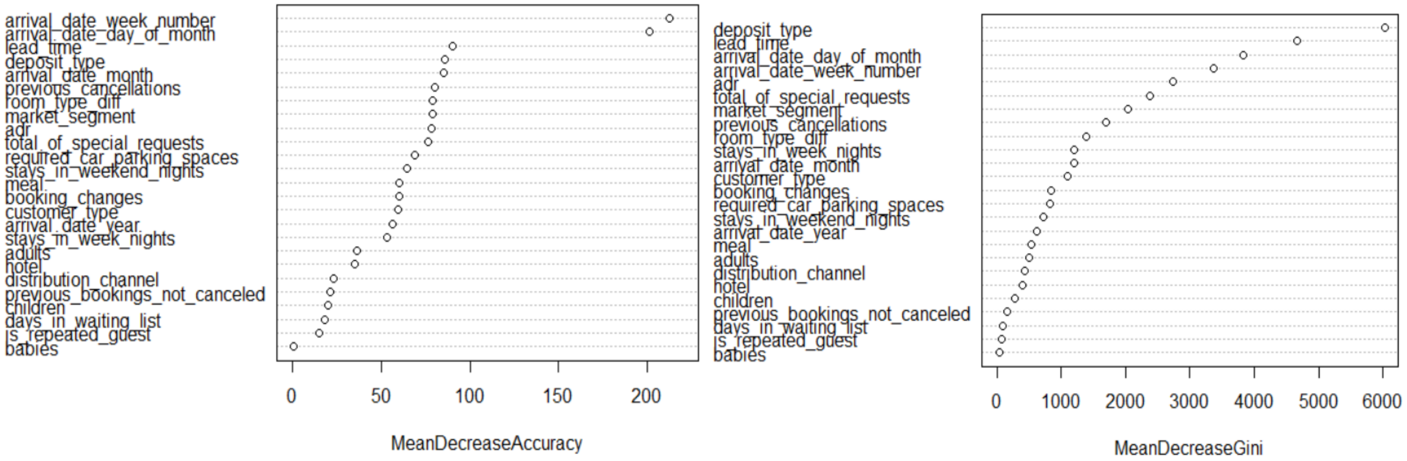
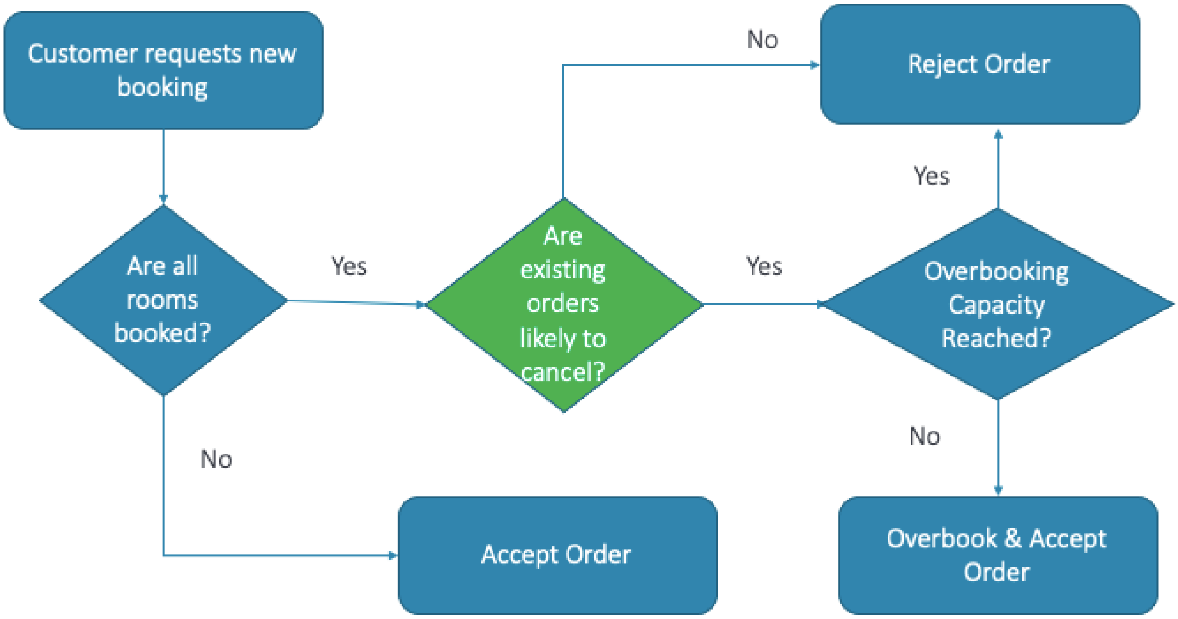


**Exhibit 4 – Model Deployment Use Case**


Customer requests new booking
Are all rooms booked?
Yes
No
Are existing orders likely to cancel?
No
Yes
Reject Order
Overbooking Capacity Reached?
Yes
No
Accept Order
Overbook & Accept Order

## Works Cited

[1] Mostipak, J. (2020). *Hotel Booking Demand*. Medium. https://www.kaggle.com/jessemostipak/hotel-booking-demand