**Team 10**
Antra Sinha
Saiyam Shah
Barnana Ganguly
Brandon Lewis
Aneerudh Ravishankar

Presentation for reference

# Booking Battles:
# Reservations vs Cancellations

## Table of Contents :

## 1. Business Problem

Hotel cancellations are a pressing concern within the hospitality sector. Our team delved into a European hotel booking and cancellation dataset, spanning July 2017 to January 2019, chosen for its relevance and comprehensiveness. This pre-COVID-19 period witnessed tourism as a major contributor to Europe's economy, with the sector directly adding 3.9% to the EU GDP in 2018 and providing 5.1% of total employment **(source)**.

Our objective is to shed light on cancellation trends and offer strategies to minimize them, benefiting both hoteliers and customers. Hotel professionals can use our insights for enhanced marketing and capacity planning, to increase revenues. Travelers can gain clarity on cancellation policies and loyalty schemes, influencing their choices.

## 2. Data Preprocessing

The Hotel Reservations dataset provides insight into determinants that influence the likelihood of a booking being canceled. The dataset contains a mix of features, capturing logistical aspects of bookings and historical guest behavior **(Appendix - Data)**.

2.1. **Class Imbalance Assessment:** The booking_status column revealed 66.75% 'Not_Canceled' and 33.25% 'Canceled' bookings. Given the relatively balanced distribution, techniques like SMOTE or oversampling aren't deemed necessary.

2.2. **Invalid Data Identification:** Observations with incorrect data (e.g.,February 29th on non-leap years, average price=0, etc.) were identified and removed to maintain the integrity of the dataset.

2.3. **Outlier Handling (Square root and Log transformation):** In our dataset, three variables – lead time, average price and the number of previous bookings not canceled – displayed pronounced outliers.
A square root transformation was initially employed for the lead time variable, normalizing its distribution. Subsequent logarithmic transformation on the other two variables reduced

average price outliers by approximately 60%. Yet, the "previous bookings not canceled" variable retained notable outliers even after this adjustment **(Appendix - Figure 2.1, Table 2.1 and Table 2.2)** Note that Outliers are determined as observations falling outside Q1-1.5IQR or beyond Q3+1.5IQR.

**Future Approach on Outliers:** Outliers were initially retained, opting to test prediction model performance. Given satisfactory results, no changes were made. However, if performance was affected by the outliers in the number of previous bookings not canceled variable, we would've considered outlier removal.

2.4. **Date Processing:** Columns for Arrival Date, Month, and Year were combined into a single datetime object for EDA. For modeling, features like Quarter, Week, Day of the Week, and Day of the Year were extracted from it.

2.5. **Categorical Variables:** Categorical variables underwent one-hot encoding to convert them to dummy variables.

## 3. EDA - Insights and Recommendations

Our EDA has highlighted key areas where interventions can lead to better occupancy rates, enhanced revenue, and improved guest experiences. The key insights and recommendations are as follows :

3.1. **Bulk Bookings and cancellations :** Our dataset often showed instances of bulk bookings with identical attributes, a trend that was mirrored in cancellations. This suggests that the hotel frequently accommodates larger events such as corporate gatherings and conferences **(Appendix - Figure 3.1)**.

***Recommendation - Operational Considerations***

The hotel should customize offerings and cancellation policies for large events, including event-specific packages and flexible cancellation terms for bulk bookings.

3.2. **Average Price for Booking cancellations-** Higher average booking prices are associated with more cancellations **(Appendix - Figure 3.2)**.

***Recommendation - Price Optimization***

To reduce cancellations, consider lowering booking prices and offering special promotions or loyalty programs to encourage non-cancellable bookings.

3.3. **Cancellation Trends Across Quarters -** Bookings made in Q4 have fewer cancellations, indicating customers plan vacations in advance during this period **(Appendix - Figure 3.3)**.

***Recommendation - Promote Early Booking Offers***

Promote early booking offers and incentives throughout the year to stabilize bookings and increase occupancy rates.

3.4. **First-Timers vs. Loyal Guests:** First-timers tend to cancel more (34%) than loyal guests (2%), despite being charged a higher rate **(Appendix - Table 3.4 and Figure 3.4)**.

***Recommendation - Deferred Discounts Strategy***

To reduce high cancellation rates among first-time guests and encourage loyalty, introduce deferred discounts. Offer discounts on subsequent stays to incentivize repeat bookings and potentially decrease initial booking cancellations.

3.5. **Overbooking :** Initial analysis shows a surge in online bookings, but about 50% are canceled, leading to potential revenue losses and operational complexities **(Appendix - Figure 3.5)**.

***Recommendation - Strategic Overbooking Approach***

To optimize occupancy and address cancellation rates, consider a strategic overbooking approach. Accept more reservations than rooms available, factoring in historical and predicted cancellations.

3.6. **Lead Time and Cancellation charges :** Room Type 1 represents the majority of bookings (~77%). Our study focussed on this room type to understand cancelation and revenue implications better **(Appendix - Table 3.6)**.

***Recommendation - Tiered cancelation Charges***

Implementing a tiered cancellation fee based on lead time quartiles could potentially recover €133k, which is approximately 15% of the total revenue lost to cancellations for Room Type 1.

3.7. **The Perks of Reserved Parking :** Guests with reserved parking showed significantly lower cancellation rates than those without ( 10% vs 34%). The price differential between them has decreased over time, converging by Q4 2018. This suggests that

the hotel may have provided incentives to those not opting for parking **(Appendix - Table 3.7 and Figure 3.7)**.

***Recommendation - Retaining Non-Parking Guests***

Introduce incentives for guests without reserved parking. By offering unique benefits to retain bookings, we can further reduce cancellations and foster loyalty.

## 4. Feature Engineering and Hyperparameter Tuning

For all models, we employed the Recursive Feature Elimination (RFE) technique to identify the most influential features for predictive modeling. RFE trained the model and discarded the least significant features, enhancing interpretability and mitigating overfitting risks. After trial and error, we selected 10 features **(Appendix - Modeling)**. Subsequently, we fine-tuned each model's performance using GridSearchCV, searching through hyperparameter combinations to optimize their predictive power **(Appendix - Hyperparameter Tuning)**.

## 5. Modeling

We progressed from simpler models, Logistic Regression, to the more complex models, Random Forest (bagging) and XGBoost (boosting). The model results are as follows :

| Model | Accuracy | Precision | Recall | AUC |
|---|---|---|---|---|
| **Logistic Regression** | 74% | 63% | 48% | 0.72 |
| **Random Forest** | 89% | 87% | 78% | 0.92 |
| **XGBoost** | 84% | 77% | 77% | 0.90 |

## 6. Conclusion

Accuracy measures the percentage of correct predictions made by the model while precision represents how many of the predicted cancellations were actually cancellations. Recall demonstrates how many of the actual cancellations were correctly identified by the model. On comparing all three models based on these metrics, we found that the Random Forest model outperformed the others, with the XGBoost model being a close second.

## 7. Appendix

### Data Preprocessing

Key variables in the Hotel Reservations dataset :

- **Booking_ID:** A unique identifier for each booking
- **no_of_adults:** The number of adults included in the booking
- **no_of_children:** The number of children included in the booking
- **no_of_weekend_nights:** The number of nights that fall on a weekend in the booking
- **no_of_week_nights:** The number of nights that fall on weekdays in the booking
- **type_of_meal_plan:** The type of meal plan selected for the booking
- **required_car_parking_space:** Indicates whether a car parking space was required for the booking (1 for required, 0 for not required)
- **room_type_reserved:** The type of room reserved for the booking
- **lead_time:** The number of days between the booking date and the arrival date
- **arrival_year:** The year of the booking's arrival date
- **arrival_month:** The month of the booking's arrival date
- **arrival_date:** The day of the booking's arrival date
- **market_segment_type:** The type of market segment the booking falls under (e.g., "Online," "Offline")
- **repeated_guest:** Indicates whether the guest is a repeated guest (1 for yes, 0 for no)
- **no_of_previous_cancellations:** The number of previous cancellations by the guest
- **no_of_previous_bookings_not_canceled:** The number of previous bookings by the guest that were not canceled
- **avg_price_per_room:** The average price per room for the booking
- **no_of_special_requests:** The number of special requests made for the booking

- **booking_status:** The status of the booking ("Not_Canceled" or "Canceled")

**Figure 2.1 : Outlier box plots**

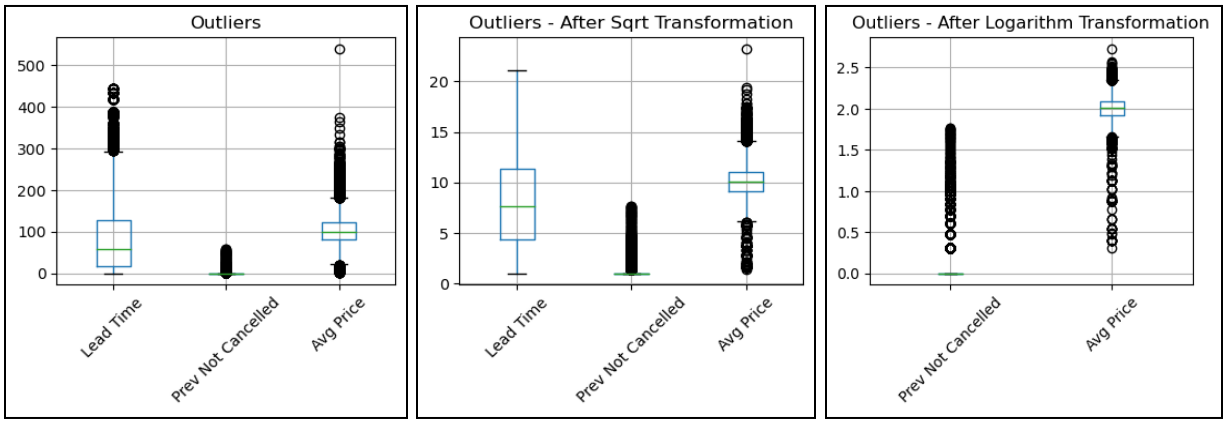


**Table 2.1: Outliers - Sqrt Transformation**

| Feature | #Outliers | # Outliers (After Sqrt Transformation) |
|---|---|---|
| **Lead Time** | 1184 | **0** |
| **No of Prev bookings not canceled** | 690 | 690 |
| **Average Price per Room** | 1101 | 711 |

**Table 2.2 : Outliers - Log Transformation**

| Feature | # Outliers | # Outliers (After Log Transformation) |
|---|---|---|
| **No of Prev bookings not canceled** | 690 | 690 |
| **Average Price per Room** | 1101 | **451 (~60% decrease from 1101)** |

Outliers are identified as observations that lie below Q1 minus 1.5 times the interquartile range (IQR) or above Q3 plus 1.5 times the IQR.

## Exploratory Data Analysis

**Figure 3.1 : Bulk bookings and Cancellations**

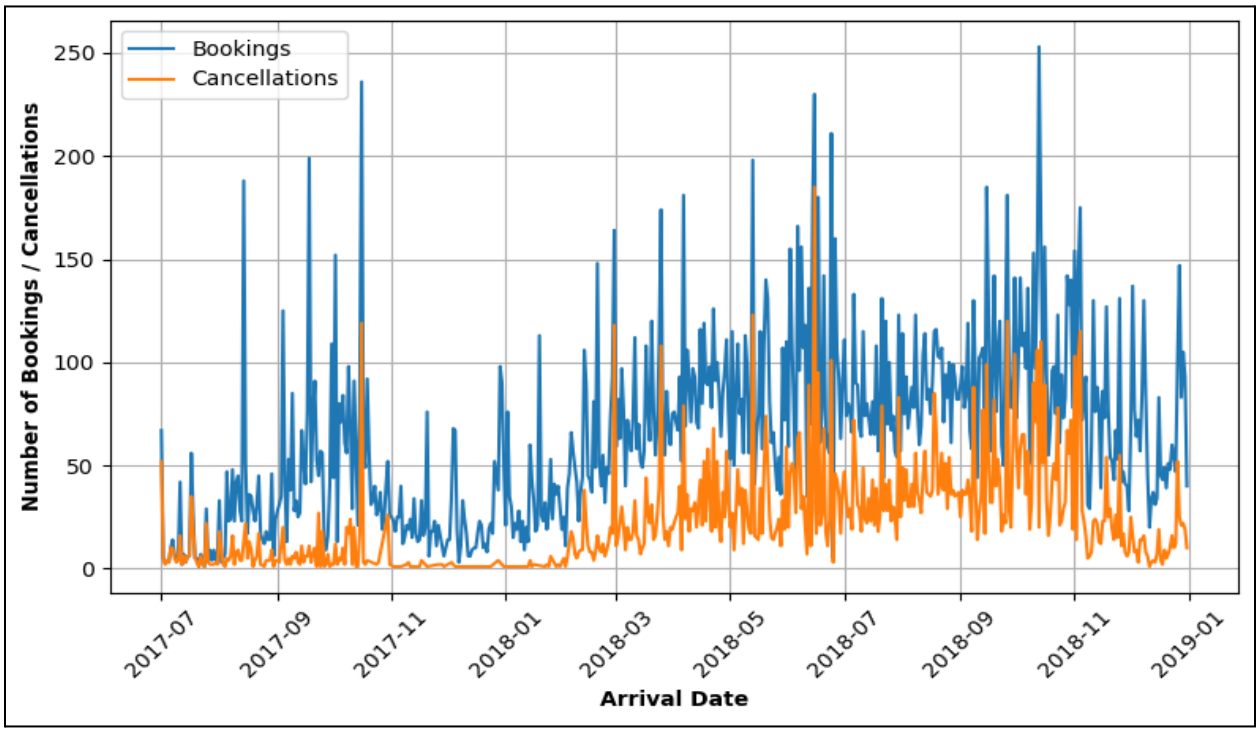


**Figure 3.2 : Average Price for Bookings and Cancellations**

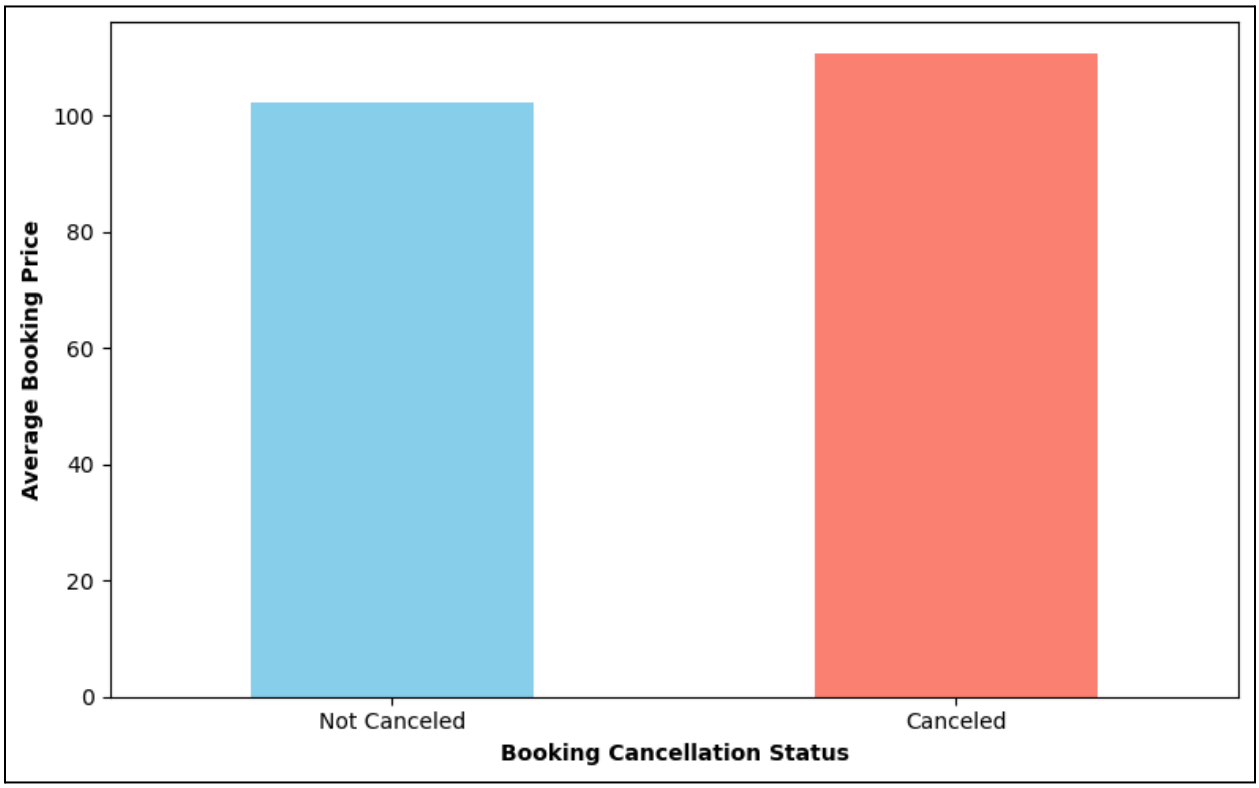

**Figure 3.3 : Canceled and Not Canceled Bookings by Quarter**

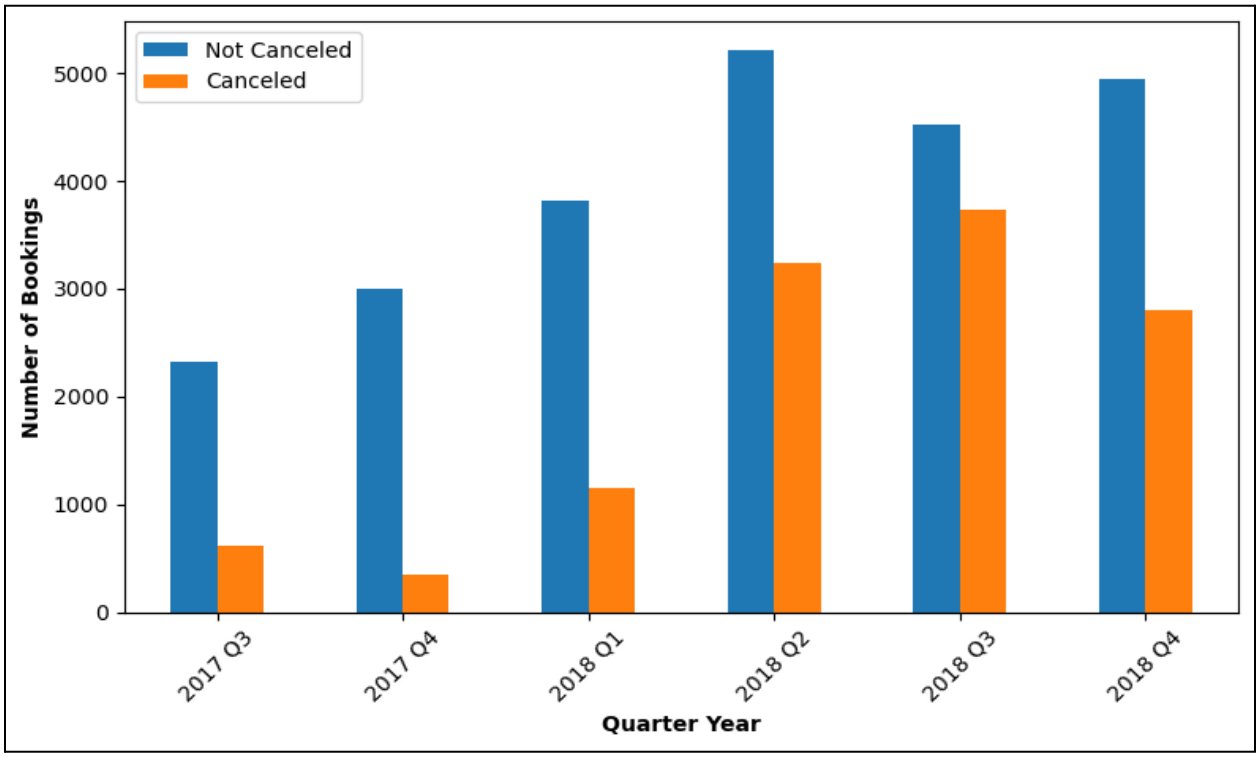


**Figure 3.4: First timers vs Loyal Guests**

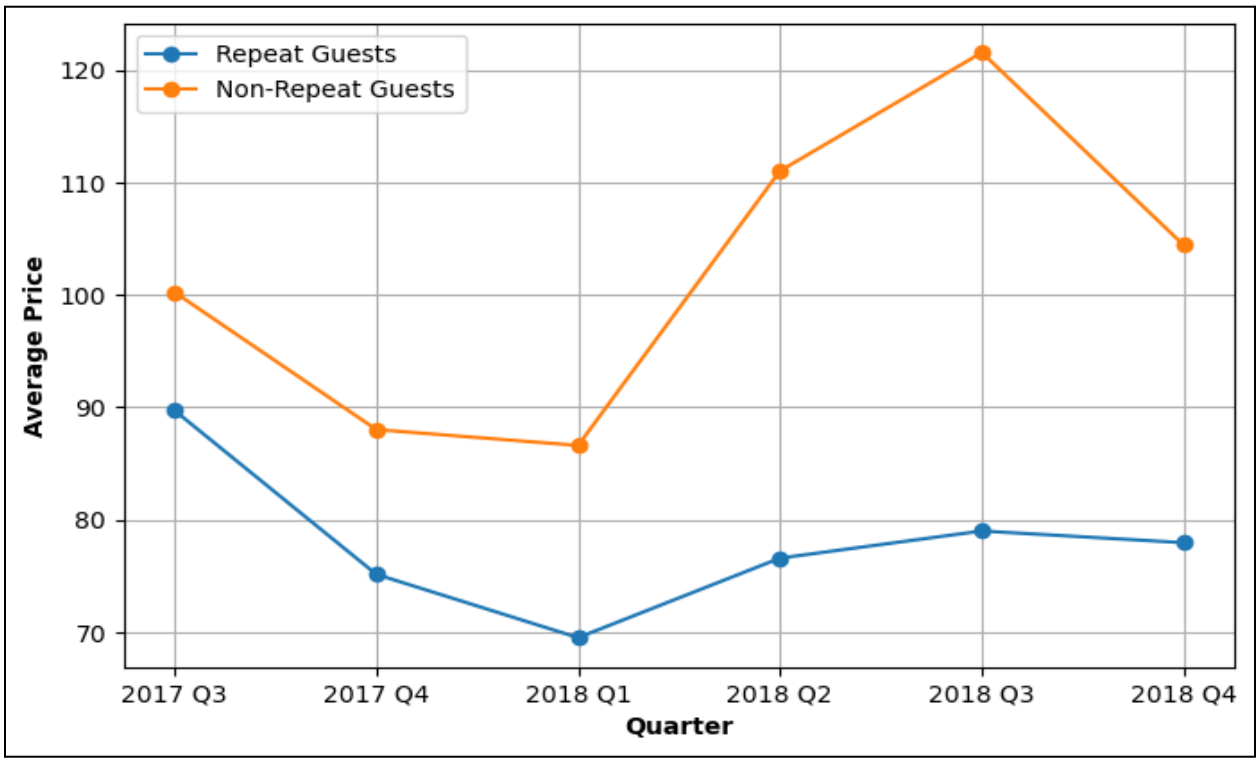

**Table 3.4: Cancellation Rates by Guest Type**

| Guest Type | Total bookings | % Canceled Bookings | % Non-Canceled Bookings |
|---|---|---|---|
| **First Timers** | 34,894 | 34% | 66% |
| **Loyal Guests** | 799 | 2% | 98% |

**Figure 3.5: Canceled and Non-canceled Bookings by Market Segment and Quarter**

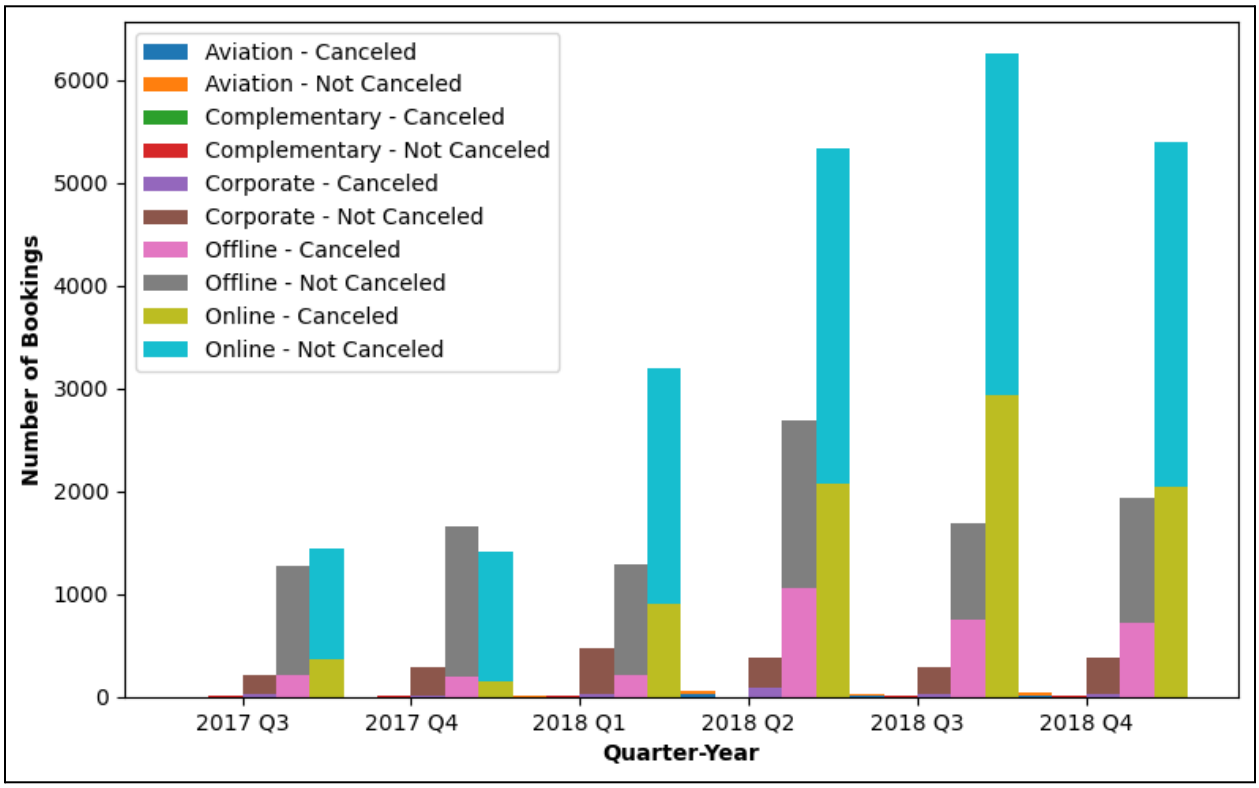

**Table 3.6: Cancellation charges and Lead Time**

| Lead Time Quartile | Proposed cancelation Charge | Number of cancellations | Recouped amount (in Euros) |
|---|---|---|---|
| 0.25 | 30% | 758 | 24,003 (18%) |
| 0.50 | 20% | 1,329 | 28,174 (21%) |
| 0.75 | 15% | 2,147 | 31,249 (24%) |
| 1 | 10% | 4,826 | 49,299 (37%) |
| **Total** | | **9,060** | **132,725** |

**Figure 3.7: Average Price by Quarter (Parking vs No Parking)**

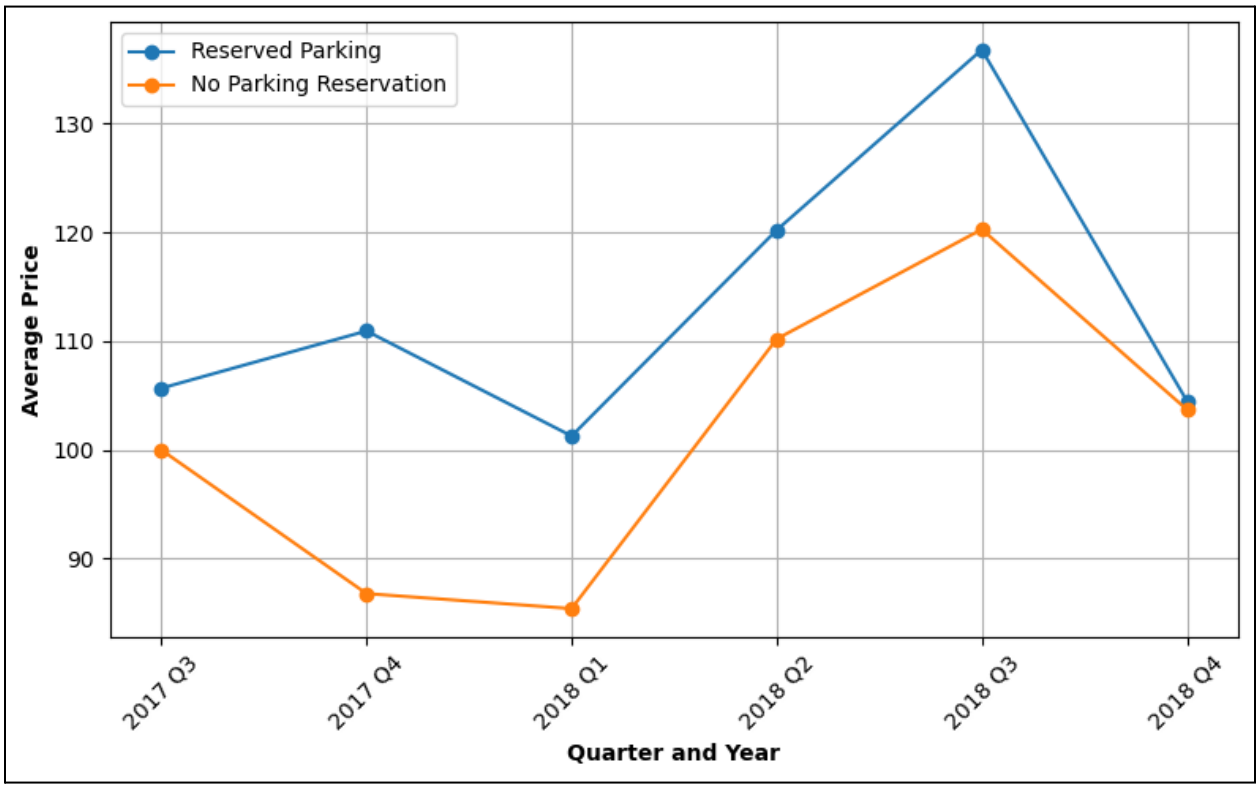

**Table 3.7: Cancellation Rates (Parking vs No Parking)**

| Guest Type | cancelation % |
|---|---|
| **Reserved Parking** | 10% |
| **Not Reserved Parking** | 34% |

## Hyperparameter Tuning

Best Parameters chosen -

1. **Logistic regression** :
    - L2 (Ridge Regression) with lambda = 12

2. **Random Forest**:
    - max_depth: None
    - Min_samples_leaf: 1
    - min_samples_split: 2
    - n_estimators: 100

3. **XGBoost :**
    - gamma: 0.1
    - learning_rate: 0.1
    - max_depth: 5
    - n_estimators: 300

## Model

**Logistic Regression:** Features selected- number of children, car parking requirements, guest repetition status, special requests, previous bookings (log), specific meal plans, room types, and market segment types

**Random Forest:** Features selected- number of weekend nights, number of week nights, special requests, lead time (square root transformed), average price per room (log transformed), booking day, week number, day of the week, day of the year, and market segment type (Online)

**XGBoost:** Features selected- number of adults, required car parking space, repeated guest status, special requests, lead time (square root

transformed), average price per room (log transformed), year, month, and market segment types (Offline and Online)