HOTEL BOOKING CANCELLATIONS ANALYSIS

Scott Fields,

Ishan Patel, &

Bhavana Reddy

Project Goals

Problem Description

In reservation-based industries, an accurate booking cancellation forecast is of foremost importance to estimate demand.

The cancellation rate for bookings in the hospitality industry is quite high. Almost 3 in 7 bookings are canceled daily. Once the reservation has been canceled, there is almost nothing to be done. This causes discomfort for these institutions and creates the need to take precautions.

Motivation

Consumers are always looking for ways to minimize their cost of buying something. When they find out that they can buy the same thing at a lower price than they paid for, they would attempt to cancel and repurchase, and that's what usually happens with hotel bookings.

A huge loss of income occurs because of unsold rooms due to last minute cancellations. Every day in the U.S alone, more than 221,000 hotel bookings are canceled, resulting in a market which sees $8.6 billion wasted annually.

Revenue per available room (RevPAR) is lower when the revenue management is done wrongly and when a canceled room is sold cheaper at the last minute. When a hotel is faced with a last-minute cancellation and then a last-minute check-in, the hotel can't do much but sell the room at a much lower price, so there goes their opportunity to earn a higher RevPAR.

Further, the COVID 19 pandemic has sent shockwaves of disruptions to travel plans worldwide. The global hospitality industry is overwhelmed by the large number of cancellations spurred by the virus. It's better to have 70% of cancellations and knowing about them in good time than 20% of cancellations that are notified at the last minute. Therefore, predicting reservations that can be canceled and preventing these cancellations will create a surplus value for the institutions.

Data Handling

Data set

The Data set consists of 119,390 rows and 32 columns. Out of the 32 columns, 12 columns have categorical values, and the remaining 20 columns have continuous values.

The response variable is 'is_canceled', which is a categorical variable indicating if the hotel booking was canceled (1) or not canceled (0). There are 32 predictor variables which include hotel type, lead time between hotel booking and arrival date, deposit type, market segment through which the hotel booking was performed, total number of guests, etc. [See Data schema]

Data Cleanup

Upon an initial look into the dataset, there seemed to be 4 categories that contained NaN values: children, country, and company. The column 'children' consisted of 4 NaN values, which were replaced with the value '0'. The column 'country' consisted of 488 NaN values, which were replaced with the country code 'PRT' since it was the highest occurring country code (40% of the countries were 'PRT'). The column 'company' consisted of 112,593 null values, and hence was removed from the data set for this analysis.

The columns 'children' and 'babies' were summed up and combined into a column called 'total_children' since there was no information about which age was the cut-off to classify the guest as a child or a baby.

Exploratory Data Analysis

Out of the 119,390 records present in the data set, 75,166 (63%) of the reservations were not canceled, and 44,224 (37%) of the reservations were canceled.

Country vs. Cancellations

Portugal (PRT) has the highest number of cancellations. This was expected since hotel reservations from PRT were 41% of the reservations from all countries. However, even though Hong Kong (HKG), Tajikistan (TJK), UAE (ARE), France (FRO), Bahrain (BHR) and Maldives (MDV), have fewer bookings with the hotel, their cancellations rates are high. On average, 85% of the bookings from these countries were canceled. [See figure]

Month vs. Cancellations

The month of August seems to have the most reservations and therefore the most cancellations with around 53% of the reservations being canceled in August. The cancellations drop off from October to November, with a spike coming towards December as the holidays roll near. [See figure]

Lead time vs. Cancellations

Most cancellations happen more than 20 days before the booking date, with the peak around 2 months in advance. When the booking are made less than 20 days in advance, there are usually no cancellations as people are desperate. [See figure]

Room rates vs. Cancellations

Except for room types L and A, the overall trend is that cancelled room rates are higher than non-cancelled room rates. [See figure]

Solutions

For this classification problem, we have looked at three different models : Logistic Regression, K-Nearest Neighbors, and Random Forest with XGBoost.

Logistic Regression

Using the 'LogisticRegressor' function from the 'sklearn' library, this model provided an 80.38% accuracy. It was determined that 'country_PRT' and 'deposit_type_Non_Refundable' were the two biggest factors that contributed to cancellations.

However, since the data was 40% "country_PRT," this can be seen as a skewing variable. "deposit_type_Non_Refundable," can be interpreted as the most important factor at a 1.73 importance value. [See figure]

K-Nearest Neighbors

Using the 'KNeighborsClassifier' function from the 'sklearn' library, this model provided an 82.30% accuracy with an n value of 15. For the purpose of knn, the categorical variables were encoded using the 'LabelEncoder' function from the 'sklearn' library.

As per the correlation matrix, the two most important factors that contributed to cancellations are 'deposit_type' (correlation factor: 0.47) and 'lead_time' (correlation factor: 0.29). [See figure]

Random Forest with XGBoost

Using the "RandomForestClassifer" function in the "sklearn" library, this model provided an accuracy of 88.20%. It was determined that "lead_time" and "adr" were the two biggest factors that contributed to cancellations. [See figure]

This outcome is coherent with the EDA as higher lead times and higher adjusted daily rates are correlated to more cancellations. The adjusted daily rate (adr) is the second most important factor, which by intuition would make sense since, when it comes to hotel rooms, rate is one of the most looked at factors. If a rate is too high, there is a high chance that a customer will search for lower prices at other competitors, and ultimately cancel this reservation.

One factor to consider is "previous_cancellations". This is interesting because it would be assumed that cancelling at the same hotel before would be independent of cancelling at the same hotel again, however, we can see that this is not true. This could be because people who have made cancellations before are more willing to cancel than those who have not or simply because these people, like the "adr" theory, booked the hotel every time as a failsafe and cancelled once a better option was found, thus putting them in "previous_cancellations" category.

The final model we used was an XGBoost. After using cross validation to find the optimal number of trees, the model produced an out-of-sample accuracy of 91.4%. Similarly, to random forest, the most important variables were lead_time and adr. [See figure]

Insights

One way to mitigate cancellations and maximize RevPAR is to allow overbooking. For example, removing "Non-Refundable Deposits" allows customers to book hassle-free. This would ensure that, even if the hotel is overbooked, there will be cancellations inevitably, ensuring that even with these cancellations, the hotel is at maximum capacity.

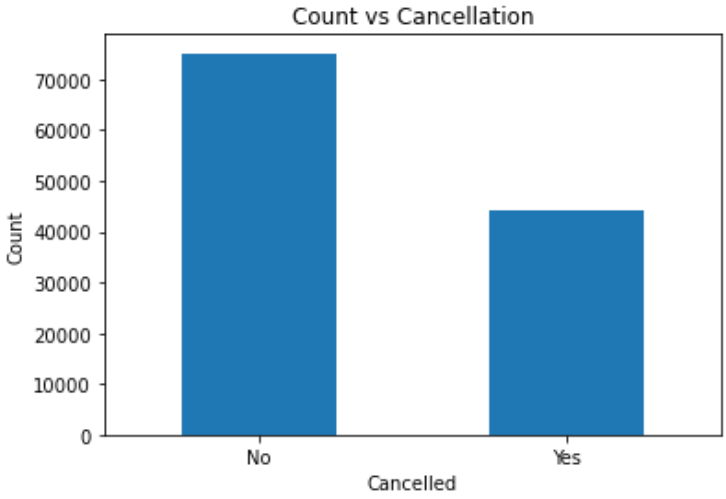
Allowing overbooking during certain months, such as August (which has the highest cancellations), also ensures maximum RevPAR. Other methods to increase the RevPAR would be re-evaluating daily rates and limiting the booking lead time to about 25-30 days.

Data Schema

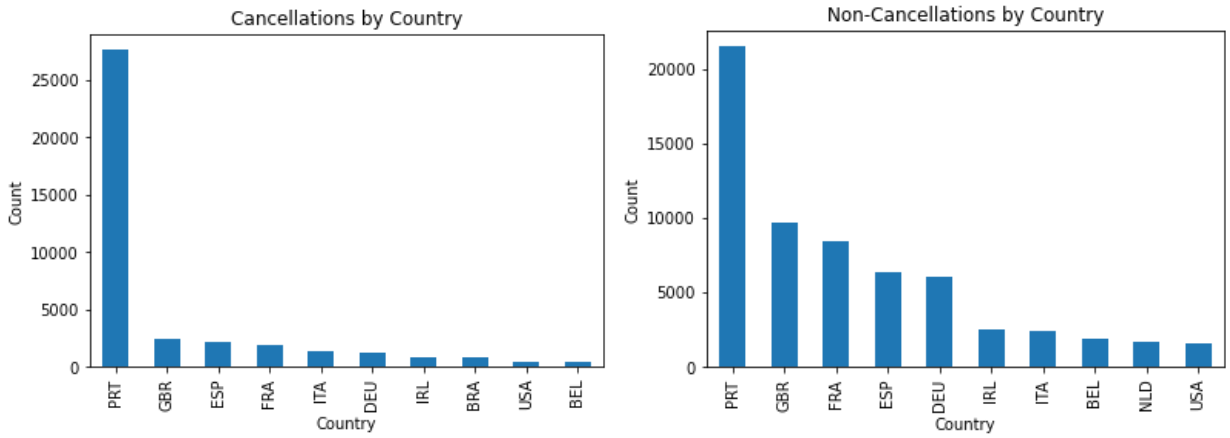
Variable	Variable Source	Description
hotel	Independent	Type of Hotel (Resort or City)
is_canceled	Independent	Reservation was canceled (1) or not canceled (0)
lead_time	Independent	Time between hotel booking and actual arrival date
arrival_date_year	Independent	Year of guest arrival date
arrival_date_month	Independent	Month of guest arrival date
arrival_date_day_of_month	Independent	Day of month of guest arrival date
stays_in_weekend_nights	Independent	Number of weekend nights in duration of stay
stays_in_week_nights	Independent	Number of weeknights in duration of stay
adults	Independent	Number of adult guests
children	Independent	Number of children guests
babies	Independent	Number of baby guests
total_children	Derived	Sum of children and baby guests
meal	Independent	Type of mean selected during booking
country	Independent	Country of origin of guest
market_segment	Independent	Market segment through which booking was made
distribution_channel	Independent	Distribution channel through which booking was made
is_repeated_guest	Independent	Has the guest stayed in hotel before (1) or no (0)
previous_cancellations	Independent	Has the guest made previous cancellations in the hotel (1) or no (0)
previous_bookings_not_canceled	Independent	Has the guest not made any previous cancellations in the hotel (1) or no (0)
reserved_room_type	Independent	Room type reserved during booking
assigned_room_type	Independent	Room type assigned during check-in
booking_changes	Independent	Changed made in the booking
deposit_type	Independent	Deposit type during booking
agent	Independent	Agent through which booking was done
company	Independent	Company through which booking was done
days_in_waiting_list	Independent	Days in waiting list before booking completion
customer_type	Independent	Hotel guest type
adr	Independent	Adjusted daily rate
required_car_parking_spaces	Independent	Number of required car parking spaces for guests
total_of_special_requests	Independent	Number of special requests made by guests
reservation_status	Independent	Reservation status of booking
reservation_date	Independent	Reservation date of booking

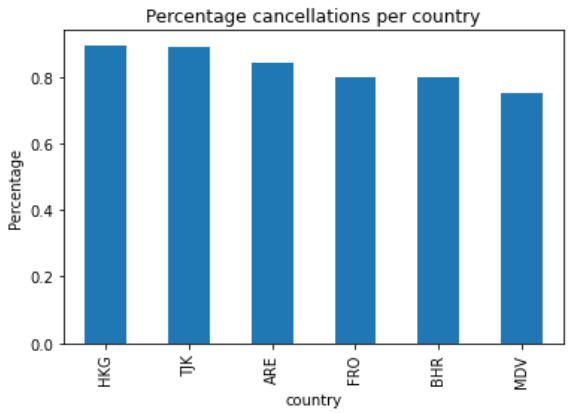
Graphs and Tables

Count vs. Cancellation

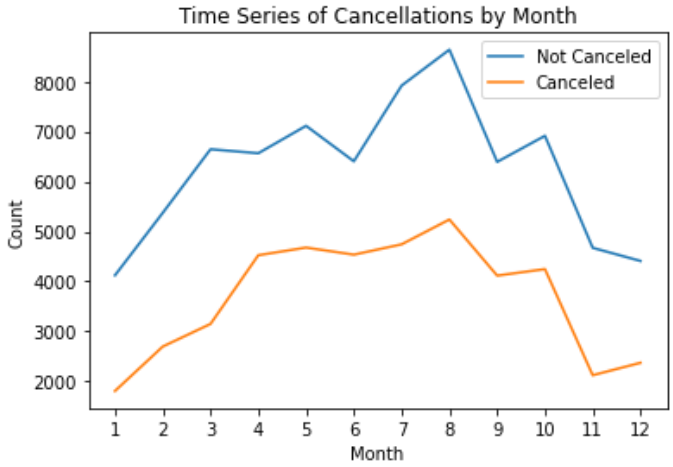


Country vs. Cancellation

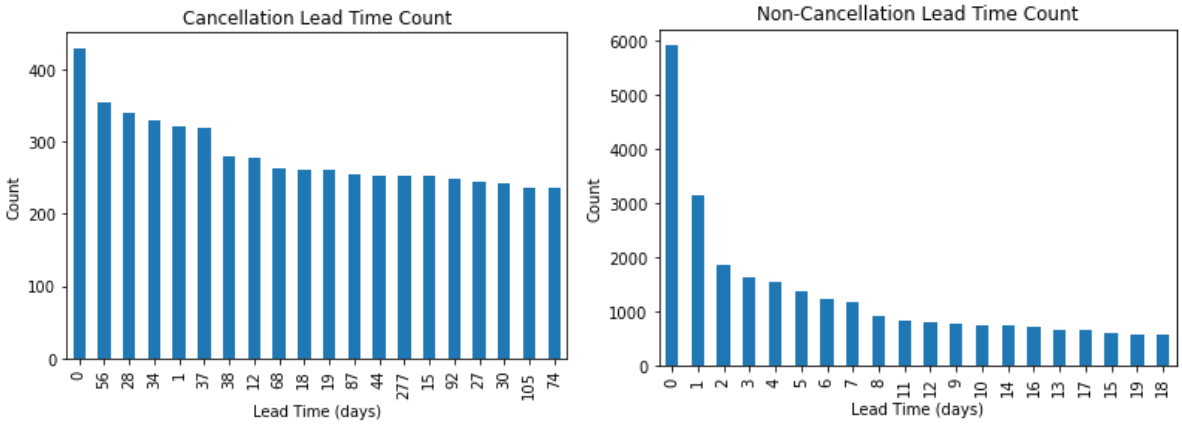




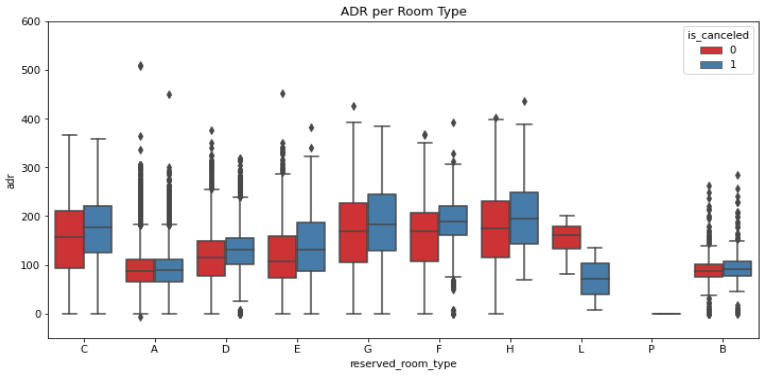
Month vs. Cancellations



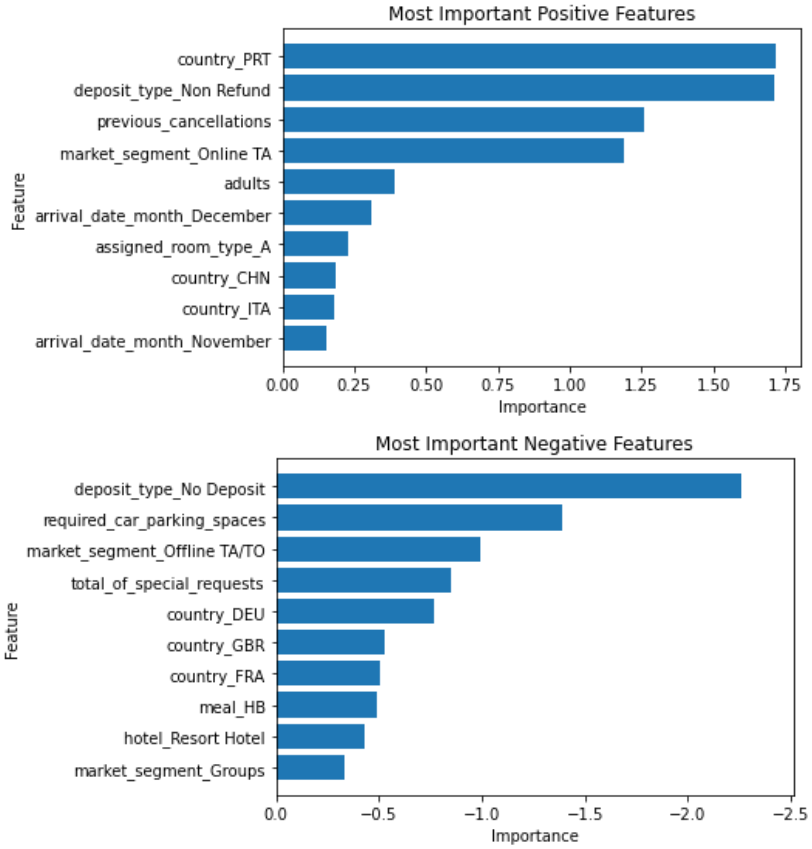
Lead time vs. Cancellation



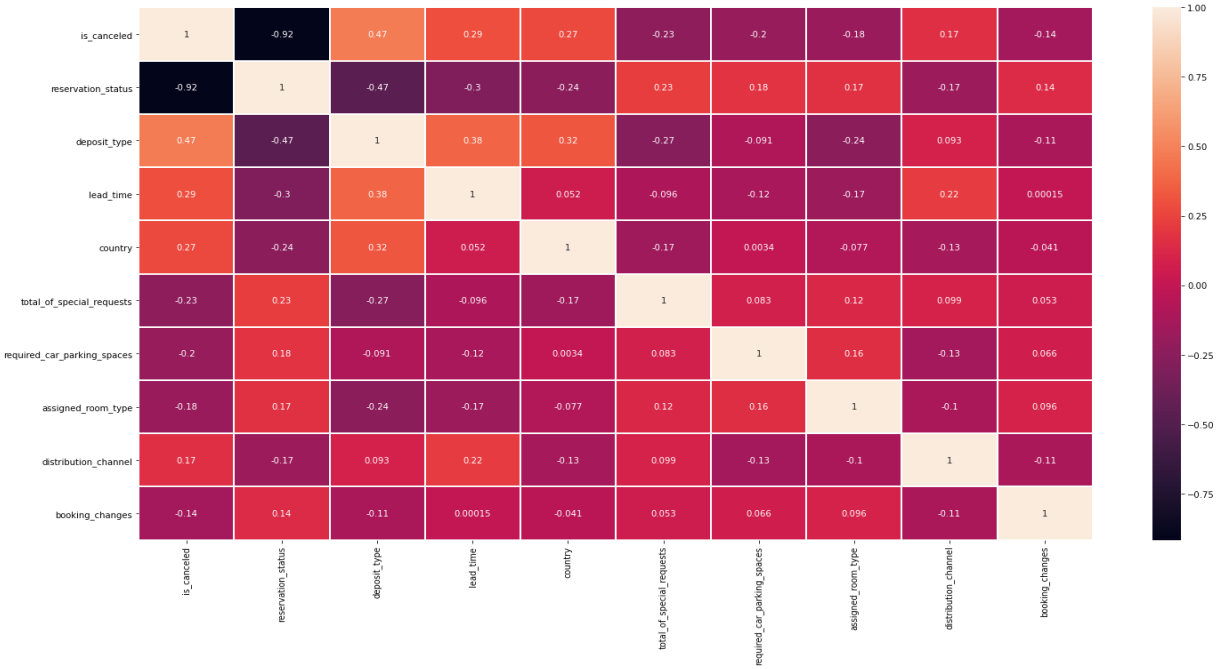
Room rates vs. Cancellations



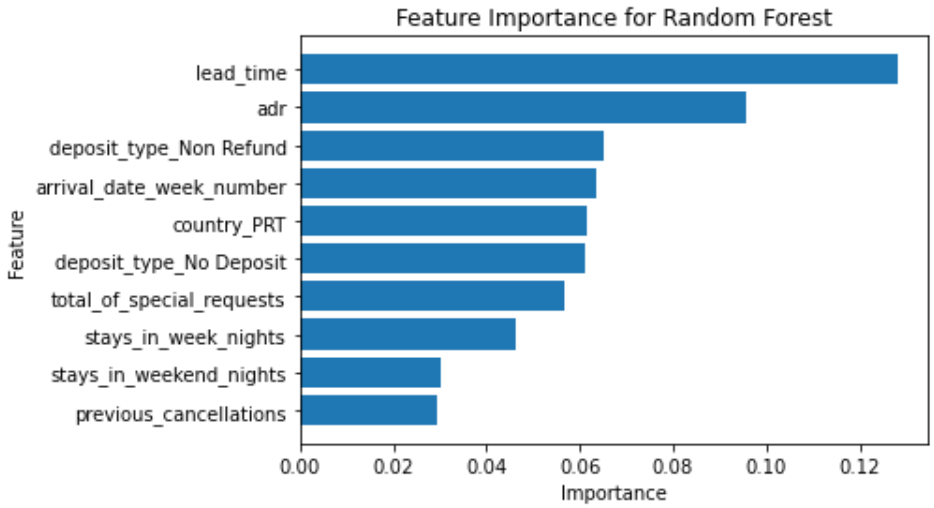
Logistic Regression



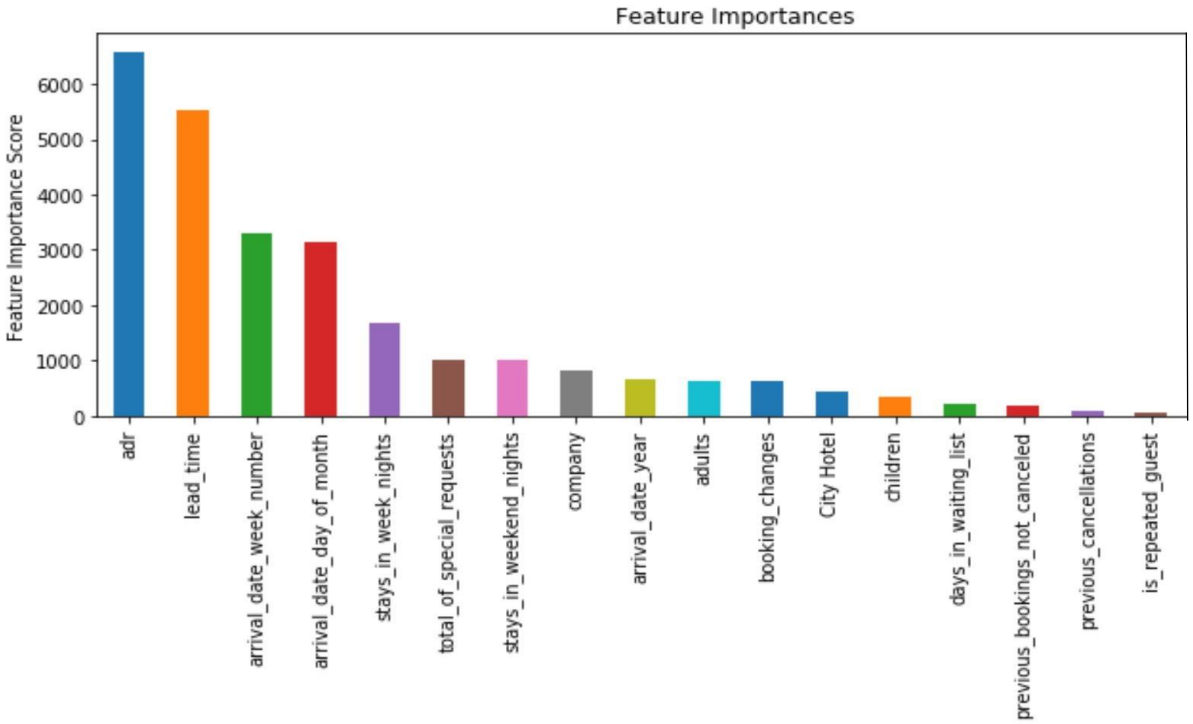
K Nearest Neighbors



Random Forest



XGBoost



Model Summary

Model	Accuracy	Most important factor
Logistic Regression	80.38%	Non-Refund Deposit
KNN	82.30%	Deposit Type
Random Forest	88.20%	Lead Time
XGBoost	91.40%	ADR