Predicting Hotel Booking Cancellations

Matt Mascarelli

Context

       Cancellations can profoundly affect revenue within the hospitality industry. To address this issue, numerous hotels employ cancellation policies or utilize overbooking strategies as a means of minimizing the impact. However, these measures can sometimes lead to adverse effects on both revenue and the hotel's reputation. By developing a machine learning model capable of predicting the likelihood of booking cancellations, hotels can enhance their strategic planning and employ more effective techniques to mitigate the financial impact caused by cancellations. By identifying bookings with a higher probability of cancellation, hotels can target them with enticing incentives, such as complimentary meals, extended stays, or other services provided by the hotel, in order to encourage guests to retain their bookings.

Problem Statement

How can a hotel reduce their revenue loss by 20% for next year by targeting bookings that are more likely to cancel with incentives to retain their reservation?

Data Wrangling

       The data is from ScienceDirect, but was originally sourced from Property Management System SQL databases. Included in the data are 31 columns from two different hotels in Portugal. Hotel 1 has 40,060 observations and hotel 2 has 79,330, where each observation represents a hotel booking. The bookings span from 2015 until 2017. The target column 'IsCanceled' tells us if the reservation was canceled (1) or not (0).

Reference: https://www.sciencedirect.com/science/article/pii/S2352340918315191#bib4

Features:

- `Hotel`: Identifies which hotel the observations belongs to (h1:Resort Hotel, h2: City Hotel)
- `IsCanceled`: Target Variable. Indicates if booking was canceled (1) or not (0)
- `LeadTime` : Number of days that elapsed between the booking data and the arrival date
- `ArrivalDateYear` : Year of arrival date
- `ArrivalDateMonth` : Month of arrival date
- `ArrivalDateWeekNumber` : Week of arrival date
- `ArrivalDateDayOfMonth` : Day of month for arrival date
- `StaysInWeekendNights` : Number of weekend nights for booking
- `StaysInWeekNights` : Number of week nights for booking
- `Adults`: Number of adults
- `Children` : Number of children
- `Babies` : Number of babies
- `Meal` : Type of meal package
  - Undefined/SC – no meal package;
  - BB – Bed & Breakfast;
  - HB – Half board (breakfast and one other meal – usually dinner);
  - FB – Full board (breakfast, lunch and dinner)
- `Country` : Country of origin. Categories are represented in the ISO 3155–3:2013 format
  - i.e. 'IRL' is Ireland, 'ESP' is Spain, etc
- `MarketSegment` : Market segment designation. In categories, the term "TA" means "Travel Agents" and "TO" means "Tour Operators"
- `DistributionChannel` : Booking distribution channel. The term "TA" means "Travel Agents" and "TO" means "Tour Operators"
- `IsRepeatedGuest` : Value indicating if the booking name was from a repeated guest (1) or not (0)
- `PreviousCancellations` : Number of previous bookings that were canceled by the customer prior to the current booking
- `PreviousBookingsNotCanceled` : Number of previous bookings not canceled by the customer prior to the current booking
- `ReservedRoomType` : Code of room type reserved. Code is presented instead of designation for anonymity reasons
- `AssignedRoomType` : Code for the type of room assigned to the booking. Sometimes the assigned room type differs from the reserved room type due to hotel operation reasons (e.g. overbooking) or by customer request. Code is presented instead of designation for anonymity reasons

2

- `BookingChanges` : Number of changes/amendments made to the booking from the moment the booking was entered until the moment of check-in or cancellation
- `DepositType` : Indication on if the customer made a deposit to guarantee the booking. This variable can assume three categories:
    - No Deposit – no deposit was made
    - Non Refund – a deposit was made in the value of the total stay cost
    - Refundable – a deposit was made with a value under the total cost of stay.
- `Agent` : ID of the travel agency that made the booking
- `Company` : ID of the company/entity that made the booking or responsible for paying the booking. ID is presented instead of designation for anonymity reasons
- `DaysInWaitingList` : Number of days the booking was in the waiting list before it was confirmed to the customer
- `CustomerType` : Type of booking, assuming one of four categories:
    - Contract - when the booking has an allotment or other type of contract associated to it;
    - Group – when the booking is associated to a group;
    - Transient – when the booking is not part of a group or contract, and is not associated to other transient booking;
    - Transient-party – when the booking is transient, but is associated to at least other transient booking
- `ADR` : Average Daily Rate
- `RequiredCarParkingSpaces` : Number of car parking spaces required by the customer
- `TotalOfSpecialRequests` : Number of special requests made by the customer (e.g. twin bed or high floor)
- `ReservationStatus` : Reservation last status, assuming one of three categories:
    - Canceled – booking was canceled by the customer;
    - Check-Out – customer has checked in but already departed;
    - No-Show – customer did not check-in and did inform the hotel of the reason why
- `ReservationStatusDate` : Date at which the last status was set. This variable can be used in conjunction with the ReservationStatus to understand when was the booking canceled or when did the customer checked-out of the hotel

Missing Values:

➔ Children: 4 rows replaced with 0
➔ Country: 488 rows replaced with 'unknown'

Duplicate Rows:

➔ **31994** duplicate rows were removed

More rows removed:

➔ **166** rows have 0 adults and 0 children
➔ **591** rows have 0 weeknights and 0 weekend nights
➔ **432** rows have 0 ADR (these exclude bookings grouped as complementary)

After data wrangling and cleaning, we are left with **86207 rows** and **34 features**.

Exploratory Data Analysis

Numerical Data:

	count	mean	std	min	25%	50%	75%	max
IsCanceled	87396.0	0.274898	0.446466	0.00	0.0	0.0	1.0	1.0
LeadTime	87396.0	79.891368	86.052325	0.00	11.0	49.0	125.0	737.0
ArrivalDateYear	87396.0	2016.210296	0.686102	2015.00	2016.0	2016.0	2017.0	2017.0
ArrivalDateWeekNumber	87396.0	26.838334	13.674572	1.00	16.0	27.0	37.0	53.0
ArrivalDateDayOfMonth	87396.0	15.815541	8.835146	1.00	8.0	16.0	23.0	31.0
StaysInWeekendNights	87396.0	1.005263	1.031921	0.00	0.0	1.0	2.0	19.0
StaysInWeekNights	87396.0	2.625395	2.053584	0.00	1.0	2.0	4.0	50.0
Adults	87396.0	1.875795	0.626500	0.00	2.0	2.0	2.0	55.0
Children	87396.0	0.138633	0.455871	0.00	0.0	0.0	0.0	10.0
Babies	87396.0	0.010824	0.113597	0.00	0.0	0.0	0.0	10.0
IsRepeatedGuest	87396.0	0.039075	0.193775	0.00	0.0	0.0	0.0	1.0
PreviousCancellations	87396.0	0.030413	0.369145	0.00	0.0	0.0	0.0	26.0
PreviousBookingsNotCanceled	87396.0	0.183990	1.731894	0.00	0.0	0.0	0.0	72.0
BookingChanges	87396.0	0.271603	0.727245	0.00	0.0	0.0	0.0	21.0
DaysInWaitingList	87396.0	0.749565	10.015731	0.00	0.0	0.0	0.0	391.0
ADR	87396.0	106.337246	55.013953	-6.38	72.0	98.1	134.0	5400.0
RequiredCarParkingSpaces	87396.0	0.084226	0.281533	0.00	0.0	0.0	0.0	8.0
TotalOfSpecialRequests	87396.0	0.698567	0.831946	0.00	0.0	0.0	1.0	5.0

➔ **ADR** has a minimum value of -6.38
  ◆ This value was replaced with 57.32 (the mean ADR for bookings within a similar group)

Categorical Data:

	count	unique	top	freq
Hotel	87396	2	h2	53428
ArrivalDateMonth	87396	12	August	11257
Meal	87396	5	BB	67978
Country	87396	178	PRT	27453
MarketSegment	87396	8	Online TA	51618
DistributionChannel	87396	5	TA/TO	69141
ReservedRoomType	87396	10	A	56552
AssignedRoomType	87396	12	A	46313
DepositType	87396	3	No Deposit	86251
Agent	87396	334	9	28759
Company	87396	353	NULL	82137
CustomerType	87396	4	Transient	71986
ReservationStatus	87396	3	Check-Out	63371
ReservationStatusDate	87396	926	2016-02-14	211

➔ The most common **Company** is NULL
  ◆ 94% of the rows are within the NULL group, so we will be dropping this column

➔ The number of unique room types differs for **AssignedRoomType** and **ReservedRoomType**
  ◆ There are 2 extra room types in AssignedRoomType. The room types are coded, so there is no way to know what the actual room types are.

6

Feature Engineering:

- **Continent**: country codes were extracted from wikipedia and were used to group the countries into continents
- **is_europe**: binary feature identifying if a booking is from europe (1) or not (0)
- **total_people**: the sum of adults, children, and babies
- **total_nights**: the sum of weekend nights and weeknights
- **cancellation_rate**: PreviousCancellations and PreviousBooksingsNotCanceled combined $PC/(PC + PBNC)$
- **res_equals_assign**: binary feature where reserved room type is the same as assigned room type (1) or not (0)
- **agent_type**: agents are grouped into 4 groups
  1. no agent
  2. uncommon agent (100 or less bookings)
  3. common agent (between 101 and 999 bookings)
  4. popular agent (1000 or more bookings)
- **wait_list**: binary feature identifying if booking is on a wait list (1) or not (0)
- **special_requests**: binary feature identifying if booking has special request (1) or not (0)
- **room_cost**: Product of ADR and total_nights

Features removed:

➜ ArrivalDateYear (irrelevant feature)

➜ Country (replaced by Continent and is_europe)

➜ Adults, Children,Babies (replaced by total_people)

➜ ReservedRoomType, AssignedRoomType (replaced by res_equals_assign)

➜ Agent (replaced by agent_type)

➜ DistributionChannel (using MarketSegment instead)

➜ Company (Mostly NULL)

➜ ReservationStatus, ReservationStatusDate

➜ DaysInWaitingList (replaced by wait_list)

➜ TotalOfSpecialRequests (replaced by special_requests)

Target: IsCanceled

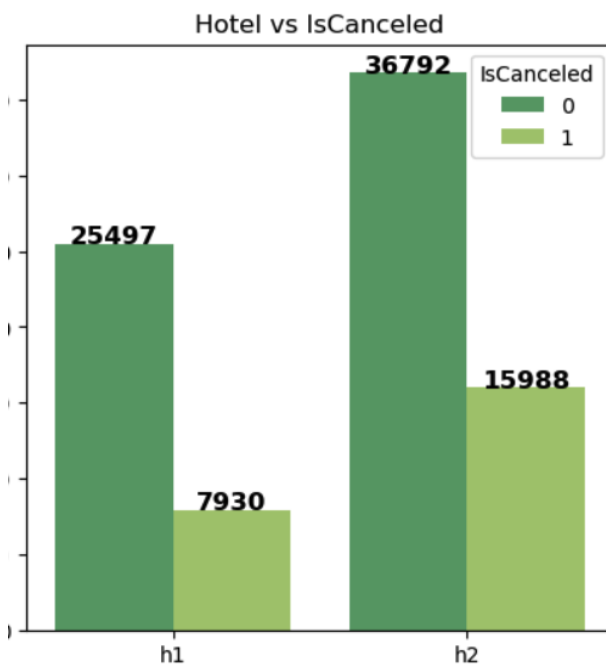**Count for Target Variable: IsCanceled**
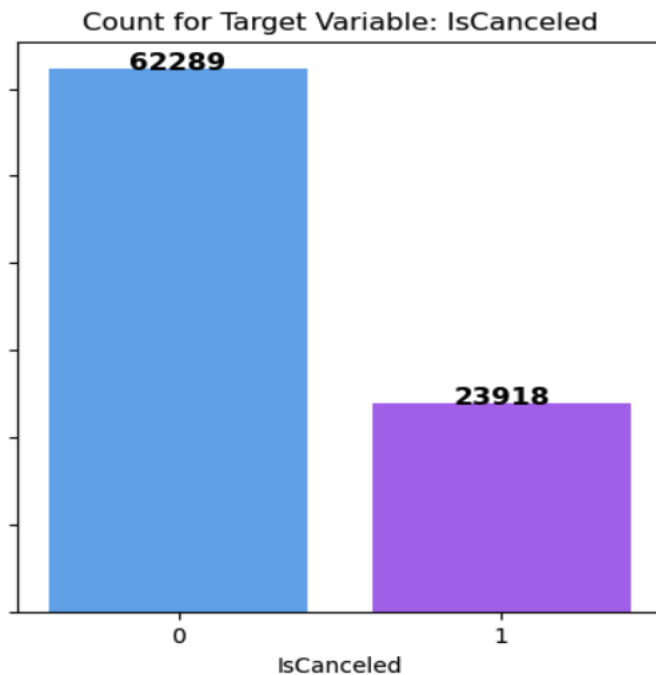
The target variable has:

➔ **62,2289 successful bookings**

➔ **23,918 canceled bookings.**

This **_class imbalance_** will be something to keep in mind during modeling.

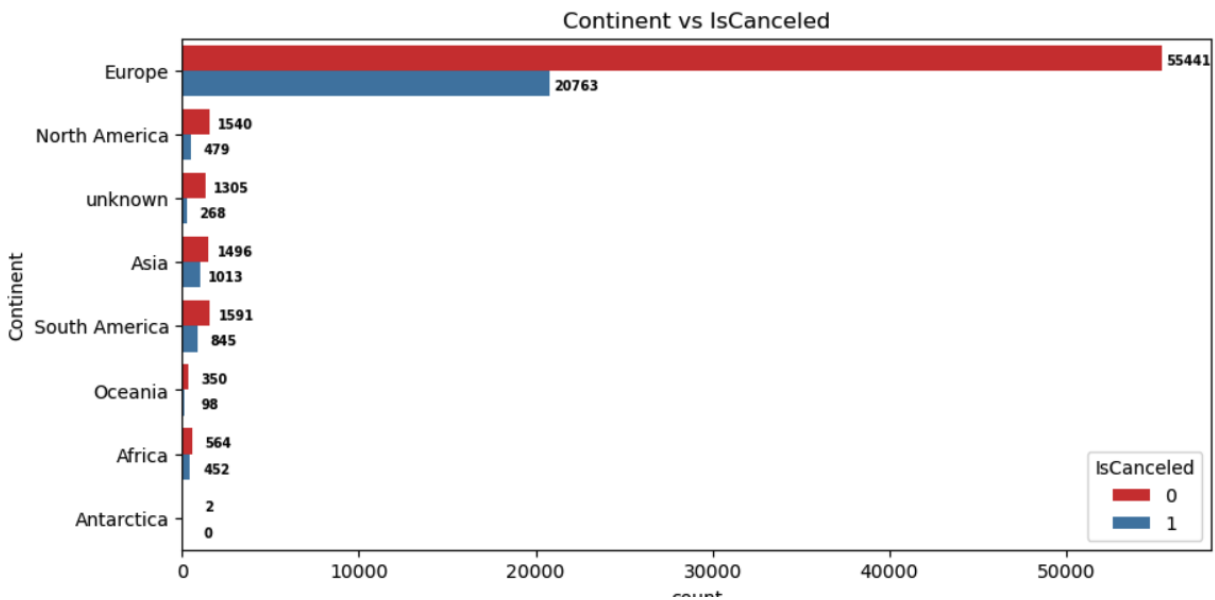## Exploring Features and Target Relationships:

1) Hotel

Hotel vs IsCanceled

➔ Hotel 1: 24% are canceled

➔ Hotel 2: 30% are canceled

Hotel 2 (city) has more bookings overall, but it also has a higher rate of cancellations than hotel 1 (resort).

2) Continent and is_europe

Continent vs IsCanceled

88% of bookings are customers from Europe, which is expected since the hotels are located in Europe (Portugal). However, the highest proportion of customer cancellations comes from other continents. **32% of customers** not from Europe are canceled, while only **27% of the customers** who are from Europe canceled.

% of cancellations within each Continent	
Africa	**44%**
Asia	**40%**
South America	34%
Europe	**27%**
North America	24%
Oceania	22%
Unknown	17%
Antarctica	0%

LeadTime for Reservations
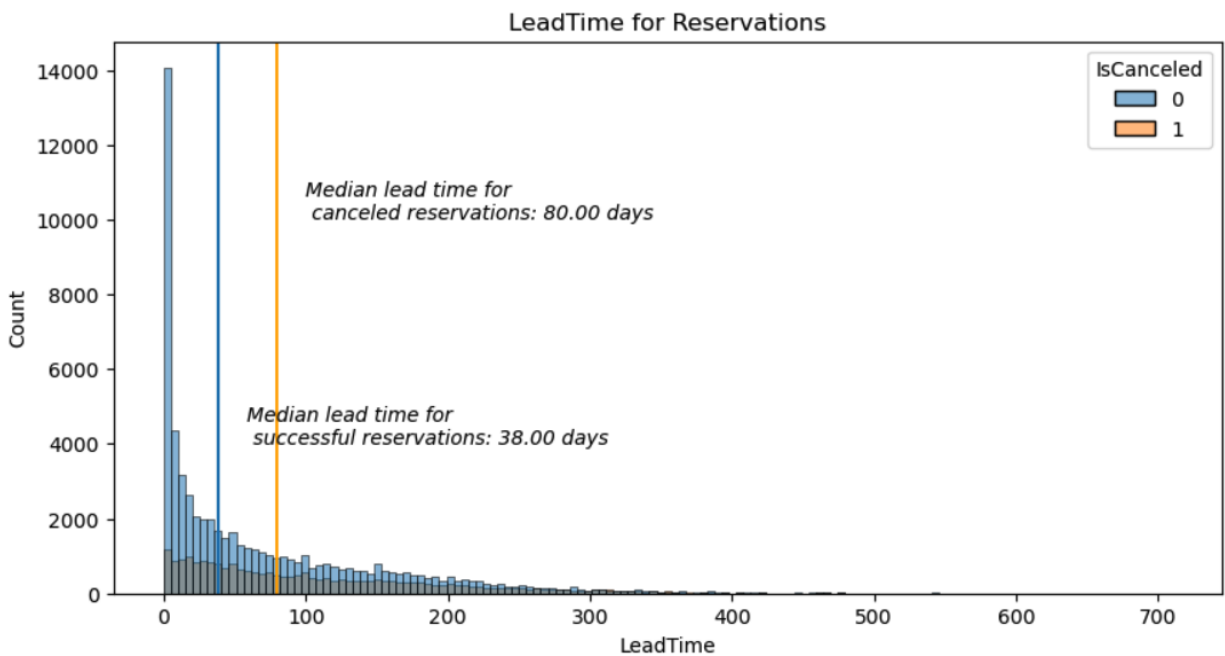
The lead time for bookings is highly skewed to the right, with the majority of bookings having a lead time between 0 and 200 days. The skew suggests some heavy outliers, with a maximum lead time of 737 days. The outliers here will not be removed, as they may provide insights into cancellations for reservations booked far in advance.

➔ For **canceled bookings**, the median lead time is **80 days**.
➔ For **successful bookings**, the median lead time is **38 days**.

This intuitively makes sense, as the more time in between the reservation and the day of arrival, the more likely something can occur that would result in having to cancel a reservation.

4) total_people

| Number of People in Booking	% of total Bookings	% of Cancellations for group
1	**18.12%**	**20%**
2	**65.62%**	**28%**
3	**11.63%**	**31%**
4	4.46%	**42%**
5	**0.15%**	**23%**
10	**0.0023%**	**0%**
12	**0.0012%**	**50%**

➔ The bookings have either 1, 2, 3, 4, 10, or 12 people with about **95%** of the bookings being in the **1, 2**, or **3 person groups**.

➔ Reservations in those groups cancel **20%**, **28%**, and **31%** of the time respectively.

➔ Bookings with **4 people** account for **4.46%** of the total bookings, but **cancel 42% of the time.**

Bookings with **more than 4 people account for less than 1% of the total bookings,** so we cannot draw many insights from them. However, we can see that as the number of people in the booking increases, the percentage of cancellation increases.

5) total_nights and is_weekend

total_nights vs IsCancled
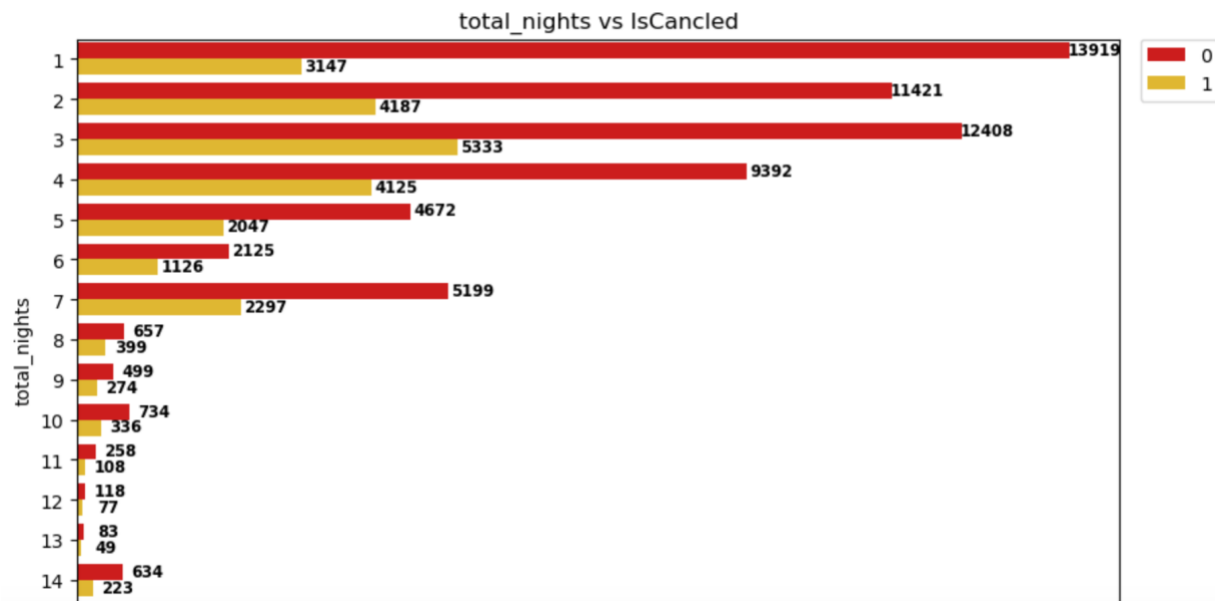
➔ The number of total nights ranges from 1 to 69, with the majority of bookings being less than 15 nights.
➔ The **median** for both the canceled and successful reservations is **3 days**.
➔ As the number of nights increases, the higher the percentage of cancellations within that group:
   ◆ 1 night = 18.4% canceled
   ◆ 2 nights = 26.8% canceled
   ◆ 3-7 nights ≃ 30% canceled
   ◆ 8-9 nights ≃ 36% canceled
   ◆ 10-11 nights ≃ 30% canceled
   ◆ **12-13 nights ≃ 38% canceled**

★ 60% of reservations are on **weekend nights**, and they **cancel 29.4% of the time.**
★ 40% of reservations are on **weeknights**, and **cancel 25.2% of the time.**

6) meal

Meal Type vs IsCanceled
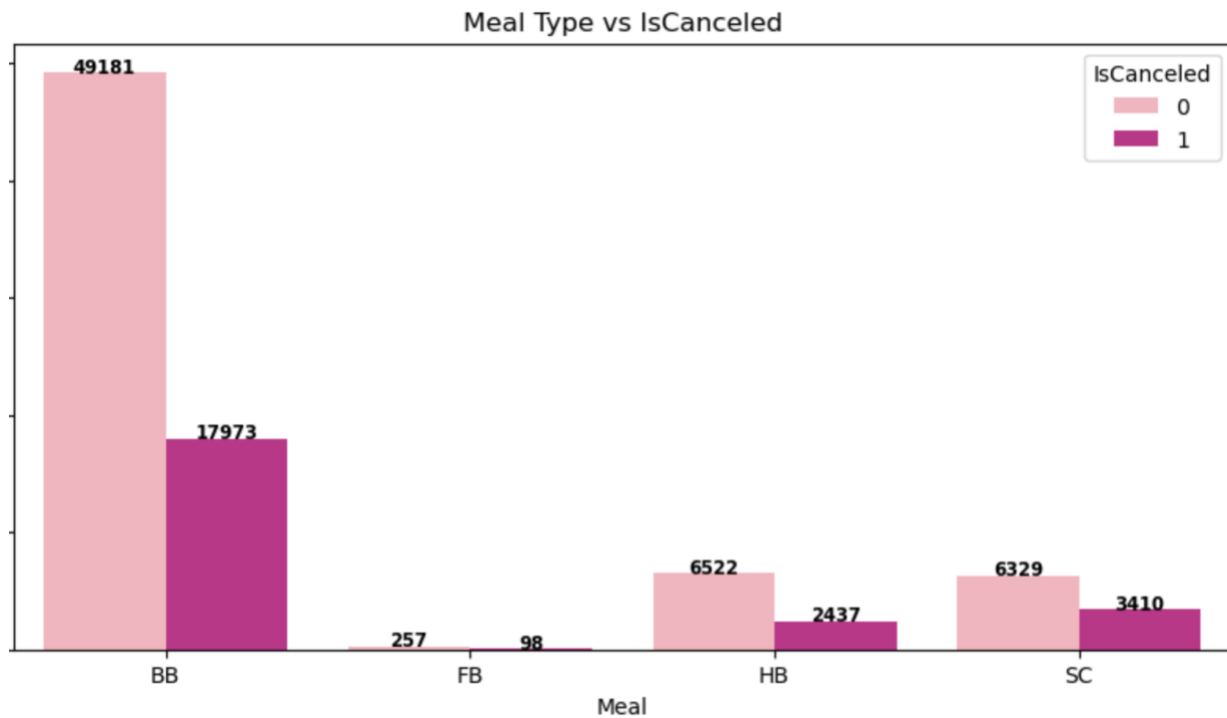
➔ The bookings with meal types of BB, FB, and HB all cancel approximately 26-27% of the time, while the **meal type of SC (no meal plan) cancels 35% of the time.**

➔ **To reduce the number of groups, we will binarize this feature (has meal = 1, no meal = 0)**

Since the market segment and customer type give almost the same information, we will only be focusing on the market segment to avoid any redundancy or collinearity between features.

Market Segment vs IsCanceled
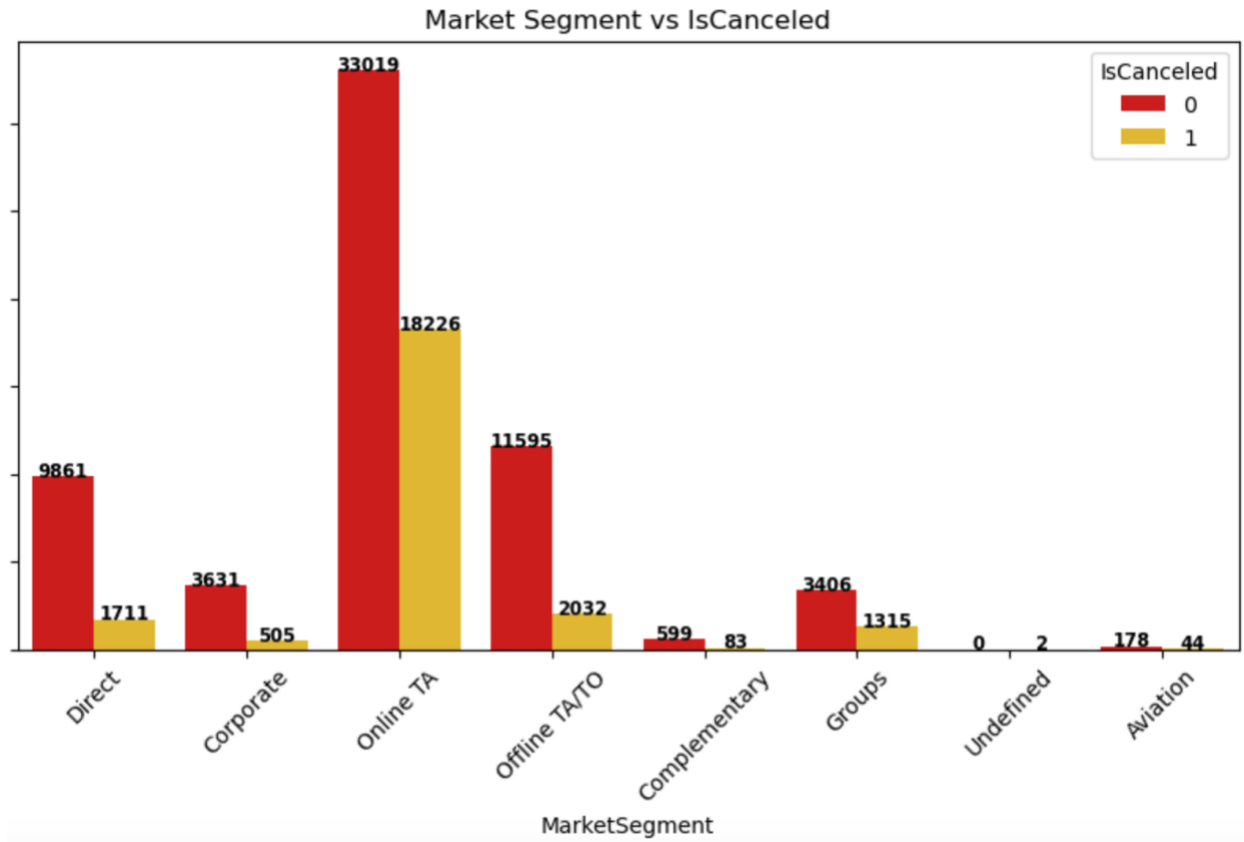
→ The biggest takeaway from analyzing market segments is that bookings in the **Online TA market segment** account for the majority of the bookings, and also have the highest percentage of cancellations, with **36% of the bookings canceled**.

→ 5% are in the **Groups** segment: **28% cancel.**

→ 16% are in the **offline TA/TO** segment: **only 15% cancel.**

8) IsRepeatedGuest and canellation_rate

IsRepeatedGuest vs IsCanceled
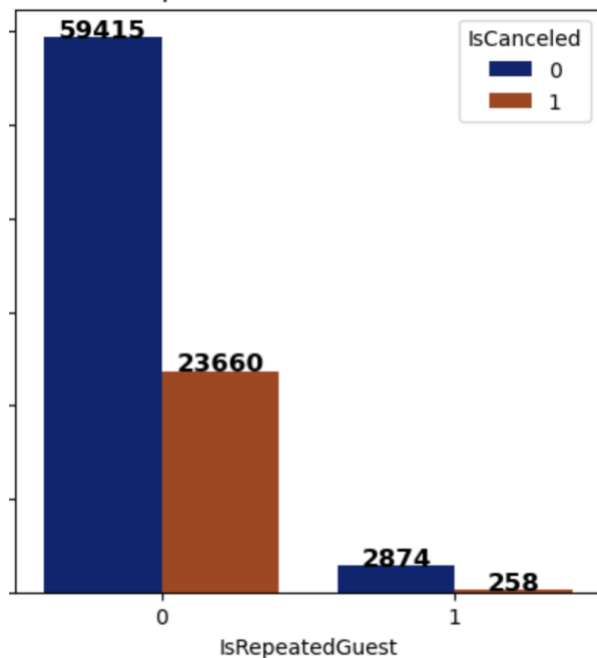
➔   96% of bookings come from **new guests.**

◆ **28% canceled**

➔   The other 4% are **repeat guests**.

◆ Only **8% canceled**

The **cancellation_rate** feature was created from previous booking information, so we will refer to this as the PCR (previous cancellation rate). To simplify the analysis of this feature, we will look three different groups:

Cancellation Rate for Previous Bookings

Group A: Bookings with 0% PCR. This group accounts for 98% of the bookings, and mostly consists of new customers. **26% canceled** in the current sample of bookings.

Group B: Bookings with 100% PCR. These are repeat guests who have canceled all of their previous reservations. **98% canceled** in the current sample of bookings.

Group C: Bookings that have a PCR between 0-100%. Only 0.8% are in this group. The histogram suggests that people with a 50% PCR, canceled very frequently again.
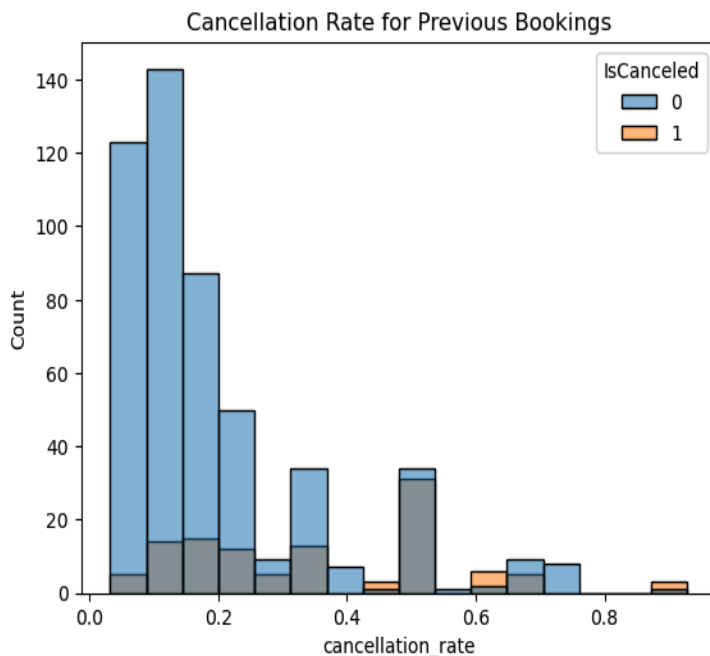
In general, people who canceled previously, tend to cancel again.

15

9) res_equals_assign

res_equals_assign vs IsCanceled
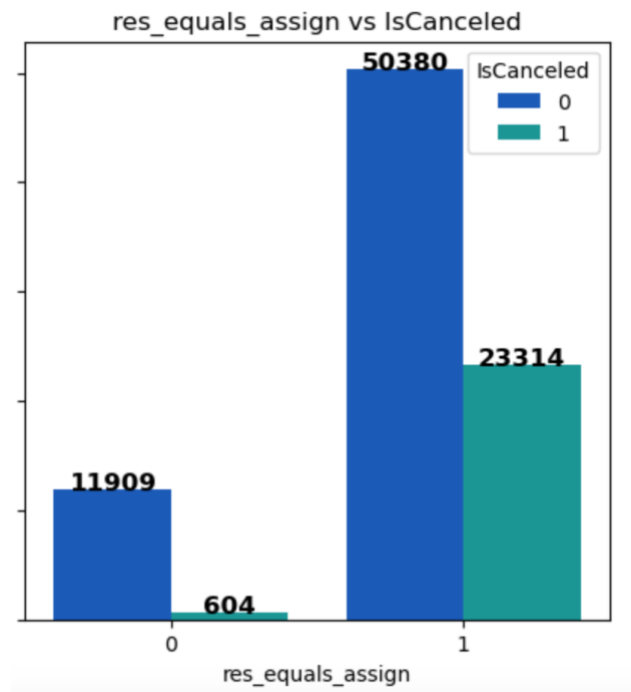
➔ 85% of the bookings received the same room type that was originally requested, and 32% canceled.

➔ **15% of bookings did not receive the same room type, and only 5% of them canceled.**

This intuitively makes sense and is quite insightful. The most likely explanation for this group canceling less is that they may have received a room upgrade, which would incentivize the customer to not cancel.

10) booking_changes

82.02% of the bookings did not request any changes to their reservation, 12.35% requested only a single change, and 3.9% requested two changes. The number of changes ranges from 1 up to a maximum of 18 changes, but everything above 2 accounts for less than 1% of the total bookings. We recoded this feature to be binary, where there is a booking change (1) or no changes (0).

BookingChanges vs IsCanceled
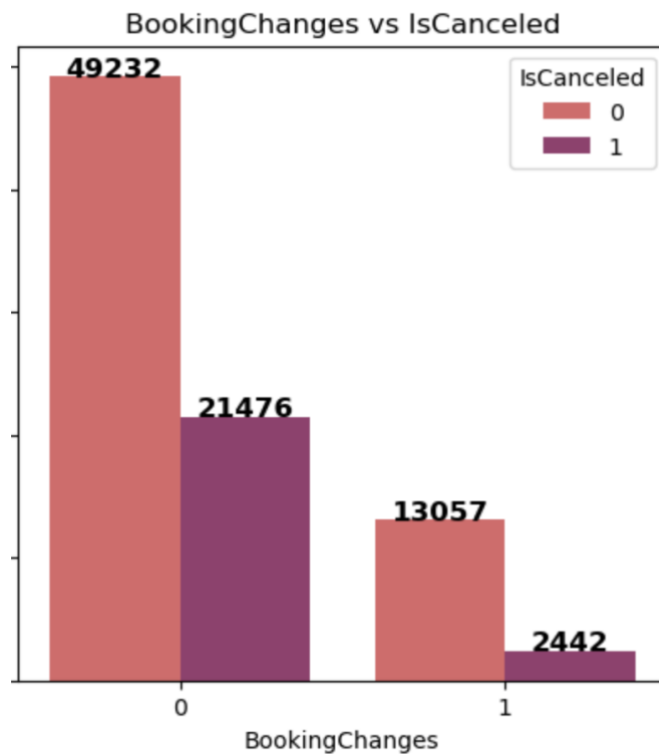
➜ Bookings with **no changes**: **30% canceled**

➜ Bookings with **1 or more changes: 16% canceled**

DepositType vs IsCanceled
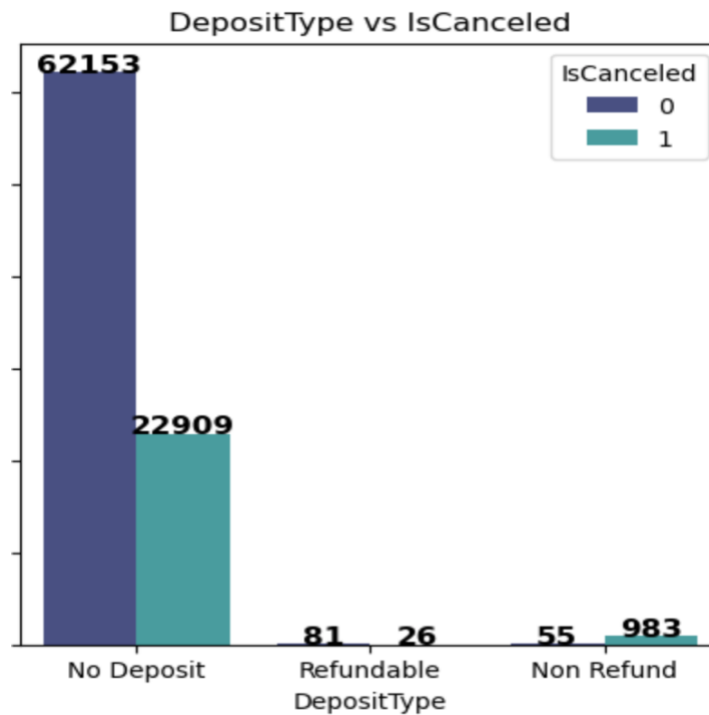
➜ 98.67% of bookings do **not have a deposit: 27% canceled**

➜ 1.2% of bookings have a **non-refundable reservation: 94.7% canceled**

➜ Less than 1% of bookings have a **refundable deposit: 24.29% canceled**

Only 1.2% of the bookings are in the non-refundable group, but it is counterintuitive that such a high percentage of these bookings are canceled. Upon further inspection, 63% of these bookings are within the 'Groups' market segment. Groups generally have more than 4 people within their party, and as we stated earlier, parties with more people generally cancel more often.
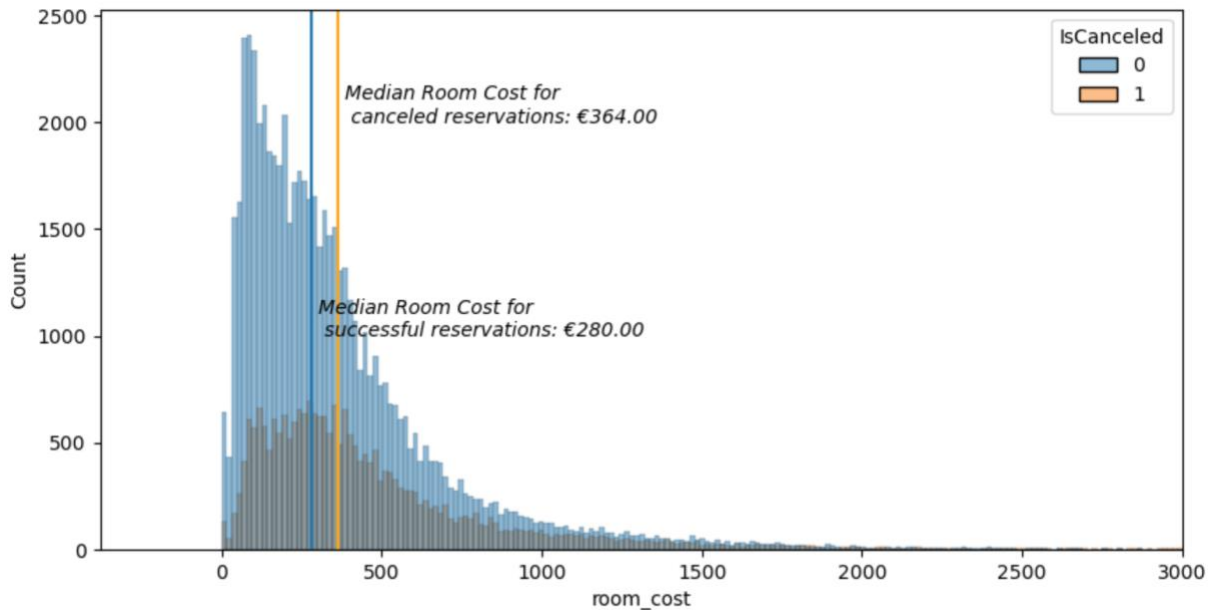
12) agent_type

Agent Type vs IsCanceled
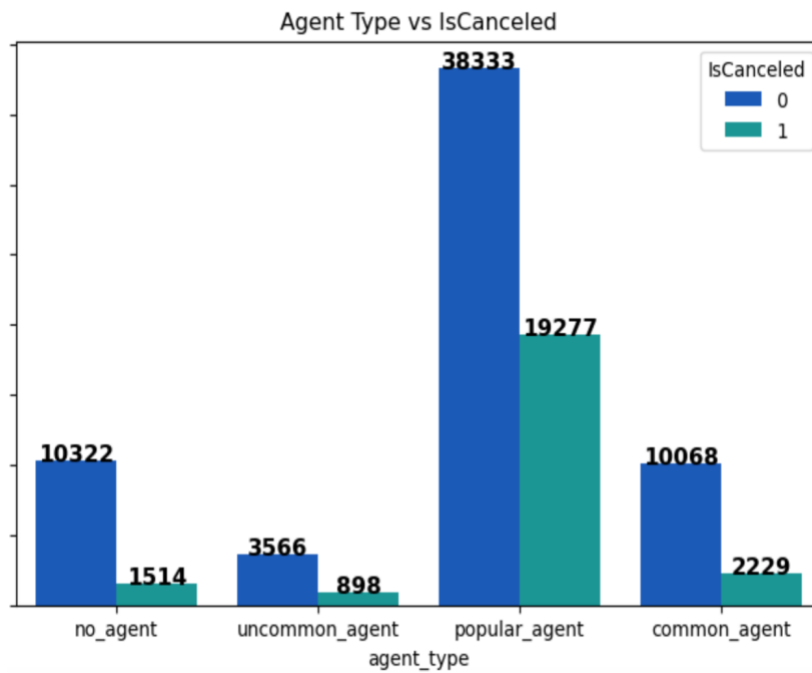
➔ 66.8% of the reservations were made by a **popular agent: 33.5% canceled**.

➔ 13.7% of reservations were made **without an agent: 12.8% canceled**.

13) wait_list

99% of the bookings are not on a wait list. Of the 834 that are on a wait list, 35% of them canceled.

14) Room Cost

The **median room cost for canceled reservations is $364**, while for **successful reservations it is $280**. This suggests that bookings that cost more tend to cancel more than cheaper bookings, which is to be expected.

There is approximately €34,446,903 worth of potential revenue if all of the bookings are successful. The cancellations account for about 33% of that, which is a significant loss.

15) RequiredCarSpaces and special_requests

These features were both coded into binary values. **RequiredCarSpaces** indicates if a parking space was requested (1) or not (0). **Special_requests** indicates whether a booking made a special request (1) or not (0).

➔ Only 8% of the bookings **required car parking** spaces: **0% canceled**.
➔ Approximately 50% of the bookings **made a special request: 22% canceled**.
➔ The other 50% **did not make a special request: 34% canceled.**

16) Arrival Month

Arrival Month vs IsCanceled
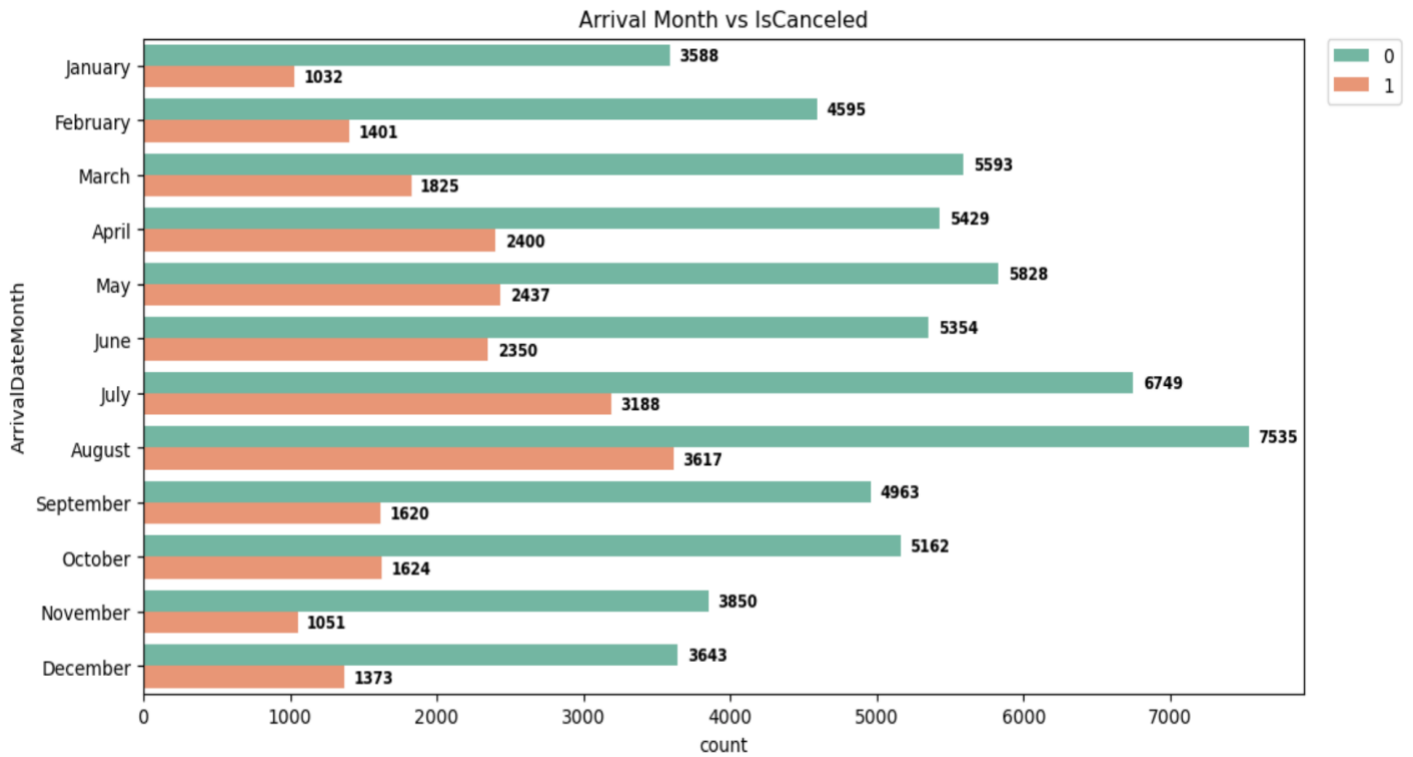
We can see that the summer months of **July and August** have the most bookings in the dataset, but they also have the most cancellations. **32% of the bookings in those months were canceled.**

Correlation and Associations

Categorical/binary Features

column	Cramer V	Association
MarketSegment	0.221915	strong
res_equals_assign	0.210918	strong
RequiredCarParkingSpaces	0.188054	strong
agent_type	0.185183	strong
DepositType	0.165105	strong
special_requests	0.131184	moderate
CustomerType	0.127644	moderate
BookingChanges	0.125365	moderate
IsRepeatedGuest	0.084595	weak
Continent	0.077996	weak
Hotel	0.071479	weak
Meal	0.057939	weak
is_weekend	0.046879	very weak
is_europe	0.030714	very weak
wait_list	0.016042	very weak

Numerical Features

	IsCanceled
LeadTime	0.183932
cancellation_rate	0.169148
room_cost	0.135016
ADR	0.126152
total_people	0.100367
total_nights	0.081672
PreviousCancellations	0.050797
ArrivalDateDayOfMonth	0.005476
ArrivalDateMonth	0.004739
PreviousBookingsNotCanceled	-0.052452

1. For the categorical/binary features, a Chi-Squared test for association was performed with the following hypotheses:

   H_0 = The feature is not associated with the target variable

   H_a = The feature is associated with the target variable

   a. With an alpha of 0.05, the corresponding p-values were all significantly smaller than it. This suggests that we can reject our null hypothesis, however it is more appropriate to utilize the Cramer's V statistic to have a better idea of the strength of these associations.

2. For the numerical features, we used Pearson correlation to check for an association with the target.

Our results from testing for associations suggest that **MarketSegment, res_equals_assign, RequiredCarParkingSpaces, agent_type, deposit_type, LeadTime, cancelation_rate,** and **room_cost** are the most strongly associated with **IsCanceled.**

Correlations between Features:
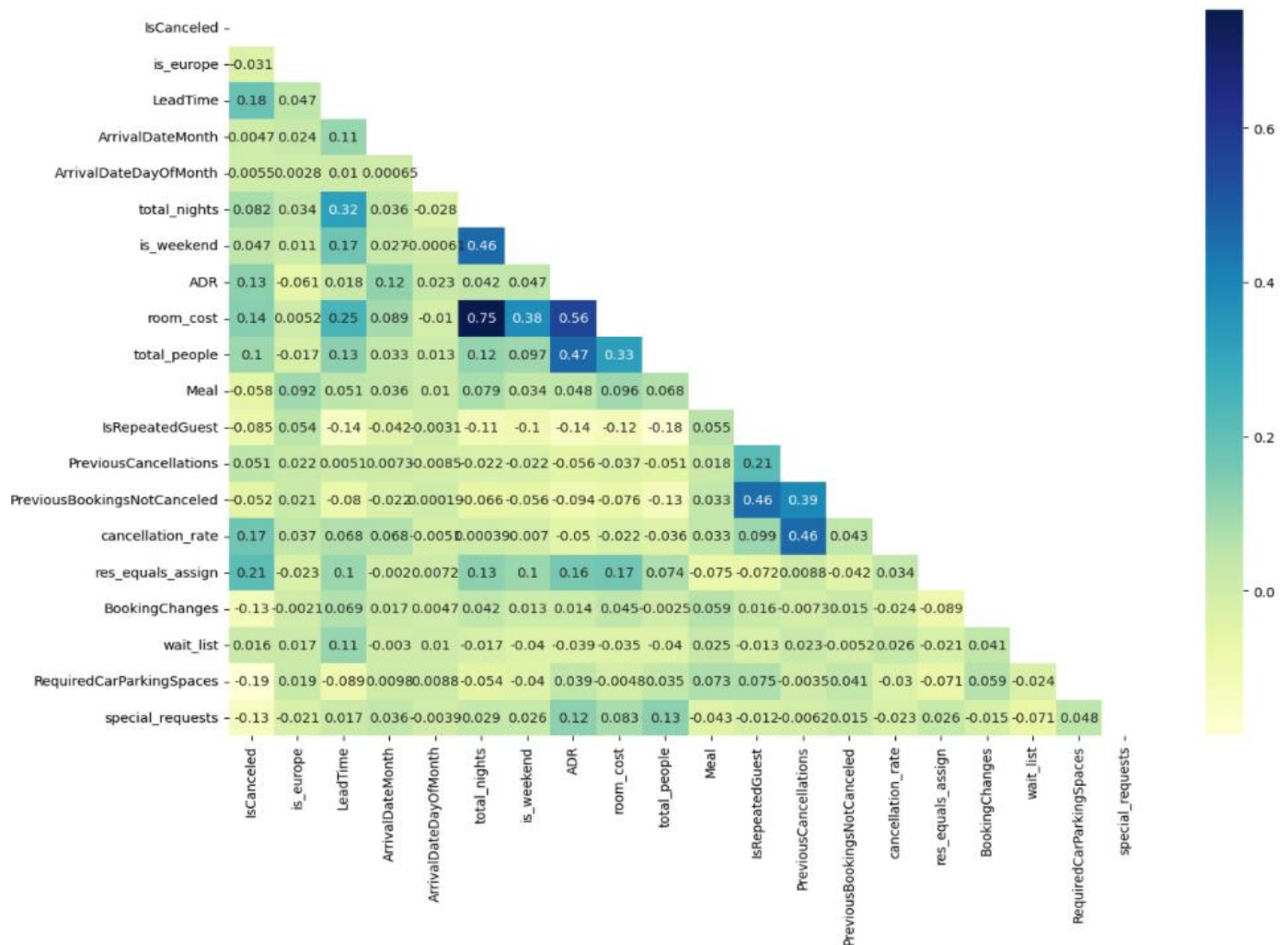
The correlation heatmap suggests the following strong associations:

- room_cost ⟺ total_nights, ADR, and total_people
- Cancelation_rate ⟺ PreviousCancellations
- IsRepeatedGuest ⟺ PreviousBookingsNotCanceled, PreviousCancellations

To avoid any redundancy or multicollinearity, we will remove **ADR, total_nights, Continent, PreviousBookingsNotCanceled, and PreviousCancellations.**
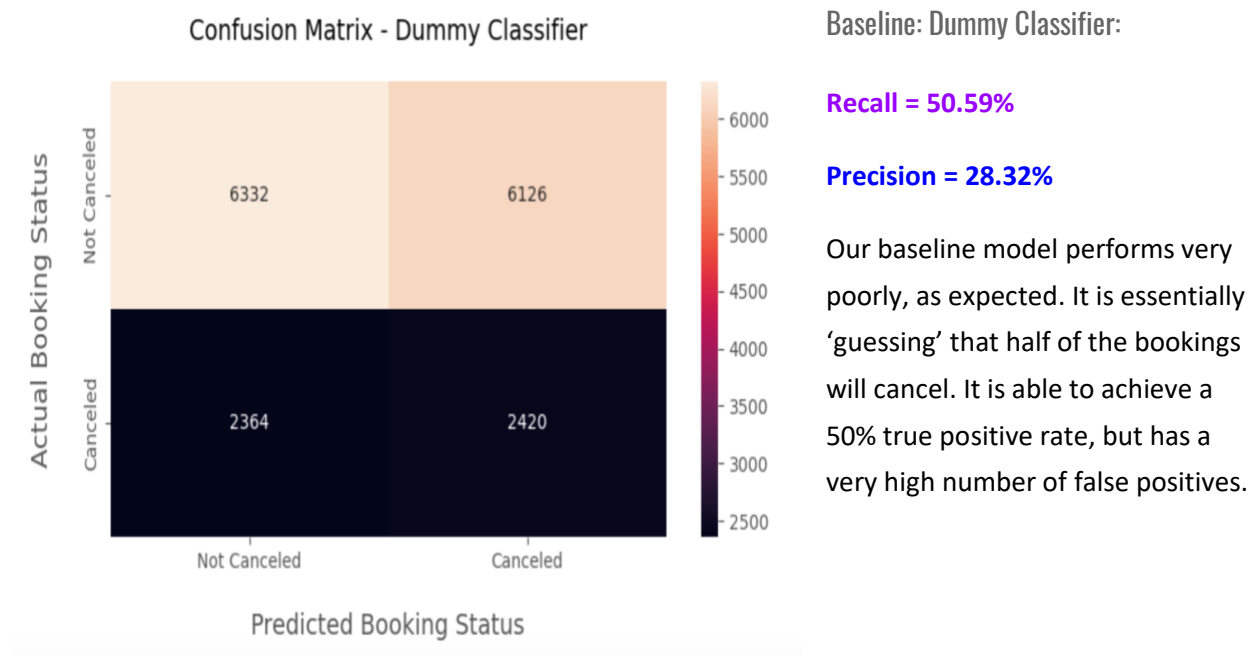
Modeling

After feature engineering and one-hot encoding our categorical features, we have **86207 rows, and 28 features**. Our data is split into training and test sets, with 20% of the data in the test set.
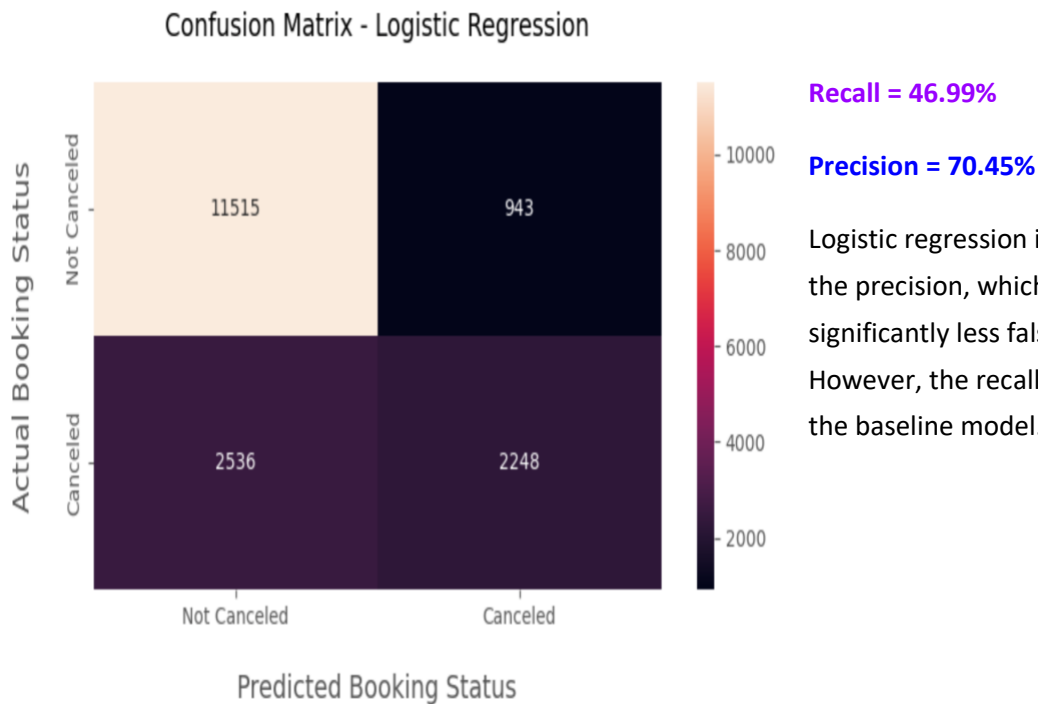
- ❖ Training set: 68965 rows
- ❖ Test set: 17242 rows

As mentioned in the EDA section, our target variable 'IsCanceled' has some significant class imbalance. The positive class (canceled) accounts for only 27.7% of the total, so we have to be careful with the metrics we choose to evaluate our models. Pure accuracy will not be appropriate here, so we will be using recall and precision to assess the quality of our models.

I.   **Recall** is the true positive rate, or the percentage of actual cancellations that were correctly predicted to cancel. This is our most important metric, as it embodies the goal of this project.
II. **Precision** tells us about the false positive rate. If our model predicts that a booking will cancel, the percentage of those predictions that are actual cancellations will be given by precision.

Confusion Matrix - Dummy Classifier

Baseline: Dummy Classifier:

Recall = 50.59%

Precision = 28.32%

Our baseline model performs very poorly, as expected. It is essentially 'guessing' that half of the bookings will cancel. It is able to achieve a 50% true positive rate, but has a very high number of false positives.

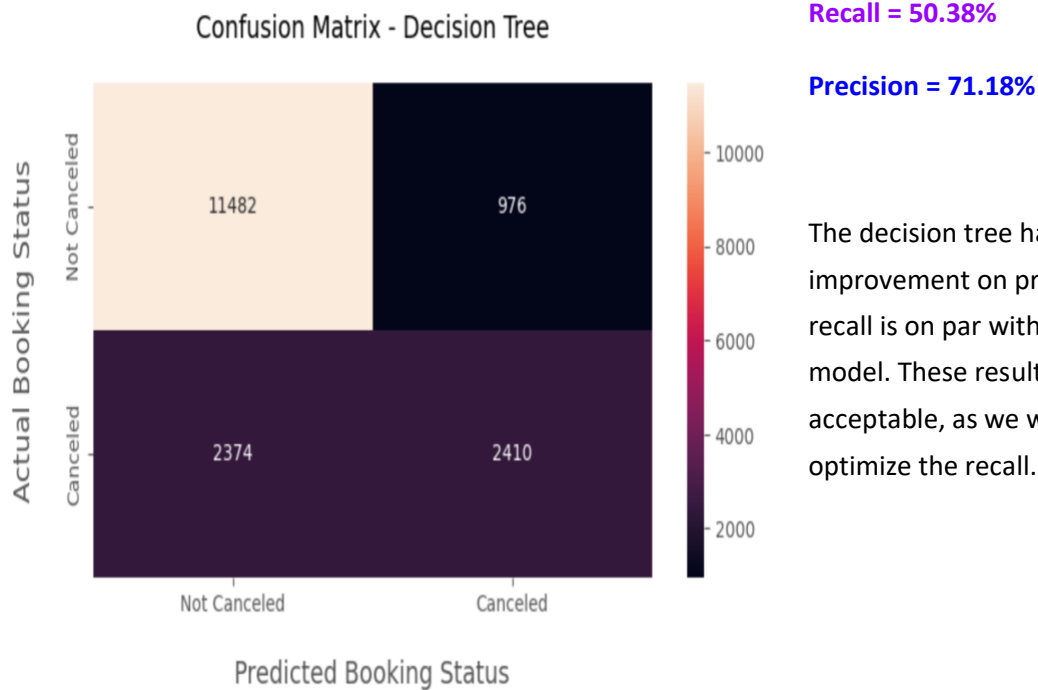## Confusion Matrix - Logistic Regression

**Recall = 46.99%**

Precision = 70.45%

Logistic regression improves upon the precision, which results in significantly less false positives. However, the recall is worse than the baseline model.

Decision Tree

Confusion Matrix - Decision Tree

Recall = 50.38%

Precision = 71.18%

The decision tree has a 1% improvement on precision, and the recall is on par with the baseline model. These results are not quite acceptable, as we would like to optimize the recall.
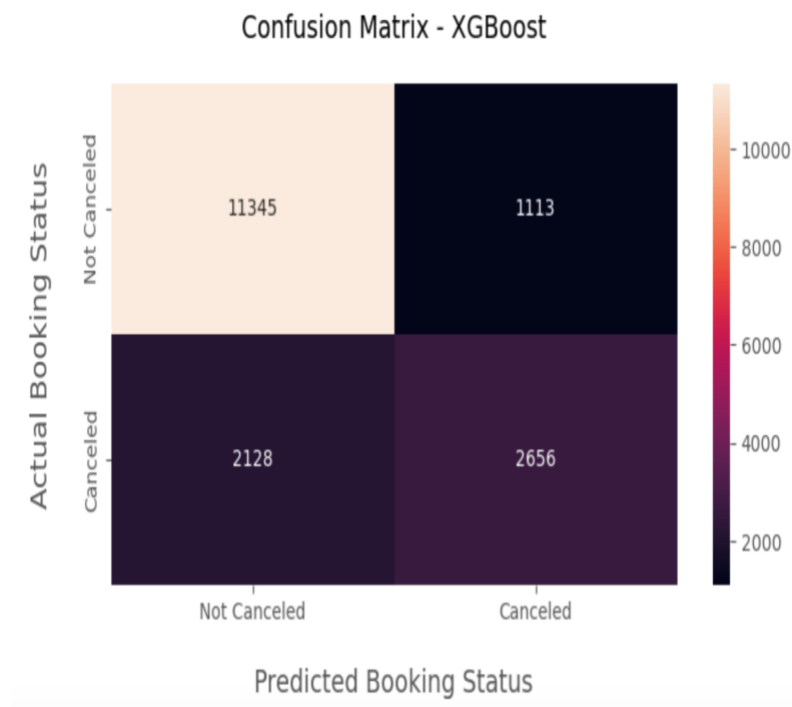
Random Forest

Confusion Matrix - Random Forest
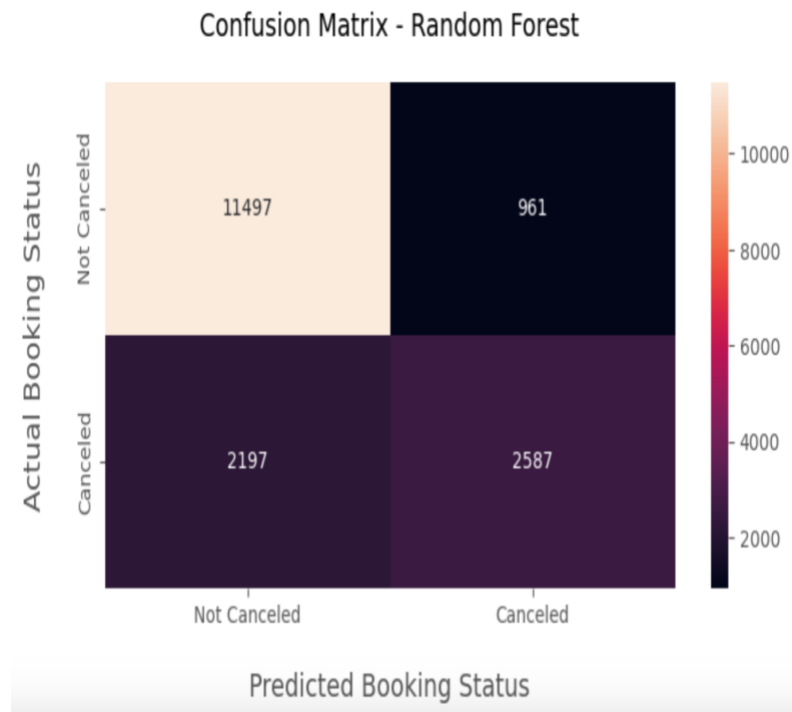
Recall = 54.08%

Precision = 72.29%

Finally, the random forest model improves on the recall of the baseline model by 4%. We will try some boosting methods to try and further optimize our recall.
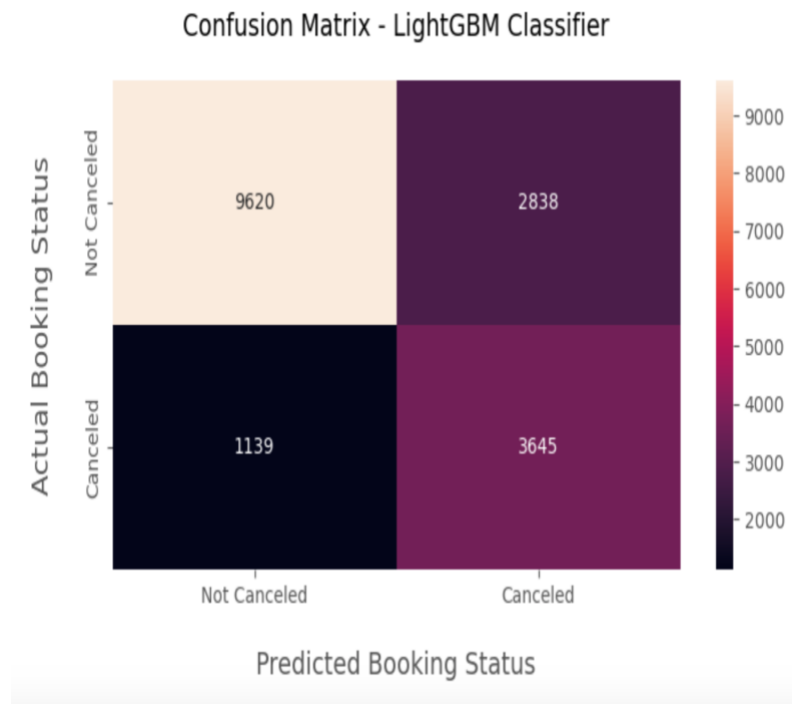
XGboost

Confusion Matrix - XGBoost

Recall = 55.52%

Precision = 70.47%

XGboost results in a slight 1% improvement in recall, but has a 2% decrease in precision.

Confusion Matrix - LightGBM Classifier

Recall = 76.19%

Precision = 56.22%

LightGBM significantly improves on the recall with a 21% increase from XGBoost. Of the 4,784 canceled bookings, the model is correctly identifying 76% of them. However, this model also has a 14% decrease in precision from the XGboost model. These results will be discussed further in the conclusion, but this will be our final model.

	Train Recall	Test Recall	Train Precision	Test Precision	Train Accuracy	Test Accuracy	Train AUC	Test AUC
Dummy Classifier	0.505906	0.505853	0.280166	0.283173	0.502284	0.507598	0.503399	0.507060
Logistic Regression	0.486987	0.469900	0.712603	0.704481	0.803176	0.798225	0.705786	0.697103
Decision Tree	0.525870	0.503763	0.725555	0.711754	0.813268	0.805707	0.724746	0.712710
Random Forest	0.822828	0.540761	0.950094	0.729143	0.938853	0.816843	0.903116	0.731811
XGBoost	0.655900	0.555184	0.806866	0.704696	0.860973	0.812029	0.797808	0.732922
LightGBM	0.818856	0.761915	0.603126	0.562240	0.800247	0.769342	0.805979	0.767055

The models have been applied to both the training and test sets to check for overfitting. Logistic regression and the decision tree classifiers both have a low variance, and thus do not have an overfitting issue. The random forest has a significantly higher variance, and while not as severe, XGBoost also has a high variance. Of the high performing models, LightGBM has the lowest variance with only a 5% difference in recall between train and test sets. To further validate these results, 5-fold cross validation was performed on LightGBM and a mean recall of 0.76 was achieved, suggesting that our model does not have an overfitting issue.

The overall accuracy was included in the table to showcase the issue with using it as a metric when class imbalance is present. Other than the baseline, it has the lowest test accuracy. This is extremely misleading, as it is the model that performs the best at predicting true positives. A more appropriate metric is the AUC, or the area under the ROC curve. The ROC curve describes the relationship between the true positive rate and false positive rate, and thus optimizing the area under the curve results in a higher performing model.

Feature Importance

Top 10 Features for LightGBM
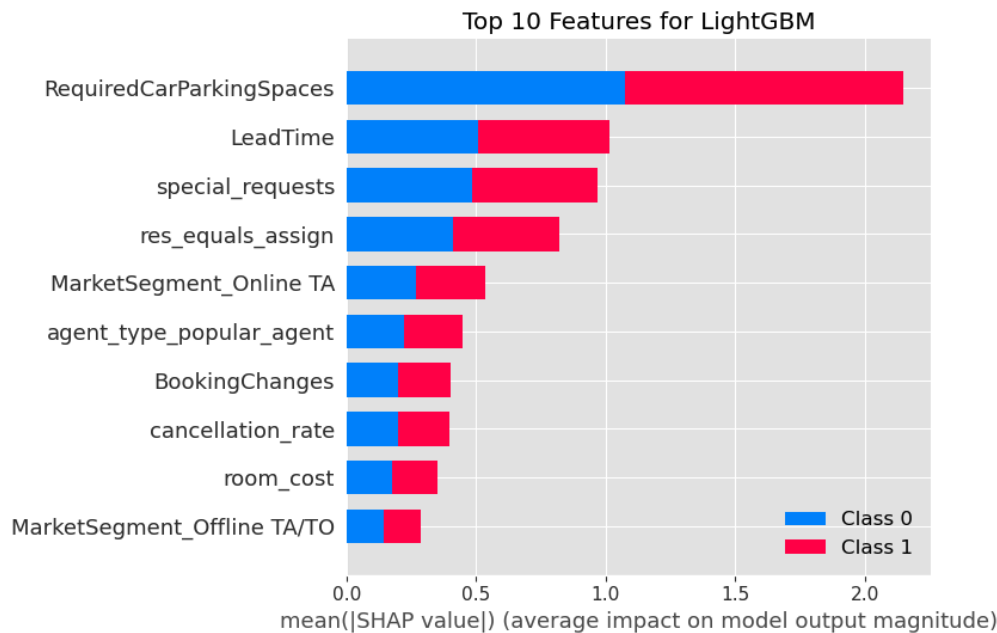
The 10 most important features used by the LightGBM model are displayed in the figure above, and these results seem to closely align with our observations from the EDA section:

1. Required car parking spaces being the most important makes sense, as we saw the bookings that requested a space never canceled.
2. For lead time we saw a significant difference in the median between the two groups, concluding that reservations made farther in advance were more likely to cancel.
3. We found that bookings with special requests were less likely to cancel.
4. Customers who did not receive the same room type that they booked canceled less often.
5. Customers from the Online TA market segment canceled more frequently.
6. Bookings made by a popular agent canceled more frequently.
7. Changes to the booking resulted in less cancellations
8. Customers that previously canceled also had more current cancellations.
9. Higher room cost results in more cancellations
10. The offline TA/TO market segment has significantly less cancellations.

Conclusions

The LightGB model has shown the most favorable overall results in relation to the specific business value we aim to provide. Out of the **4,784 canceled bookings** in our test set, the model **accurately identifies 76% of them**. However, it is worth noting that this model exhibits a notable decrease in precision. Nevertheless, this decrease in precision is not necessarily problematic since it refers to bookings that were predicted to cancel but actually did not.

To reiterate the project's objective, our goal is to identify bookings that are likely to cancel in order to make efforts to retain them. If a customer was not planning to cancel in the first place, misclassifying them does not have a significant negative impact. Among the **12,458 successful bookings** in the test set, only **22.7% are mistakenly classified as cancellations**.

It's important to consider that the chosen strategy to retain bookings may potentially affect profitability. For instance, if the hotel decides to offer complimentary nights to every booking predicted to be at risk of cancellation, it could result in substantial costs. However, since we lack access to the financial data of these hotels, we cannot accurately assess the actual expenses. One approach the business can adopt is to base decisions not solely on the model's classifications, but also on the corresponding probabilities.

IsCanceled	predictions	probability_sucessful	probability_canceled
0	1	0.412427	0.587573
1	1	0.118317	0.881683
0	0	0.949426	0.050574
1	0	0.635925	0.364075
0	0	0.701680	0.298320

The table above shows a sample of 5 bookings. Row 1 was a successful booking, but was predicted to be canceled. The model made that decision with only a 58.7% probability. Row 2 on the other hand was a canceled booking that was correctly predicted to cancel with 88% probability. A probability threshold may be needed when making the final decision on which bookings should be targeted.

Ideas for Implementation:

I. Bookings predicted to cancel with **at least 70% probability** will be considered **high risk**. They will receive the following:

    A. An email

    B. A phone call

    C. Complementary nights, meals, other services (if costs allow)

II. Bookings predicted to cancel with **less than a 70% probability** will be considered **low risk** and receive the following:

    A. An email

    B. A phone call

The purpose of this strategy is to provide valuable incentives to high-risk bookings, even though they may be expensive, because the potential revenue generated from successful bookings is expected to outweigh those costs. On the other hand, for low-risk bookings, simply contacting the customer may be sufficient to retain their booking, assuming they were contemplating cancellation in the first place. By implementing this plan, we can proactively reach out to all predicted cancellations without adversely affecting profits.

Further Work

Without financial data related to the hotel revenue and profit, there is no way of knowing how many incentives the hotel can realistically offer. Of the 6,483 bookings predicted by the model to cancel, how many complementary rooms or meals can be given without hurting the profits? With financial data this can be explored further, and a more accurate business decision can be made.

In terms of improving the model, one feature that is currently unknown to us is weather information. It is intuitive to expect that many cancellations are weather related, and so having weather forecast data may be helpful. For instance, if a storm is expected over the North Atlantic Ocean, then customers traveling from North America will likely cancel.