Star Hotels
Business Presentation

Contents

The presentation consists of three Four Sections:

1. Background and Business Problem Overview
2. Data Overview
3. Exploratory Data Analysis
4. Model overview and performance summary
5. Key findings and insights
6. Business recommendations

Background of the Business and Objective

Background

A significant number of hotel bookings with STAR Hotels are called off due to cancellations or no-shows.
This is often made easier by the option to do so free of charge or preferably at a low cost which is beneficial to hotel guests, but it is a less desirable and possibly revenue-diminishing factor for hotels to deal with.

Such losses are particularly high on last-minute cancellations.

Objective

Develop a Machine Learning based solution that can help in predicting which booking is likely to be canceled.

Analyze the data provided to find which factors have a high influence on booking cancellations, build a predictive model that can predict which booking is going to be canceled in advance, and help in formulating profitable policies for cancellations and refunds.

Data Overview

Data	Description
no_of_adults	Number of adults
no_of_children	Number of Children
no_of_weekend_nights	Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
no_of_week_nights	Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel
type_of_meal_plan	Type of meal plan booked by the customer
required_car_parking_space	Does the customer require a car parking space? (0 - No, 1- Yes)
room_type_reserved	Type of room reserved by the customer. The values are ciphered (encoded) by Star Hotels Group
lead_time	Number of days between the date of booking and the arrival date
arrival_year	Year of arrival date
arrival_month	Month of arrival date
arrival_date	Date of the month
market_segment_type	Market segment designation.
repeated_guest	Is the customer a repeated guest? (0 - No, 1- Yes)
no_of_previous_cancellations	Number of previous bookings that were canceled by the customer prior to the current booking
no_of_previous_bookings_not_canceled	Number of previous bookings not canceled by the customer prior to the current booking

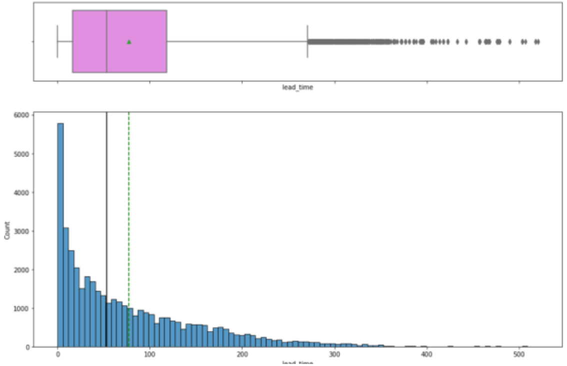
Data Overview

- There are 56926 rows and 18 columns.
- There are 14350 duplicate values in the dataset, We removed the duplicate data before processing.
- Most of the data-types are either int64 or float64.
- type_of_meal_plan, room_type_reserved, market_segment_type and booking_status having data-types as an object, this means we need to convert these into suitable data-type before we feed our data into the model.
- There are no missing values in the dataset.
- Unique values shows that
  - There are 4 types of meal plans
  - There are 7 types of room types
  -There are 5 types of market segment type

Exploratory Data Analysis

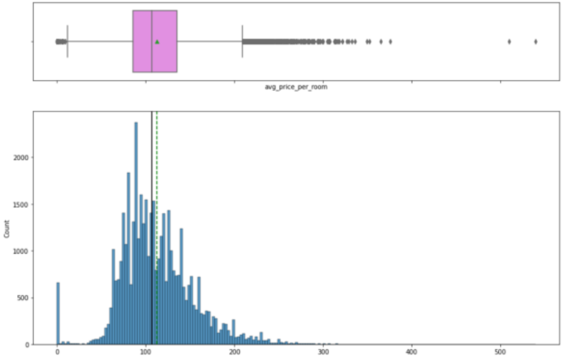
Lead Time .

```
<Figure size 504x288 with 0 Axes>
```



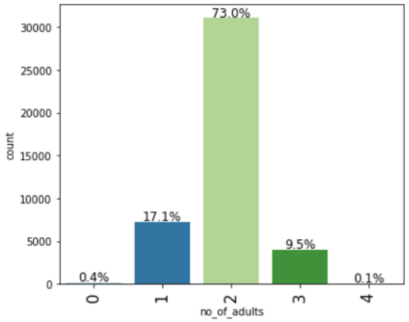
Average nightly price

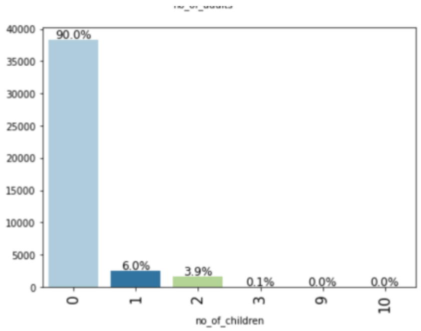
```
<Figure size 504x288 with 0 Axes>
```

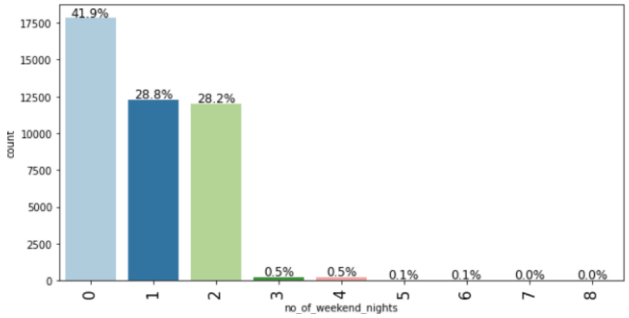


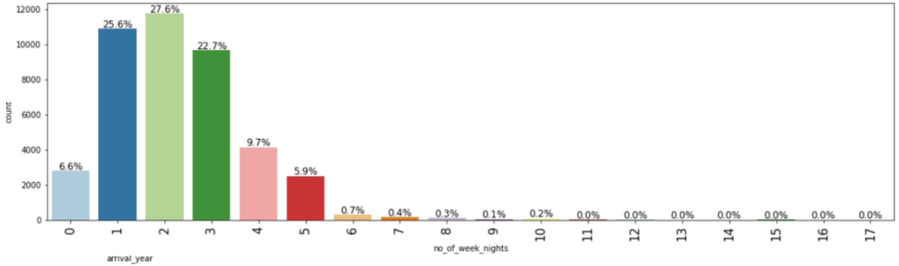
lead_time is right skewed.
There are outliers exists in avg_price_per_room.

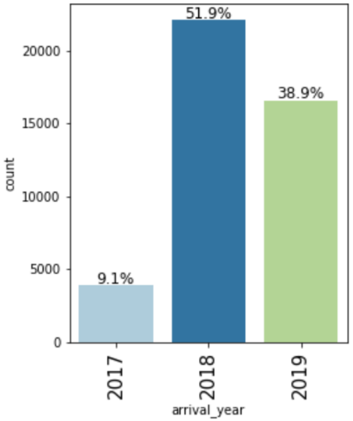
Exploratory Data Analysis

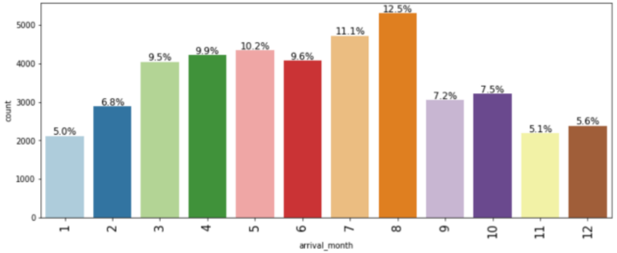




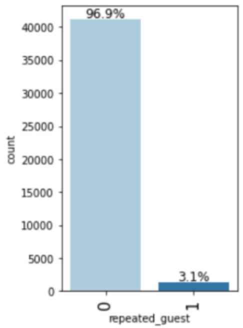


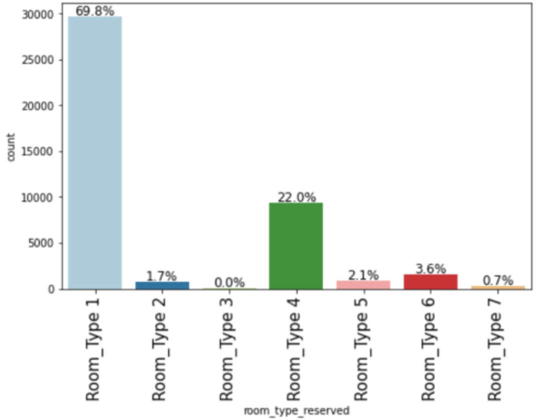


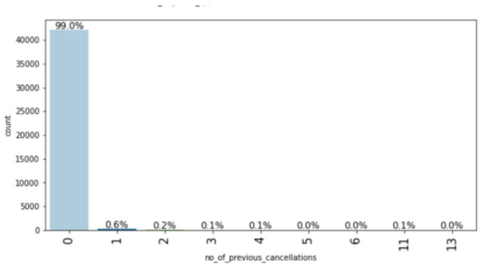


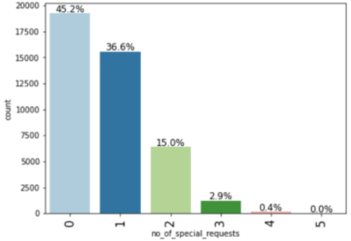


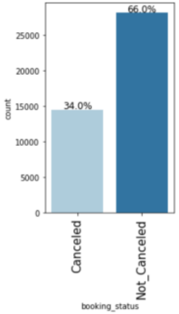
Exploratory Data Analysis - Univariate analysis

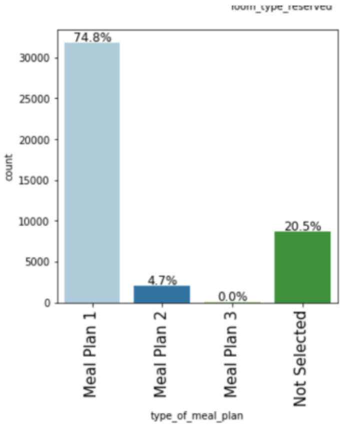


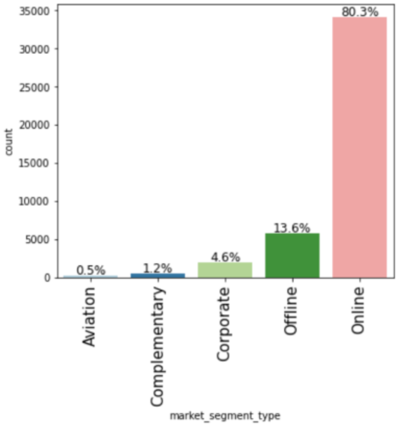




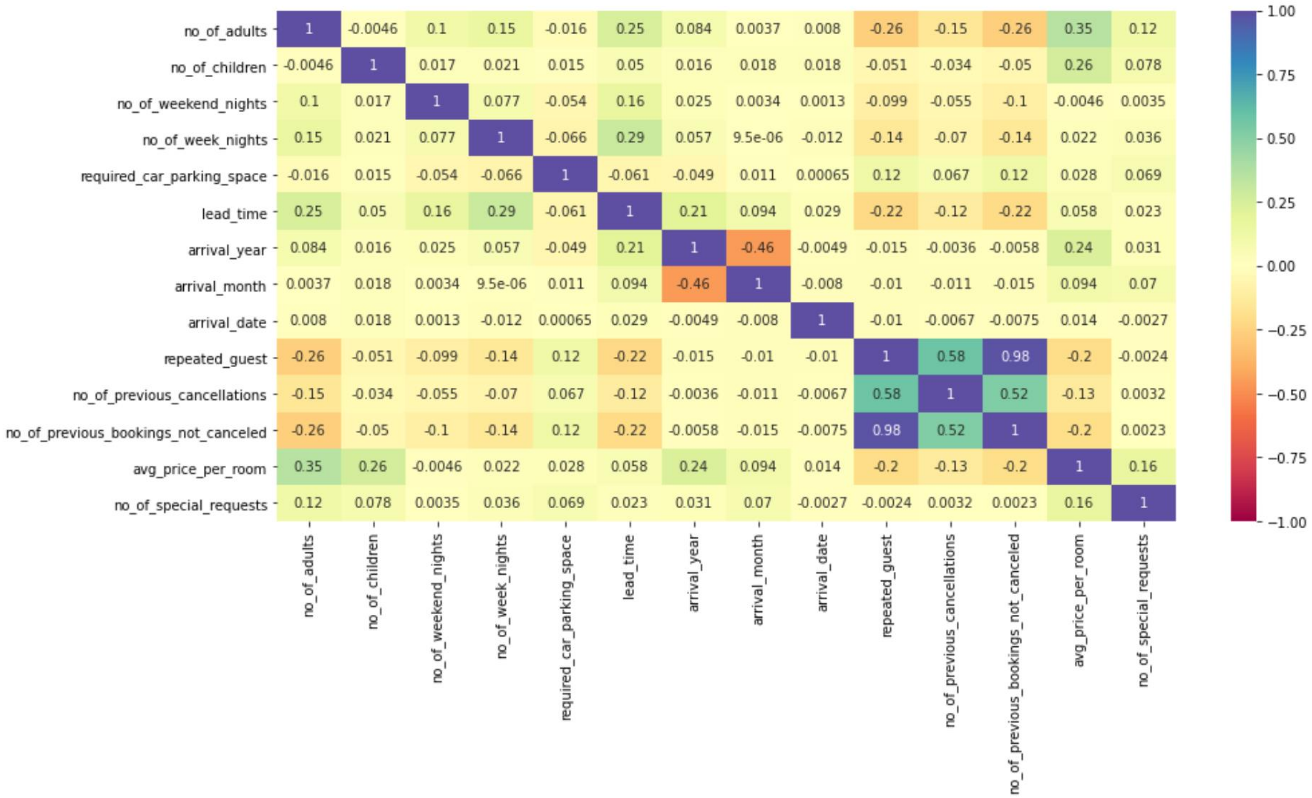




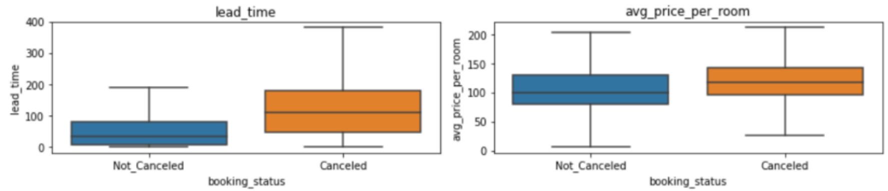


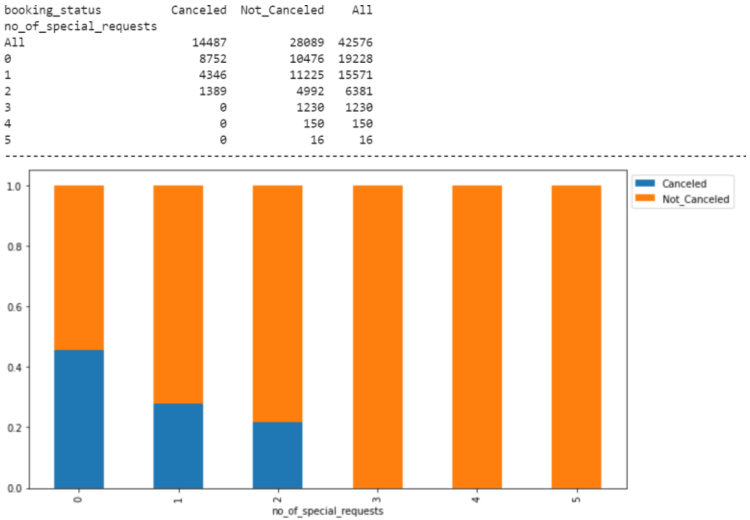


Exploratory Data Analysis - Bivariate Analysis

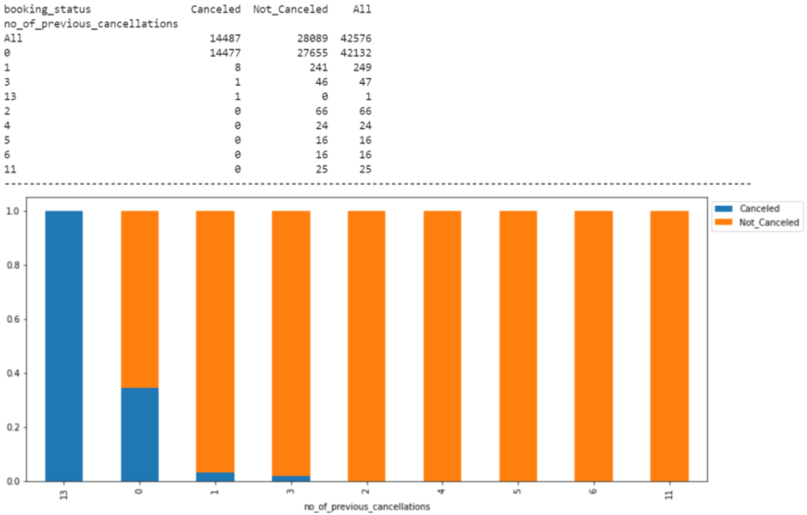


Exploratory Data Analysis - Bivariate Analysis graphs



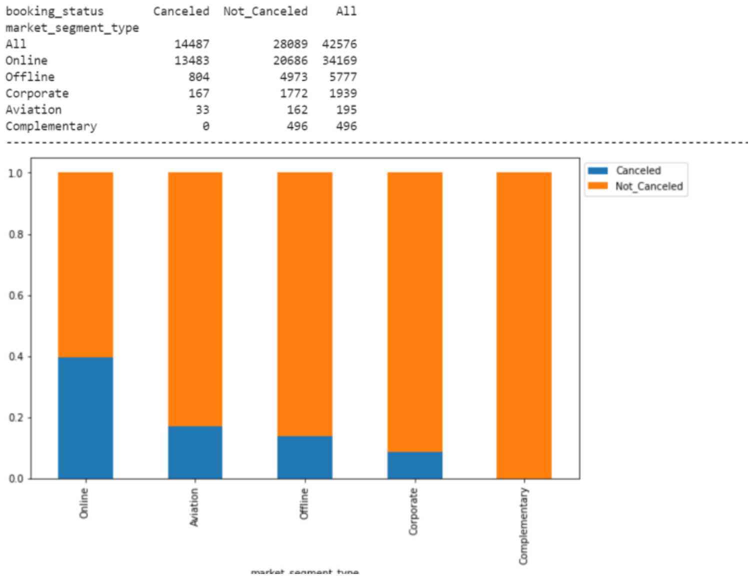


```
booking_status          Canceled  Not_Canceled    All
no_of_special_requests
All                        14487         28089  42576
0                           8752         10476  19228
1                           4346         11225  15571
2                           1389          4992   6381
3                              0          1230   1230
4                              0           150    150
5                              0            16     16
------------------------------------------------------------------------------------------------------------------------
```

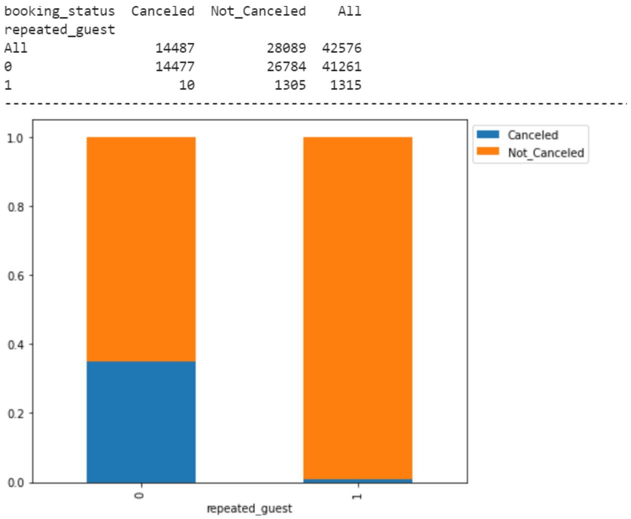


```
booking_status                 Canceled  Not_Canceled    All
no_of_previous_cancellations
All                               14487         28089  42576
0                                 14477         27655  42132
1                                     8           241    249
3                                     1            46     47
13                                    1             0      1
2                                     0            66     66
4                                     0            24     24
5                                     0            16     16
6                                     0            16     16
11                                    0            25     25
------------------------------------------------------------------------------------------------------------------------
```

Exploratory Data Analysis - Bivariate Analysis graphs



```
booking_status       Canceled  Not_Canceled    All
market_segment_type
All                     14487         28089  42576
Online                  13483         20686  34169
Offline                   804          4973   5777
Corporate                 167          1772   1939
Aviation                   33           162    195
Complementary               0           496    496
------------------------------------------------------------------------------------------------------------------------
```



```
booking_status  Canceled  Not_Canceled    All
repeated_guest
All                14487         28089  42576
0                  14477         26784  41261
1                     10          1305   1315
------------------------------------------------------------------------------------------------------------------------
```

Exploratory Data Analysis - Observations

Meal Plan 1 is mostly selected, Also 10072 bookings not selected for any meal plans.
Room_Type 1 is highest slected among bookings 42807.
Room_Type 3 is least selected among bookings.
Highest market_segment type is through online booking.
35378 bookings are not cancelled vs 21548 Cancelled.
There are outliers exists in the lead time
lead_time is right skewed.
avg_price_per_room is also right skewed.
average nightly price is 112.37
73% of the booking is done where occupants are 2 adults in the room
90% room booked with by the occupants with no kids.
41% of the time week ends nights are not booked.
weekends nights occupency for 1 night and 2 nights both are 28.8% and 28.2%
highest occupency are for 2 weeknights which is 27.6% followed by 1 night which is 25.6%
2018 and 2019 were among the busiest year with 51.9% and 38.9% arrivals.
July and August are busiest months with 11.1% and 12.5% arrivals.
96.9% guest were not a repeated guest only 3.1% guest were repeated guest.
Out of 7 catagories room types, Room catagory type 4 has highest booking.

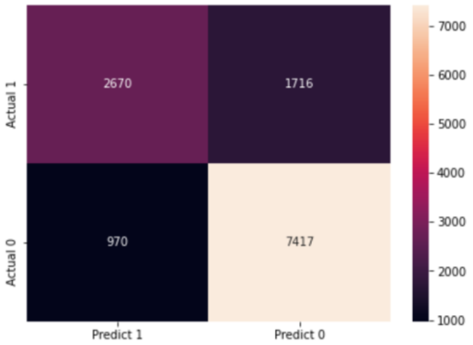
Exploratory Data Analysis - Observations

74.8% customers selected meal plan 1 and 20.5% customers did not selected any meal plan.
80.3 % bookins is done through online.
36% guest has some special request
66% customers not cancelled the booking v/s 34% customers cancelled.
no_of_repeated_guests is highly co-related with no_of_previous_bookings_not cancelled
max number of booking cancelled by customers which are booked for 2 adults stay in the room.
max number of booking cancelled by customers is in August month and also max booking took place in August.
max number of booking cancelled by customers who were not repeated guests, Repeated guests hardly cancelled any bookings.
max number of booking cancelled by customers who booked Room_Type 1 which is also had highest room type booked.
max number of booking cancelled by customers who opted for meal plan 1
max number of booking cancelled by customers who booked online
max number of booking cancelled by customers which got the rooms booked for an average_price_per_room of $115

Logistic regression overview and performance summary

- We used regular Logistic regression using sklearn and statsmodels library to build our machine learning model.
- We replaced the categorical data into the continuous format to fit into the models using the onehot columns method.
- We also split the data into the training and testing set in a ration of 70-30%.
- We build confusion matrix for both the models on and also checked the performance on train and test data.

Logistic regression model sklearn and performance summary



True Positives (TP): we correctly predicted that customer cancelled for 2670 bookings.

True Negatives (TN): we correctly predicted that customer will not cancel 7417

False Positives (FP): we incorrectly predicted that customer would cancel (a "Type I error") 970 Falsely predicted Type I error.

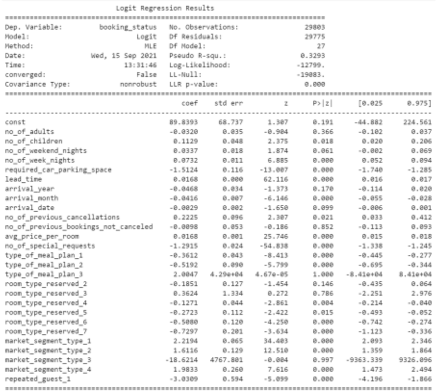
False Negatives (FN): we incorrectly predicted that customer will not cancel (a "Type II error") 1718 Falsely predicted negative Type II error.

The training and testing recall are 61.3% and 60.98% respectively.

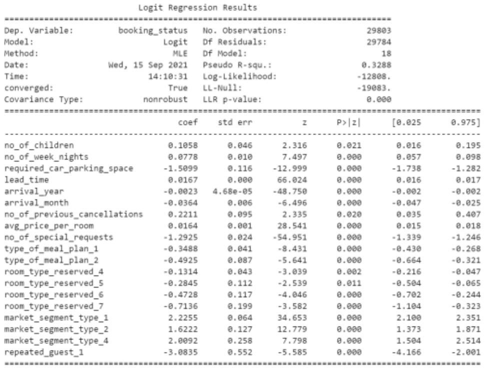
The training and testing f1_scores are 0.66

Recall on the train and test sets are comparable. f1_score on the train and test sets are comparable.

Logistic regression statsmodel performance summary



```
                           Logit Regression Results
==============================================================================
Dep. Variable:         booking_status   No. Observations:                29803
Model:                          Logit   Df Residuals:                    29775
Method:                           MLE   Df Model:                           27
Date:                Wed, 15 Sep 2021   Pseudo R-squ.:                  0.3293
Time:                        13:31:46   Log-Likelihood:                -12799.
converged:                      False   LL-Null:                       -19083.
Covariance Type:            nonrobust   LLR p-value:                     0.000
==================================================================================================
                                          coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------------------
const                                  89.8393     68.737      1.307      0.191     -44.882     224.561
no_of_adults                           -0.0320      0.035     -0.904      0.366      -0.102       0.037
no_of_children                          0.1129      0.048      2.375      0.018       0.020       0.206
no_of_weekend_nights                    0.0337      0.018      1.874      0.061      -0.002       0.069
no_of_week_nights                       0.0732      0.011      6.885      0.000       0.052       0.094
required_car_parking_space             -1.5124      0.116    -13.007      0.000      -1.740      -1.285
lead_time                               0.0168      0.000     62.116      0.000       0.016       0.017
arrival_year                           -0.0468      0.034     -1.373      0.170      -0.114       0.020
arrival_month                          -0.0416      0.007     -6.146      0.000      -0.055      -0.028
arrival_date                           -0.0029      0.002     -1.650      0.099      -0.006       0.001
no_of_previous_cancellations            0.2225      0.096      2.307      0.021       0.033       0.412
no_of_previous_bookings_not_canceled   -0.0098      0.053     -0.186      0.852      -0.113       0.093
avg_price_per_room                      0.0168      0.001     25.746      0.000       0.015       0.018
no_of_special_requests                 -1.2915      0.024    -54.838      0.000      -1.338      -1.245
type_of_meal_plan_1                    -0.3612      0.043     -8.413      0.000      -0.445      -0.277
type_of_meal_plan_2                    -0.5192      0.090     -5.799      0.000      -0.695      -0.344
type_of_meal_plan_3                     2.0047   4.29e+04   4.67e-05      1.000   -8.41e+04    8.41e+04
room_type_reserved_2                   -0.1851      0.127     -1.454      0.146      -0.435       0.064
room_type_reserved_3                    0.3624      1.334      0.272      0.786      -2.251       2.976
room_type_reserved_4                   -0.1271      0.044     -2.861      0.004      -0.214      -0.040
room_type_reserved_5                   -0.2723      0.112     -2.422      0.015      -0.493      -0.052
room_type_reserved_6                   -0.5080      0.120     -4.250      0.000      -0.742      -0.274
room_type_reserved_7                   -0.7297      0.201     -3.634      0.000      -1.123      -0.336
market_segment_type_1                   2.2194      0.065     34.403      0.000       2.093       2.346
market_segment_type_2                   1.6116      0.129     12.510      0.000       1.359       1.864
market_segment_type_3                 -18.6214   4767.801     -0.004      0.997   -9363.339    9326.096
market_segment_type_4                   1.9833      0.260      7.616      0.000       1.473       2.494
repeated_guest_1                       -3.0309      0.594     -5.099      0.000      -4.196      -1.866
==================================================================================================
```



```
                           Logit Regression Results
==============================================================================
Dep. Variable:         booking_status   No. Observations:                29803
Model:                          Logit   Df Residuals:                    29784
Method:                           MLE   Df Model:                           18
Date:                Wed, 15 Sep 2021   Pseudo R-squ.:                  0.3288
Time:                        14:10:31   Log-Likelihood:                -12808.
converged:                       True   LL-Null:                       -19083.
Covariance Type:            nonrobust   LLR p-value:                     0.000
================================================================================================
                                  coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------
no_of_children                  0.1058      0.046      2.316      0.021       0.016       0.195
no_of_week_nights               0.0778      0.010      7.497      0.000       0.057       0.098
required_car_parking_space     -1.5099      0.116    -12.999      0.000      -1.738      -1.282
lead_time                       0.0167      0.000     66.024      0.000       0.016       0.017
arrival_year                   -0.0023   4.68e-05    -48.750      0.000      -0.002      -0.002
arrival_month                  -0.0364      0.006     -6.496      0.000      -0.047      -0.025
no_of_previous_cancellations    0.2211      0.095      2.335      0.020       0.035       0.407
avg_price_per_room              0.0164      0.001     28.541      0.000       0.015       0.018
no_of_special_requests         -1.2925      0.024    -54.951      0.000      -1.339      -1.246
type_of_meal_plan_1            -0.3488      0.041     -8.431      0.000      -0.430      -0.268
type_of_meal_plan_2            -0.4925      0.087     -5.641      0.000      -0.664      -0.321
room_type_reserved_4           -0.1314      0.043     -3.039      0.002      -0.216      -0.047
room_type_reserved_5           -0.2845      0.112     -2.539      0.011      -0.504      -0.065
room_type_reserved_6           -0.4728      0.117     -4.046      0.000      -0.702      -0.244
room_type_reserved_7           -0.7136      0.199     -3.582      0.000      -1.104      -0.323
market_segment_type_1           2.2255      0.064     34.653      0.000       2.100       2.351
market_segment_type_2           1.6222      0.127     12.779      0.000       1.373       1.871
market_segment_type_4           2.0092      0.258      7.798      0.000       1.504       2.514
repeated_guest_1               -3.0835      0.552     -5.585      0.000      -4.166      -2.001
================================================================================================
```

None of the variables exhibit high multicollinearity, so the values in the summary are reliable.

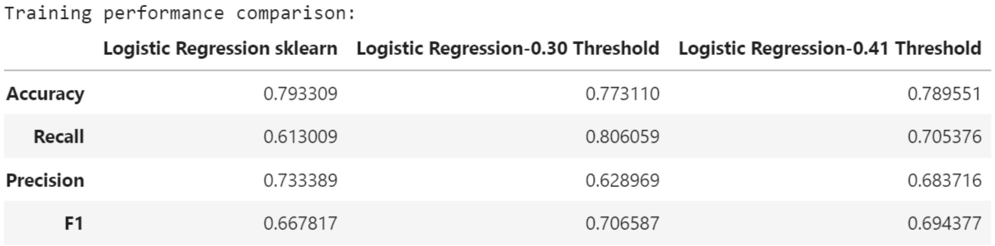
We removed the features which had p-value > 0.05, both the results are attached.

None of the variables exhibit high multicollinearity, so the values in the summary are reliable.

Coefficient of no_of_children,

no_of_week_nights,lead_time, no_of_previous_cancellations, avg_price_per_room, market_segment_type_1, market_segment_type_2, market_segment_type_4 are positive an increase in these will lead to increase in chances of a person cancelling the bookings.

Coefficient of required_car_parking_space, arrival_year,arrival_month, no_of_special_requests, type_of_meal_plan, room_type_reserved and repeated_guest are negative and an increase in these will lead to decrease in chances of a person cancelling booking



Training performance comparison:

	Logistic Regression sklearn	Logistic Regression-0.30 Threshold	Logistic Regression-0.41 Threshold
Accuracy	0.793309	0.773110	0.789551
Recall	0.613009	0.806059	0.705376
Precision	0.733389	0.628969	0.683716
F1	0.667817	0.706587	0.694377

Model overview and performance summary

Conclusion

We kept improving on our model and reached to the model which provided a generalized performance on training and test set.
The highest recall was 80% on the training set.

Using the model with default threshold the model gave low recall but good precision scores - This model will achieve the hotel save resources but lose on potential customers.

Using the model with 0.30 threshold the model gave a high recall but low precision scores - This model will help the hotel identify potential cancellations effectively, but the cost of resources will be high.

Using the model with 0.41 threshold the model will give a balance recall and precision score - This model will help the hotel to maintain a balance in identifying potential customer and the cost of resources.

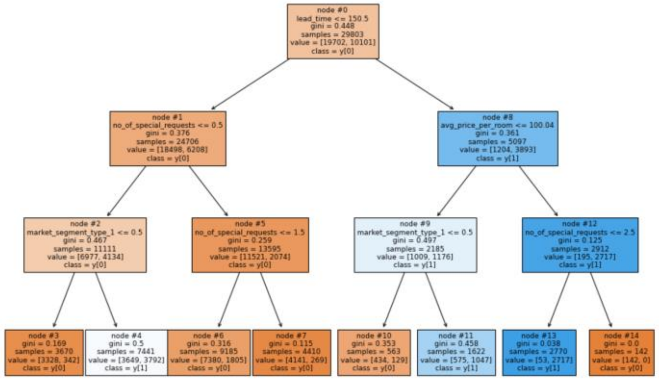
Decision Tree Model overview and performance summary

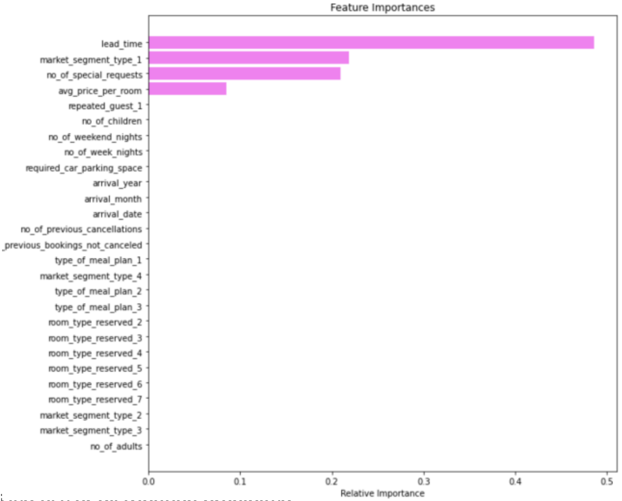
We also evaluated the data with the decision tree model and figured out that with the default depth the model was overfitted.

We then pruned the decision tree with maximum depth 3 and 5 respectively and strived to get the best fitted model with ideal recall score on train and test data.

Recall on training set was reduced from 0.99 to 0.75 so we reached to the conclusion that this model is not overfitting and we have a generalized model.

Finally, we got the most important feature with the feature of importance which can main effective features in hotel booking cancellation.





Decision Tree Model overview and performance summary

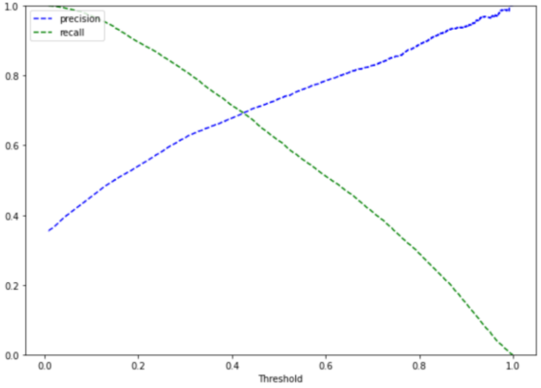
We also evaluated the data with the decision tree model and figured out that with the default depth the model was overfitted.

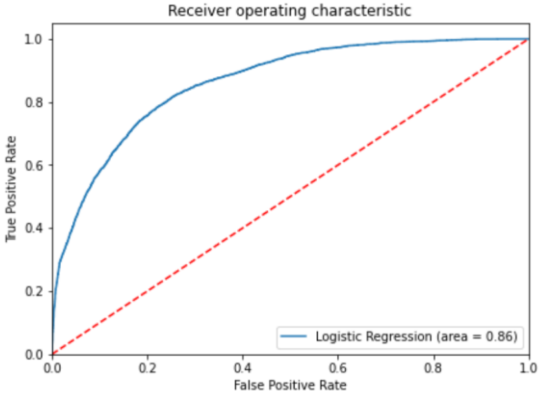
We then pruned the decision tree with maximum depth 3 and 5 respectively and strived to get the best fitted model with ideal recall score on train and test data.

Recall on training set was reduced from 0.99 to 0.75 so we reached to the conclusion that this model is not overfitting and we have a generalized model.

Finally, we got the most important feature with the feature of importance which can main effective features in hotel booking cancellation.

Precision Recall curve at threshold around 0.41 shows that we got a balanced recall and precision.





Key Findings and Insights

We analyzed the "StarHotels Group" booking data using different techniques and used Decision Tree Classifier to build a predictive model for the same.

The model built can be used to predict if a customer is going to cancel the booking or not.

We visualized different trees and their confusion matrix to get a better understanding of the model. Easy interpretation is one of the key benefits of Decision Trees.

We observed the fact that much lesser data preparation is needed for Decision Trees and such a simple model gave good results even with no outliers treatment which shows the robustness of Decision Trees.

Lead Time, Market Segment Type, Avg Price per room, and No_Of_Special Requests are the most important variable in predicting the customers cancellation behavior

We established the importance of hyper-parameters/ pruning to reduce overfitting.

Key Findings and Insights

We can make observations such as from decision tree that gives us an indication of kind of bookings we need to observe:

Lead Time > 9.5 and lead_time <= 150.50
no_of_special_requests <= 0.50 ,
market_segment_type_1 > 0.50 then the booking is more likely to cancel

If lead_time > 150.50 but avg_price_per_room <= 100.04, customer has a higher chance of cancelling.
if lead_time > 150.50 and even if avg_price_per_room > 100.04, The customer has lesser special requests than 2.5, then booking is likely to cancel.

If lead_time > 150.50 but avg_price_per_room <= 100.04, customer has a higher chance of cancelling
if lead_time > 150.50 and even if avg_price_per_room > 100.04, if the customer has lesser special requests than 2.5, she is very likely to cancel

Recommendations

Customer booking Lead Time is one of the most important feature to determine the cancellation odds i.e.if the customer will be cancelling the booking

Market_Segment_Type is another important field for determining the customer cancellation trend

Avg_Price Per room is indicated by more than one tree models as an important field for determining the customer cancellation trend

No_Of_Special Request is a factor in determining the customer cancellation trend.

Lead Time is a key indicator . Stat Hotel can run marketing schemes to incentivize customers to book in advance along with special offers and requests to lock the booking and make them aware through channels to ensure that they are more engaged and less likely to cancel

It is observed that 36% guests had special requests and were less likely to cancel the bookings. This depicts clear trend across Customers who do adequate planning with room details are less likely to cancel.

Star Hotel should engage with the customers and increase this data collection resulting in deeper engagement.

Repeat guests % is lower (3.1%) but they were least likely to cancel. Star Hotel should run special discounts for customers with schemes for Repeat guests as it may improve the non-cancellation trends.

max number of booking cancelled by customers is in August month and also max booking took place in August so to compensate this loss Hotel can come up with a Hotel Credit which can be used for future bookings.

Recommendations

Repeated guests hardly cancelled any bookings, but repeated guest bookings are also very less, Hotel should focus on this and they should improve the count on repeated guest by providing them more intensives by offering them memberships.

Online Bookings are the maximum counts which are getting cancelled. Hotel can use engagement techniques such as booking confirmation calls, follow-ups etc. to ensure that there is deeper engagement with the customer.

Average nightly rates is an important factor influencing cancellation. We observe that Room Type 1 (potentially the lowest segment type) is cancelled maximum times. Hotel can deploy an algorithm which ensure that cancellation factor is marginally included in the pricing algorithm.

Most booking is cancelled by customers couples (2 adults) with zero children. Star Hotels pricing algorithm should take this into account while pricing for couple vs families. That way it can recover some of the costs. The model can follow grades for cancellation charges with lowest refunds on last minute cancellations, 20-50% refunds within a week of booking vs 70% or more for cancellations done prior to 3 weeks of booking start date

Since there is 41% vacancy over weekends, the Hotel can offer weekend bookings instead of cancellation refunds/charges. That way, customer would be locked

greatlearning
Power Ahead

Happy Learning !