PuriCloud

Hotel Booking Cancellation Predictor

5/17/2023

Author: Bradley D. Castle

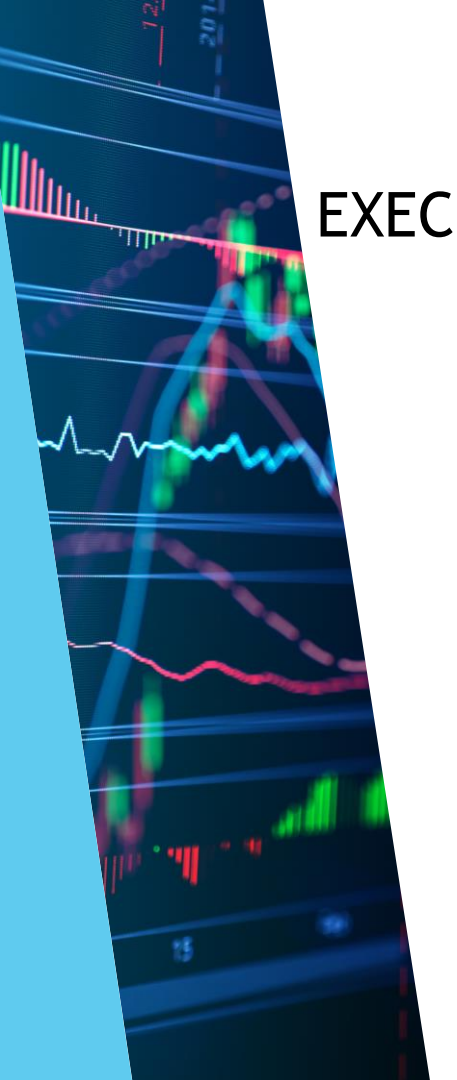
PuriCloud

CONTENTS/AGENDA

PuriCloud

EXECUTIVE SUMMARY: ACTIONABLE INSIGHTS

- **Key Factors:** There are several key factors that drive cancellations. These include the lead time between booking and arrival, and the number of previous cancellations made by the guest, and the type of meal plan selected.
- **Key Insights:**
  - The majority of bookings were made online, with only a few made offline.
  - The most common meal plan selected was Meal Plan 1.
  - The average price per room for bookings that were not canceled was around $100-$130.
  - There were several instances of guests canceling multiple bookings in the past.
  - The lead time between booking and arrival were significantly different.



PuriCloud

EXECUTIVE SUMMARY: RECOMMENDATIONS

Non-Refundable Rate Incentives: We recommend that the hotel offer incentives such as discounts or upgrades for guests who choose non-refundable rates which may encourage them to keep their reservations.

Increase Communication: Increase and improve communication with guests during the lead time between booking and arrival. This could include sending reminder emails or text messages closer to the arrival date, as well as providing more detailed information about local attractions and events that may be of interest to guests.

Flexible Meal Plans: Offer more flexible meal plan options. Allowing guests to switch between meal plans up until a certain point before their arrival date may help reduce cancellations due to changes in travel plans or preferences.

PuriCloud

RESEARCH OVERVIEW

The introduction of advanced technologies and online platforms has influenced the landscape of customer booking capabilities and behavioral patterns. This has contributed to challenges and complexity with hotel cancellation management.

The cancellation of bookings impact the hotel on various fronts:

- Loss of resources (revenue) when the hotel cannot resell the room.
- Additional costs of distribution channels by increasing commissions or paying for publicity to help sell these rooms.
- Lowering prices last minute, so the hotel can resell a room, resulting in reducing the profit margin.
- Human resources to make arrangements for the guests.

PuriCloud

METHOD OVERVIEW

Objective: The increase of cancellations necessitates a Machine Learning (ML) solution to predict and reduce potential booking cancellations. The hospitality industry is grappling with excessive booking cancellations and could benefit from a data driven methodology. As a data scientist, the provided data will be critically analyzed to identify factors with high impact on cancellations, with the objective of developing a highly accurate predictive model to reduce potential cancellations and generate profitability.

Methodology:

1. Analyze data to identify key factors influencing cancellations.
2. Preprocess the data to prepare for model training.
3. Select and train the model to predict cancellations.
4. Evaluate and refine the model for accuracy.
5. Devise profit driven cancellation and refund policies.

PuriCloud

DATA OVERVIEW

The data provided is of various hotel bookings and attributes containing information on hotel bookings, including booking IDs, number of guests, meal plans, room types, arrival dates, market segments, and booking statuses. It also includes details on cancellations and previous bookings.

Data Attributes:

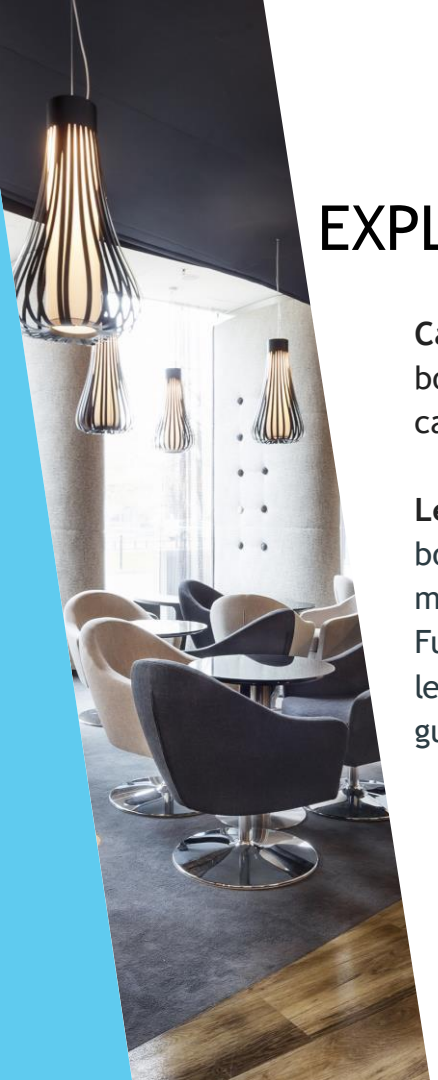
- **booking_ID:** the unique identifier of each booking
- **no_of_adults:** Number of adults
- **no_of_children:** Number of Children
- **no_of_weekend_nights:** Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
- **no_of_week_nights:** Number of weeknights (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:
- **Not Selected:** No meal plan selected
- **Meal Plan 1:** Breakfast
- **Meal Plan 2:** Half board (breakfast and one other meal)
- **Meal Plan 3:** Full board (breakfast, lunch, and dinner)

PuriCloud

DATA OVERVIEW

Additional Data Attributes:

- **required_car_parking_space:** Customer parking space? (0 - No, 1- Yes)
- **room_type_reserved:** Type of room reserved. The values are ciphered
- **lead_time:** 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
- **no_of_previous_bookings_not_canceled:** Number of previous bookings kept
- **avg_price_per_room:** Average price per day of the reservation
- **no_of_special_requests:** Total number of special requests
- **booking_status:** Flag indicating if the booking was canceled or not.

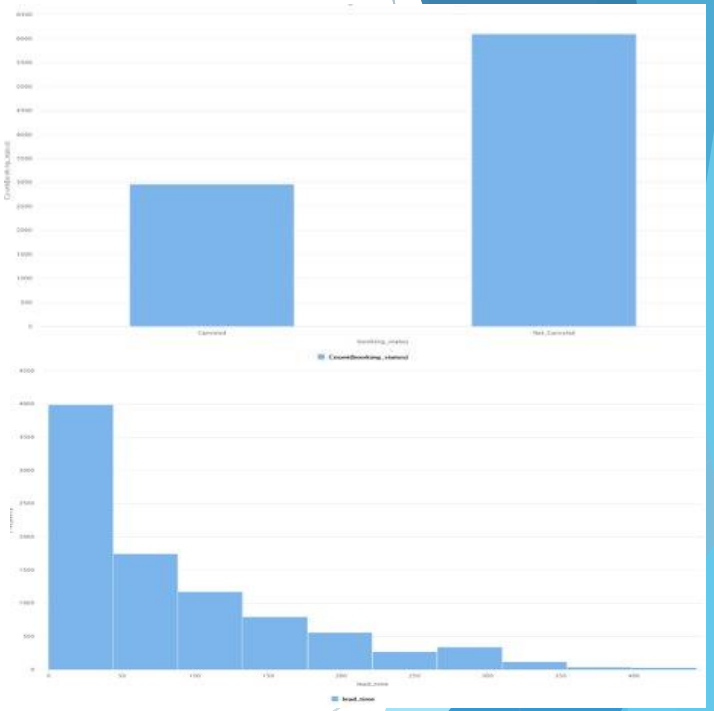


PuriCloud

EXPLORATORY DATA ANALYSIS

Cancelled Bookings: Out of 6,098 bookings, 2,971 bookings were cancelled (33% cancelled)

Lead Time: The lead time indicates that the bookings made far in advance or at the last minute may be more likely to be canceled. Further analyzing cancellation rates based on lead time may provide additional insights into guest behavior and preferences.

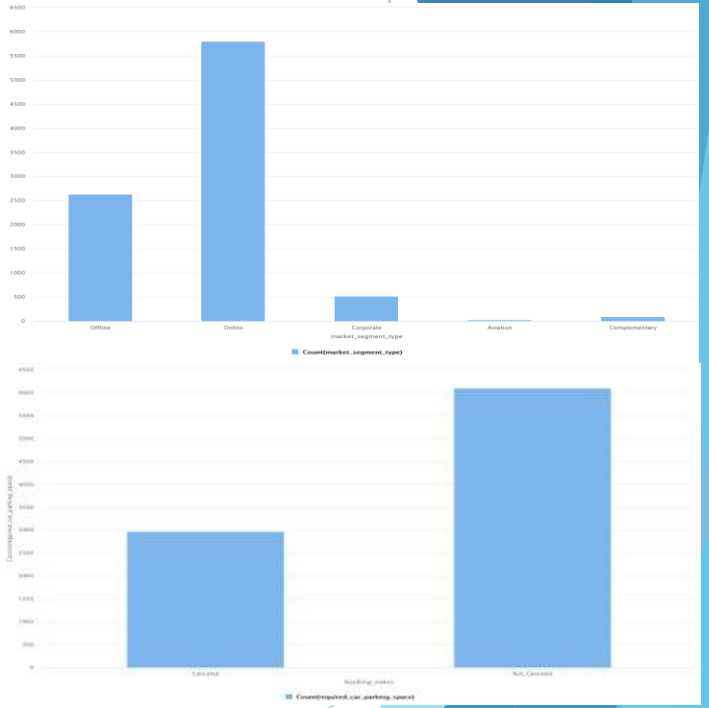


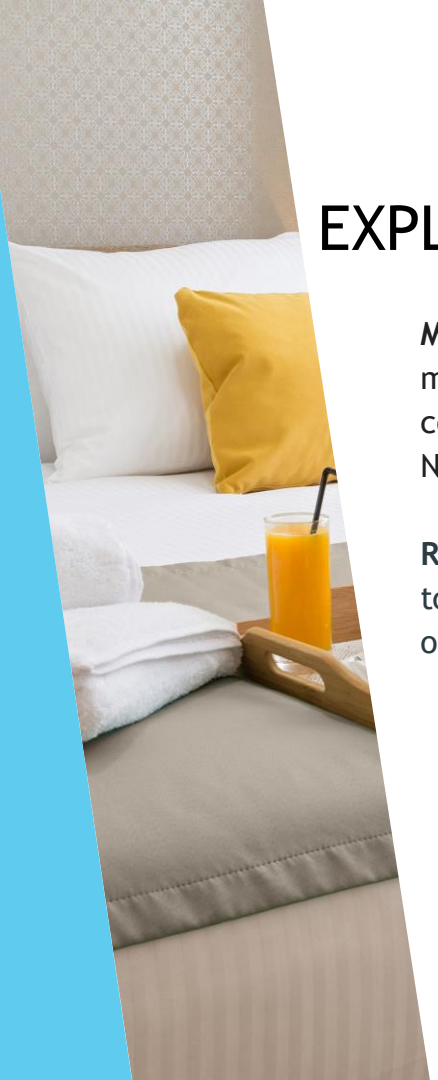
PuriCloud

EXPLORATORY DATA ANALYSIS

Market Segment: The type of market segment that the booking belongs to, such as corporate, group, or leisure, may have different cancellation patterns. Majority cancelled online.

Required Parking: Bookings that require parking may be less likely to cancel as they have already made arrangements for transportation.



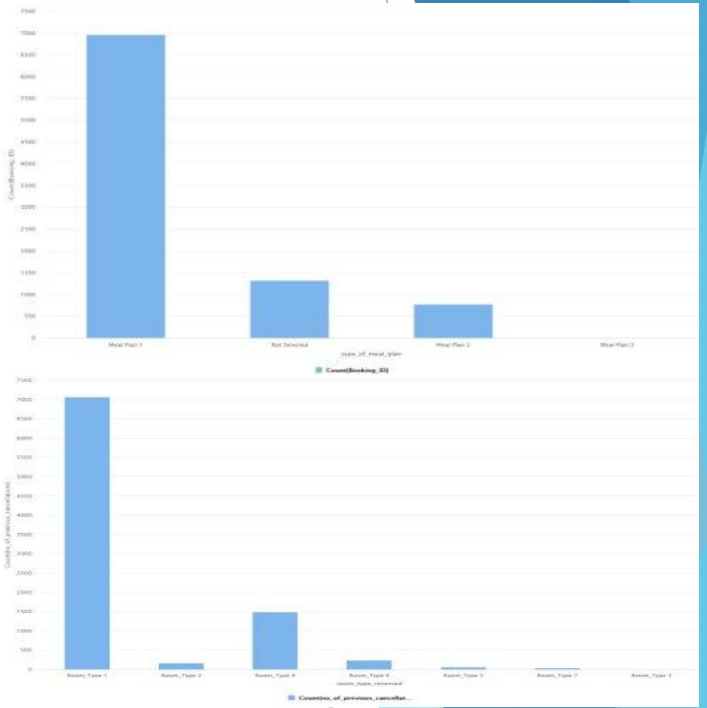


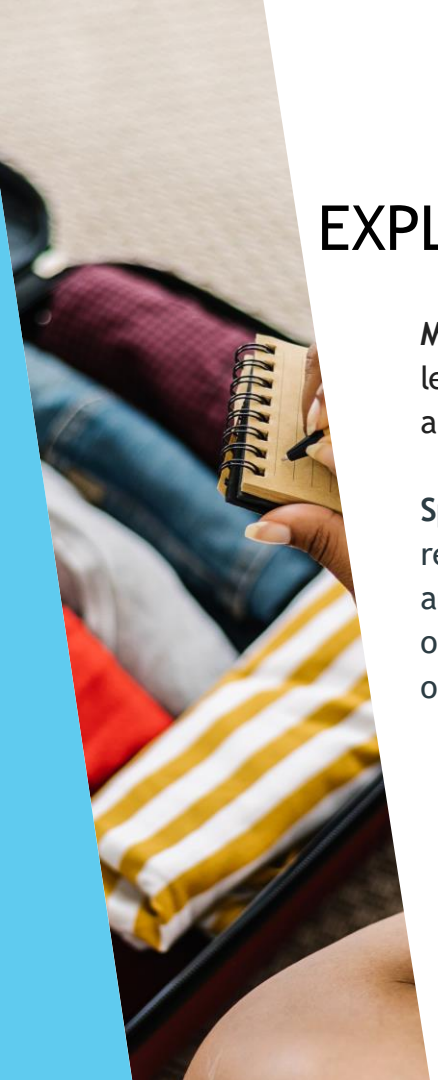
PuriCloud

EXPLORATORY DATA ANALYSIS

Meal Plan: Bookings with Meal Plan 1 appear more likely to cancel than Meal Plan 2. A filter condition excluding Meal Plan 3 and Not Selected may or may not be appropriate.

Room Type: Bookings for Room Type 1 appear to be more likely to be canceled than those for other room types.



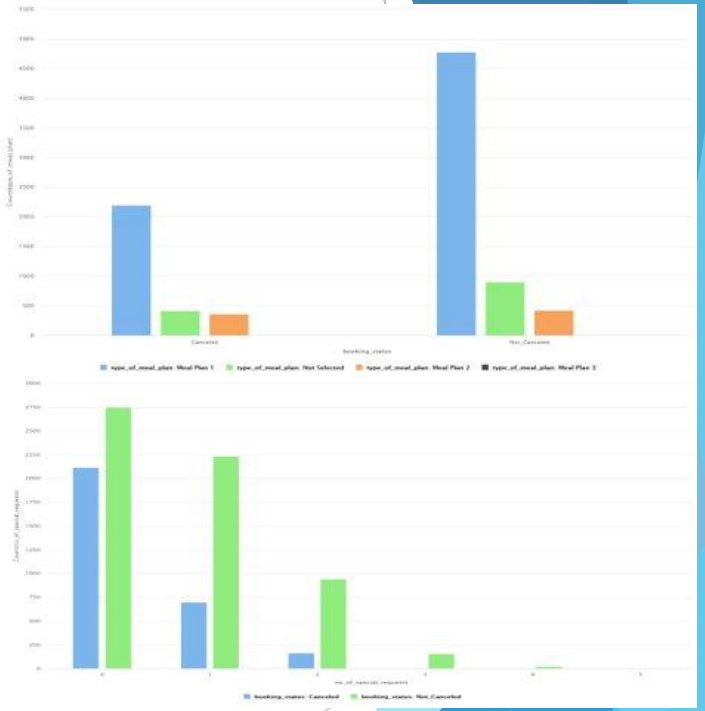


PuriCloud

EXPLORATORY DATA ANALYSIS

Meal Plan: Bookings with meal plans may be less likely to be canceled as customers have already paid for their meal plan.

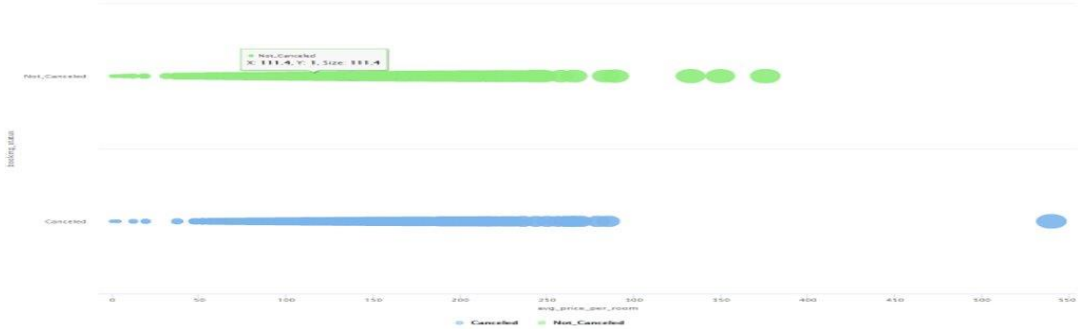
Special Requests: Bookings with more special requests may be more invested in their stay and less likely to cancel as the fewer number of special requests the greater percentage of cancellations to bookings not cancelled.



PuriCloud

EXPLORATORY DATA ANALYSIS

Price Per Room: Exploring each booking with the average price per room on the x-axis and a binary indicator showing canceled vs. not cancelled bookings on the y-axis illustrates the relationship between price and cancellation rate. The higher priced bookings appear to be less likely to be canceled, and bookings with higher prices seem to be more likely to fall in the not canceled category. Possible outlier in cancelled booking.



PuriCloud

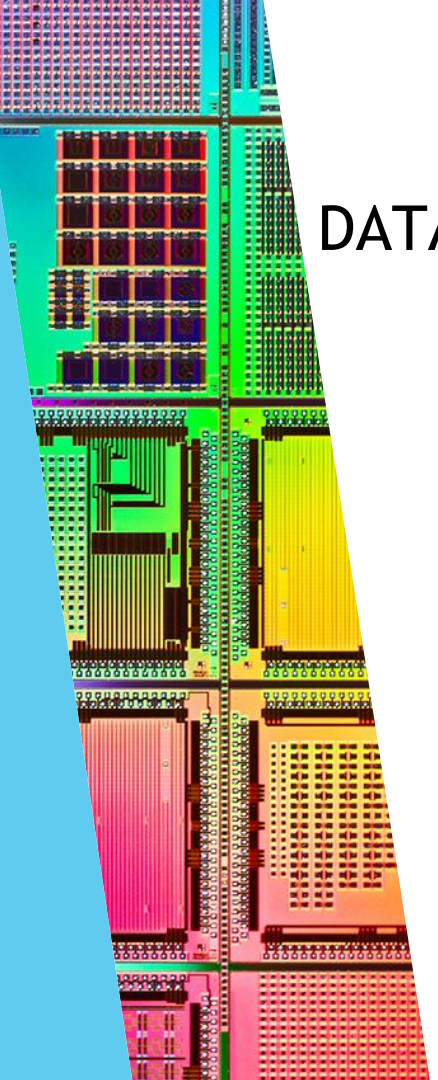
CORRELATION MATRIX

- A significant positive correlation (0.509) between repeated_guests and no_of _previous_bookings_not_canceled in dicates likely predictive value for avoiding cancellation behavior.
- A significant positive correlation (0.499) between no_of_previous cancellations and no_of_previous_bookin gs_not_cancelled suggests a useful predictive measure for lower cancellation rates.
- The moderate negative correlation (-0.441) between booking status and lead time indicates longer lead times may reduce booking cancellation probability.
- The negative correlation (-0.022) between the number of adults and children suggests an insignificant inverse relationship, providing minimal predictive utility for hotel cancellation trends among families.
- A moderate positive correlation (0.290) between adult count and average room price implies a possible association, yet insufficiently strong enough to predict cancellations definitively.

PuriCloud

CORRELATION MATRIX

Attributes	no_of_adults	no_of_children	no_of_weekend_nights	no_of_week_nights	required_car_parking_space	arrival_month	arrival_date	repeated_guest	arrival_date	repeated_guest	no_of_previous_cancellations	no_of_previous_bookings_not_canceled	avg_price_per_room	no_of_special_requests	booking_status	lead_time
no_of_adults	1	-0.022	0.118	0.099	-0.011	0.014	0.011	-0.193	0.011	-0.193	-0.045	-0.118	0.290	0.191	-0.090	0.104
no_of_children	-0.022	1	0.019	0.015	0.047	0.001	0.022	-0.036	0.022	-0.036	-0.017	-0.020	0.330	0.116	-0.029	-0.044
no_of_weekend_nights	0.118	0.019	1	0.197	-0.043	-0.029	0.019	-0.060	0.019	-0.060	-0.015	-0.015	-0.005	0.079	-0.058	0.052
no_of_week_nights	0.099	0.015	0.197	1	-0.064	0.032	-0.003	-0.090	-0.003	-0.090	-0.015	-0.023	0.013	0.052	-0.106	0.160
required_car_parking_space	-0.011	0.047	-0.043	-0.064	1	-0.015	-0.004	0.115	-0.004	0.115	0.028	0.063	0.053	0.084	0.090	-0.072
arrival_month	0.014	0.001	-0.029	0.032	-0.015	1	-0.032	0.009	-0.032	0.009	-0.044	0.004	0.054	0.103	0.023	0.128
arrival_date	0.011	0.022	0.019	-0.003	-0.004	-0.032	1	-0.022	1	-0.022	-0.007	0.002	0.016	0.016	-0.008	0.001
repeated_guest	-0.193	-0.036	-0.060	-0.090	0.115	0.009	-0.022	1	-0.022	1	0.397	0.509	-0.162	-0.022	0.111	-0.143
no_of_previous_cancellations	-0.045	-0.017	-0.015	-0.015	0.028	-0.044	-0.007	0.397	-0.007	0.397	1	0.499	-0.059	0.002	0.037	-0.052
no_of_previous_bookings_not_ca...	-0.118	-0.020	-0.015	-0.023	0.063	0.004	0.002	0.509	0.002	0.509	0.499	1	-0.097	0.019	0.058	-0.077
avg_price_per_room	0.290	0.330	-0.005	0.013	0.053	0.054	0.016	-0.162	0.016	-0.162	-0.059	-0.097	1	0.183	-0.131	-0.069
no_of_special_requests	0.191	0.116	0.079	0.052	0.084	0.103	0.016	-0.022	0.016	-0.022	0.002	0.019	0.183	1	0.254	-0.099
booking_status	-0.090	-0.029	-0.058	-0.106	0.090	0.023	-0.008	0.111	-0.008	0.111	0.037	0.058	-0.131	0.254	1	-0.441
lead_time	0.104	-0.044	0.052	0.160	-0.072	0.128	0.001	-0.143	0.001	-0.143	-0.052	-0.077	-0.069	-0.099	-0.441	1



PuriCloud

DATA PREPROCESSING

Missing data: There are some missing values in the dataset, such as missing values for the "required_car_parking_space" variable.

Irrelevancy: The following variables were removed based on relevancy to my use case of the model: Booking_ID, arrival_year.

Categorical Variable Reduction: Reduced by using principal component analysis (PCA).

Split: Dataset has been split into training and testing set in the ration of 70:30.

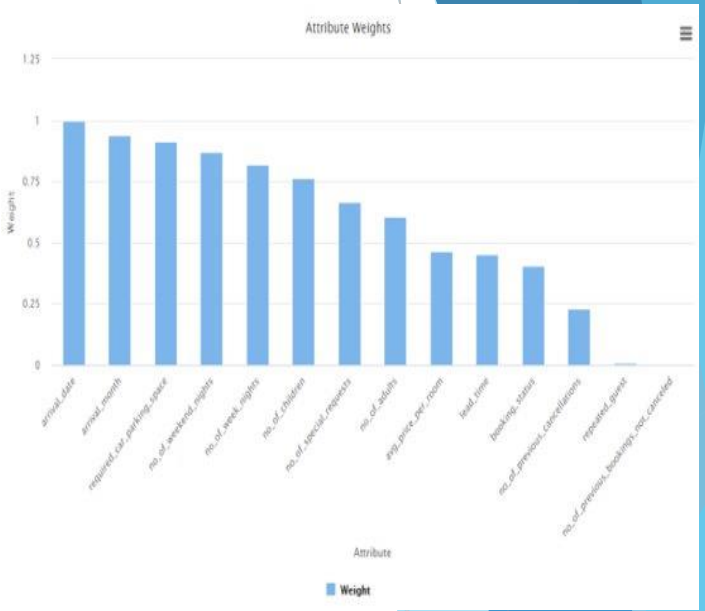
Outliers: Several variables were filtered to exclude outliers.

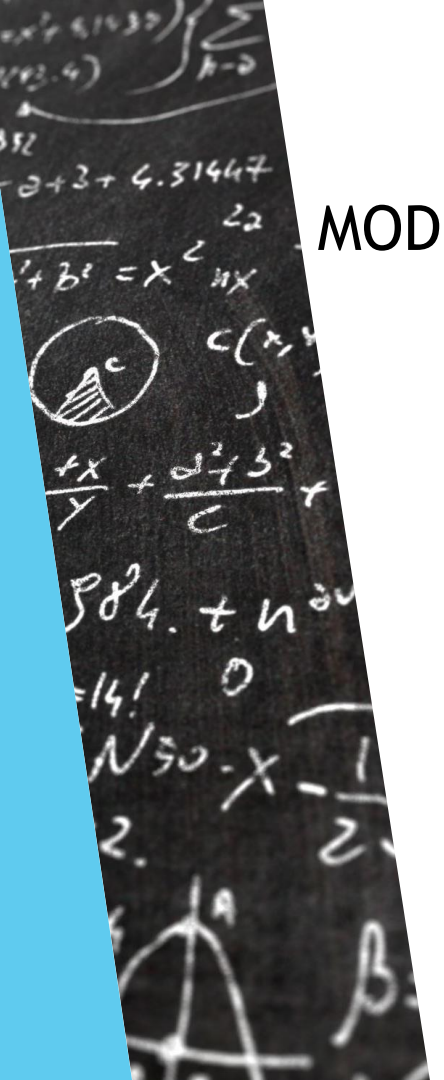
PuriCloud

MODEL BUILDING

The following steps were used to build the models:

1. Data preparation
2. Model building
3. Evaluate the performance on training set
4. Evaluate the performance on test set
5. Check for important features





PuriCloud

MODEL BUILDING

Decision Tree vs Random Forest

In comparing the four comparative Machine Learning models constructed:

1. The **Decision Tree** showed 79.89% training accuracy and 79.89% test accuracy, with 74.33% recall respectively.
2. The **Pruned Decision Tree** showed 93.74 training accuracy and 86.69% test accuracy, with 92.41% training recall and 86.67% test recall.
3. The **Random Forest** model had 77.70% training accuracy, 73.49% test accuracy, 66.17% training recall, and 60.29% test recall.
4. The **Pruned Random Forest** outperformed other models with 95.64% training accuracy, 93.60% test accuracy, 90.53% training recall, and 89.26% test recall.

PuriCloud

MODEL BUILDING

Model	Train Accuracy	Test Accuracy	Train Recall	Test Recall
Decision Tree	79.89 %	79.89 %	74.33 %	74.33 %
Decision Tree Pruned	93.74 %	86.69 %	92.41 %	86.67 %
Random Forest	77.70 %	73.49 %	66.17 %	60.29 %
Random Forest Pruned	95.64 %	93.60 %	90.53 %	89.26 %

PuriCloud

MODEL BUILDING

Selection: Random Forest

The Decision Tree provided straightforward interpretation, while the pruning reduced overfitting. The Random Forest models enhanced generalization through ensemble learning, with pruning limiting complexity. The sequential development of the models allowed for the comparison of complexity, interpretability, and predictive performance.

While the Decision Tree model provided high interpretability, the performance was weaker than the Random Forest models. In considering the performance vs interoperability, the Random Forest model with pruning was the preferred model for predictive accuracy and overall performance.

The Random Forest model better handled the large and complex dataset and reduced overfitting by combining multiple decision trees using random subsets of features for each tree.

PuriCloud

MODEL PERFORMANCE SUMMARY

Methods used to evaluate the importance of each feature were Information Gain and Correlation methods.

- The Information Gain quantified entropy reduction when splitting data.
- The arrival_date (1) feature was showing as the most important, while arrival_month (0.940), and required_car_parking_space (.914) are the other important features due to high values, which indicate they significantly reduce uncertainty when used to split data.
- The no_of_previous_bookings_not_canceled (0) is the least important, with repeated_guest (.010) also causing minimal uncertainty reduction.
- The correlation coefficients measured linear relationships with values near ±1 indicating strong relationships.

PuriCloud

MODEL PERFORMANCE SUMMARY

attribute	weight
no_of_adults	0.606
no_of_children	0.765
no_of_weekend_nights	0.872
no_of_week_nights	0.817
required_car_parking_space	0.914
arrival_month	0.940
arrival_date	1
repeated_guest	0.010
no_of_previous_cancellations	0.232
no_of_previous_bookings_not_canceled	0
avg_price_per_room	0.464
no_of_special_requests	0.664
booking_status	0.407
lead_time	0.452

PuriCloud

MODEL PERFORMANCE SUMMARY

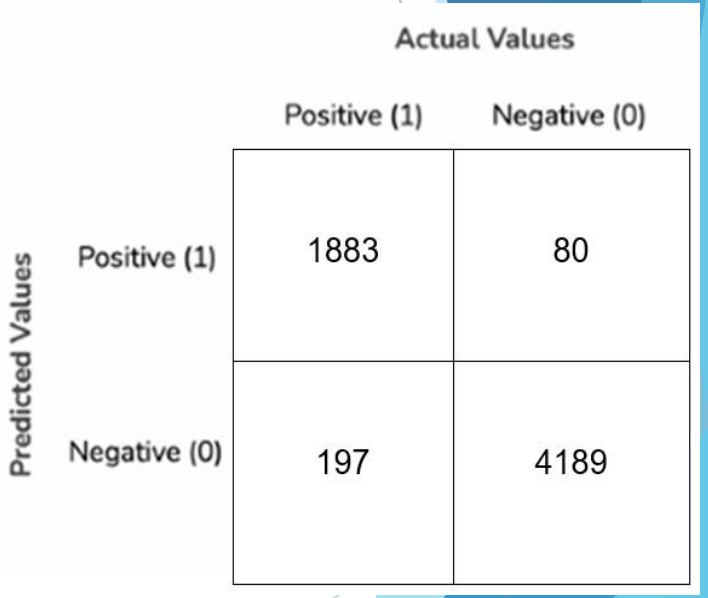
Model predictions compared to outcomes:

True Positive (TP): Correctly predicted cancellations = 1883

False Positives (FP): Incorrectly predicted cancellations (actually not cancelled) = 80

True Negatives (TN): Correctly predicted not cancelled = 4189

False Negatives (FN): Incorrectly predicted not cancelled (actually cancelled)=197



PuriCloud

MODEL PEFORMANCE SUMMARY

Confusion Matrix calculations:

Accuracy: Proportion of correctly predicted results among the total number of observations:

- (TP + TN) / (TP + FP + FN + TN) = (1883 + 4189) / (1883 + 80 + 197 + 4189) = 95.64%

Precision: Proportion of true positives to all the predicted positives, i.e. how valid the predictions are:

- TP / (TP + FP) = 1883 / (1883 + 80) = 95.92%

Recall: Proportion of true positives to all the actual positives, i.e., how complete the predictions are:

- TP / (TP + FN) = 1883 / (1883 + 197) = 90.53%

PuriCloud

MODEL PEFORMANCE SUMMARY

The performance vector confirms the above calculations:

- Accuracy of 95.64%
- Precision (weighted mean precision) of 95.92%
- Recall (weighted mean recall) of 90.53%

PuriCloud

APPENDIX

PuriCloud

DATA BACKGROUND AND CONTENTS

9,069 Rows: Each row represents information about a booking made from a customer.

20 Columns: Every column includes an attribute/feature related to an individual booking.

Row No.	Booking_ID	no_of_adults	no_of_childr...	no_of_week...	no_of_week...	type_of_me...	required_ca...	room_type_...	lead_time	arrival_year	arrival_month	arrival_date	market_seg...	repeated
1	INN23152	1	0	0	2	Meal Plan 1	0	Room_Type 1	188	2018	6	15	Offline	0
2	INN21915	1	0	0	2	Meal Plan 1	0	Room_Type 1	103	2018	4	19	Offline	0
3	INN24290	2	0	1	4	Not Selected	0	Room_Type 1	33	2018	4	18	Online	0
4	INN31921	2	0	0	3	Meal Plan 1	0	Room_Type 1	64	2018	11	22	Online	0
5	INN34718	2	0	1	1	Meal Plan 2	0	Room_Type 1	247	2018	6	6	Offline	0
6	INN31303	2	0	0	3	Meal Plan 1	0	Room_Type 1	304	2018	11	3	Offline	0
7	INN34963	1	0	3	5	Not Selected	0	Room_Type 1	275	2018	10	9	Online	0
8	INN14729	2	0	2	0	Meal Plan 1	0	Room_Type 1	146	2018	4	24	Offline	0
9	INN06771	2	2	2	3	Meal Plan 1	0	Room_Type 2	41	2018	9	4	Online	0
10	INN34053	2	0	2	1	Meal Plan 1	0	Room_Type 4	41	2018	9	18	Online	0
11	INN12335	2	0	1	0	Not Selected	0	Room_Type 1	10	2018	3	13	Online	0
12	INN32422	3	0	2	1	Meal Plan 1	0	Room_Type 4	128	2018	10	29	Online	0
13	INN07271	1	0	0	1	Meal Plan 1	0	Room_Type 1	177	2018	7	30	Online	0

PuriCloud

LINEAR RELATIONSHIPS

The correlation matrix was used to measure the linear relationships between variables in the dataset. The following are some key equations related to generating the correlation matrix:

The Pearson correlation coefficeint (r) measures the linear relationship between two variables, X and Y: $r = \Sigma((x_i - \bar{x})(y_i - \bar{y})) / (\sqrt{\Sigma}(x_i - \bar{x})^2 * \sqrt{\Sigma}(y_i - \bar{y})^2)$

Spearman's Rank Correlation Coefficient (ρ) measures the monotonic relationship between x and Y: $\rho = 1 - (6 * \Sigma d_i^2) / (n * (n^2 - 1))$

Kendall's rank correlation coefficient (τ) measures the ordinal association between two variables, X and Y: $\tau = (n_c - n_d) / (n * (n - 1) / 2)$