Decision Systems Presentation

Hotel Booking Cancellation Prediction

MIT- No Code AI and Machine Learning

Scott Ungar
May 15, 2023

Contents / Agenda

Executive Summary

- Cancellations are most significantly influenced by Lead Time. The longer the lead time of guest bookings, the more likely the guest is to cancel their reservation.
  - Management should consider implementing a cancellation policy that addresses the cancelation of guest bookings based on the number of days to their reservation date. A sample model may look like:

Lead Time to Booking Date	*CancellationRefund*
> 99 days	100%
30 - 98 days	50%
Within 30 days	0%

- There is no significant correlation between Market Segment & Cancellations, so our cancellation policy would be indifferent to Market Type

- There is a strong correlation between guest cancellations and the number of Special Requests the hotel can accommodate. It is strongly advised that we implement communication to identify any Special Requests and do everything we can to accommodate those requests.

Business Problem Overview

Hotel bookings are very volatile to cancellations or no-shows. Reasons for cancellations include change of plans, scheduling conflicts, etc. The problem of cancellations is exacerbated because hotels offer their guests the option to cancel free of charge or at a low cost, which leads to a decline in hotel revenue. Revenue losses are exceptionally high on last-minute cancellations.

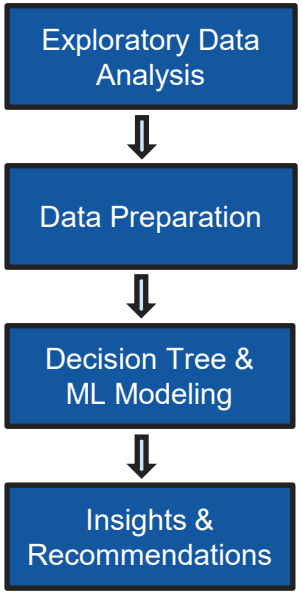
The cancellation of bookings impacts a hotel in many ways:

- Loss of revenue if the hotel cannot resell the room
- Additional costs in marketing & advertising to help sell these rooms
- Last-minute price reductions so the hotel can resell a canceled room, reducing the profit margin
- An increase in costs in human resources to accommodate guest cancellations and additional bookings

Business Solution Approach

The solution is to utilize the hotel's historical data in a machine-learning approach to predict which bookings are more likely than others to cancel. This analysis can then lead to management's change in policies and how they promote their offerings to the public.

The Approach



Analyze the data provided to find which factors have a significant influence on booking cancellations

Prepare the data for modeling utilizing dimensionality reduction techniques, handle missing values, and remove any outliers

Build a predictive model that can predict which type of booking will be canceled in advance

Formulate profitable policies for cancellations and refunds

Exploratory Data Analysis (EDA)

- **9,069 Records** *(Rows)*: Each record represents a hotel guest booking
- **19 Regular Attributes** *(Columns)*
- **Of the 9,069 bookings, there were 2,971 cancellations (32.7%)**
- **64% of all bookings were completed online, and canceled bookings & revenue lost from cancellations remained relatively consistent**

Market Segment	Total Bookings	Total Bookings %	Canceled Bookings	Canceled Booking %	Revenue Canceled	Revenue Not Canceled	Total Revenue	TOTAL Revenue %	% Revenue Canceled within Category	% of TOTAL Canceled Revenue
Aviation	26	0.29%	7	0.24%	$ 710.00	$ 891.00	$ 1,601.00	0.22%	44%	0.22%
Complementary	93	1.03%	-	0.00%	$ -	$ 19.00	$ 19.00	0.00%	0%	0.00%
Corporate	517	5.70%	56	1.88%	$ 4,847.33	$ 25,197.83	$ 30,045.16	4.09%	16%	1.49%
Offline	2,635	29.06%	805	27.10%	$ 80,925.27	$ 110,035.22	$ 190,960.49	25.97%	42%	24.80%
Online	5,798	63.93%	2,103	70.78%	$ 239,875.13	$ 272,761.03	$ 512,636.16	69.72%	47%	73.50%
TOTAL	**9,069**	**100.00%**	**2,971**	**100.00%**	**$ 326,357.73**	**$ 408,904.08**	**$ 735,261.81**	**100%**		**100.00%**

Link to Appendix slide on data background check

Model Performance Summary

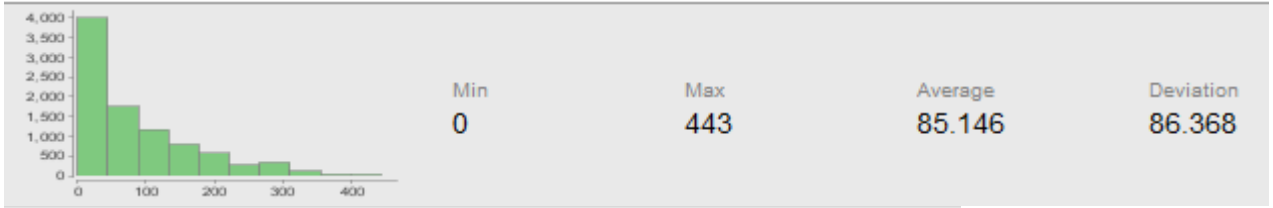
Weighted Attributes

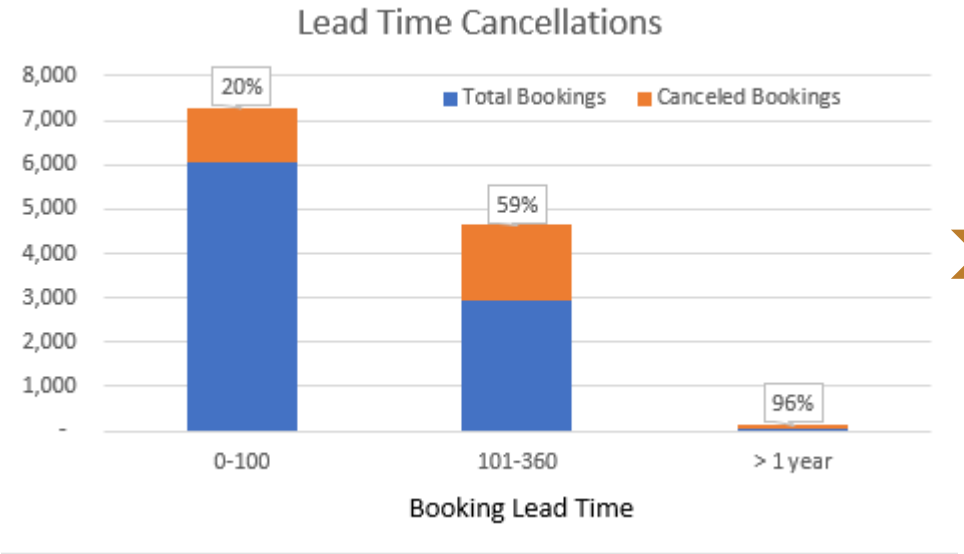
attribute	weight ↓
lead_time	0.237
arrival_date	0.137
no_of_weekend_nights	0.107
avg_price_per_room	0.086
no_of_week_nights	0.086
no_of_adults	0.083
arrival_month	0.075
type_of_meal_plan	0.072
room_type_reserved	0.044
market_segment_type	0.024
arrival_year	0.021
no_of_children	0.012
no_of_special_requests	0.010
required_car_parking_space	0.006
repeated_guest	0.000

- Lead Time has the greatest influence on guest cancellations.
- Arrival Date is far less of an influencer & is directly correlated to Lead Time

Exploratory Data Analysis (EDA)

- **The average lead time per booking is 85 days**





Guests who book with less lead time are far more likely not to cancel their reservation. Lead times over one year have a 96% cancellation rate.

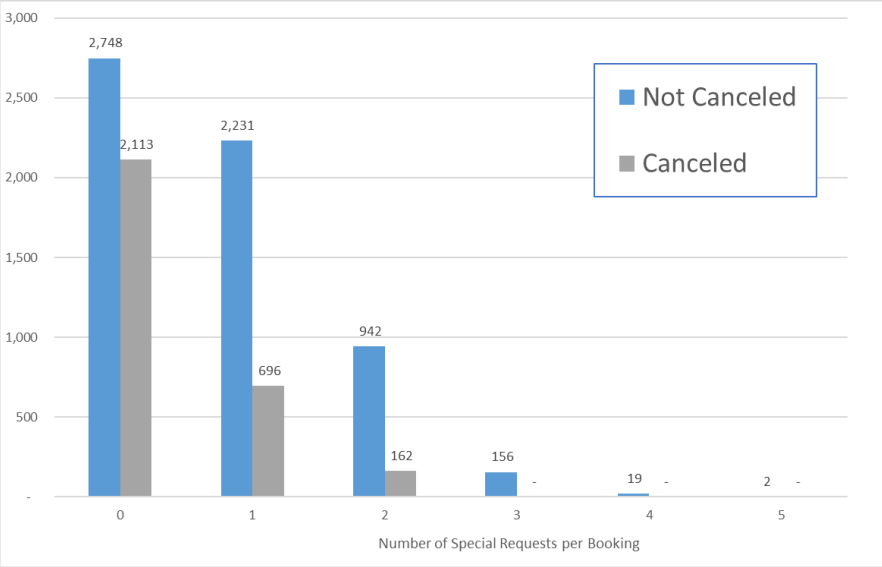
Exploratory Data Analysis (EDA)

Correlation Matrix

Attributes	booking_status ↓	no_of_special_requests	repeat...	requir...	no_of_...	mark...	no_of...	arrival...	arrival...	room_...	no_of_...	no_of...	type_...	no_of...	no_of...	arriv...	lead_...	avg_...
booking_status	1	0.254	0.111	0.090	0.058	0.055	0.037	0.023	-0.008	-0.016	-0.029	-0.058	-0.073	-0.090	-0.106	-0.191	-0.441	-0.131
no_of_special_requests	0.254	1	-0.022	0.084	0.019	0.206	0.002	0.103	0.016	0.139	0.116	0.079	-0.052	0.191	0.052	0.059	-0.099	0.183
repeated_guest	0.111	-0.022	1	0.115	0.509	0.283	0.397	0.009	-0.022	-0.022	-0.036	-0.060	-0.071	-0.193	-0.090	-0.032	-0.143	-0.162
required_car_parking_space	0.090	0.084	0.115	1	0.063	0.133	0.028	-0.015	-0.004	0.036	0.047	-0.043	-0.028	-0.011	-0.064	0.029	-0.072	0.053
no_of_previous_bookings_not_canceled	0.058	0.019	0.509	0.063	1	0.175	0.499	0.004	0.002	-0.000	-0.020	-0.015	-0.040	-0.118	-0.023	0.027	-0.077	-0.097
market_segment_type	0.055	0.206	0.283	0.133	0.175	1	0.067	-0.020	0.009	0.176	0.071	-0.006	-0.154	-0.101	-0.062	0.072	-0.315	-0.042
no_of_previous_cancellations	0.037	0.002	0.397	0.028	0.499	0.067	1	-0.044	-0.007	-0.007	-0.017	-0.015	-0.023	-0.045	-0.015	0.003	-0.052	-0.059
arrival_month	0.023	0.103	0.009	-0.015	0.004	-0.020	-0.044	1	-0.032	-0.008	0.001	-0.029	0.035	0.014	0.032	-0.355	0.128	0.054
arrival_date	-0.008	0.016	-0.022	-0.004	0.002	0.009	-0.007	-0.032	1	0.031	0.022	0.019	0.030	0.011	-0.003	0.015	0.001	0.016
room_type_reserved	-0.016	0.139	-0.022	0.036	-0.000	0.176	-0.007	-0.008	0.031	1	0.342	0.055	-0.161	0.248	0.086	0.094	-0.098	0.427
no_of_children	-0.029	0.116	-0.036	0.047	-0.020	0.071	-0.017	0.001	0.022	0.342	1	0.019	-0.066	-0.022	0.015	0.048	-0.044	0.330
no_of_weekend_nights	-0.058	0.079	-0.060	-0.043	-0.015	-0.006	-0.015	-0.029	0.019	0.055	0.019	1	-0.055	0.118	0.197	0.062	0.052	-0.005
type_of_meal_plan	-0.073	-0.052	-0.071	-0.028	-0.040	-0.154	-0.023	0.035	0.030	-0.161	-0.066	-0.055	1	0.027	-0.072	-0.085	0.119	0.046
no_of_adults	-0.090	0.191	-0.193	-0.011	-0.118	-0.101	-0.045	0.014	0.011	0.248	-0.022	0.118	0.027	1	0.099	0.070	0.104	0.290
no_of_week_nights	-0.106	0.052	-0.090	-0.064	-0.023	-0.062	-0.015	0.032	-0.003	0.086	0.015	0.197	-0.072	0.099	1	0.034	0.160	0.013
avg_price_per_room	-0.131	0.183	-0.162	0.053	-0.097	-0.042	-0.059	0.054	0.016	0.427	0.330	-0.005	0.046	0.290	0.013	0.180	-0.069	1
arrival_year	-0.191	0.059	-0.032	0.029	0.027	0.072	0.003	-0.355	0.015	0.094	0.048	0.062	-0.085	0.070	0.034	1	0.162	0.180
lead_time	-0.441	-0.099	-0.143	-0.072	-0.077	-0.315	-0.052	0.128	0.001	-0.098	-0.044	0.052	0.119	0.104	0.160	0.162	1	-0.069

A relatively high positive correlation exists between the number of Special Requests and whether or not the hotel guest cancels their booking.

Exploratory Data Analysis (EDA)



The more the hotel can accommodate a guest's special requests, the less likely the guest is to cancel.

Data Preprocessing

- We are starting with a clean data set and there are no missing values
- For all variations of modeling, the following were consistently applied:
  - Omitted Booking ID, the unique identifier that would have no bearing on the analysis
  - Converted all Nominal Attributes to Numerical Values using unique integers, except for Booking Status, the ***Target Variable***
  - The data was split into Training and Test subsets using partitions set at 80/20%, respectively
  - Both the Training and Test Performace criterion was set to Accuracy
- The variables changed to perform various test scenarios:
  - Decision Tree vs. Random Forest
  - With Pruning & without Pruning
  - Model criterion of Gini Index vs. Information Gain
  - Maximum Depth

Model Performance Summary

MODEL	DECISIOON TREE CRITERION	Max Depth	TRAIN ACCURACY (%)	TEST ACCURACY (%)	TRAIN WEIGHTED MEAN PRECISION (%)	TEST WEIGHTED MEAN PRECISION (%)	TRAIN WEIGHTED MEAN RECALL (%)	TEST WEIGHTED MEAN RECALL (%)
Decision Tree	Gini_Index	11	90.32%	86.55%	89.28%	85.08%	88.60%	84.00%
Decision Tree	Gini_Index	12	91.26%	87.32%	90.22%	85.79%	89.87%	85.22%
Decision Tree	Gini_Index	13	92.45%	85.94%	91.35%	84.02%	91.55%	84.11%
Decision Tree	Information_Gain	11	89.28%	87.21%	88.15%	85.94%	87.29%	84.62%
Decision Tree	Information_Gain	12	90.03%	86.77%	89.01%	85.27%	88.18%	84.38%
Decision Tree	Information_Gain	13	91.11%	86.77%	90.14%	85.15%	89.55%	84.59%
Decision Tree - Pruned	Gini_Index	11	89.69%	87.32%	88.27%	85.57%	88.35%	85.69%
Decision Tree - Pruned	Gini_Index	12	90.92%	87.10%	89.73%	85.24%	89.63%	85.66%
Decision Tree - Pruned	Gini_Index	13	92.07%	86.60%	90.92%	84.67%	91.14%	85.16%
Decision Tree - Pruned	Information_Gain	11	89.01%	86.77%	87.69%	85.00%	87.24%	84.94%
Decision Tree - Pruned	Information_Gain	12	90.20%	86.55%	89.05%	84.71%	88.59%	84.77%
Decision Tree - Pruned	Information_Gain	13	91.29%	86.55%	90.53%	84.80%	89.51%	84.56%
Random Forest	Gini_Index	11	89.28%	86.88%	89.47%	86.90%	85.81%	82.69%
Random Forest	Gini_Index	12	90.31%	87.16%	90.36%	86.89%	87.34%	83.32%
Random Forest	Gini_Index	13	91.34%	87.21%	91.36%	86.94%	88.75%	83.41%
Random Forest	Information_Gain	13	90.53%	87.21%	90.42%	86.85%	87.78%	83.49%
Random Forest	Information_Gain	14	91.48%	87.49%	91.32%	87.05%	89.10%	83.96%
Random Forest	Information_Gain	15	92.45%	87.76%	92.37%	87.29%	90.31%	84.38%
Random Forest - Pruned	Gini_Index	13	91.23%	88.53%	91.20%	88.19%	88.65%	85.30%
Random Forest - Pruned	Gini_Index	14	92.13%	88.53%	92.06%	88.19%	89.88%	85.30%
Random Forest - Pruned	Gini_Index	15	93.05%	88.59%	92.89%	88.16%	91.19%	85.47%
Random Forest - Pruned	Information_Gain	14	91.55%	88.37%	91.40%	87.97%	89.19%	85.13%
Random Forest - Pruned	Information_Gain	15	92.16%	88.64%	92.03%	88.16%	89.97%	85.60%
Random Forest - Pruned	Information_Gain	16	92.93%	88.42%	92.82%	87.89%	90.97%	85.35%

- The model that yielded the best results was Random Forest, Pruned, using the Information Gain criterion & a maximum depth of 15.
- The results of this model yielded the Highest Test Weighted Mean Precision + Highest Training & Test Set Accuracy without overfitting

APPENDIX

Data Background and Contents

Data Definition

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)*
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 *Values are ciphered by INN 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
avg_price_per_room	Average price per day of the reservation *Prices of rooms are dynamic (in euros)*
no_of_special_requests	Total number of special requests made by the customer *(e.g. high floor, view from the room, etc)*
booking_status	Flag indicating if the booking was canceled or not

THANK YOU