Hotel Reservation Cancellation

BA305 A1 Team 1

Sahir Doshi, Sophia Stearn, Andrew Trapp, Daniela Wong

Presentation Agenda

01

Introduction

02

Exploratory Findings

03

Pre-processing

04

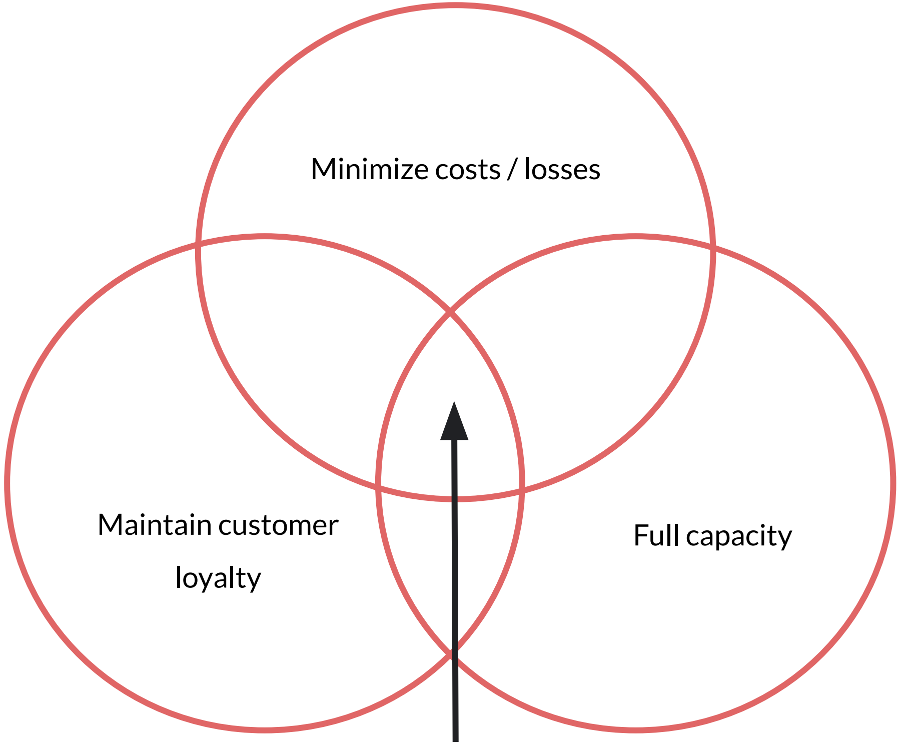
Modeling

05

Evaluation

01

Introduction



Solution: finding a model which predicts the likelihood of cancellation.

Why hotel cancellations?

The hotel and tourism industry typically accounts for about **10%** of worldwide GDP.

Our Dataset

36,275
rows

19
columns

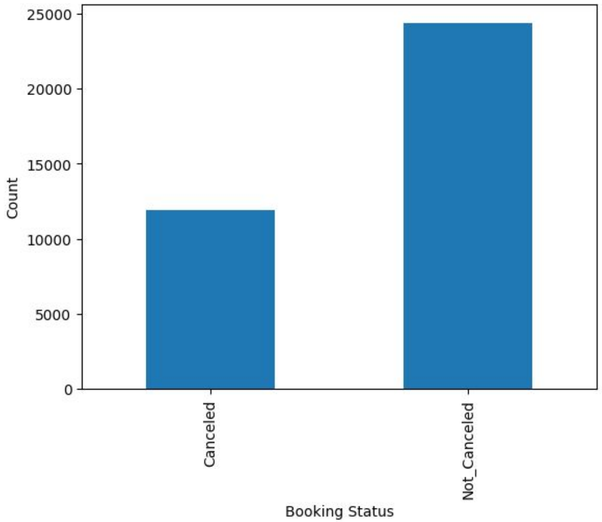
0
null-values

02

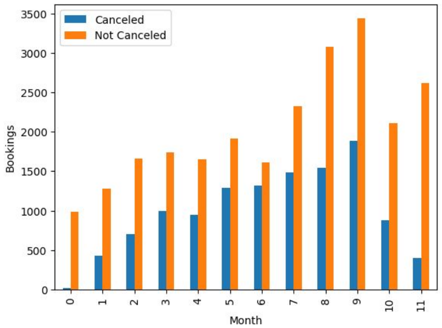
Exploratory Findings

Exploratory Findings

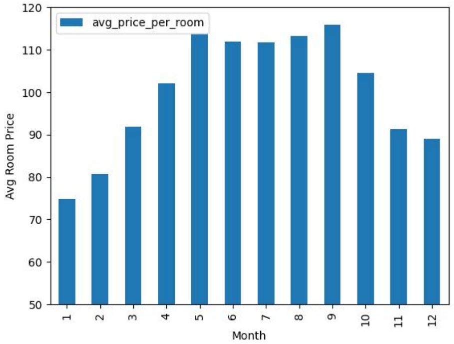
On average, **32.7%** of reservations get cancelled



October has most reservations, with **35.4%** getting cancelled

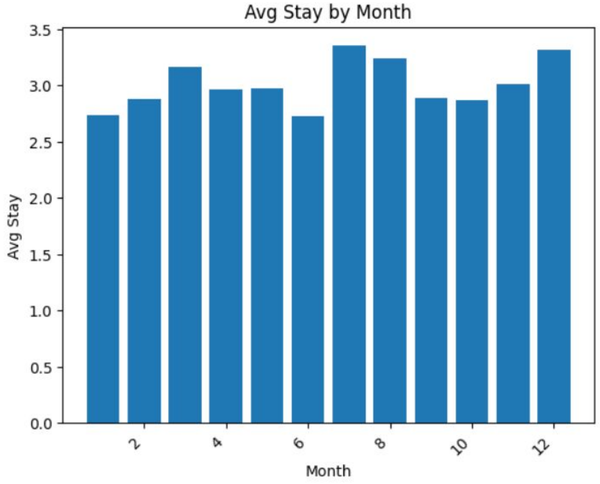


September is **most expensive** due to increased popularity

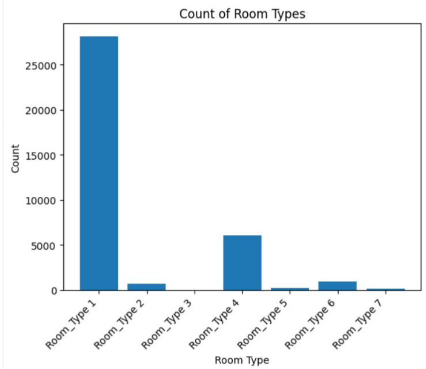


Exploratory Findings

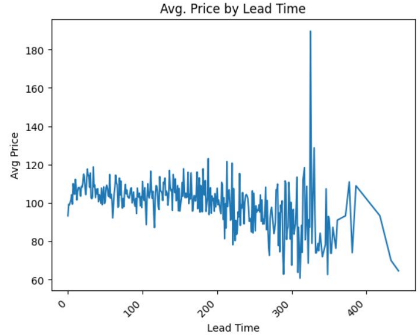
The Average Stay is Roughly
3 Days, all Year Long



95% of Rooms Booked are
of **Type 1 or 4**



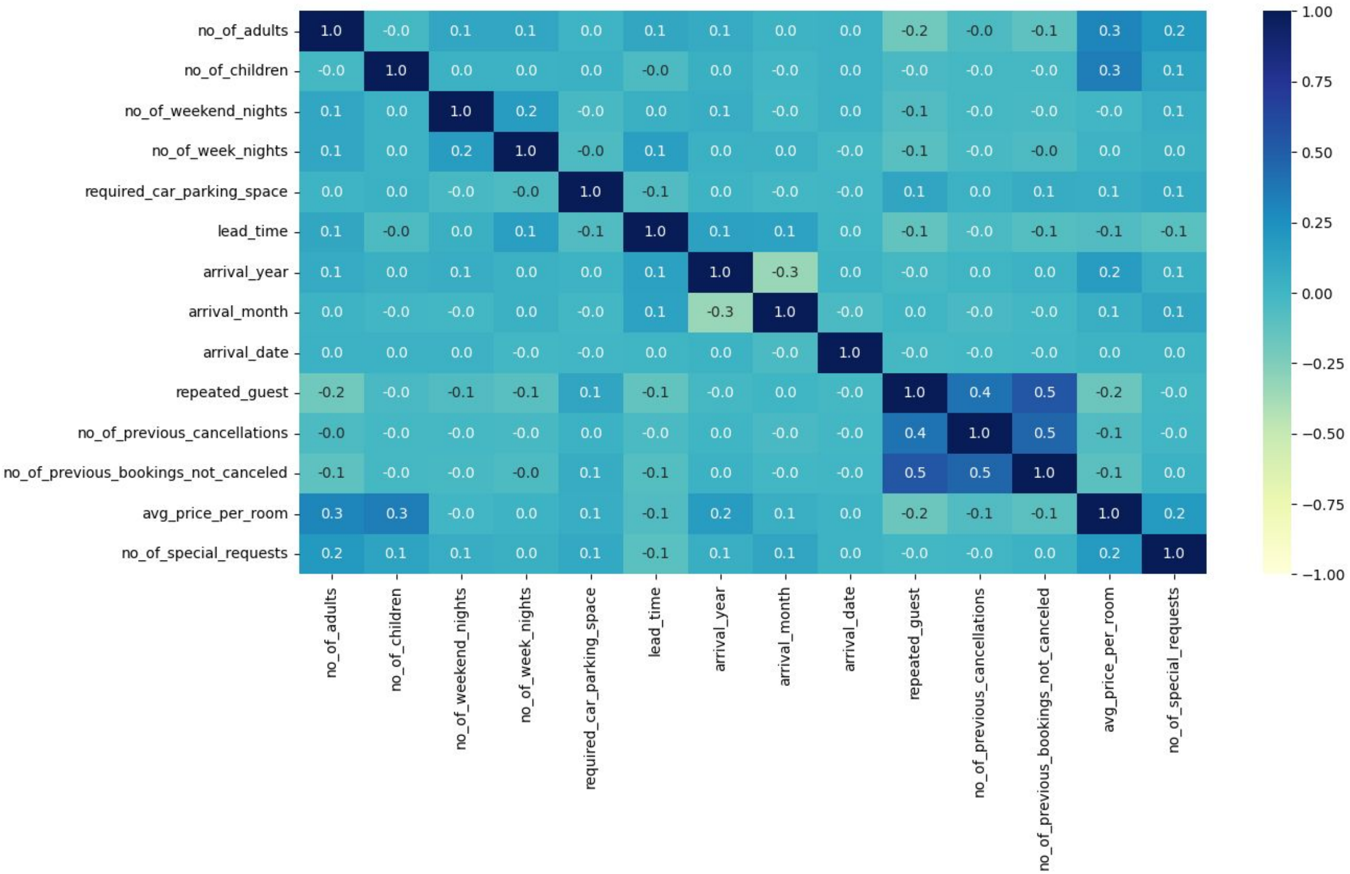
The Average Lead Time is **85 Days**



Lead Time does not Appear to Affect the Average Price of Booking

No correlation

> 0.5



03

Pre-processing

Pre-Processing

Renamed

Remove unuseful predictors

- Booking ID
- Arrival Year
- Arrival Date

Bucket predictors

- Booking Status
- Meal Plan
- Room Type

Check for outliers

- Lead time was too close

Normalize/standardize

- Sklearn.preprocessing scale()

Run PCA

- List item

PCA Breakdown

	0	1	2	3	4
num__no_of_adults	-0.43	0.01	0.26	-0.16	**-0.57**
num__no_of_children	-0.32	0.29	-0.14	**0.56**	0.42
num__no_of_weekend_nights	-0.18	-0.07	0.45	-0.27	**0.51**
num__no_of_week_nights	-0.19	-0.17	**0.52**	0.01	0.31
num__lead_time	-0.03	-0.41	0.40	**0.46**	-0.28
num__no_of_previous_cancellations	0.30	**0.47**	0.38	0.16	-0.16
num__no_of_previous_bookings_not_canceled	0.35	**0.49**	0.33	0.10	-0.05
num__avg_price_per_room	**-0.51**	0.25	-0.10	0.31	-0.16
num__no_of_special_requests	-0.34	**0.36**	0.07	-0.42	-0.06

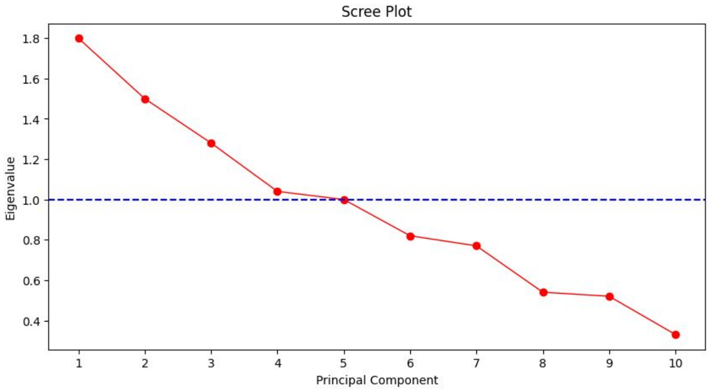
	% of variance explained	Cumulative % explained
0	0.166	0.166
1	0.138	0.304
2	0.118	0.422
3	0.096	0.518
4	0.092	0.610
5	0.076	0.685
6	0.070	0.756
7	0.050	0.805
8	0.048	0.854
9	0.030	0.884

PCA Breakdown

5 components account for **61%**

No correlation

Used comps. Above 1 eig According to the Latent Root Criterion

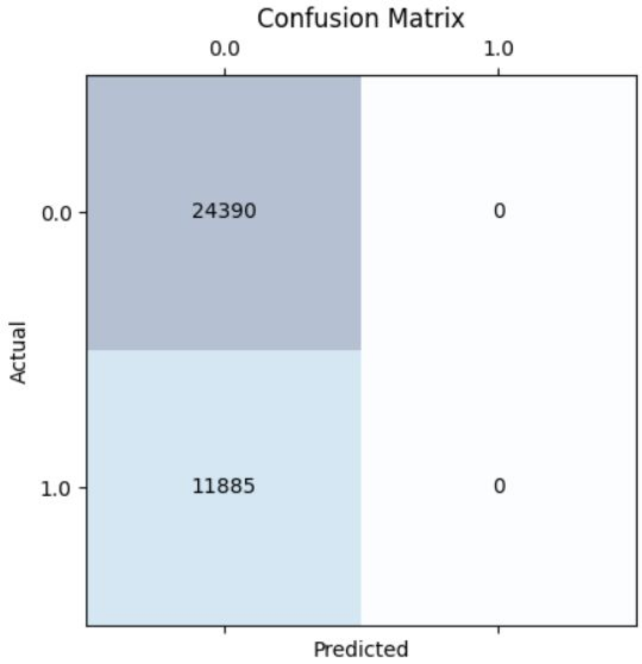


	0	1	2	3	4
0	1.0	-0.0	0.0	-0.0	0.0
1	-0.0	1.0	0.0	0.0	0.0
2	0.0	0.0	1.0	0.0	0.0
3	0.0	0.0	0.0	1.0	0.0
4	0.0	0.0	0.0	0.0	1.0

	% of variance explained	Cumulative % explained
0	0.166	0.166
1	0.138	0.304
2	0.118	0.422
3	0.096	0.518
4	0.092	0.610
5	0.076	0.685
6	0.070	0.756
7	0.050	0.805
8	0.048	0.854
9	0.030	0.884

67% Baseline Accuracy if always predict not cancelled

F1 score:	0%
Precision:	0%
Recall:	0%
Split data:	67/33



04

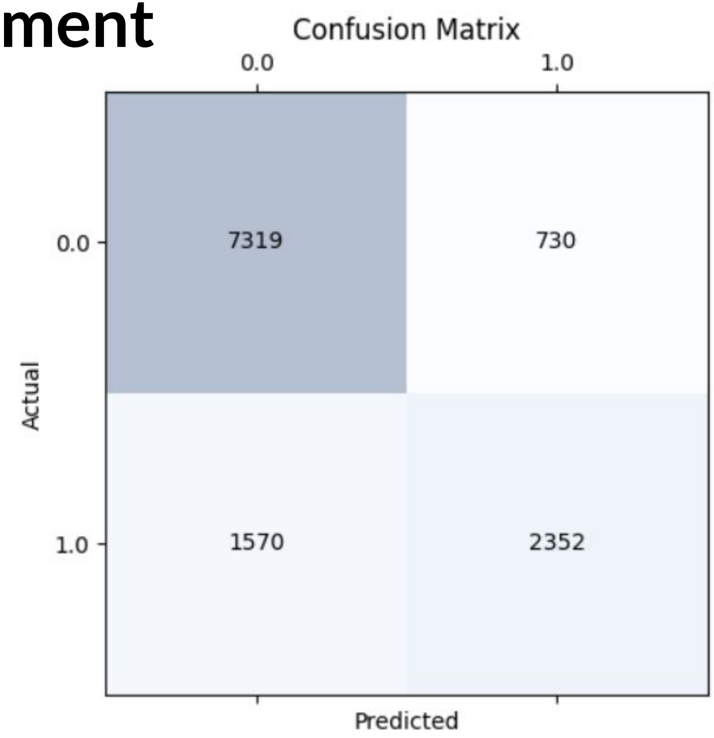
Modeling

Model Analysis

	Accuracy	Improvement from Baseline
Logistic Regression	80.5%	20.15%
Logistic with Optimal Threshold	80.8%	**20.60%**
Logistic Regression Elastic Net	80.4%	20.0%
KNN Classifier	85.5%	27.61%
KNN Classifier PCS	88.0%	**31.34%**
Decision Tree w/o Pruning	86.4%	28.96%
Decision Tree w/ Optimal Penalty	85.0%	26.87%
Decision Tree w/ Grid Search & Random Search	86.1%	28.51%
Decision Tree w/ Bagging	89.2%	**33.13%**
Random Forests w/ Grid Search	89.1%	**32.99%**
Neural Networks	86.0%	**28.36%**
Neural Networks w/ PCS	84.0%	25.37%

Logistic Regression with optimal threshold: **20.6%** improvement

Accuracy:	**80.8%**
Best Threshold:	**54%**
Precision:	**73%**
Recall:	**56%**



Logistic Regression with optimal threshold: **20.6%** improvement

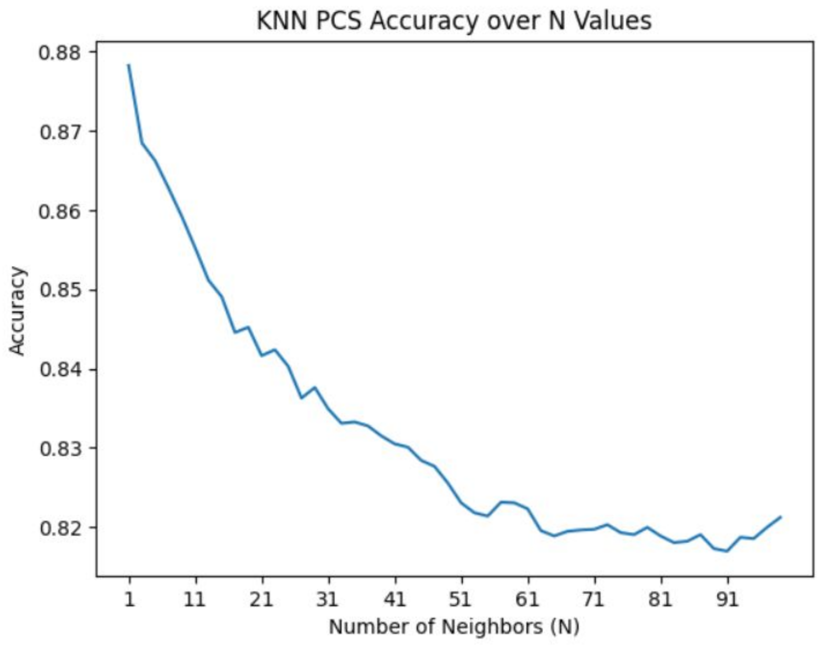
Arrival Month (March):	**0.967**
Segment Type (Aviation)	**1.022**
Segment Type (Online):	**1.044**
Numerical Lead Time:	**1.423**
Intercept:	**-2.057**

Feature Importance:

	Feature	Coefficient	Abs_Coefficient
0	Intercept	-2.057306	2.057306
1	cat__arrival_month_1	-1.784433	1.784433
2	remainder__repeated_guest	-1.712577	1.712577
3	remainder__required_car_parking_space	-1.597155	1.597155
4	cat__market_segment_type_Complementary	-1.556831	1.556831
5	num__lead_time	1.423397	1.423397
6	cat__arrival_month_12	-1.357797	1.357797
7	num__no_of_special_requests	-1.201861	1.201861
8	cat__market_segment_type_Online	1.044196	1.044196
9	cat__market_segment_type_Aviation	1.022324	1.022324
10	cat__arrival_month_2	0.967786	0.967786
11	cat__market_segment_type_Offline	-0.767327	0.767327
12	remainder__frequent_room	0.753398	0.753398
13	num__avg_price_per_room	0.640742	0.640742
14	cat__arrival_month_11	0.621021	0.621021
15	cat__arrival_month_3	0.534550	0.534550
16	cat__arrival_month_4	0.371541	0.371541
17	remainder__meal_plan_selected	-0.369301	0.369301
18	cat__arrival_month_6	0.321989	0.321989
19	cat__market_segment_type_Corporate	0.232196	0.232196
20	num__no_of_previous_bookings_not_canceled	-0.222719	0.222719
21	cat__arrival_month_10	0.219303	0.219303
22	num__no_of_weekend_nights	0.135385	0.135385
23	num__no_of_previous_cancellations	0.108541	0.108541
24	cat__arrival_month_7	0.092832	0.092832
25	num__no_of_children	0.076607	0.076607
26	cat__arrival_month_9	-0.069856	0.069856
27	num__no_of_week_nights	0.067284	0.067284
28	cat__arrival_month_8	0.059583	0.059583
29	num__no_of_adults	0.026528	0.026528
30	cat__arrival_month_5	-0.001962	0.001962

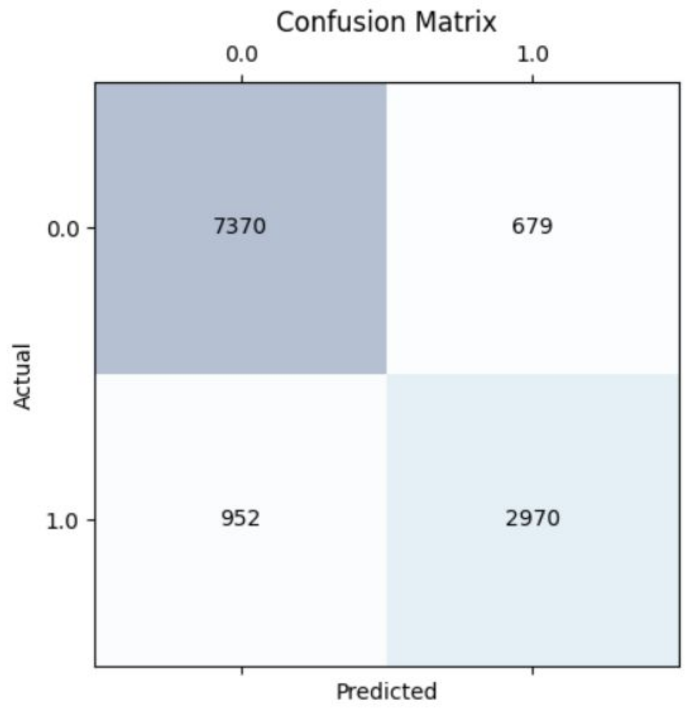
KNN Classifier PCS: **31.34% improvement**

Accuracy:	**88%**
Best Neighbors:	**1**
Precision:	**82%**
Recall:	**76%**



KNN Classifier PCS: **31.34%** improvement

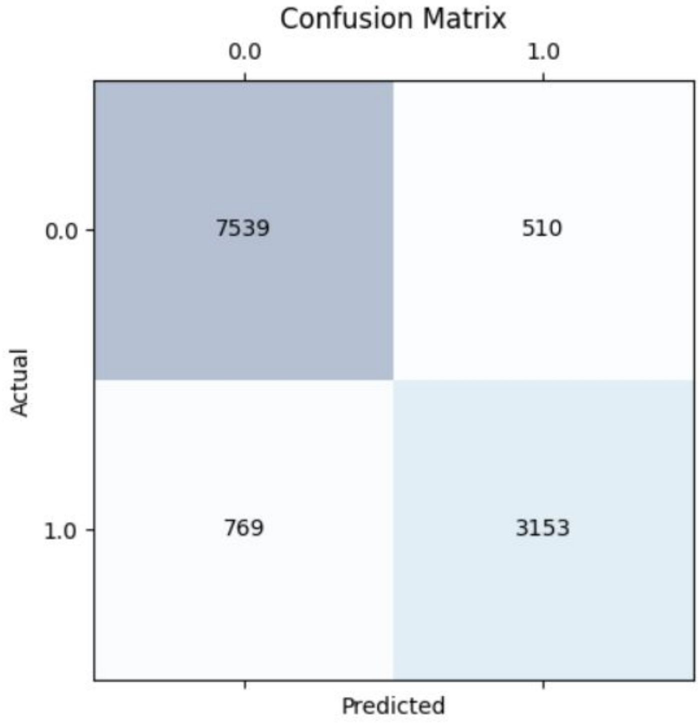
Accuracy:	**86%**
Chosen Neighbors:	**11**
Precision:	**81%**
Recall:	**76%**



Decision Tree with bagging: 33.13% improvement

Accuracy:	**89.0%**
F1- Score:	**83%**
Precision:	**86%**
Recall:	**80%**

max_depth:	7
learning rate:	1
n_estimators:	1735
Algorithm:	'SAMME'



Random Forest with Grid Search: **32.99%** improvement

Parameters

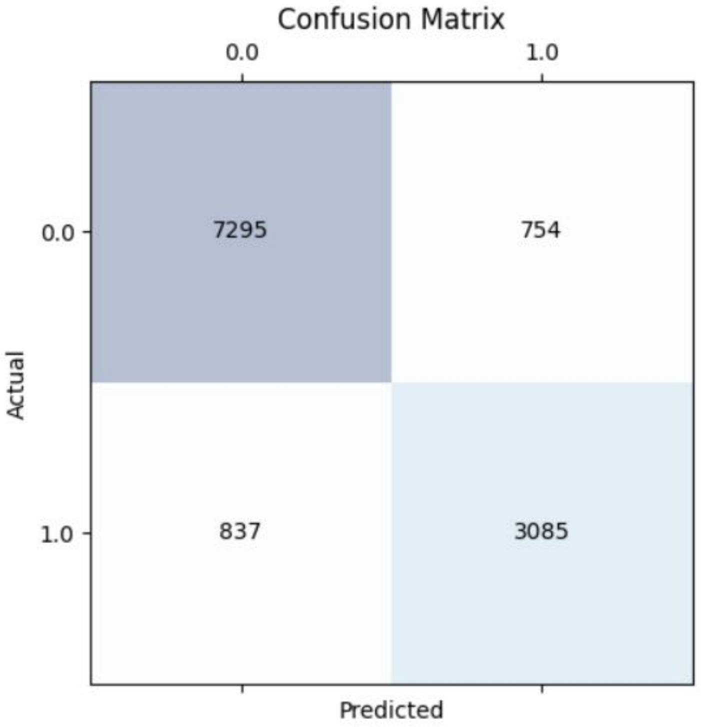
N_estimators:	1400
criterion:	“gini”

Feature	Importance
Lead Time	37%
Avg Price per Room	20%
Num of Special Requests	11%
Number of Week Nights	6%
Number of Weekend Nights	5%

Neural Network: 28.36% improvement

Accuracy:	**86.0%**
F1- Score:	**78%**
Precision:	**83%**
Recall:	**74%**

Activation:	'ReLU'
Alpha:	0.1
Hidden_layer_sizes:	(50, 50)
learning_rate:	'adaptive'
max_iter:	2000
solver:	'lbfgs'



05

Evaluation

Tuned Optimal Threshold Decision-Tree with Costs Assigned yielded the best precision score and lowered 'costs'

Cost of Misclassification:

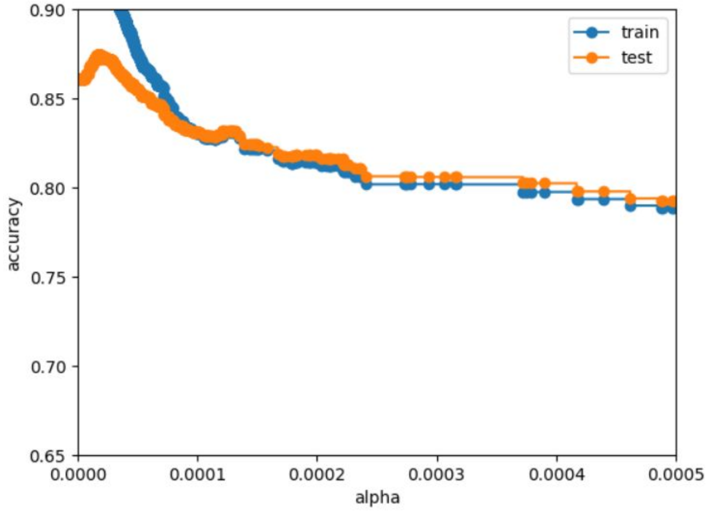
- Classify as not canceled when guest actually does cancel - False Negative
- Classify as canceled when guest actually shows - False Positive

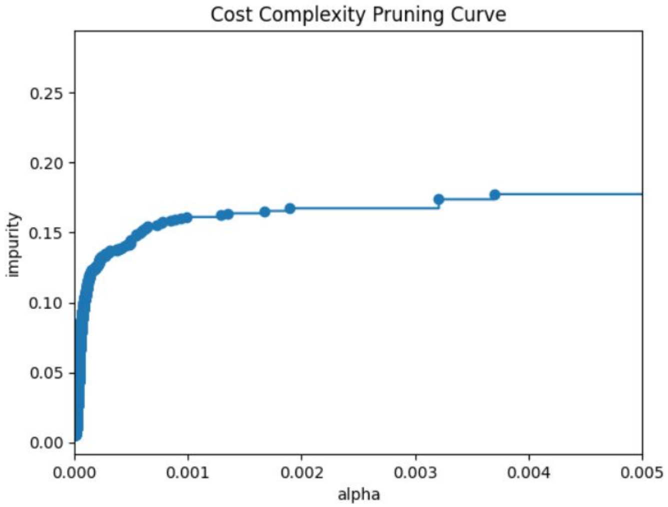
Goal is to minimize cost by:

- **Emphasizing precision**, as FP is more costly, while **keeping recall in mind**
- **Higher accuracy than baseline**

Assumptions courtesy of Courtyard by Marriott Boston Brookline

Tuned Optimal Threshold Decision-Tree with Costs Assigned yielded the best precision score and lowered 'costs'



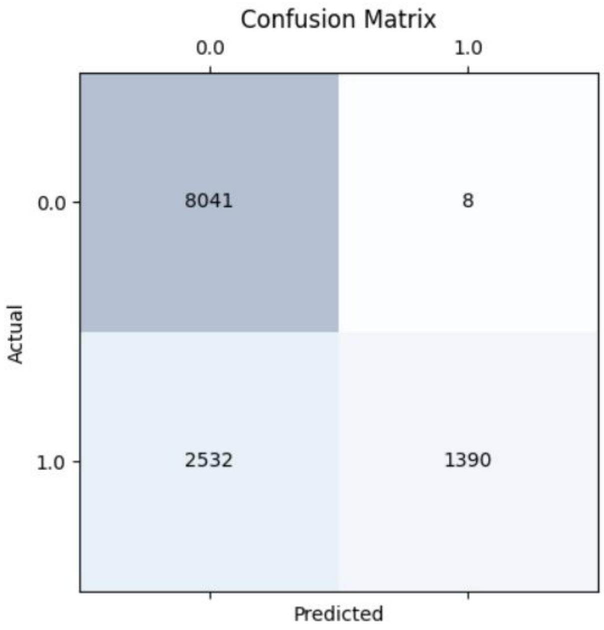


Optimal Alpha: **.001**

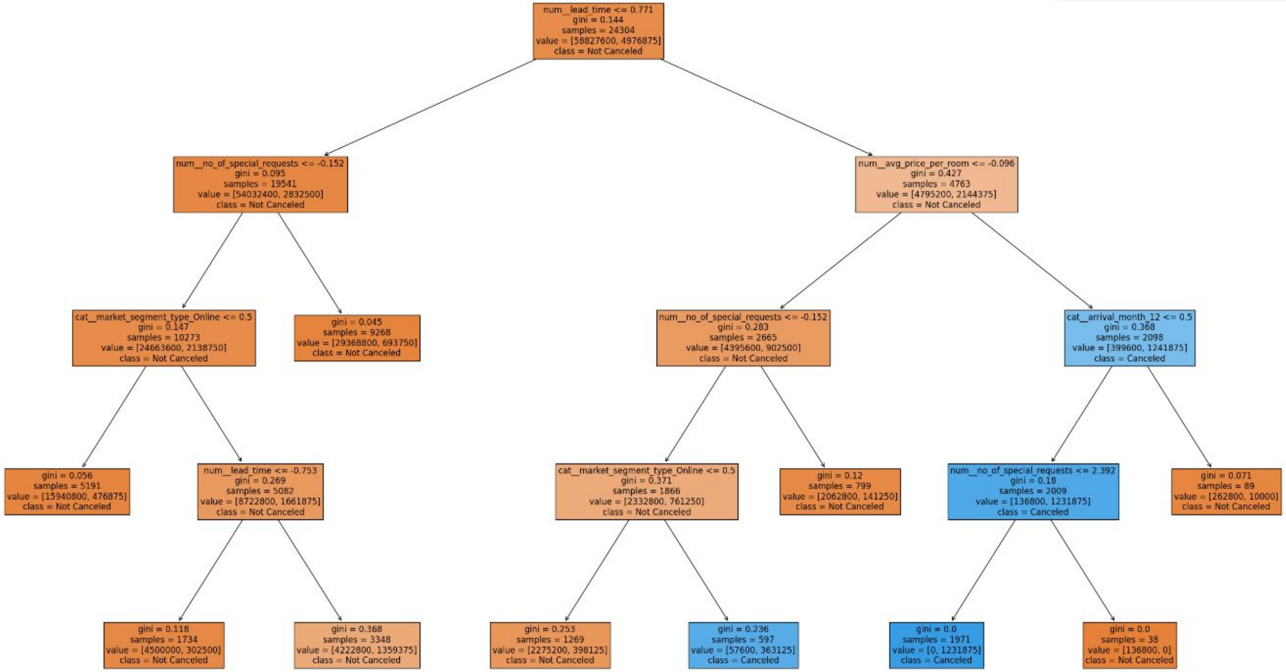
Tuned Optimal Threshold Decision-Tree with Costs Assigned yielded the best precision score and lowered 'costs'

Feature	Importance
Lead Time	31%
Avg Price per Room	27%
Market Segment Type Online	18%
Number of Special Requests	13%
Arrived in December	10%
Others	1%

Accuracy:	**78%**
F1-Score:	**52%**
Precision:	**99%**
Recall:	**35%**
Optimal Threshold:	**25%**



Tuned Optimal Threshold Decision-Tree with Costs Assigned yielded the best precision score and lowered 'costs'



Model Evaluation

Result	Evidence
Increases Accuracy of Prediction	Accuracy increases by 11% over naïve model
Accounts for costs by directing most mistakes to the "cheaper" option which provides time to adjust	Results in total annual loss of €36,000 relating to Type I and II errors, nearly the lowest, while…
Interpretable results compared to higher accuracy models	Decision tree can be interpreted by staff up to management
Low Cost Complexity and High Flexibility	Model can be run quickly and parameters can be tuned fairly simply

Challenges

Challenges	Future Improvements
Don’t know when cancellations occur	Dataset could include the type of cancellation
Don’t know the location of the hotel	Dataset could include type of hotel or specific region
Computational costs when doing parameter searches	Get a better computer or supercomputer lol