Hotel Cancellation Problem and Overbooking Tactics Analysis

Contents

Business Overview

Hotel Background Information

Provide services for customers from **20+** countries

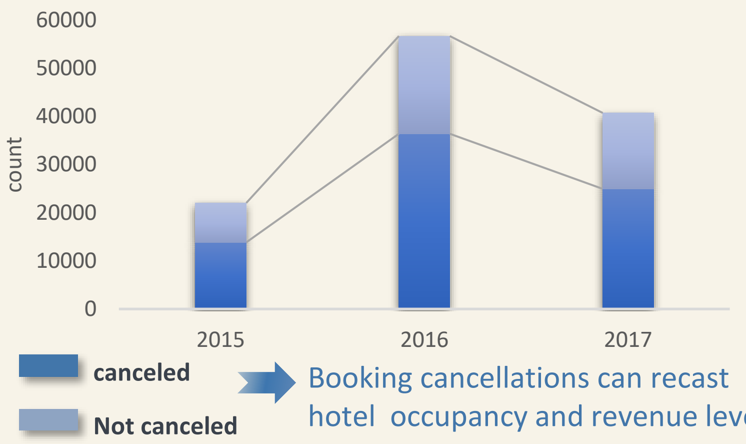
Average annual booking: **40k**
Average total customers: **100k+**

42% customer book reservations **3 months** ahead

Provide Resort&City Hotel Reservations

July to November is the **peak** of hotel passenger flow

High Hotel Cancelation Rate



Booking cancellations can recast hotel occupancy and revenue levels.

Cancelation rate over the past three years

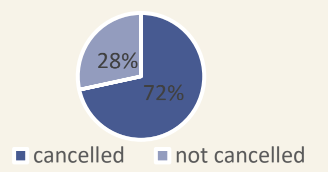
2015: 63.0% **2016:** 64.1%. **2017:** 61.3%

2016 Hotel industry cancellation rate

- The average percentage of canceled reservations, is currently **24%.**

No Deposit Booking Policy

- No Deposit
- No refund
- Refundable



- Among all the cancelation cases, **88%** customer are in "No deposit policy" type

"The revenue that hotels received for **cancellations and no-shows** increased by almost **12 percent** annually on average from 2012 to 2016 "

-- *Lodging Magazine*

(e.g., using deposits policy and overbooking strategy)

Overbooking as Industry Tactics

Overbooking in hotel management is a confidence strategy which accept more reservations than rooms you have available and anticipate some will cancel

No-show

Last-minute cancellation

Advantages of overbooking in hotels

- Minimizes losses by creating a **backup plan for cancelled reservations**
- **Achieving full occupancy** as no financial potential is wasted

Potential damage of overbooking in hotels

- Negative guest experiences when it comes to **aggressive overbooking**
- Bad reviews from customers will harm the hotel **reputation and have long-term negative effect**

1 What is the most suitable overbooking rate to reduce vacancy?

2 Can overbooking perfectly solve the cancelation problem?

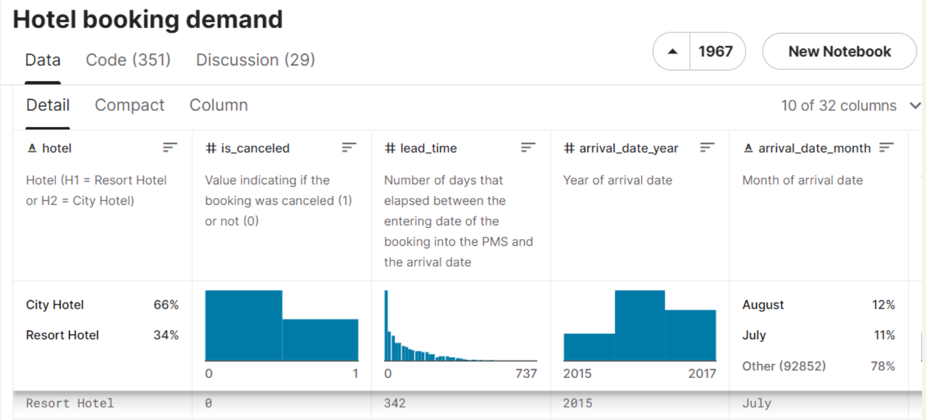
3 How can any other options tackle high cancelation problem?

Optimizing Overbooking Rate using Random Forest

- Exploring the data and select significant features.
- Build up different models and decide the most suitable one.
- Improve the random forest model and estimate the overbooking rate.

Data source and data cleaning

1. Data source



"Hotel Booking Demand" from *Kaggle*

- **119,390** data points
- **32** attributes

2. Data cleaning

- Filled up missing value
- Remove abnormal data
- Remove outlier
- Remove duplicate data

3. Split data

¾ training	¼ test

Exploratory Data Analysis

Remove features

Reason 1:
Features must can be obtained at reservation.

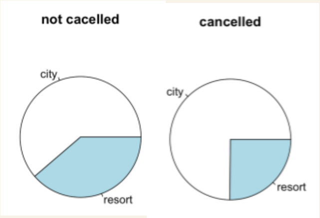
- booking_changes
- reservation_status
- assigned_room_type

Reason 2:
Data masking.

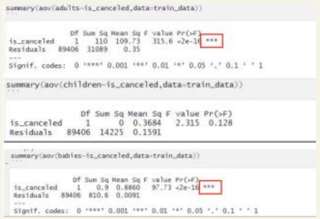
- country

Select features

Visualization

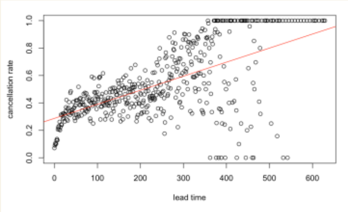


ANOVA test



```
summary(aov(adults~is_canceled,data=train_data))

              Df Sum Sq Mean Sq F value Pr(>F)
is_canceled    1    110  109.73   315.6 <2e-16 ***
Residuals  89406  31089    0.35
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

summary(aov(children~is_canceled,data=train_data))

              Df Sum Sq Mean Sq F value Pr(>F)
is_canceled    1      0  0.3684   2.315  0.128
Residuals  89406  14225  0.1591

summary(aov(babies~is_canceled,data=train_data))

              Df Sum Sq Mean Sq F value Pr(>F)
is_canceled    1    0.9  0.8860   97.73 <2e-16 ***
Residuals  89406  810.6  0.0091
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Linear regression



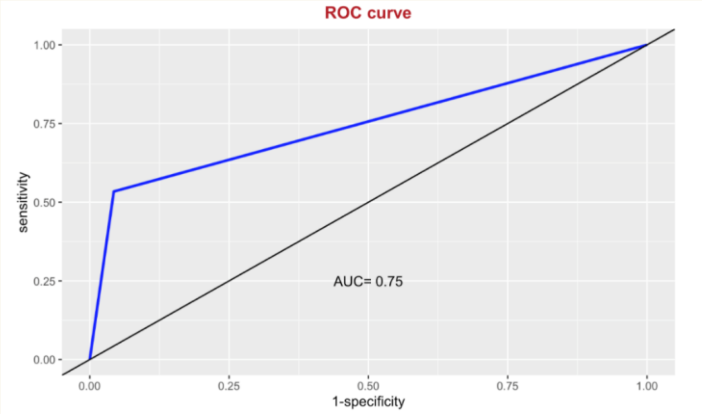
14 features selected

- whether_equal_room_type
- adults
- babies
- hotel
- meal
- is_repeated_guest
- lead_time
- previous_cancellations
- customer_type
- required_car_parking_spaces
- arrival_date_month
- distribution_channel
- agent
- deposit_type

Build Random Forest Model and Get Importance Table

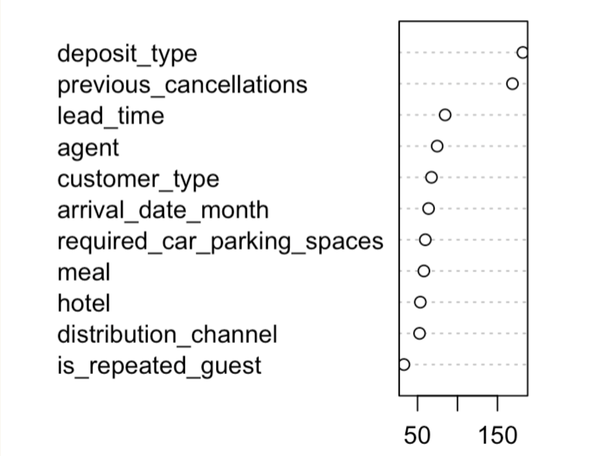
Random Forest

```
rf.fit <- randomForest(is_canceled~whether_equal_room_type+adults+babies+hotel+meal+is_repeated_guest+lead_time+previous_cancellations+customer_type+required_car_parking_spaces+arrival_date_month+distribution_channel+agent+deposit_type,data=train_data,importance=TRUE)
```



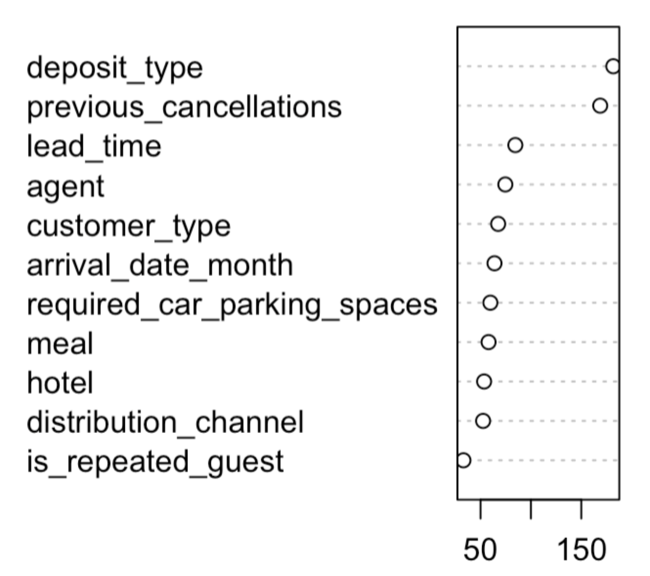
Error rate=19.45%

Importance table



Mean decrease of accuracy

Select Random Forest as Our Model



	number_of_feature	error.lda	error.qda	error.log
[1,]	1	0.2646361	0.2646361	0.2646361
[2,]	2	0.2651254	0.2604009	0.2654771
[3,]	3	0.2687033	0.2602786	0.2682139
[4,]	4	0.2654618	0.2628319	0.2646973
[5,]	5	0.2618075	0.2614558	0.2612417
[6,]	6	0.2600491	0.2780297	0.2591317
[7,]	7	0.2589177	0.2779535	0.2591470
[8,]	8	0.2589789	0.3076312	0.2583980

Cross-validation error of LDA, QDA, and logistic regression from 1 to 8 features

Random forest error rate=19.45%

Evaluate and revise the random forest model

Real

Predict		Positive	Negative
	Positive	TP 24,022	FP 7,892
	Negative	FN 9,126	TN 48,370

FP → Actual shows up

FN → Actual no shows

The model should:

- predict the cancelation rate with high accuracy
- avoid having actual shows up more than actual no shows

Threshold – a crucial parameter

Loss of failing to provide rooms due to aggressive overbooking

- Negative customer experience
- Negative word of mouth
- Loss of customer loyalty
- Costs of compensation

Profits generated from overbooking

- Revenue from utilizing cancelled rooms

Security rate = 1 - (actual shows up / actual no shows)

The security rate should:

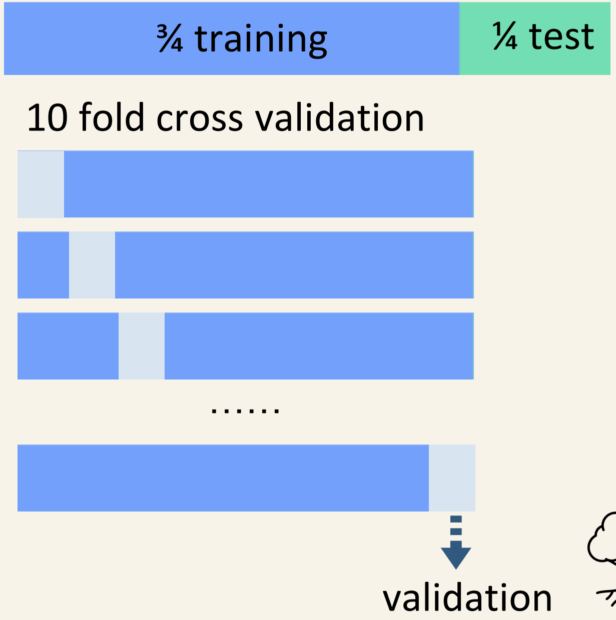
- Larger than 0
- As small as possible

We set the security rate at 10%

When 9 customers are actually staying, 10 customers will be actually leaving.

Adjust threshold and generate the final model

Set the threshold from 0.05 to 0.5 to build up several new models.



Threshold	Actual shows up	Actual no shows	Security rate	Total error rate
0.05	34.8%	9.9%	-496.8%	25.53%
0.1	25.1%	15.1%	-181.0%	21.37%
0.2	17.0%	23.5%	-22.6%	19.41%
0.25	14.8%	26.4%	4.6%	19.11%
0.27	**14.0%**	**27.5%**	**13.5%**	**19.03%**
0.3	23.3%	28.7%	21.4%	19.03%
0.4	9.7%	35.1%	52.8%	19.13%
0.5	7.8%	39.3%	66.4%	19.45%

Final model: Random Forest with threshold equals to 0.27

Generate overbooking rate based on test set

¾ training | ¼ test

Real

Predict

	Positive	Negative
Positive	TP 5,827	FP 1,089
Negative	FN 5,224	TN 17,660

Accuracy rate = 78.87%

- Cancellation rate = 23.2%
- Overbooking rate =

$$Cancellation\ rate \times \frac{Current\ booking}{Total\ capacity}$$

- Overbooking rate = 23.2%

Increase revenue $1,237,272 per year

Optimal Overbooking Rate Model Testing

☑ **What the model can help with**

- Accurately estimate the cancellation rate
- Set a red line for overbooking

☒ **What the model can't help with**

- Stably offset large percentage of the revenue lost caused by no-shows

A need to tackle the problem from the root cause

- Generally accurate prediction on cancellation rate

21.18%
Total error rate

- Minimized overbooking risk

5.81%
False positive rate

Increasing but not satisfying utilization vacancy

➢May due to the data's nature of high variance

47.27%
False negative rate

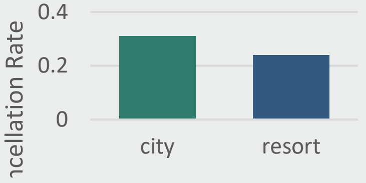
20.85%
Vacancy utilization rate

Cancellation Improvement with Various Approaches

- **<u>EDA & Logistic Regression</u>**: Discover the extent of different factors’ effect on customer‘s decision of reservation cancelling. → Focus on these features to provide a better service
- **<u>Clustering</u>:** Discover features that highly affect customer‘s decision of reservation canceling → Implement “Stratified Deposit Plan”

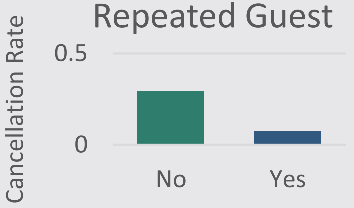
Reviewing EDA and Logistic Regression

Factor effect through the lens of *EDA*

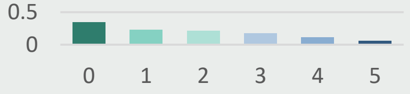


- City hotel faces more severe cancellation problem

- Lowest cancellation rates for both hotels appear in August



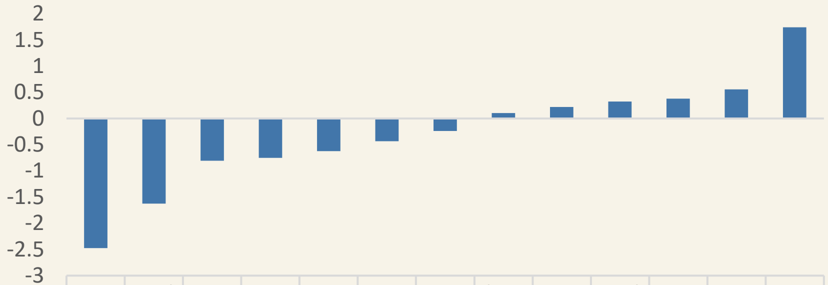
- Higher special request number seems to improve stickiness and less cancellation



- First-time customers have higher tendency to cancel their booking

Factor effect through *logistic regression*

$$P(Cancellation) = \frac{e^{\beta_0+\beta_1 LeadTime+\beta_2\ PreviousCancellation+\ \ldots\ldots}}{1+e^{\beta_0+\beta_1 LeadTime+\beta_2\ PreviousCancellation+\ \ldots\ldots}}$$



required_car_parking_spaces	market_segment_Offline TA/TO	is_repeated_guest	distribution_channel_Direct	total_of_special_requests	customer_type_Group	previous_bookings_not_canceled	distribution_channel_Corporate	average_daily_rate	market_segment_Complementary	lead_time	customer_type_Transient	previous_cancellations
-2.48	-1.62	-0.81	-0.75	-0.63	-0.44	-0.24	0.1	0.219	0.323	0.376	0.555	1.744

Possible strategies inspired by EDA observation and LR model

Initiative 1

Observation

- Longer leading time—higher cancellation

Possible measures

- Deliver ***directed push*** to customers after booking
- Content of push: reminder message, ads of featured service, festival greetings

Initiative 2

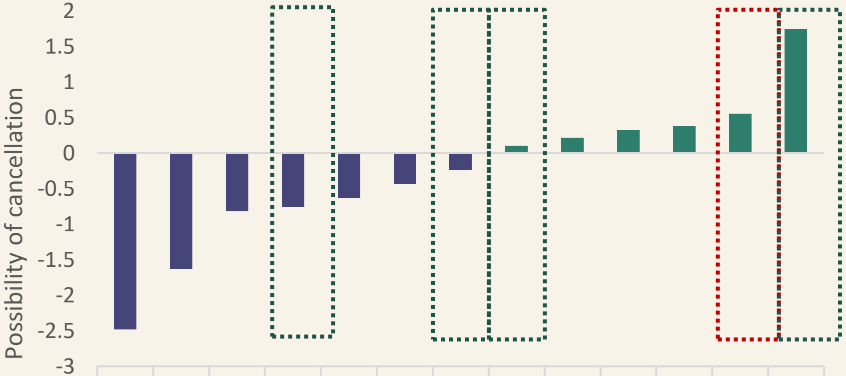
Observation

- Cancellation possibility varies with Distribution channels and customer types

Possible measures

- Reshape **customer target** and **distribution strategy**
- For example, reserve more rooms to direct distribution channel and direct ads with more focus on Group customer.

Coefficients of the Logistic Regression



required_car_parking_spaces	market_segment_Offline TA/TO	is_repeated_guest	distribution_channel_Direct	total_of_special_requests	customer_type_Group	previous_bookings_not_canceled	distribution_channel_Corporate	average_daily_rate	market_segment_Complementary	lead_time	customer_type_Transient	previous_cancellations
-2.5	-1.6	-0.8	-0.8	-0.6	-0.4	-0.2	0.1	0.22	0.32	0.38	0.56	1.74

More initiatives can be formulated with our data-driven observation & Complete strategy can be formed based on the board’s interest…

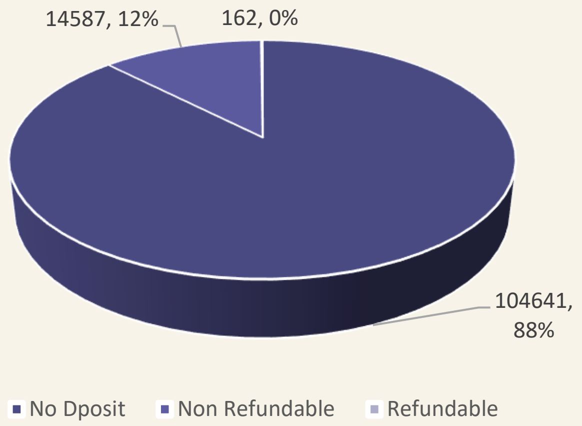
Potential Problem & Potential Strategy

! "Problematic Deposit Plan"

Majority are "No Deposit"

Very likely to cancel due to

No Extra Expense



✓ "Stratified Deposit Plan"

Industry Evidence:

Deposit can lower the cancellation rate.

Strategy:

Different deposit for different clusters of customers

Sector	Predicted cancel rate	Mean/centroid of each variable	Number of datapoint
1	......	......	......
2	......	......	......
3	......	......	......

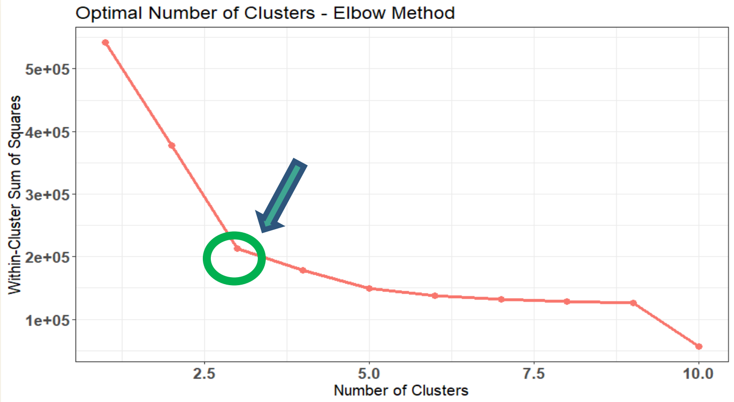
Clustering with Customer Behavior Data

Chosen Variables	Selection Reason
Lead time Previous_booking not cancelled	Most influential factors in Logistic Regression & Random Forest
Price	Monetary in RFM Rare Numerical Data type in Dataset

Number of Sectors

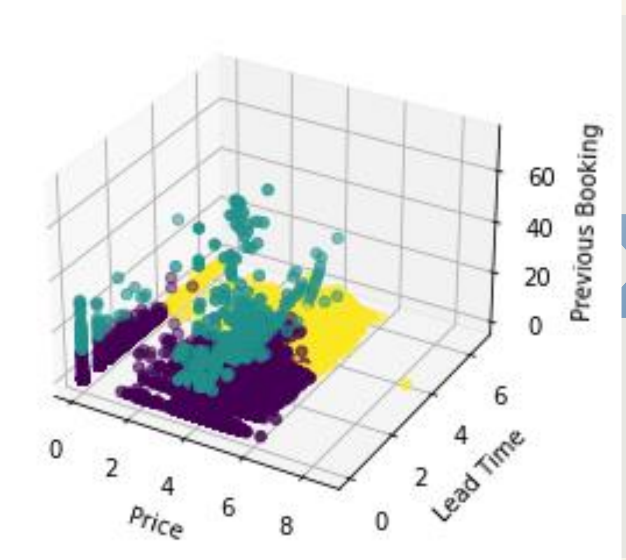
Elbow Method

K=3 is the "Elbow"



Clustering with Customer Behavior Data

3D Clustering Graph



Clustering Data Result

Sector	Cancel rate	Number of customers
1	0.128948275	26119
2	0.059561129	319
3	0.353251046	58089

Sector	lead_time	previous_bookings_not_canceled	price
1	1.455220592	0.292239366	4.27062846
2	1.583392646	22.76489028	3.55557916
3	4.451020014	0.020021002	4.60584244

Clustering with Customer Behavior Data

Clustering Data Result

Sector	Number of customers	Cancel rate	Cancel level
1	26119	0.129	Middle
2	319	0.060	Low
3	58089	0.353	High

sector	lead_time	previous_bookings_not_canceled	price
1	Low	Middle	Middle
2	Middle	High	Low
3	High	Low	High

Stratified Deposit Plan

Diamond Customer

- Predicted Cancellation Rate = 6.0%
- Largest mean num of Fulfilling Appointment
- Middle mean Lead Time, Low mean Price

50% Room Fee as Deposit

Golden Customer

- Predicted Cancellation Rate = 12.9%
- Middle mean num of Fulfilling Appointment
- Low mean Lead Time, Middle Price

70% Room Fee as Deposit

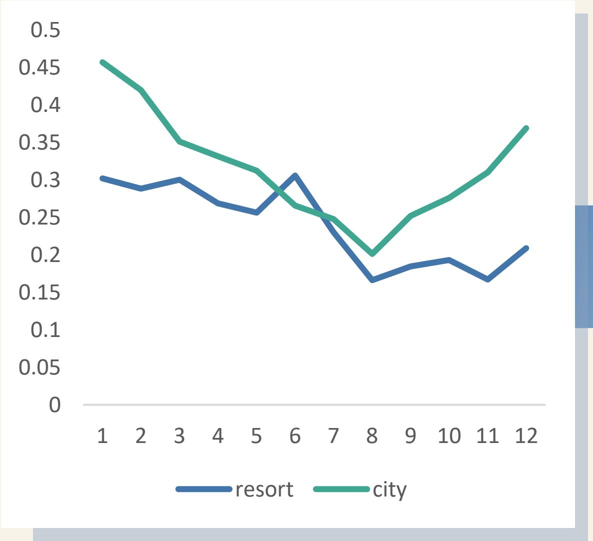
Copper Customer

- Predicted Cancellation Rate = 35.3%
- Low mean num of Fulfilling Appointment
- High mean Lead Time, High Price

Full Room Fee as Deposit

Other related Factors

Cancellation Rate of Different Month & Hotel Type



Month

Summer & Fall Low ✓

Winter & Spring High

10% Off

(Jul. – Nov.)

Hotel type

Resort Low ✓

City High

10% Off

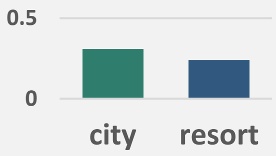
Summary &
Answers to Business Questions

Summary

Exploratory Data Analysis

– for cancellation rate

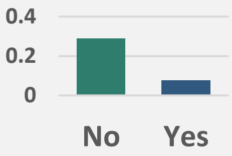
- **Hotel type**



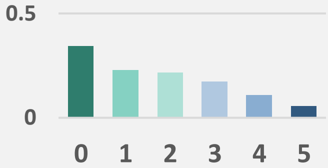
- **Residence season**

Jul ~ Nov	Low
Dec ~ Jun	High

- **Repeated gust**



- **Special request number**



Random Forest

- Optimal model compared to LDA and QDA
- Select 14 out of 31 features
- Adopt a security rate of 10% due to unbalance between FP and FN rate

Overbooking rate	23.2%
Accuracy	79%

Logistic Regression

$$P(Cancellation) = \frac{e^{\beta_0+\beta_1 LeadTime+\beta_2\ PreviousCancellation+\ \ldots\ldots}}{1+e^{\beta_0+\beta_1 LeadTime+\beta_2\ PreviousCancellation+\ \ldots\ldots}}$$

required_car_parking_spaces	-2.48	distribution_channel_Corporate	0.10
market_segment_Offline TA/TO	-1.62	adr	0.22
is_repeated_guest	-0.81	market_segment_Complementary	0.32
distribution_channel_Direct	-0.75	lead_time	0.38
total_of_special_requests	-0.63	customer_type_Transient	0.56
customer_type_Group	-0.44	previous_cancellations	1.74
previous_bookings_not_canceled	-0.24		

Clustering

- 3 key attributes：
  Lead time & Price & Previous booking not cancelled
- “Elbow” K=3
- 3 customer groups

- **Stratified Deposit Plan**

Predicted cancellation	Membership	Deposit discount
5.96%	Diamond Customer	50%
12.89%	Golden Customer	30%
35.33%	Copper Customer	0

Answers to Business Questions

1. What is the most suitable overbooking rate to reduce vacancy?

Random forest model → Predicted cancellation →

Overbooking rate = $Cancellation\ rate \times \frac{Current\ booking}{Total\ capacity}$ **= 23.2%**

2. Can overbooking perfectly solve cancellation problem?

- **Benefits** – improve the utilization of customers' cancellation
- **Imperfection** – plenty of vacancies remain

3. How can any other options tackle high cancellation problem?

Motivation → Tackle the root: to improve customers' credit on cancellation

Strategies:

- **Stratified Deposit Plan** – different deposit discounts for customers
- **More Initiative** – Upgraded Promotion Mechanism adjusting the distribution channel, and more based on data pattern

Appendix

Revenue Increased

hotel	arri_year	arri_month	rooms	prc	over	rev	total
City Hotel	2016	June	3923	108.8876	1	427165.9	5333071
City Hotel	2016	May	3676	108.64	1	399360.6	5333071
City Hotel	2016	October	4219	108.4667	1	457621	5333071
City Hotel	2016	September	3871	118.155	1	457378	5333071
City Hotel	2017	April	3919	121.885	1	477667.3	5333071
City Hotel	2017	June	3971	129.138	1	512806.9	5333071
City Hotel	2017	May	4556	132.1264	1	601968.1	5333071
Resort Hotel	2016	April	1867	68.64242	1	128155.4	5333071
Resort Hotel	2016	August	1685	190.9587	1	321765.4	5333071
Resort Hotel	2016	March	1778	57.08722	1	101501.1	5333071
Resort Hotel	2016	May	1802	71.42881	1	128714.7	5333071
Resort Hotel	2016	October	1984	66.71244	1	132357.5	5333071
Resort Hotel	2017	April	1742	87.71724	1	152803.4	5333071
Resort Hotel	2017	August	1800	207.3455	1	373221.9	5333071
Resort Hotel	2017	July	1754	177.6805	1	311651.5	5333071
Resort Hotel	2017	June	1676	117.7484	1	197346.3	5333071
Resort Hotel	2017	May	1757	86.27518	1	151585.5	5333071