Hotel Booking Cancellations

G1 Group 2
Brandon, Clara, Gugan, Kelvin,
Raam, Rong Jun, Vengka

BACKGROUND – WHY?

Direct impact on bottom line

Throws off any **demand forecasting**, **advanced costing** that hotel accounted for in the cancelled bookings

Workaround solutions

Revenue management strategies like overbooking, strict cancellation policies and dynamic pricing based on **intuition** and "**experience**"

PROBLEM STATEMENT

Problem

Overbooking and cancellation policies based on **intuition** and **"experience"** <u>**ONLY**</u>

Consequences

Loss sales

Fall in business reputation

Fall in customer loyalty

Value Generation

#1 Better forecasting and planning

#2 Constant improvement because of feedback loop

GOAL OF ANALYSIS

#1

Understand

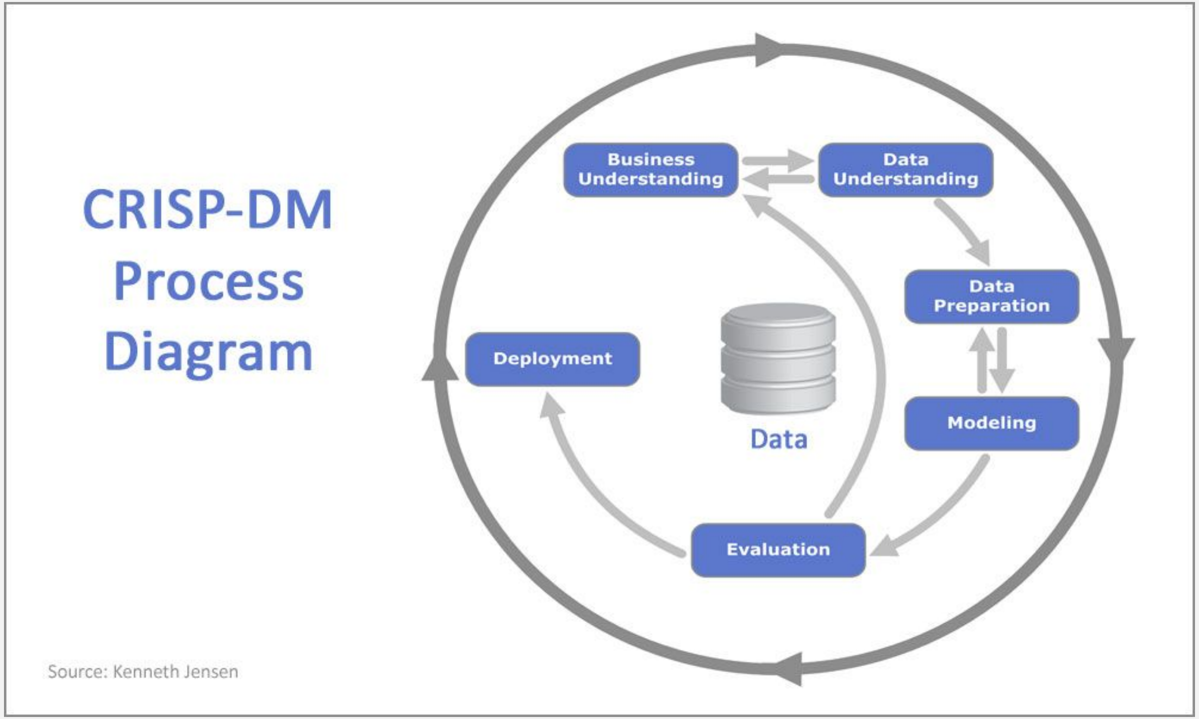
Is booking cancellations **correlated** to factors like lead time and market segment?

#2

Predict

Can we **predict with reasonable accuracy** if a booking will be canceled, based on its attributes?

CRISP-DM APPROACH



Cross-Industry Standard Process for Data Mining

DATA SET

Key Characteristics

- 120K rows, 32 columns
- **Context**: Portugal
- City hotel AND Resort hotel**
- **Source**: hotel's database
- Mostly processed, no duplicates

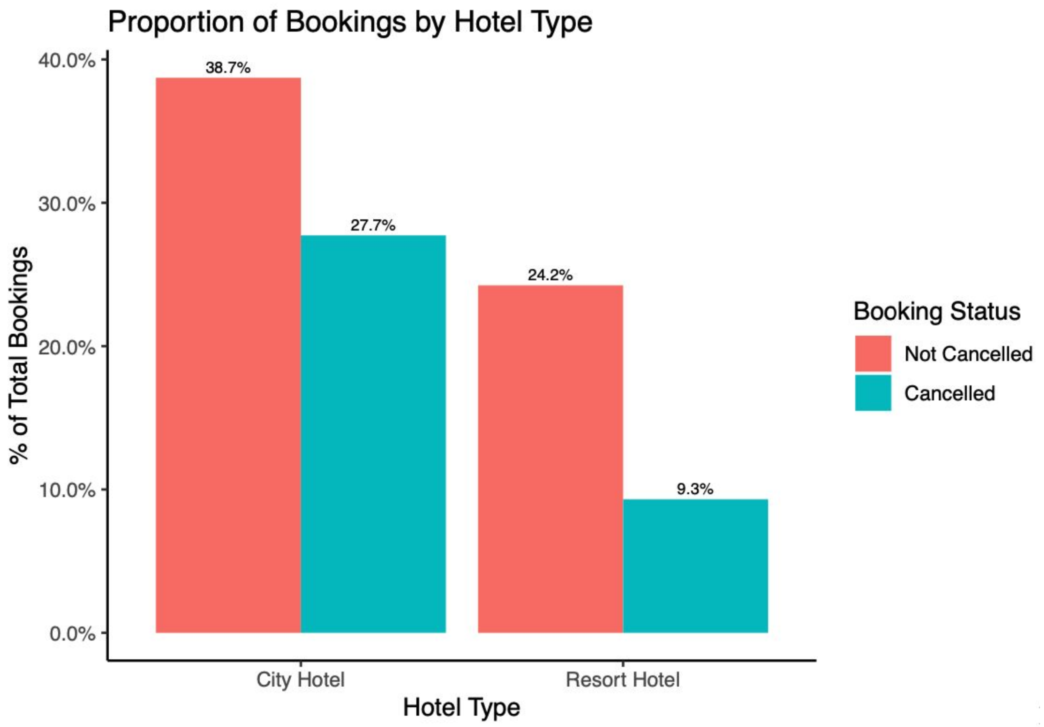
DATA EXPLORATION

Proportion of Total* Dataset

1. City Hotel (66%), 2x of Resort hotel (33%)

2. More cancelled bookings from City than Resort

Separate analysis for City/Resort Hotel?



DATA PRE-PROCESSING

- **N.A. columns**: Replace using corresponding column (“baby” -> “children”)
- **Direct indicators**: removed “reservation_status” and “reservation_status_date” columns
- **As.Factor**: Numerical variables that are ACTUALLY categories (is_repeated_guest, arrival_date_day_of_month)

DESCRIPTIVE ANALYTICS

- Top 3 Discoveries
- Their applications ("So-what?")
- **Complete dataset*

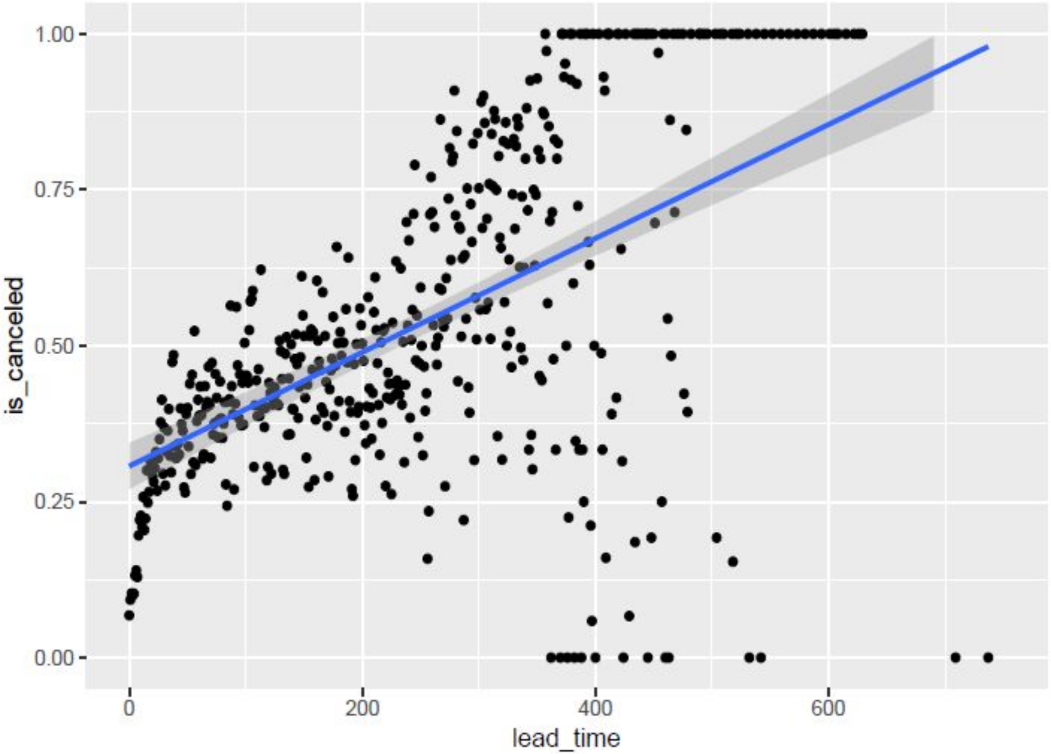
INSIGHT #1

Lead Time on Cancellations

- Scatterplot is quite concentrated
- Potential correlation

"So what?"

- For customers that book with large lead times, how might hotels incentivise customers not to cancel?



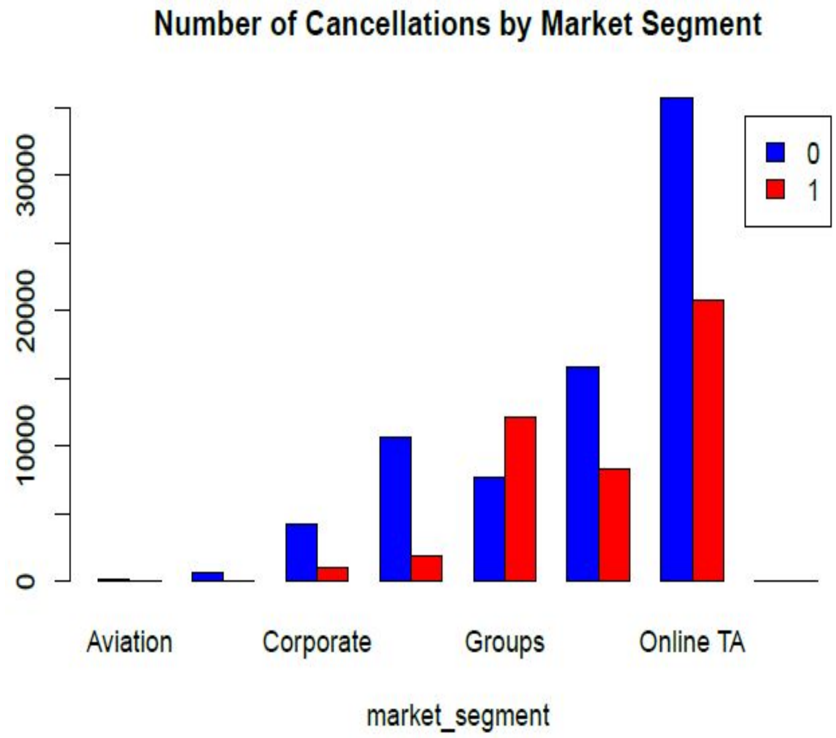
INSIGHT #2

Market Segment on Cancellations

- More bookings cancelled than booked for customer segment: "Groups"
- Related to business context, they have a lot of requirements that tend to change

"So what?"

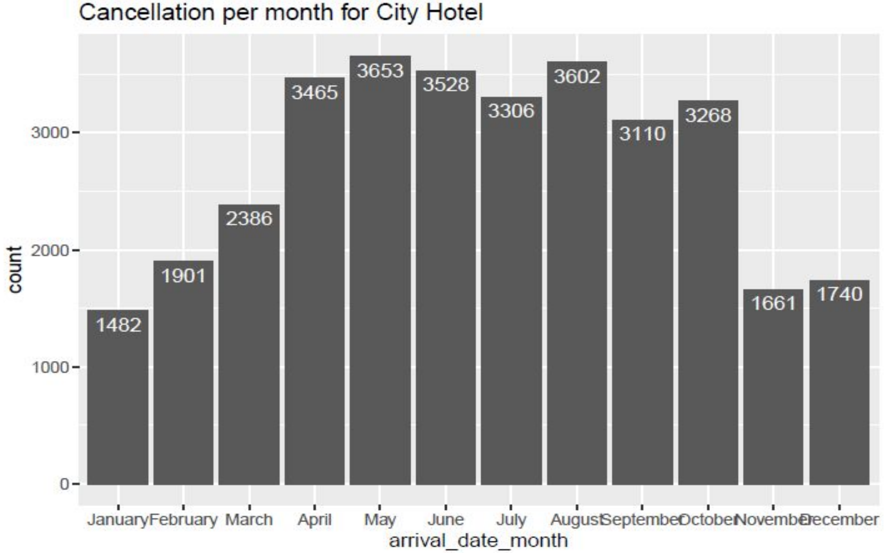
- More sensitive to Group bookings
- Special booking process to facilitate group bookings



INSIGHT #3

Seasonality of Cancellations

- **City**: April - October (work trips?)
- **Resort**: July AND August (summer?)



"So what"

- Different strategies for dealing with cancellations in City-Resort hotel
- Separate analysis for City-Resort

PREDICTIVE ANALYTICS

- CART
- Random Forest

DATA PREPARATION - CART

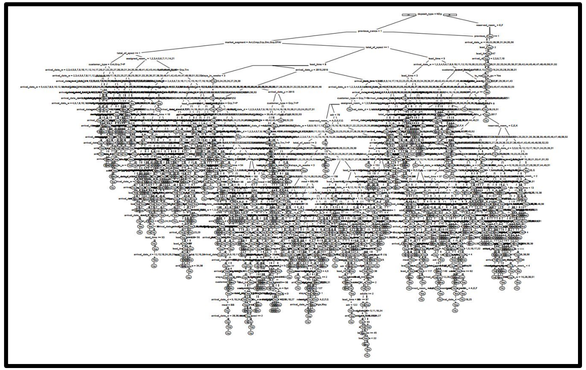
Built predictive model based on **City Hotel data only**

Excluded **reservation_status** and **reservation_status_date**

MODEL 1: CART

Cross-validation with **k=5** for a **80% to 20% training-test split**

Optimal cp = **3e^-0.5**



AUC = **0.88**

Benefits

- All variables can be considered
- Identify first few key factors in the decision tree

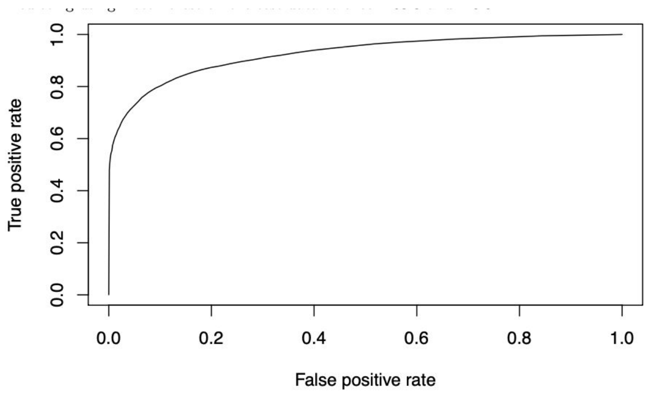
DATA PREPARATION - RANDOM FOREST

Model cannot handle variables with >53 levels

- Attempted to create new column up to 52nd level, and 53rd level as "others"
- Not scalable, just removed variables (3 of them) with >53 levels

MODEL 2: RANDOM FOREST

Manually tuned to get **OPTIMAL** parameters → **"ntree" = 50**, **"nodesize" = 1** and **"mtry" = 8**



AUC = **0.925**

Results

- *Built with City data ONLY
- Better AUC as compared to CART model
- Re-ran model with Resort data, similar outcomes

Evaluation of Models

CART Model

1. Revealed interesting variables for further analysis:
    a. deposit_type
    b. previous_cancellations
    c. total_special_requests

Random Forest Model

- Better prediction model than CART
- Might not be an issue to use City and Resort hotel data together

Next Steps:

- Further validate the three *customer behaviour variables* pointed out by CART

DIAGNOSTIC ANALYTICS

Test for Causality

Lead_time

Market_segment

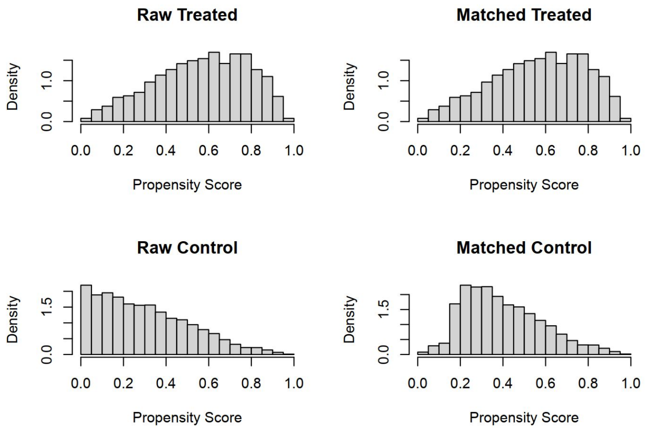
Deposit_type

Lead Time

Step 1: Created new column “lead_bin”

Step 2: Decided on value of split at lead_time == 100

Step 3: Matched data based on chosen control variables less than and more than lead_time == 100



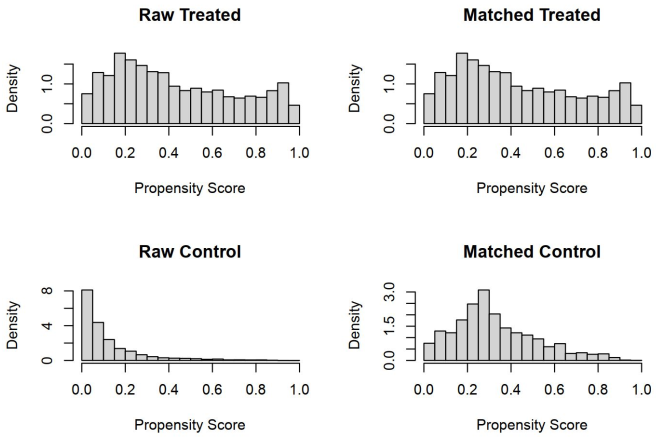
- Probability of cancellations did not change much
- Logistic regression to test significance

- Matching of data was not strong
- Probability differed more with split value of 100 changing to other values

Market Segment

Step 1: Created new column “group”

Step 2: Matched data based on market_segment == “Group”



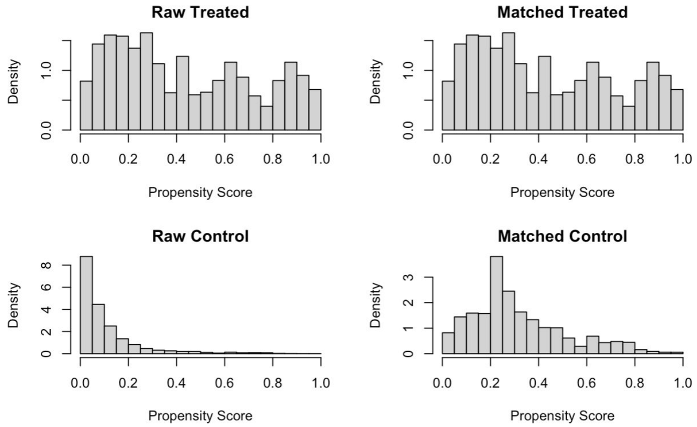
- Better matched data
- Probability of cancellations did not change much for both groups
- Logistic regression to test significance

Deposit Type

Step 1: Created new column “refund”

Step 2: Matched data based on deposit_type == “Non Refund”

No Deposit / Refundable	No Refund
0.3048058	0.9981349



- Better matched data
- Probability of cancellations did not change much for both groups
- Logistic regression to test significance

<u>Counter-intuitive Insight</u>:

- Implementing a penalty on bookings (non-refundable rooms) do not prevent customers from cancelling their bookings

LIMITATIONS AND NEXT STEPS

LIMITATION #1

Dataset limited to mostly customer demographic data

Demographic Data

Variables like # adults, # children, arrival dates, segments etc. don't seem to be as useful

Customer Behavior data

Variables relating to the behavior of customer bookings revealed to be more significant and insightful

LIMITATION #2

Scope of our dataset: City/Resort hotels in Portugal

Non-homogeneity of customers

Customer behavior for cancellations in Portugal may not be reflective of the rest of world

Non-homogeneity of location

Context of each booking/cancellation might differ based on location (Vacation? Business?)

NEXT STEPS #1

Include more **customer behaviour data**

Browser sessions data might be useful

Identify other **good predictors** of cancellations

Look for sources of **customer behaviour** data

Understand **customer booking journey** better

NEXT STEPS #2

Cross-reference with results and findings with data from other countries

Replicate model for data in other cities and continents that are different in: (Examples but not limited to)

- **Type of tourists they attract**: Cities that attract more business or leisure customers
- **Different customer demographics**: Cities with younger populations

Found another study on data from a hotel chain in Finland, revealed similar insights

THANK YOU

Q&A

ALL THE BEST FOR FINALS :)