Hotel Booking

Group 2:
Anan Song, Jiaci Jiang,
Jiayi Hao, Shuyue Yang

Overview

Summary

01

Research question,
Dataset,
Variables, Measures

EDA

02

Descriptive analysis,
Data cleaning

Models

03

Classification model,
Parameter selection,
Model performance

Conclusion

04

Conclusion,
Business
recommendation

01

Summary

Research question, Dataset, Variables, Measures

Research question

- **Which hotel reservations, given data on the booking and customer information, are most likely to be canceled?**

- Predict whether a new reservation will be canceled
- Modify hotel policies to reduce the cancellation rate and prevent losses. (implement an overbooking strategy)

Dataset

- **119390 observations** for a City Hotel and a Resort Hotel
- Each observation represents a hotel booking between July 1st, 2015 and August 31st, 2017
- Source: https://www.kaggle.com/datasets/mojtaba142/hotel-booking

Variables

- A mix of **36** quantitative and categorical variables
- **Booking information**: city/resort hotel, is_canceled, deposit type, arrival date, stays in weekend/weekday nights, etc.
- **Customer information**: adults, children, babies, country, customer type (Contract/Group/Transient/Transient-party), is_repeated_guest, etc.

Measures

- Target variable: **Is_Canceled** (if the booking was canceled (1) or not (0))

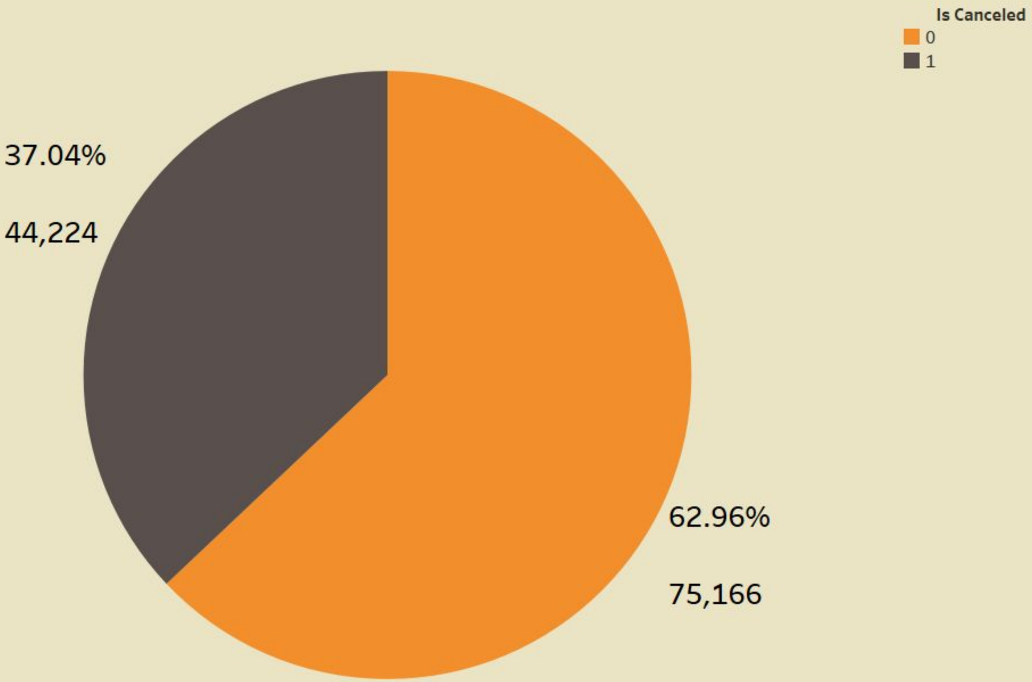
- Accuracy: How many bookings did we correctly predict among all test set?
- **Recall**: How many bookings were predicted to be canceled out of all the bookings that were canceled in real situation?
- **Precision**: How many bookings were actually canceled out of all the bookings predicted to be canceled?

02

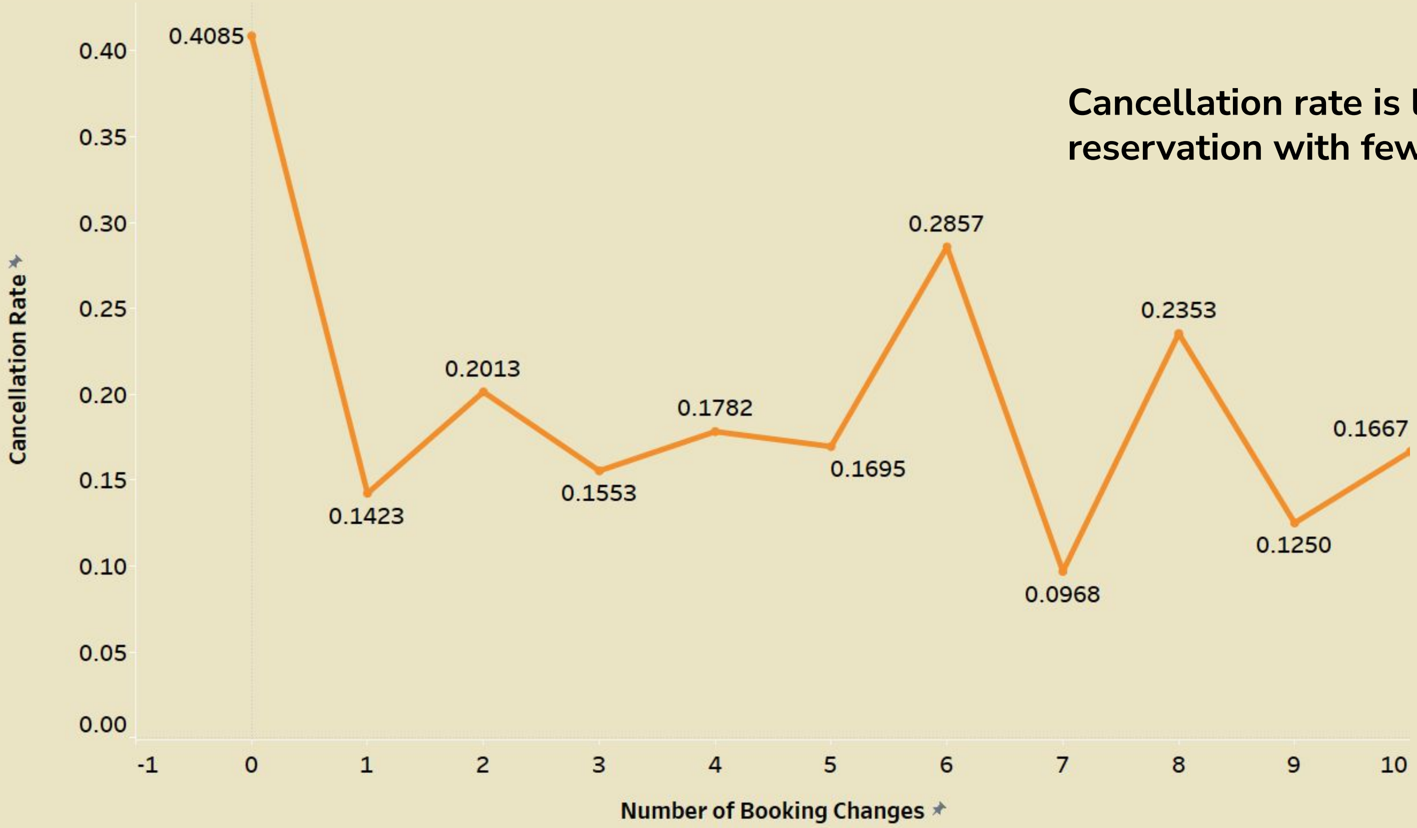
EDA

Descriptive analysis,
Data cleaning

Target Variable

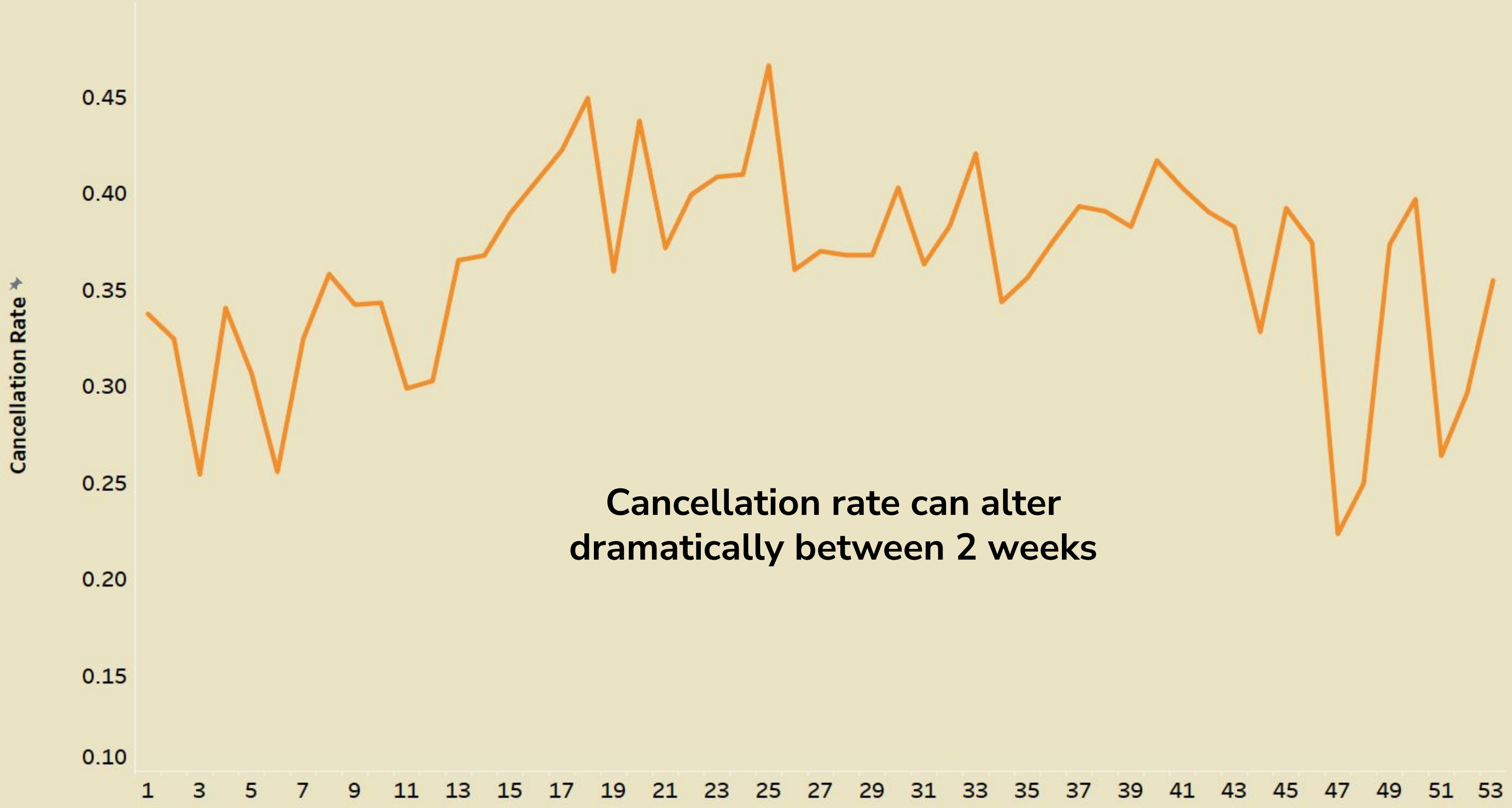


Booking Changes



Cancellation rate is lower in reservation with few changes

Week Number



Cancellation rate can alter dramatically between 2 weeks

Cleaning - Attribute

There are 3 groups of attributes we want to remove from the dataset:

1. Attributes with private information of hotel customers
   - ❑ Name; Email; Phone-number; Credit card number
2. Attributes that overlap with other attributes
   - ❑ Repeated: reservation_status; assigned_room_type
   - ❑ Overlap: reservation_status_date; arrival_date_month; arrival_date_day_of_month; (Keep) arrival_date_week_number
   - ❑ Highly Correlated: distribution_channel
3. Other attributes that are not very helpful to answer our question
   - ❑ Missing values: agent, company
   - ❑ Biased Distribution: country

Cleaning - Data

1. Found 4 missing data in column ‘children’

   - filled with 0

2. No invalid/unreasonable values found

3. Turn all the categorical variables into dummy variables.

03

Models

Classification models,
Parameter selection,
Model performance

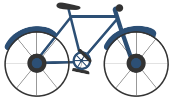
Classification Models

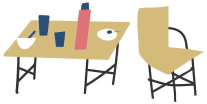
Multinomial Naive Bayes

1. Discrete Attributes (e.g. dummy variables)

Decision Tree

1. criterion = "entropy"
2. ccp_alpha = .001



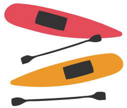


K-Nearest Neighbor

1. Standard Normalization
2. # of neighbor: 345

Random Forest

1. criterion = "entropy"
2. # of trees = 200



Model Performance on cleaned dataset

Model Type	Accuracy	Recall (1)	Precision (1)
Multinomial Naive Bayes	65%	52%	52%
K-Nearest Neighbor	79%	58%	81%
Decision Tree	81%	62%	83%
Random Forest	**86%**	**76%**	**86%**

Model Performance on more relevant attributes

Model Type	Accuracy	Recall (1)	Precision (1)
Multinomial Naive Bayes	65%	52%	53%
K-Nearest Neighbor	**80%**	**59%**	**83%**
Decision Tree	81%	62%	83%
Random Forest	**86%**	**75%**	**85%**

04

Conclusion

Conclusion,
Business recommendation

Conclusion

- **Best model: Random Forest**
  - Highest accuracy(86%) → being correct overall
  - Highest precision(85%) → high so customers won’t have no rooms
  - Highest recall(75%) → high so we can make overbook decisions to minimize risks and losses
- **Model improvements**
  - Peak seasons
  - Pandemic
  - Other relevant attributes: weather, hotel location, star category, etc.

Business Recommendation

- **Goal: maximize profit**
  - Advantages of overbooking: mitigating loss, full occupancy, compensation is cheaper than having empty rooms
  - Disadvantages: harms guest experience, reputation, and long-term profit
- **Implementations**
  - Use random forest model to estimate the right number of overbookings
  - Predictive model can manipulate and benefit large scale of data
  - Don’t book out the loyal customers and highest-priced reservations
  - Determine the ideal compensation
  - Have overbooking partnerships with neighboring hotels

Reference: mews.com, “What is an overbooking strategy in hotels and what are its advantages?”
https://www.mews.com/en/blog/hotel-overbooking-strategy

Thanks!