SC1015 Mini-Project

FCE3 Group 8

Joel Tan Xin Wei
Han Sheng Jie, Philip

Table of contents

01

Problem Definition

Introduction

In 2019, global cancellation rate of hotel reservations was almost 40%

“Book now, pay later” practice

Zero booking fee

High cancellation rates → Lower room occupancy rates → Lower revenue for hotels

Dataset

Hotel Reservations Dataset
by Ahzan Raza

Found on Kaggle

The Variables

Data Dictionary

- **Booking_ID**: 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 week nights (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:
- **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. The values are ciphered (encoded) by INN Hotels.
- **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 the 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.

19 variables

Problem Definition

How do variables such as room type, lead time, and number of previous cancellations affect the likelihood of a booking being cancelled?

02

Data Cleaning

Data Cleaning

Remove Invalid Data

Invalid dates
Adults + children is 0
Weekdays + weeknights is 0

Arrival Date and Time

Combining arrival year, month, date into single variable

Variable Encoding

Encoding into machine-readable categorical values
E.g. One Hot Encoding
Label Encoding

Data Cleaning

115 invalid entries removed

```python
#drop any invalid dates (NaT)
train = train.dropna(subset=['arrival_datetime'])
train.reset_index(drop=True, inplace=True)
train.shape
```

✓ 0.0s

```
(36238, 17)
```

Number of rows has been reduced from 36275 to 36238, thus there were 37 cases of invalid dates that was removed

```python
#drop any rows where both adults and children are 0 (nobody booked)
train = train[~((train['no_of_adults'] == 0) & (train['no_of_children'] == 0))]
train.reset_index(drop=True, inplace=True)
train.shape
```

✓ 0.0s

```
(36238, 17)
```

There was no change in rows after this, so there was at least one person per booking.

```python
#drop any rows where both weekend nights and week nights are 0 (did not stay a night)
train = train[~((train['no_of_weekend_nights'] == 0) & (train['no_of_week_nights'] == 0))]
train.reset_index(drop=True, inplace=True)
train.shape
```

✓ 0.0s

```
(36160, 17)
```

Number of rows has been reduced from 36238 to 36160, thus there were 78 cases of invalid dates that was removed

03

Exploratory Data Analysis

Variables

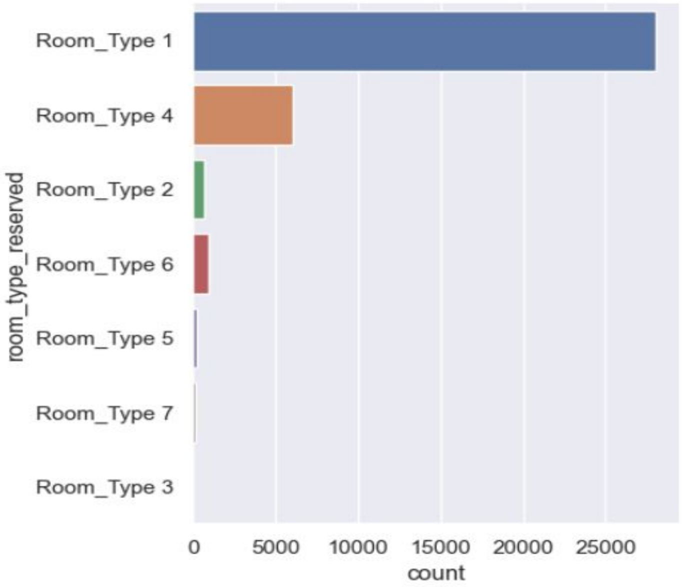
```
In [2]:  1  clean = pd.read_csv('CleanedEDA.csv')
          2  clean.info()
```

```
 ---  ------                                --------------  -----
 0    Unnamed: 0                            36160 non-null  int64
 1    Booking_ID                            36160 non-null  object
 2    no_of_adults                          36160 non-null  int64
 3    no_of_children                        36160 non-null  int64
 4    no_of_weekend_nights                  36160 non-null  int64
 5    no_of_week_nights                     36160 non-null  int64
 6    type_of_meal_plan                     36160 non-null  object
 7    required_car_parking_space            36160 non-null  int64
 8    room_type_reserved                    36160 non-null  object
 9    lead_time                             36160 non-null  int64
 10   market_segment_type                   36160 non-null  object
 11   repeated_guest                        36160 non-null  int64
 12   no_of_previous_cancellations          36160 non-null  int64
 13   no_of_previous_bookings_not_canceled  36160 non-null  int64
 14   avg_price_per_room                    36160 non-null  float64
 15   no_of_special_requests                36160 non-null  int64
 16   booking_status                        36160 non-null  object
 17   arrival_datetime                      36160 non-null  object
dtypes: float64(1), int64(11), object(6)
```

15 variables for consideration

Uni-Variate Analysis

Room Type Reserved



Room Type 1 is most commonly chosen room

- Standard room

Room Type 4 has a sizeable number of bookings

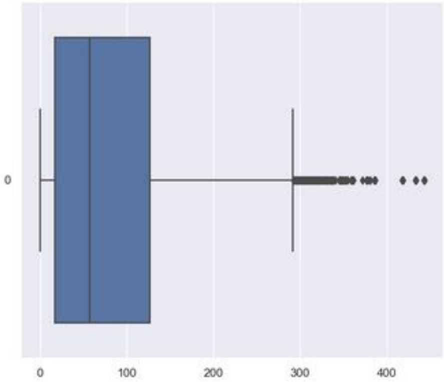
- Larger room for families

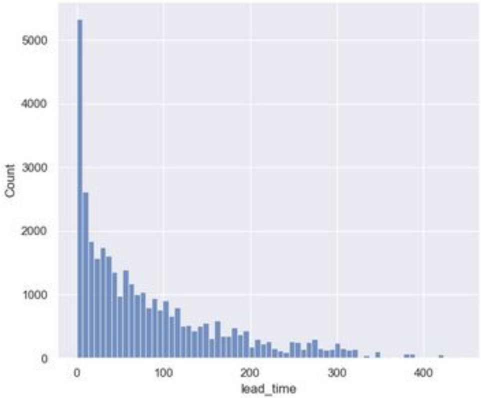
Other types have small number of bookings

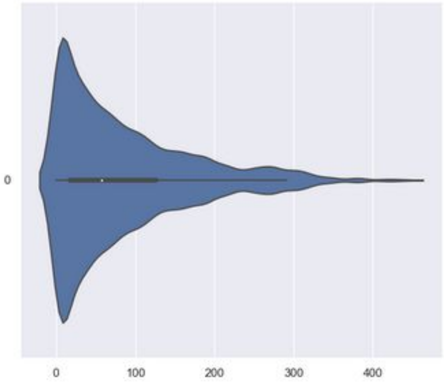
- Luxury room types e.g. suites

Uni-Variate Analysis

Lead Time (Days booked in advance)







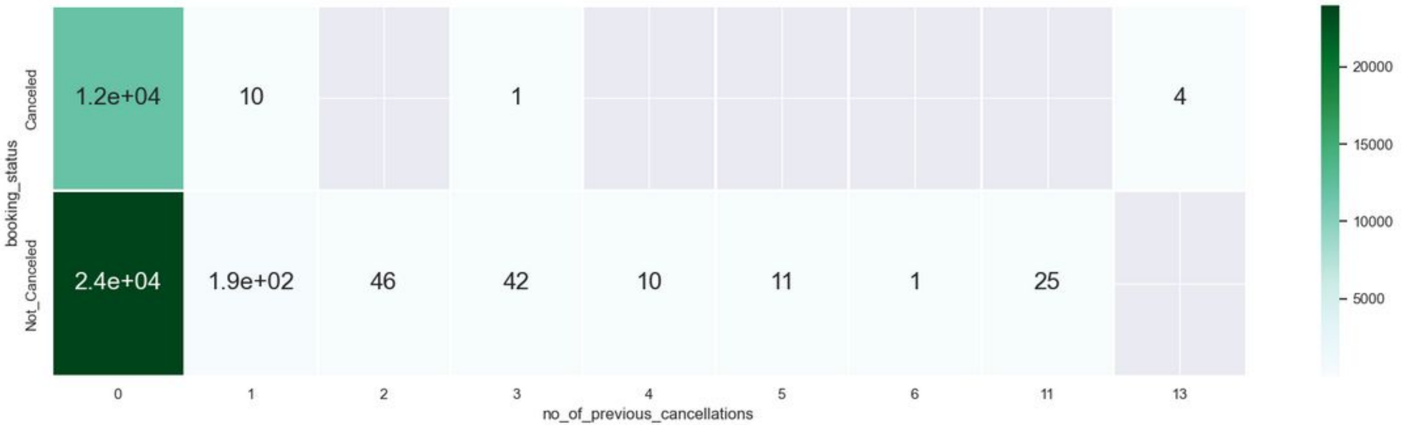
Mean > Median
85 57

Small number of bookings made many days in advance
→ Right-skewed graph

Bi-Variate Analysis

Number of Previous Cancellations and Booking Status

```
booking_status
Not_Canceled     24284
Canceled         11876
Name: count, dtype: int64
```

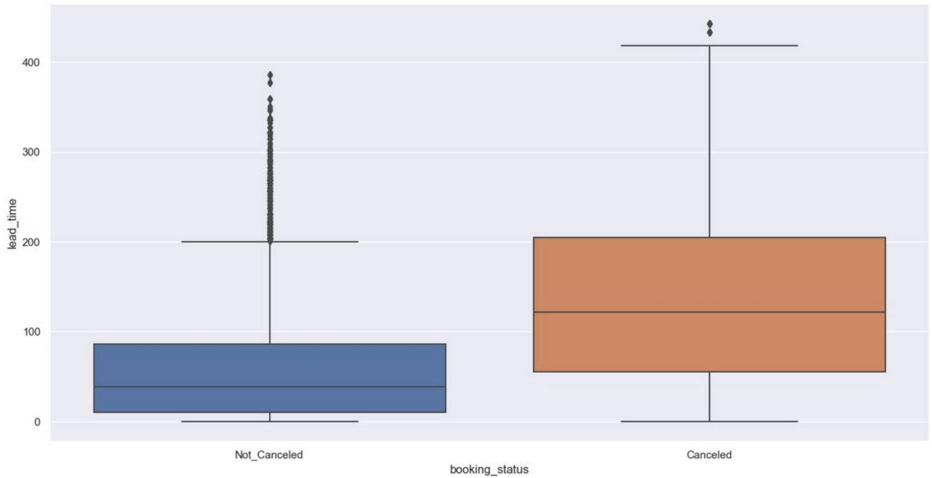


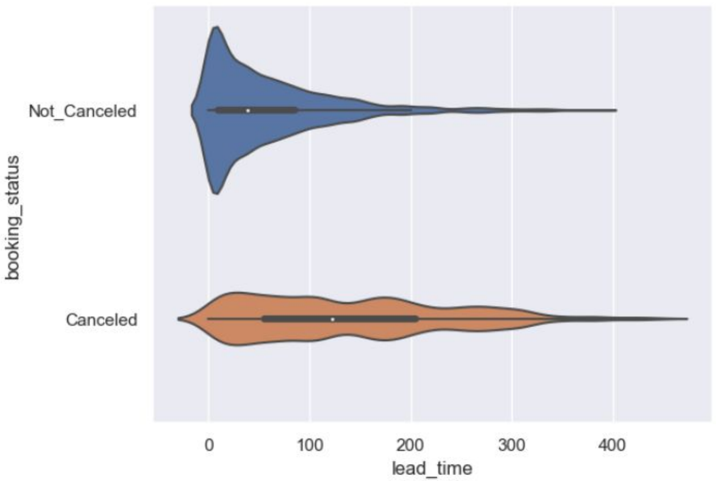
Bookings with >= 1 previous cancellation → Lower ratio of cancelled bookings compared to total

People who previously cancelled their bookings are less likely to cancel them again

Bi-Variate Analysis

Lead Time and Booking Status





Cancelled bookings had a greater median lead time → Possibly due to sudden issues coming up

Cancelled bookings have more spread out lead times

04 Machine Learning

Identifying the top predictors

- Started with Decision Tree Classifier
- Used all individual variables to predict booking_status
- Majority was around 67% accurate

Classification Score	Predictor
0.665597	no_of_adults
0.669912	no_of_children
0.675664	no_of_weekend_nights
0.672013	no_of_week_nights
0.673009	type_of_meal_plan
0.675111	required_car_parking_space
0.665265	room_type_reserved
0.769580	lead_time
0.670907	repeated_guest
0.676549	no_of_previous_cancellations
0.673341	no_of_previous_bookings_not_canceled
0.704093	avg_price_per_room
0.668142	no_of_special_requests
0.668584	market_segment_type_Aviation
0.667146	market_segment_type_Complementary
0.672456	market_segment_type_Corporate
0.669801	market_segment_type_Offline
0.674115	market_segment_type_Online

Using top predictors / all predictors

- Used top 4 predictors: 80% accuracy
- Using all predictors: 86% accuracy

```
Top Predictors: 0.8
Everything:  0.86
```

Random Forest Classifier

- Ensemble Learning
  - Creates multiple decision tree
  - Each tree predicts independently
  - Final prediction averages the predictions of all trees

- Hyperparameters for performance
  - N_estimators: number of trees in forest
  - max_depth

GridSearchCV

- Assist in finding best hyperparameters for Random Forest
  - n_estimators
  - Max_depth
- Cross validates 5 times

```python
# Random Tree Classifier using Train Data
# Define the Parameter Grid
param_grid = {
    'n_estimators': [100, 500, 1000],
    'max_depth': [None, 5, 10, 15]
}
rtree = RandomForestClassifier()
#Use Grid Search to Find the Best Parameters
grid_search = GridSearchCV(estimator=rtree, param_grid=param_grid, cv=5)
grid_search.fit(X_train, y_train)

#Use the best tree from the grid search
best_rtree = grid_search.best_estimator_
```

Logistic Regression

Usually used for binary classification and prediction

Why not Linear Regression?

Handling categorical variables e.g. meal plan type, room type

Linear regression is usually used for numeric variables only

05 Outcomes

3 Models Compared

86%

Decision Tree

Quite accurate

88%

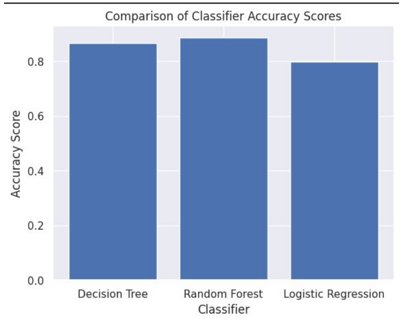
Random Forest

Most accurate model, can be recommended to predict booking status

79%

Logistic Regression

Not as accurate as the other two models



06
Final Insights

Learnings

- Cleaning invalid values of dataset
- Different encoders and their use cases
  - One Hot Encoding
  - Label Encoding
- How to analyse variables, both individually and in relation to the variable of interest, booking status
- New machine learning techniques
  - Logistic Regression
  - Random Forest Classifier
    - GridSearchCV
- Advantages/Disadvantages of each learning model

Data-driven insight

- Reliable Models (88% accuracy)

- Recommendations
  - Dataset shows that 1 in 3 people cancels their booking
  - Using our model allows the hotel to
    - Overbook properly
    - Predict cancelled bookings earlier

Thank you!

CREDITS: This presentation template was created by **Slidesgo**, and includes icons by **Flaticon**, and infographics & images by **Freepik**

Please keep this slide for attribution