Hotel Booking Demand

Xingyue Fang
Xinyi Gu
Guillermo Trefogli
2022.03.17

CONTENTS

01 / Background: The Business Problem

02 / Goal & Approach

03 / Data Processing

04 / Modeling

05 / Conclusions

PART ONE 1

Background:
The Business Problem

Background: The Business Problem

Booking cancellations in Hotel Industry

- Hotel industry give customers the possibility of choose in advance the services they want to consume; this feature in the industry is referred as booking.
- Companies deal with the risk of facing looses due to booking cancellations. This drives to facing the economic problem of not maximizing profits.
- A key input is critical to deal with this problem: information about the **probability of booking cancellations**.
- Having this input will allow business in the industry to improve the management of bookings. Basically, they would be able to apply a set of effective policies/procedures based on the information of the probability of booking cancellations for the characteristics of bookings. This ultimately will allow them to minimizing looses (or **maximize profits**).

PART TWO 2

Goal & Approach

Goal & Approach

Goal

The main goal of this paper is to provide business in the industry with recommendations to optimize their profits. To do so, we will offer an analytical strategy that builds a set of models and choose the best one to predict Hotel bookings cancellation for clients.

Approach

We apply machine learning techniques to approach this problem. We perform this analysis using a real business dataset (available at kaggle.com) containing 32 variables and almost 120,000 observations.

We start briefly describing the data at hand. Then, we built the following machine learning models: 1.Decision Tree, 2. Random Forest, 3. Boosting, and 4. Neural Networks. We finally compare their performance and choose the best one to predict the probability of booking cancellations.

PART THREE 3

Data Processing

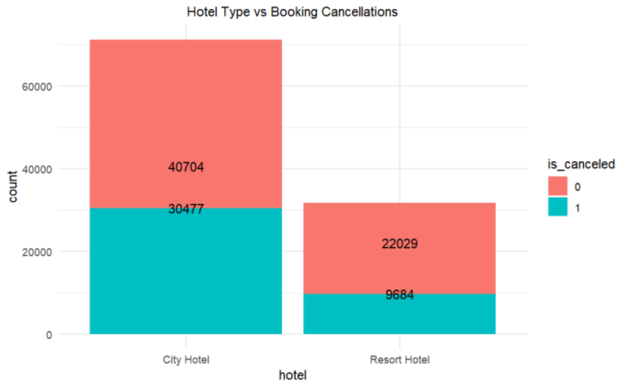
Data Introduction

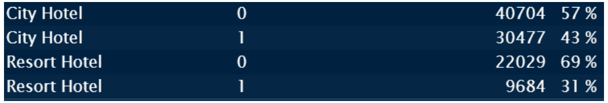
Exploring the data and the business problem

Main features of the dataset:

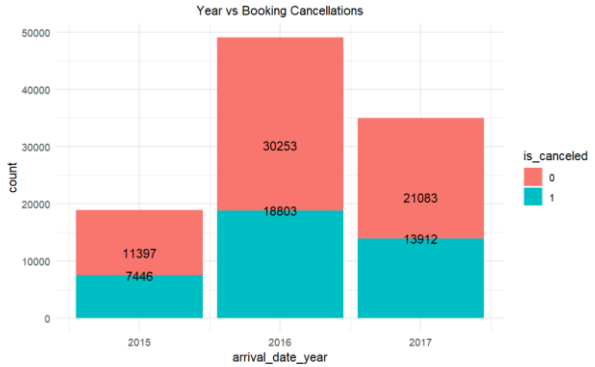
- Bookings due to arrive between July 01, 2015 and August 31, 2017.
- Two type of hotels: City and Resort.
- The unit of observation is a booking.
- Outcome of interest: a dummy variable for cancellation or not.
- Among the 31 potential predictors:
  - ❖ Characteristics of the booking (hotel type, date, stays type of night, type of room, parking option)
  - ❖ Characteristics of customer (number of adults, children, country),
  - ❖ Records for the customer (repeated guest, previous cancellations),
  - ❖ Others.

Characteristics of bookings in the dataset:



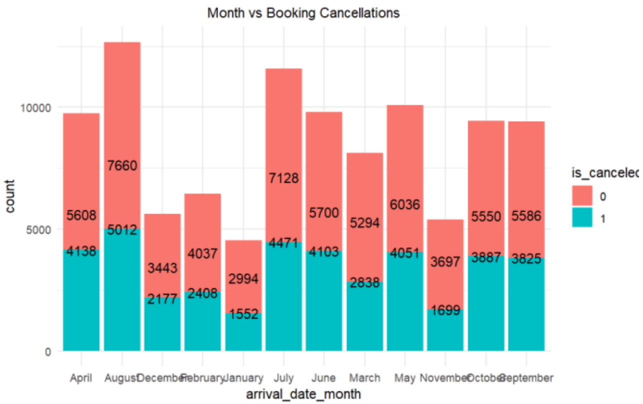


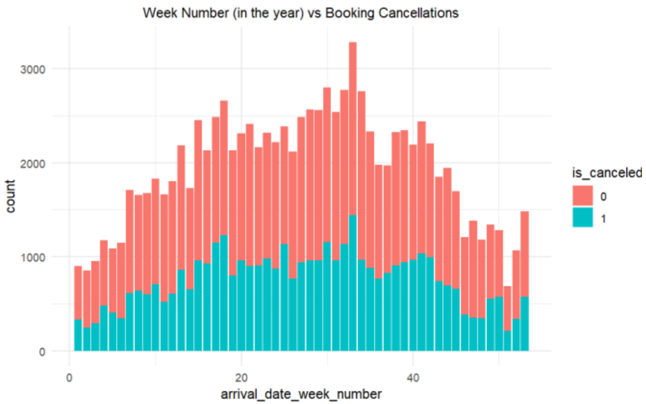
City Hotel	0	40704	57 %
City Hotel	1	30477	43 %
Resort Hotel	0	22029	69 %
Resort Hotel	1	9684	31 %

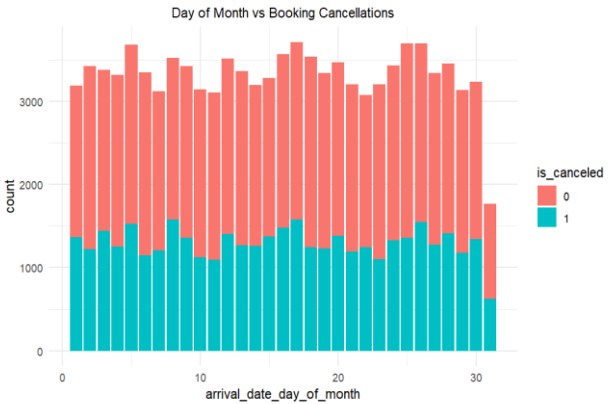


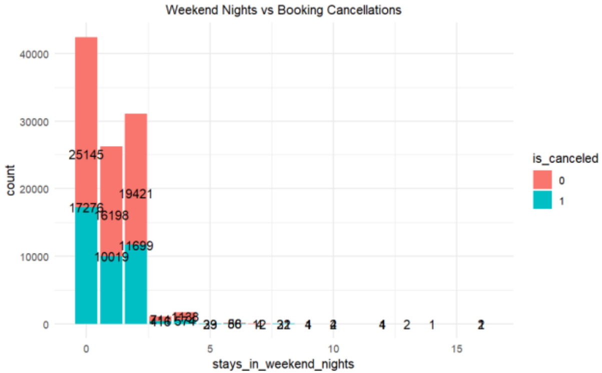
Data Introduction

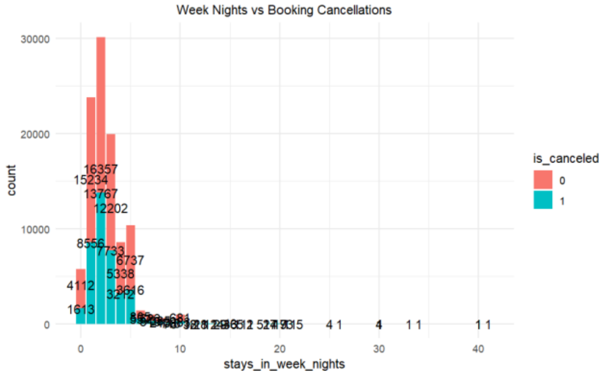
Characteristics of bookings in the dataset:





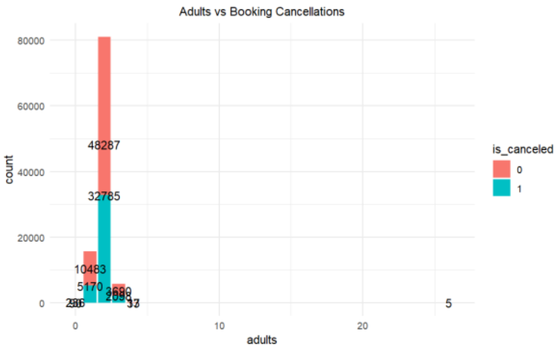


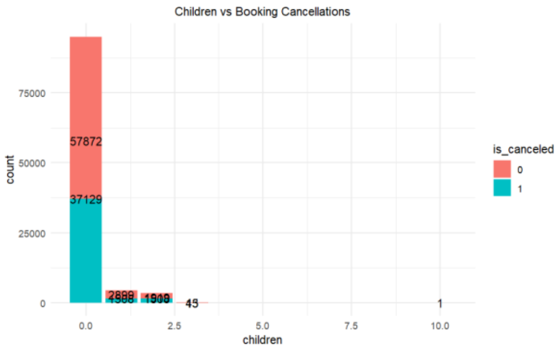


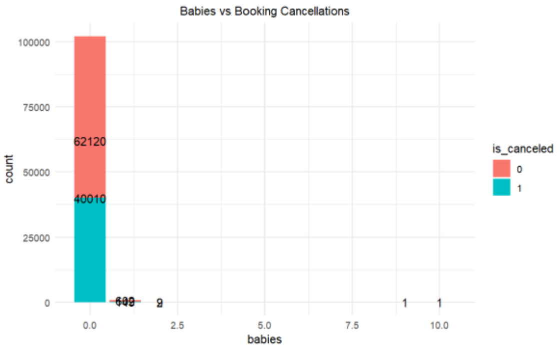


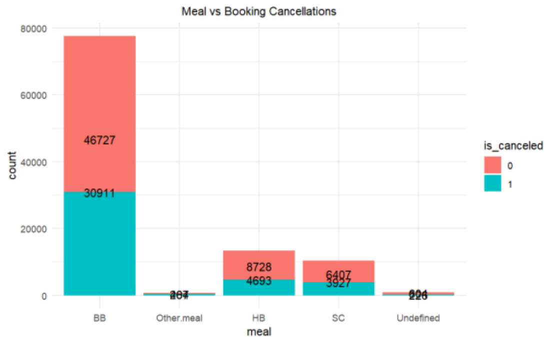
Data Introduction

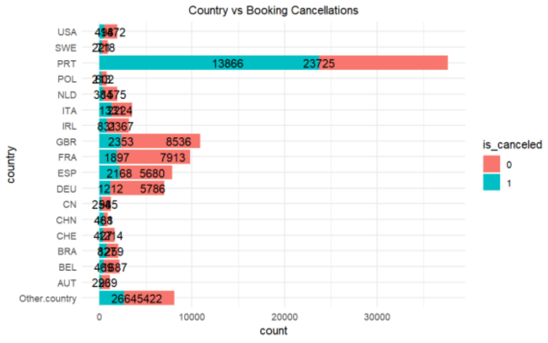
Characteristics of bookings in the dataset:

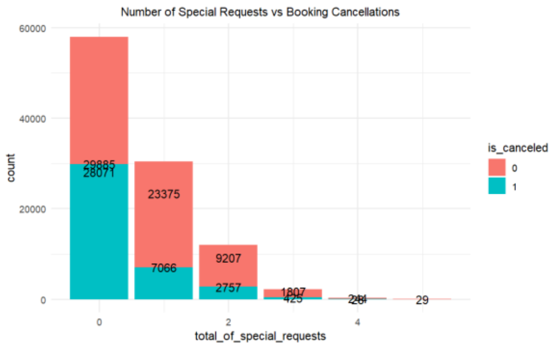

Adults vs Booking Cancellations


Children vs Booking Cancellations


Babies vs Booking Cancellations

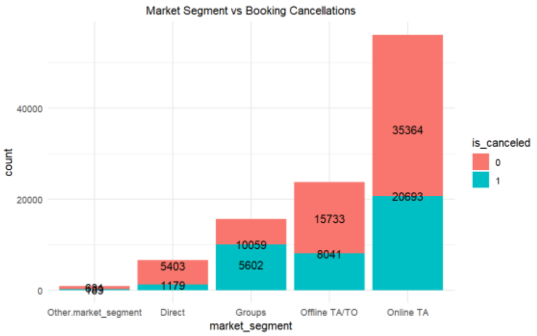

Meal vs Booking Cancellations

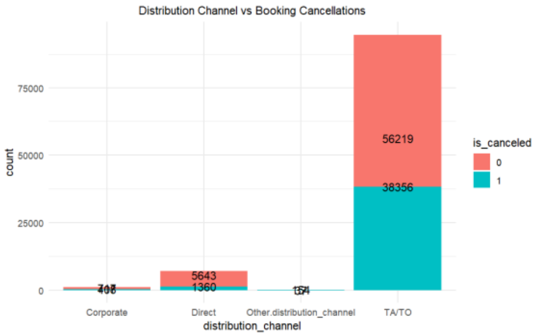

Country vs Booking Cancellations

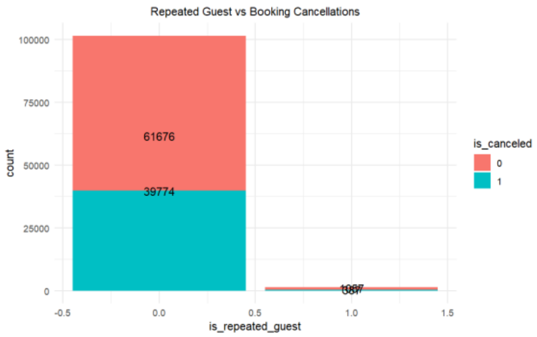

Number of Special Requests vs Booking Cancellations

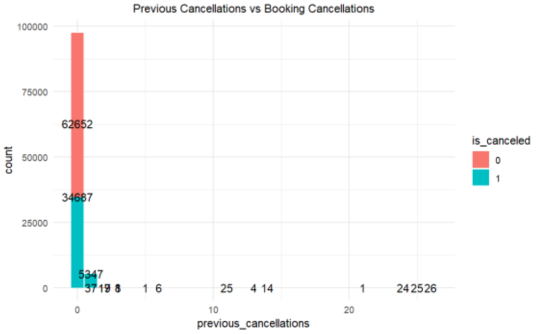
Data Introduction

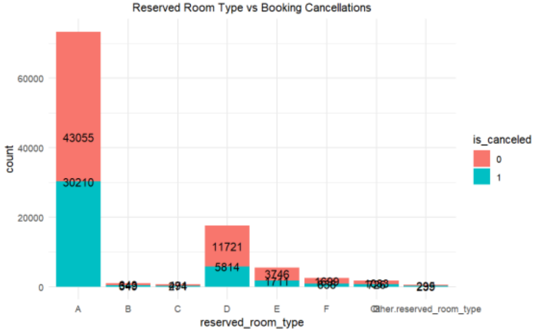
Characteristics of bookings in the dataset:

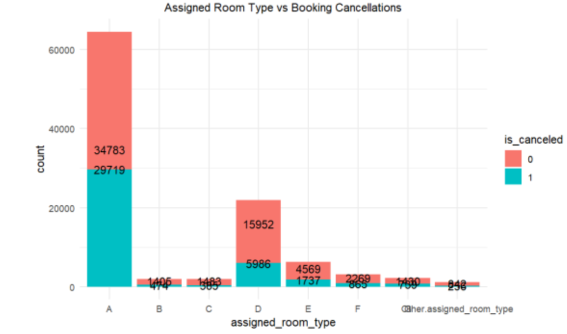






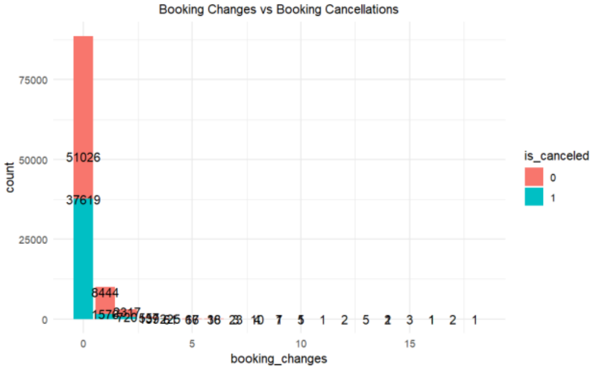


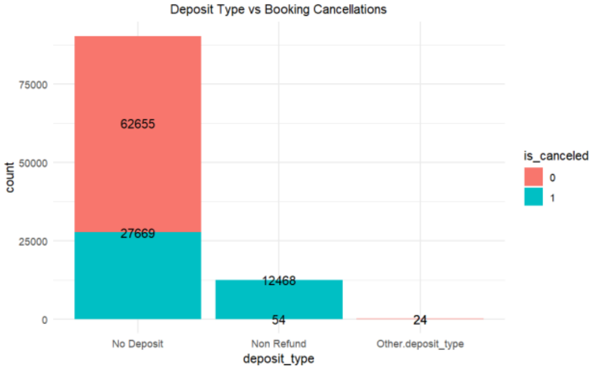


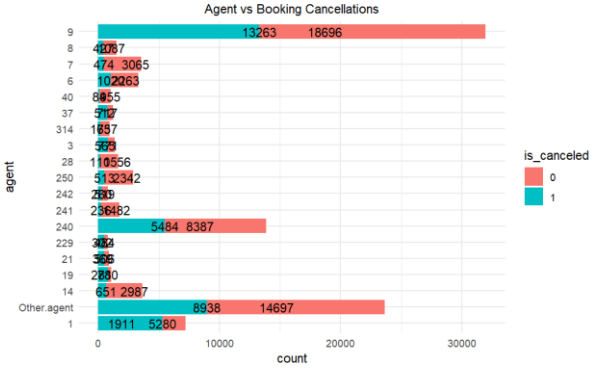


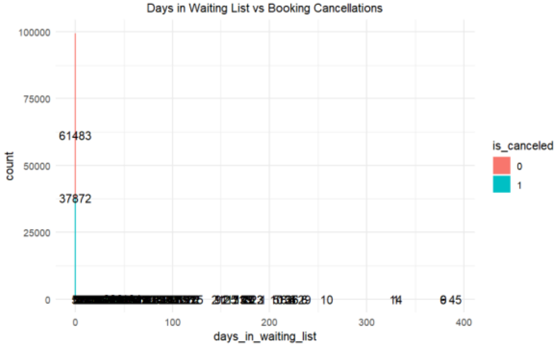
Data Introduction

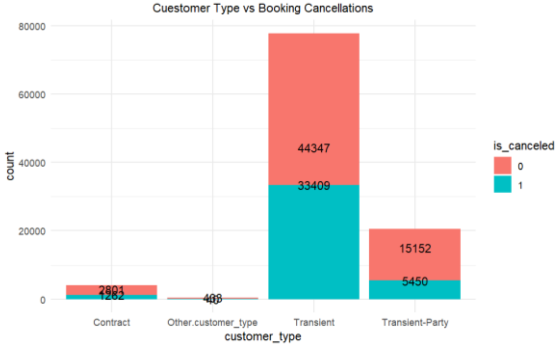
Characteristics of bookings in the dataset:

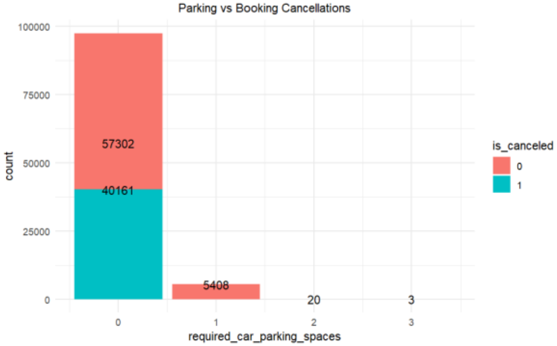








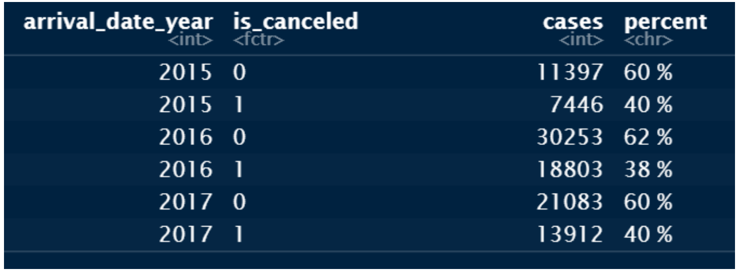




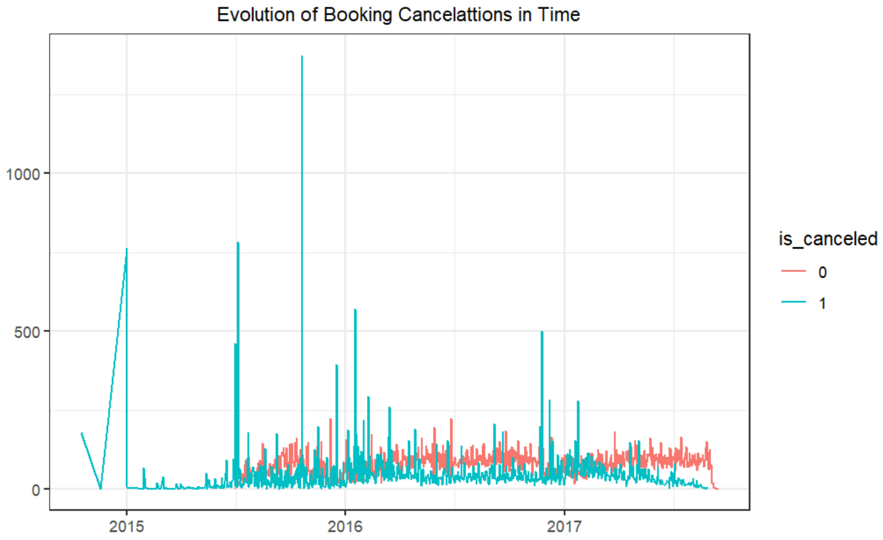
Data Introduction

Exploring the data and the business problem

Booking cancellations seems to be a relevant topic to analyze based on its prominence in the business cycle
(not impact on profits):



arrival_date_year	is_canceled	cases	percent
<int>	<fctr>	<int>	<chr>
2015	0	11397	60 %
2015	1	7446	40 %
2016	0	30253	62 %
2016	1	18803	38 %
2017	0	21083	60 %
2017	1	13912	40 %



Data Processing

1. Deal with NA

Drop NA rows; drop columns with too muvh NA.

2. Select variables as predictors

There is one variabe that is almost equal to y, so we delete that.

3. Deal with different types of data

Use as.factor to change nominal variable into factor.

PART FOUR 4

Modeling

Modeling

Models

1. Decision Tree
2. Random Forest
3. Boosting
4. Neural Networks

Details

Using grid search to tune parameters

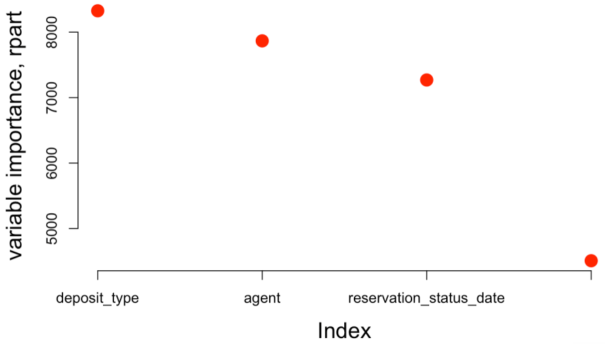
Modeling - Decision Tree

Confusion Matrix (on test set) :

	Predict 0	Predict 1
Actual 0	12221	385
Actual 1	721	7252

Accuracy: 94.63%

Variable Importance



Three most important features are deposit type, agent, and reservation status date.

Modeling - Random Forest

Confusion Matrix (on test set) :

	Predict 0	Predict 1
Actual 0	12464	94
Actual 1	737	7284

Accuracy: 95.96%

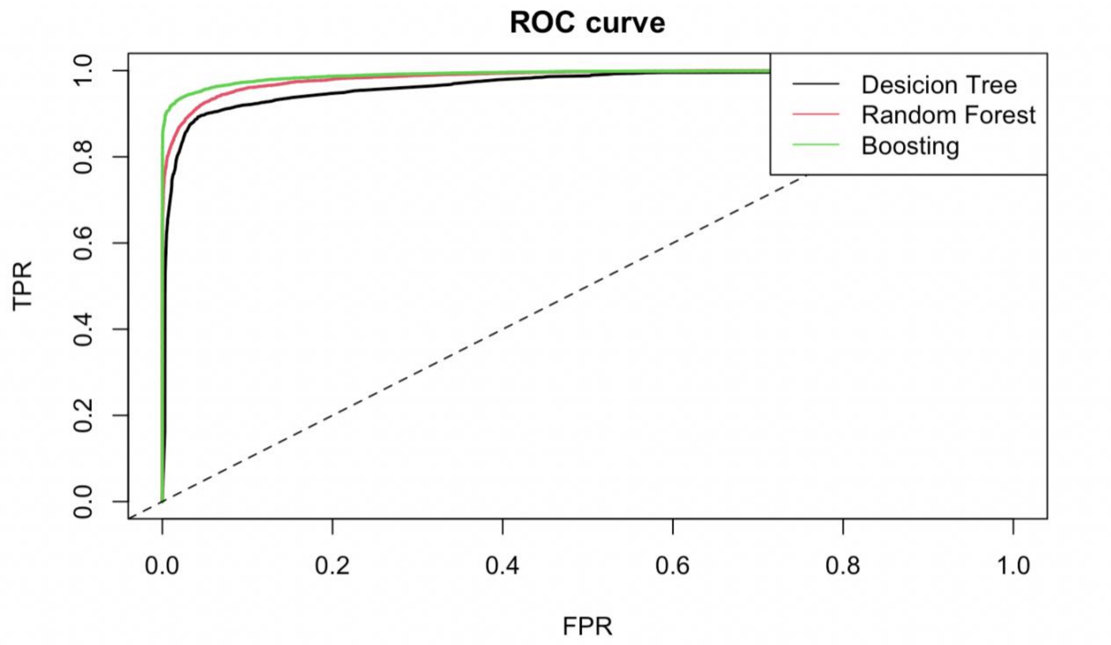
Modeling - Boosting

Confusion Matrix (on test set) :

	Predict 0	**Predict 1**
Actual 0	12446	160
Actual 1	612	7361

Accuracy: 96.25%

Modeling - ROC Curve



Boosting model has the highest AUC of 0.98, while AUC of Random Forest is 0.972 and 0.967.

Modeling - Neural Network

We found the best parameters with perfect prediction on validation set using Grid Search. But not that good on test set.

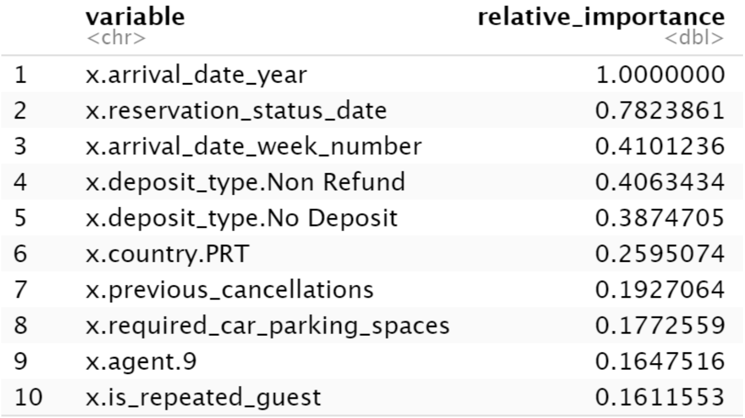
Confusion Matrix (on test set) :

	Predict 0	**Predict 1**
Actual 0	8612	3918
Actual 1	2001	6048

Accuracy = 0.712

AUC value = 0.749

Variable importance:



	variable <chr>	relative_importance <dbl>
1	x.arrival_date_year	1.0000000
2	x.reservation_status_date	0.7823861
3	x.arrival_date_week_number	0.4101236
4	x.deposit_type.Non Refund	0.4063434
5	x.deposit_type.No Deposit	0.3874705
6	x.country.PRT	0.2595074
7	x.previous_cancellations	0.1927064
8	x.required_car_parking_spaces	0.1772559
9	x.agent.9	0.1647516
10	x.is_repeated_guest	0.1611553

PART FIVE 5

Conclusion

Conclusion and Business Meaning

1 Final Model

- We are showing hotel industry business the benefits of applying machine learning tools to get useful information to make better decisions regarding business problems, in this case booking cancellations.
- Particularly, we find that **our final model is boosting model**, with **accuracy of 96%** on test set.

2 Recommendations

- Hotel industry companies can choose an optimal level of threshold to maximize the profits according to its own revenues and costs using this model. In other words, **the best model will be the one that allow each company to increase its profits.** The discussion based on level of accuracy of models is an initial step for this analysis.
- This strategy will allow you to better approach the **administration of bookings**, for example, by creating set of policies, according to the probability of booking cancellations. By doing so, you will minimize losses or **maximize profits** in your business.

THANK YOU