HOTEL BUSINESS CASE

By: Aaron Xiao

INTRODUCTION

- Hotels must be able to predict their cancelation rates reasonably well to ensure that they do not overbook or under-book rooms

BACKGROUND

- A city hotel chain approached our data analytics team and would like to know the number of possible cancelations the hotel could receive, in order to overbook their rooms appropriately.

- The owner of the city hotel wants to know by how much they should overbook their facilities to ensure that the hotel can operate at maximum capacity.

ANALYTICAL QUESTION

- Understanding business question:
  - Predict cancelation rate
  - Cancelation rate varies based on characteristics of bookings

Analytical question: Who is likely to cancel their hotel bookings?

DATA

Hotel Booking Demands Dataset:

- July 2015 to August 2017
- 119,320 records
- 32 columns
- 2 types of hotels: City & Resort

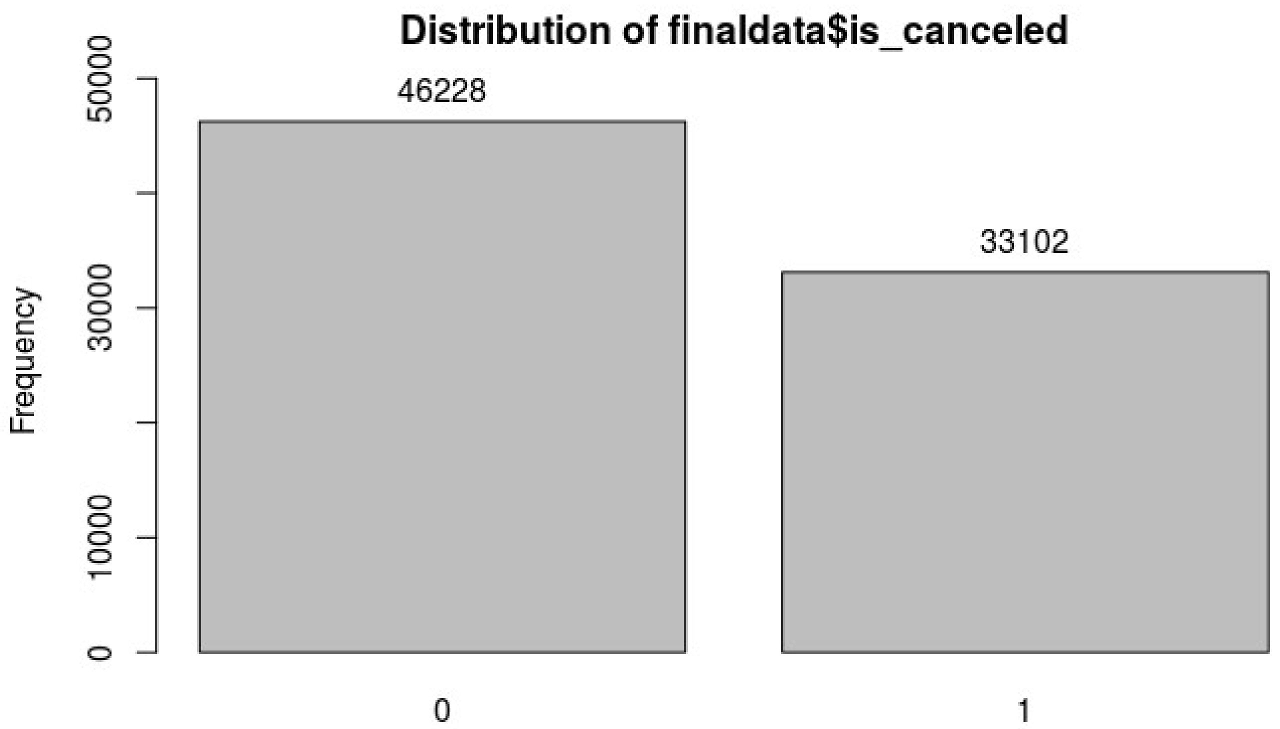
Final Data:

- City Hotel
- Eliminate redundant or irrelevant columns
- 79,330 records, 12 columns

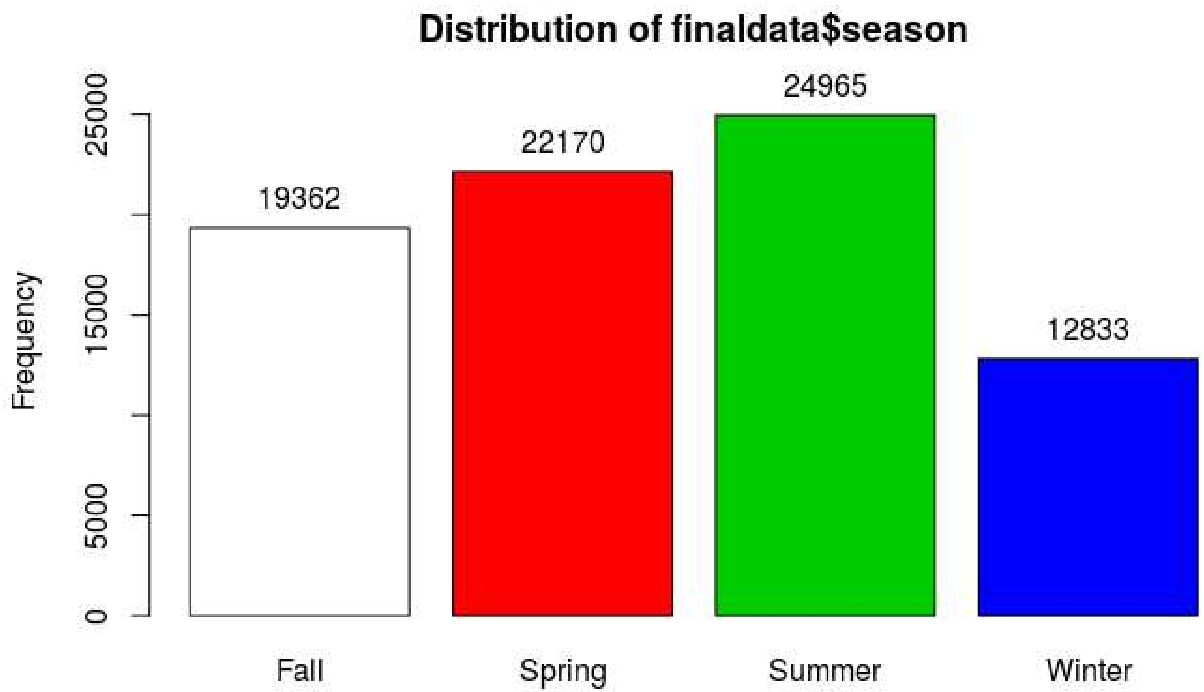
COLUMNS FINAL DATASET

Column Name	Type of Variable	Coding	Description
is_canceled	Numeric	Dummy Coded (0,1)	Target variable; whether booking was canceled
lead_time	Numeric	Continuous (numeric)	Days booked in advance of arrival
is_repeated_guest	Numeric	Dummy Coded (0,1)	Whether guest is a repeated guest
previous_cancelations	Numeric	Dummy Coded (0,1)	Whether there were any previous cancelations
booking_changes	Numeric	Dummy Coded (0,1)	Whether there were any booking changes
required_car_parking_spaces	Numeric	Dummy Coded (0,1)	Whether parking spaces were required
total_of_special_requests	Numeric	Dummy Coded (0,1)	Whether special requests were made
season	Character	4 Categories: Spring, Summer, Fall, Winter	Season of arrival
haskid	Numeric	Dummy Coded (0,1)	Whether booking has children or babies
countrygrp	Character	12 Categories: Portugal, France, Great Britain, Germany, Spain, Other Europe, Africa, Asia, Latin America, Middle Easy, North America, Other	Country guests are from
distribution	Character	4 Categories: Direct, Corporate, Travel Agent/Tour Operator, Other	Channel through which the booking was made
deposit	Numeric	Dummy Coded (0,1)	Whether deposit was made

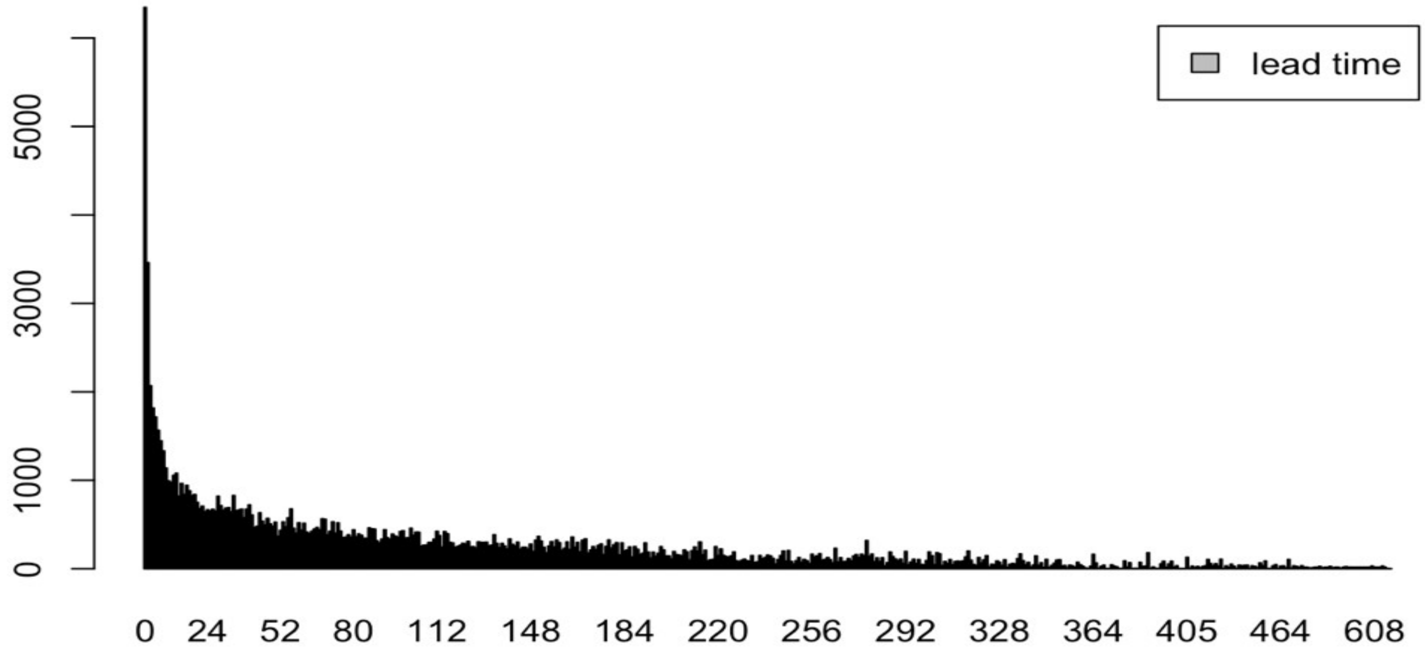
DATA EXPLORATION



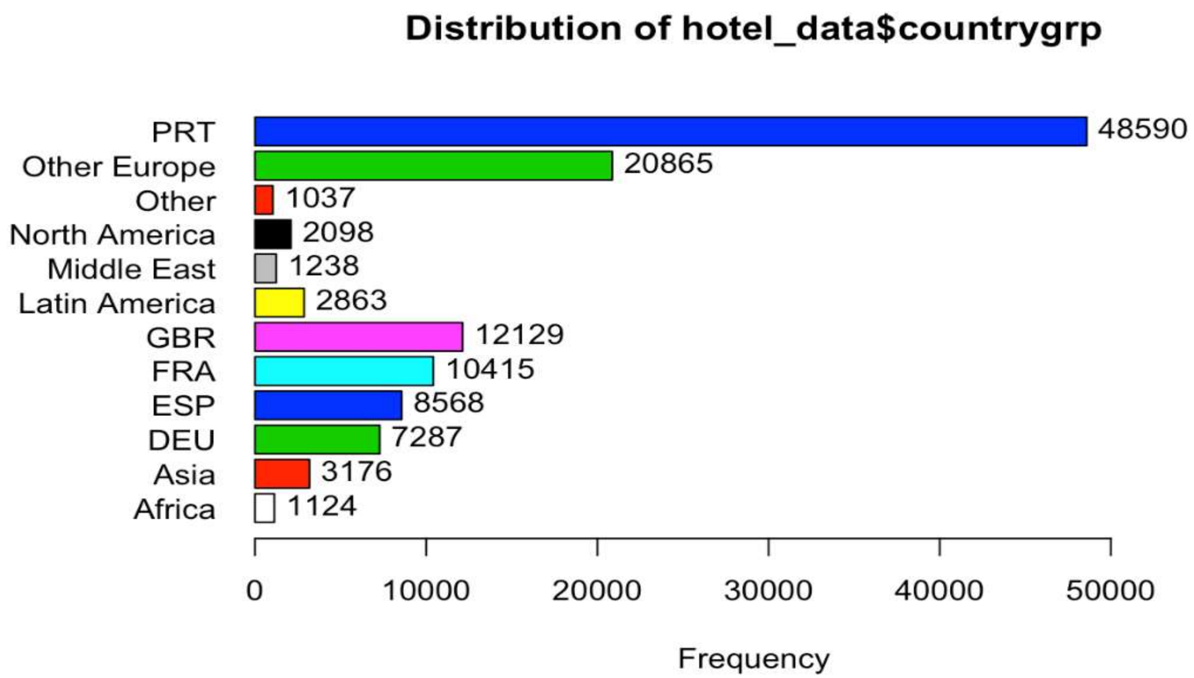
DATA EXPLORATION



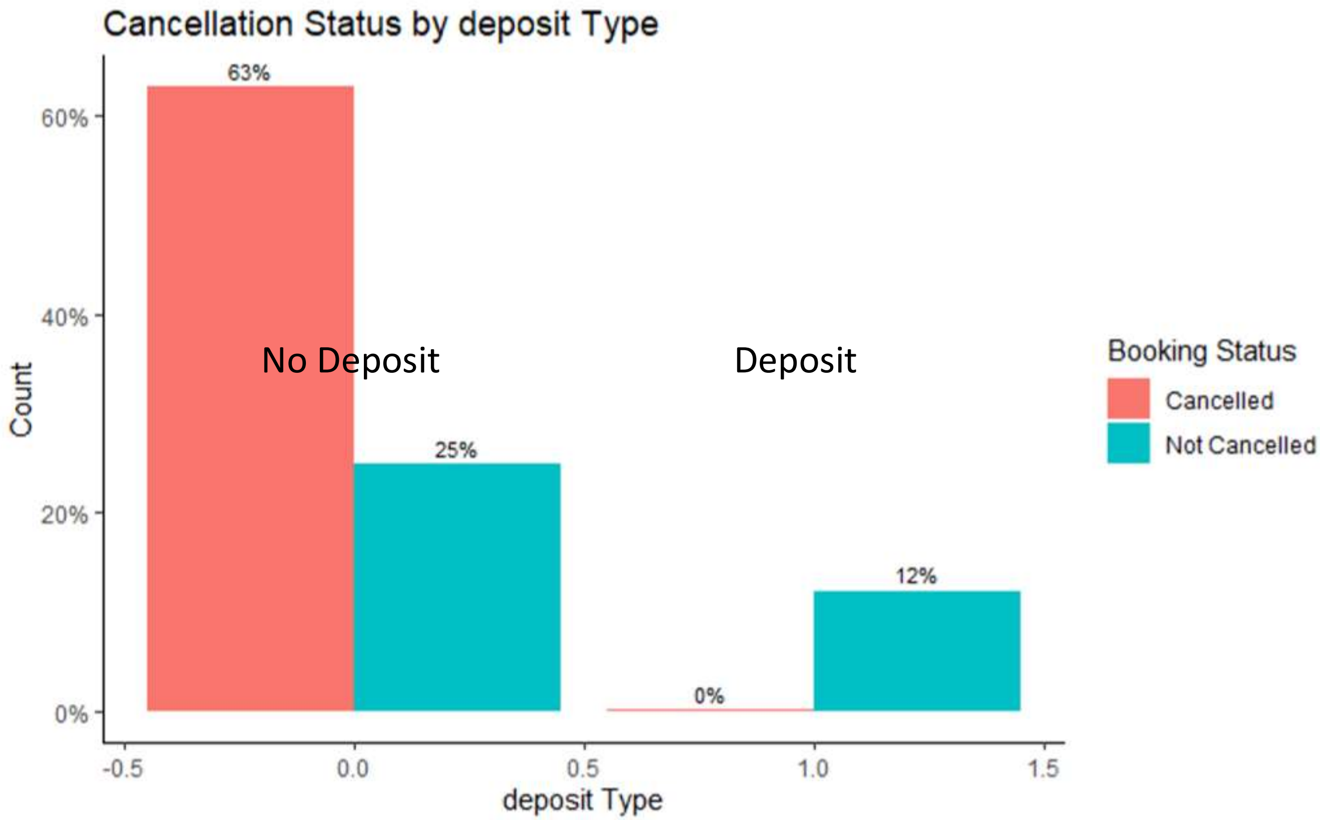
DATA EXPLORATION



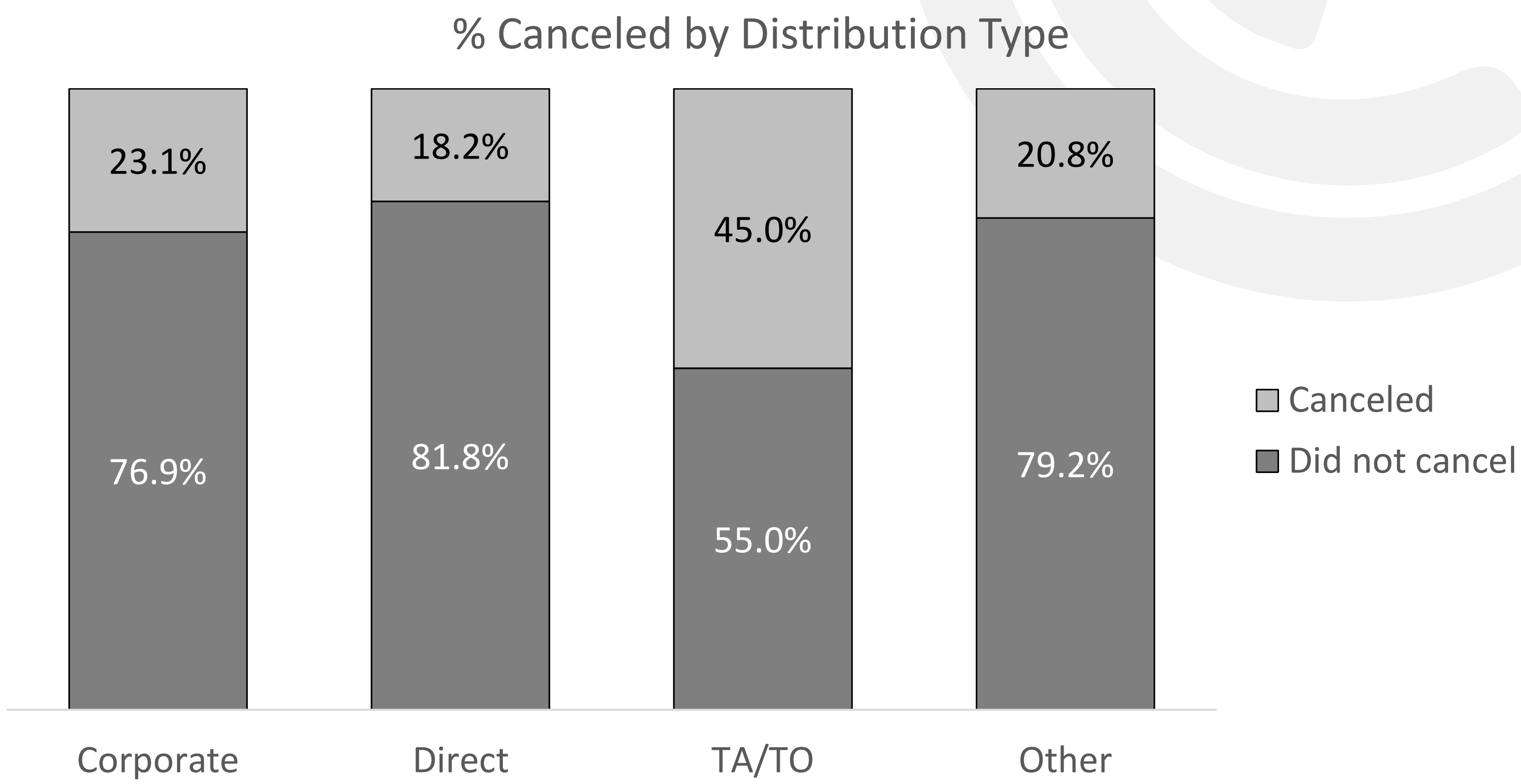
DATA EXPLORATION



DATA EXPLORATION



DATA EXPLORATION



TRAIN/TEST SPLIT

- Selected the attributes in our final columns shown in slide 6 of this presentation
- We split them into 80% training data and 20% testing data and check the distribution of the “is_canceled” attribute
- In situations where we have categorical variables (factors) but need to use them in analytical methods that require numbers (in our case is Logistic Regression), we need to create dummy variables. We did this for “season”, “countrygrp”, and “distribution”

LOGISTIC REGRESSION

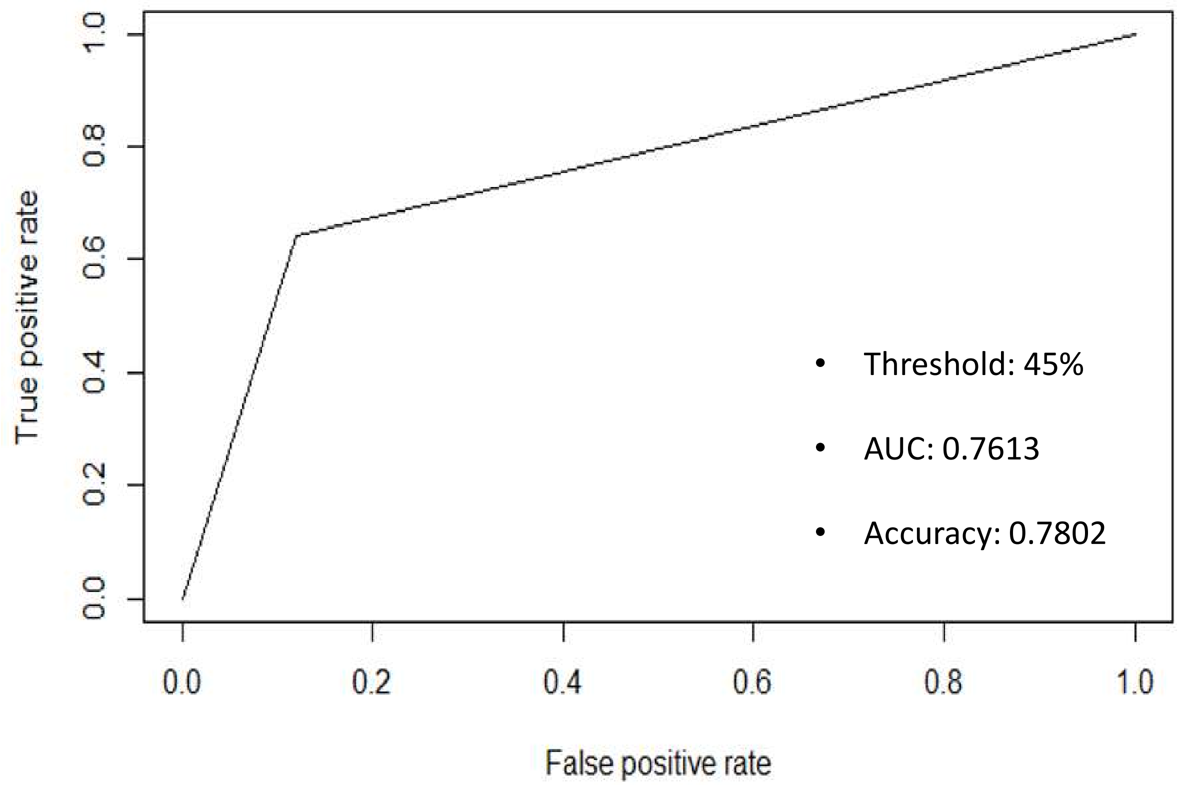
```
summary(mod_lr)
mod_lr <- glm(is_canceled ~ ., data = training2, family = "binomial")

Deviance Residuals:
    Min       1Q      Median     3Q      Max
-3.5266  -0.7725  -0.4509   0.1538   3.3992

Coefficients:
                               Estimate    Std. Error  z-value   Pr(>|z|)
(Intercept)                 -1.000e+00  1.057e-01  -9.463  < 2e-16  ***
lead_time                    2.870e-03  1.226e-04  23.421  < 2e-16  ***
Season                       4.352e-01  2.993e-02  14.543  < 2e-16  ***
Haskid                       5.476e-01  3.721e-02  14.719  < 2e-16  ***
Countrygrp                  -3.737e-01  9.429e-02  -3.963  7.39e-05 ***
Distribution                 1.122e+00  6.758e-02  16.597  < 2e-16  ***
is_repeated_guest           -2.530e+00  1.561e-01 -16.207  < 2e-16  ***
previous_cancellations       3.849e+00  1.463e-01  26.302  < 2e-16  ***
booking_changes             -1.063e+00  3.474e-02 -30.595  < 2e-16  ***
deposit_typeNon Refund       5.676e+00  2.202e-01  25.783  < 2e-16  ***
deposit_typeRefundable       2.035e+00  5.763e-01   3.532  0.000412 ***
car_parking_spaces          -1.527e+01  5.641e+01  -0.271  0.786655
total_of_special_requests   -8.645e-01  2.170e-02 -39.846  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 86196  on 63463  degrees of freedom
Residual deviance: 55658  on 63437  degrees of freedom
AIC: 55712
```

```
Prob_lm <- predict(object = mod_lr, newdata = testing2, type = "response")
Pred_lm <- ifelse(Prob_lm >= 0.45, "yes", "no")
```



- Threshold: 45%
- AUC: 0.7613
- Accuracy: 0.7802

DECISION TREE MODELING

Summary(Decision Tree)
mod_tree <- rpart(is_canceled ~ ., data = training, method = "class")

Variables actually used in tree construction:
[1] deposit_type previous_cancellations

Root node error: 26424/63464 = 0.41636

n= 63464

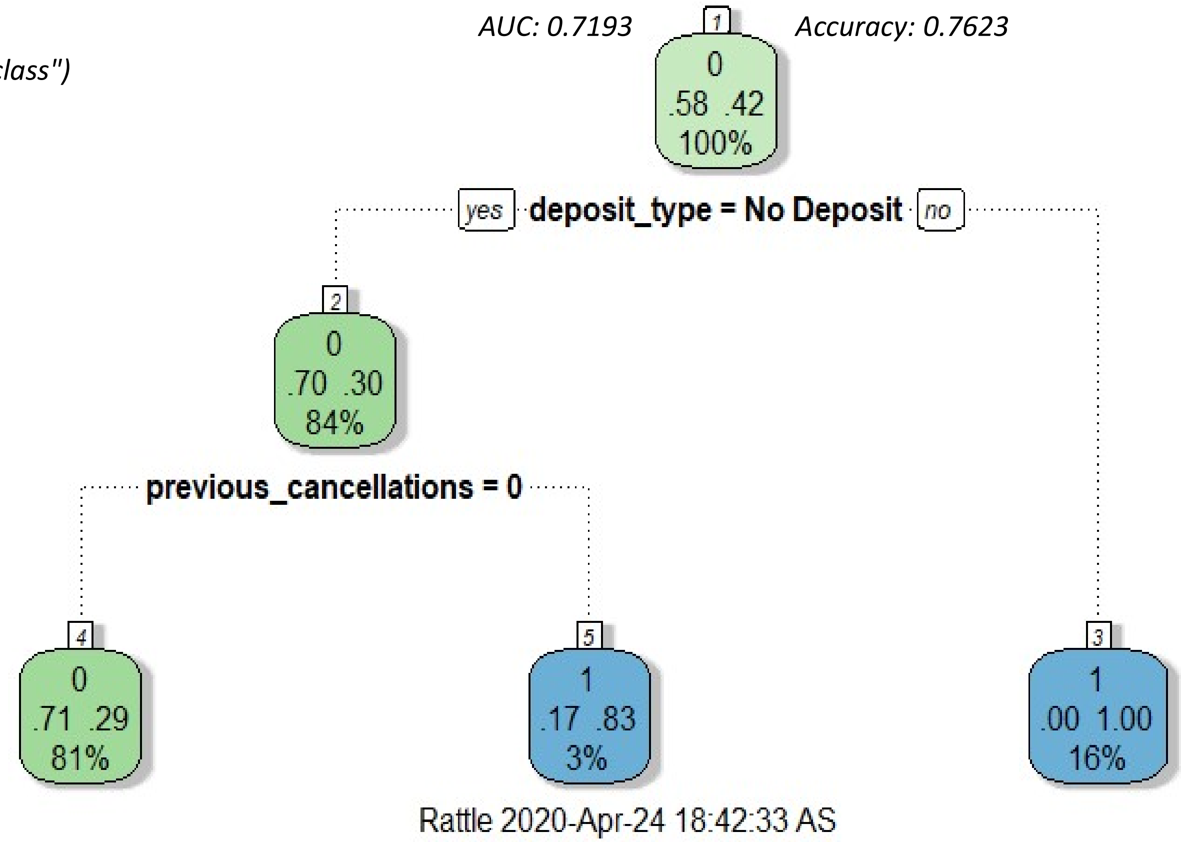
	CP	nsplit	rel error	xerror	xstd
1	0.38642900	0	1.0000000	1.0000000	0.004699725
2	0.04166667	1	0.6135710	0.6138359	0.004158498
3	0.01000000	2	0.5719043	0.5721692	0.004061393

Variable importance

deposit_type	previous_cancellations	lead_time
77	16	8

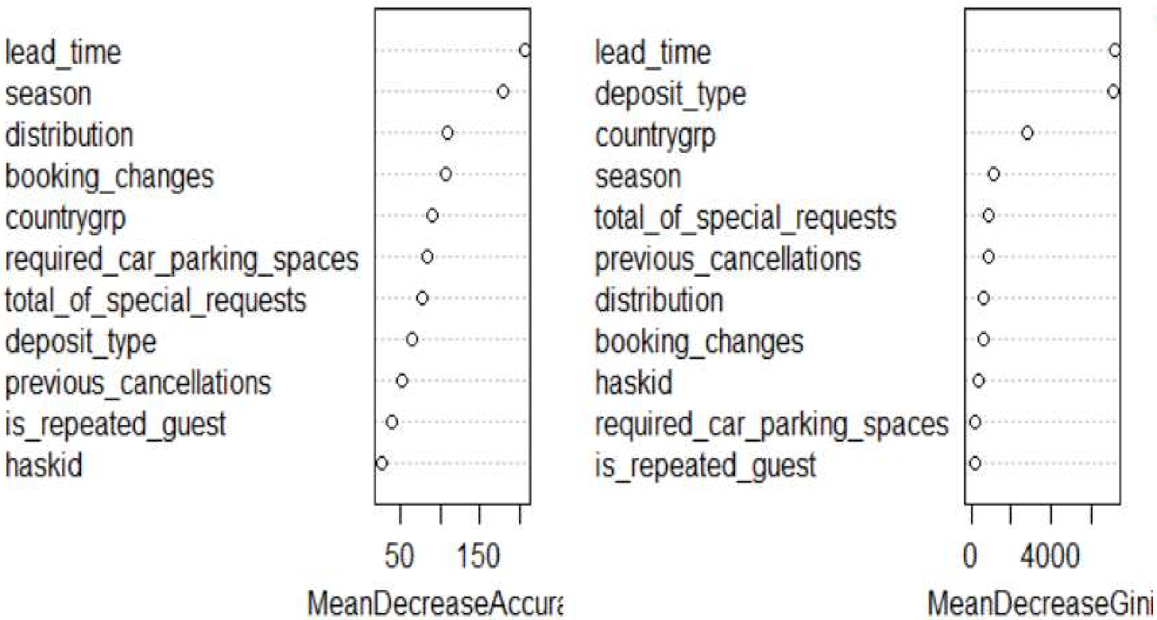
Pred_dt <- predict(mod_tree, testing, type = "class")

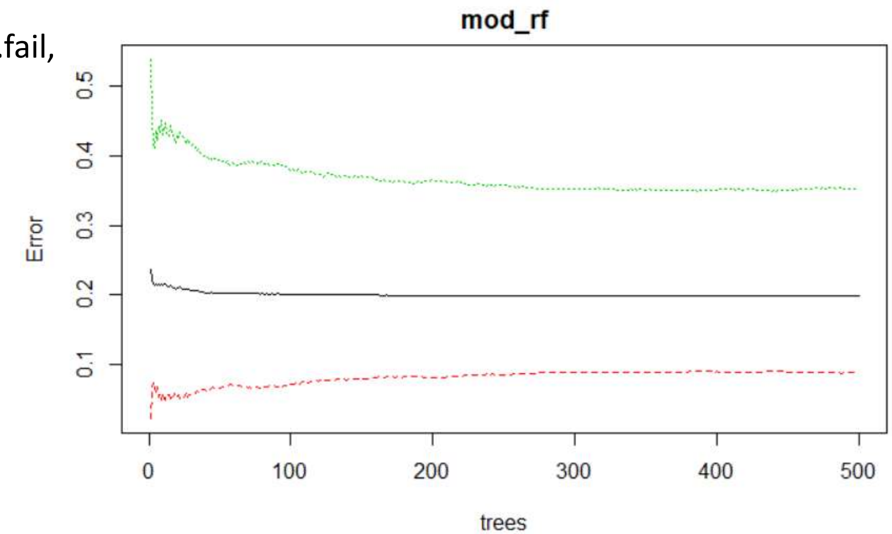
AUC: 0.7193 *Accuracy: 0.7623*



RANDOM FOREST MODELING

randomForest(is_canceled ~ ., method = "class", data = training, na.action = na.fail, nodesize = 1, importance = TRUE, ntree = 200, ntry = 8)



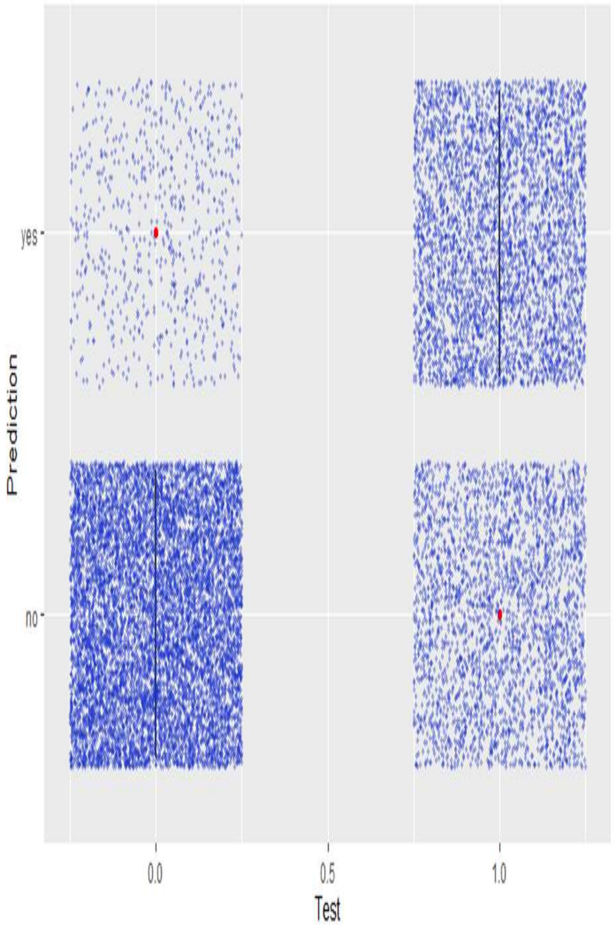


- AUC: 0.8094
- Accuracy: 0.8222

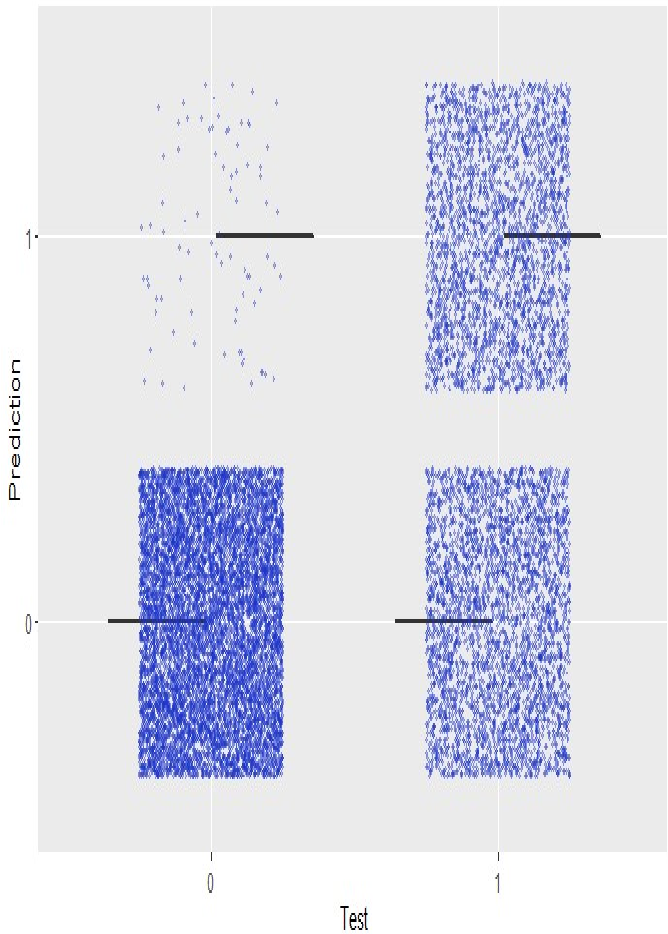
BEST MODEL
- Highest AUC
- Highest Accuracy

RESULT SUMMARY

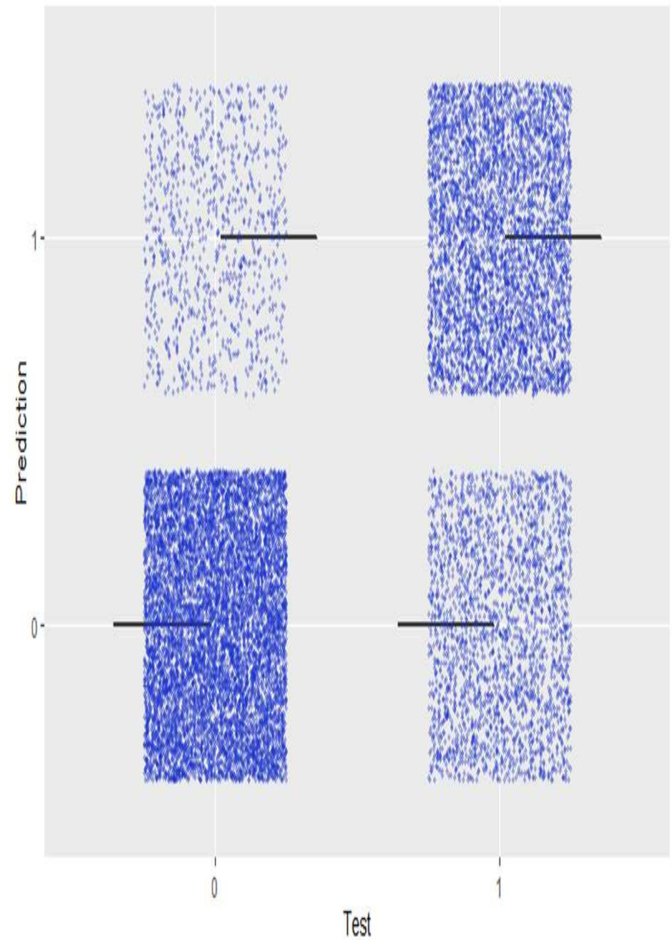
LOGISTIC REGRESSION



DECISION TREE



RANDOM FOREST



DEPLOYMENT AND RECOMMENDATIONS

- Feed booking data directly into model
  - Real-time predictions on cancelation rates
- Allow overbooking until hotel capacity is reached based on "no-cancel" predictions → then STOP overbooking
- Cancelation rates determined on case-by-case basis
  - Allows more accuracy and flexibility based on booking characteristics

THANK YOU