Hotel Booking

Group 8

Liang Yangkai	21500371
HUANG Jianxin	18251870
MAO Yitao	18251315
SONG Yijin	19251599
ZHANG Haolin	19251238

Table of contents

01

Business Understanding

Business Understanding

- **Goal:** to predict whether the customer will cancel the booking based on the model.

- **Business meaning:** By predicting the cancellation of customers, the hotel can arrange properly to reduce the vacancy rate.

- **Method:** We use the given variables, such as days_in_waiting list to build SVM mode, Tree model and Regression model. And by evaluating these three model, we can find the best model to predict the cancellation.

Here are some additional business meaning:

- **Off-season** and **peak season** of the hotel. Hotel can allocate resources based on this to reduce unnecessary expenses.

- According to the given article, the data also has significant meaning for **tourism management** and **education.**

02

Data Understanding & Preparation

Data understanding

Strength:

1. The data set is huge.
2. The data has various variables.

Cost: The data is provided by hotel.

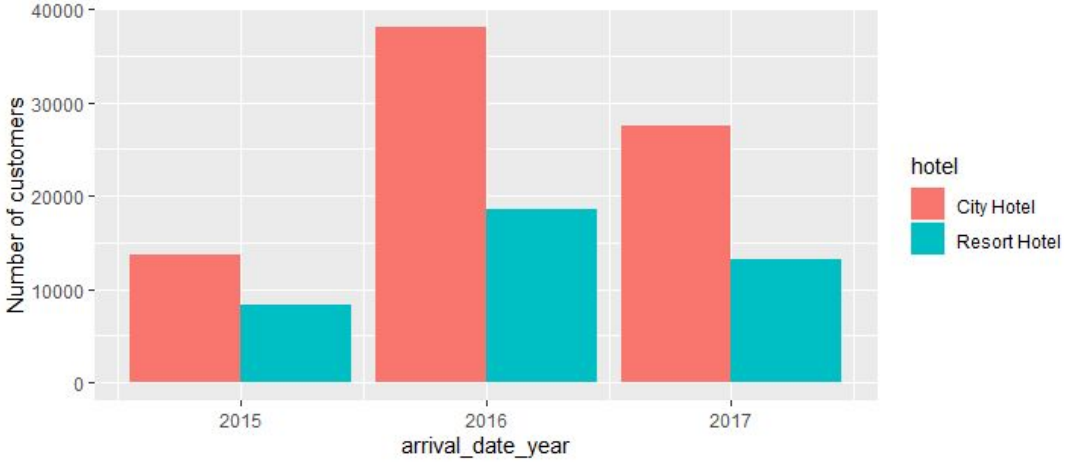
Limitation:

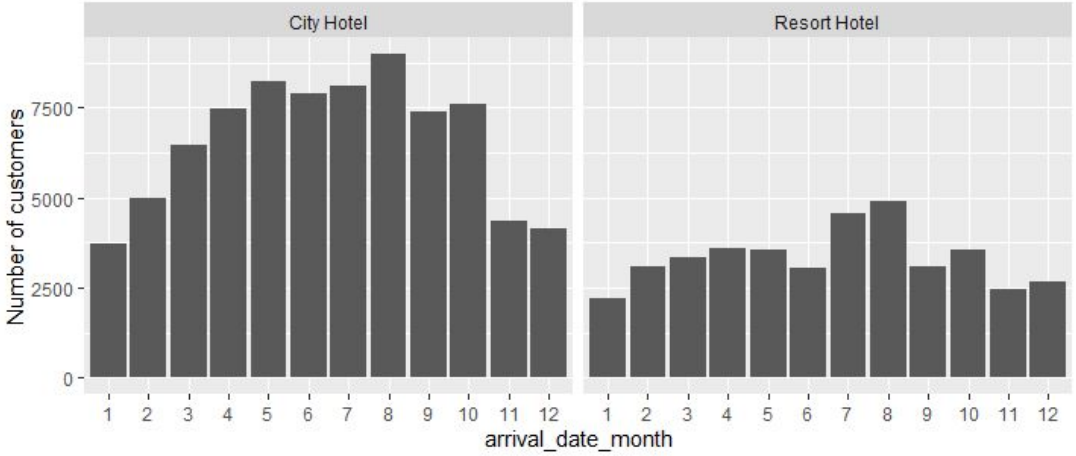
1. The data is customers real data, therefore, due to the safety reason, some data is missing.
2. People are cautious. Therefore, sometimes their decision is not related to the variables.
3. The cost of empty room is not given, therefore, we can not get the accurate profit curve.

Variable	Type	Description
is_canceled	Categorical	**1**: the booking was canceled; **0**: the booking was not canceled
lead_time	Interger	Number of days that elapsed between date of the booking and the arrival date
Adults	Interger	Number of adults
Children	Interger	Number of children
Babies	Interger	Number of babies
Meal	Categorical	Type of meal booked. **Undefined/SC**: no meal package; **BB**: Bed & Breakfast; **HB**: Half board (breakfast and one other meal – usually dinner); **FB**: Full board (breakfast, lunch and dinner)
MarketSegment	Categorical	Market segment designation. **TA**: Travel Agents; **TO**: Tour Operators
DistributionChannel	Categorical	Booking distribution channel. **TA**: Travel Agents; **TO**: Tour Operators
IsRepeatedGuest	Categorical	**1**: the booking name was from a repeated guest; **0**: the booking name was not from a repeated guest

Variable	Type	Description
PreviousCancellations	Interger	Number of previous bookings that were cancelled by the customer prior to the current booking
PreviousBookingsNotCanceled	Interger	Number of previous bookings not cancelled by the customer prior to the current booking
ReservedRoomType	Categorical	Code of room type reserved.
DepositType	Categorical	Indication on if the customer made a deposit to guarantee the booking. **No Deposit**: no deposit was made; **Non Refund**: a deposit was made in the value of the total stay cost; **Refundable**: a deposit was made with a value under the total stay
DaysInWaitingList	Interger	Number of days the booking was in the waiting list before it was confirmed to the customer
CustomerType	Categorical	Type of booking. **Contract**: when the booking has an allotment or other type of contract associated to it; **Group**: when the booking is associated to a group; **Transient**: when the booking is not part of a group or contract, and is not associated to other transient booking; **Transient-party**: when the booking is transient, but is associated to at least other transien
ADR	Numeric	Average Daily Rate
RequiredCardParkingSpaces	Interger	Number of car parking spaces required by the customer

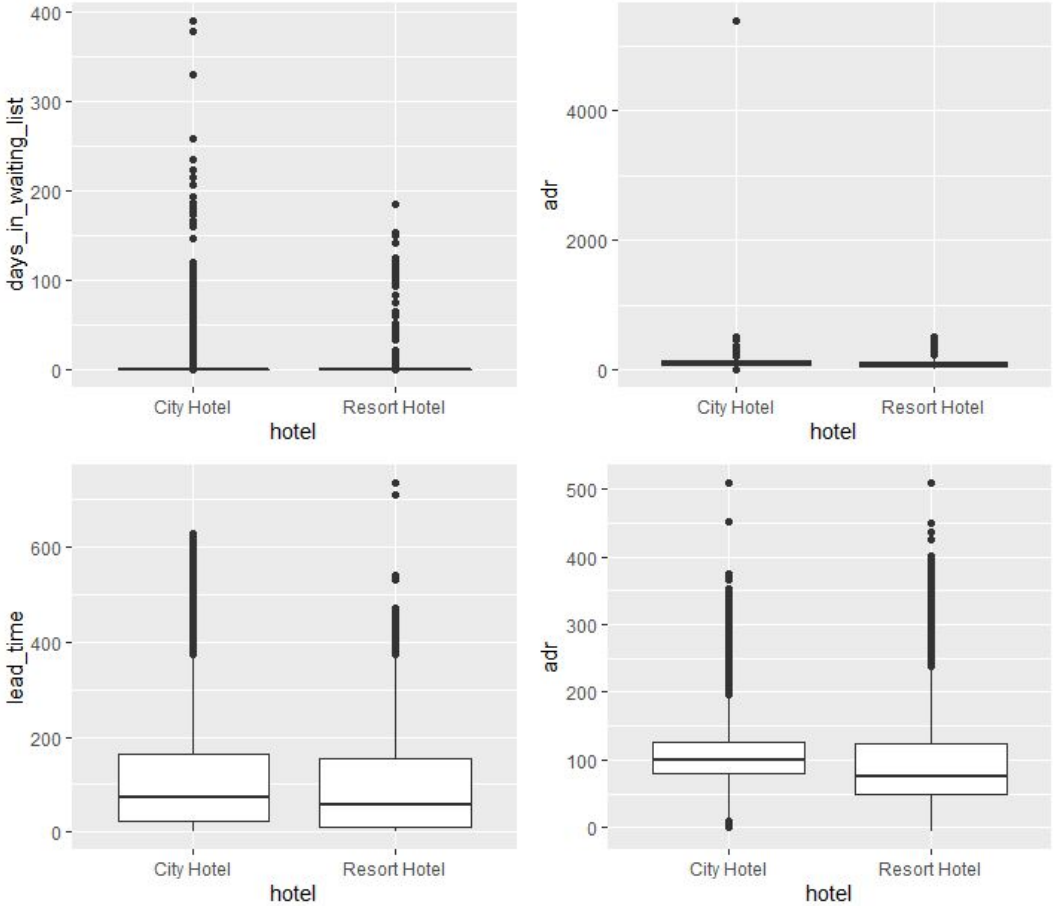
Number of customers among three years





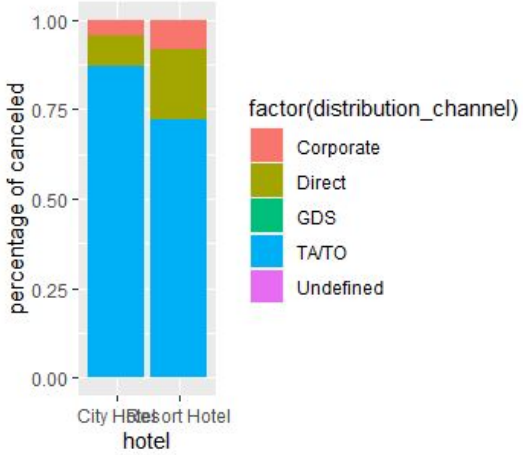
- City Hotel has a larger volume.
- In 2016, both hotels experienced an explosive growth, and also both declined slightly in 2017.
- July and August are hot months for both hotels. City Hotels perform better in the summer.

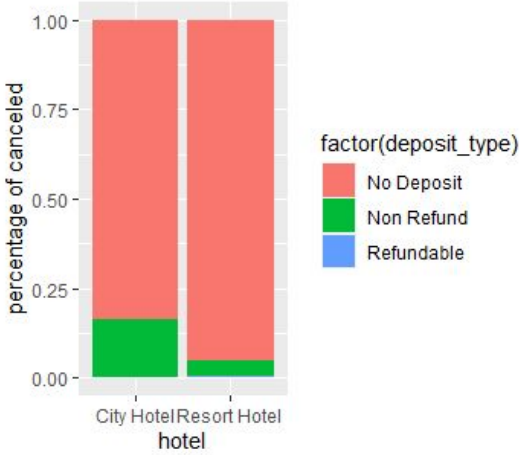
Variables between the hotels



- Most customers at both hotels need not to wait on waiting lists.
- The advance booking time of customers at the two hotels is basically equal.
- The average profit at city hotel is slightly higher.

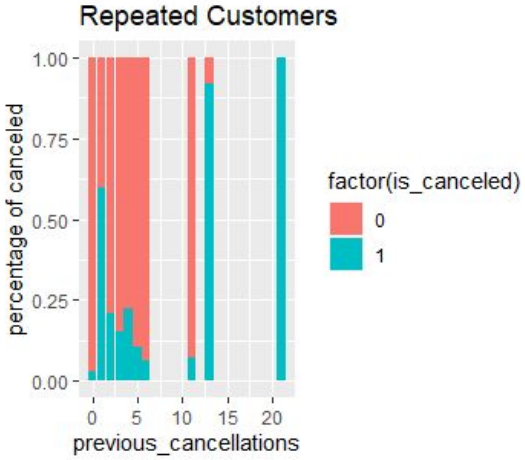
Categorical variables of hotels





- For market segment and distribution channel, travel agents and tour operators, including online and offline, take the most proportion.
- Most customers are transient in both hotels.
- No deposit orders account for more than 75% in both hotels, but the proportion of non refund orders in city hotel is higher.

Relationship between cancels and repeat customers



- The percentage of canceled of City Hotel is higher, reaching nearly 45%.
- The proportion of repeated customers of both hotels are not high.
- The repeated customers have less probability to cancel the order.

Menu

03

Modeling

Outline of SVM

- Statement of SVM
- SVM Modeling
- Modeling performance
  - ROC Curve & AUC
  - CRC Curve
  - Lift Curve
- Summary of SVM

Statement of the SVM

1.Data

- Train data 21490 (18% of the total data)
- Test data 11939 (10% of the total data)

The reason of decreasing of scale of train data

Because kernel Matrix describes the similarity between samples using the kernel Matrix of the dataset, the number of matrix elements increases squared as the size of the data increases. This makes SVM computing unprocessable as the size of the data increases.

Statement of the SVM

2. kernel

- Linear
- polynomial
- radial
- sigmoid

- Various classification variables
- Large amount of training data sufficient to fit complex nonlinear models
- Nonlinear models fit better in most of the situations.

kernel = radial

SVM Modeling

Tuna

Cost: 0.01, 5, 10, 100

Best parameters: cost=100
best generalization acc on val: 78.69%

```
Parameter tuning of 'svm':

- sampling method: 10-fold cross validation

- best parameters:
 cost
  100

- best performance: 0.2141319

- Detailed performance results:
   cost     error  dispersion
1 1e-02 0.2547213 0.010148955
2 5e+00 0.2173483 0.010965931
3 1e+01 0.2163617 0.009501884
4 1e+02 0.2141319 0.009730320
```

SVM Modeling

Summary of the best_mod:

- SVM-Type: C-classification
- Kennel: radial
- Cost: 100
- SV: 8832

```
Call:
best.tune(method = svm, train.x = is_canceled ~ ., data = train_data, ranges = list(cost = c(0.01,
    5, 10, 100)), probability = TRUE)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  100

Number of Support Vectors:  8832

 ( 4561 4271 )


Number of Classes:  2

Levels:
 0 1
```

Modeling Performance

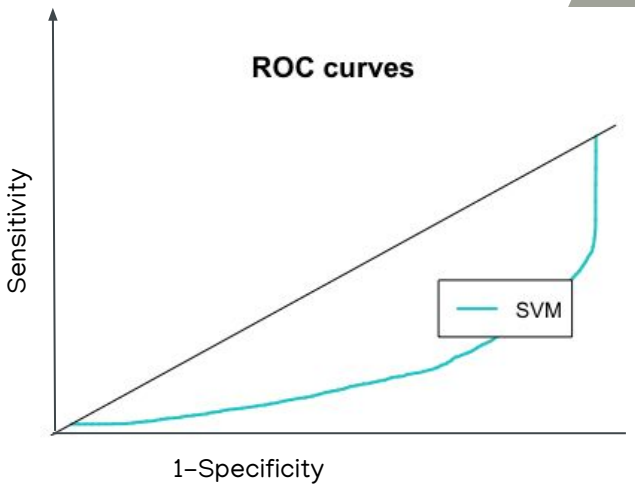
- Accuracy 0.2128319
- Sensitivity 0.4860858
- Specificity 0.0542687

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0  410 2253
         1 7145 2131

               Accuracy : 0.2128
                 95% CI : (0.2055, 0.2203)
    No Information Rate : 0.6328
    P-Value [Acc > NIR] : 1

                  Kappa : -0.3724

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.48609
            Specificity : 0.05427
         Pos Pred Value : 0.22973
         Neg Pred Value : 0.15396
             Prevalence : 0.36720
         Detection Rate : 0.17849
   Detection Prevalence : 0.77695
      Balanced Accuracy : 0.27018

       'Positive' Class : 1
```

Modeling Performance

- ROC & AUC

By using randomness as a reference for comparison, we find that the SVM model fits worse than randomness, and also its AUC(is the area under ROC) value is less than 0.5 (the AUC value of randomness).



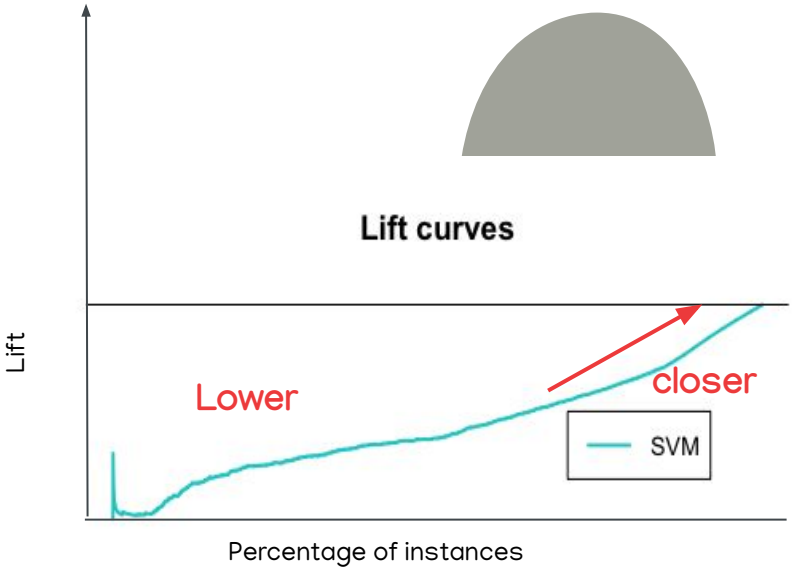

```
> auc_svm@y.values
[[1]]
[1] 0.1613585
```

Modeling Performance

- **Lift**

The lift represents the advantage a model provides over random guessing.

Which means the model is worse than random classifier in the beginning and in the end it is still keep the same level as random classifier and do not show any advantage.



Summary of SVM

Performance

All of the curve ROC, AUC, CRC, Lift show the bad performance

Modeling

Calculate delay
Low Accuracy

Data

Scale limitation

SVM model is not very suitable for application to this problem(hotel booking)

Outline of Logit regression

1. Model training & outcome of model on training set
2. Outcome on test set
3. Graphs explained

Model training & outcome on training set

```
m1 = glm(is_canceled~lead_time+adults+children+babies+meal+market_segment+distribution_channel+
         is_repeated_guest+previous_cancellations+previous_bookings_not_canceled+reserved_room_type+
         deposit_type+days_in_waiting_list+customer_type+adr+required_car_parking_spaces,
       data=train,
       family=binomial(link='logit')
)
```

General accuracy: 66.14%

Outcome on test set

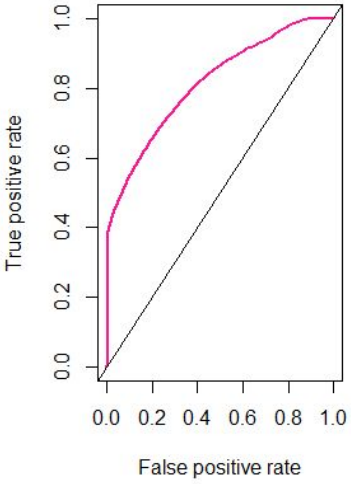
```
          Reference
Prediction     0     1
         0 22464  8217
         1    91  5045

               Accuracy : 0.768
```

Sensitivity:0.3804
Specificity:0.9960

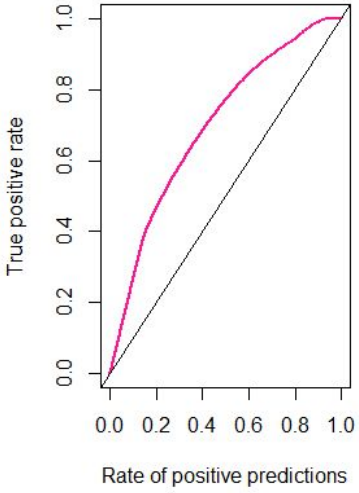
0 refers to negative(not cancel)
1 refers to positive(cancel)

1. General accuracy: 76.8%
2. More accurate predicting actual negatives(less type 1 errors),high sensitivity
3. Less accurate predicting actual positives(more type 2 errors), low specificity

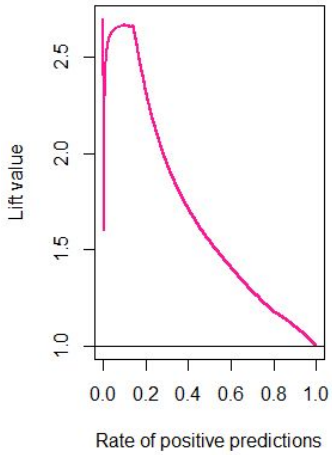
Curves explained



Curve above random classifier(Performance better than random classifier)
AUC:0.814



Curve above random classifier

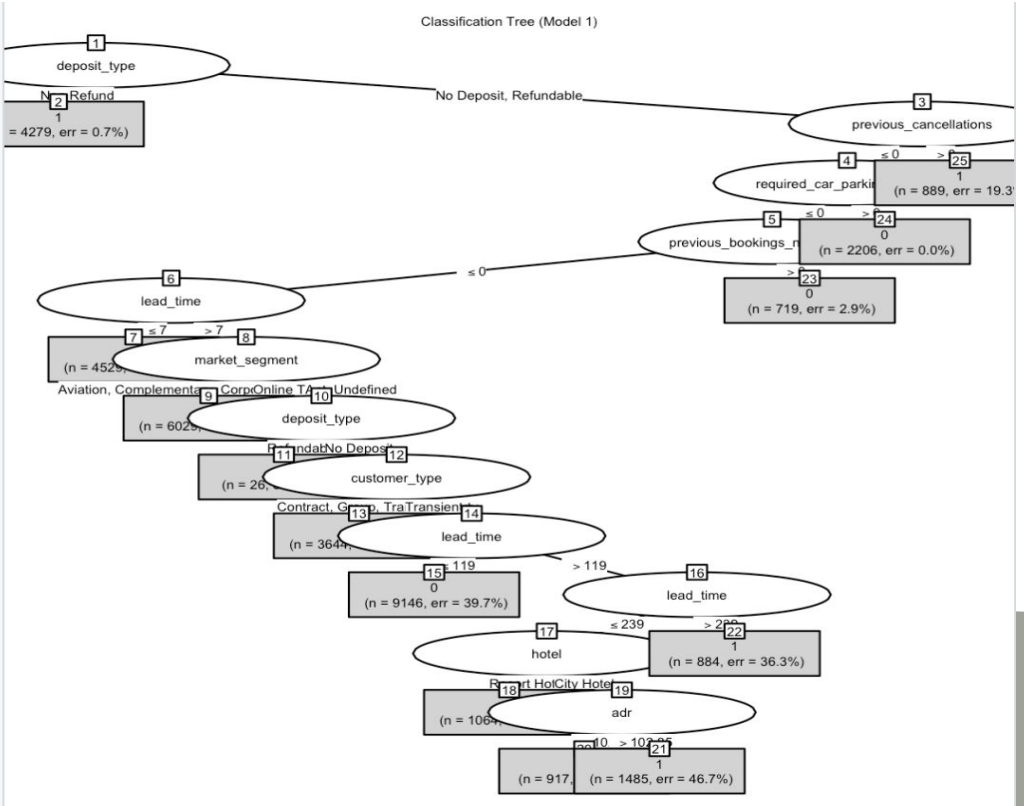


Curve above random classifier

Establish Tree model——full tree

att1 == All variable in training data set except is_canceled

```
mod1 = C5.0(x=training_2[,att1], y=training_2$is_canceled,
            control=C5.0Control(minCases=500))
```



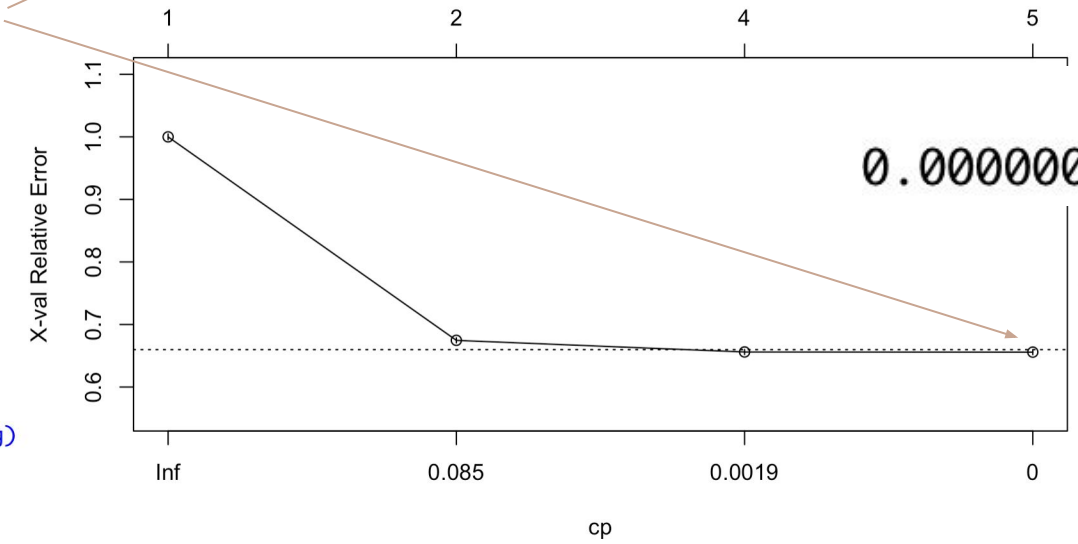
Prune Tree to avoid overfitting

Regardless of cost in order to get Maximum predict probability

	CP	nsplit	rel error	xerror	xstd
1	0.32530938	0	1.00000	1.00000	0.0045106
2	0.02200394	1	0.67469	0.67469	0.0040439
3	0.00016156	3	0.63068	0.65608	0.0040060
4	0.00000000	4	0.63052	0.65582	0.0040055

size of tree

The less Complexity, the less cross error, the more branches in model



CP	nsplit
0.000000000	4.000000000

	actual	
cted	0	1
No	22523	8284
Yes	20	4989

Accuracy:

```
> (22486+4964)/nrow(testing)
[1] 0.7663958
```

Compare with logistic regression

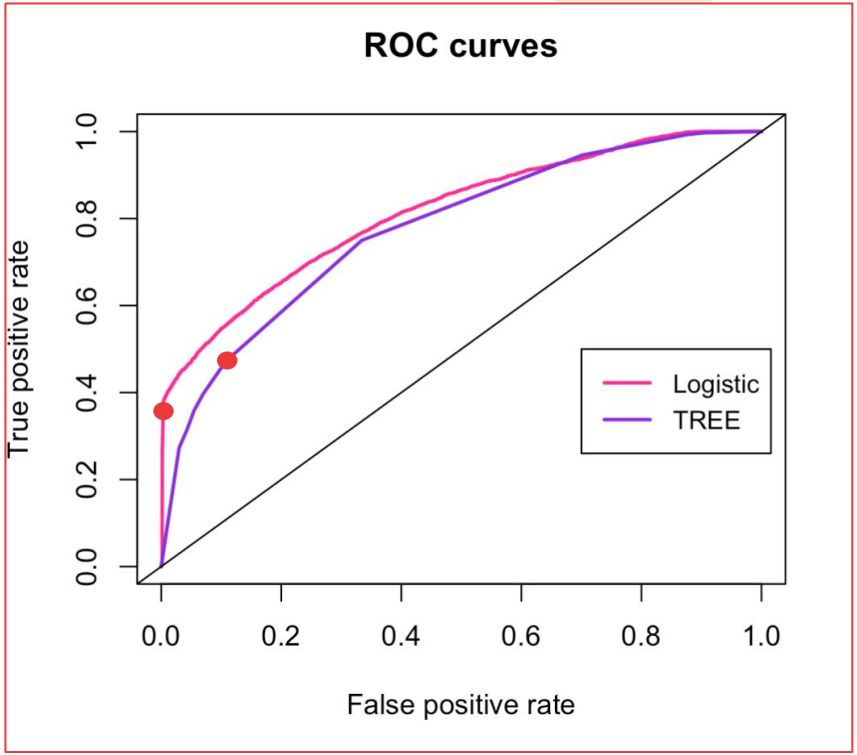
Goal of Modeling: Prevent Empty room when unpredict canceled order happened.

Model Criteria: Prediction of False Negative as small as possible

Statistic angle : Sensitivity as high as posible(1−Sensitivity = False Negative rate)

Logistic Model
Sensitivity:0.3804
Specificity:0.9960

Tree Model
Sensitivity:0.5151
Specificity:0.9339



Summary of SVM and Logit, Tree model

1. Logit and Tree model Performance are better than random classifier, SVM out.

2. Tree model performs better in reducing FN rate.

04
Evaluation & Deployment

Final Model

Sensitivity	
Logistic Regression	0.3804
Tree	0.515
SVM	Out of table

When we choose the final model:

Our goal is to avoid false negative. That means the customers was predicted to not cancel the booking. However, the customer cancel it in the last. Therefore, we need to maximize true positive. That means, we need to compare sensitivity of three model. Larger sensitivity lead to better avoiding false negative. Therefore, we should choose the Tree model.

Evaluation

Assess the result: The tree model has an accuracy of 0.778. The model can predict the data. It is valid.

Support decision making: It can predict the customer decision and lower the vacancy rate,therefore it can maximize the profit.

The cost of false alarm:

(1) The customers may be not cancelled, but the hotel predict is cancel. That may cause some customers may not able to stay in hotel. That may cause the lower evaluation of the hotel.

(2) The hotel would lose potential profit if the customers who were predicted to not to cancel the bookings actually did, as this type of error would increase vacancy rate, therefore inducing losses.

Deployment

When the hotel rooms are not enough for all customers,the hotel can use this model to predict whether the customers will cancel or not. If a customer has a high chance to cancel the booking, the hotel can arrange other customers to the waiting list instead of reject them.

Thanks!