Cornell University

ENMGT 5930 Project One Report

Using PCA and K-Means for Hotel Customers Segmentation

Yu Pan (yp425)
Yajing Wang (yw832)
Xinyue Zhao (xz332)
Ziyin Wang (zw252)

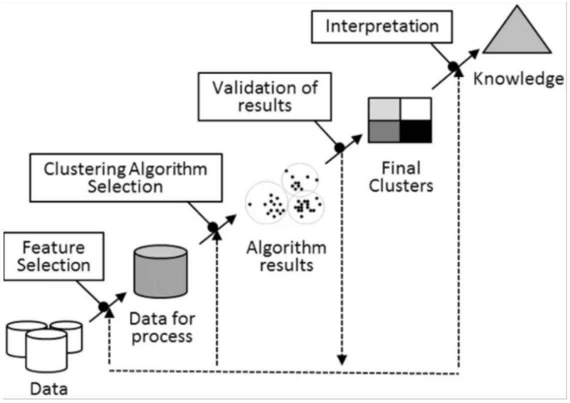
11/5/2021

Problem Statement

The manager of a city hotel would like to plan out a marketing strategy to target different customer segmentations for the next year 2022 to increase the revenue.

The dataset of hotel booking demand and the tools we learned in this course listed below will be used to extract information and insights for making suggestions and recommendations for the hotel manager.

Project Procedure



Data -> Information -> Insights -> Decisions -> Actions

Data Cleaning and Variable Selection

In this project, the major methods we used are PCA and K-means. The main idea of PCA is to map n-dimensional features to k-dimensions. This k-dimension is a brand-new orthogonal feature, also called principal component, which is a k-dimensional feature reconstructed on the basis of the original n-dimensional feature. The job of PCA is to sequentially find a set of mutually orthogonal coordinate axes from the original space. The choice of new coordinate axes is closely related to the data itself. The first new coordinate axis selection is the direction with the largest variance in the original data, the second new coordinate axis selection is the plane orthogonal to the first coordinate axis that maximizes the variance, and the third axis is related to the first and second axes. The plane with the two axes orthogonal to the plane has the largest variance. By analogy, n such coordinate axes can be obtained. Therefore, the essence of the PCA method is the analysis of data variance. So, when we used the PCA method, only numerical variables remained.

Those numerical variables include booking_id,lead_time,stays_in_weekend_nights, stays_in_week_nights,adults,children,babies,previous_cancellations,previous_bookings_not_canceled, booking_changes, adr, required_car_parking_spaces, and total_of_special_requests.

The reason why we select these variables is not only because they are numerical variables, but also because these variables can meet our objectives for analyzing. For example, the variable stays_in_weekend_nights means that the number of weekends during the stay, and by this variable we can estimate if this guest book the hotel for vacation. And for stays_in_week_nights, we can estimate if this guest was businesspeople. For the variable babies, we can reduce the number of family tours. For booking_changes, it represents the revising times from booking to check in, and it can also reveal if this guest is capricious. Like adr, it means the daily average prices of rooms, it can reveal the cost level of the guests. And for required_car_parking_spaces, we can conclude if the guests have cars, or how many. Based on these variables, we can give some suggestions to hotel for different kinds of guests.

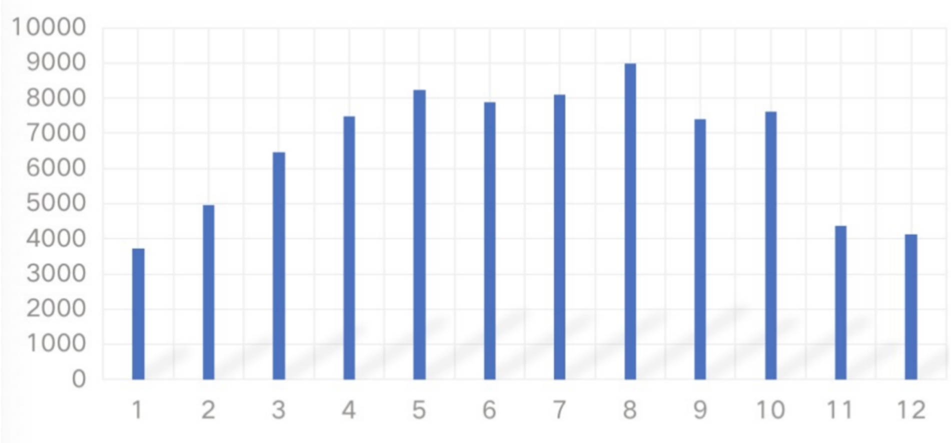
For the categorical variables, such as is_cancelled, arrival_date_year, arrival_date_month, arrival_date_week_number, arrival_date_day_of_month, meal, country, market_segement, we decided to analyze them separately, which meant that we could use EDA instead. Exploratory data analysis (EDA) is an investigation process. We can use summary statistics and graphical tools to analyze data and discover patterns in data, understand potential relationships between variables or find abnormal phenomena (such as outliers or anomalies observation). And we can also check the average arrival date after classified the numerical variables into different groups. Then, for different arrival dates of different groups, we can offer suggestions about marketing strategies. This part will be discussed in detail later.

Data Exploration and Normalization

We conduct some data exploration analysis on the above continuous variables and take several representative attributes as follows.

1. arrival_date_month

Arrival_date_month refers to the month of arrival date, we analyze the frequency distribution of the hotel booking amount in different months through the following histogram.



The top four months with the most booking orders are from May to August, while it has small number of orders from November to February, which indicates that people tend to go for traveling in summer rather than winter, thus a strategy could be made on the price of hotel based on the potential number of guests.

2. adr

Adr refers to the Average Daily Rate, which is derived from dividing the sum of all lodging transactions by the total number of nights staying.

The minimum value of adr is 0, we have 1208 entries that have the 0 adr value, we may consider the zero rate as a cancellation or a bonus of the previous booking records. The maximum value of adr is 5400, since the second largest value is 510, we consider the 5400 entry as an outlier and delete it. The mean value is 105.3044654, the median value is 99.9.

3. stays_in_weekend_nights, stays_in_week_nights

Stays_in_weekend_nights refers to the number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel. Stays_in_week_nights refers to the number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel. Adding these two variables, we could obtain the total number of days.

The minimum value of stays_in_weekend_nights is 0, it indicates these guests stay in the hotel on weekdays, they are more likely to come for business. The maximum value of stays_in_weekend_nights is 16, it means the guest stayed in the hotel for almost 8 weeks, this hotel could have better service and be recommended to other guests. The mean value is 0.795184672, the median value is 1, so most people take on a trip for less than one week.

The minimum value of stays_in_week_nights is 0, it indicates these guests stay in the hotel on weekends, they are more likely to come here to travel. The maximum value of stays_in_week_nights is 41, it means the guest stayed in the hotel for almost 8 weeks, this value fits the maximum value in stays_in_weekend_nights. The mean value is 2.182957267, the median value is 2, so most people stayed in the hotel for about 2 days.

4. adults, children, babies

For those guests who bring children or babies and inform the hotel ahead of time, the hotel will offer special service for them. First look at the number of adults, the value ranges from 0 to 4, the mean value is 1.850976932, the median value is 2, so in most of cases, 2 adults book the hotel together.

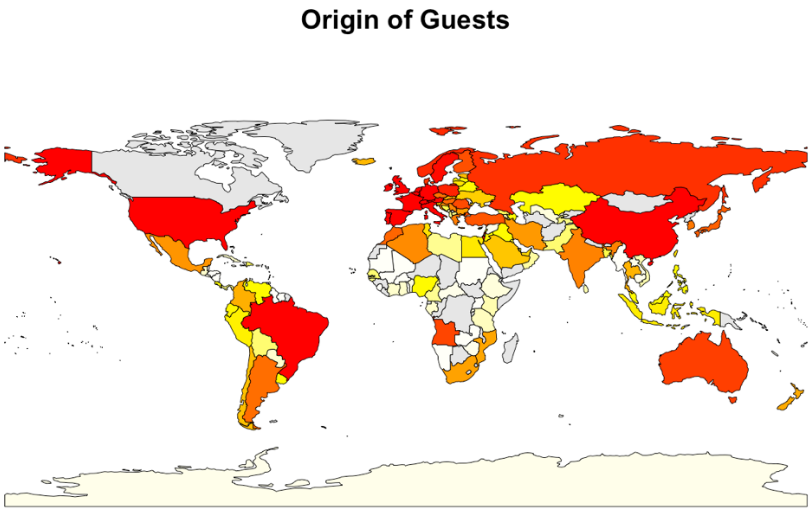
As for the number of children, the value ranges from 0 to 3, it makes sense for a family travel. To be specific, there are 4 entries that do not have this field, and 59 entries of 3 children, 11 of them are not accompanied by an adult, the hotel should have the responsibility to guarantee the safety of children. Besides, there are 2024 entries of 2 children, 3023 entries of 1 child, the remaining are 0 children. The mean value is 0.09136979, most of the guests do not bring children.

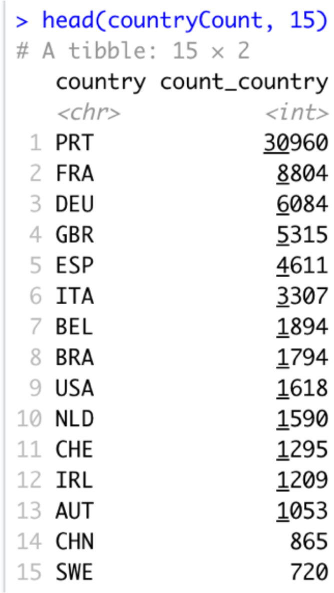
As for the number of babies, the largest value is 10 with 2 adults, the second largest value is 9 with 1 adult, and these two entries would be special cases. Besides, there are 6 entries of 2 babies, all of them are with 2 adults, 361 entries of 1 baby, the remaining are 0 baby. The mean value is 0.004941384 that approximates to 0, so most guests do not bring babies.

Data Visualization and Patterns

I. Categorical Data Analysis

a. Geospatial Data Analysis





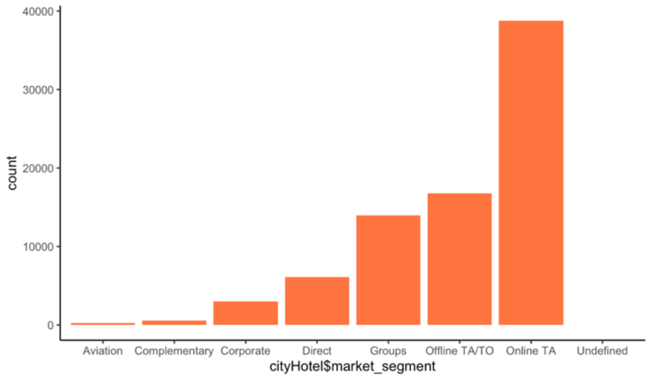
```
> head(countryCount, 15)
# A tibble: 15 × 2
   country count_country
   <chr>           <int>
 1 PRT             30960
 2 FRA              8804
 3 DEU              6084
 4 GBR              5315
 5 ESP              4611
 6 ITA              3307
 7 BEL              1894
 8 BRA              1794
 9 USA              1618
10 NLD              1590
11 CHE              1295
12 IRL              1209
13 AUT              1053
14 CHN               865
15 SWE               720
```

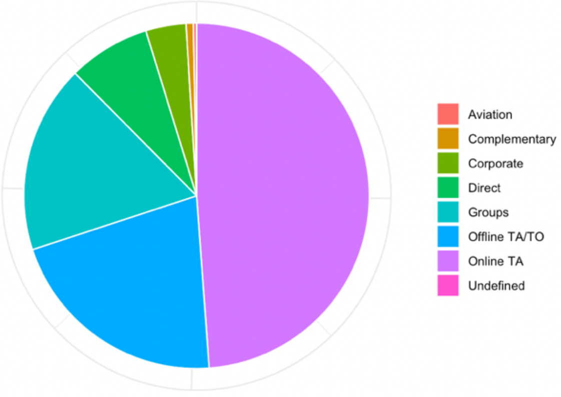
The above graph indicates where the guests of the city hotel are coming from. The color scale from red, orange, yellow, and all the way to white and grey represents the level of number of guests from each country. The darker red indicates a high volume of guests, whereas the grey means that there are no guests from that area.

The list on the right provides the top 15 countries that the guests are from. They are Portugal, France, Germany, United Kingdom, Spain, Italy, Belgium, Brazil, United States, Netherlands, Switzerland, Ireland, Austria, China, and Sweden. It is not hard to tell that the majority of the guests are from European Union with additions of the States and China.

To gain a high retention rate of existing guests, the city hotel should continue to focus on marketing the EU area. There is more potential for bringing in more customers from the United States, China, Brazil, Australia and any other areas with orange, yellow, and light-yellow color. New markets can be explored or targeted for the white and grey countries, including Canada, Congo, Chad, Afghan, Mongolia, Greenland, etc. A creative marketing strategy for these specific countries may be required to satisfy the real need of guests of these areas.

b. Market Segment Analysis

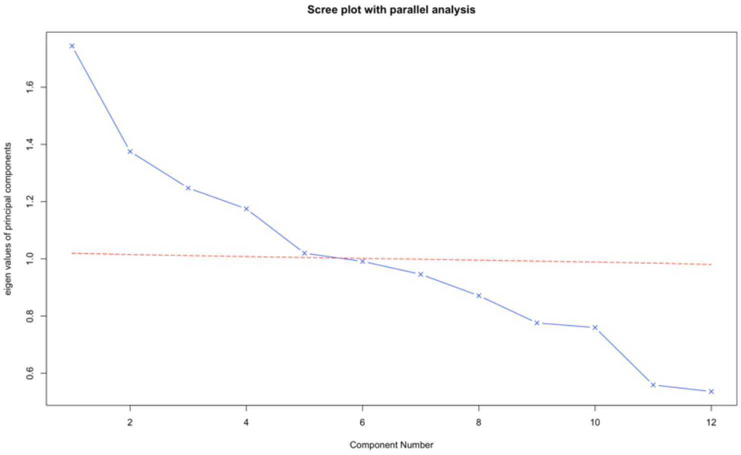


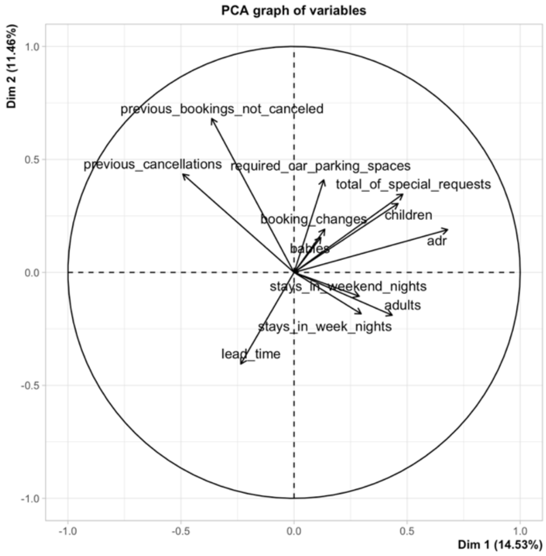


From the histogram and the pie chart we can tell that approximately half of all guests book the city hotel through online travel agencies, and about 20% of them order from offline travel agencies or travel operators. The figures also illustrate that the direct booking with the city hotel is not many, which means that the city hotel should cooperate and champaign more with online and offline travel agencies and operators to sell its rooms. A large amount of groups and corporate bookings may also imply that collaborating with companies to accommodate for more business trips can help with revenue growth.

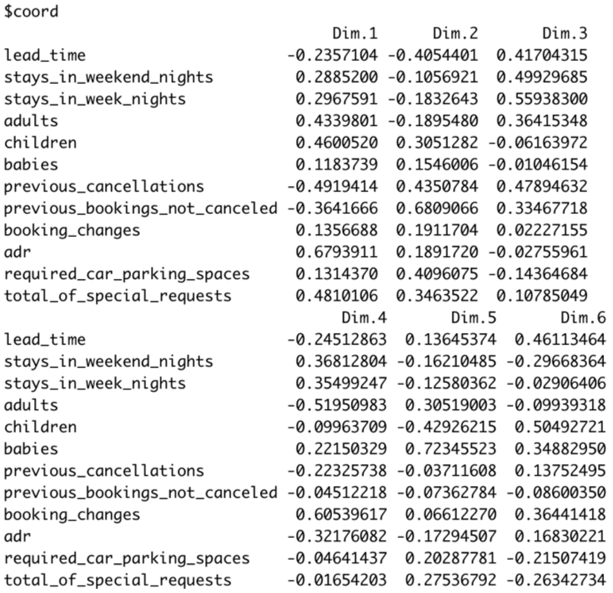
II. Numeric Data Analysis

a. PCA

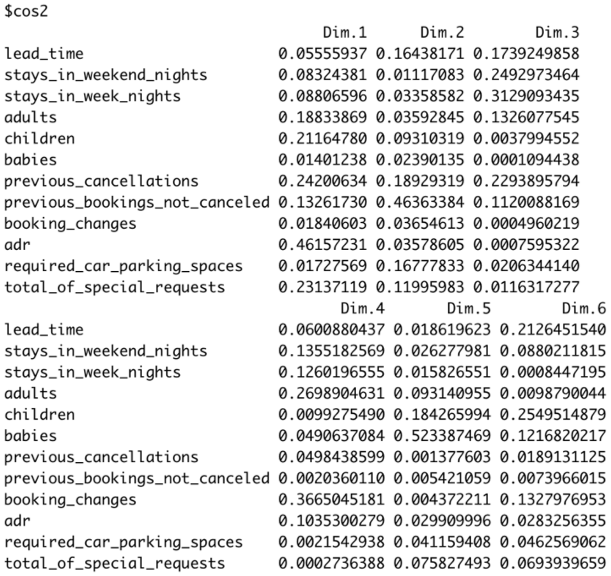




According to the parallel analysis plot for PCA, we chose the number of principal components to be 6.



```
$coord
                                  Dim.1      Dim.2       Dim.3
lead_time                    -0.2357104 -0.4054401  0.41704315
stays_in_weekend_nights       0.2885200 -0.1056921  0.49929685
stays_in_week_nights          0.2967591 -0.1832643  0.55938300
adults                        0.4339801 -0.1895480  0.36415348
children                      0.4600520  0.3051282 -0.06163972
babies                        0.1183739  0.1546006 -0.01046154
previous_cancellations       -0.4919414  0.4350784  0.47894632
previous_bookings_not_canceled -0.3641666  0.6809066  0.33467718
booking_changes               0.1356688  0.1911704  0.02227155
adr                           0.6793911  0.1891720 -0.02755961
required_car_parking_spaces   0.1314370  0.4096075 -0.14364684
total_of_special_requests     0.4810106  0.3463522  0.10785049
                                    Dim.4       Dim.5       Dim.6
lead_time                    -0.24512863  0.13645374  0.46113464
stays_in_weekend_nights       0.36812804 -0.16210485 -0.29668364
stays_in_week_nights          0.35499247 -0.12580362 -0.02906406
adults                       -0.51950983  0.30519003 -0.09939318
children                     -0.09963709 -0.42926215  0.50492721
babies                        0.22150329  0.72345523  0.34882950
previous_cancellations       -0.22325738 -0.03711608  0.13752495
previous_bookings_not_canceled -0.04512218 -0.07362784 -0.08600350
booking_changes               0.60539617  0.06612270  0.36441418
adr                          -0.32176082 -0.17294507  0.16830221
required_car_parking_spaces  -0.04641437  0.20287781 -0.21507419
total_of_special_requests    -0.01654203  0.27536792 -0.26342734
```



```
$cos2
                                    Dim.1      Dim.2        Dim.3
lead_time                      0.05555937 0.16438171 0.1739249858
stays_in_weekend_nights        0.08324381 0.01117083 0.2492973464
stays_in_week_nights           0.08806596 0.03358582 0.3129093435
adults                         0.18833869 0.03592845 0.1326077545
children                       0.21164780 0.09310319 0.0037994552
babies                         0.01401238 0.02390135 0.0001094438
previous_cancellations         0.24200634 0.18929319 0.2293895794
previous_bookings_not_canceled 0.13261730 0.46363384 0.1120088169
booking_changes                0.01840603 0.03654613 0.0004960219
adr                            0.46157231 0.03578605 0.0007595322
required_car_parking_spaces    0.01727569 0.16777833 0.0206344140
total_of_special_requests      0.23137119 0.11995983 0.0116317277
                                      Dim.4       Dim.5        Dim.6
lead_time                      0.0600880437 0.018619623 0.2126451540
stays_in_weekend_nights        0.1355182569 0.026277981 0.0880211815
stays_in_week_nights           0.1260196555 0.015826551 0.0008447195
adults                         0.2698904631 0.093140955 0.0098790044
children                       0.0099275490 0.184265994 0.2549514879
babies                         0.0490637084 0.523387469 0.1216820217
previous_cancellations         0.0498438599 0.001377603 0.0189131125
previous_bookings_not_canceled 0.0020360110 0.005421059 0.0073966015
booking_changes                0.3665045181 0.004372211 0.1327976953
adr                            0.1035300279 0.029909996 0.0283256355
required_car_parking_spaces    0.0021542938 0.041159408 0.0462569062
total_of_special_requests      0.0002736388 0.075827493 0.0693939659
```

$var$coord projects the variables to the PC-coordinates, while $var$cos2 represents the amount of variability captured by the component. The more variance of a variable explained by a PC, the stronger the relationship between that variable and the PC is. It turned out that the 13 original variables were represented by the 6 new principal components, and through the results we can reidentify the PCs with their new implications.

PC1: Captures a large amount of variability of adr, special requests and cancellation amount, but has a negative correlation with cancellation → **PC for high-demand & high-consumption & low-cancellation-rate customers**

PC2: Largely explains the variability of previous booking orders that have not been cancelled, also negative correlation with leading time → **PC for spontaneous checked-in orders**

PC3: Has relatively stronger relationships with staying time (both weekdays & weekends), also quite relates to cancellation amount → **PC for long stay customers**

PC4: Has strong correlation with number of adults and changes/amendments made to the booking → **PC for capricious customers**

PC5: Represents the number of babies → **PC for families with babies**

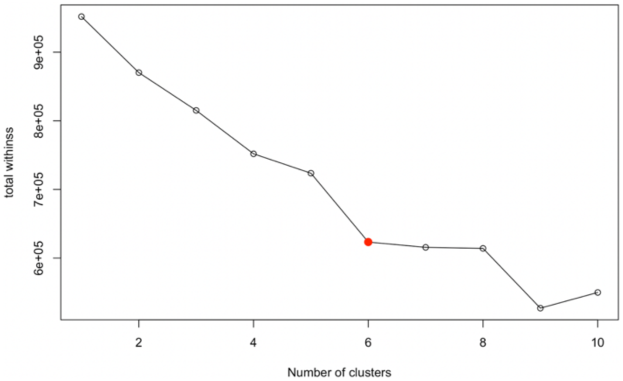
PC6: Weighs high on children and number of days between booking date and arrival date planning ahead of time → **PC for farseeing families with children**

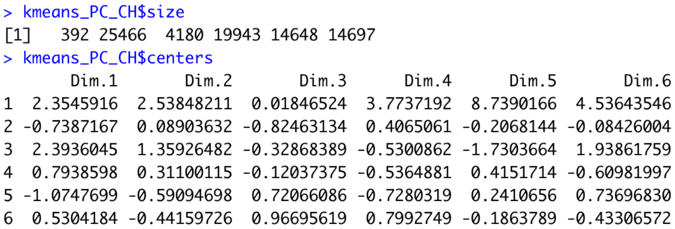
PC1	PC2	PC3	PC4	PC5	PC6
high-demand & high-consumption &low-cancellation-rate customers	spontaneous checked-in orders	long stay customers	capricious customers	families with babies	farseeing families with children

b. K-Means

After mapping the original data set to the new PC space, we can cluster the samples into different groups using k-means algorithm.

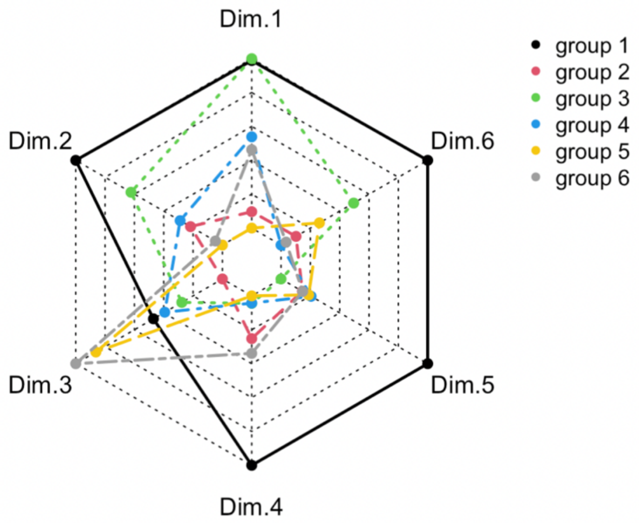
We used the Elbow Method to get the optimal value of k. Ideally we want a clustering that has the properties of internal cohesion and external separation. The total within-clusters sum of square indicates how close samples in each cluster are. The broken line graph shows that the decreasing trend of total withinss slows down at k = 6. Thus, k=4 was chosen to be the number of clusters.




```
> kmeans_PC_CH$size
[1]   392 25466  4180 19943 14648 14697
> kmeans_PC_CH$centers
       Dim.1       Dim.2       Dim.3      Dim.4      Dim.5       Dim.6
1  2.3545916  2.53848211  0.01846524  3.7737192  8.7390166  4.53643546
2 -0.7387167  0.08903632 -0.82463134  0.4065061 -0.2068144 -0.08426004
3  2.3936045  1.35926482 -0.32868389 -0.5300862 -1.7303664  1.93861759
4  0.7938598  0.31100115 -0.12037375 -0.5364881  0.4151714 -0.60981997
5 -1.0747699 -0.59094698  0.72066086 -0.7280319  0.2410656  0.73696830
6  0.5304184 -0.44159726  0.96695619  0.7992749 -0.1863789 -0.43306572
```

The results of clustering the 79,326 observations are as follows: There were 392 samples in group 1, 25,466 samples in group 2, 4,180 samples in group 3, 19,943 samples in group 4, 14,648 samples in group 5, 14,697 samples in group 6.

The coordinates of centroids for each cluster are also shown in the table. Since Z-Score standardization has been carried out on the data set before, it can be known that the mean value of data in each column is 0. By comparing the centers with 0, we can infer at what level each group is in terms of different PCs. The radar chart below also displays the weight on each PC of each group.



Targeted Marketing Strategies

Group 1: Key Customers. This group has high demand & high consumption & low cancellation rate, and always book hotels spontaneously. Also, they are at the highest levels of owning children or babies. They can be assumed to be well-to-do families with who usually travel. For these customers, hotels could advocate the "80/20 rule", which means that 80% of turnover comes from repeated purchases or consumption by 20% of loyal customers, while the other 20% comes from those 80% of free customers. Therefore, for these key customers, hotels can carefully investigate their preferences and launch VIP customized services. Such as, baby care services, etc. This type of customers are usually successful in their field, the hotel information can be advertised in magazines of various fields to attract our target customers.

Group 2: The biggest feature of this group is that they don't stay long, and they are not regarded as "high-level" customers according to the low level on PC1. Besides, every other metric is at rather average level. They have indistinct features but have a large population. Thus, for this part of customers, hotels don't need to spend a lot of time or energy to develop specific strategies, but can choose some general high performance-to-price ratio methods instead. And the marketing strategies for these customers we give are that hotels can implement some return visit questionnaires to figure out the customers' wishes, and can also make more interactions with the customers through social media. Avoid using platitudes and empty promises in ads, try to make the ads more attractive.

Group 3: This group has the highest levels of demand & consumption. It also quite resembles group 1 in terms of spontaneous bookings, except that this group has a relatively low level of owning children. We can speculate that they are mostly businesspeople who frequently go on business trips. For this customer group, the hotel can put ads on the APPs they usually use. For businesspeople, the most common apps are Google maps, LinkedIn, TripIt, WeatherPro, iTranslate, Uber, and Mynd Calender, etc. Meanwhile, the hotel can provide business suites, high-end business banquets and gyms. This group of people are business oriented; hotels can be recommended to the company's employees through the company's full-time personnel.

Group 4: This group is also at an average level in terms of every principal component. They don’t stay long, don’t have many requirements, nor do they plan ahead of time. Hence, this group can be regarded as potential customers like Group 2, and aforementioned general strategies can be made in both these two groups.

Group 5: This group stay long, plan ahead of time, not high-level group. Can be merged with group 6 as stated below.

Group 6: This group is high-level and stay relatively the longest among all groups, which can be merged with group 5 to conduct marketing. For these customers who prefer to order ahead, hotels can advance notice of low-priced room types by ads. Or, the hotel can recommend and sell discount coupons to these customers in some promotional seasons. This group of people tend to book the hotel ahead of time and search for relevant information actively, so a large amount of advertisement could be published on the Internet in advance.

Cornell University

ENMGT 5930 Project Two Report

Using Three Classification Methods for Predicting Hotel Cancellation

Yu Pan (yp425)
Yajing Wang (yw832)
Xinyue Zhao (xz332)
Ziyin Wang (zw252)

12/14/2021

Problem Statement

For resort hotel reservations, customers often cancel orders, and they are accustomed to hotels with free cancellation policies when booking hotels. The cancellation rate of orders in 2018 was as high as 40%, which had large losses to hotel revenues. There are many reasons for this, for example, customers canceled the order due to inaccurate weather forecasts, and some customers booked the order that has no charge to cancel the order, so they can cancel the reservation because of the change of schedule flexibly. Reviews of hotel services and the hotel reputation score also affect customers' decisions towards cancellation, for example, in 2017 reviews affected travel expenses of $546 billion. In order to improve the occupancy rate of the hotel and increase the revenue, the hotel manager needs to analyze the historical customer cancellation data and put forward a marketing strategy to deal with the potential cancellations and bring economic benefits for the hotel.

The following three classification methods could be used to predict the cancellation probability of new customers.

1. Logistic Regression

Logistic regression is a form of binary regression used to model the probability of a certain event. The amount of calculation is relatively small, and a large sample space can finish training in a short period of time. When sampled have nonlinear characteristics, transformation is required. In our case, the dependent variable is a binary categorical variable about whether a customer canceled the hotel reservation. Therefore, logistic regression can be conducted to predict a customer’s cancellation probability.

2. Classification Tree

Classification tree is used for classification and regression analysis. It basically works by classifying or predicting outcomes based on a set of variables. The classification tree is intuitive to explain and does not consider the correlation between samples, when the tree has a deep depth, it's prone to overfit. The branches of trees correspond to the features of samples, and the nodes represent the predicted classes.

3. Neural Network

Neural network is a nonlinear supervised learning model. Through the operation of many nodes, neural networks can obtain high classification accuracy and have strong learning abilities when choosing a reasonable number of hidden layers and hidden nodes. It is suitable for both linear data and nonlinear data. However, the logic behind the model is not intuitive and it is difficult to carry out clear logical analysis, it also requires a large amount of computation.

The input resort hotel cancellation data set was divided into two parts: 70% as training set and 30% as testing set. We used logistic regression, classification tree, and neural network model for analysis respectively. According to the prediction results, the two models with higher accuracy are selected as the final models of our project.

Exploratory Data Analysis (EDA)

I. Autocorrelation among numeric variables

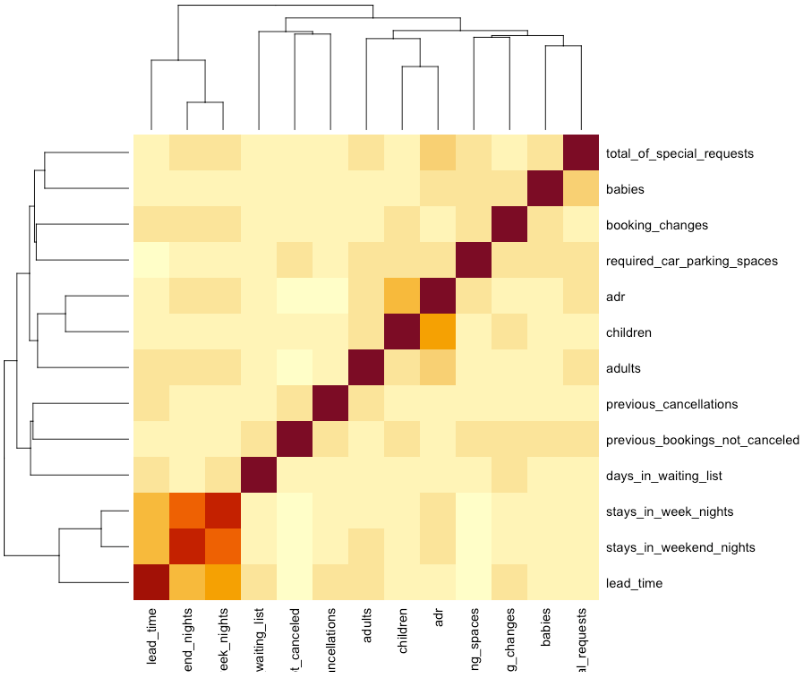


Figure 1. Correlation between each numeric variable

Firstly, we examined the correlation between each numeric variable. The correlation matrix and heatmap showed that each variable pair has a tolerable correlation that is no more than 0.4, except for that the correlation between stays_in_week_nights and stays_in_weekend_nights is 0.72. Hence, we consider eliminating the variables.

II. Correlation between independent & dependent variable

Since the target variable is categorical (is_canceled), the correlation matrix is not applicable for measuring the relationship between each independent variable. Therefore, we instead examined the scatter plots to see the distribution of each independent variable in the two circumstances of the dependent variable. Some variables are distributed differently in is_canceled = 0 vs is_canceled = 1, which indicates that it may be an important explanatory variable for predicting the cancellation.

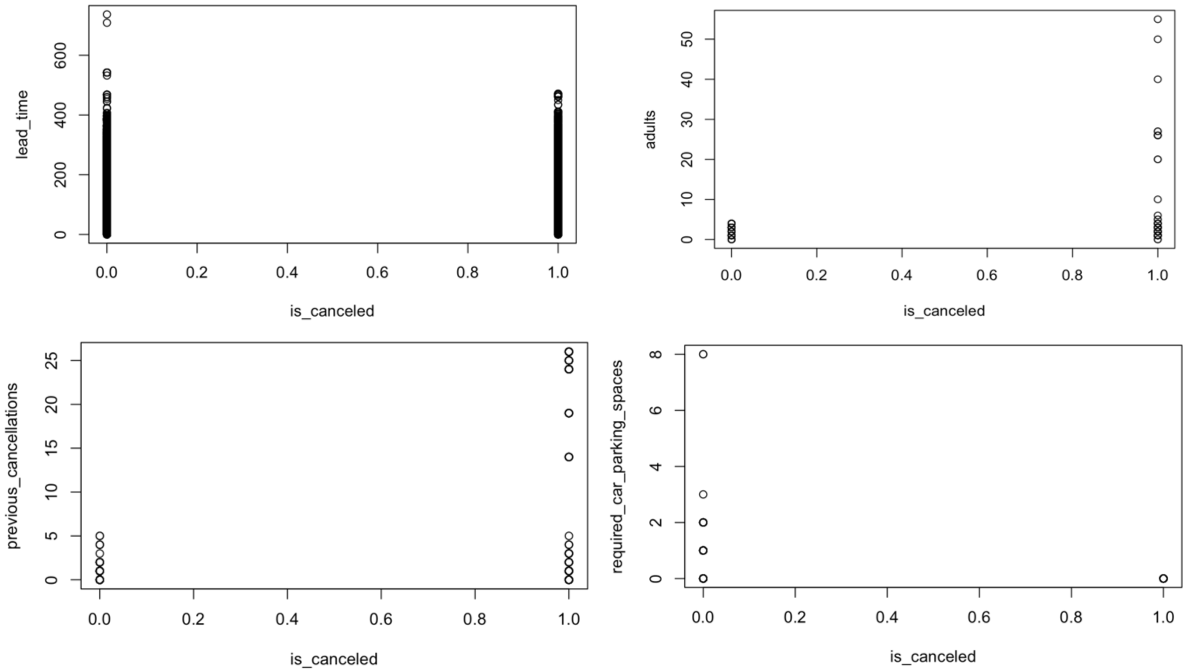
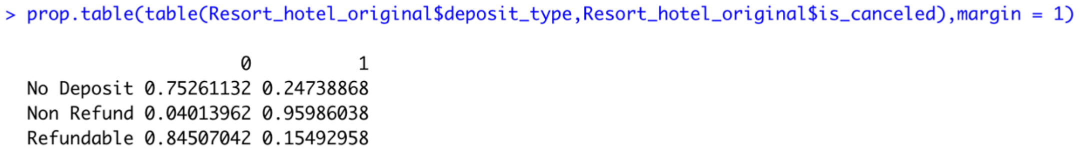


Figure 2. Correlation between independent & dependent variable

For example, higher values of lead_time and required_car_parking_spaces only appear when is_canceled = 0 but not when is_canceled = 1, showing that people with a longer interval between booking time and arrival time as well as more parking spots required would prone to keep their reservations. Similarly, relatively more adults and previous cancellations may lead to a higher cancellation probability. However, these are merely first-step estimations and their specific significance for the model will be displayed in the modeling part, evaluated by p-value and other metrics.

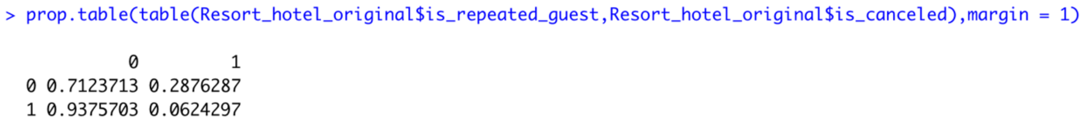
III. Choosing categorical independent variables

There are several original features that are categorical, such as arrival month, agent, type of meal booked, room type reserved. We created frequency tables and found that some variables' percentages of different values that correspond to different cancellation statuses vary a lot, indicating that these variables may have some sort of correlation with the target variable. For example, deposit type and repeated guest have obvious differences of distribution among each value in two cancellation statuses while room type has not.



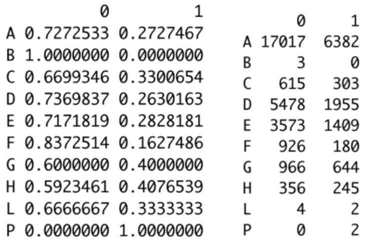
```
> prop.table(table(Resort_hotel_original$deposit_type,Resort_hotel_original$is_canceled),margin = 1)

                       0          1
  No Deposit 0.75261132 0.24738868
  Non Refund 0.04013962 0.95986038
  Refundable 0.84507042 0.15492958
```

Figure 3. Deposit type VS Cancel or not



```
> prop.table(table(Resort_hotel_original$is_repeated_guest,Resort_hotel_original$is_canceled),margin = 1)

            0         1
  0 0.7123713 0.2876287
  1 0.9375703 0.0624297
```

Figure 4. Repeated guest or not VS Cancel or not

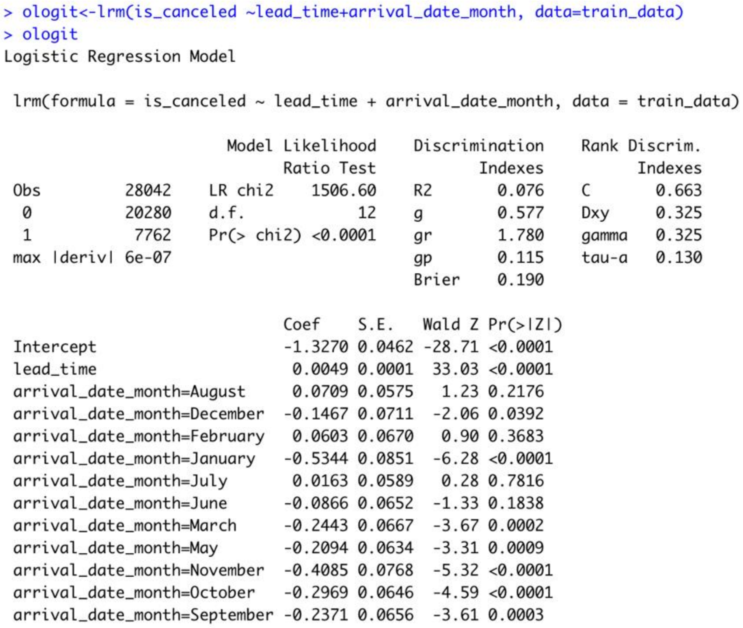


```
          0         1
A 0.7272533 0.2727467
B 1.0000000 0.0000000
C 0.6699346 0.3300654
D 0.7369837 0.2630163
E 0.7171819 0.2828181
F 0.8372514 0.1627486
G 0.6000000 0.4000000
H 0.5923461 0.4076539
L 0.6666667 0.3333333
P 0.0000000 1.0000000
```

```
      0     1
A 17017  6382
B     3     0
C   615   303
D  5478  1955
E  3573  1409
F   926   180
G   966   644
H   356   245
L     4     2
P     0     2
```

Figure 5. Room type VS Cancel or not

Furthermore, we considered the feasibility and complexity of model fitting. When a categorical variable with k different values is included in the regression, R will automatically create k-1 dummy variables. In that case, if we include categorical variables like arrival month in classification, too many dummy variables will be generated and make the model quite complex and less interpretable. Hence, we only chose those categorical variables with fewer than four values as input variables.



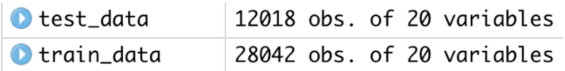
```
> ologit<-lrm(is_canceled ~lead_time+arrival_date_month, data=train_data)
> ologit
Logistic Regression Model

 lrm(formula = is_canceled ~ lead_time + arrival_date_month, data = train_data)

                       Model Likelihood     Discrimination     Rank Discrim.
                             Ratio Test            Indexes           Indexes
 Obs         28042    LR chi2    1506.60     R2       0.076     C       0.663
  0          20280    d.f.            12     g        0.577     Dxy     0.325
  1           7762    Pr(> chi2) <0.0001     gr       1.780     gamma   0.325
 max |deriv| 6e-07                           gp       0.115     tau-a   0.130
                                             Brier    0.190

                             Coef    S.E.   Wald Z Pr(>|Z|)
 Intercept                   -1.3270 0.0462 -28.71 <0.0001
 lead_time                    0.0049 0.0001  33.03 <0.0001
 arrival_date_month=August    0.0709 0.0575   1.23 0.2176
 arrival_date_month=December -0.1467 0.0711  -2.06 0.0392
 arrival_date_month=February  0.0603 0.0670   0.90 0.3683
 arrival_date_month=January  -0.5344 0.0851  -6.28 <0.0001
 arrival_date_month=July      0.0163 0.0589   0.28 0.7816
 arrival_date_month=June     -0.0866 0.0652  -1.33 0.1838
 arrival_date_month=March    -0.2443 0.0667  -3.67 0.0002
 arrival_date_month=May      -0.2094 0.0634  -3.31 0.0009
 arrival_date_month=November -0.4085 0.0768  -5.32 <0.0001
 arrival_date_month=October  -0.2969 0.0646  -4.59 <0.0001
 arrival_date_month=September -0.2371 0.0656  -3.61 0.0003
```

Figure 6. Dummy variables example

IV. Deciding variables & Splitting dataset

Eventually, we decided to use the following variables: lead_time, adults, children, is_repeated_guests, previous_cancellations, previous_bookings_not_canceled, booking_changes, deposit_type_non_refund, deposit_type_refundable, adr, and total_of_special_requests.

The entire dataset is split into a training dataset and a testing dataset. The ratio we used is 70% and 30% of the entire dataset respectively.



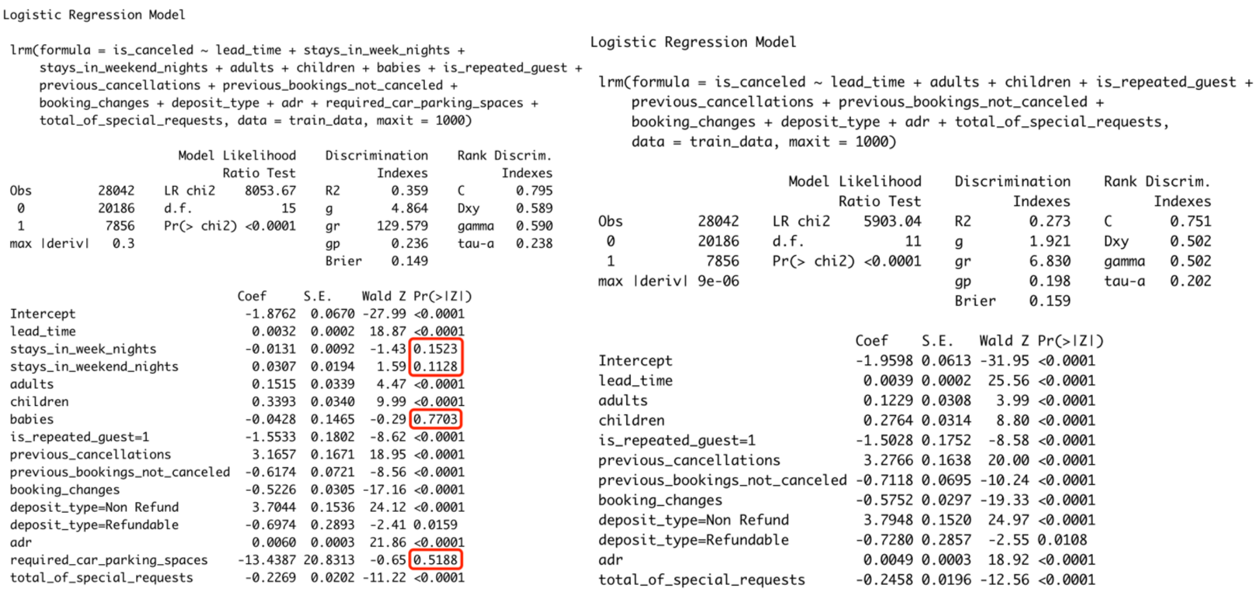
test_data	12018 obs. of 20 variables
train_data	28042 obs. of 20 variables

Figure 7. Data set deviation

The training dataset ended up with 28,042 observations, and the testing dataset has 12,018 observations. The same data are trained using different models with the same set of variables listed above to compare the effectiveness of among models.

Classification Modelling

I. Logistic Regression



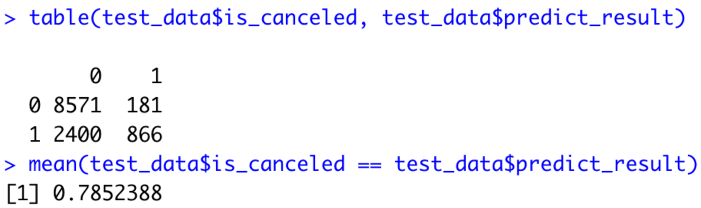
```
Logistic Regression Model

 lrm(formula = is_canceled ~ lead_time + stays_in_week_nights +
     stays_in_weekend_nights + adults + children + babies + is_repeated_guest +
     previous_cancellations + previous_bookings_not_canceled +
     booking_changes + deposit_type + adr + required_car_parking_spaces +
     total_of_special_requests, data = train_data, maxit = 1000)

                       Model Likelihood    Discrimination    Rank Discrim.
                          Ratio Test           Indexes           Indexes
 Obs          28042    LR chi2    8053.67    R2       0.359    C       0.795
  0           20186    d.f.            15    g        4.864    Dxy     0.589
  1            7856    Pr(> chi2) <0.0001    gr     129.579    gamma   0.590
 max |deriv|    0.3                          gp       0.236    tau-a   0.238
                                             Brier    0.149

                                 Coef     S.E.    Wald Z Pr(>|Z|)
 Intercept                        -1.8762  0.0670 -27.99 <0.0001
 lead_time                         0.0032  0.0002  18.87 <0.0001
 stays_in_week_nights             -0.0131  0.0092  -1.43 0.1523
 stays_in_weekend_nights           0.0307  0.0194   1.59 0.1128
 adults                            0.1515  0.0339   4.47 <0.0001
 children                          0.3393  0.0340   9.99 <0.0001
 babies                           -0.0428  0.1465  -0.29 0.7703
 is_repeated_guest=1              -1.5533  0.1802  -8.62 <0.0001
 previous_cancellations            3.1657  0.1671  18.95 <0.0001
 previous_bookings_not_canceled   -0.6174  0.0721  -8.56 <0.0001
 booking_changes                  -0.5226  0.0305 -17.16 <0.0001
 deposit_type=Non Refund           3.7044  0.1536  24.12 <0.0001
 deposit_type=Refundable          -0.6974  0.2893  -2.41 0.0159
 adr                               0.0060  0.0003  21.86 <0.0001
 required_car_parking_spaces     -13.4387 20.8313  -0.65 0.5188
 total_of_special_requests        -0.2269  0.0202 -11.22 <0.0001
```

```
Logistic Regression Model

 lrm(formula = is_canceled ~ lead_time + adults + children + is_repeated_guest +
     previous_cancellations + previous_bookings_not_canceled +
     booking_changes + deposit_type + adr + total_of_special_requests,
     data = train_data, maxit = 1000)

                       Model Likelihood    Discrimination    Rank Discrim.
                          Ratio Test           Indexes           Indexes
 Obs          28042    LR chi2    5903.04    R2       0.273    C       0.751
  0           20186    d.f.            11    g        1.921    Dxy     0.502
  1            7856    Pr(> chi2) <0.0001    gr       6.830    gamma   0.502
 max |deriv| 9e-06                           gp       0.198    tau-a   0.202
                                             Brier    0.159

                                Coef    S.E.   Wald Z Pr(>|Z|)
 Intercept                      -1.9598 0.0613 -31.95 <0.0001
 lead_time                       0.0039 0.0002  25.56 <0.0001
 adults                          0.1229 0.0308   3.99 <0.0001
 children                        0.2764 0.0314   8.80 <0.0001
 is_repeated_guest=1            -1.5028 0.1752  -8.58 <0.0001
 previous_cancellations          3.2766 0.1638  20.00 <0.0001
 previous_bookings_not_canceled -0.7118 0.0695 -10.24 <0.0001
 booking_changes                -0.5752 0.0297 -19.33 <0.0001
 deposit_type=Non Refund         3.7948 0.1520  24.97 <0.0001
 deposit_type=Refundable        -0.7280 0.2857  -2.55 0.0108
 adr                             0.0049 0.0003  18.92 <0.0001
 total_of_special_requests      -0.2458 0.0196 -12.56 <0.0001
```

Figure 8. logistic regression model

We first ran logistic regression based on 15 selected variables. After discarding independent variables that are not statistically significant, we kept 11 remaining variables to fit the logistic model. The revised model turned out to be fully statistically significant, and the rest of the classification methods that followed all used these variables as a standard.

Next, we made a prediction in the test set. It's worth mentioning that after using the "predict" function, we also needed to add a "Sigmoid" function to get the real probabilities of cancellation. Then we set a threshold of 0.5 which lets any probability that exceeds it be predicted as "cancel", and the probability that is less than 0.5 be predicted as "not cancel". The contingency table below shows that the accuracy rate of the prediction model is **78.52%**.



```
> table(test_data$is_canceled, test_data$predict_result)

       0    1
  0 8571  181
  1 2400  866
> mean(test_data$is_canceled == test_data$predict_result)
[1] 0.7852388
```

Figure 9. Results

II. Classification Tree

We used the 11 remaining variables which were selected before to build a classification tree model to predict ‘is_canceled’. Then, we cut out the tree model and trained the model on the training set. And we used the trained model to make predictions on the test set, the results are as follows:

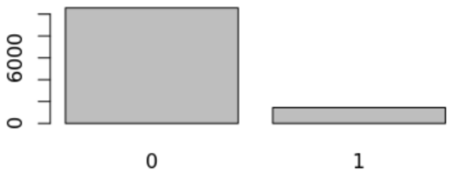
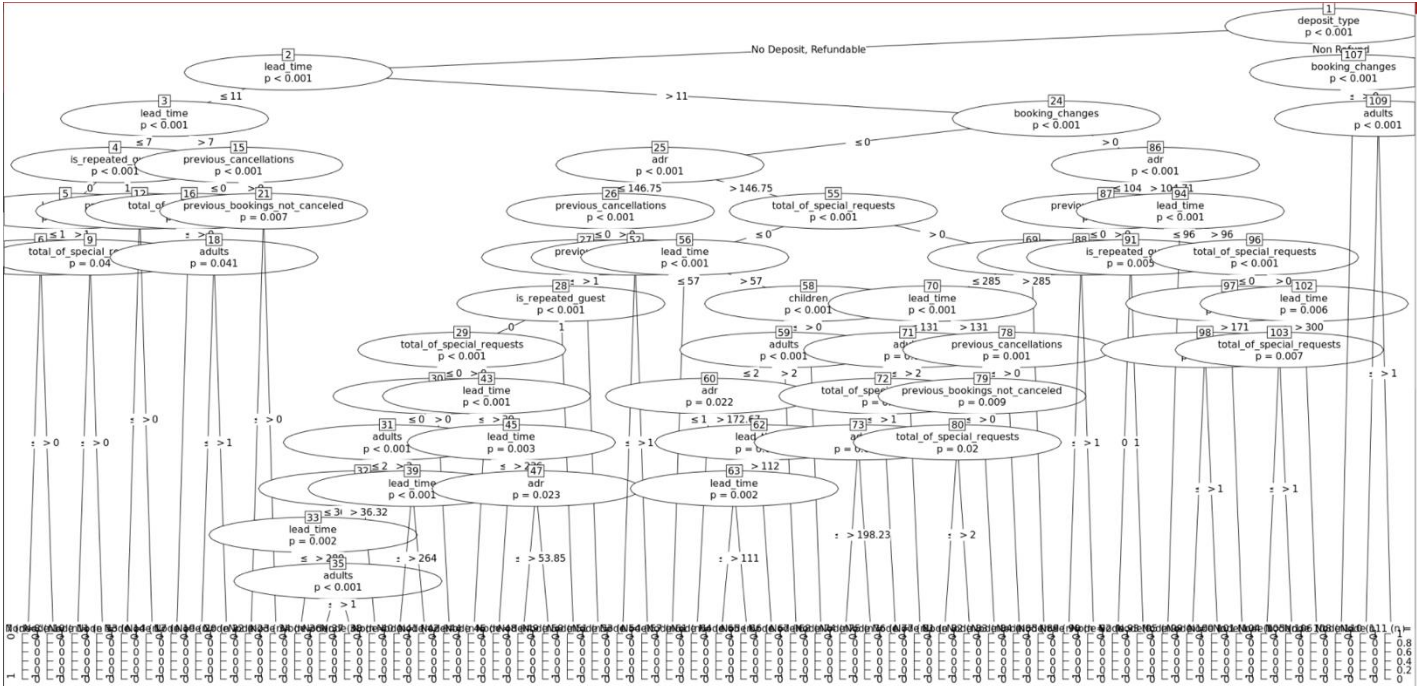


Figure 10. Prediction results

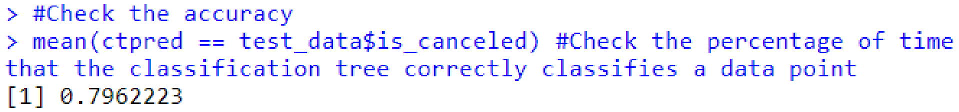
And the resulting tree model is as follows:



1 deposit_type p < 0.001
No Deposit, Refundable
Non Refund
2 lead_time p < 0.001
≤ 11
> 11
3 lead_time p < 0.001
≤ 7
> 7
4 is_repeated_g p < 0.001
15 previous_cancellations p < 0.001
16
21 previous_bookings_not_canceled p = 0.007
9 total_of_special_r p = 0.04
18 adults p = 0.041
24 booking_changes p < 0.001
≤ 0
> 0
25 adr p < 0.001
≤ 146.75
> 146.75
26 previous_cancellations p < 0.001
55 total_of_special_requests p < 0.001
56 lead_time p < 0.001
≤ 57
> 57
28 is_repeated_guest p < 0.001
29 total_of_special_requests p < 0.001
43 lead_time p < 0.001
31 adults p < 0.001
45 lead_time p = 0.003
39 lead_time p < 0.001
47 adr p = 0.023
33 lead_time p = 0.002
35 adults p < 0.001
58 children p < 0.001
59 adults p < 0.001
60 adr p = 0.022
63 lead_time p = 0.002
70 lead_time p < 0.001
78 previous_cancellations p = 0.001
79 previous_bookings_not_canceled p = 0.009
80 total_of_special_requests p = 0.02
86 adr p < 0.001
94 lead_time p < 0.001
91 is_repeated_g p = 0.005
96 total_of_special_requests p < 0.001
102 lead_time p = 0.006
103 total_of_special_requests p = 0.007
107 booking_changes p < 0.001
109 adults p < 0.001

Figure 11. Classification Tree

By mean function, the prediction accuracy can be estimated, which is **79.62%**.



```
> #Check the accuracy
> mean(ctpred == test_data$is_canceled) #Check the percentage of time
that the classification tree correctly classifies a data point
[1] 0.7962223
```

Figure 12. Results

III. Neural Network

With the same datasets and independent variables, the model was also built with neural network. First, the datasets are normalized using function(x) {return ((x-min(x))/(max(x)-min(x)))}. Neuralnet() function is applied, and number of hidden layers, number of nodes, and activation function are adjusted to obtain the better accuracy.

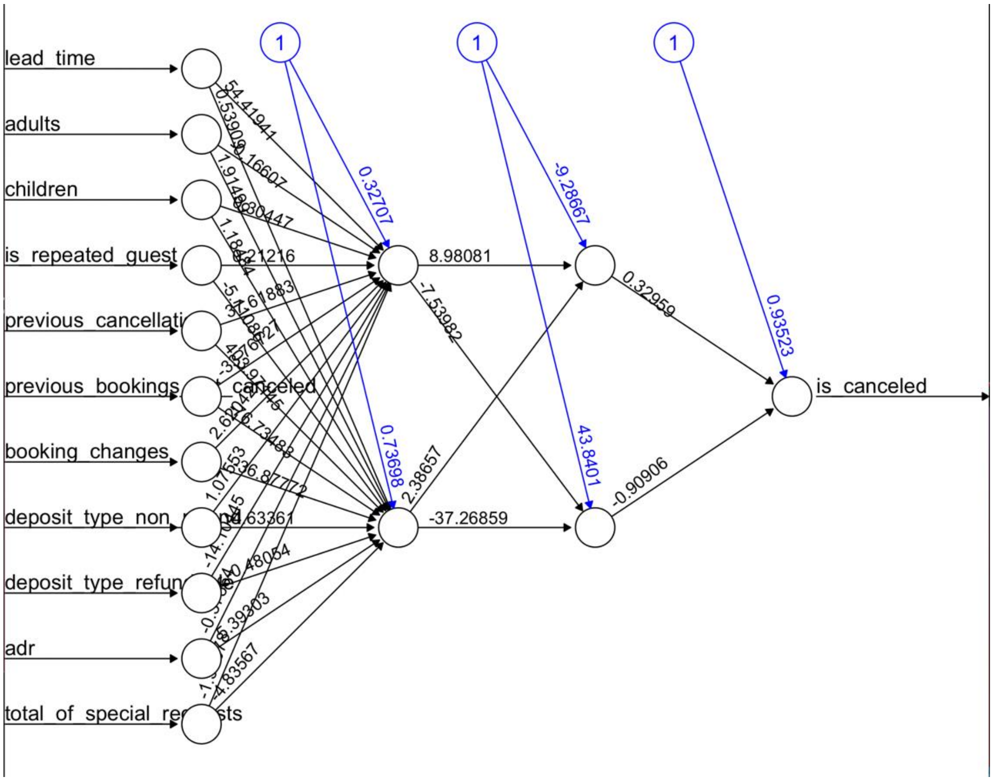
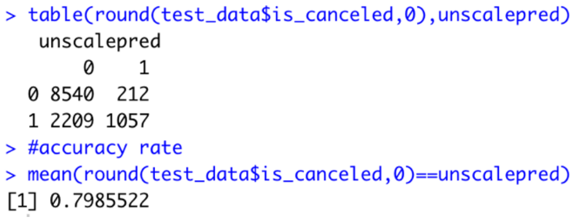


Figure 13. Neural Network

The best result we obtained is using logistic function, with 2 hidden layers and 2 nodes for each hidden layer. Three bias nodes were also inserted to gain better accuracy as shown.



```
> table(round(test_data$is_canceled,0),unscalepred)
   unscalepred
       0    1
  0 8540  212
  1 2209 1057
> #accuracy rate
> mean(round(test_data$is_canceled,0)==unscalepred)
[1] 0.7985522
```

Figure 14. Results

The overall result we obtained using neural network modeling is **79.855%.**

Results Comparison

On the data set of 28,042 observations, we trained the models of logistic regression, classification tree, and neural network, and evaluated the accuracy on the test set of 12,018 observations, the accuracy rate reaches 0.785, 0.796, and 0.798 respectively. The accuracy of these models is roughly at the same level, among which the neural network performs best and has the highest accuracy due to its super learning ability and various possible structures, classification tree, and logistic regression have slightly lower accuracy.

Conclusion

Throughout our EDA and modeling, we found that some important features that could be used to predict cancellation rate are lead time, whether the payment is refundable, total number of special requests, previous cancellation, booking changes, number of guests, etc. We figured these utilizing these independent variables were reasonable not just because they were statistically significant, but also realistic. For instance, guests tended not to cancel when the bookings were nonrefundable, they asked for special requests from the resort hotels, or low average daily rate they got at the time of book, etc. According to our selected variables, logistic regression, classification tree, and neural network can be used to predict the cancellation of resort hotels and facilitate managers to carry out effective management and marketing strategy. At present, we chose 11 variables from the original data set that can better characterize samples. From the result of the experiment, a neural network has the best performance based on designing a suitable structure. In the neural network model, two hidden layers, two nodes for each layer, and three bias nodes are added. The activation function used was logistic. An overall accuracy on the testing dataset is 79.855%. In order to further improve accuracy, more factors outside this data set can be considered in practice and added to each sample.

Discussion and Future Work

In order to compare the prediction performance of the three prediction models, we set the training set and the test set in advance and selected the same variables for the three models to facilitate comparison. But in fact, the choice of these variables may not be suitable for all models. For example: these 11 variables may be suitable for logit regression model, but if the same variables are used to build neural network and classification tree models, their prediction performance cannot be improved mostly. By using the same variables, training set and data set, we have reached the conclusion that the neural network has the best predictive performance, but this conclusion may have certain errors.

Because in general, the prediction success rate of classification tree is better than logistic regression model and neural network model, but it is more difficult to use classification tree structure in practice, this is according to Handan et al. [2007] Therefore, in the future, we can try to optimize and adjust the data set. At the same time, in this data set, we discarded a lot of categorical data and dummy variables, because this kind of data is not suitable for our model analysis. For example: there are 1 to 12 months, adding these variables will make our model difficult to explain and complicated. By discarding these variables, it may also affect our predicted results.

Reference:

Handan Ankarali Camdeviren, A. C. Y. Z. A., Resul Bugdayci, Mehmet Ali Sungur. "Comparison of logistic regression model and classification tree: An application to postpartum depression data." Elsevier 32: 987–994. [2007]

Appendix: Different Methods Pros & Cons Comparison

	Logit Regression	**Classification Tree**	**Neural Network**
Pros	·Easy and efficient to implement and interpret. ·It makes no assumptions about distributions of classes in feature space. ·It can easily extend to multiple classes (multinomial regression) and a natural probabilistic view of class predictions.	· Easy to understand and implement · No need to prepare data · Ability to handle both data type and regular type attributes at the same time · It can make feasible and effective results for large data sources in a relatively short time · Insensitive to missing values · Can handle irrelevant characteristic data · High efficiency, only need to build once and use repeatedly	·It can deal with nonlinear and complex relationships. ·Easy to conceptualize and generalize ·Prediction is relatively fast
Cons	·The dependent variable of Logistic Regression is bound to the discrete number set. ·Nonlinear problems can't be solved with logistic regression since it has a linear decision surface.	· It is difficult to predict continuous fields · For chronological data, a lot of preprocessing work is required · When there are too many categories, errors may increase faster · Doesn't perform well when dealing with data with relatively strong feature correlation	·The process is very time-consuming when training with a large amount of data with normal CPUs. ·With many layers in between the input layer and the output layer and also many nodes, it is hard to tell how much each independent

			variable is influencing the dependent variables.