Brown-Project

Tyler Brown

2022-12-02

Abstract

The Hotel booking dataset describes the hotel demand between July of 2015 and August of 2017. The data has two different types of hotels: resort hotel and city hotel. There are 31 variables describing 40,060 observations of resort hotels and 79,330 observations of city hotels. Each of the observations represents a hotel booking. All data pertaining to hotel or customer identification were removed for confidentiality.

Overview

The data is collected by Hotel Booking Demand Datasets available through ScienceDirect.com, written by Nuno Antonio, Ana Almeida, and Luis Nunes in February 2019. Additionally, the observational study is openly available from Kaggle.com. The data was cleaned by Thomas Mock and Antoine Bichat on February 11, 2020.

The quantitative dependent variables within the dataset are Children and Babies. The qualitative independent variable is the Hotel as it facilitates the name of the hotel. Additionally, Is_Canceled and Adults are quantitative independent variables as they provide the total number of adults under the reservation. But what is the probability of the customer's that cancel their reservation have a child or baby between the two different hotels? The variables will assist in answering this research question.

Summary Statistics

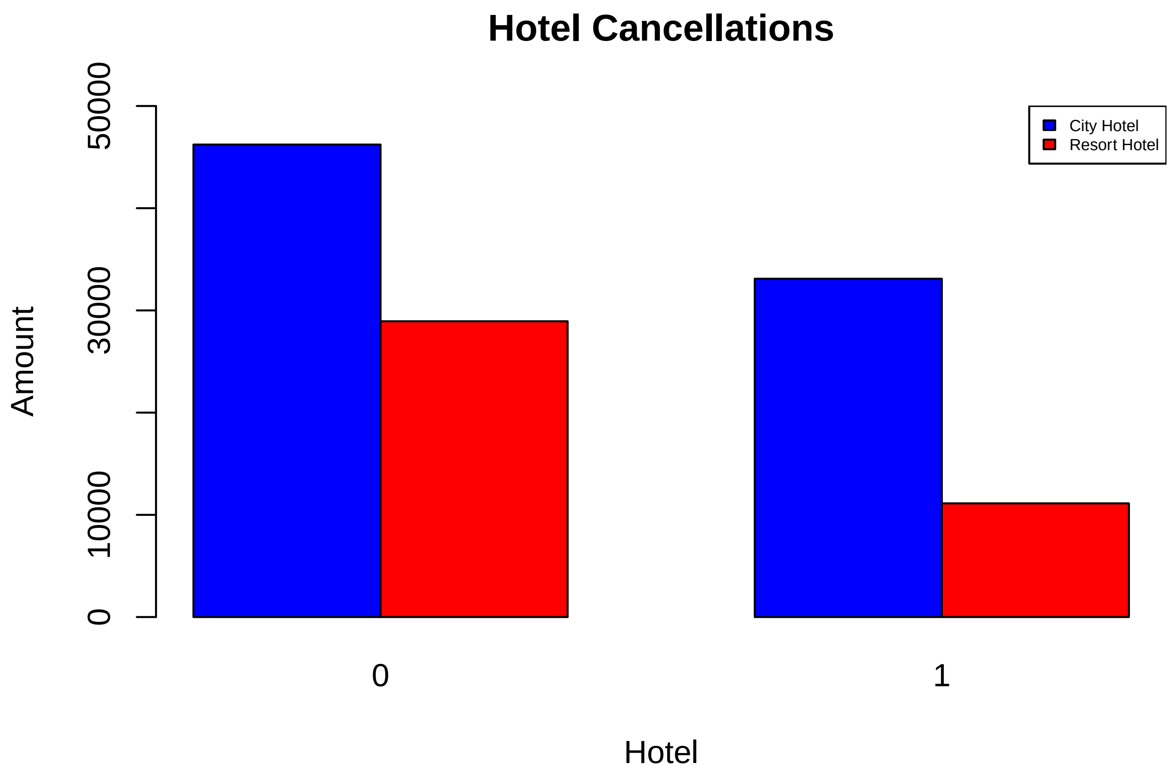
The Hotel show the total amount of reservations made. The Is_Canceled show the average cancellations being 0.37. The Adults show the average number of adults per reservation being 1.856 as well as the median amount of adults per reservation at 2. The Children show the average number of children being 0.1039 over all reservations. The Babies show the average number of babies being 0.0079 over all reservations.

```
summary(df)
```

```
##     hotel            is_canceled         adults          children
##  Length:119390      Min.   :0.0000   Min.   : 0.000   Min.   : 0.0000
##  Class :character   1st Qu.:0.0000   1st Qu.: 2.000   1st Qu.: 0.0000
##  Mode  :character   Median :0.0000   Median : 2.000   Median : 0.0000
##                     Mean   :0.3704   Mean   : 1.856   Mean   : 0.1039
##                     3rd Qu.:1.0000   3rd Qu.: 2.000   3rd Qu.: 0.0000
##                     Max.   :1.0000   Max.   :55.000   Max.   :10.0000
##      babies
##  Min.   : 0.000000
##  1st Qu.: 0.000000
##  Median : 0.000000
##  Mean   : 0.007949
##  3rd Qu.: 0.000000
##  Max.   :10.000000
```

Data Visualizations

The BarPlot below shows cancellations between the two hotels, where 1 represents a cancellation and 0 does not. It can be concluded that city hotels had more reservations, the most cancellations and not.



Due to this inequality among the hotel volume, let's identify the percentage of cancellations per hotel.

```
hotel_group = df %>%
  group_by(hotel,is_canceled) %>%
  count(is_canceled)
hotel_group
```

```
## # A tibble: 4 x 3
## # Groups:   hotel, is_canceled [4]
##   hotel        is_canceled     n
##   <chr>              <int> <int>
## 1 City Hotel             0 46228
## 2 City Hotel             1 33102
## 3 Resort Hotel           0 28938
## 4 Resort Hotel           1 11122
```

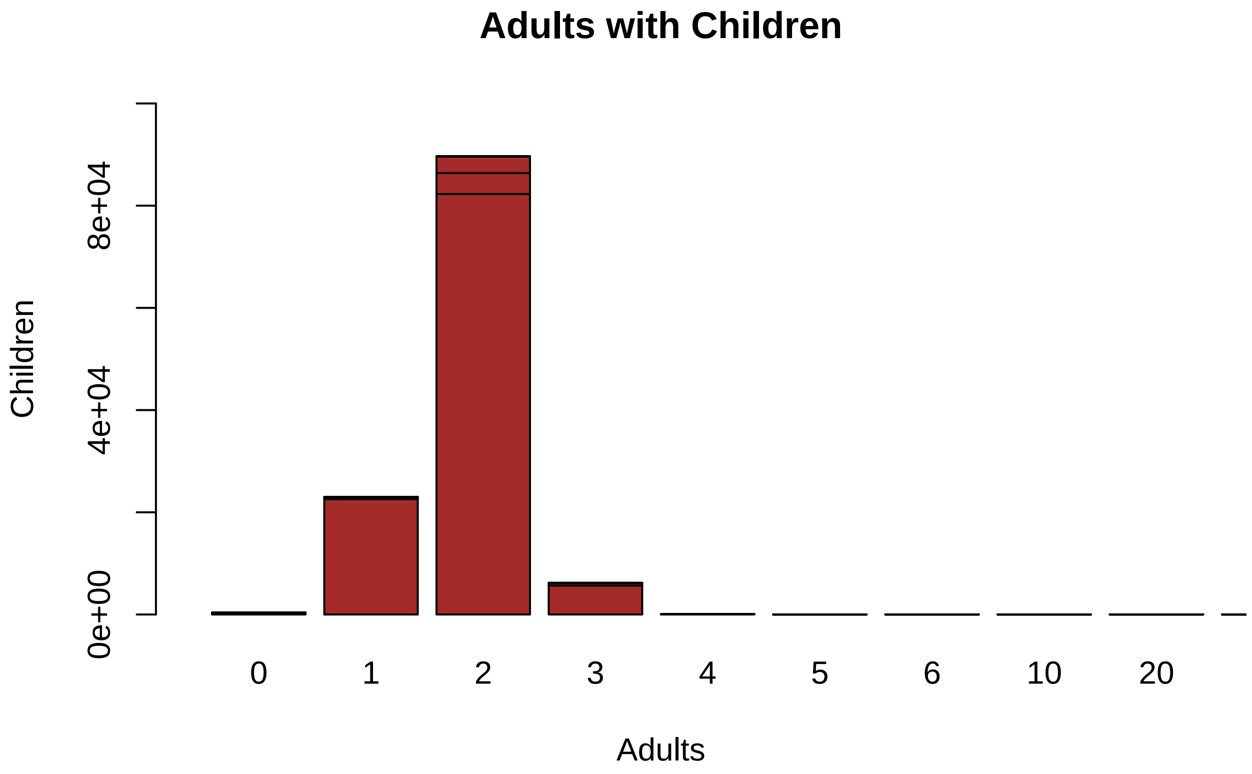
```
city = round((33102/(46228+33102))*100,2)
print(paste0("City Hotels have a ", city,"% cancellation rate."))
```

```
## [1] "City Hotels have a 41.73% cancellation rate."
```

```
resort = round((11122/(11122+28938))*100,2)
print(paste0("Resort Hotels have a ", resort,"% cancellation rate."))
```

```
## [1] "Resort Hotels have a 27.76% cancellation rate."
```

Now lets take a look at the total reservations of children with or without adults. As we can see, most of the reservations with children were made with 2 adults as well. It is shown there are a few reservations with children made without an adult on the reservation.



We can analyze this further and see numerical values of the total reservations made with the amount of adults, children, and the corresponding total.

```
adult_child = df %>%
  group_by(adults,children) %>%
  count(children)
adult_child
```

```
## # A tibble: 27 x 3
## # Groups:   adults, children [27]
##    adults children     n
##     <int>    <dbl> <int>
##  1      0        0   180
##  2      0        1     4
##  3      0        2   208
##  4      0        3    11
##  5      1        0 22587
##  6      1        1   279
##  7      1        2   157
##  8      1        3     4
##  9      2        0 82281
## 10      2        1  4089
## # ... with 17 more rows
```

Statistical Output

Logistic Regression Model

```
##
## Call:  glm(formula = is_canceled ~ children + babies, family = "binomial",
##     data = train)
##
## Coefficients:
## (Intercept)     children       babies
##    -0.52719      0.03137     -0.89886
##
## Degrees of Freedom: 95511 Total (i.e. Null);  95509 Residual
## Null Deviance:       125900
## Residual Deviance: 125800    AIC: 125800
```

From the Logistic Regression model, we have an R-squared of 0.001 and an accuracy of 0. The Confusion Matrix table shows a low percentage of successful true positives and true negatives. These results are dramatically low, therefore the model did not fit the data well and has a low predictive power.

```
#Evaluating Accuracy/Confusion Matrix
print(paste0("The R-squared is ",with(summary(logistic_model), 1 - deviance/null.deviance)))
```

```
## [1] "The R-squared is 0.000930292323550685"
```

```
missing_class = mean(predict_reg != test$is_canceled)
print(paste0('Accuracy = ', 1 - missing_class))
```

```
## [1] "Accuracy = 0"
```

```
table(test$is_canceled, predict_reg)
```

```
##    predict_reg
##     0.0890815343592638 0.0916604759708653 0.19371566813849 0.198663164284924
##   0                  3                  1             116                24
##   1                  0                  0              21                 3
##    predict_reg
##     0.203705095243963 0.371172826704161 0.378524827860103 0.385933078768364
##   0                 6             13822               654               407
##   1                 1              8193               303               308
##    predict_reg
##     0.393394540700865
##   0                12
##   1                 4
```

Conclusion

To conclude, the probability of the customer's that cancel their reservation have a child or baby between the hotels cannot be accurately determined using the Logistic Regression model. The Hotel Booking dataset has an abundance of data but the data is not diverse enough to use for predictive modeling. The R-squared and accuracy are too low, therefore I would not recommend implementing this model.

Limitations

The analysis is important to draw a correlation between reservations with children and reservation cancellations. Hotel analysts could use this data to prepare employees for cancellations due to children or babies. This analysis was limited due to the data provided and the lack of transparency of hotel information.