Data Science Project Report & Findings

Abstract:

The results of this analysis were broken into segments based on the findings the dataset provided. The segments which are four different analytics, contributed to our approach to the data & gave answers to questions. The four different analytics are Descriptive, Diagnostic, Predictive, and Prescriptive. The report will reflect our thoughts for each individual approach.

People stay in hotels for a variety of reasons. Some stay because they are traveling for business, while others stay because they are on vacation. Every traveler has a certain amount of money they can afford to spend, each time they visit your hotel. Guests can spend money on a room, on food and drinks, as well as amenities and services. The total amount of money the hotel receives for all the goods and services it provides to its guests is defined as the Hotel's Revenue.

Submitted by: Avani Rasikbhai Patel, Reid Yaworski, Chensheng Ma, Issak Hernandez
IST 687 Introduction to Data Science
Professor: Stephan Wallace, Jasmina Techeva
05/9/2021

Exploratory Analysis

In this section, we were given the project details and were instructed to provide the project deliverables. Our clients asked us to provide an analysis on the effects that certain values had in relation to Average Daily Rate (ADR). Along with the analysis, our team will also provide recommendations from that analysis.

Prior to developing the answers to the questions our client required, our team approached the dataset in an exploratory form. Utilizing R Studio, our team read, manipulated, and researched the attributes from the dataset. Once the dataset was understood, the team was able to articulate and generate questions from the provided overview.

The approach taken from the problem set and our exploratory analysis guided the team to conduct the following steps:

- Data Cleansing
- Descriptive Analysis
- Diagnostic Analysis
- Predictive Analysis
- Prescriptive Analysis

Data Cleansing

During our exploration, the team encountered several problems with the data. To work with a dataset that would benefit the team for a complete thorough analysis, the data must be cleansed. A few problems with the 'dirty' data revolved around redundancies, and lack of additional variable insight. For example, determining which variables to add in addition to the original dataset.

Another problem we encountered was determining the validity behind the dates within our datasets. Our team had to understand and research the need for specific dates in the current dataset to create additional variables that would provide information on cancelled hotel rooms. For example, the following code (**Excerpt1**) allowed us to solo specific values for the column "Reservation Status Date". This date represents the day the last status was set. That variable along with "StaysinWeekend" and "StaysinWeekdays" allowed us to apply simple addition and create a variable that represents the reservation checkout dates. Utilizing this information provided our team with substantial information that applied to our analysis.

Excerpt1

```
ReservationCheckOutDates <- H2_City[which(H2_City$ReservationStatus=='Check-Out'
&H2_City$ReservationStatus!='No-Show' &H2_City$ReservationStatus!='Canceled'),]

ReservationCheckOutDates <- ReservationCheckOutDates$ReservationStatusDate

NightsStayed <- H2_City[H2_City$ReservationStatus=='Check-Out',]


NightsStayed <- NightsStayed$StaysInWeekendNights + NightsStayed$StaysInWeekNights

H2_City[which(H2_City$ReservationStatus=='Check-Out'),]$`Arrival Date` <-
(as.Date(ReservationCheckOutDates) - NightsStayed)

sum(is.na(H2_City[H2_City$ReservationStatus=='Canceled',]$`Arrival Date`)) +
sum(is.na(H2_City[H2_City$ReservationStatus=='No-Show',]$`Arrival Date`))
```

The following steps were conducted in cleaning the data

- Add Missing Variables
- Verify Data Types
- Check for Unexpected Characters in Data
- Check for Wrong Data (Logically)
- Fix(impute)/Omitting Inconsistencies within the Data
- Checked for Outliers
- String Normalization
- Remove Redundant Variables (Columns)
- Remove Records (Rows)

Descriptive Analysis

In this stage of our analysis, our data is starting to paint a picture and give us insight into what it is conveying. As stated, we were required to provide substance to how ADR could be influenced between two hotel locations. In this step, there was a considerable change in the team's mindset and the interpretation of the data. After cleansing the data, the team understood the importance of cleansing the data first. Initially, an exploratory analysis was done. It helps provide an overview of the data, but cleansing provided a clearer sense of what our team could

do and build off with the clean data. During this step it was abundantly clear that things would need to be done to our code to approach it with an abstract eye to see further into the data and client's request.

While processing and running our initial code to understand what the code provided to our team, one aspect that we had to consider was attributes from the data that did not make sense or should be utilized as attributes combined from other attributes to make one column. For example, understanding the relationship between guest that were staying at the hotels. A sense of normalization and logic was utilized to make our columns less autonomous. To utilize the "Visitor Type" to the full extent we used conditionals to capture whether a guest was single, couple, or family. (**Excerpt 2)** is the logic our team used to validate the "Visitor Type". A key factor to consider was the cleansing of NAs in this column.

Excerpt2

```
H1_Resort$VisitorType <- ifelse(H1_Resort$Adults+H1_Resort$Children+H1_Resort$Babies==1, 'Single',
                        ifelse(H1_Resort$Adults+H1_Resort$Children+H1_Resort$Babies==2, 'Couple',
                        ifelse(H1_Resort$Adults+H1_Resort$Children+H1_Resort$Babies>=3, 'Family', NA)))
```

Additionally, while cleansing, certain columns were removed to help with our analysis and silo our primary objective. Creating the picture for our analysis was contributed by the various amount of descriptive code that was sought out of the dataset. A brief example was given by (**Excerpt2**). In combination with that code, it was also especially important to understand how our dates would be considered. Time frames, seasons, holidays would help our team understand the customer base and what times of the year their guest are arriving to the hotels. It may be safe to generalize that most guest travel during the summer but establishing the structure for the season was a particularly important part to our analysis.

During our descriptive analysis no definite answer could be found or concluded to justify to our client that any reasonable insight could be provided.

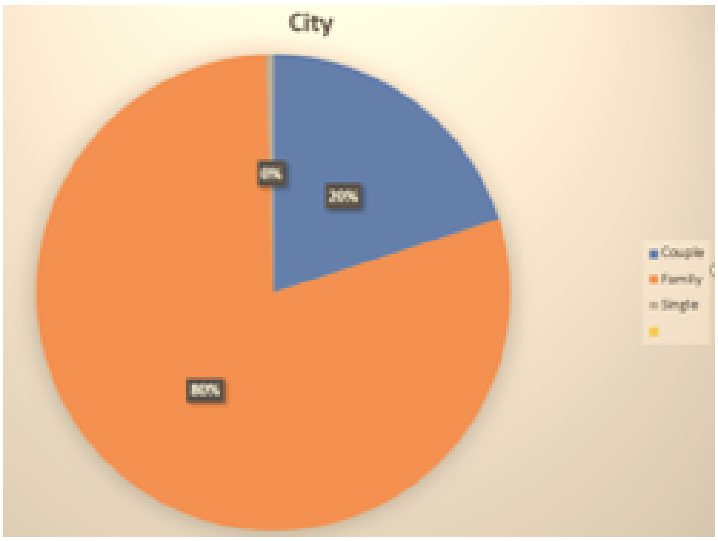
Visualizations

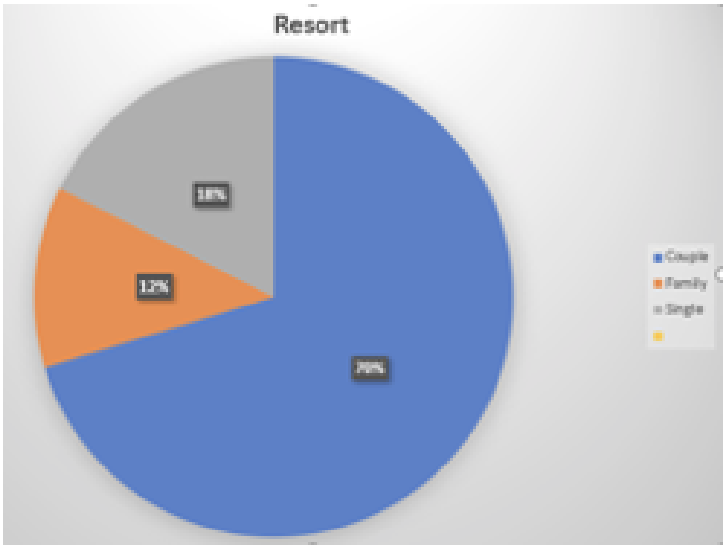
Within our analysis we have provided several visualizations to showcase to our clients a representation of what the data is showing.

Typical Visitor Pie Chart

We found that our typical visitor for city hotels were numbers that resembled that of a family (3 or more). The visitors traveled from Portugal and during the summer. The visitors also primarily booked their reservation online and was a transient customer.

There are many similarities between the city and resort dataset. The only difference from the city hotels is the visitor type is couples rather than family size visitors.



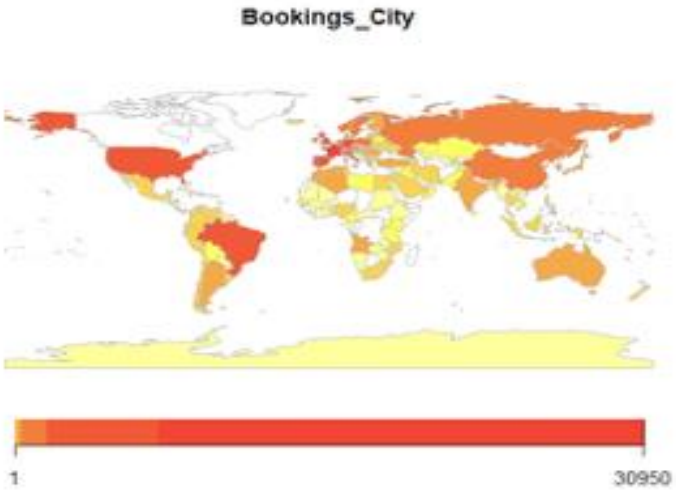


Seasonal Market

To help visualize the seasonal importance of travel for customers between both locations our team created bar plots to capture which season brings in the greatest ADR. The bar plots below show that seasonal reservations are high in summer and lowest in winter. According to the data, this has been a typical trend throughout the years. This trend can be linked to the fluctuation in ADR throughout the course of a year. One important thing to note is that the resort hotel has in increase in ADR for the additional seasons as well.

Popularity Global Map

These map visualizations represent the global origins from visitors in respect to both hotel locations. Both maps show a concentrated coloration from Portugal, but the map also provides a great visual of where other booking are coming from. The darker the color equates to a larger number of customers that booked from that country.



Diagnostic Analysis

In search of insight that could provide our analysis additional information, we turn to diagnostic analytics. Based on the dataset, which is historical data, we understand what factors our client have had to maneuver to understand their customer base.

Based on this historical data our team has been able to establish attributes for most frequent visitors for our clients. In addition, we have developed our analysis to provide where

their customers are coming from, whether they are single, couples, or families, & what season they visit the hotel based on the dates.

One column we paid close attention to was whether the customer cancelled and what factors lead to that cancellation as well the lead time for those reservations. From there we were able to gain more insight on where the revenue is coming from. Our team considered the deposit type and the season to generate visuals to provide structure to our analysis.

Continuing with diagnostic analysis, we were able to understand what other factors contributed to cancellation or what attributes played a role in a customer canceling. This will be discussed in Predictive Analytics. First our team needed to understand where our largest contributors came from to justify which models, we would endorse. For example, (**Excerpt3)** utilizes the package "ggplot2" our team created several barplots to understand what season had the most cancellations for our customers hotels. In addition, (**Excerpt4**) will show what the largest cancellations contributor by market segment.

Excerpt3

```
library('ggplot2')
barSeason<-ggplot(H1_Resort, aes(Season, IsCanceled)) +
  geom_bar(stat="identity")

barSeason #bar plot illustrating cancellation per season

barSeason2<-ggplot(H2_City, aes(Season, IsCanceled)) +
  geom_bar(stat="identity")

barSeason2 #bar plot illustrating cancellation per season
```

Excerpt4

```
## LARGEST CANCELLER BY SEGMENT ##
barSegment<-ggplot(H1_Resort, aes(MarketSegment, IsCanceled)) +
  geom_bar(stat="identity")

barSegment #cancellation by MarketSegment

barSegment2<-ggplot(H2_City, aes(MarketSegment, IsCanceled)) +
  geom_bar(stat="identity")

barSegment2 #cancellation by MarketSegment
```

This phase did provide our team a greater understanding of the characteristics within the data. We were able to manipulate the data to show the value and full capability of the customer base and its general relation to ADR. Although diagnostic analysis did contribute to a major gain for our client, this step did not generate enough insight to provide a full analysis.

Predictive Analysis

During this section we utilized linear models, associated rules mining and support vector machines techniques to support our statistical analysis of the client's customer data. While each provides its own value, it is important to note each have considerations with respect to each other.

Linear Models

Our team has provided 6 linear models to represent the predictions from the data. It was challenging to justify which attribute/predictor to utilize to generate the best possible linear relationship between the dependent variable, ADR. If it weren't for substantial time taken in exploration, cleansing, descriptive analysis, and diagnostic analysis our team would not have been able to silo out the best predictors.

The following linear models were created for Resorts & City Hotels

- Linear Model Predicting ADR by Season
- Linear Model Predicting ADR by Market Segment
- Linear Model Predicting Cancellation by Market Segment
- Linear Model Predicting Cancellation by Market Segment with Multiple Variables
- Linear Model Predicting ADR by Room Type with Multiple Variables

One consideration to note for our linear model was the possibility of analyzing the price deference between weekend and weekday customers based on the visitor type. Taking that information and applying it may influence ADR.

What are these linear models representing? Each of our models represents the positive relationship between the dependent variables (ADR) and the predictor variables. For example, (**Excerpt5**) will show just how strong that relationship is. Alongside these models, (**Excerpt5**) will also include linear models for predicting cancellation.

Excerpt5

```
Call:
lm(formula = ADR ~ AssignedRoomType + Season + MarketSegment + 
    Meal, data = H1_Resort)

Residuals:
     Min       1Q   Median       3Q      Max 
-236.160  -18.632   -1.938   15.586  288.595 

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)    
(Intercept)                -24.6922     2.5628  -9.635  < 2e-16 ***
AssignedRoomTypeB            5.3212     2.8488   1.868 0.061788 .  
AssignedRoomTypeC           21.3292     0.8133  26.225  < 2e-16 ***
AssignedRoomTypeD           10.2148     0.4549  22.455  < 2e-16 ***
AssignedRoomTypeE           21.8961     0.5576  39.268  < 2e-16 ***
AssignedRoomTypeF           29.4923     0.9234  31.939  < 2e-16 ***
AssignedRoomTypeG           61.7691     0.8929  69.182  < 2e-16 ***
AssignedRoomTypeH           70.8981     1.3861  51.148  < 2e-16 ***
AssignedRoomTypeI          -29.4271     2.1113 -13.938  < 2e-16 ***
AssignedRoomTypeL         -134.6318    35.7323  -3.768 0.000165 ***
SeasonSpring                -0.8867     0.5141  -1.725 0.084585 .  
SeasonSummer                76.0344     0.5119 148.524  < 2e-16 ***
SeasonWinter               -16.6935     0.5622 -29.692  < 2e-16 ***
MarketSegmentCorporate      61.5598     2.6352  23.361  < 2e-16 ***
MarketSegmentDirect         91.2896     2.5633  35.614  < 2e-16 ***
MarketSegmentGroups         66.7358     2.5835  25.832  < 2e-16 ***
MarketSegmentOffline TA/TO  55.7998     2.5651  21.754  < 2e-16 ***
MarketSegmentOnline TA      91.3240     2.5429  35.914  < 2e-16 ***
MealFB                      43.3256     1.3582  31.898  < 2e-16 ***
MealHB                      31.7252     0.4669  67.950  < 2e-16 ***
MealSC                     -24.8798     4.6089  -5.398 6.77e-08 ***
MealUndefined               45.5388     1.1393  39.972  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 35.73 on 39975 degrees of freedom
Multiple R-squared:  0.6609,    Adjusted R-squared:  0.6607 
F-statistic:  3710 on 21 and 39975 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = ADR ~ AssignedRoomType + Season + MarketSegment + 
    Meal + Agent, data = H2_City)

Residuals:
    Min      1Q  Median      3Q     Max 
-196.74  -13.90   -2.46   12.69  332.26 

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)               52.4453     1.9398  27.036  < 2e-16 ***
AssignedRoomTypeB         -6.8718     0.6094 -11.277  < 2e-16 ***
AssignedRoomTypeC          3.4385     2.0952   1.641 0.100783    
AssignedRoomTypeD         17.0203     0.2681  63.483  < 2e-16 ***
AssignedRoomTypeE         37.8737     0.5940  63.762  < 2e-16 ***
AssignedRoomTypeF         68.7121     0.6146 111.806  < 2e-16 ***
AssignedRoomTypeG         89.7969     1.0291  87.255  < 2e-16 ***
AssignedRoomTypeK        -40.1208     1.6064 -24.976  < 2e-16 ***
 [ reached getOption("max.print") -- omitted 238 rows ]
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26.48 on 79050 degrees of freedom
Multiple R-squared:  0.5477,    Adjusted R-squared:  0.5463 
F-statistic: 390.7 on 245 and 79050 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = IsCanceled ~ LeadTime + Season + MarketSegment + 
    Country + CustomerType + Agent + VisitorType, data = H1_Resort)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.97213 -0.25799 -0.08695  0.22889  1.30713 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)              3.260e-01  9.787e-02   3.332 0.000864 ***
LeadTime                 1.155e-03  2.478e-05  46.622  < 2e-16 ***
SeasonSpring             3.419e-02  5.748e-03   5.948 2.74e-09 ***
SeasonSummer             1.898e-02  5.684e-03   3.338 0.000843 ***
SeasonWinter            -6.395e-03  6.264e-03  -1.021 0.307346    
MarketSegmentCorporate   1.474e-02  2.836e-02   0.520 0.603247    
MarketSegmentDirect     -4.524e-02  2.775e-02  -1.630 0.103019    
MarketSegmentGroups      2.461e-01  2.826e-02   8.708  < 2e-16 ***
 [ reached getOption("max.print") -- omitted 317 rows ]
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3758 on 39672 degrees of freedom
Multiple R-squared:  0.3019,    Adjusted R-squared:  0.2962 
F-statistic: 52.95 on 324 and 39672 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = IsCanceled ~ LeadTime + Season + MarketSegment + 
    Country + CustomerType + Agent + VisitorType, data = H2_City)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.07357 -0.31815 -0.06041  0.30976  1.27530 

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)    
(Intercept)                -4.686e-01  2.778e-01  -1.687 0.091646 .  
LeadTime                    9.815e-04  1.665e-05  58.959  < 2e-16 ***
SeasonNA                    3.507e-01  1.126e-02  31.156  < 2e-16 ***
SeasonSpring                3.224e-02  4.105e-03   7.852 4.13e-15 ***
SeasonSummer               -8.634e-04  3.988e-03  -0.217 0.828585    
SeasonWinter                3.694e-03  4.839e-03   0.763 0.445277    
MarketSegmentComplementary -2.995e-01  3.215e-02  -9.315  < 2e-16 ***
MarketSegmentCorporate     -6.525e-02  2.805e-02  -2.327 0.019991 *  
 [ reached getOption("max.print") -- omitted 399 rows ]
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3899 on 78889 degrees of freedom
Multiple R-squared:  0.378,     Adjusted R-squared:  0.3748 
F-statistic: 118.1 on 406 and 78889 DF,  p-value: < 2.2e-16
```

Associated Rules Mining

Creating our rules for this section took a bit longer than anticipated due to not converting our variables to factors. After that was determined, the biggest challenge was determining which support and confidence levels we would use for our rules. At some points while adjusting the support and confidence level we would get thousands of rules. With that information it was difficult to determine the strongest rules within the dataset. Our rules were primarily based on cancellation and discovering the patterns that existed surrounding that. (**Excerpt6**) shows our conversion and rules created.

Excerpt6

```
H1_Resort_New <- data.frame(cancellation=as.factor(H1_Resort$IsCanceled),
                 assignedRoom=as.factor(H1_Resort$AssignedRoomType),
                 season=as.factor(H1_Resort$Season),
                 visitorType=as.factor(H1_Resort$VisitorType),
                 customerType=as.factor(H1_Resort$CustomerType),
                 distributionChannel=as.factor(H1_Resort$DistributionChannel),
                 marketSegment=as.factor(H1_Resort$MarketSegment),
                 mealType=as.factor(H1_Resort$Meal))

H1_Resort_Transactions <- as(H1_Resort_New, "transactions") #this code assigns the above factor variables to a transactions data frame
```

```
Freq_H1_Resort_Tranx <- itemFrequency(H1_Resort_Transactions)

head(sort(Freq_H1_Resort_Tranx))
tail(sort(Freq_H1_Resort_Tranx))

itemFrequencyPlot((H1_Resort_Transactions))
inspect(H1_Resort_Transactions[1:10])

### RULES 1 MODEL ####
rules1 <- apriori(H1_Resort_Transactions,
            parameter=list(supp=0.01, conf=0.55), #support and confidence
            control=list(verbose=F),
            appearance=list(default="lhs",rhs=("cancellation=1")))

inspect(rules1)
plot(rules1)

#### RULES 2 MODEL ####
rules2 <- apriori(H1_Resort_Transactions,
            parameter=list(supp=0.005, conf=0.3), #support and confidence
            control=list(verbose=F),
            appearance=list(default="lhs",rhs=("cancellation=1")))

inspect(rules2[1:10])
plot(rules2)

#### RULES 3 MODEL ####
rules3 <- apriori(H1_Resort_Transactions,
            parameter=list(supp=0.008, conf=.95),
            control=list(verbose=F), appearance=list(default="lhs",rhs=('cancellation=1')))

inspect(rules3) ### the rules and their variables are related to cancellation
plot(rules3)
```

Support Vector Machines

In this section, our team made a prediction based on the hotel data we have. We could use support vector machines (or SVM) to predict visitors will cancel the booking or not. We would use the SVM algorithm so that R will learn how to predict based on the previous hotel history. Also, gave it additional data to see how well the algorithm is performing. The challenging part in this part is the preparation of cleansing, descriptive analysis. The single NAs may directly cause less accuracy.

At first, our team trained the SVM algorithm with the whole dataset for both city and resort datasets. The (Excerpt7) will show how we used the whole dataset without NAs rows.

- SVM predicts variable IsCanceled by the whole dataset.

Excerpt7

```
resort_num <-
data.frame(IsCanceled = as.factor(H1_Resort_VisitorType_without_NA$IsCanceled),
        LeadTime = as.integer(H1_Resort_VisitorType_without_NA$LeadTime),
        StaysInWeekendNights = as.integer(H1_Resort_VisitorType_without_NA$StaysInWeekendNights),
        StaysInWeekNights = as.integer(H1_Resort_VisitorType_without_NA$StaysInWeekNights),
        Adults = as.integer(H1_Resort_VisitorType_without_NA$Adults),
        Children = as.integer(H1_Resort_VisitorType_without_NA$Children),
        Babies = as.integer(H1_Resort_VisitorType_without_NA$Babies),
        BookingChanges = as.integer(H1_Resort_VisitorType_without_NA$BookingChanges),
        ADR = as.integer(H1_Resort_VisitorType_without_NA$ADR),
        TotalOfSpecialRequests = as.integer(H1_Resort_VisitorType_without_NA$TotalOfSpecialRequests)
)

str(resort_num )
trainList <- createDataPartition(y=resort_num$IsCanceled,p=.40,list=FALSE)
trainSet <- resort_num [trainList,]
testSet <- resort_num [-trainList,]
svmOut <- ksvm(IsCanceled~., data = trainSet, kernel= "rbfdot", kpar = "automatic", C = 5, cross = 3,
prob.model = TRUE)
svmOut
svmPred <- predict(svmOut, newdata = testSet, type = "response")
confusionMatrix(svmPred, testSet$IsCanceled)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     0     1
##          0 16125  4506
##          1  1231  2166
##
##                Accuracy : 0.7612
##                  95% CI : (0.7558, 0.7666)
##     No Information Rate : 0.7223
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.2989
##
##  Mcnemar's Test P-Value : < 2.2e-16
```

From the above graph, we saw our SVM predict quite accurate results compared with the previous history. However, we spent quite a long time running the algorithm within a large amount of data. Also, our team also thought hotels may consider different types of visitors to develop their services. Based on the descriptive analysis, we separated the original dataset into three different visitor types. These three sub-datasets of visitor types can also reduce computing time and enhance accuracy.

- SVM predicts variable IsCanceled by the sub-dataset of single visitor type.
- SVM predicts variable IsCanceled by the sub-dataset of couple visitors' type.
- SVM predicts variable IsCanceled by the sub-dataset of family groups' type.

The below algorithm could run less computing time and achieve more accurate results base on the visitor types.

Excerpt8

```
resort_num <- data.frame(IsCanceled = as.factor(H1_Resort_VisitorType_Single$IsCanceled),
                         LeadTime = as.integer(H1_Resort_VisitorType_Single$LeadTime),
                         StaysInWeekendNights = as.integer(H1_Resort_VisitorType_Single$StaysInWeekendNights),
                         StaysInWeekNights = as.integer(H1_Resort_VisitorType_Single$StaysInWeekNights),
                         Adults = as.integer(H1_Resort_VisitorType_Single$Adults),
                         Children = as.integer(H1_Resort_VisitorType_Single$Children),
                         Babies = as.integer(H1_Resort_VisitorType_Single$Babies),
                         IsRepeatedGuest = as.integer(H1_Resort_VisitorType_Single$IsRepeatedGuest),
                         PreviousCancellations = as.integer(H1_Resort_VisitorType_Single$PreviousCancellations),
                         PreviousBookingsNotCanceled = as.integer(H1_Resort_VisitorType_Single$PreviousBookingsNotCanceled),

                         BookingChanges = as.integer(H1_Resort_VisitorType_Single$BookingChanges),
                         DaysInWaitingList = as.integer(H1_Resort_VisitorType_Single$DaysInWaitingList),
                         ADR = as.integer(H1_Resort_VisitorType_Single$ADR),
                         RequiredCarParkingSpaces = as.integer(H1_Resort_VisitorType_Single$RequiredCarParkingSpaces),
                         TotalOfSpecialRequests = as.integer(H1_Resort_VisitorType_Single$TotalOfSpecialRequests
)

str(resort_num, H1_Resort_VisitorType_Single)
trainList <- createDataPartition(y=resort_num$IsCanceled,p=.40,list=FALSE)
trainSet <- resort_num [trainList,]
testSet <- resort_num [-trainList,]
svmOut <- ksvm(IsCanceled~., data = trainSet, kernel= "rbfdot", kpar = "automatic", C = 5, cross = 3,
prob.model = TRUE)
svmOut
svmPred <- predict(svmOut, newdata = testSet, type = "response")
confusionMatrix(svmPred, testSet$IsCanceled)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 3391  545
##          1  107  164
##
##                Accuracy : 0.845
##                  95% CI : (0.8337, 0.8558)
##     No Information Rate : 0.8315
##     P-Value [Acc > NIR] : 0.009438
##
##                   Kappa : 0.2663
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
```

As we can see, our SVMs are quite stable and have high accuracy. The hotel could use our SVM algorithm code to make further predictions and warrant further investigation.

Prescriptive Analysis

Based primarily on the machine learning techniques used by our team, we have developed high-level recommendations to take full advantage of future revenue. Our recommendations are in line with evidence that our overall data analysis has shown.

Recommendation 1: Dynamic Pricing Based on Seasons

The datasets provided a range of dates that travelers visited our clients' hotels. Creating dynamic pricing based on seasons would generate greater revenue based on peak traveling season. Dynamic pricing signifies a step toward a more efficient revenue strategy.

Recommendation 2: Package Discounts

Our analysis has shown that there are considerable number of customers who are frequent visitors combined with less likely to cancel their reservations. Providing those customers with package discounts would contribute to more stays in the future.

Recommendation 3: Global Marketing

Both locations have demonstrated customer exposure around the world. In fact, the country that has more visitors are coming from international locations. By concentrating on these countries, we can solidify our relationship with our visitors as well as expand our influence to neighboring countries.

Recommendation 4: Overbook

With the large number of cancellations, it will be smart to overbook to an extent to help ease the value gap for those customers with less lead team between cancellations. This approach gives each hotel the advantage of anticipating cancellations while also extending our marketing foothold.

Recommendation 5: Staff Training

The analysis shows agent consideration matters. Many of the cancelled reservations were linked to several unique agent id numbers. This reenforces the notion that each employee has the power to make an impact, either negatively or positively. Preparing additional training for travel agents will work to increase additional confirmed reservations as well as prevent more future cancellations.

Team Management

Over the course of the project our team communication was a large benefit toward this analysis. Virtual meetings started mid-April on Thursdays at 3:30. As the project approached additional meetings were scheduled on a need's basis.

Summary

Over the duration of this project & primary focus was to provide an analysis that would determine what factors would generate greater revenue for our client. The Average Daily Rate was our focal point in making that happen. By conducting exploratory analysis, descriptive analysis, diagnostic analysis, predictive analysis, and prescriptive analysis we were able to provide more and more detailed information in every step. With these steps, we were able to consider the main contributors in relation to its effectiveness on ADR. Linear Models, Associated Rules Mining, and Support Vector Machines were the three machine learning techniques used to help support our recommendations.