Similarity: Classification

Sujay Vadlakonda

2023 Mar 18

I am using a dataset about hotel reservations I found here. The classification should predict whether a reservation will be cancelled or not.

Divide the data into train and test

```
df <- read.csv("hotel-reservations.csv", header=TRUE)

df$booking_status <- factor(df$booking_status)
df$type_of_meal_plan <- factor(df$type_of_meal_plan)
df$room_type_reserved <- factor(df$room_type_reserved)
df$market_segment_type <- factor(df$market_segment_type)

# Remove column "Booking_ID"
df <- df[,-1]

set.seed(1234)
i <- sample(1:nrow(df), 0.8*nrow(df), replace=FALSE)
train <- df[i,]
test <- df[-i,]
```

Explore the data statistically

Note that there is a 2:1 ratio between not cancelled reservations and cancelled reservations in the training data.

For a lot of the quantitative columns, more than 75% of the observations are 0 and have a very small mean. This means that these columns are only meaningful for a small number of observations and should be used sparingly. Some examples are: - no_of_children - required_car_parking_space - repeated_guest - no_of_previous_cancellations - no_of_previous_bookings_not_canceled

```
str(train)
```

```
## 'data.frame':    29020 obs. of  18 variables:
##  $ no_of_adults                      : int  2 2 2 1 1 2 1 2 3 2 ...
##  $ no_of_children                    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ no_of_weekend_nights              : int  2 2 2 0 1 0 1 0 0 2 ...
##  $ no_of_week_nights                 : int  5 1 2 2 0 2 2 2 1 1 ...
##  $ type_of_meal_plan                 : Factor w/ 4 levels "Meal Plan 1",..: 1 4 4 2 4 4 1 1 1 1 ..
##  $ required_car_parking_space        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ room_type_reserved                : Factor w/ 7 levels "Room_Type 1",..: 4 1 1 1 1 1 1 1 4 1 ..
##  $ lead_time                         : int  106 148 68 320 131 2 152 51 65 23 ...
##  $ arrival_year                      : int  2018 2018 2018 2018 2018 2018 2018 2018 2018 2018 ...
##  $ arrival_month                     : int  7 4 2 8 10 3 8 11 8 10 ...
##  $ arrival_date                      : int  19 23 6 18 10 24 26 4 16 9 ...
```

```
##  $ market_segment_type                : Factor w/ 5 levels "Aviation","Complementary",..: 5 5 5 4 5
##  $ repeated_guest                     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ no_of_previous_cancellations       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ no_of_previous_bookings_not_canceled: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ avg_price_per_room                 : num  121.4 61.6 51.1 90 108 ...
##  $ no_of_special_requests             : int  0 0 0 0 0 1 0 0 2 0 ...
##  $ booking_status                     : Factor w/ 2 levels "Canceled","Not_Canceled": 1 2 2 2 1 2 1
```

```
names(train)
```

```
##  [1] "no_of_adults"
##  [2] "no_of_children"
##  [3] "no_of_weekend_nights"
##  [4] "no_of_week_nights"
##  [5] "type_of_meal_plan"
##  [6] "required_car_parking_space"
##  [7] "room_type_reserved"
##  [8] "lead_time"
##  [9] "arrival_year"
## [10] "arrival_month"
## [11] "arrival_date"
## [12] "market_segment_type"
## [13] "repeated_guest"
## [14] "no_of_previous_cancellations"
## [15] "no_of_previous_bookings_not_canceled"
## [16] "avg_price_per_room"
## [17] "no_of_special_requests"
## [18] "booking_status"
```

```
dim(train)
```

```
## [1] 29020    18
```

```
head(train)
```

```
##       no_of_adults no_of_children no_of_weekend_nights no_of_week_nights
## 15241            2              0                    2                 5
## 33702            2              0                    2                 1
## 35716            2              0                    2                 2
## 17487            1              0                    0                 2
## 15220            1              0                    1                 0
## 19838            2              0                    0                 2
##       type_of_meal_plan required_car_parking_space room_type_reserved lead_time
## 15241       Meal Plan 1                          0        Room_Type 4       106
## 33702      Not Selected                          0        Room_Type 1       148
## 35716      Not Selected                          0        Room_Type 1        68
## 17487       Meal Plan 2                          0        Room_Type 1       320
## 15220      Not Selected                          0        Room_Type 1       131
## 19838      Not Selected                          0        Room_Type 1         2
##       arrival_year arrival_month arrival_date market_segment_type
## 15241         2018             7           19              Online
## 33702         2018             4           23              Online
## 35716         2018             2            6              Online
## 17487         2018             8           18             Offline
## 15220         2018            10           10              Online
## 19838         2018             3           24              Online
```

```
##       repeated_guest no_of_previous_cancellations
## 15241              0                            0
## 33702              0                            0
## 35716              0                            0
## 17487              0                            0
## 15220              0                            0
## 19838              0                            0
##       no_of_previous_bookings_not_canceled avg_price_per_room
## 15241                                    0             121.37
## 33702                                    0              61.56
## 35716                                    0              51.09
## 17487                                    0              90.00
## 15220                                    0             108.00
## 19838                                    0             134.00
##       no_of_special_requests booking_status
## 15241                      0       Canceled
## 33702                      0   Not_Canceled
## 35716                      0   Not_Canceled
## 17487                      0   Not_Canceled
## 15220                      0       Canceled
## 19838                      1   Not_Canceled
```

```
summary(train)
```

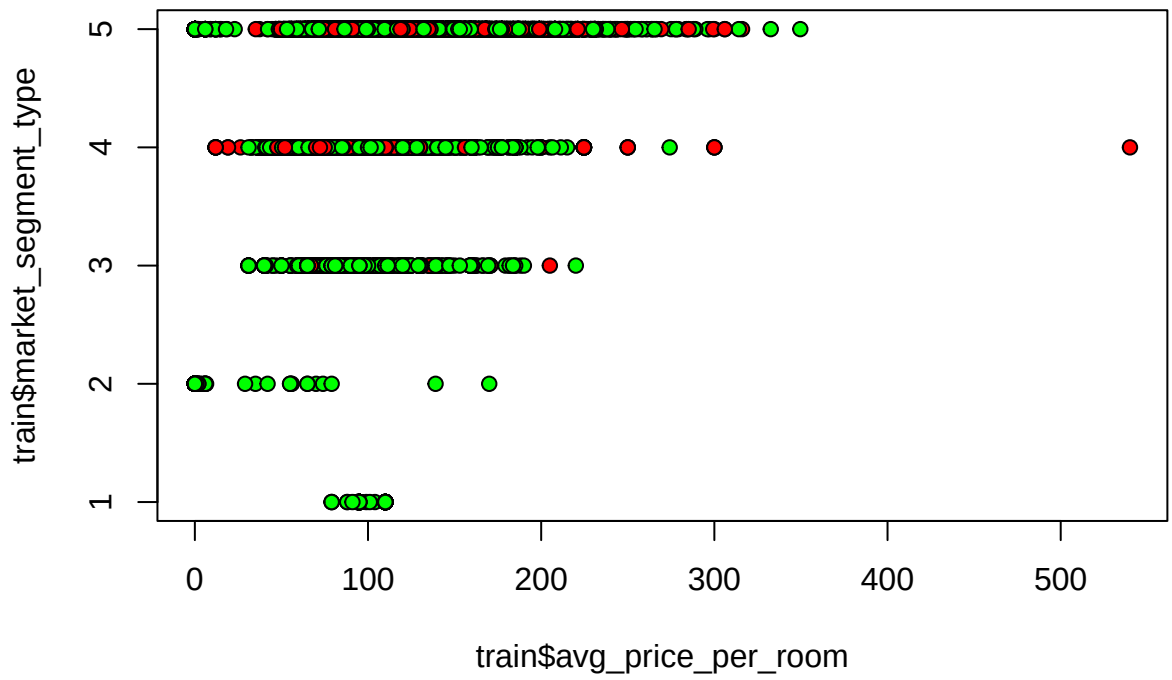
```
##   no_of_adults   no_of_children   no_of_weekend_nights no_of_week_nights
##  Min.   :0.000   Min.   :0.0000   Min.   :0.0000       Min.   : 0.000
##  1st Qu.:2.000   1st Qu.:0.0000   1st Qu.:0.0000       1st Qu.: 1.000
##  Median :2.000   Median :0.0000   Median :1.0000       Median : 2.000
##  Mean   :1.845   Mean   :0.1063   Mean   :0.8106       Mean   : 2.206
##  3rd Qu.:2.000   3rd Qu.:0.0000   3rd Qu.:2.0000       3rd Qu.: 3.000
##  Max.   :4.000   Max.   :9.0000   Max.   :7.0000       Max.   :17.000
##
##     type_of_meal_plan required_car_parking_space   room_type_reserved
##  Meal Plan 1 :22245   Min.   :0.00000            Room_Type 1:22541
##  Meal Plan 2 : 2674   1st Qu.:0.00000            Room_Type 2:  548
##  Meal Plan 3 :    3   Median :0.00000            Room_Type 3:    6
##  Not Selected: 4098   Mean   :0.03032            Room_Type 4: 4814
##                       3rd Qu.:0.00000            Room_Type 5:  214
##                       Max.   :1.00000            Room_Type 6:  772
##                                                  Room_Type 7:  125
##    lead_time       arrival_year  arrival_month     arrival_date
##  Min.   :  0.00   Min.   :2017   Min.   : 1.000   Min.   : 1.00
##  1st Qu.: 17.00   1st Qu.:2018   1st Qu.: 5.000   1st Qu.: 8.00
##  Median : 57.00   Median :2018   Median : 8.000   Median :16.00
##  Mean   : 85.08   Mean   :2018   Mean   : 7.434   Mean   :15.59
##  3rd Qu.:126.00   3rd Qu.:2018   3rd Qu.:10.000   3rd Qu.:23.00
##  Max.   :443.00   Max.   :2018   Max.   :12.000   Max.   :31.00
##
##     market_segment_type repeated_guest    no_of_previous_cancellations
##  Aviation     :  101    Min.   :0.00000   Min.   : 0.00000
##  Complementary:  313    1st Qu.:0.00000   1st Qu.: 0.00000
##  Corporate    : 1625    Median :0.00000   Median : 0.00000
##  Offline      : 8457    Mean   :0.02564   Mean   : 0.02123
##  Online       :18524    3rd Qu.:0.00000   3rd Qu.: 0.00000
##                         Max.   :1.00000   Max.   :13.00000
```

```
##
##  no_of_previous_bookings_not_canceled avg_price_per_room no_of_special_requests
##  Min.   : 0.0000                      Min.   :  0.00     Min.   :0.0000
##  1st Qu.: 0.0000                      1st Qu.: 80.30     1st Qu.:0.0000
##  Median : 0.0000                      Median : 99.45     Median :0.0000
##  Mean   : 0.1537                      Mean   :103.40     Mean   :0.6167
##  3rd Qu.: 0.0000                      3rd Qu.:120.00     3rd Qu.:1.0000
##  Max.   :58.0000                      Max.   :540.00     Max.   :5.0000
##
##       booking_status
##  Canceled    : 9507
##  Not_Canceled:19513
##
##
##
##
##
```

Explore the data graphically

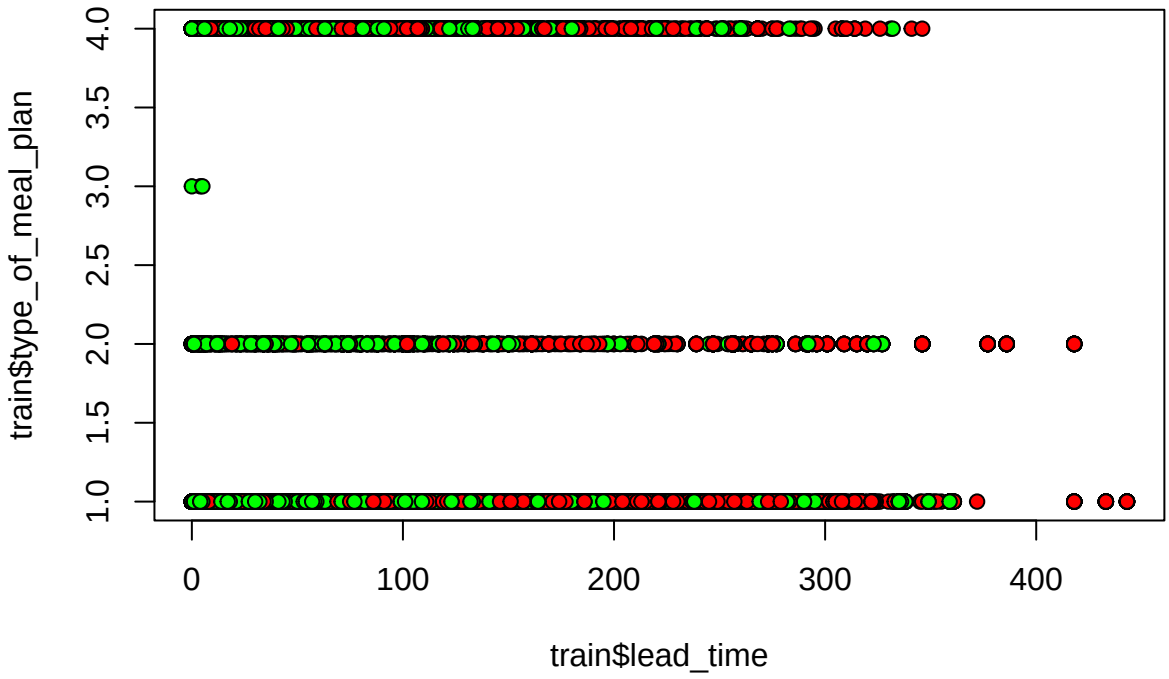
Market Segments 1-3 to almost always do not cancel Average Room Price does not seem to affect booking status

```
plot(train$avg_price_per_room, train$market_segment_type, pch=21, bg=c("red", "green")[train$booking_sta
```



A larger lead time correlates with booking cancellation Not selecting a meal plan (Meal Plan 4) also correlates with cancellation

```
plot(train$lead_time, train$type_of_meal_plan, pch=21, bg=c("red", "green")[train$booking_status])
```



Logistic Regression

```
logistic_regression_model <- glm(booking_status ~ avg_price_per_room
                                                + market_segment_type
                                                + lead_time
                                                + type_of_meal_plan,
                                                data=train,
                                                family="binomial")
logistic_regression_model
```

```
##
## Call:  glm(formula = booking_status ~ avg_price_per_room + market_segment_type +
##     lead_time + type_of_meal_plan, family = "binomial", data = train)
##
## Coefficients:
##                       (Intercept)                avg_price_per_room
##                           2.05690                           -0.01159
## market_segment_typeComplementary      market_segment_typeCorporate
##                          13.75913                            1.46674
##       market_segment_typeOffline         market_segment_typeOnline
##                           1.91383                           1.01004
##                         lead_time      type_of_meal_planMeal Plan 2
##                          -0.01410                           -0.12933
##     type_of_meal_planMeal Plan 3     type_of_meal_planNot Selected
##                          -0.20744                           -0.38803
##
## Degrees of Freedom: 29019 Total (i.e. Null);  29010 Residual
## Null Deviance:       36710
## Residual Deviance: 29120     AIC: 29140
```

```
summary(logistic_regression_model)
```

```
##
```

```
## Call:
## glm(formula = booking_status ~ avg_price_per_room + market_segment_type +
##     lead_time + type_of_meal_plan, family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5685  -0.7998   0.5050   0.7705   2.4223
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       2.057e+00  2.215e-01   9.288  < 2e-16 ***
## avg_price_per_room               -1.159e-02  4.769e-04 -24.311  < 2e-16 ***
## market_segment_typeComplementary  1.376e+01  8.103e+01   0.170   0.8652
## market_segment_typeCorporate      1.467e+00  2.317e-01   6.332 2.43e-10 ***
## market_segment_typeOffline        1.914e+00  2.204e-01   8.685  < 2e-16 ***
## market_segment_typeOnline         1.010e+00  2.179e-01   4.636 3.55e-06 ***
## lead_time                        -1.410e-02  2.064e-04 -68.346  < 2e-16 ***
## type_of_meal_planMeal Plan 2     -1.293e-01  5.548e-02  -2.331   0.0197 *
## type_of_meal_planMeal Plan 3     -2.074e-01  8.441e+02   0.000   0.9998
## type_of_meal_planNot Selected    -3.880e-01  4.257e-02  -9.115  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 36708  on 29019  degrees of freedom
## Residual deviance: 29125  on 29010  degrees of freedom
## AIC: 29145
##
## Number of Fisher Scoring iterations: 14
```

Our logistic regression model has an accuracy of 17%

```
library(caret)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```
logistic_probabilities <- predict(logistic_regression_model, newdata=test, type="response")
logistic_predictions <- ifelse(logistic_probabilities>0.5, 1, 0)
logistic_accuracy <- mean(logistic_predictions==as.integer(test$booking_status))
print(paste("logisitic accuracy = ", logistic_accuracy))
```

```
## [1] "logisitic accuracy =  0.172846312887664"
```

kNN Classification

```
library(class)

# Remove target columns
knn_train = train[, 1:17]
knn_test = test[, 1:17]

for(column in 1:17) {
  knn_train[, column] <- as.integer(knn_train[, column])
```

```
  knn_test[, column] <- as.integer(knn_test[, column])
}

# Scale
means <- sapply(knn_train, mean)
stdevs <- sapply(knn_train, sd)
knn_train <- scale(knn_train, center=means, scale=stdevs)
knn_test <- scale(knn_test, center=means, scale=stdevs)

knn_train_labels = train[, 18]
knn_test_labels = test[, 18]

knn_predictions <- knn(train = knn_train,
                       test = knn_test,
                       cl = knn_train_labels,
                       k=3)
```

Our scaled kNN has an accuracy of 84.4%. Unscaled gave 80.5%. k=5 made the results worse.

```
knn_results <- knn_predictions == knn_test_labels
knn_accuracy <- length(which(knn_results == TRUE)) / length(knn_results)
knn_accuracy
```

```
## [1] 0.8443832
```

Decision Trees Classification

The decision trees show that a market_segment_type of "Online" greatly increases the chances of "not canceled". Additionally, having 1 or more special requests greatly increases the chances of "not canceled".

```
library(tree)
tree_bookings <- tree(booking_status~., data=train)
tree_bookings
```

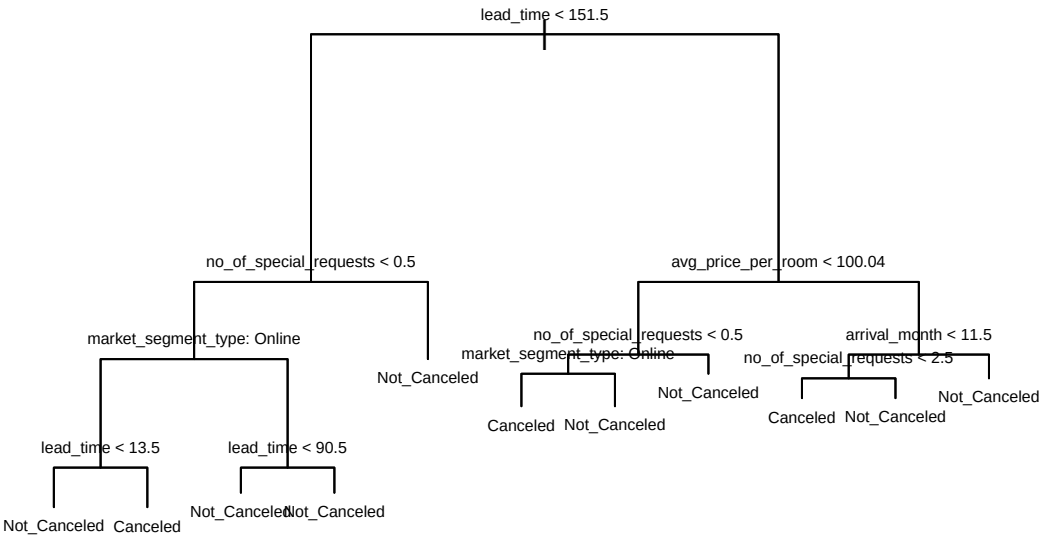
```
## node), split, n, deviance, yval, (yprob)
##       * denotes terminal node
##
##  1) root 29020 36710.00 Not_Canceled ( 0.32760 0.67240 )
##    2) lead_time < 151.5 23315 25190.00 Not_Canceled ( 0.23084 0.76916 )
##      4) no_of_special_requests < 0.5 12241 15530.00 Not_Canceled ( 0.33012 0.66988 )
##        8) market_segment_type: Online 6002  8306.00 Canceled ( 0.52433 0.47567 )
##         16) lead_time < 13.5 1635  1794.00 Not_Canceled ( 0.23792 0.76208 ) *
##         17) lead_time > 13.5 4367  5748.00 Canceled ( 0.63155 0.36845 ) *
##        9) market_segment_type: Aviation,Complementary,Corporate,Offline 6239  5127.00 Not_Canceled (
##         18) lead_time < 90.5 4844  2852.00 Not_Canceled ( 0.08650 0.91350 ) *
##         19) lead_time > 90.5 1395  1789.00 Not_Canceled ( 0.34050 0.65950 ) *
##      5) no_of_special_requests > 0.5 11074  8175.00 Not_Canceled ( 0.12109 0.87891 ) *
##    3) lead_time > 151.5 5705  6732.00 Canceled ( 0.72305 0.27695 )
##      6) avg_price_per_room < 100.04 3180  4388.00 Canceled ( 0.53962 0.46038 )
##       12) no_of_special_requests < 0.5 2214  2865.00 Canceled ( 0.65086 0.34914 )
##         24) market_segment_type: Online 711   153.00 Canceled ( 0.97750 0.02250 ) *
##         25) market_segment_type: Corporate,Offline 1503  2084.00 Not_Canceled ( 0.49634 0.50366 ) *
##       13) no_of_special_requests > 0.5 966  1154.00 Not_Canceled ( 0.28468 0.71532 ) *
##      7) avg_price_per_room > 100.04 2525   941.20 Canceled ( 0.95406 0.04594 )
##       14) arrival_month < 11.5 2426   383.00 Canceled ( 0.98475 0.01525 )
##         28) no_of_special_requests < 2.5 2389     0.00 Canceled ( 1.00000 0.00000 ) *
```

```
##         29) no_of_special_requests > 2.5 37     0.00 Not_Canceled ( 0.00000 1.00000 ) *
##       15) arrival_month > 11.5 99    99.63 Not_Canceled ( 0.20202 0.79798 ) *
```

```
summary(tree_bookings)
```

```
##
## Classification tree:
## tree(formula = booking_status ~ ., data = train)
## Variables actually used in tree construction:
## [1] "lead_time"              "no_of_special_requests" "market_segment_type"
## [4] "avg_price_per_room"     "arrival_month"
## Number of terminal nodes:  11
## Residual mean deviance:  0.8221 = 23850 / 29010
## Misclassification error rate: 0.1823 = 5290 / 29020
```

```
plot(tree_bookings)
text(tree_bookings, cex=0.5, pretty=0)
```



Decision Trees resulted in a 82% accuracy.

```
tree_predictions <- predict(tree_bookings, newdata=test, type="class")
table(tree_predictions, test$booking_status)
```

```
##
## tree_predictions Canceled Not_Canceled
##     Canceled         1480          408
##     Not_Canceled      898         4469
```

```
mean(tree_predictions==test$booking_status)
```

```
## [1] 0.8199862
```

Result Comparison

Logisitic Regression: 17% k Nearest Neighbors: 84.4% Decision Trees: 82%

Logistic Regression performed by far the worst on this data. kNN had the highest accuracy, which was closely followed by Decision Trees.

Result Analysis

Logistic Regression performed the worst on this dataset. This is because logistic regression is a high bias algorithm that provides great results on a dataset that can be divided by a line. Simply dividing all the data points by a line will not be effective for this dataset because the distribution of cancelled reservations are not linear.

kNN performed the best on this dataset. kNN is a high variance algorithm that can tightly fit datasets. It works especially well for this dataset, because the cancelled reservations occur at specific parts of the multidimensional graph. Decision Trees performed a little bit worse than kNN. This occurred because decision trees generally sacrifice some accuracy for increased interpretability. The decision trees accuracy could have been further increased by overfitting and adding more decisions, but this would reduce its interpretability.