Classification.Rmd

```
hotel_bookings <- read.csv('C:/Users/colto/Documents/RegressionClassification/hotel_bookings.csv')
```

Source: https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand

a. First we split the data set into 80/20 train/test.

```
# dividing the dataset into 80/20
n = nrow(hotel_bookings)
trainIndex = sample(1:n, size=round(0.8*n), replace=FALSE)
# assign train/test 80/20
train = hotel_bookings[trainIndex ,]
test = hotel_bookings[-trainIndex ,]
```

b. We can now explore the training data to verify it's contents.

```
head(train) # Print the first six rows
```

```
##              hotel is_canceled lead_time arrival_date_year arrival_date_month
## 45705    City Hotel           0       140              2015           November
## 57527    City Hotel           1        22              2016          September
## 61183    City Hotel           1        32              2016           December
## 68778    City Hotel           1       113              2017                May
## 18901  Resort Hotel           0         0              2016            January
## 108338   City Hotel           0        17              2017              March
##        arrival_date_week_number arrival_date_day_of_month
## 45705                        47                         18
## 57527                        40                         26
## 61183                        49                          3
## 68778                        20                         20
## 18901                         3                         13
## 108338                       12                         19
##        stays_in_weekend_nights stays_in_week_nights adults children babies meal
## 45705                        0                    3      2        0      0   BB
## 57527                        1                    3      1        0      0   BB
## 61183                        2                    2      2        0      0   BB
## 68778                        2                    1      1        0      0   BB
## 18901                        0                    2      2        0      0   BB
## 108338                       2                    5      2        0      0   BB
##        country market_segment distribution_channel is_repeated_guest
## 45705      PRT         Groups                TA/TO                 0
## 57527      PRT      Online TA                TA/TO                 0
## 61183      BEL      Online TA                TA/TO                 0
## 68778      PRT  Offline TA/TO                TA/TO                 0
```

```
## 18901      PRT         Direct               Direct                 1
## 108338     FRA  Offline TA/TO                TA/TO                 0
##        previous_cancellations previous_bookings_not_canceled reserved_room_type
## 45705                       0                               0                 A
## 57527                       0                               0                 A
## 61183                       0                               0                 A
## 68778                       0                               0                 A
## 18901                       0                               2                 A
## 108338                      0                               0                 A
##        assigned_room_type booking_changes deposit_type agent company
## 45705                   A               0   No Deposit    29    NULL
## 57527                   A               0   No Deposit    86    NULL
## 61183                   A               0   No Deposit     9    NULL
## 68778                   A               0   Non Refund    34    NULL
## 18901                   F               1   No Deposit  NULL     110
## 108338                  A               1   No Deposit    20    NULL
##        days_in_waiting_list   customer_type   adr required_car_parking_spaces
## 45705                    87 Transient-Party  75.0                           0
## 57527                     0       Transient 106.4                           0
## 61183                     0       Transient 104.0                           0
## 68778                     0       Transient 190.0                           0
## 18901                     0       Transient  38.0                           0
## 108338                    0 Transient-Party  75.0                           0
##        total_of_special_requests reservation_status reservation_status_date
## 45705                          1          Check-Out              2015-11-21
## 57527                          0           Canceled              2016-09-07
## 61183                          0           Canceled              2016-11-17
## 68778                          0           Canceled              2017-01-27
## 18901                          0          Check-Out              2016-01-15
## 108338                         1          Check-Out              2017-03-26
```

```
names(train) # Get column names
```

```
##  [1] "hotel"                          "is_canceled"
##  [3] "lead_time"                      "arrival_date_year"
##  [5] "arrival_date_month"             "arrival_date_week_number"
##  [7] "arrival_date_day_of_month"      "stays_in_weekend_nights"
##  [9] "stays_in_week_nights"           "adults"
## [11] "children"                       "babies"
## [13] "meal"                           "country"
## [15] "market_segment"                 "distribution_channel"
## [17] "is_repeated_guest"              "previous_cancellations"
## [19] "previous_bookings_not_canceled" "reserved_room_type"
## [21] "assigned_room_type"             "booking_changes"
## [23] "deposit_type"                   "agent"
## [25] "company"                        "days_in_waiting_list"
## [27] "customer_type"                  "adr"
## [29] "required_car_parking_spaces"    "total_of_special_requests"
## [31] "reservation_status"             "reservation_status_date"
```

```
dim(train) # Get number of rows & columns
```

```
## [1] 95512    32
```

```
colSums(is.na(train)) # Get number of missing values
```

```
##                          hotel                    is_canceled
##                              0                              0
##                      lead_time              arrival_date_year
##                              0                              0
##             arrival_date_month       arrival_date_week_number
##                              0                              0
##      arrival_date_day_of_month        stays_in_weekend_nights
##                              0                              0
##           stays_in_week_nights                         adults
##                              0                              0
##                       children                         babies
##                              4                              0
##                           meal                        country
##                              0                              0
##                 market_segment           distribution_channel
##                              0                              0
##              is_repeated_guest         previous_cancellations
##                              0                              0
## previous_bookings_not_canceled             reserved_room_type
##                              0                              0
##             assigned_room_type                booking_changes
##                              0                              0
##                   deposit_type                          agent
##                              0                              0
##                        company           days_in_waiting_list
##                              0                              0
##                 customer_type                            adr
##                              0                              0
##    required_car_parking_spaces       total_of_special_requests
##                              0                              0
##             reservation_status         reservation_status_date
##                              0                              0
```

```
str(train) # Get structure of variables
```

```
## 'data.frame':    95512 obs. of  32 variables:
##  $ hotel                         : chr  "City Hotel" "City Hotel" "City Hotel" "City Hotel" ...
##  $ is_canceled                   : int  0 1 1 1 0 0 1 0 1 0 ...
##  $ lead_time                     : int  140 22 32 113 0 17 60 20 279 6 ...
##  $ arrival_date_year             : int  2015 2016 2016 2017 2016 2017 2016 2017 2015 2015 ...
##  $ arrival_date_month            : chr  "November" "September" "December" "May" ...
##  $ arrival_date_week_number      : int  47 40 49 20 3 12 52 25 41 40 ...
##  $ arrival_date_day_of_month     : int  18 26 3 20 13 19 23 19 8 28 ...
##  $ stays_in_weekend_nights       : int  0 1 2 2 0 2 0 1 2 1 ...
##  $ stays_in_week_nights          : int  3 3 2 1 2 5 1 0 5 0 ...
##  $ adults                        : int  2 1 2 1 2 2 2 1 2 1 ...
##  $ children                      : int  0 0 0 0 0 0 2 0 0 0 ...
##  $ babies                        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ meal                          : chr  "BB" "BB" "BB" "BB" ...
##  $ country                       : chr  "PRT" "PRT" "BEL" "PRT" ...
```

```
##  $ market_segment                : chr  "Groups" "Online TA" "Online TA" "Offline TA/TO" ...
##  $ distribution_channel          : chr  "TA/TO" "TA/TO" "TA/TO" "TA/TO" ...
##  $ is_repeated_guest             : int  0 0 0 0 1 0 0 0 0 0 ...
##  $ previous_cancellations        : int  0 0 0 0 0 0 0 0 1 0 ...
##  $ previous_bookings_not_canceled: int  0 0 0 0 2 0 0 0 0 0 ...
##  $ reserved_room_type            : chr  "A" "A" "A" "A" ...
##  $ assigned_room_type            : chr  "A" "A" "A" "A" ...
##  $ booking_changes               : int  0 0 0 0 1 1 0 0 0 0 ...
##  $ deposit_type                  : chr  "No Deposit" "No Deposit" "No Deposit" "Non Refund" ...
##  $ agent                         : chr  "29" "86" "9" "34" ...
##  $ company                       : chr  "NULL" "NULL" "NULL" "NULL" ...
##  $ days_in_waiting_list          : int  87 0 0 0 0 0 0 0 0 0 ...
##  $ customer_type                 : chr  "Transient-Party" "Transient" "Transient" "Transient" ...
##  $ adr                           : num  75 106 104 190 38 ...
##  $ required_car_parking_spaces   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ total_of_special_requests     : int  1 0 0 0 0 1 1 0 0 0 ...
##  $ reservation_status            : chr  "Check-Out" "Canceled" "Canceled" "Canceled" ...
##  $ reservation_status_date       : chr  "2015-11-21" "2016-09-07" "2016-11-17" "2017-01-27" ...
```

```
summary(train) # Basic descriptive statistics
```

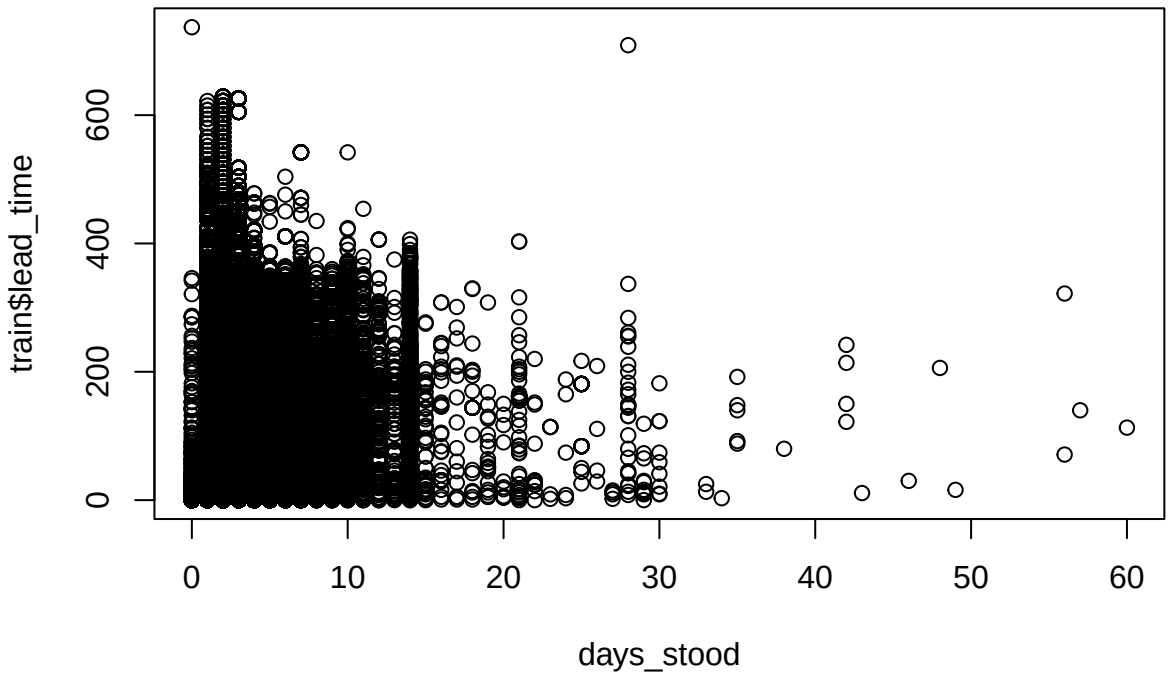
```
##     hotel            is_canceled       lead_time   arrival_date_year
##  Length:95512       Min.   :0.0000   Min.   :  0   Min.   :2015
##  Class :character   1st Qu.:0.0000   1st Qu.: 18   1st Qu.:2016
##  Mode  :character   Median :0.0000   Median : 69   Median :2016
##                     Mean   :0.3708   Mean   :104   Mean   :2016
##                     3rd Qu.:1.0000   3rd Qu.:161   3rd Qu.:2017
##                     Max.   :1.0000   Max.   :737   Max.   :2017
##
##  arrival_date_month arrival_date_week_number arrival_date_day_of_month
##  Length:95512       Min.   : 1.00            Min.   : 1.00
##  Class :character   1st Qu.:16.00            1st Qu.: 8.00
##  Mode  :character   Median :28.00            Median :16.00
##                     Mean   :27.18            Mean   :15.81
##                     3rd Qu.:38.00            3rd Qu.:23.00
##                     Max.   :53.00            Max.   :31.00
##
##  stays_in_weekend_nights stays_in_week_nights     adults          children
##  Min.   : 0.0000         Min.   : 0.000       Min.   : 0.000   Min.   : 0.000
##  1st Qu.: 0.0000         1st Qu.: 1.000       1st Qu.: 2.000   1st Qu.: 0.000
##  Median : 1.0000         Median : 2.000       Median : 2.000   Median : 0.000
##  Mean   : 0.9287         Mean   : 2.502       Mean   : 1.856   Mean   : 0.103
##  3rd Qu.: 2.0000         3rd Qu.: 3.000       3rd Qu.: 2.000   3rd Qu.: 0.000
##  Max.   :18.0000         Max.   :42.000       Max.   :50.000   Max.   :10.000
##                                                                NA's   :4
##      babies              meal              country          market_segment
##  Min.   : 0.000000   Length:95512       Length:95512       Length:95512
##  1st Qu.: 0.000000   Class :character   Class :character   Class :character
##  Median : 0.000000   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 0.007915
##  3rd Qu.: 0.000000
##  Max.   :10.000000
##
##  distribution_channel is_repeated_guest previous_cancellations
```

```
##  Length:95512         Min.   :0.0000    Min.   : 0.00000
##  Class :character     1st Qu.:0.0000    1st Qu.: 0.00000
##  Mode  :character     Median :0.0000    Median : 0.00000
##                       Mean   :0.0315    Mean   : 0.08621
##                       3rd Qu.:0.0000    3rd Qu.: 0.00000
##                       Max.   :1.0000    Max.   :26.00000
##
##  previous_bookings_not_canceled reserved_room_type assigned_room_type
##  Min.   : 0.0000                Length:95512       Length:95512
##  1st Qu.: 0.0000                Class :character   Class :character
##  Median : 0.0000                Mode  :character   Mode  :character
##  Mean   : 0.1352
##  3rd Qu.: 0.0000
##  Max.   :72.0000
##
##  booking_changes   deposit_type          agent             company
##  Min.   : 0.0000   Length:95512       Length:95512       Length:95512
##  1st Qu.: 0.0000   Class :character   Class :character   Class :character
##  Median : 0.0000   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 0.2199
##  3rd Qu.: 0.0000
##  Max.   :21.0000
##
##  days_in_waiting_list customer_type           adr
##  Min.   :  0.000      Length:95512       Min.   : -6.38
##  1st Qu.:  0.000      Class :character   1st Qu.: 69.30
##  Median :  0.000      Mode  :character   Median : 94.80
##  Mean   :  2.315                         Mean   :101.79
##  3rd Qu.:  0.000                         3rd Qu.:126.00
##  Max.   :391.000                         Max.   :510.00
##
##  required_car_parking_spaces total_of_special_requests reservation_status
##  Min.   :0.0000              Min.   :0.0000            Length:95512
##  1st Qu.:0.0000              1st Qu.:0.0000            Class :character
##  Median :0.0000              Median :0.0000            Mode  :character
##  Mean   :0.0633              Mean   :0.5712
##  3rd Qu.:0.0000              3rd Qu.:1.0000
##  Max.   :8.0000              Max.   :5.0000
##
##  reservation_status_date
##  Length:95512
##  Class :character
##  Mode  :character
##
##
##
##
```

c. We can also create informative graphs using the data.

Is there some sort of relationship between the number of days stood and the lead time?

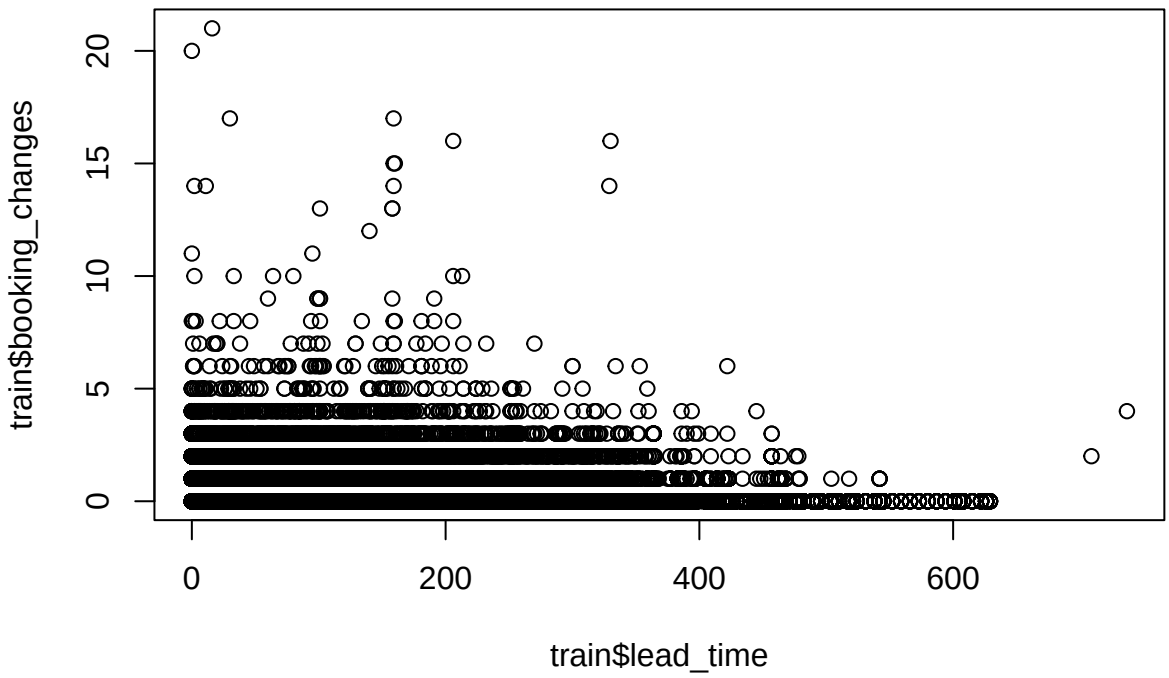
```
# Add stays_in_week_nights and stays_in_weekend_nights
days_stood = train$stays_in_week_nights+train$stays_in_weekend_nights
# Plot days_stood v lead_time
plot(days_stood, train$lead_time)
```



Right now, it's difficult to tell if there is due to the density of the points. This isn't suprising, since we can expect the vast majority of points to be distributed towards x=0 as extremely long stays at a hotel are unusual and uncommon. However, we can observe that those who give 400+ days of lead time tend to not stay for very long.(Less than 10 days) We can also observe that those who stay for very long (10 or more days) tend to give a year or less of lead time. My hypothesis is that there is an inverse correlation between the number of days stood and the lead time. We will build a linear regression model to test this hypothesis later.

Is there some sort of relationship between the lead time and the number of booking changes?

```
# Add stays_in_week_nights and stays_in_weekend_nights
stays = train$stays_in_week_nights+train$stays_in_weekend_nights
# Plot lead_time v booking_changes
plot(train$lead_time, train$booking_changes)
```



Right now, it's difficult to tell if there is due to the density of the points, but the outliers do provide a hint that there may be an inverse correlation between the lead time and the number of booking changes. We won't analyze this relationship any further as it is not required, but my hypothesis is that there is no correlation between these two attributes.

d. We'll make a logistic regression model for the relationship between the number of days stood and is_repeated_guest.

```
# Add stays_in_week_nights and stays_in_weekend_nights
stays = train$stays_in_week_nights+train$stays_in_weekend_nights
# lm between lead_time and stays
results1 <- glm(is_repeated_guest ~ stays, data=train)
# Display summary of the results
summary(results1)
```

```
##
## Call:
## glm(formula = is_repeated_guest ~ stays, data = train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.05634  -0.04186  -0.03462  -0.02738   1.14636
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  0.0563396  0.0009392   59.98   <2e-16 ***
## stays       -0.0072392  0.0002193  -33.00   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.030168)
##
##     Null deviance: 2914.2  on 95511  degrees of freedom
## Residual deviance: 2881.3  on 95510  degrees of freedom
## AIC: -63330
##
## Number of Fisher Scoring iterations: 2
```

Although the sum of our residuals approaches 0, our residuals do not reflect eachother. We can see that there's a negative correlation between the length of a booking's stay and the likelihood someone is a repeated guest. According to the coefficient estimate, each day of the stay decreases the likelihood that someone is a repeated guest by 0.7%. Since the P-values is considerably lower than 0.05, we can say that the amount of days stood is a significant factor to likelihood someone is a repeated guest, though the low magnitude of the coefficient estimate makes me think that this might not be true.

e. We'll make a naive Bayes model for the relationship between the number of days stood and is_repeated_guest. First we need to remove the non-numeric columns from our data set.

```
# removing non-numeric columns
test_copy = subset(test, select=-c(hotel,arrival_date_month,meal,country,market_segment,distribution_cha
train_copy = subset(train, select=-c(hotel,arrival_date_month,meal,country,market_segment,distribution_
library(e1071) #e1071 library required for naiveBayes
# Add stays_in_week_nights and stays_in_weekend_nights
stays = train_copy$stays_in_week_nights+train$stays_in_weekend_nights
# creating naiveBayes matrix
model=naiveBayes(is_repeated_guest~., data=train_copy)
pred = predict(model,test_copy)
table(pred, test_copy$is_repeated_guest, dnn=c("Predic","Actual"))
```

```
##       Actual
## Predic     0     1
##      0 22157   279
##      1   920   522
```

The model is predicting 22096 + 509 observations correctly. The model is predicting 275 + 998 observations incorrectly. The model has a bias towards predicting 1's incorrectly more than 1's correctly, but this could be attributed to the scarcity of 1's in the data set.

TP = 22096, FP = 275, FN = 998, TN = 509

f. Classification Metrics:

```
# The algorithm for Kappa implies we know the actual agreement, but we do not. This would be simple to
# ROC and AUC algorithms not given during class.
print("Sensitivity:")
```

```
## [1] "Sensitivity:"
```

```
print(22096/(22096+998))
```

```
## [1] 0.9567853
```

```
print("Specificity:")
```

```
## [1] "Specificity:"
```

```
print(509/(509+275))
```

```
## [1] 0.6492347
```

```
print("Precision:")
```

```
## [1] "Precision:"
```

```
print(22096/(22096+275))
```

```
## [1] 0.9877073
```

```
print("Recall:")
```

```
## [1] "Recall:"
```

```
print(22096/(22096+998))
```

```
## [1] 0.9567853
```

```
print("MCC:")
```

```
## [1] "MCC:"
```

```
print(((22096*509)-(275*998))/sqrt((22096+275)*(22096+998)*(22096+275)*(22096+509)))
```

```
## [1] 0.02146668
```

We have high values for sensitivity, specificity, precision, and recall. MCC and Specificity suffers because our model tends to mispredict 1's. This is to be expected because the duration of the stay isn't a strong predictor of repeated guests.

g. Both algorithms are great for classification problems, but Naive Bayes's weakness lies in the assumption that each feature is independent, which is not always the case. In such cases, linear regression is preferable because it splits feature space linearly which is ideal for dealing with correlated variables. However, logistic regression struggles with small training data as it may over fit the data. However, Naive Bayes works well with small training data because estimates are based on the joint density function.

h. Accuracy is a straightforward metric that measures the percentage of correctly classified instances. However, it can be heavily influenced by the majority class.

Precision measures the percentage of correctly predicted positive instances out of all positive predictions. This method of measurement is great when the cost of false positives is high, but it fails to take into account false negatives.

Recall measures the percentage of correctly predicted positive instances out of all positive instances. This method of measurement is great when the cost of false positives is negatives, but it fails to take into account false positives.

ROC AUC measures the area under the receiver operating characteristic curve, which plots the true positive rate against the false positive rate. It is great for imbalanced data sets, but it is not the best metric when the cost of false positives and false negatives are inequal.

The MCC classification metric measures the quality of binary classifications. One of the benefits of the MCC metric is that it is not affected by imbalanced data sets, MCC can also be used to compare the performance of different machine learning models as it takes into account both true and false predictions. However, MCC is only applicable to binary classification problems and cannot be used for multicast classification problems. MCC can also be affected by a lack of variability in data, leading to biased predictions.

The Kappa classification measures the agreement between predicted and actual classes. It is great for imbalanced data sets, as well as evaluating the performance of classifiers when the distribution is not uniform. It is widely considered to be more robust than other classifications, but it is not ideal for cases where classes are not mutually exclusive.