Classification

Sujay Vadlakonda

2023 Feb 18

1. Linear Models for Classification

Write a paragraph explaining in general terms how linear models for classification work, and what are the strengths and weaknesses of these linear models. A linear regression attempts to find a line that minimizes the distance between each of the data points graphed with predictors on the x axis and target column on the y axis. A linear regression seeks to predict a quantitative target. Linear regression is a high bias algorithm, which means it is prone to underfitting. A linear regression wants to see a linear relationship between the target and its predictors and cannot observe other relationships.

2. Load Data

I am using a dataset about hotel reservations I found here.

```
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)
```

2a. Create Test and Train Data

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

2b. Training Data Exploration

```
str(train)
```

```
## 'data.frame':    29020 obs. of  19 variables:
##  $ Booking_ID                          : chr  "INN15241" "INN33702" "INN35716" "INN17487" ...
##  $ 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] "Booking_ID"
##  [2] "no_of_adults"
##  [3] "no_of_children"
##  [4] "no_of_weekend_nights"
##  [5] "no_of_week_nights"
##  [6] "type_of_meal_plan"
##  [7] "required_car_parking_space"
##  [8] "room_type_reserved"
##  [9] "lead_time"
## [10] "arrival_year"
## [11] "arrival_month"
## [12] "arrival_date"
## [13] "market_segment_type"
## [14] "repeated_guest"
## [15] "no_of_previous_cancellations"
## [16] "no_of_previous_bookings_not_canceled"
## [17] "avg_price_per_room"
## [18] "no_of_special_requests"
## [19] "booking_status"
```

```
dim(train)
```

```
## [1] 29020    19
```

```
head(train)
```

```
##       Booking_ID no_of_adults no_of_children no_of_weekend_nights
## 15241   INN15241            2              0                    2
## 33702   INN33702            2              0                    2
## 35716   INN35716            2              0                    2
## 17487   INN17487            1              0                    0
## 15220   INN15220            1              0                    1
## 19838   INN19838            2              0                    0
##       no_of_week_nights type_of_meal_plan required_car_parking_space
## 15241                 5       Meal Plan 1                          0
## 33702                 1      Not Selected                          0
## 35716                 2      Not Selected                          0
## 17487                 2       Meal Plan 2                          0
## 15220                 0      Not Selected                          0
## 19838                 2      Not Selected                          0
##       room_type_reserved lead_time arrival_year arrival_month arrival_date
## 15241        Room_Type 4       106         2018             7           19
## 33702        Room_Type 1       148         2018             4           23
## 35716        Room_Type 1        68         2018             2            6
## 17487        Room_Type 1       320         2018             8           18
## 15220        Room_Type 1       131         2018            10           10
```

```
## 19838       Room_Type 1         2        2018             3            24
##       market_segment_type repeated_guest no_of_previous_cancellations
## 15241              Online              0                            0
## 33702              Online              0                            0
## 35716              Online              0                            0
## 17487             Offline              0                            0
## 15220              Online              0                            0
## 19838              Online              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
```

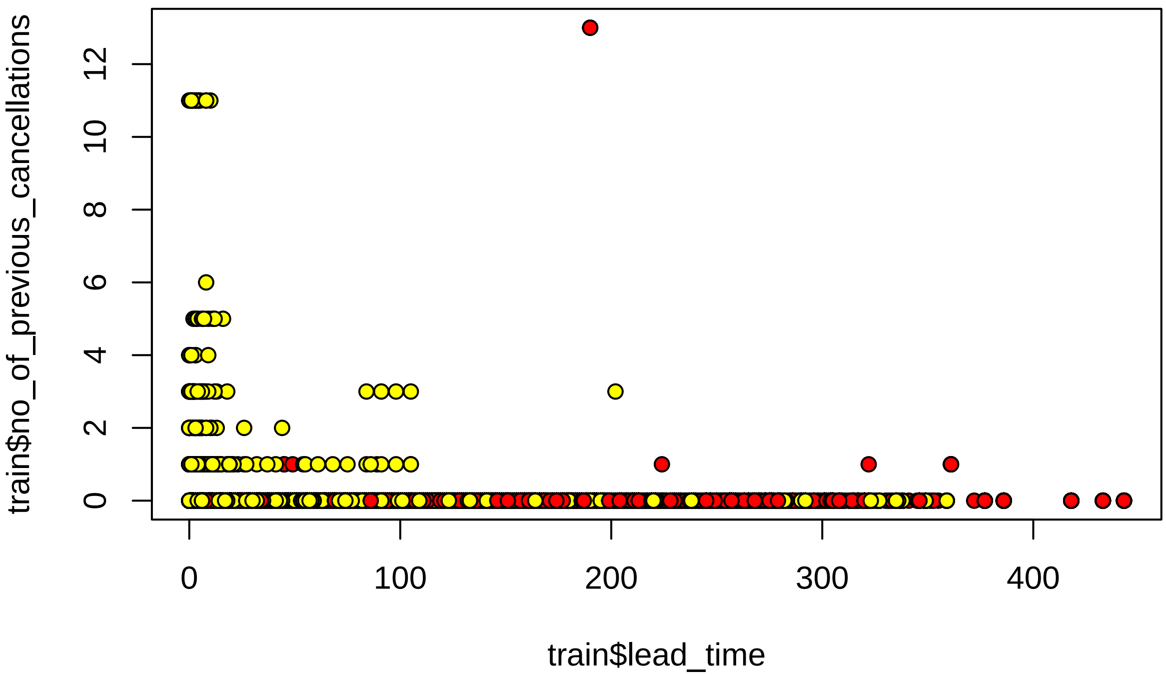
```
summary(train)
```

```
##   Booking_ID         no_of_adults   no_of_children   no_of_weekend_nights
##  Length:29020       Min.   :0.000   Min.   :0.0000   Min.   :0.0000
##  Class :character   1st Qu.:2.000   1st Qu.:0.0000   1st Qu.:0.0000
##  Mode  :character   Median :2.000   Median :0.0000   Median :1.0000
##                     Mean   :1.845   Mean   :0.1063   Mean   :0.8106
##                     3rd Qu.:2.000   3rd Qu.:0.0000   3rd Qu.:2.0000
##                     Max.   :4.000   Max.   :9.0000   Max.   :7.0000
##
##  no_of_week_nights    type_of_meal_plan required_car_parking_space
##  Min.   : 0.000    Meal Plan 1 :22245   Min.   :0.00000
##  1st Qu.: 1.000    Meal Plan 2 : 2674   1st Qu.:0.00000
##  Median : 2.000    Meal Plan 3 :    3   Median :0.00000
##  Mean   : 2.206    Not Selected: 4098   Mean   :0.03032
##  3rd Qu.: 3.000                         3rd Qu.:0.00000
##  Max.   :17.000                         Max.   :1.00000
##
##    room_type_reserved   lead_time       arrival_year  arrival_month
##  Room_Type 1:22541    Min.   :  0.00   Min.   :2017   Min.   : 1.000
##  Room_Type 2:  548    1st Qu.: 17.00   1st Qu.:2018   1st Qu.: 5.000
##  Room_Type 3:    6    Median : 57.00   Median :2018   Median : 8.000
##  Room_Type 4: 4814    Mean   : 85.08   Mean   :2018   Mean   : 7.434
##  Room_Type 5:  214    3rd Qu.:126.00   3rd Qu.:2018   3rd Qu.:10.000
##  Room_Type 6:  772    Max.   :443.00   Max.   :2018   Max.   :12.000
##  Room_Type 7:  125
##   arrival_date      market_segment_type repeated_guest
##  Min.   : 1.00   Aviation     :  101    Min.   :0.00000
##  1st Qu.: 8.00   Complementary:  313    1st Qu.:0.00000
##  Median :16.00   Corporate    : 1625    Median :0.00000
##  Mean   :15.59   Offline      : 8457    Mean   :0.02564
##  3rd Qu.:23.00   Online       :18524    3rd Qu.:0.00000
```

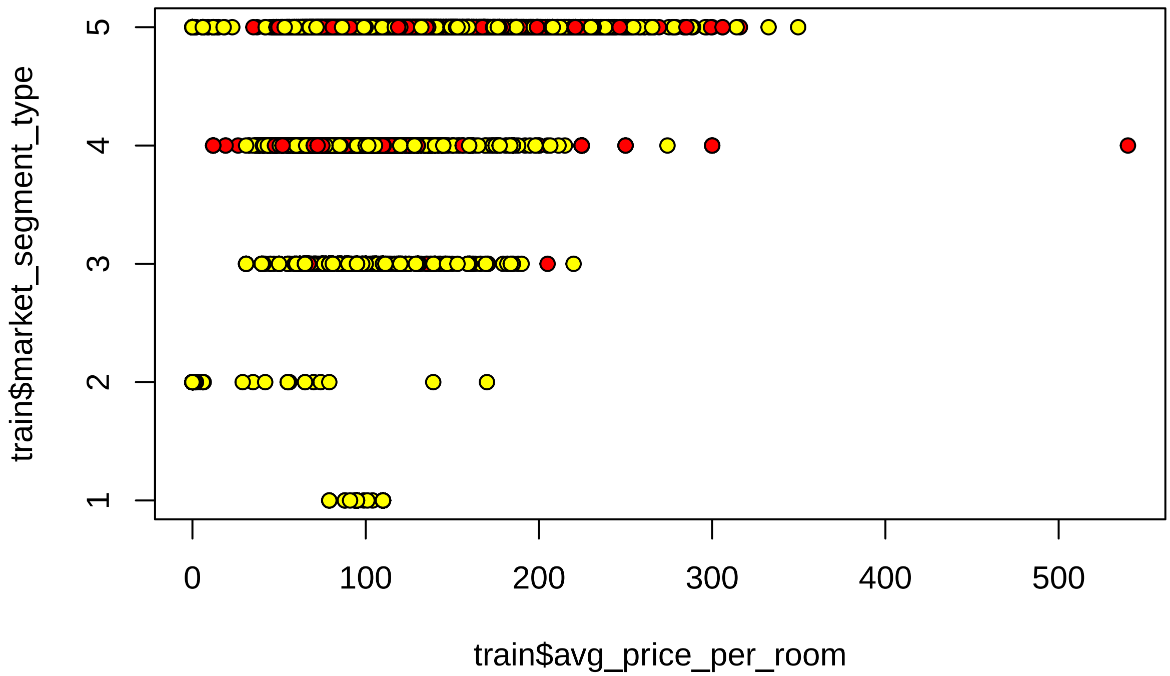
```
##   Max.   :31.00                          Max.   :1.00000
##
##   no_of_previous_cancellations no_of_previous_bookings_not_canceled
##   Min.   : 0.00000             Min.   : 0.0000
##   1st Qu.: 0.00000             1st Qu.: 0.0000
##   Median : 0.00000             Median : 0.0000
##   Mean   : 0.02123             Mean   : 0.1537
##   3rd Qu.: 0.00000             3rd Qu.: 0.0000
##   Max.   :13.00000             Max.   :58.0000
##
##   avg_price_per_room no_of_special_requests      booking_status
##   Min.   :  0.00     Min.   :0.0000         Canceled    : 9507
##   1st Qu.: 80.30     1st Qu.:0.0000         Not_Canceled:19513
##   Median : 99.45     Median :0.0000
##   Mean   :103.40     Mean   :0.6167
##   3rd Qu.:120.00     3rd Qu.:1.0000
##   Max.   :540.00     Max.   :5.0000
##
```

2c. Data Graphing

```
plot(train$lead_time, train$no_of_previous_cancellations, pch=21, bg=c("red","yellow")[train$booking_sta
```




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



2d. Logistic Regression

```
logistic_regression_model <- glm(booking_status~avg_price_per_room+market_segment_type+lead_time+no_of_p
logistic_regression_model
```

```
##
## Call:  glm(formula = booking_status ~ avg_price_per_room + market_segment_type +
##     lead_time + no_of_previous_cancellations, family = "binomial",
##     data = train)
##
## Coefficients:
##                      (Intercept)                avg_price_per_room
##                          1.95638                          -0.01068
## market_segment_typeComplementary      market_segment_typeCorporate
##                         13.81454                           1.45709
##       market_segment_typeOffline         market_segment_typeOnline
##                          1.86132                           0.90434
##                        lead_time      no_of_previous_cancellations
##                         -0.01390                           0.15317
##
## Degrees of Freedom: 29019 Total (i.e. Null);  29012 Residual
## Null Deviance:       36710
## Residual Deviance: 29210     AIC: 29220
```

```
summary(logistic_regression_model)
```

```
##
## Call:
## glm(formula = booking_status ~ avg_price_per_room + market_segment_type +
##     lead_time + no_of_previous_cancellations, family = "binomial",
##     data = train)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -2.5315  -0.8028   0.5116   0.7650   2.4446
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       1.9563788  0.2208736   8.857  < 2e-16 ***
## avg_price_per_room               -0.0106845  0.0004495 -23.770  < 2e-16 ***
## market_segment_typeComplementary 13.8145393 80.6568571   0.171   0.8640
## market_segment_typeCorporate      1.4570934  0.2316137   6.291 3.15e-10 ***
## market_segment_typeOffline        1.8613151  0.2199547   8.462  < 2e-16 ***
## market_segment_typeOnline         0.9043441  0.2174333   4.159 3.19e-05 ***
## lead_time                        -0.0139021  0.0002027 -68.598  < 2e-16 ***
## no_of_previous_cancellations      0.1531653  0.0926760   1.653   0.0984 .
## ---
## 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: 29207  on 29012  degrees of freedom
## AIC: 29223
##
## Number of Fisher Scoring iterations: 14
```

The residual deviance has a sharp decrease from the null deviance, which suggests that the logistic model has good correlation to the training data. We can see that offline bookings increase the log odds of a reservation not being cancelled because the coefficient is greater than the coefficient of online bookings.

I was not able to run the logistic regression on all available predictors because rstudio ran out of memory before the knitting was complete.

2e. Naive Bayes

```
library(e1071)
naive_bayes_model <- naiveBayes(booking_status~avg_price_per_room+market_segment_type+lead_time+no_of_pr
naive_bayes_model
```

```
##
## Naive Bayes Classifier for Discrete Predictors
##
## Call:
## naiveBayes.default(x = X, y = Y, laplace = laplace)
##
## A-priori probabilities:
## Y
##     Canceled Not_Canceled
##    0.3276017    0.6723983
##
## Conditional probabilities:
##               avg_price_per_room
## Y                     [,1]     [,2]
##   Canceled     110.60689 32.19462
##   Not_Canceled  99.88407 35.77694
##
##               market_segment_type
```

```
## Y                  Aviation Complementary   Corporate     Offline      Online
##   Canceled     0.003260755   0.000000000 0.018302304 0.264121174 0.714315767
##   Not_Canceled 0.003587352   0.016040588 0.074360683 0.304719930 0.601291447
##
##               lead_time
## Y                    [,1]     [,2]
##   Canceled     138.96634 99.02558
##   Not_Canceled  58.82755 63.89847
##
##               no_of_previous_cancellations
## Y                      [,1]      [,2]
##   Canceled     0.003786683 0.1912959
##   Not_Canceled 0.029723774 0.3853057
```

It seems that the complementary market segment never cancels their reservations. The data shows that larger lead times result in a higher probability of being cancelled. The average price of the room does not seem to affect room cancellation.

I was not able to run the logistic regression on all available predictors because rstudio ran out of memory before the knitting was complete.

2f. Test Data

```
library(caret)
```

Logisitic Regression Predictions and Evaluation

```
## 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 =  0.17381116471399"
```

```
library(caret)
naive_bayes_predictions <- predict(naive_bayes_model, newdata=test, type="class")
confusion_matrix <- table(naive_bayes_predictions, test$booking_status)
mean(naive_bayes_predictions==test$booking_status)
```

Naive Bayes Predictions and Evaluation

```
## [1] 0.7399035
```

```
sensitivity(confusion_matrix)
```

```
## [1] 0.5420521
```

```
specificity(confusion_matrix)
```

```
## [1] 0.8363748
```

2g. Strengths and Weaknesses of Logistic Regression and Naive Bayes

Naive Bayes, generally speaking, is better with smaller datasets compared to logistic regression. Naive Bayes has a higher bias and a lower variance than logistic regression. This means that Naive Bayes is more prone to underfitting, whereas logistic regression is more prone to overfitting. Naive Bayes assumes that all predictors are independent of each other, which means that it can be innaccurate if the independence of the predictors is not verified.

2h. Classification Metrics: Description, Benefits, Drawbacks

The mean classification metric is the percentage of the test observations that the model got accurately. It is a base level indicator for the accuracy of a classification model but does not provide sophisticated understanding of the accuracy of the model, because it does not account for skew in the test data. Sensitivity measures the true positive rate of the model and is useful when the model is primarily concerned with determining the positive classification. Specificity measures the true negative rate of the model and is useful when the model is primarily concerned with determining the negative classification.