

Statstical Analysis using Logistic regression on Dataset of 2012 passenger survey from san francisco airport (sfo)



Contents

[Descriptive Statistics: 2](#_Toc125127806)

[Visualisation of the variables - wait and usa: 4](#_Toc125127807)

[Logistic regression model using dirty, wait, last year and usa as predictors: 5](#_Toc125127808)

[Best Model: 6](#_Toc125127809)

[Odds ratio and 95% for best model: 7](#_Toc125127810)

[Change in predictor variable effecting risk: 8](#_Toc125127811)

[Confusion Matrix: 9](#_Toc125127812)

List of Figures, Tables and Outputs:

|  |  |
| --- | --- |
| Fig 1 | Summary of descriptive statistics for all variables in data |
| Fig 2 | Pie chart representing percentages of ‘good’ variable |
| Fig 3 | Box plot of ‘lastyear’ variable |
| Fig 4 | Percentages of ‘usa’ variable’s categories |
| Fig 5 | Boxplot of wait, usa and good variables |
| Fig 6 | Confusion Matrix using Python |
| Table 1 | Percentages of ‘dirty’ variable’s categories |
| Table 2 | Summary of logistic regression model using 4 variables |
| Table 3a to 3m | 13 logistic models with different predictors |
| Table 4 | Summary to explain odds ration and confidence intervals of best model |
| Table 5 | Dy/dx values from marginal effects for variables |
| Table 6 | Confusion matrix table for calculations |
| Table 7 | TN, TP, FN, FP descriptions |
| Output 1 | Mean predictor values |
| Output 2 | Predicted Odds of acceptance at mean values |
| Output 3 | Predicted risk at mean |

# Descriptive Statistics:

Table

Description automatically generated

Fig 1

The first variable ‘good’ is the dependent variable and binary. 58.5% of passengers have rated the airport good (Max-1) and 41.5% as not good (Min-0).

Chart, pie chart

Description automatically generated

Fig 2

The ‘dirty’ variable is categorical, the percentages associated are the number of places in the airport that the passengers thought are dirty. From Table 1, 93.2% (Max) of passengers found 0 places dirty, hence the airport can be inferred as clean.

|  |  |
| --- | --- |
| Number of dirty places | Percentage |
| 0 | 93.2% |
| 1 | 4.9% |
| 2 | 1.45% |
| 3 | 0.3% |
| 4 | 0.02% |
| 5 | 0.05% |
| 6 | 0.02% |

Table 1

The ‘wait’ variable is numerical. From fig1 the mean for this variable tells us that, on an average a passenger spent 2.44 hours in the airport between arrival and flying. On an average, passengers deviate from this value by 1.83 hours which is the standard deviation.

‘last year’ variable is skewed, quartiles are considered to explain this. From fig3 the blue box denotes 50% of passengers travelling from SFO have travelled between 1 time (q1) and 4  
 times (q2) in the last 12 months. The top 25% of passengers travelled between 4 times to approximately 9 times. Anything above 9 times are a rare occurrence for the given data or an outlier.

Chart, histogram

Description automatically generated

Fig 3

The ‘usa’ variable, is binary. From fig4 75.6% of passengers have travelled domestic (1-Max) 24.4% travelled international (0-Min).

Chart, pie chart

Description automatically generated

Fig 4

# Visualisation of the variables - wait and usa:

Chart, box and whisker chart

Description automatically generated

Fig 5

From Fig5, we focus on the orange boxes as they represent positive outcome that the airport is good. Considering passengers travelling to destinations outside the USA, for 50% of them the wait time is between 1.30 hours to 2.50 hours, for top 25% its between 2.50 to 4.40 hours and bottom 25% 36 minutes to 1.3 hours. Now, for passengers travelling in USA, 50% passengers spend between 40 minutes to 4.70 hours and top 25% 4.70 to 5.37 hours and bottom 25% 36-46 minutes.

Now, for the passengers with destination outside USA, and rating as not good, the maximum wait time (2.60 hours) is higher for the mid 50% of data when compared to maximum wait time of 50% and rating as good (2.50 hours). Similar trend is followed in the 50% data of passengers travelling in USA with 4.70 and 4.80 hours for good and not good respectively. However, the maximum hours for overall data excluding outliers, for destination outside USA and rating as not good is at 5 hours and good is at 4.3 hours and that of passengers travelling in USA the maximum wait time is same at 5.37 hours.

The green triangles in Fig5 denote mean. For passengers travelling within and outside USA , the average wait time is always high for a not good rating when compared to good. To conclude, as the wait time increases the rating is lowered by the passengers.

Note: all the time values above are approximations.

# Logistic regression model using dirty, wait, last year and usa as predictors:

The model took 5 iterations to arrive at the coefficient values that maximize the log likelihood of the model.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variables | coef | Odds Ratios | P>|z|/  P Value | Confidence Interval |
| lastyear | -0.00 | 0.99 | 0.75 | -0.01 to  0.00 |
| dirty | -0.79 | 0.44 | 0.00 | -1.00 to  -0.59 |
| wait | -0.10 | 0.90 | 0.00 | -0.14 to -0.06 |
| usa | 0.07 | 1.07 | 0.33 | -0.08 to 0.23 |

Table 2

For lastyear and USA, the odds ratio is 0.99 (lastyear) which is pretty much close to 1 and 1.07 (usa) which is greater than 1 indicating that the events are more likely to occur if predictors increase. For every extra time that the passengers travel from the SFO, the odds for a good rating goes down by 0.99. Also, the odds that the rating will be good is 1.07 times likely that the passengers fly to a destination in the USA. The coef is negative for lastyear which implies that the outcome and predictor are inversely proportional and positive for usa hence the outcome and predictor are directly proportional (good 1, usa 1) . P values are greater than 0.05 making the result insignificant. The confidence interval goes from negative to positive which includes 0, that is the true effect on the variable on the population could be zero making it insignificant too.

For dirty and wait, the odds ratio is 0.44 and 0.90 which less than 1 indicating that the event is less likely to occur if predictor increases. For every extra hour that the passenger spends in SFO, the odds for a good rating goes down by 0.90. Also, for every extra dirty place that the passenger finds at the airport the odds of good rating goes down by 0.44. The coef is negative which implies that the outcome and predictor are inversely proportional. P values for both is less than 0.05 making the result significant. The confidence interval lies in negative making the predictors significant.

# Best Model:

The logistic models below are taken on the basis of predictors. Each variable from the data set is added and removed as a predictor with different combinations by trial and error for finding a best model.

Model 1

|  |  |  |  |
| --- | --- | --- | --- |
| Variables | AIC | R squared | P values |
| wait, lastyear, usa | 4933 | 0.006 | 0.00, 0.3, 0.5 |

Model 2

|  |  |  |  |
| --- | --- | --- | --- |
| Variables | AIC | R squared | P values |
| wait, lastyear, dirty | 4864 | 0.02 | 0.00, 0.8, 0.00 |

Model 3

|  |  |  |  |
| --- | --- | --- | --- |
| Variables | AIC | R squared | P values |
| wait, lastyear, dirty, usa | 4865 | 0.02 | 0.00, 0.7, 0.00, 0.3 |

Model 4

|  |  |  |  |
| --- | --- | --- | --- |
| Variables | AIC | R squared | P values |
| wait, lastyear | 4931 | 0.006 | 0.00, 0.3 |

Model 5

|  |  |  |  |
| --- | --- | --- | --- |
| Variables | AIC | R squared | P values |
| wait, usa | 4932 | 0.006 | 0.00, 0.5 |

Model 6

|  |  |  |  |
| --- | --- | --- | --- |
| Variables | AIC | R squared | P values |
| wait, dirty | 4862 | 0.02 | 0.00, 0.00 |

Model 7

|  |  |  |  |
| --- | --- | --- | --- |
| Variables | AIC | R squared | P values |
| lastyear, dirty | 4892 | 0.01 | 0.5, 0.00 |

Model 8

|  |  |  |  |
| --- | --- | --- | --- |
| Variables | AIC | R squared | P values |
| lastyear, usa | 4960 | 0.00 | 0.2, 0.9 |

Model 9

|  |  |  |  |
| --- | --- | --- | --- |
| Variables | AIC | R squared | P values |
| dirty, usa | 4892 | 0.01 | 0.0, 0.7 |

Model 10

|  |  |  |  |
| --- | --- | --- | --- |
| Variables | AIC | R squared | P values |
| dirty | 4890 | 0.01 | 0.0 |

Model 11:

|  |  |  |  |
| --- | --- | --- | --- |
| Variables | AIC | R squared | P values |
| usa | 4960 | 0.00 | 0.9 |

Model 12:

|  |  |  |  |
| --- | --- | --- | --- |
| Variables | AIC | R squared | P values |
| wait | 4930 | 0.00 | 0.0 |

Model 13:

|  |  |  |  |
| --- | --- | --- | --- |
| Variables | AIC | R squared | P values |
| lastyear | 4958 | 0.00 | 0.2 |

Tables 3a to 3m

From tables 3a-3m, the chosen best model is Model 6 with predictor variables as wait and dirty, it has least AIC score of 4862 when compared to all other models. Also, model 6 has the highest Pseudo R squared score at 0.2 compared to all the other models. The p values for this model are less than 0.05. These 3 factors: low AIC, high Pseudo R squared and p values less than 0.05, result in best possible model and based on values from the Tables 3a to 3m, Model 6 has all 3 making it the best amongst all the others.

# Odds ratio and 95% Confidence Intervals for best model:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variables | Coeff | Variable’s Confidence intervals | Odd’s Ratio | Confidence interval of Odd’s Ratios |
| wait | -0.10 | -0.13 to  -0.06 | 0.90 | 0.87 to 0.93 |
| dirty | -0.79 | -1.00 to  -0.59 | 0.45 | 0.33 to 0.61 |

Table 4

The confidence intervals for both predictor variables lies in negative making the predictors significant as it doesn’t include a zero effect on population. Wait’s odds ratio is 0.90 < 1 indicating that the event is less likely to occur if predictor increases and for every extra hour that the passenger spends in SFO, the odds for positive good rating goes down by 0.90.

Dirty’s odds ratio is 0.45< 1, indicating that the event is less likely to occur if predictor increases. Also, for every extra dirty place at the airport the odds of positive good rating goes down by 0.44.

The confidence intervals for wait’s odds ratio infer that for 95% of the dataset, for every extra hour spent at the airport by the passenger the odds for a positive good rating goes down by the interval of 0.87-0.93. The confidence intervals for dirty’s odds ratio infer that for 95% of the dataset, for every extra point increase in dirty places at the airport the odds for a positive good rating goes down by the interval of 0.87-0.93.

# Change in predictor variable effecting risk:

----- the mean predictor values -----

2.43

0.09

Output 1

From the output1 it can be understood that average (mean) wait time is 2.43 hours and average (mean) of number dirty places found in airport by passengers are 0.

----- the predicted odds of acceptance at these mean values ----

1.71

Output 2

From output2 it can be inferred that for every bad rating on the airport there is a chance of getting 1.7 good rating for the SFO airport.

----- the predicted risk at the mean -----

0.63

Output 3

From output3 it can be understood that for an average passenger (considering the average values of wait and usa) there is a 63% chance of getting a good rating on SFO.

# Marginal effects for predicting the risk of event when the predictor value changes:

dy/dx is one of the values that comes up when you run a marginal effect code, it is the slope which denotes the curve and we look at how does the predicted risk goes down if the wait time (by hour) and number of (1 place) dirty places goes up.

|  |  |
| --- | --- |
| Variable | dy/dx |
| wait | -0.02 |
| dirty | -0.19 |

Table 5

2% decrease of predicted risk when the wait time increases by one hour and 19% decrease of predicted risk if a passenger finds an extra dirty place in SFO.

For an average passenger in the data set with 2.43 hours wait time and finds 0 places dirty, one hour increase in wait time will decrease the probability of getting a good rating from 63% to 61% and one point increase in number of places the passenger finds dirty will decrease the probability to 44%.

# Confusion Matrix:

Graphical user interface, text

Description automatically generated

Fig 6

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predicted  0 | Predicted  1 | Sum |
| Actual  0 | TN=859 | FP=657 | 1516 |
| Actual  1 | FN= 930 | TP=1205 | 2135 |
|  |  |  |  |

Table 6

|  |  |
| --- | --- |
| TN | True negative (cases in which the actual value and predicted value were negative) |
| TP | True positive (cases in which actual value and predicted values were positive) |
| FN | False negative(cases in which actual value is positive and predicted value is negative) |
| FP | False positives (cases in which actual value is negative and predicted values is positive) |

Table 7

Percentage of Predicted Accuracy risk= TN+TP/Total Values\*100= 2064/3651\*100= 56.5%

The percentage value from above calculation means that the classifier is correct 56.5% of the times and since the value is greater than 50% the outcome is good.