## **Logistic Regression Model Analysis**

We first started our analysis via logistic regression model. Our goal was to see if we could get any valuable insight into predicting customer satisfaction using the variables provided via the datasets. We decided to use a logistic regression model because our dependent variable satisfaction (satisfaction\_v2) is a qualitative/classification variable that is binary, which is satisfied and not satisfied (neutral or dissatisfied). Additionally, the logistic regression method is a simple way for binary classification analysis and a very efficient way to run analysis to see significant variables for predicting the dependent variables. Therefore, We will use a logistic regression model to understand the relationship between the dependent variables and multiple independent variables.

We used multiple inputs for our logistic regressions. As stated earlier, we have Satisfaction as the dependent variable. After removing variables that are not integral to our analysis which are gender, customer type, age, type of travel, and class, we have used the remaining variables as independent variables: Flight Distance, Seat comfort, Departure arrival time convenient, Food and drink, Gate location, Inflight wifi service, Inflight entertainment, Online support, Ease of Online booking, Onboard service, Legroom service, Baggage handling, Checkin service, Cleanliness, Online boarding, Departure Delay in Minutes, and Arrival Delay in Minutes. After running the logistic regression, the data shows:

| **Term** | **estimate** | **std.error** | **statistic** | **p.value** |
| --- | --- | --- | --- | --- |
| (Intercept) | -9.171017464 | 0.065194384 | -140.6718943 | < 2e-16 |
| Flight.Distance | -1.79E-05 | 8.39E-06 | -2.134269623 | 0.032820716 |
| Seat.comfort | 0.497574576 | 0.010805206 | 46.04952074 | < 2e-16 |
| Departure.Arrival.time.convenient | -0.505955884 | 0.009398088 | -53.83604481 | < 2e-16 |
| Food.and.drink | 0.125149287 | 0.011278374 | 11.09639488 | < 2e-16 |
| Gate.location | 0.039713765 | 0.009069914 | 4.378626156 | 1.19E-05 |
| Inflight.wifi.service | -0.149319662 | 0.009475077 | -15.75920314 | < 2e-16 |
| Inflight.entertainment | 1.017447568 | 0.009558245 | 106.4471057 | < 2e-16 |
| Online.support | 0.194092775 | 0.009590014 | 20.2390511 | < 2e-16 |
| Ease.of.Online.booking | 0.447102385 | 0.012418397 | 36.00322804 | < 2e-16 |
| On.board.service | 0.393201804 | 0.008914286 | 44.10917696 | < 2e-16 |
| Leg.room.service | 0.324642715 | 0.00771868 | 42.05935747 | < 2e-16 |
| Baggage.handling | 0.047747058 | 0.010146098 | 4.705952753 | 2.53E-06 |
| Checkin.service | 0.275532406 | 0.007420185 | 37.13282227 | < 2e-16 |
| Cleanliness | 0.001259366 | 0.010463637 | 0.120356417 | 0.904200814 |
| Online.boarding | 0.046458164 | 0.010513734 | 4.418807337 | 9.92E-06 |
| Departure.Delay.in.Minutes | 0.00254993 | 0.000853609 | 2.987234281 | 0.002815139 |
| Arrival.Delay.in.Minutes | -0.006737248 | 0.000839982 | -8.020707282 | 1.05E-15 |

Based on the above summary, you can see the following variables: seat comfort, food and drink, gate location,inflight entertainment, online support, ease of online booking, on board service, leg room service, baggage handling, check in service, online boarding departure delay have a positive and significant effect on customer satisfaction. Meanwhile, flight distance, departure arrival time convenient, inflight wifi service, and arrival delays have negative and significant effects on customer satisfaction. Additionally, the overall fraction of correct prediction is 0.83826, precision is 0.8413158, and recall rate is 0.8643002, which is a good result for predicting customer satisfaction.

However, The result can be considered misleading since we trained and tested the model on the same set of 119255 rows of input. Training error rate often underestimates the test error rate. Thus, to develop better accuracy of logistic regression models, we will be using cross-validation. We created subsets of training and testing data sets divided into 2/3 to training and remaining under testing. We ran the logistic regression using the subset and excluding the non-significant variable (Cleanliness).

| **term** | **estimate** | **std.error** | **statistic** | **p.value** |
| --- | --- | --- | --- | --- |
| (Intercept) | -9.122882107 | 0.078582535 | -116.0929982 | < 2e-16 |
| Flight.Distance | -1.98E-05 | 1.03E-05 | -1.922887333 | 0.054494201 |
| Seat.comfort | 0.49390676 | 0.013251583 | 37.27152902 | < 2e-16 |
| Departure.Arrival.time.convenient | -0.526578008 | 0.011568628 | -45.51775919 | < 2e-16 |
| Food.and.drink | 0.13186732 | 0.013846148 | 9.52375491 | < 2e-16 |
| Gate.location | 0.052856759 | 0.011130268 | 4.748920717 | 2.05E-06 |
| Inflight.wifi.service | -0.149072094 | 0.011596734 | -12.85466232 | < 2e-16 |
| Inflight.entertainment | 1.023061659 | 0.011750955 | 87.06200432 | < 2e-16 |
| Online.support | 0.192617414 | 0.011729342 | 16.42184238 | < 2e-16 |
| Ease.of.Online.booking | 0.465263576 | 0.014835962 | 31.3605253 | < 2e-16 |
| On.board.service | 0.385792592 | 0.010645921 | 36.23853476 | < 2e-16 |
| Leg.room.service | 0.322947195 | 0.009400938 | 34.35265726 | < 2e-16 |
| Baggage.handling | 0.041032838 | 0.011526712 | 3.559804104 | 0.000371132 |
| Checkin.service | 0.277197692 | 0.009036473 | 30.67542828 | < 2e-16 |
| Online.boarding | 0.030168098 | 0.01285245 | 2.34726439 | 0.01891183 |
| Departure.Delay.in.Minutes | 0.003817181 | 0.001056421 | 3.613313123 | 0.000302309 |
| Arrival.Delay.in.Minutes | -0.007747888 | 0.001042189 | -7.434247591 | 1.05E-13 |

We found it almost similar to our previous logistic regression with no significant changes. Additionally, the overall fraction of prediction is the same .83823 and precision to be .8420267, and recall rate to be .8639127.

Lastly, we decided to run the statistic with a smaller model with specific independent variables that we think are going to be significant for predicting the dependent variables. We picked the IV: Seat comfort, Leg room service, Arrival Delay in Minutes, Food and drink, and Onboard service, and the result shows:

| **term** | **estimate** | **std.error** | **statistic** | **p.value** |
| --- | --- | --- | --- | --- |
| (Intercept) | -4.6740555 | 0.04141038 | -112.8715913 | <2e-16 |
| Seat.comfort | 0.644650417 | 0.009520731 | 67.71018341 | <2e-16 |
| Leg.room.service | 0.417885936 | 0.007232078 | 57.78227913 | <2e-16 |
| Arrival.Delay.in.Minutes | -0.003771605 | 0.000230647 | -16.35226752 | <2e-16 |
| Food.and.drink | -0.103820995 | 0.008954996 | -11.59363991 | <2e-16 |
| On.board.service | 0.535731443 | 0.007380296 | 72.58942189 | <2e-16 |

Seat comfort, leg room service, and onboard service have a positive and significant effect on customer satisfaction; meanwhile, arrival delay and food and drink have significant negative effects on customer satisfaction which is very different from earlier tests we run. Additionally, when computing the overall fraction of correct prediction, it is only .72610, while precision is .73541 and recall rate is .7734, which is significantly lower than the original. However, it is not an outstanding rate but is still a decent rate to predict customer satisfaction.

In summary, based on our analysis, multiple variables are significant to customer satisfaction. Thus, not one specific variable can completely predict the dependent variable as our prediction rate has decreased when making our logistic regression model smaller. Therefore, in our logistic regression model, it is vital to use all the variables to see the variables that play a positive and significant role and what plays a significant negative role in customer satisfaction.