# Project: Prediction of Passenger Satisfaction

By: Miguel Mitra

Objective

To predict airline passenger satisfaction based on passenger survey results and determine the which services have the most effect on satisfaction results.

Data Overview

* Source: <https://www.kaggle.com/datasets/binaryjoker/airline-passenger-satisfaction>
* Summary: The dataset has 129,880 rows and 24 columns. Exploratory data analysis of the feature columns shows one column had 393 null values, but this specific column is not part of the analysis, so all rows are kept.
* Target Variable: Satisfaction (“Satisfied” or “Neutral or Dissatisfied”) – converted to binary variables Satisfied: 1, Neutral or Dissatisfied: 0
* Features: passenger scores on service-related features
  + 'Departure and Arrival Time Convenience', 'Ease of Online Booking', 'Check-in Service', 'Online Boarding', 'Gate Location', 'On-board Service', 'Seat Comfort', 'Leg Room Service', 'Cleanliness', 'Food and Drink', 'In-flight Service', 'In-flight Wifi Service', 'In-flight Entertainment', 'Baggage Handling'.

# Classification Model

* Logistic Regression
* Support Vector Machine (SVM)

# Results

Results are analyzed for the whole dataset, for Business Class Passengers, and for Economy Class Passengers. Additionally, analysis was performed on the Satisfaction scores for short and long haul flights.

## Full Dataset

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | | TP | FP | TN | FN | accuracy |
| Full Dataset | Logistic Regression | 18545 | 3649 | 12783 | 4244 | 0.7987 |
| SVM | 18466 | 3728 | 12863 | 4164 | 0. 7987 |

Top 5 Features (Using Logistic Regression)

* In-flight Wifi Service: 0.7384563490400841
* Online Boarding: -0.42096242957472013
* Leg Room Service: -0.4125980220392266
* In-flight Entertainment: -0.39354130810011734
* Seat Comfort: -0.29153334995150954

Overall, Logistic Regression and SVM are good models to predict passenger satisfaction. Logistic Regression and SVM are mathematically similar in terms of prediction accuracy. Both models seems to have struggled with False Negatives than False Positives.

Additionally, Logistic Regression was used to extract to features that affected the Satisfaction Score the most, and In-Flight WiFi has the most positive effect on the satisfaction scores. It is worth noting 4 of the top 5 features have negative coefficients, suggesting that as these features receive lower scores or poorer ratings, passengers are more likely to be dissatisfied.

## Business Class Passengers

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | | TP | FP | TN | FN | accuracy |
| Business Class Dataset | Logistic Regression | 3881 | 1727 | 11978 | 1076 | 0.8498 |
| SVM | 3901 | 1707 | 11995 | 1059 | 0.8517 |

Top 5 Features (Using SVM)

* Online Boarding: -0.4996227604793608
* In-flight Wifi Service: 0.49771279575747057
* Leg Room Service: -0.3557254002802218
* In-flight Entertainment: -0.2993393497720516
* On-board Service: -0.20738457252067696

Logistic Regression and SVM models to predicted passenger satisfaction more accurately when using just the Business Class passengers compared to using the full dataset. SVM performed better than Logistic Regression in terms of prediction accuracy. Unlike the full dataset, both models seem to have struggled with False Positives than False Negatives when using Business Class passenger data only.

Unlike the full dataset, SVM was used for feature selection on Business Class passenger as SVM has the higher accuracy score. In-Flight WiFi has the most positive effect on the satisfaction scores, but the Online Boarding is the most important feature. Again, 4 of the top 5 features have negative coefficients, suggesting that as these features receive lower scores or poorer ratings, passengers are more likely to be dissatisfied. On-board Service is also a top 5 important feature for Business Class passengers.

## Economy Class

The data set has “Economy” and “Economy Plus” label. For the purposes of this analysis, these passenger classes are combined.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | | TP | FP | TN | FN | accuracy |
| Economy Class Data Set | Logistic Regression | 15630 | 642 | 1914 | 2093 | 0.8651 |
| SVM | 15851 | 421 | 1803 | 2204 | 0.8705 |

Top 5 Features (Using SVM)

* In-flight Wifi Service: 0.9136404467995236
* Ease of Online Booking: -0.23643215720361285
* In-flight Entertainment: -0.22052113171736717
* Gate Location: 0.18539734057137183
* Online Boarding: 0.1158424822689948

Logistic Regression and SVM models to predicted passenger satisfaction more accurately when using just the Economy Class passengers. SVM performed better than Logistic Regression in terms of prediction accuracy. Like the full dataset (and unlike Business Class only data), both models seem to have struggled with False Negative than False Positives.

Like the full dataset, In-Flight WiFi has the most positive effect on the satisfaction scores for Economy Class passengers. An improvement in the In-flight Wifi Service will significantly and positively affect satisfaction scores. Interestingly, Gate Location and Online Boarding also have positive coefficients. It is also worth noting only 2 of the top 5 features have negative coefficients, indicating areas for improvement for the airlines to avoid dissatisfaction amongst passengers in Economy Class.

## Short Haul

Short Haul flights are considered as flights with flight distance of less than 3000 miles.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | | TP | FP | TN | FN | accuracy |
| Short Haul Dataset | Logistic Regression | 18204 | 3250 | 10379 | 4246 | 0.7922 |
| SVM | 18117 | 3337 | 10527 | 4098 | 0.7939 |

Top 5 Features (Using SVM)

* In-flight Wifi Service: 0.5944848431000441
* In-flight Entertainment: -0.4082119526370745
* Online Boarding: -0.36271722527063577
* Leg Room Service: -0.31614005893687525
* Seat Comfort: -0.18903517147058507

Logistic Regression and SVM models to predicted passenger satisfaction less accurately when using just the Short Haul Flight dataset. SVM performed slightly better than Logistic Regression in terms of prediction accuracy. Both models seem to have struggled with False Negative than False Positives. SVM was used for feature analysis.

In-Flight WiFi has the most positive effect on the satisfaction scores for Short Haul passengers, which is consistent across the analysis so far. It is also interesting to note that the top 5 features when using Logistic regression is like the top 5 features above and in the same order, indicating that the two models are not very dissimilar from one another. Again, 4 of the top 5 features have negative correlation to Satisfaction.

## Long Haul

Long Haul flights are considered as flights with flight distance of more than 3000 miles.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | | TP | FP | TN | FN | accuracy |
| Long Haul Class Dataset | Logistic Regression | 499 | 225 | 2225 | 85 | 0.8978 |
| SVM | 505 | 219 | 2230 | 80 | 0.9014 |

Top 5 Features (Using SVM)

* In-flight Wifi Service: 0.3751951088641056
* In-flight Entertainment: -0.34685797127750384
* Leg Room Service: -0.2757694273372306
* Gate Location: -0.2134663064340157
* On-board Service: -0.1796722819948001

Logistic Regression and SVM models to predicted passenger satisfaction the best out of all the variations of the dataset when using just the Long Haul Flight dataset. It is worth noting this is the small of all the dataset variations so it there might be less variations in the data. SVM also performed slightly better than Logistic Regression in terms of prediction accuracy and was used for feature analysis. Both models seem to have struggled with False Positives than False Negatives.

In-Flight WiFi has the most positive effect on the satisfaction scores for Long Haul passengers. Apart from the Gate Location, these top 5 services are what are typically assumed long haul passengers would find important, as they are features that are about comfort while in the plane. Again, 4 of the top 5 features have negative correlation to Satisfaction, which airlines should focus on refining to improve satisfaction scores.

## Additional Analysis

An additional analysis was performed for the top 5 features for the full dataset; however, the accuracy of the predictions significantly dropped and so this was not repeated to the rest of the analysis.

An additional analysis was also performed that bifurcated Long and Short Haul flights into the passenger classes (Business and Economy). Short Haul flights result generally follow the top features found in the results in Business and Economy datasets, suggesting that the results are mostly driven by passenger classes and less on flight distance. For Long Haul flights followed the same trend for Long Haul Business passengers, but Long Haul Economy did not. Upon further analysis there are only 26 passengers in the test set that belonged in this group so the result for this group is found inconclusive.

# Conclusion

The analysis showed that SVM is a better classifier than Logistic Regression. Additionally, bifurcating the dataset by passenger class (Business Class and Economy Class) can achieve better accuracy. The analysis also showed In-Flight Wifi has highest positive correlation to Satisfaction scores.

The results generally make sense as any flight passenger can attest. It is interesting note In-Flight Wifi Service has the highest positive correlation in every analysis that was performed, which *can* suggest that the airlines have a very good service already, but correlation does not necessarily mean causation, but any additional improvements on this service will certainly improve satisfaction scores. Other obvious important features that the analysis supported are the importance of Leg Room Service to Business Class passengers but was not an important feature to Economy Passengers (presumably because lack of leg room comes with the price of the ticket). There are also surprising results, like gate location being a in important feature in several times during this analysis. This is interesting to note given gate location is not a service the airline can necessarily improve easily as gate location is sometimes dictated by airport layout or how busy an airport is on any given day.

Although these analysis results generally make sense in the context of passenger classes or whether the flight is a short or long haul, further analysis can be performed to calculate or quantify how much an improvement to each feature can affect satisfaction scores (i.e, using log odds), which is very useful when advising an airline company which service it can focus on improving. Additionally, further analysis on passenger characteristics (gender, age group, type of traveler, etc.) can bring more insights that can help determine which services can improve satisfaction scores, which is more catered, targeted, and specific to a passenger market / group.