Colin Michael

DSC 550 Data Mining

Term Project Final Submission

November 18th, 2022

Airline travel is a rapidly growing industry and is something that impacts the lives of millions of people across the world. It is estimated that 674 million people flew on U.S. airlines in 2021 (bts.gov). This number increased from 369 million people in 2020, which is an 82% increase (it is important to note that this dramatic spike might have been influenced by COVID-19) (bts.gov). The International Air Travel Association projects that air travel will continue to skyrocket into 2023 (simpleflying.com). Data shows that the airline travel industry impacts millions of people and will continue to be a major part of life in the United States of America.

I will be focusing my DSC 550 Term Project specifically on customer satisfaction data from an anonymous airline company. I chose to study the airline industry because of its economic impact and how prevalent air travel is in American society. I am fascinated by the airline travel industry because of how lucrative of a sector of the economy it is, but also how low customers generally rank their experiences on airlines.

The air transportation sector produces $779 billion dollars, or 4.2%, of The United States of America’s GDP (iata.org). Approximately 6.5 million jobs are supported by the airline industry(iata.org). The United States is the largest commercial flyer in the world, with China being a distant second (zippia.com). The sheer economic impact that airline travel has on the United States is incredible.

A 2022 Gallup study found that 27 percent of Americans have a very/somewhat positive view of the airline industry, compared to 37 percent that have a very/somewhat negative view of the airline industry (gallup.com). I am genuinely shocked that a multi-billion-dollar consumer industry has 10 percent more negative reviews than positive reviews. Furthermore, these rates have dropped significantly over recent years. In 2018, 43 percent of Americans had very/somewhat positive views of the airline industry, compared to 24 percent that had very/somewhat negative views of the airline industry. Common complaints include smaller seats, fewer in-flight entertainment options, higher priced tickets, and more expensive baggage options.

From a business perspective, it is deeply concerning that Americans are developing increasingly lower opinions of airline industry. It is reasonable to believe that higher disapproval rates may lead to lower demand for airline travel. Consumers may look to alternative means of travel, such as trains, cars, and buses. Businesses have reason to fear that their profits will diminish if negative sentiment towards airlines continues.

I believe that data can be utilized to better understand and improve customer satisfaction on flights. The insights produced by these findings would be invaluable to all types of airlines. An airline that improves its customer feedback could financially benefit greatly. Customers would be inclined to pick flights on airlines that have significantly higher satisfaction rates than their competitors.

I pulled an anonymous dataset from Kaggle that includes approximately 27,000 flight experience survey responses (Kaggle.com). The dataset is structured such that each row of data represents one survey response. The extensive dataset will allow me to identify which variables are most significantly related to customer feedback. This will allow me to product insights for airlines on which specific areas of the airline travel process customers are most impacted by. Accordingly, airlines will be able to use these insights to improve their flights most efficiently.

I began my data analysis by creating visualizations from the data to get an understanding of the general trends in the data. The charts are not meant to answer any specific questions, but rather to get a rough understanding of the massive dataset I am working with. First, I created a bar chart comparing gender of survey respondents and found that there were approximately the same number of female and male respondents:

Chart, bar chart

Description automatically generated

Next, I created a bar chart showing the average ease of booking rating for business travel vs. personal travel and found that business travelers have a slightly higher average ease of booking rating:

Chart, bar chart

Description automatically generated

I then created a scatterplot comparing flight distance to departure delay in minutes, and found a slightly negative correlation between the two variables:

Chart, scatter chart

Description automatically generated

Last, I created a stacked bar chart comparing the distribution of neutral/dissatisfied vs. positive reviews amongst loyal and disloyal customers. I found that loyal customers have a higher ratio of positive to neutral/dissatisfied reviews compared to disloyal customers:

Chart, bar chart

Description automatically generated

These charts are all very useful towards better understanding the dataset, but do not necessarily create any business insights. Graphs alone only reveal correlations in the data. I created a logarithmic regression model to identify causation within the data. First, I converted my categorical variables to dummy variables of 0 or 1 to meet the requirements of a logarithmic regression. Gender\_Female, Gender\_Male, Customer Type\_Loyal Customer, Customer Type\_Disloyal Customer, Type of Travel\_Business Travel, Type of Travel\_Personal Travel, Class\_Business, Class\_Eco, and Class-\_Eco Plus were my dependent variables. A value of 1 represents that the row of data is categorized as that variable and a value of 1 represents the opposite. Satisfaction was my dependent variable, with values being satisfied or neutral/dissatisfied.

My logarithmic regression produced the following results:

Table

Description automatically generated

My model produced some interesting results. The precision columns show what percent of classifications were correct. The model accurately predicted 80% of neutral or dissatisfied responses and 75% of satisfied. The recall columns show what percent of positives cases the model caught. The model caught 81% of neutral or dissatisfied and 74% of satisfied results. The F1 score shows the percent of positive predictions that were correct. The model correctly predicts positive results for 81% of neutral or dissatisfied reviews and 75% of satisfied reviews.

The confusion matrix shows that exact amount of true positive (3948), false positive (907), false negative (963) and true negative (2755) results in the model. Overall, the model has a 78% accuracy rating, which I am extremely happy with.

The aim of my data analysis was to create actionable insights backed by data for airline companies to utilize. I successfully did this by creating a logarithmic model based on customer survey data to predict customer feedback ratings. The model I created has a 78% accuracy rating, based on a dataset with 27,000 rows. I believe this accuracy could be improved if an airline embraced my model and provided more consumer data. Understanding customer feedback sentiments is crucial towards continued success for airlines and represents a potential key advantage in a competitive industry.