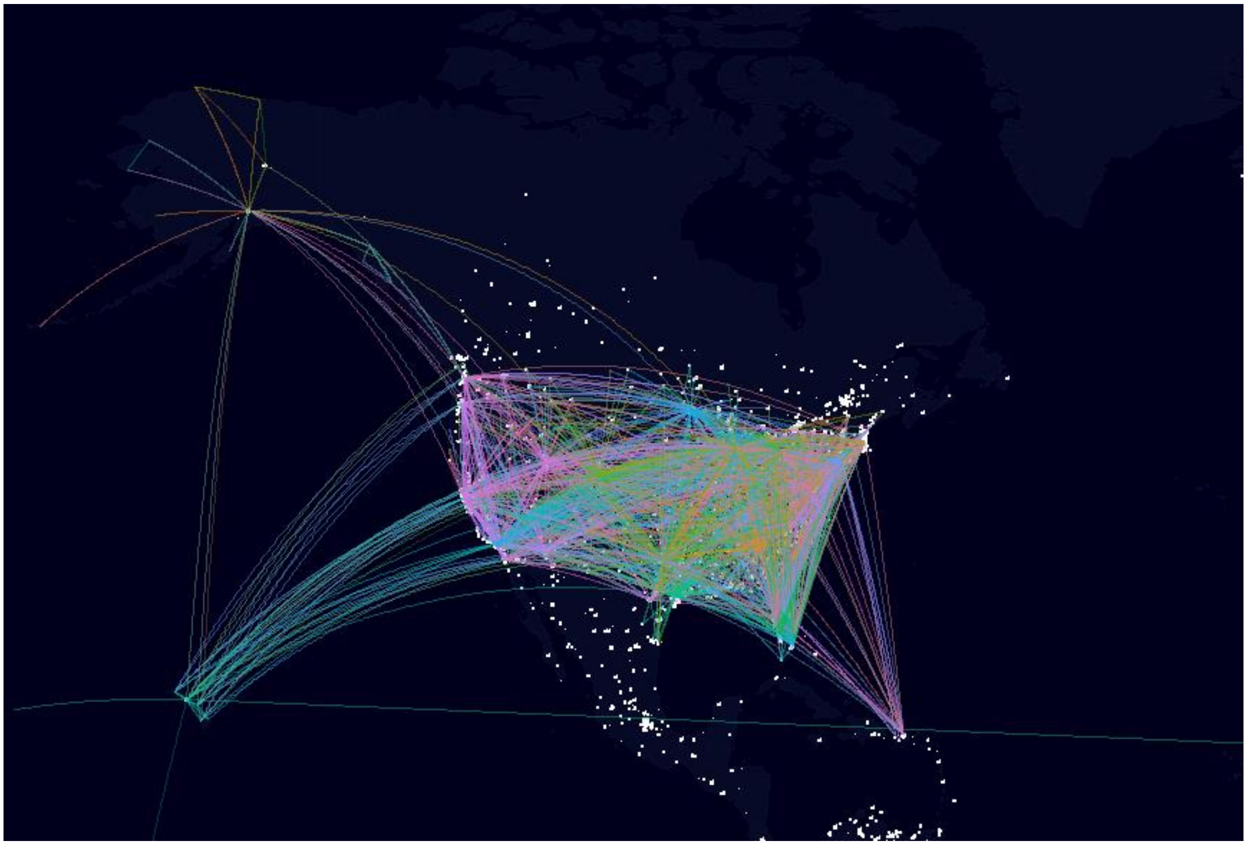
Syracuse University

School of Information Studies

Fall 2020

Flight Satisfaction Survey

IST687: Introduction to Data Science



Instructor: Gary Krudys

Presenter: Alec Schneider afschnei@syr.edu

Table of Contents

Introduction

Background

Airlines have been pioneers in the analytics community regarding pricing and route optimization, however they have long struggled with understanding customer satisfaction. It is possible they have struggled with this due to company bias, and lack of customer data from other airlines. To tackle this problem an industry funded customer survey gathered ~130,000 responses each with 25 attributes. I was tasked with analyzing this data thoroughly enough to provide actionable insights for the businesses to implement in their company strategies.

After cleaning and exploring the survey data, creating informative visualizations to assist in understanding the data, and applying machine learning algorithms to it, airlines are now closer to fully understanding what their customers value most, thus improving customer satisfaction. This will be a win for the airlines, as increased satisfaction is believed to increase demand of air travel, and the consumers who choose to travel on their aircrafts.

Business Questions

1. Which airlines and origin cities have the highest rates of departure delays and cancellations?
2. Do vacation destinations have a higher customer satisfaction score?
3. Which demographics are most price sensitive?
4. Which airlines have the highest satisfaction scores?
5. Which factors affect the customer satisfaction score most?

Data Cleaning & Feature Engineering

The analysis began with first understanding the tidiness of the dataset and understand what logic should be used to clean any naming conventions and missing values, before engineering any additional features. After all cleaning logic and feature engineering logic was decided upon, each step was coded as function that could be imported and reused, to save storage space by not having to create an additional file of the cleaned dataset. These functions were then combined in one function so the raw data could be transformed with one line of code inside an R script (see appendix XXX).

Data Cleaning

Any column names that contained a space were substituted with an underscore to prevent any future difficulties in column extraction. In the raw data there were four columns that had missing data. The following logic was applied to clean up these missing values:

1. Satisfaction column
    a. Impute the missing values with the mean satisfaction score, excluding the missing values
2. 337 rows of data were removed from the data set as they were survey responses where the flight was not cancelled, but there was no arrival delay, flight time, or departure delay information. The decision was made to remove these rows since it as these data points could be reliably imputed with a mean value

3. Departure_Delay_in_Minutes column
    a. Since all remaining missing values in this column were associated with a cancelled flight, the assumption applied was there was no delay before cancellation. A value of zero was used to replace the missing value
4. Arrival_Delay_in_Minutes column
    a. Since all remaining missing values in this column were associated with a cancelled flight, the assumption applied was there was no delay before cancellation. A value of zero was used to replace the missing value
5. Flight_time_in_minutes column
    a. Since all remaining missing values in this column were associated with a cancelled flight, the assumption applied was there was no delay before cancellation. A value of zero was used to replace the missing value

Feature Engineering

In order to extract additional information from the existing variables, new variables were created to categorize discrete variables and one hot encode category variables so they can be used in a machine learning algorithm.

Categorization:

One Hot Encoding:

Additional Features:

Exploratory Data Analysis

Modeling

XGBoost Classification

Support Vector Machine Classification

Naive Bayes Classificaton

Conclusion

Appendix