Business Report: Natural Language Processing of Airline Customer Reviews

Johann Manukulasuriya

Fall 2021, BrainStation Cohort

Problem Statement

By utilizing Natural Language Processing, how can I help potential travellers pick the right airline for their purpose of travel by predicting whether a past customer's review led to a yes, or no, recommendation? This would help future travellers navigate the large and clustered world of competing airlines based on where they are and where they're flying to by making their selection of airline choice that little bit easier based on past reviews.

Background

The business use case for creating a model that can accurately predict whether a customer review led to a yes/no recommendation will allow travellers to pick the airline that they feel meets their checklist when it comes to flying. It can also prompt airlines to airlines to actively monitor the quality of service on all their routes, ensuring that their travellers experience the best overall service, boosting their reputation, but also increase the retention rates of past travellers. Data has always played an integral role in the airline industry. It helps airlines prepare for peak seasons – summer and holidays. Airlines can use data to create new routes based on popularity and customer demand. The goal of this project is to travellers pick the right airline, but airlines themselves can also use the analysis to evaluate how they, and their competitors, are performing when it comes to flying passengers.

Data Source and Information

The dataset was downloaded from Kaggle. However, it was web-scraped by Divyansh Agrawal from the Skytrax website. The shape of the dataset was ~69,000 rows by 17 columns, consisting of categorical and numeric variables. The numeric columns contained score data of various categories (e.g., cabin service, seat comfort, value for money), and were scored on a rating of 1 – 5. The categorical variables consisted of variables such as traveller type, author, route, customer review and the target column – recommended. After looking over and familiarizing myself with the dataset and its features, I began the process of data cleaning and feature engineering.

Data Cleaning/Preprocessing and EDA

The cleaning process involved a combination of dropping rows that missing categorical data and imputing missing numerical values. Columns that were not necessary for the analysis or modelling section, such as the author column, were dropped. After cleaning, I was left with a dataset of ~34000 rows by 18 columns. The target column, recommended, was also converted from categorical to binary values in preparation for the modelling portion.

The EDA section consisted of creating a smaller data-frame consisting of airlines of a specific region – Middle Eastern carriers in this case. Creating a smaller dataset would allow for a more in-depth analysis and comparison of airlines of the same region. This would also mean less cluttering of the visuals as we would only be examining a small number of airlines.

An example is given below of the Overall Score performance of a select few Middle Eastern carriers:

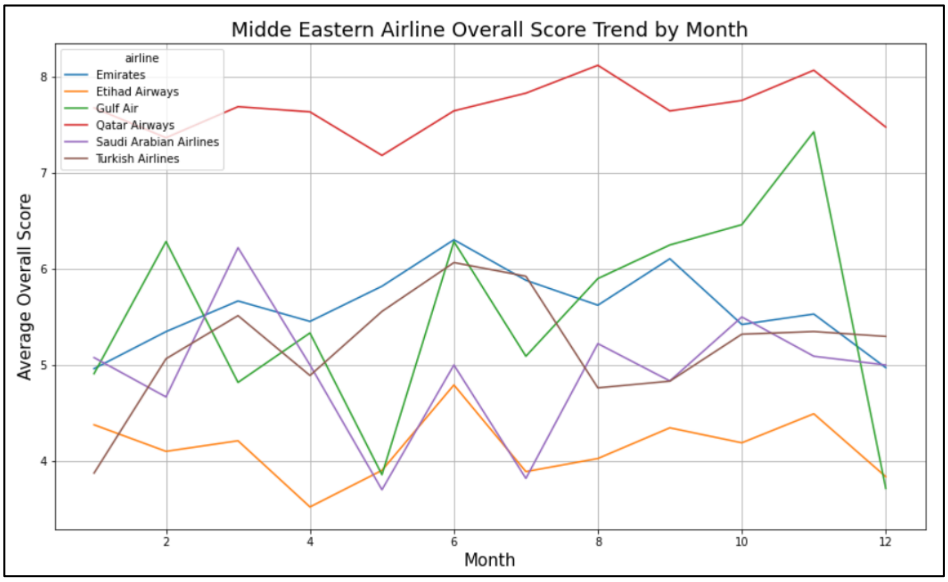

Image 1: Average Overall Score Trend for Middle Eastern Airlines

A general comparison of the Middle Eastern airlines was then followed up with a detailed exploration of the yes recommendation % for one airline – Emirates. This analysis involved examining the yes recommendation % breakdown by traveller type, as well as month, to gather any insight as to when Emirates generated the highest yes recommendation %. Once the analysis was completed, we moved onto modelling.

Modelling Summary

The different machine learning models used to accurately predict whether customer reviews resulted in a yes/no recommendation consisted of the Logistic Regression, Decision Trees, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM) models. I selected to run these models as they were relatively simple and sufficient for the task at hand. The KNN and Decision Tree models produced accuracy scores of 70% and 76% respectively. Both the Logistic Regression and SVM models generated an accuracy of 86%. However, I decided to go with the Logistic Regression model as it had a slightly better test accuracy.

Since accuracy alone can be misleading depending on the class balance of the data, I utilized alternative metrics which consider the relative proportions of correct and incorrect classifications in both the positive and negative classes. By employing the confusion matrix and classification report, I was able to output precision and recall rates of 87% for each. A high precision score indicates that the model correctly identifies a point as actually belonging to that class. And a high recall score indicates that we can be confident that the model is not missing many class members.

Below is a visual of the Logistic Regression train and validation datasets displaying how the two datasets perform over time.

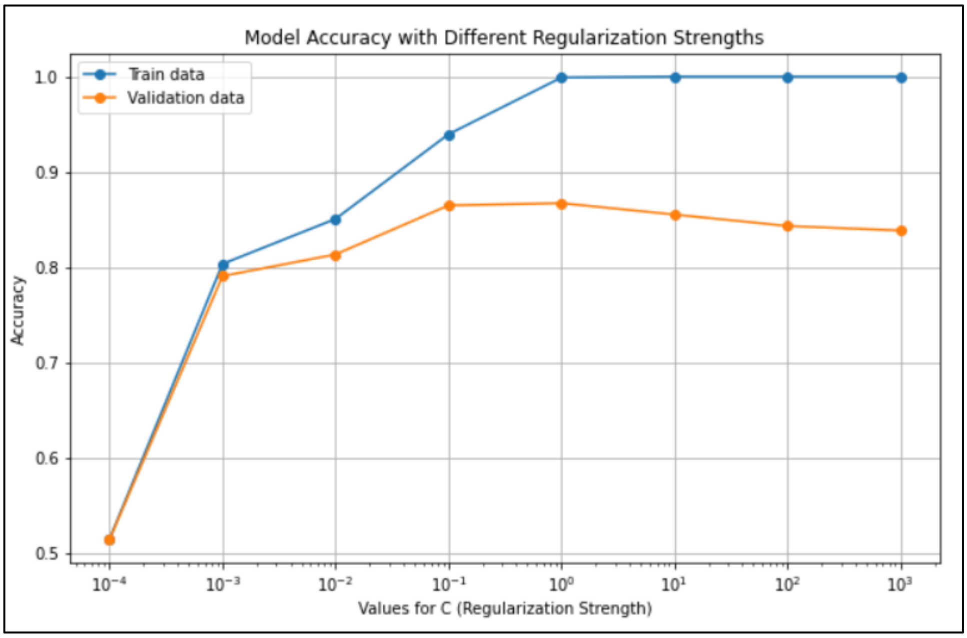

Image 2: Logistic Regression Model of Different Regularization Strengths (C)

Conclusions

Based on the EDA of Middle Eastern carriers, Qatar Airways emerged as the clear-cut airline of choice based on the performance of its individual categories (i.e., cabin service) receiving high scores, with Emirates and Turkish Airlines coming in a close second. This can be a clear indication for any travellers flying to/through/from the Middle East to fly Qatar Airways. However, both Emirates and Turkish Airlines can use the results of the analysis to implement improvements that can close the gap with Qatar, making them significantly more competitive. The Logistic Regression producing the highest test accuracy, along with highest precision and recall, was a bit of a surprise since Logistic Regression takes a probabilistic approach whereas SVM searches for the boundary in the middle of the two classes. However, less overfitting implies less bias towards the training data, resulting in the Logistic Regression method over SVM.

Business Application and Next Steps

The EDA portion presents valuable insight as to how each airline is performing as well as what airlines travellers can select based on their scores. Even airline higher-ups can use this information to create detailed plans for what services to improve and where they stand when it comes to their competition. With the Logistic Regression model, travellers will be able to find the peace of mind that the Logistic Regression model can accurately predict which airlines received yes/no recommendations based on past reviews, thereby allowing them to select that same airline. Airlines can use that information to monitor their service on every individual flight in their route network and can stay ahead of its competition as those ‘yes’ recommendations will likely lead to new customers flying their airline.

There are a couple of features that can be implemented to improve upon the work completed. The first step is building an airline recommender system. The current model can only predict whether the reviews led to Yes/No recommendation. By creating a recommender system, I could provide travellers with the top two to five airlines for a region, or of their preference based on what specific attributes they are looking for in an airline. Some of the other steps involve experimenting with a time series analysis as well as neural networks.