ORIE 4741
Midterm Report

Airline Satisfaction Data Analysis

Alan Lee, Jason Miao

November 7, 2019

aml326, jjm467

1 Data Set Description

For reference, the data set we are analyzing is accessible here.

Our dataset consists of 20 features and 41,396 examples. Each example represents a review submitted between 2002 and 2015 on Skytrax, an airline and airport rating/review site. The first half of the features represent nominal values pertaining to the review. This includes the name of the airline being reviewed, the link to the review on the Skytrax website, the title of the review, author of the review, the country that author is from, the date of the review, the text body of the review, as well as corresponding aircraft flown, traveler type (i.e. solo leisure, business), cabin flown (i.e. economy class, business class), and the route flown. These features are followed by discrete valued features representing the numerical ratings (out of 10) in various categories. This includes the overall rating, seat comfort rating, cabin staff rating, food/beverages rating, in-flight entertainment rating, ground service rating, WiFi connectivity rating, and value/money rating. The last feature is a boolean representing whether or not the reviewer would recommend that airline.

2 Exploratory Data Analysis

A preliminary exploratory data analysis was undertaken in order to familiarize ourselves with the dataset. Several features were found to have exactly 41,396 entries. These features were the airline name, a link to the review, the title of the review, author of the review, date the review was posted, text body of the review, and whether or not the reviewer would recommend that airline. Checking the airline name feature, we found 362 unique airlines. This is roughly on par with the number of airlines currently linked on the Skytrax website (the discrepancies can be attributed to airline bankruptcies). Conducting a visual inspection, we found that this feature is clean in the sense that there exist no double entries due to spelling or formatting differences. One strange entry did appear in the date feature with a review dating back to 1970. This point has been omitted from our analysis as it is likely the result of an error when compiling the dataset.

All other features have missing entries. Features such as the country the author is from, cabin flown, overall rating, seat comfort rating, cabin staff rating, food/beverages rating, in-flight entertainment rating, and value/money rating all have 31,000+ entries. These features are usable in our analysis since they still have a sizeable number of examples. The source of these missing entries is likely due to the fact that entering this information was optional.

Unfortunately, we found that some features had a majority of the entries missing. The aircraft type, type of traveler, route, ground service rating, and WiFi connectivity rating all had less than 3000 entries. Again, since these fields are optional this was likely the result of reviewer omission. Since there are so few examples in these features, they are likely unusable.

To gain a better picture of the data and a sense of how individual airlines are performing, we found the average overall rating aggregated across each unique airline, as well as the percentage of how many people recommended the airline for each unique airline. The results can be seen in Figures 1 and 2 below. The first histogram shows that the majority of airlines have overall ratings falling somewhere between a 5 to 7. The distribution of these ratings looks vaguely normal, but with a skew towards the higher end of the scale. This is supported by the fact that the average of the averaged ratings for each unique airline is 6.05. The second histogram shows a similar distribution for the percentage of reviewers who would recommend each airline. The mean in this case was 51.2%. It is worth noting, however, that there are an abnormally high number of very low percentage recommendations and very high percentage recommendations. This led us to the discovery that there were some airlines with very few reviews that were either all positive or all negative. If we want to conduct future work that sorts the dataset via airline, this is certainly something to keep note of.

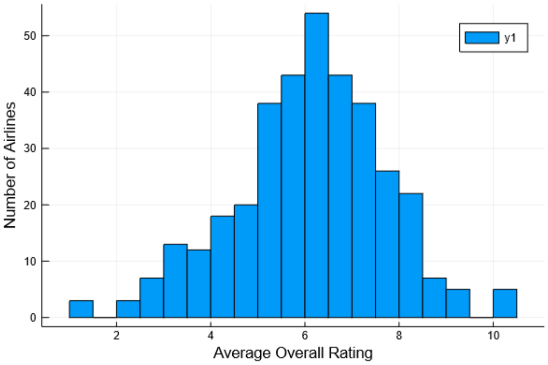

Figure 1: True vs. Predicted Airline Ratings data points (blue) with ideal trend line (black) for two rating features

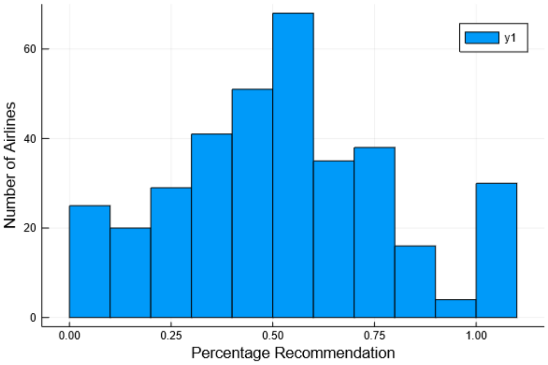

Figure 2: True vs. Predicted Airline Ratings data points (blue) with ideal trend line (black) for two rating features

3 Supervised Models

Several linear regression models were generated to predict an airline's overall rating as an introduction to better understanding the posed problem and provided data set. The data set was split into a training set and testing set which composed of 80% and 20% of the data set, respectively. The first linear regression model generated included only two features: seat comfort rating and cabin staff rating. These two features were chosen based on our personal opinion on which features would be most influential to a reviewer's overall score of an airline. After producing 1000 models with randomized training and testing data sets, the average testing mean squared error (MSE) and training MSE were 3.126 and 3.131. The relation between the true and predicted ratings can be observed in Fig. 3. Based on the provided plot, it is very difficult to see where the majority of the data points are located since many of the points are exist in the same exact same location. The concentration of true vs. predicted rating data points can be better visualized by referring to the marginal histogram as seen in Fig. 4, where the data points are grouped more closely towards the center of the trend line diagonal as shown by the brighter bins. As observed by the brighter, but yet still faint bins, the majority of data points are somewhat close to the ideal trend diagonal.

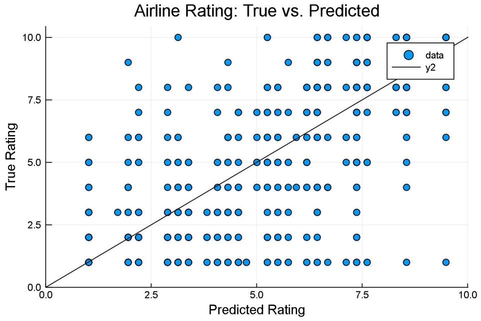

Figure 3: Histogram showing the average overall rating given to each unique airline.

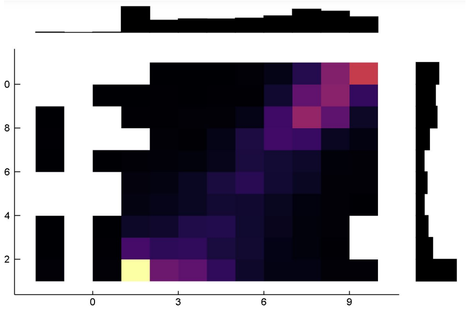

Figure 4: Histogram showing the percentage recommendation given to each unique airline.

More complex models were generated to increase the prediction accuracy by including additional features, such as the food beverages rating, inflight entertainment rating, and value money rating. After obtaining linear models for 1000 randomized testing and training sets, the average training MSE and testing MSE were 2.236 and 2.240, respectively. This new model's MSE improved by almost 30% in comparison to the regression model generated using only two features. Similar to the first model with only two features, this new model's training and testing MSE that are also very close. This can be explained by assuming that the testing and training data points are obtained from an identical distribution and shows that the five features chosen to train the model generalizes the data set. As seen in Fig. 6, the diagonal is much brighter than the initial generated model in Fig. 4, describing the increased accuracy of the regression model. However, the MSE of both the training and testing sets are still relatively large (¿2.0) which results in an average rating error larger than 1.0.

The final solution generated with these five features with an included offset was $w = [0.425, 0.662, 0.167, 0.032, 0.976, -1.55]$. This suggests that the largest factor that positively affects the overall rating is the value money rating, followed by the cabin staff rating and seat comfort rating. The last value corresponds to the included offset for the regression model.

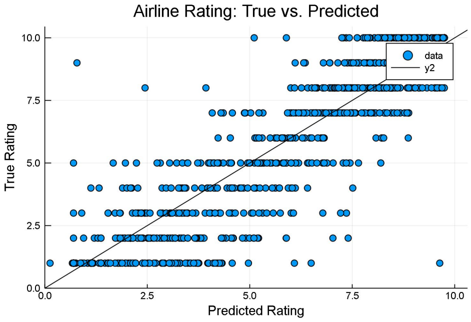

Figure 5: True vs. Predicted Airline Ratings data points (blue) with ideal trend line (black) for five rating features

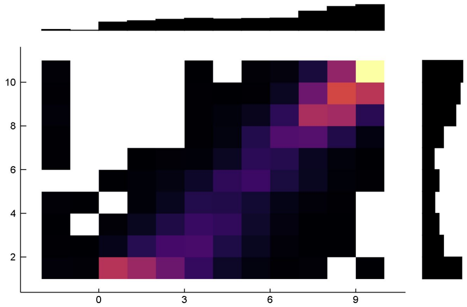

Figure 6: True vs. Predicted Airline Ratings Marginal Histogram for five rating features. Higher concentrations of data points are represented by brighter markers.

4 Future Work

The prediction model can be further improved by adding in more features to better characterize the data set. Other ratings have yet to be incorporated, although we believe that these additions will show little improvements to the result. We can also include text features to correspond positively written reviews to higher overall airline ratings by incorporating a ”bag of words” or similar method. The current models generated can also be experimentally fitted with a polynomial fit as well as the addition of a regularization parameter to account for existing outliers and prevent overfitting.

Another next step is to develop a classification model to recommend or not recommend a certain airline. After understanding which features might be the most important to an airline's overall rating, we can manipulate the features and develop a classified model that can minimize the number of classifications.

If time permits, it may be interesting to scrape the text body of the reviews using regular expressions for commonly omitted information such as the route flown and the aircraft type. In other words, though there were many missing entries for route flown and aircraft type, this information actually appeared in a significant amount of reviews in the text body.