ORIE 4741
Final Report

Airline Satisfaction Data Analysis

Alan Lee, Jason Miao

December 10, 2019

aml326, jjm467

Contents

Abstract

The airline industry is one characterized by thin margins and high competition. As such, winning customers over is key. Skytrax, the most popular airline review site, offers a glimpse into the factors that lead to customer satisfaction. Given a data set of all customer reviews on the Skytrax site as of August 2nd, 2015, we investigate the factors that contribute most directly to a customer's overall rating of an airline, as well as whether or not they would recommend that airline to others. To generate more features, we analyzed the text-body of the customer review to scrape for route information, make a many-hot encoding for key words such as "delay," and perform a sentiment analysis. For predicting the overall rating of an airline, we found our best results via linear regression with l_2 loss and no regularizer. For predicting customer recommendation, we obtained results by applying binary classifiers with hinge loss and logistic loss.

1 Introduction

1.1 Background

The aviation industry is intrinsic to the functioning of the global economy. Yet, in recent times airline profit margins have become increasingly slim. With that, more and more airlines are failing. 2019 alone has already seen high profile bankruptcies of key airlines such as Thomas Cook and WOW Air. At the same time, it has become increasingly difficult for new airlines to enter the market. This has major implications on the options for customers as well as the aviation industry as a whole.

1.2 Objective

In such a time, there is little margin for error within the aviation industry. As such, we sought to determine: *Can we predict airline customer satisfaction? What decisions influence which airline a customer prefer to take and why?* We did so by performing a data analysis on a data set of all customer reviews from Skytrax, an airline airport review and ranking site. Ultimately, an airline is driven by its customer base. Thus, through this data analysis, we hope we can improve the airline industry by discussing both what is working and what is not.

2 Data Set

For reference, the data set we are analyzing is accessible here (link may not work on GitHub previewer).

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.

2.1 Features

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 5) in various categories. This includes the seat comfort rating, cabin staff rating, food/beverages rating, in-flight entertainment rating, ground service rating, WiFi connectivity rating, and value/money rating. The overall rating is rated ou of 10. The last feature is a boolean representing whether or not the reviewer would recommend that airline.

2.2 Exploratory Data Analysis

2.2.1 Examination of Data Fields

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.

In the end, since our dataset is rather large, we ended up dealing with missing overall rating, seat comfort rating, cabin staff rating food/beverage rating, in-flight entertainment rating, and value/money rating by simply omitting any review that was missing any one of the above. This yielded a dataset that still has 28,341 examples. In a later part of the analysis, we were also able to recover airport/route information using information from the text body. This is covered further in the method section.

2.2.2 Data Visualization

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.

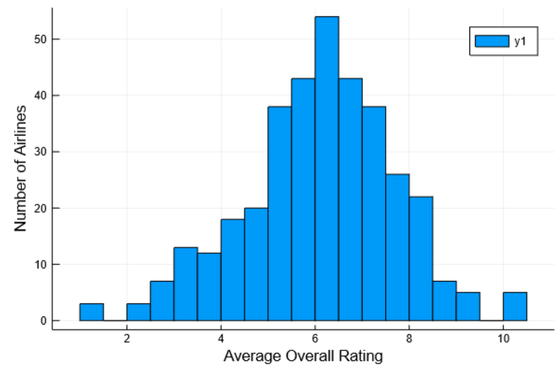

Figure 1: Histogram showing distribution of average overall rating for each airline.

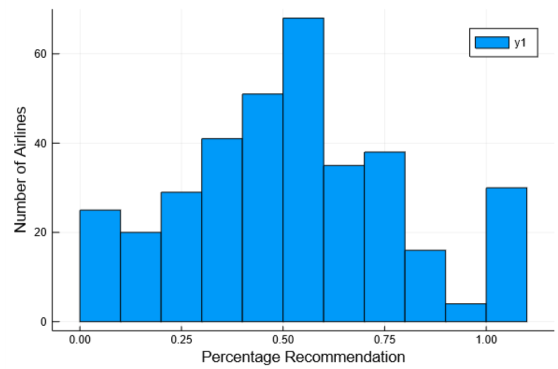

Figure 2: Histogram showing the distribution of recommendation percentage for each airline.

3 Methods

3.1 Regression Model

Several linear regression models were generated to predict an airline's overall rating to better understand 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 five features that contained the following user sub-ratings: seat comfort rating, cabin staff rating, food and beverage rating, in-flight entertainment rating, and value money rating. These five features were chosen based on our personal opinion on which features would be most influential to a reviewer's overall score of an airline. Both the L1 loss function and L2 loss function were used to predict the overall rating of review.

With a specified randomized seed, the training MSE and testing MSE were calculated and shown in Table 1 and the ratings predicted by the L2 loss function model are plotted against the true rating data points in Figure 3. The majority of the blue data points exist near the diagonal linear black line that represents the ideal truth data trend. However, it is difficult to see where the density of data points is largest due to the blue data points overlapping one another in the plot. The data is more clearly represented by a marginal histogram as shown in Figure 4. The marginal histogram shows where the true vs. predicted data points are located. The bins that are brighter (lighter in shade) represent a larger density of data points in comparison to bins that are darker. This first iteration of the supervised model performs best along the extreme sides of the ratings, where the ratings are predicted to be close to 1-3 and 7-10 while predictions require large improvements towards the middle range of the ratings.

Table 1: Training and Testing MSE Using Survey Sub-Ratings

Loss Function	Train MSE	Test MSE
L1 Loss	2.409	2.383
Quad Loss	2.273	2.095

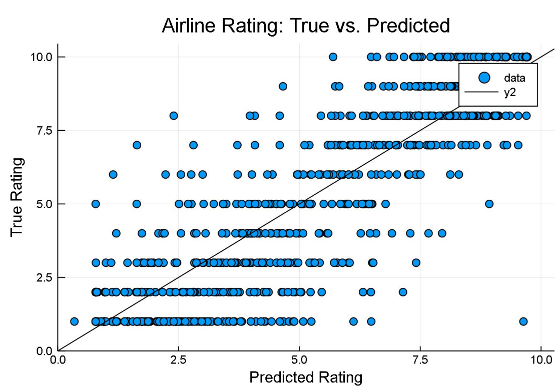

Figure 3: True vs. Predicted Airline Ratings data points (blue) calculated by L2 loss function with ideal trend line (black) using sub-rating features

3.1.1 Bag of Words

To improve the rating prediction model, more features were extracted from user text reviews. Although the sub-rating survey that the customers

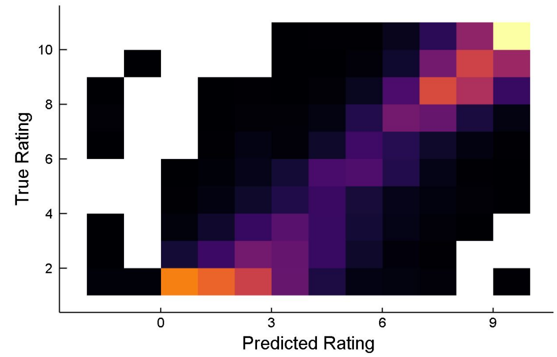

Figure 4: True vs. Predicted Airline Ratings Marginal Histogram for five sub-rating features calculated by L2 loss function. Higher concentrations of data points are represented by lighter markers.

rated contained important information for airlines, there exists many other factors that were not considered. For example, flight delays, cancelled flights, and comfortable traveling experiences would also significantly affect the user's rating. To account for these potential key factors, these text words (full list in Appendix A.2) were chosen to be included in our next linear regression models. First, a many-hot encoding method was performed to create bag-of-words features containing thirty commonly used words in the text reviews. These chosen texts were word(s) that would most likely positively or negatively impact the overall rating, such as "good", "comfortable", "delayed", and "cancelled". These text features were implemented in the data model and trained against the training set. Like the previous trained model, the L1 loss function and L2 loss function were used to achieve the following training and testing MSE as shown in Table 2. Including the bag-of-words features improved the prediction for the L1 loss function by 15% and the prediction for the L2 loss function by 5%.

Table 2: Training and Testing MSE Using Survey Sub-Ratings and Text Bag-of-Words

Loss Function	Train MSE	Test MSE
L1 Loss	2.207	2.029
Quad Loss	2.149	1.984

In Fig. 6, the bins across the diagonal of the marginal histogram became lighter in shade in comparison to the first marginal histogram in Fig. 4. This visual shows that the data points gravitated towards the truth diagonal that represents

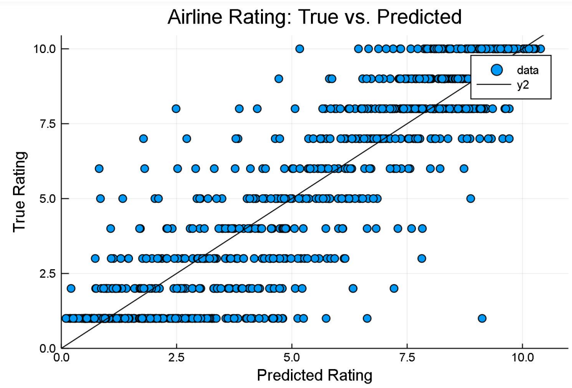

Figure 5: True vs. Predicted Airline Ratings data points (blue) calculated by L2 loss function with ideal trend line (black) using sub-rating and bag-of-words features

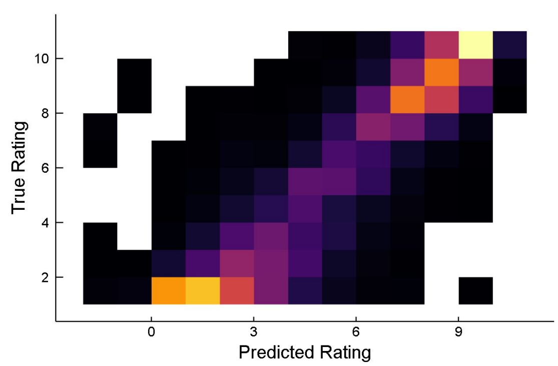

Figure 6: True vs. Predicted Airline Ratings Marginal Histogram for five sub-rating and bag-of-words features calculated by L2 loss function. Higher concentrations of data points are represented by lighter markers.

the ideal truth vs. prediction trend line.

3.1.2 Sentiment Analysis

In addition to bag-of-words, sentiment analysis was performed on all text reviews. This was done using version 3.4.5 Natural Language Toolkit for Python (NLTK). Specifically, we used the Valence Aware Dictionary and sEntiment Reasoner (VADER) Lexicon, a rule-based tool for sentiment analysis that was trained on social media. Since social media is a decent representation of colloquial writing, we saw it fit to run the text body of the review through such a tool. The function we used was the VADER sentiment intensity analyzer. This function returns a positive, negative, and neutral score between 0 and 1. The positive and negative scores indicate how much of the analyzed text has positive or negative sentiment (with 0 being the lowest). The neutral score indicates the polarity of the statement (again with 0 being the lowest). The compound score is an aggregate score computed between -1 and 1 which captures both polarity and sentiment. The more negative a number is the more negative the overall text was, and vice versa.

We applied the sentiment intensity analyzer to the text in two different ways. In the first way, we put the entire body of the text review for each review through the analyzer and found the positive, neutral, negative, and compound scores for the overall. In the second way, we used the NTLK tokenize tool to tokenize the text body of the reviews into separate sentences. We then ran each individual sentence through the analyzer and then averaged the scores over the number of sentences in the review.

Incorporating the positive, neutral, negative, and compound scores for both the sentenced and tokenized text structures improved the prediction for the L1 loss function by 6% and the prediction for the L2 loss function by 5%.

Table 3: Training and Testing MSE Using Survey Sub-Ratings, Text Bag-of-Words, and Sentiment Analysis

Loss Function	**Train MSE**	**Test MSE**
L1 Loss	2.010	1.909
Quad Loss	1.934	1.894

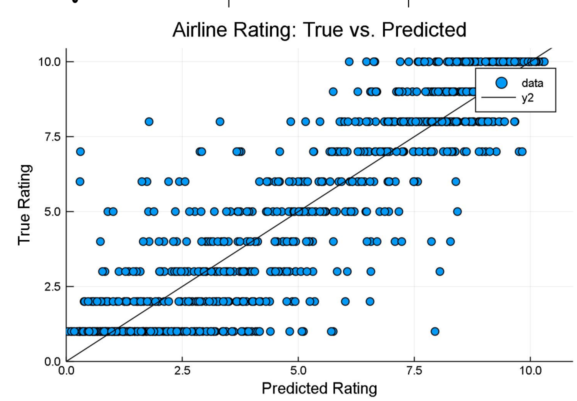

Figure 7: True vs. Predicted Airline Ratings data points (blue) calculated by L2 loss function with ideal trend line (black) using sub-rating, bag-of-words, and text sentiment analysis features

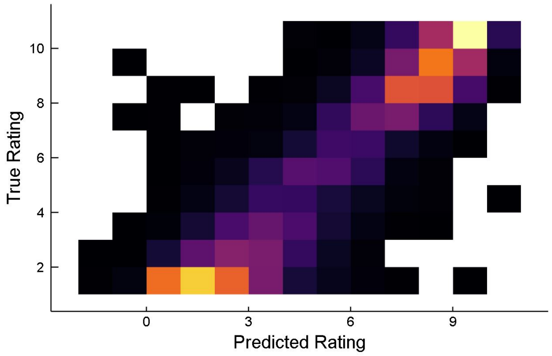

Figure 8: True vs. Predicted Airline Ratings Marginal Histogram for five sub-rating, bag-of-words, and text sentiment analysis features calculated by L2 loss function. Higher concentrations of data points are represented by lighter markers.

3.1.3 Extracting Route Information

Note: The features were generated in this section were too large to push to Github. Thus, we have made the .CSV available here (link may not work on GitHub previewer).

Finally, we looked to airports along the routes of the reviewers. These features were extracted from the text reviews to understand if specific airports helped with predicting the user's overall rating. We noticed that in the majority of reviews, this information was mentioned using the airports' International Air Transport Association (IATA) codes. These codes all consist of three capital letters and are unique to each airport. We wrote a regular expression that would essentially find three capital letters in a row and either proceeded or followed by a space, dash, or parentheses. We also made a list of exceptions such as "USD" (US dollars) and "CRJ" (used to designate planes made by Canadair Regional Jet). The names of all unique IATA airport codes were stored and these features were encoded using a one-hot encoding scheme. This encoding method created a binary column for each airport category for the user's flight, in which columns correspond to different airports, such as "SFO", "SJC", and "IAD". Due to the immense diversity of user reviews, the number of airports recorded was very large with 1914 unique airport labels.

Training the new model with the added airport features produced the training and testing MSE shown in Table 4. The insignificant change in test MSE could be due to the combination of two reasons. The number of unique airport labels was incredibly large and therefore the data for each individual airport was very small in comparison to the overall testing data set. As a result, the airport name features had a very small impact on the overall training model. Also, the insignificant change in MSE for both the training and testing data suggests that the specific airport has very little influence on the user's overall rating. Due to its low impact and significant increase in computation time, the airport features were not included in the final supervised learning model.

Table 4: Training and Testing MSE Using Survey Sub-Ratings, Text Bag-of-Words, Sentiment Analysis, and Airport

Loss Function	Train MSE	Test MSE
L1 Loss	2.007	1.910
Quad Loss	1.919	1.894

3.1.4 Ordinal Encoding

One of the final feature engineering methods we exercised was viewing the sub-ratings as ordinal features and transformed them with a binary coding scheme. The sub-ratings were ranked from 0 to 5 and these sub-ratings were converted to binary features for our new model. This was done to prevent any unwanted influences that the sub-rating scheme provided to the model (e.g. a rating of 5 isn't necessarily five times better than a rating of 1). This new feature significantly improved the training and testing MSE and ultimately used in our final model shown in Figure 8. However, although the success of the model improved, the solution weights are much more difficult to interpret in comparison to interpreting the weights of each individual sub-rating because the ordinal encoded weights no longer match to an interpretable label.

3.2 Classification Model

In addition to predicting the customer's overall rating from the customer's text reviews and experience sub-ratings, classification models were developed to predict whether the customer would recommend or would not recommend the airline. The progression of added features followed the same evolution that was taken with the linear regression model mentioned in the "Regression Model" section. Classification models were trained by incorporating a combination of the customer sub-ratings, many-hot encoding with commonly used word(s), text review sentiment analysis, and airport labels and shown in Tables 5, 6, and 7. The final models are shown in Table 9. The logistic loss and hinge loss functions were implemented to predict the user's recommendation.

Table 5: Training and Testing MSE Using Survey Sub-Ratings

Loss Function	Train Error	Test Error
Log Loss	9.41%	8.73%
Hinge Loss	8.95%	8.52%

Table 6: Training and Testing MSE Using Survey Sub-Ratings and Text Bag-of-Words

Loss Function	Train Error	Test Error
Log Loss	9.88%	9.70%
Hinge Loss	8.17%	7.94%

Table 7: Training and Testing MSE Using Survey Sub-Ratings, Text Bag-of-Words, and Sentiment Analysis

Loss Function	Train Error	Test Error
Log Loss	8.17%	7.93%
Hinge Loss	7.98%	7.50%

The addition of the text bag-of-words and sentiment analysis features generally the misclassification rate error by 12% while the airport word features made no significant impact. The final logistic loss model can be visualized by referring to Figure 9. Although it seems that the logistic loss prediction model poorly classifies the data in Figure 9, the visual itself is slightly misleading due to the high number of data points. Due to the 93% success classification rate, as shown in Table. 7, it can be assumed that the large majority of data points are along the correct side of the logistic prediction.

4 Results/Discussion

4.1 Obtaining the Result

To obtain our final model results, we cross-validated the results using the L2 loss function for our regression model and the hinge loss function for our classification models based on our analysis

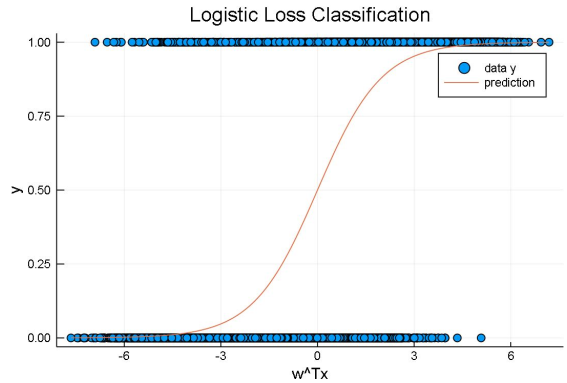

Figure 9: Logistic Loss Classification with sub-rating, bag-of-words, and text sentiment analysis features. Blue data points represent boolean recommended data and orange line represents recommended prediction.

results. The data set was randomly split into testing and training sets, 80% and 20%, respectively, 50 times and both our interpretable model and highest performing model were trained. The major difference between the two models is that the interpretable model contains direct features with the user's sub-ratings while the performance model implements the ordinal encoding scheme for the user's sub-ratings. For each learning model, the resulting w_{L2} and w_{hinge} and their respective training MSE and misclassification rates were stored in separate arrays. All values were then averaged. The final results are shown in Tables 8 and 9.

Figure 10: Training and Test MSE vs. λ for supervised learning regression model

For these models, we did not implement a regu-

Table 8: Final Training and Testing MSE Using Survey Sub-Ratings, Text Bag-of-Words, and Sentiment Analysis. Interpretable Model directly attaches to sub-ratings and Performance Model exercises ordinal encoding scheme

Model	**Train MSE**	**Test MSE**
Interpretable	1.927	1.916
Performance	1.577	1.592

Table 9: Final Training and Testing Misclassifation Rates Using Survey Sub-Ratings, Text Bag-of-Words, and Sentiment Analysis. Interpretable Model directly attaches to sub-ratings and Performance Model exercises ordinal encoding scheme

Loss Function	**Train Error**	**Test Error**
Interpretable	7.98%	7.53%
Performance	6.70%	6.75%

larization parameter because our model was most accurate when $\lambda = 0$ regardless of the regularization function (L1, L2, NonNegativeConstraint). Multiple models were trained with increasing λ values and the MSE was lowest when $\lambda = 0$ as shown in Fig. 10. Due to the relatively low number of features in comparison to the number of data points, the optimal λ equaling zero suggests that our training model is dominated by bias rather than variance. This discovery was first made when the training model only included the data set's sub-rating reviews. To counter this challenge, a large part of our analysis efforts was to extract more features from the text reviews provided in the data set. As mentioned previously, text features included bag-of-words, many-hot encoding, sentiment analysis, and airport labels. Although these additional features improved both the regression and classification predictions on the training data sets, the optimal regularization parameter, λ, was still zero.

4.2 Interpretation of Results

For our initial l_1 and l_2 regression models, we only included seat comfort rating, cabin staff rating, food and beverage rating, in-flight entertainment rating, and value money rating. The resulting w looked like the following array of weights (including an offset):

$$w_{L1} = [0.436, 0.545, 0.326, 0.068, 0.599, -0.051]$$

$$w_{L2} = [0.427, 0.658, 0.171, 0.0205, 0.957, -1.448]$$

Both the l_1 and l_2 came to the same conclusions in terms of order of weight importance, although the emphasis was more spread out for the l_1 model. The resulting weights of w suggest that the money value rating has the largest influence over the user's overall rating, followed by cabin staff rating and seat comfort rating. Least important were food/beverage rating, and in-flight entertainment rating.

After adding in text features in our final model, we determined that l_2 was the most interpretable model with the lowest test MSE of 1.916. We thus examine the weights of the final l_2 model in a similar fashion, though there are many more features and thus many more weights to look at. We find that value/money rating still has by far the largest weight, at about 0.87. Cabin staff rating and seat comfort rating also follow with weights of 0.61 and 0.40, respectively.

Looking at our key words/phrases, we find that the phrase with the largest weight is "on time" with a weight of 0.20. The next most influential weight is for the phrase "delay" which has a weight of -0.19. This makes sense since both terms seem to show a consistent emphasis on timeliness of an airline, with "on time" creating a positive change in overall rating and "delay" resulting in a negative change in overall rating. Outside of timeliness words, we find the next most influential to be "seat" with a weight of -0.062 and "crew" with a weight of 0.029. Thus, it seems like when an airline passenger notices their seat in a review, it is likely a complaint. However, when a passenger notice their crew in a review, it is likely a positive experience.

4.3 Confidence in Results

We are rather confident in our results. This is because the results we found are consistent with both our intuition and what has been generally reported of the airline market. *However*, this also means that we, unfortunately, did not find anything too ground shaking in this analysis. We know that often the biggest factor in which airline to take is simply the money/value of the ticket. So it is no surprise that that is the biggest factor in airline satisfaction. Next to that, some of the biggest talking points of an airline's service regard the friendliness and culture of cabin crew as well as the comfort of the seats, amount of legroom, etc. We have also established that no one likes it when a flight is delayed or cancelled.

Because our results appear to be accurate and intuitive (and frankly somewhat obvious for some), it wouldn't hurt to use our model to make company decisions. An interesting use for our model may be gathering information on a customer's perception of an airline's cabin crew and seat comfort through other means (since not everyone leaves a review on Skytrax) then attempting to predict their overall satisfaction level and making company decisions accordingly.

5 Ethical Assessment

5.1 Is our project a "weapon of math destruction?"

We have determined that our project is likely not a "weapon of math destruction." For one, its outcome is easy to measure. We can apply our model to any new reviews on Skytrax and measure its performance by examining its error in predicting the overall rating of a review.

The predictions are also not likely to harm anyone. We predict the overall rating of an airline based on customer satisfaction factors such as seat comfort and food/beverage. However, an airline's safety reputation reigns supreme over any sort of seat comfort or food it can provide. Therefore, the factors that we deem to be important are merely factors that may give an airline a competitive edge in the market, but cannot supersede the importance of safety.

Lastly, we do not think our analysis will create a feedback loop. This is because we are not using our analysis to generate a new rating for airlines. Rather, we are trying to understand what factors contribute most to why an airline gets the rating that it gets. As such, as customer needs change, that will be reflected in their very own review, and that trend will be picked up by the model.

5.2 Is fairness an issue for our project?

After careful assessment, we have determined that fairness is not a major issue for our project. This is because we do not believe we have features that vary drastically per a given demographic trait. We do not look at anything pertaining to demographic information of the poster of the review, nor anything having to do with the status, national origins, alliance membership of the airline. Rather, we are simply looking at the factors that all customers use to rate satisfaction of all airlines on the Skytrax website. In essence, our models are trying to study the airline market as a whole and make comments on the trends in that market without assigning any sort of result to any specific group of customers or airlines.

6 Conclusion/Future Work

In conclusion, we found that by far the most important factor to customers' airline satisfaction is how much value is in the ticket price. Beyond that, customers place a much higher emphasis on quality of the cabin staff and seat comfort over quality of food/beverage and in-flight entertainment.

Our text analysis revealed that the timeliness of the airline is also critical. Beyond that, it seems that the seat is a more of a "hygiene" factor. If it's bad, it's more likely to be discussed in the report. The friendliness or professionalism of the crew is likely a "motivational" factor that motivates people to write good reviews about their experiences.

Though our model didn't uncover anything that would alter the field, we do believe that with some future work, our model could be quite useful in examining trends in the airline industry. It would be entirely feasible to scrape future reviews on Skytrax apply our model once again. One can then examine trends by studying the relative change in weights in our model to see if customer preferences are shifting. It is our hope that this will allow airlines to understand their customers better and stay relevant in the market.

A Appendix

Dataset Detail

Table A.1: Detailed description of the different features in our dataset.

Feature Name	Type	Number of Entries
airline_name	Nominal	41396
link	Nominal	41396
title	Nominal	41396
author	Nominal	41396
author_country	Nominal	39805
date	Nominal	41396
content	Text	41396
aircraft	Nominal	1278
type_traveller	Nominal	2378
cabin_flown	Nominal	38520
route	Nominal	2341
overall_rating	Ordinal	36861
seat_comfort_rating	Ordinal	33706
cabin_staff_rating	Ordinal	33708
food_beverages_rating	Ordinal	33264
inflight_entertainment_rating	Ordinal	31114
ground_service_rating	Ordinal	2203
wifi_connectivity_rating	Ordinal	565
value_money_rating	Ordinal	39723
recommended	Binary	41396

Bag of Words

The following is the list of key words we identified for our bag of words analysis:

- service
- food
- seat
- crew
- staff
- entertainment
- boarding
- experience
- leg
- attendants
- offered

- free
- delay
- water
- attentive
- cancelled
- waiting
- on time
- business class
- cabin crew
- flight attendants