Airline Satisfaction Analysis

Yoyo Lu (yl3397), Xinhao He (xh245), Bonju Koo (bk368)
Department of Operations Research and Information Engineering, Cornell University
ORIE 4741: Learning with Big Messy Data
May 10, 2024

Table of Contents

1. Introduction

a. Problem

As the global air travel industry recovers and evolves post-pandemic, airlines are seeking to maximize passenger demand and expand their influence in the market. With high competition across the most popular routes around the world, airlines need to identify the most crucial factors to provide them with the advantage. For example, are passengers sensitive to ticket prices and therefore, airlines should focus on providing the best value to their customers? Or alternatively, are passengers more interested with the best flight services and amenities, with greater flexibility towards pricing? With escalating competition and substantial revenue implications, understanding and refining the factors that provide the best passenger experience is an essential business strategy for airlines. This study aims to address these concerns and provide preliminary insights for airlines to allocate resources effectively and provide the best customer satisfaction.

b. Dataset

We selected a dataset of airline reviews from 2013 to 2024, compiled by Kaggle user Sujal Suthar. It includes feedback on the world's top 10 rated airlines in 2023, including Singapore Airlines, Qatar Airways, All Nippon Airways (ANA), Emirates, Japan Airlines, Turkish Airlines, Air France, Cathay Pacific Airways, EVA Air, and Korean Air according to Skytrax's World Airline Award. The dataset provides insights for each passenger review, covering airline, flight class, travel type, flight period, and ratings on multiple service categories like seat comfort, staff service, food and beverages, inflight entertainment, and value for money. A full table of the dataset features can be found in the Appendix, under Dataset Description. Using these features, our analysis will identify the factors that most significantly affect the passenger's overall rating and provide a blueprint for airlines to capitalize on to continue improving upon their quality customer experience.

2. Data Processing

Prior to performing data analysis, feature engineering and preliminary data exploration were performed to prepare the dataset and identify potentially significant variables for analysis. Initially containing 17 features ranging from textual, categorical, ordinal, time series, and continuous types, we limited our analysis to comprise of 6,216 verified reviews from passengers confirmed to have traveled with the airlines.

a. Feature Engineering

<u>Time Series Data</u>

Each review included both a travel date and a review date. While review date is specified in a datetime format, travel dates only include month and year descriptions, in the format "Month Year". To properly work with time series data, we mapped travel dates from the textual description to a numerical representation. To ensure that the reviews within the dataset convey the most accurate and up-to-date information for each flight, we decided to limit our analysis to reviews that were posted within 1 year of the passenger's travel date. This ultimately only excluded 5 reviews from our analysis, where the reviewers took over a year to leave a review.

Categorical Data

Categorical data, including a passenger's travel type and airline, were transformed using one-hot encodings. Since there were only 4 distinct travel types (Solo Leisure, Family Leisure, Couple Leisure, and Business) and 10 distinct airlines (top 10 airlines as rated by 2023 World Airline Awards), using one-hot encodings would not increase dataset's feature space too drastically. To avoid multicollinearity issues when performing data analysis using these categorical variables, we chose one category for each variable as the reference level and dropped that category. For travel type, we chose Business as the reference category. For airline travel, we chose the lowest-ranked airline among the top 10 awarded airlines as the reference category, which was Korean Air.

Another categorical variable that we chose to encode differently was the passenger's Flight Class. Flight class in the dataset included Economy, Premium Economy, Business, and First Class. Since there is a natural ordinal ranking to these categories, given different ticket prices and service expectations between the classes, these categories were encoded using ordinal values instead. Economy class was chosen as the reference level, with an ordinal value of 0. Each class upgrade was then assigned a higher ordinal value, with Premium Economy at 1, Business at 2, and First Class at 3.

Textual Data

Beyond the ratings and categorical data that each review includes, a major component that we also want to investigate is the textual reviews that passengers provide. For this objective, we employed the use of the Universal Sentence Encoder (USE), a pre-trained language model that encodes arbitrary-length text into a 512-length feature vector. The dimensionality of the resulting feature vectors would likely result in overfitting during data analysis, since we initially started with 17 features. To alleviate this problem, we will use Principal Component Analysis (PCA) in our analysis to reduce the feature space of the encodings while still retaining most of the information found in the texts.

b. Data Visualization

After preparing the dataset and making the necessary feature transformations, we explored the data distributions and identified useful features in performing our analysis.

Service Categories

For each review, passengers rated the five service categories (Seat Comfort, Staff Service, Food & Beverage, Inflight Entertainment, and Value For Money) from a scale of 1 to 5. To examine the data distribution of these ratings, we plotted a histogram of the data, found in Figure 1. From this histogram, we observed that most categories have a good distribution balance, with each category containing over 500 data samples in each rating. Only Staff Service appears to be exceptionally left-skewed, with 42% of all reviews rating Staff Service at 5. To correct for skewness in the data distribution, we performed standardization in our analysis so that the service categories are more evenly distributed among each rating.

Since it is possible for an airline to provide similar services across the service categories, a possible concern is high correlation between these variables. To explore this concern, we plotted a heatmap of the correlation between the five service categories. We additionally included flight class (Economy, Premium Economy, Business, and First Class) in the correlation matrix, since it is reasonable to expect that higher flight class should correlate with better

services. In Figure 2, the correlation heatmap indicated that such concerns were not prominent within the dataset, as the variables were not highly correlated. Although there were positive relationships between the variables as expected, the strength of these relationships are relatively weak (highest correlation of 0.2), suggesting that these variables do not affect each other too drastically.

Data over Time

Since flight patterns tend to follow seasonality trends, we wanted to look for skewness over time within the dataset. In Figure 3, we plotted a histogram of both the number of reviews and number of flights over the months to examine the data distributions. From this histogram, we found that there were no obvious outliers among the months, with each month containing a good portion of the total dataset. Interestingly, the number of reviews and flights appear to increase at the start and end of the year, with a dip in the middle around June, which contrasts the seasonality trends that were expected in flight data. This could suggest that the passengers who leave reviews do not follow the same trends as flights themselves, leading to a more balanced dataset for reviews.

3. Data Analysis

For convenience, all model evaluation metrics discussed in this section can be collectively found in the Appendix, under Model Performance Summary.

a. Linear Regression

To first identify the variables that best predict passenger ratings for an airline, our objective was to model a linear relationship between the features and the outcome, overall ratings. This was accomplished through Ordinary Least Squares (OLS) regression, using the mean squared error (MSE) as the model evaluation metric.

Numerical & Categorical Variables

The first linear regression model we fit combined the numerical and categorical variables in the dataset; this included variables such as month flown, passenger ratings for each of the five service categories, the passenger's flight class, the passenger's travel type, and airline. Using the one-hot encodings and dropping a category as the reference level, as described in the Feature Engineering section, we have a total of 6,211 data samples and 19 features, plus a constant coefficient. Since the dataset is not too large and as a result, the computational costs are not too expensive, we opted to use 5-fold cross validation to evaluate model performance. With MSE as the evaluation metric, our OLS model achieved a mean MSE of 2.513 with a standard deviation of 0.5634. Since overall rating ranges from a scale of 1 to 10, a mean MSE of 2.513 suggests that the model is able to capture a lot of the relationship between the features and the overall rating without much error; the standard deviation of 0.5634 also indicates that the model is not too sensitive to training data, as the errors do not vary significantly across the splits.

Feature Significance and L1 Regularization

One factor that we want to consider is whether all the features used in the OLS model are significant in predicting the outcome. Using the full model summary as presented in Figure 4, we can perform a multiple-hypothesis test on the model features using the p-values and a

significance level of 0.05. As a result, only 9 features were found to have statistically significant relationships with the overall rating: Seat Comfort, Staff Service, Value For Money, and Class; the 4 airlines All Nippon Airways (ANA), Qatar Airways, Singapore Airlines, and Turkish Airlines; and the travel type Solo Leisure. Among this result, most of these outcomes could be expected; it is reasonable that passengers factor staff service, value for money, and their flight class more greatly into their travel experience than other categories, such as inflight entertainment. 3 of the significantly relevant airlines were listed as the top three airlines of 2023 (ANA, Qatar Airways, and Singapore Airlines), so it is also reasonable that they would have a statistically significant effect on passengers' overall ratings.

In addition to performing hypothesis testing, another method to identify feature importance is to apply L1 regularization to our OLS model, also known as Lasso Regression. L1 regularization adds an absolute penalty to coefficient values, which allows the model to perform feature selection and push unnecessary feature weights to zero. Since lasso regression is not scale-invariant like OLS, we first standardized all features to have mean 0 and variance 1. Next, we used grid search to identify the optimal regularization parameter λ, searching over possible values from 0.01 to 1. With the optimal λ of 0.02, we used 5-fold cross validation to find the mean MSE of 2.503 and standard deviation of 0.5643. This performance is slightly better than OLS, which indicates that regularization does help improve model robustness and generalization. We found that there are 12 features with non-zero coefficients in lasso regression, as shown in Figure 5. Nine of these features are the same as the significant features found in the OLS model, with the 3 additional features being Food & Beverages and two airlines, Air France and Emirates. Since L1 regularization did not fully push these coefficients to zero, it is possible that these additional features have some positive but miniscule effect on passengers' overall ratings.

Principal Component Analysis

As we identified during feature engineering, using the full textual embeddings with 512 features could lead to overfitting. We therefore chose to reduce the dimensionality of the feature vectors by using PCA to identify a lower-rank approximation of the embeddings. To compare the results of using different dimensions, we first used 5-fold cross validation to compute the MSE using the full embeddings, together with the other features mentioned above. Then we tested low-rank approximations of the full embedding by intervals of 50 and similarly used 5-fold cross validation to compute the resulting MSE. In Figure 6, we see that the MSE is minimized by using a rank 250 approximation, with better performance than any other approximation and the full textual embedding. We therefore chose to use a 250-rank approximation of the full textual embeddings, together with the relevant features identified through hypothesis testing and lasso regression, to create our final linear model. Using the same cross-validation technique as before, we found that including the textual embeddings yielded a mean MSE of 1.958 and a standard deviation of 0.5035, outperforming both the OLS and lasso regression models. This outcome suggests that textual reviews are a significant factor as well in determining passenger ratings and would be a relevant area to focus on to achieve optimal customer satisfaction.

b. Decision Tree

In addition to linear regression, which captures the linear relationship between the overall rating and features, we wanted to explore whether there existed non-linear relationships in the data. To capture non-linearities, we chose a decision tree model, which partitions the feature space into distinct regions by minimizing impurity. To reduce the possibility of deep trees

overfitting the data, we first used 5-fold cross validation to identify the optimal depth of the trees. Since the MSE is minimized at a max depth of 4, as Figure 7 illustrates, we opted for a regression tree with a max depth of 4. Using cross validation to evaluate the performance of the depth-4 regression tree, we achieved a mean MSE of 2.446 and a standard deviation of 0.5921. Surprisingly, this model performed worse than the full OLS model, while also having higher variance compared to any of the linear models. We opted for two common ensemble techniques, random forests and tree boosting, to address the concern of high variance and improve model performance.

Random Forest

The first ensemble model we looked at is the random forest, which aggregates a collection of simple regression trees to form a collective prediction. Since each tree is likely to make different types of errors, we would expect the random forest to outperform the single regression tree from above. Using the same max depth as the regression tree, we fit 100 random trees on the full dataset and used cross validation to evaluate the performance of the aggregated random forest. The random forest achieved a mean MSE of 2.249 and a standard deviation of 0.5626, which aligns with the expectation of an ensemble model. We see that the standard deviation is indeed lower than the single regression tree, while the random forest also achieved a better performance in terms of mean MSE. However, the random forest is still unable to outperform the full OLS model, which may indicate that non-linearities in the data were not as prevalent as we initially expected. To further improve upon the decision tree performance, we pivoted to another ensemble approach involving gradient boosting.

Tree Boosting

In contrast to random forests, which train a collection of decision trees in parallel, tree boosting as an ensemble method trains trees in sequence. By approximating the process of gradient descent and having each tree iteratively improve upon previous trees' errors, the aggregated model would also be expected to outperform the single regression tree. Using the same hyperparameters as the random forest, we fit 100 boosted trees with max depth of 4 on the full dataset and used cross validation to evaluate the performance of the aggregate model. Tree boosting achieved a mean MSE of 2.154 and a standard deviation of 0.5479. We therefore see that tree boosting is able to outperform the random forest in both the mean MSE and the model variance. However, we still have not achieved a performance similar to the full OLS model, which as we recall achieved a mean MSE of 1.958 and a standard deviation of 0.5035. Also taking into consideration the additional computational complexity of decision trees, it therefore appears that deploying the full OLS model would be better suited for an airline's everyday operations.

Feature Importance

Although our decision trees have been unable to match the performance of our linear regression model, we can use the models to provide additional insight into which features are most significant in affecting passenger ratings. For decision trees, we evaluated a feature's importance by computing its Mean Decrease in Impurity (MDI); simply put, a feature with higher MDI provided much greater improvement across all the splits using that feature, indicating that that feature is significant for the model. Figures 8 and 9 show the MDI for the top 10 most significant features in the Random Forest and Tree Boosting respectively. As both

figures prominently indicate, Value For Money was by far the most significant feature in determining passenger ratings. This aligns with the results from our linear regression model, where Value For Money similarly had the largest coefficient. Also relevant, though to a lesser degree, is the passenger's flight class, which again is related to the value that passengers expect from their flight experience. As a result, we see a common trend that focusing on customer value would likely have the greatest impact on improving passenger satisfaction.

4. Conclusion

The coefficients and results derived from our models provide actionable insights that can inform specific business decisions within the airline industry. In this section, all results mentioned are based solely on results from linear models for congruency.

While aspects like inflight entertainment, food & beverages, seat comfort, and staff service are intuitively important, our analysis shows that these four features do not contribute a lot to the customers' overall satisfaction. Inflight entertainment and food & beverages have a nonsignificant P-value and do not exhibit a significant relationship with overall ratings. Seat comfort and staff service are both significant predictors with similar coefficients around 0.03. This means that while higher scores in both are statistically correlated to a higher overall rating, their influence on overall rating is relatively small.

By far, the most influential feature is value for money, with a high coefficient of 1.88. This means that passengers place significant emphasis on the perceived value-for-money of their flight. Airlines looking to improve passenger satisfaction and overall ratings should prioritize efforts to enhance the perceived value of their flights relative to their pricing. This could involve offering competitive pricing, providing added amenities or services that justify the cost of the ticket, and ensuring transparency and fairness in pricing practices.

When it comes to class, our analysis shows that for each unit increase in the class (e.g. Economy to Business, Business to First), we expect an increase in overall satisfaction by approximately 0.224 units, all else held equal. This suggests that passengers flying in higher classes tend to report higher levels of overall satisfaction compared to those in lower classes. It aligns with the intuitive expectation that passengers paying for premium services and amenities are likely to have higher expectations and subsequently rate their overall experience more positively.

The coefficients of different airlines flown offer insight into the performances of certain aviation companies. As mentioned before, All Nippon Airways (coefficient = 0.47), Qatar Airways (0.38), Singapore Airlines (0.41), and Turkish Airlines (-0.36) are the four significant predictors for overall satisfaction. These findings enable airlines to conduct peer benchmarking and identify best practices or areas for improvement based on the performance of their competitors. An aviation enterprise aiming to improve its customer satisfaction could delve into what ANA, Qatar Airways, or Singapore Airlines are doing well and investigate what Turkish Airline has been lacking.

Our analysis shows that amongst three types of travelers (solo, family, or couple leisure), only solo leisure is a statistically significant predictor with a coefficient of 0.13. Solo travelers tend to rate their overall flight experience more positively. This could be due to a variety of factors, such as the flexibility and independence associated with solo travel, or the fact that solo travelers may have different expectations or standards compared to families or couples. However, without further information, it is hard for a business to change its practices solely based on this result.

Discussion of Fairness

Fairness refers to the unbiased treatment of individuals or groups by an algorithm. It means that the algorithm's decisions should not be influenced by certain sensitive attributes such as race, sex, color, religion, disability, familial status, or nationality. These are the groups or individuals that should be protected from discrimination.

In our analysis, we have been careful to ensure that our model does not use information that could reveal sensitive attributes. This includes not only explicit attributes like race or sex, but also proxies that can indirectly reveal these attributes. But even seemingly innocuous information, like the type of traveler (solo, couple, or family leisure), could potentially be sensitive in certain contexts. For example, if families or couples tend to belong to certain protected groups more than others, then using traveler type as a feature could indirectly lead to discrimination against these groups. However, there is no evidence for such accidental discrimination in our model.

Appendix

Dataset Description

Feature	**Description**
Title	A short summary of the review
Name	Reviewer's name
Review Date	Date the review was made
Airline	The airline being reviewed. The dataset tracks the top 10 airlines as ranked by Skytrax World Airlines Award 2023.
Verified	Whether the information in the review has been verified. We limit our analysis to only verified reviews.
Reviews	A full text review of the airline
Month Flown	"Month Year" description of when the passenger's flight took place
Route	Passenger's flight route
Class	Passenger's flight class. Includes 4 classes in total: Economy, Premium Economy, Business, First Class
Seat Comfort	Rating on 1-5 scale for flight's seat comfort
Staff Service	Rating on 1-5 scale for flight's service
Food & Beverages	Rating on 1-5 scale for flight's food and beverages
Inflight Entertainment	Rating on 1-5 scale for flight's inflight entertainment
Value For Money	Rating on 1-5 scale for flight's value for money
Traveler Type	Purpose of the passenger's travel. Includes 4 categories in total: Business, Couple Leisure, Family Leisure, Solo Leisure
Overall Rating	Rating on 1-10 scale for passenger's overall experience
Recommended	Boolean for whether the passenger would recommend the airline

Model Performance Summary

	Mean MSE	MSE Standard Deviation
OLS (numerical & categorical features only)	2.513	0.5637
Lasso Regression	2.503	0.5643
Full OLS, with PCA for textual embeddings	1.958	0.5035
Regression Tree, depth = 4	2.446	0.5921
Random Forest, depth = 4 and n_trees = 100	2.249	0.5626
Tree Boosting, depth = 4 and n_trees = 100	2.154	0.5479

List of Figures

Figure 1: A histogram showing the data distribution of passenger ratings for each service category. Passengers rate each of the 5 service categories on a scale between 1 to 5 when they leave a review.

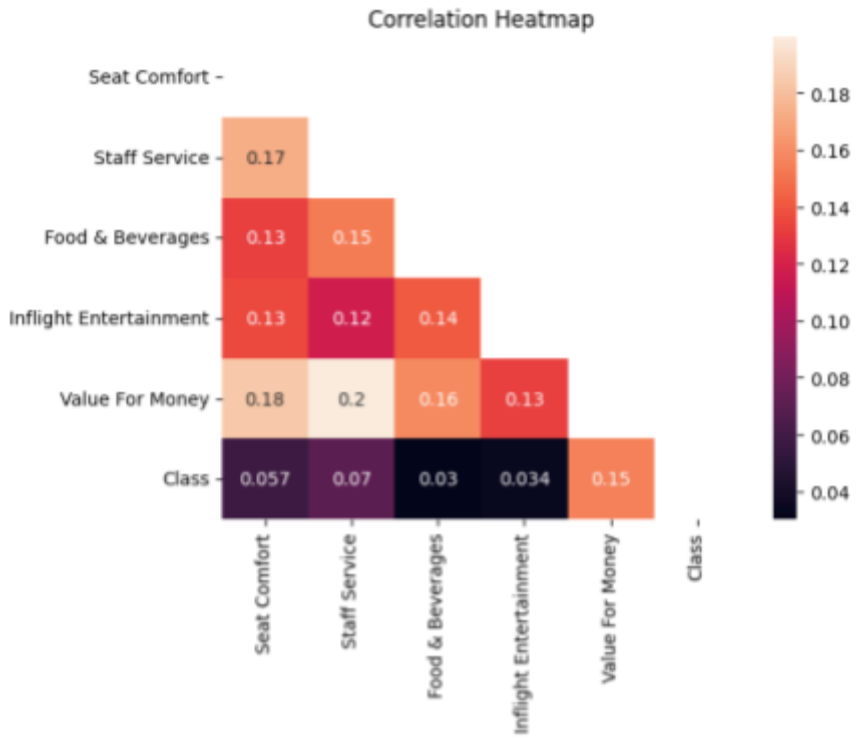

Figure 2: Correlation heatmap between each of the 5 service categories, as well as flight class. All variables have slightly positive correlations, indicating positive but weak relationships.

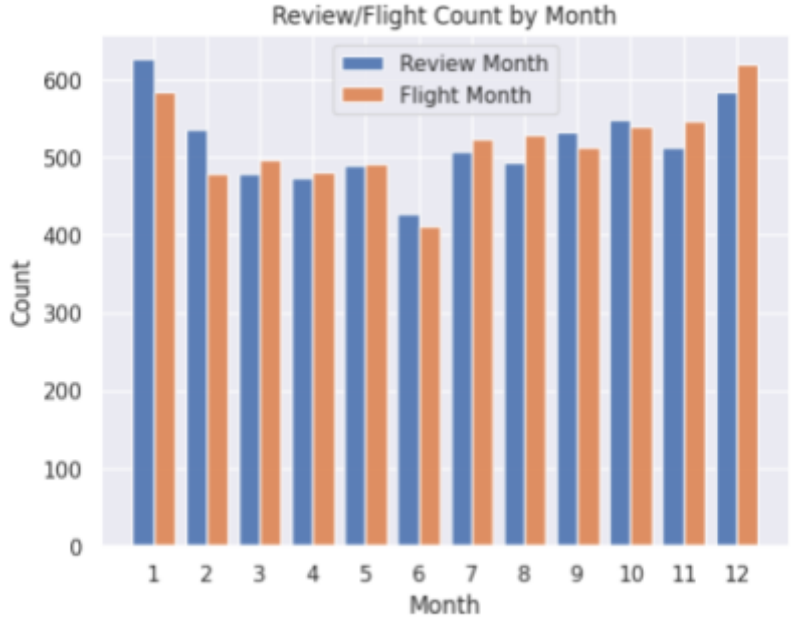

Figure 3: A histogram showing the distributions of reviews (in blue) and flights (in orange) per month. The distributions appear to be relatively uniform, with a slight decrease during June and slight increases in January and December.

3.a Linear Regression

OLS Regression Results

Dep. Variable:	Overall Rating	R-squared:	0.804
Model:	OLS	Adj. R-squared:	0.803
Method:	Least Squares	F-statistic:	1334.
Date:	Fri, 10 May 2024	Prob (F-statistic):	0.00
Time:	05:17:41	Log-Likelihood:	-11600.
No. Observations:	6211	AIC:	2.324e+04
Df Residuals:	6191	BIC:	2.338e+04
Df Model:	19		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.9153	0.183	-5.002	0.000	-1.274	-0.557
Seat Comfort	0.0363	0.015	2.384	0.017	0.006	0.066
Staff Service	0.0385	0.014	2.711	0.007	0.011	0.066
Food & Beverages	0.0152	0.014	1.060	0.289	-0.013	0.043
Inflight Entertainment	-0.0063	0.016	-0.396	0.692	-0.038	0.025
Value For Money	1.8836	0.014	137.865	0.000	1.857	1.910
Class	0.2248	0.022	9.990	0.000	0.181	0.269
Month Flown	0.0039	0.006	0.699	0.484	-0.007	0.015
Air France	0.0725	0.153	0.475	0.635	-0.227	0.372
All Nippon Airways	0.4667	0.178	2.627	0.009	0.118	0.815
Cathay Pacific Airways	0.2113	0.153	1.380	0.168	-0.089	0.511
EVA Air	0.1251	0.175	0.715	0.475	-0.218	0.468
Emirates	-0.0617	0.148	-0.418	0.676	-0.351	0.228
Japan Airlines	0.2383	0.185	1.289	0.197	-0.124	0.601
Qatar Airways	0.3814	0.145	2.624	0.009	0.097	0.666
Singapore Airlines	0.4064	0.150	2.718	0.007	0.113	0.700
Turkish Airlines	-0.3603	0.146	-2.462	0.014	-0.647	-0.073
Couple Leisure	0.0520	0.064	0.818	0.413	-0.073	0.177
Family Leisure	0.0225	0.068	0.332	0.740	-0.111	0.156
Solo Leisure	0.1299	0.058	2.249	0.025	0.017	0.243

Figure 4: Full model summary of OLS using numerical and categorical features to predict passenger's Overall Rating for an airline.

	Feature	Coefficient
0	Seat Comfort	0.044891
1	Staff Service	0.054887
2	Food & Beverages	0.011833
4	Value For Money	2.991452
5	Class	0.190246
7	Air France	-0.008173
8	All Nippon Airways	0.037303
11	Emirates	-0.067242
13	Qatar Airways	0.079391
14	Singapore Airlines	0.064269
15	Turkish Airlines	-0.196942
18	Solo Leisure	0.031681

Figure 5: Features with non-zero coefficients using Lasso Regression.

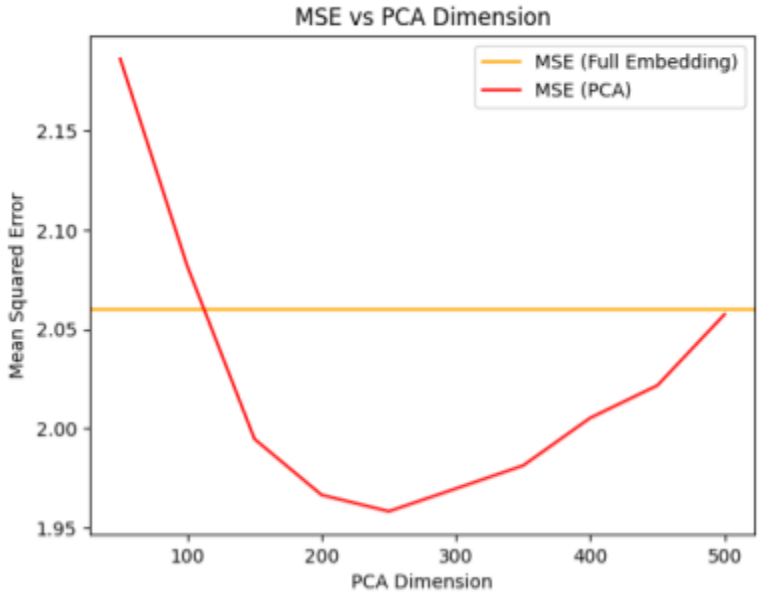

Figure 6: MSE for different low-dimensional approximations of the full textual embeddings. The full embedding contains 512 features, and the PCA approximations range from 50 to 500 at intervals of 50. 5-fold cross validation was used to obtain the MSE for each approximation.

3.b Decision Tree

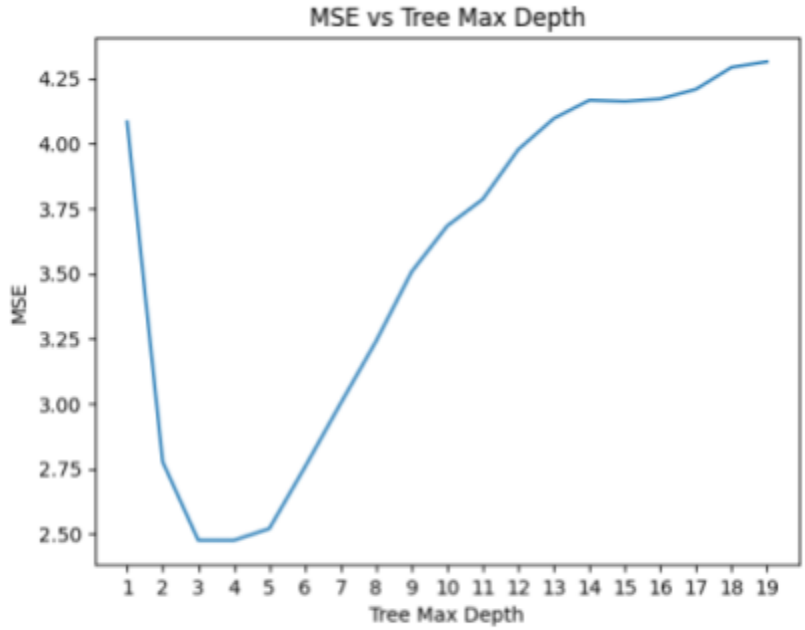

Figure 7: Cross-validation MSE for each tree depth from depth 1 to 20. The MSE is minimized around a max depth of 4.

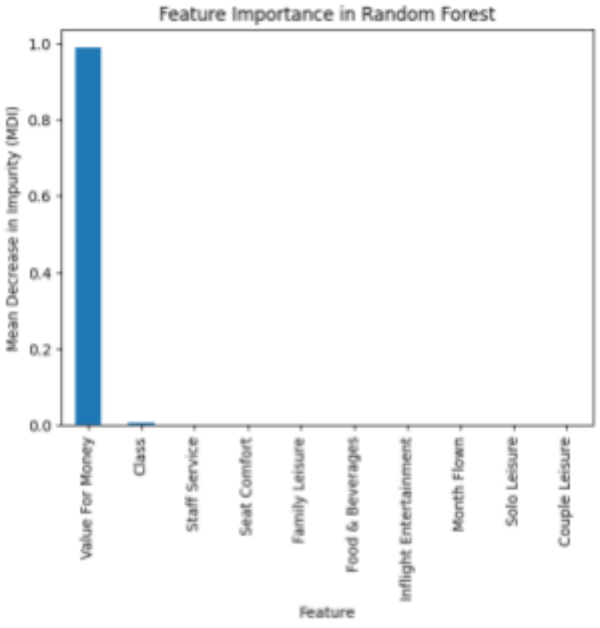

Figure 8: The Mean Decrease in Impurity (MDI) for the top 10 features in Random Forest.

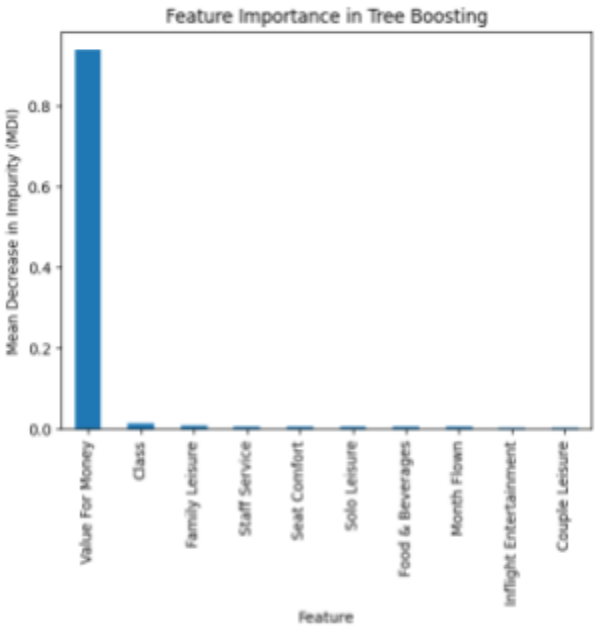

Figure 9: The Mean Decrease in Impurity (MDI) for the top 10 features in Tree Boosting.

References

Dataset Source: https://www.kaggle.com/datasets/sujalsuthar/airlines-reviews
Airline Reviews: https://www.airlinequality.com/review-pages/a-z-airline-reviews/
World Airline Awards 2023: https://www.worldairlineawards.com/worlds-top-10-airlines-2023/

External Links

Project Github Link: https://github.com/xinhaohe245/4741-project
(finished): https://github.com/yoyolu1124/ORIE4741_Final_Project/tree/main
Fairness: people.orie.cornell.edu/mru8/orie4741/lectures/fairness.pdf