Assignment 2: Flight Rating Prediction

Nan Huang, Baoni Li,
Wenqian Xu, Yi Zhang

Abstract

Our project utilizes a Kaggle dataset containing various airline reviews to predict overall ratings. The dataset reveals the prevalence of negative reviews and highlights specific areas that airlines need to address for improvement. An ensemble model combining TF-IDF and a random forest regressor outperforms other models, with the key to its success being feature representation and optimization.

The model combines TF-IDF transformed reviews and one-hot encoding to achieve optimal results through synergy. Hyperparameter tuning, especially for "n_estimators", balances complexity and performance. Efficient scaling, avoidance of overfitting, and robust handling of missing values all contribute to the success of the model.

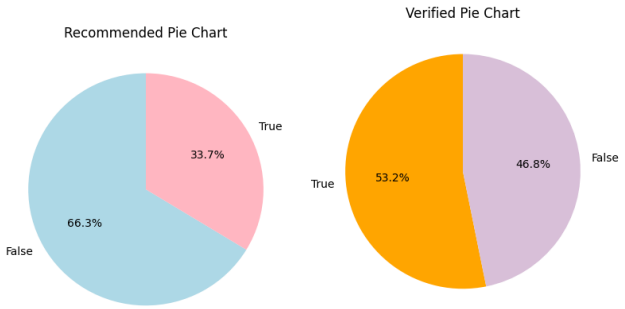
The literature review explores the origins of the dataset, previous studies on flight characteristics, attempts at sentiment analysis, and the impact of service information on ratings. The findings of existing studies are consistent with our findings, emphasizing the relevance of service-related features in predicting ratings.

Our ensemble model provides valuable insights into the airline industry, emphasizing the importance of specific service-related features in predicting overall ratings.

PART 1: Identify a Dataset to Study

The dataset we are using is: (https://www.kaggle.com/datasets/juhibhojani/airline-reviews/data)

The given file consists of 20 columns(including the index column) and 23171 rows.



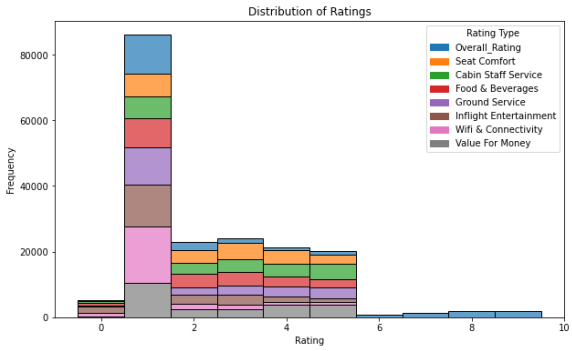
About 66% of reviewers do not recommend the airline, while the remaining 34% recommend the airline. This indicates that a significant portion of reviewers had a negative experience or were dissatisfied with the airline they took.

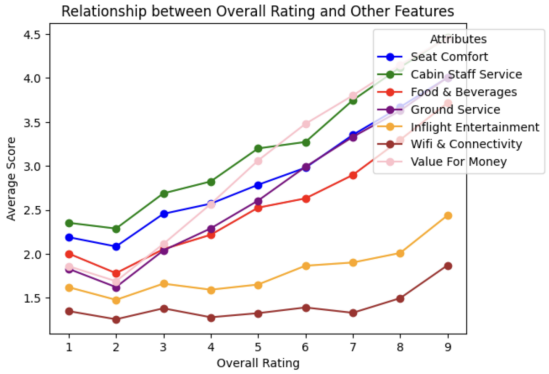
In terms of verification, 53% of the reviewers were verified, which suggests that a slight majority of the reviews could be considered more reliable or credible. However, since 47% of the reviewers were not verified, there is not much difference between the two. This suggests the mix of verified and unverified comments contributed to the overall feedback. This feature may not be a significant, influential factor in rating prediction.

Features of the Dataset look like the following:

Features of the Dataset:

Column	Description
Airline Name	Name of Airline (str)
Overall Rating	Rating given by user (str)
Review Title	Title of review (str)
Review Date	The date when the review was entered(str)
Verified (whether the review is verified or not)	Whether the reviewer is verified or not (boolean)
Review	Detailed review (str)
Aircraft	Aircraft type and Number (str)
Type of Traveller	Categorical class type (str)
Seat Type	Categorical class type (str)
Route	Flight source and destination (str)
Date Flown	Month and Year of flight (str)
Seat Comfort	Rating out of 5 (float)
Cabin Staff Service	Rating out of 5 (float)
Food & Beverages	Rating out of 5 (float)
Ground Service	Rating out of 5 (float)
Inflight Entertainment	Rating out of 5 (float)
Wifi & Connectivity	Rating out of 5 (float)
Value for Money	Rating out of 5 (float)
Recommended	Whether the flight is recommended or not (str, True or False)



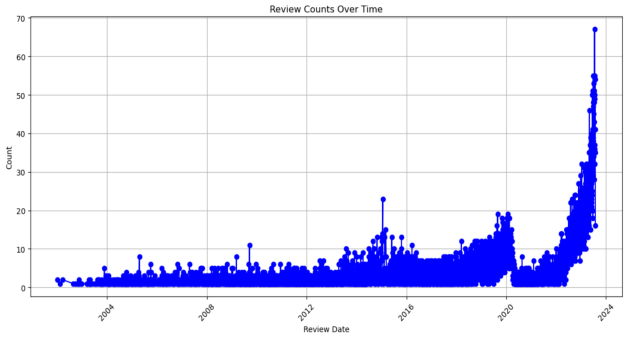


Overall ratings on a scale of 1 to 10 show a trend where a significant number of reviewers gave a rating of 1.

Similarly, there is a common trend of the lowest rating (1) peaking in specific areas such as "Seat Comfort", "Flight Attendant Service", "Food and Beverage", "Ground Services", "In-flight Entertainment", "Wifi and Connectivity", and "Value for Money" (all out of 5). This consistent trend suggests that reviewers tend to be dissatisfied with certain aspects of their flight experience, as reflected in the individual ratings.

In summary, the overall analysis indicates a high percentage of negative reviews and dissatisfaction across all aspects of the flight experience. This information is valuable for airlines to identify areas for improvement and to address issues raised by reviewers, especially for the areas with the most negative feedback. In addition, the mix of verified and unverified comments emphasizes the importance of considering both types of feedback in understanding overall customer sentiment.

In this line chart, we employed the average scores of various features corresponding to each Overall Rating to examine potential relationships between these features and the target variable, Overall Rating, which is the column we aim to predict. The analysis revealed a positive correlation between the Overall Rating and all the float features included in the chart. Consequently, we will incorporate all these float features into our predictive model.



In the analysis of review counts over time, we examined the number of received ratings and illustrated the trends. The dataset spans from the earliest recorded rating in 2002 to the most recent in 2023. Notably, a discernible upward trend indicates a progressive increase in the number of ratings over time.

PART 2: PREDICTIVE TASK

We want to predict the Airline's Overall rating. We will evaluate our model's performance by comparing the MSE, MAE, RMSE values between the actual value and the predicted value, and also R^2 values. We will compare the baseline model (linear regression) with the random forest regressor model and the XGBoost regressor model. We will assess the validity of our model by fitting them on the training set and predicting them on the test set to avoid overfitting problems.

We will use features: 'Review', 'Seat Comfort', 'Cabin Staff Service', 'Food & Beverages', ' Ground Service', 'Inflight Entertainment', 'Wifi & Connectivity ', 'Value For Money ', 'Type Of Traveller', 'Seat Type', and 'Recommend.' These are the features highly correlated with the overall rating we are trying to predict. For 'Seat Comfort', 'Cabin Staff Service', 'Food & Beverages', '

Ground Service', 'Inflight Entertainment', 'Wifi & Connectivity ', and 'Value For Money,' we left them as their original form, which is float. According to our Relationship Between Overall Rating and Other Features line chart, we can see a positive correlation between overall ratings and the float features we are using, so we believe these float columns are important features to predict Overall_Rating.

Model	MSE	FVU	MAE	RMSE	R2
Linear Regressor (Baseline)	5.134	0.639	1.575	2.266	0.361
Linear Regressor (+One hot encoding feature)	4.390	0.546	1.530	2.095	0.454
Random Forest Regressor (+One hot encoding feature)	3.482	0.433	1.154	1.866	0.567
XGBoost Regressor (+One hot encoding features)	3.280	0.408	1.138	1.811	0.592
TF-IDF + Linear Regressor	4.798	0.597	1.632	2.190	0.403
TF-IDF + Random Forest Regressor	3.045	0.379	1.142	1.780	0.621
TF-IDF + XGBoost Regressor	3.168	0.394	1.089	1.745	0.606

We used one-hot encoding on 'Type of Traveller', 'Seat Type', and 'Recommended' because they are categorical features. For the 'Review' column, we used TF-IDF to include it in the model because it's a string feature. We also handled missing values in the following columns: 'Seat Comfort', 'Cabin Staff Service', 'Food & Beverages', 'Ground Service', 'Inflight Entertainment', 'Wifi & Connectivity', and 'Value For Money' by employing imputation methods. For each missing value, we utilized the median value of the respective airline in that specific column for imputation. In cases where all values for a given airline in a particular column were missing, we imputed them with a value of 0.

PART 3: SELECT/DESIGN MODEL

Our model uses an ensemble of TF-IDF and Random Forest Regressor. The reason we used is that we get the lowest MSE in this model compared to our linear regression model, ensemble TF-IDF and Linear Regression model, and ensemble TF-IDF and Random Forest Regressor model.

Unsuccessful Attempts

One of the unsuccessful attempts we have is the TF-IDF + Linear Regression model. The overall performance is worse compared to the performance of Linear Regression by itself with the same features. This is because TF-IDF may introduce high-dimensional and sparse feature vectors, potentially leading to multicollinearity issues. This occurs when features are highly correlated, making it difficult for the model to distinguish their individual effects.

Feature Representations

Our exploratory analysis, as depicted in the Relationship Between Overall Rating and Other Features line chart provided in Part 1, reveals a strong positive correlation between overall ratings and the float features we have employed ('Seat Comfort', 'Cabin Staff Service', 'Food & Beverages', 'Ground Service', 'Inflight Entertainment', 'Wifi & Connectivity', 'Value For Money'), specifically 'Value For Money'. This correlation reinforces our belief that these float columns play a crucial role as important features in predicting Overall_Rating.

Upon analyzing the performance metrics detailed in the table, the inclusion of one-hot encoding features ('Type Of Traveller', 'Seat Type') notably enhanced the efficacy of the baseline Linear Regression model. Particularly noteworthy is the superior performance achieved by integrating 'reviews' transformed using the TF-IDF model.

The comparison across various models underscores a consistent trend: the amalgamation of both feature types — TF-IDF and One-Hot Encoding — within ensemble models, namely Random Forest and XGBoost, consistently outperforms other configurations.
This robust combination proves to be instrumental in achieving optimal predictive results. The inclusion of diverse features, encompassing both the TF-IDF-transformed 'reviews' and one-hot encoding,

stands out as a pivotal factor contributing to superior model performance.

Model Comparison

We also have Linear Regressor, Random Forest regressor, XGBoost regressor by themselves, and an ensemble of TF-IDF with Linear Regressor, an ensemble of TF-IDF with XGBoost regressor as our comparison.

Scaling/Overfitting

Scaling our model to the required size was executed without significant challenges. The computational efficiency of our TF-IDF and Random Forest Regressor pipeline contributed to seamless scalability. The ensemble nature of Random Forest mitigated overfitting concerns, and hyperparameter tuning, particularly focusing on "n_estimators," ensured optimal performance without sacrificing efficiency. The dataset's adequacy, coupled with regularization techniques, further prevented overfitting. In summary, the model's thoughtful design and optimization strategies successfully addressed potential issues related to scaling and overfitting.

Missing Values

In our dataset, we encountered missing values in several columns, including 'Seat Comfort', 'Cabin Staff Service', 'Food & Beverages', 'Ground Service', 'Inflight Entertainment', 'Wifi & Connectivity', and 'Value For Money'. To address these missing values, we employed a systematic imputation strategy.

Considering that different airlines may exhibit different characteristics and service quality, we utilized the median value of the respective airline in that specific column for the imputation of each missing value, with the aim of maintaining a degree of granularity and precision in the calculation of missing values.

In cases where all values for a given airline in a particular column were missing, we applied a more generalized imputation by replacing the missing values with a constant value of 0. This decision was made under the assumption that the absence of data might signify a lack of information rather than a genuine numerical value of zero. Imputing with zero allows us to preserve the structure of the dataset while acknowledging the absence of specific feedback for that airline in the respective service quality dimension.

Strength and Weaknesses of Different Models

For Linear Regression, the strength is it is easier to interpret and is computationally efficient. The weakness is that it assumes a linear relationship between features, which might not be the case sometimes, and it can only have limited complexity.

For Random Forest regression, the strength is that it can capture the nonlinear relationship between features and outputs, and they are not easily affected by outliers. The weakness is that it is less understandable, can cause overfitting to the data, and has a longer runtime.

For XGBoost regression, the strength is that it can have better performance and can work well if the data is non-linear. The weakness is that it is very complicated and may require careful tuning to avoid overfitting.

Model Optimization

Following the model selection phase, we employed the gp_minimize function for hyperparameter tuning to optimize the chosen model. The objective was to identify the parameter values that yield the lowest Mean Squared Error (MSE) for our model.

During this optimization process, the model pipeline, incorporating TF-IDF and Random Forest, underwent fine-tuning. Specifically, the hyperparameter "n_estimators" was targeted, ranging from 50 to 150. This range was chosen based on an initial observation that a relatively low MSE was achieved when "n_estimators" was set to 100. The gp_minimize function systematically explored parameter values and selected the configuration that minimized the MSE.

The outcome of the hyperparameter tuning indicated that the optimal setting for "n_estimators" was found to be 110. This configuration led to the lowest MSE value of 3.034, as illustrated in the accompanying results.

```
Current Hyperparameters: n_estimators = 130
MSE_rfW: 3.0371994791367456
----------------------
Current Hyperparameters: n_estimators = 68
MSE_rfW: 3.0527326121767957
----------------------
Current Hyperparameters: n_estimators = 128
MSE_rfW: 3.037293402251045
----------------------
Current Hyperparameters: n_estimators = 110
MSE_rfW: 3.0345684381324274
----------------------
Current Hyperparameters: n_estimators = 95
MSE_rfW: 3.035191852907691
----------------------
Current Hyperparameters: n_estimators = 60
MSE_rfW: 3.0532977346278316
----------------------
Current Hyperparameters: n_estimators = 96
MSE_rfW: 3.0361719077744818
----------------------
Current Hyperparameters: n_estimators = 83
MSE_rfW: 3.0449468165483706
----------------------
Current Hyperparameters: n_estimators = 64
MSE_rfW: 3.052466531401699
----------------------
Current Hyperparameters: n_estimators = 115
MSE_rfW: 3.034821875962258
----------------------
Best Hyperparameters: n_estimators = 110
Best MSE_rfW: -3.0345684381324274
```

PART 4: RELATED LITERATURE

Our project drew upon a dataset sourced from airlinequality.com, a resource initially organized and shared on Kaggle by GitHub users. This dataset encompasses diverse analyses, such as sentiment analysis and predictive tasks, with the ultimate goal of unraveling underlying patterns to enhance the overall airline customer experience.

Similar datasets in the past have often included features related to flight delays, cancellations, and detailed information on departure and arrival times. One noteworthy example involved the examination of accidents associated with an airline company, studying the effects of fatalities on public perception and the industrial rank of the airline. These datasets aimed to facilitate research practices and analyses, contributing to a deeper understanding of the aviation industry and better user experience.

During the literature exploration, we encountered an existing flight price prediction task and utilized the XGB Regressor model. This model explores the efficacy of dataset features such as source and destination cities, departure time, number of stops during a flight, arrival time, flight class, and duration. Inspired by these findings, our feature selection process focused on flight service-related features, including seat type and cabin staff service, to ensure our model captured relevant aspects of the airline customer experience.

Previously, our exploration involved sentiment analysis, particularly in the context of reviews associated with airline services. In line with existing literature, we attempted to assign sentiment scores to each review using the VADER sentiment analysis tool. Our analysis revealed that the overall sentiment of the dataset tended to be positive. While this correlation was observed with features like "Value for Money," we concluded that sentiment scores did not significantly contribute to the prediction of review ratings. Consequently, sentiment analysis was excluded from our final model.

As we delved into the literature, we encountered a variety of models and techniques for EDA and feature engineering, each contributing unique insights into effective ways to handle similar datasets. Some of the features highlighted in the literature significantly improved our model's performance. However, the applicability of certain methods, such as sentiment analysis, was contingent on the specific characteristics of the studied dataset. The literature served as a guiding framework, helping us navigate the complexities of feature selection, model building, and evaluation metrics.

The existing work helps us have a general understanding of the flights and what features might be important. Our model found that the service information was closely related to the review rating, which aligned with the external resources. However, the sentiment analysis does not necessarily work.

PART 5: CONCLUSION

In conclusion, our ensemble model combining TF-IDF and Random Forest Regressor outperformed alternative models, showcasing its superiority, particularly

concerning the 'Value For Money' feature, which significantly influenced overall model performance. The TF-IDF + Linear Regression model exhibited diminished effectiveness, highlighting the limitations of linear models in capturing the non-linear nature of certain features specific to this predictive task, consistent with the inherent weaknesses of linear regression.

Performance metrics comparisons across models revealed that the TF-IDF + Random Forest Regressor ensemble achieved the lowest Mean Squared Error (MSE) and other evaluation metrics, indicating superior predictive accuracy. The Random Forest Regressor, known for its proficiency in handling non-linearities and robustness against outliers, emerged as the top-performing model overall.

Through hyperparameter tuning, our model pipeline's performance was further optimized. The emphasis on fine-tuning the "n_estimators" hyperparameter within the Random Forest algorithm, representing the number of trees in the forest, yielded an optimal setting of 110. This optimization led to the lowest MSE value of 3.034, emphasizing the significance of increasing the number of trees in enhancing the model's ability to capture intricate relationships within the data, resulting in improved predictive accuracy.

Interpreting the parameters of our optimized model, the reduced MSE signifies more accurate predictions of Overall Ratings. The lower MSE indicates minimized squared differences between predicted and actual values, underscoring enhanced precision in the model's predictions.

The success of our proposed model can be attributed to the ensemble nature of Random Forest, which is adept at handling non-linear relationships and capturing intricate patterns in the data. The fine-tuning of hyperparameters, specifically "n_estimators," allowed us to strike a balance between model complexity and predictive performance.

In contrast, models like TF-IDF + Linear Regression faced challenges due to the linear regression's assumption of linearity, inadequately capturing the non-linear relationships present in the data. The ensemble of TF-IDF and Random Forest Regressor showcased superior performance by synergizing the strengths of both techniques, resulting in a more robust and accurate predictive model.

REFERENCE

1. Airline twitter sentiment - dataset by Crowdflower. (2016, November 21). data.world. Retrieved December 5, 2023, from https://data.world/crowdflower/airline-twitter-sentiment
2. Quankiquanki. (n.d.). Quankiquanki/Skytrax-reviews-dataset: An air travel dataset consisting of user reviews from Skytrax (www.airlinequality.com). GitHub. Retrieved December 5, 2023, from https://github.com/quankiquanki/skytrax-reviews-dataset
3. Python: Sentiment analysis using Vader. (2021, October 7). GeeksforGeeks. GeeksforGeeks. Retrieved December 5, 2023, from https://www.geeksforgeeks.org/python-sentiment-analysis-using-vader/
4. Airlines - dataset by NDSOUZA. (2016, October 15). data.world. Retrieved December 5, 2023, from https://data.world/ndsouza/airlines
5. Damodarabarbosa. (2023, June 7). Flight price prediction: Xgbregressor: 97.56%. Kaggle. Kaggle. Retrieved December 5, 2023, from https://www.kaggle.com/code/damodarabarbosa/flight-price-prediction-xgbregressor-97-56