berlin international UNIVERSITY OF APPLIED SCIENCES

Prediction of overall rating, sentiment analysis, and topic modeling of airline reviews

A Project Report for Topics in Machine Learning

Prof. Dr. Florian Artinger

Jaeden Capinig

2202987

Contents

Introduction

Customer reviews are now vital for a business's reputation and success. In the airline industry, travelers rely more on reviews from other travelers, rather than advertisements when choosing flights. Reviews highlight key aspects like seat comfort, cabin staff service, food quality, and entertainment, helping potential customers make informed decisions.

For airline companies, understanding customer satisfaction is critical in staying competitive. By analyzing reviews and flight experiences, companies can identify what drives satisfaction or dissatisfaction. Machine learning models made it possible to predict overall ratings, enabling data-driven improvements.

Sentiment analysis of reviews also offers deeper insights into travelers' emotions and opinions. By analyzing text, companies can uncover pain points and improve services. This report uses machine learning models and techniques to predict airline overall ratings and perform sentiment analysis, providing actionable insights for service improvement.

Dataset

Sourced from Kaggle.com, entitled "Airline Reviews Dataset" and is useful for discovering customer sentiments, analyzing trends over time and predicting ratings.

In data cleaning and preparation, I performed the following tasks:

1. Renamed columns for efficient coding
2. Changed variable type (i.e. `overall_rating` to numeric)
3. Replaced NaN values with 0
4. Converting dates to datetime

5. Repositioned variables within a table (i.e. by importance)
6. Merged `review_title` to `review`
7. Dropped unnecessary columns (i.e. `aircraft type, flight_route, flight_recommended, full_review, review_title`)

After data cleaning, it contains the following features:

Feature	Description	Type
airline_name	Name of airline being reviewed	Categorical
overall_rating	Overall rating given by the reviewer (e.g., 1-10)	Numerical
review_date	Date when the review was posted	Categorical
seat_comfort	Rating for seat comfort	Numerical
cabin_staff	Rating for cabin crew service	Numerical
food_beverages	Rating for food and beverage quality	Numerical
ground_service	Rating for ground service experience	Numerical
inflight_ent	Rating for inflight entertainment	Numerical
wifi_connect	Rating for in-flight WiFi connectivity	Numerical
value_money	Rating for value for money	Numerical
verified	Whether the review is verified (1 = Yes, 0 = No)	Categorical
review	Full text of the review written by the customer, merged the title	Categorical
traveller_type	Type of traveler (e.g., Business, Family, Solo)	Categorical
seat_type	Class of travel (e.g., Economy, Business, First)	Categorical
date_flown	Date of the flight being reviewed	Categorical

Libraries used are:

- os – Operating system interactions
- json – JSON data handling
- random – Random number generation
- glob – File path pattern matching
- typing – Type hints for Python
- pandas – Data manipulation and analysis
- numpy – Numerical computations
- seaborn – Data visualization
- matplotlib – Plotting and visualization

- string – String manipulation
- nltk – Natural Language Processing (NLP)
- datasets – Hugging Face datasets for NLP
- transformers – Hugging Face transformers for NLP tasks
- huggingface_hub – Hugging Face Hub integration
- langchain – Document processing and embeddings
- sklearn (scikit-learn) – Ensemble learning, linear models, decision trees, support vector regression, model selection and hyperparameter tuning, model evaluation, data preprocessing, handling missing values, data transformation pipeline, machine learning workflow, and text feature extraction
- torch (PyTorch) – Deep learning framework
- torchmetrics – Metrics for PyTorch models
- torchinfo – Model summary and visualization
- pytorch_lightning – Training and logging for PyTorch models
- torchvision – Image datasets and transformations
- ISLP – Machine learning with standard datasets
- openai – OpenAI API integration
- sentence_transformers – Sentence embeddings and similarity

Summary statistics

Descriptive statistics of `overall_rating` and other numerical variables

	overall_rating	**seat_comfort**	**cabin_staff**	**food_beverages**	**ground_service**	**inflight_ent**	**wifi_connect**	**value_money**
count	2000.0	2000.0	2000.0	2000.0	2000.0	2000.0	2000.0	2000.0
mean	2,700	2,413	2,551	1,868	22065,000	1507,000	0.64	2,317
std	2,668	16314,000	1,765	1,763	1,674	1,717	1,201	1,585
min	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
25%	1.0	1.0	1.0	0.0	1.0	0.0	0.0	1.0

50%	1.0	2.0	2.0	1.0	1.0	1.0	0.0	1.0
75%	3.0	4.0	4.0	3.0	4.0	3.0	1.0	4.0
max	9.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0

Table 1.1

Airlines to work with was chosen based on the top 20 most frequently reviewed.

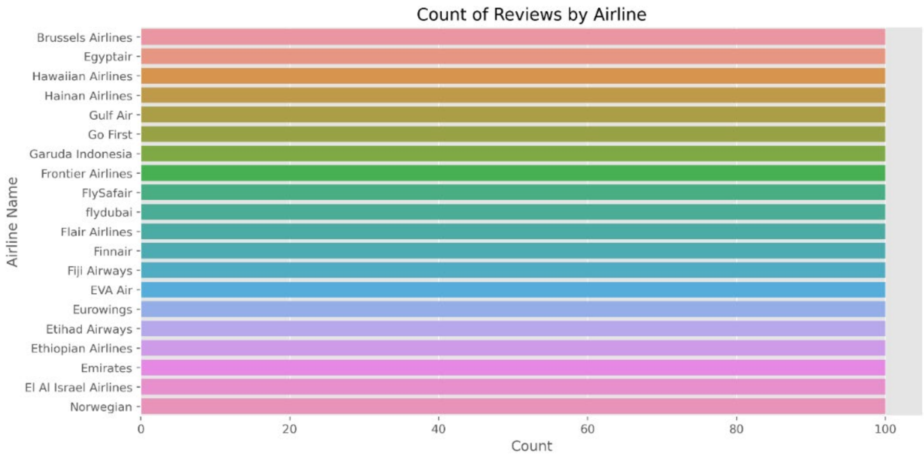

Figure 1.1

Below is the distribution of `overall_rating` and other key features.

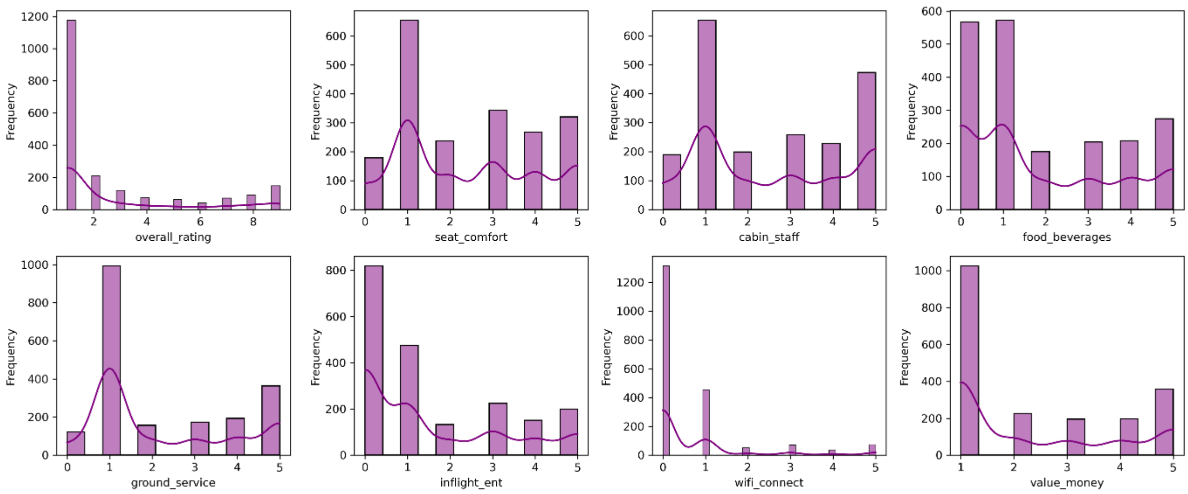

Figure 1.2

1. `overall_rating`, `seat_comfort`, `cabin_staff`, `food_beverages`, `ground_service` and `value_money` significantly peak at 1, and heavily skewed towards the lower end, suggesting that majority of the reviews has a low rating among these variables.

2. Smaller frequencies at higher ratings could be seen in `seat_comfort`, `cabin_staff`, `food_beverages`, `ground_service` and `value_money`.
3. The distributions indicate general dissatisfaction among customers.

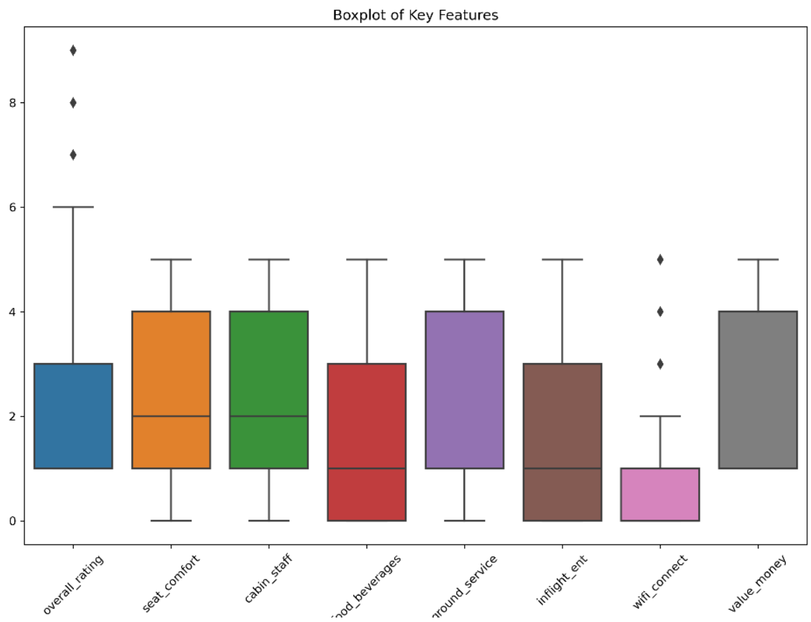

Figure 1.3

Median ratings sit around 2 or 3, suggesting low to moderate satisfaction on features. Wifi_connect having the lowest rating.

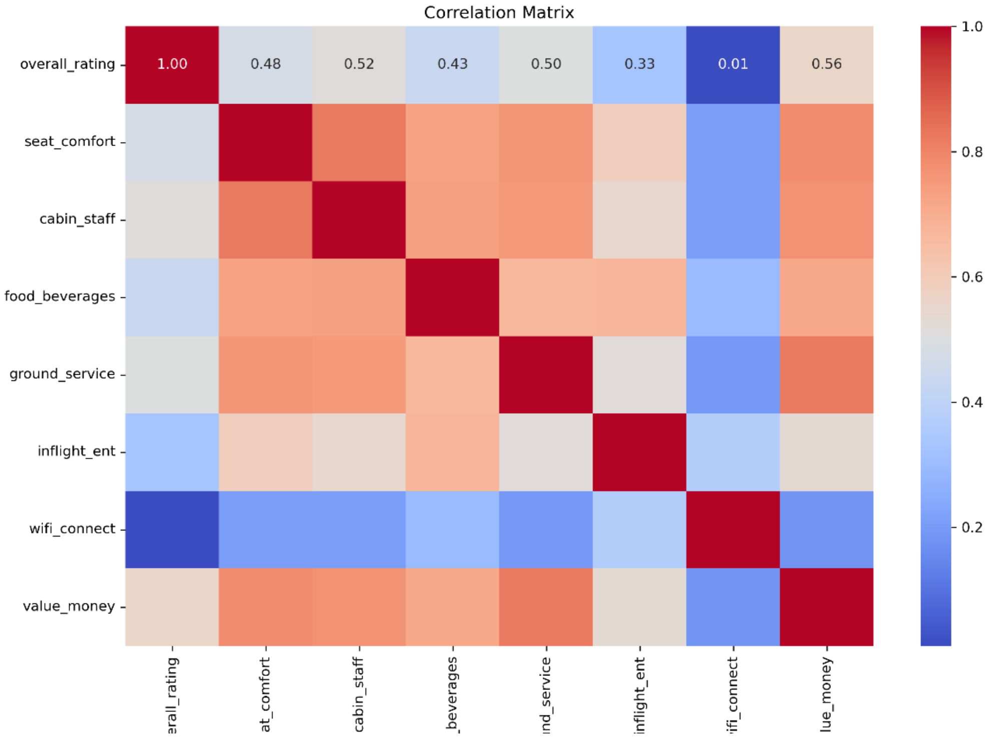

Figure 1.4

`Overall_rating` has a strong positive correlation with `seat_comfort` (0.48), `cabin_staff` (0.52), `food_beverages` (0.43), `ground_service` (0.50), and `value_money` (0.56), suggesting better experiences from these areas trigger overall satisfaction.

Below are bar charts for the counts of categorical features:

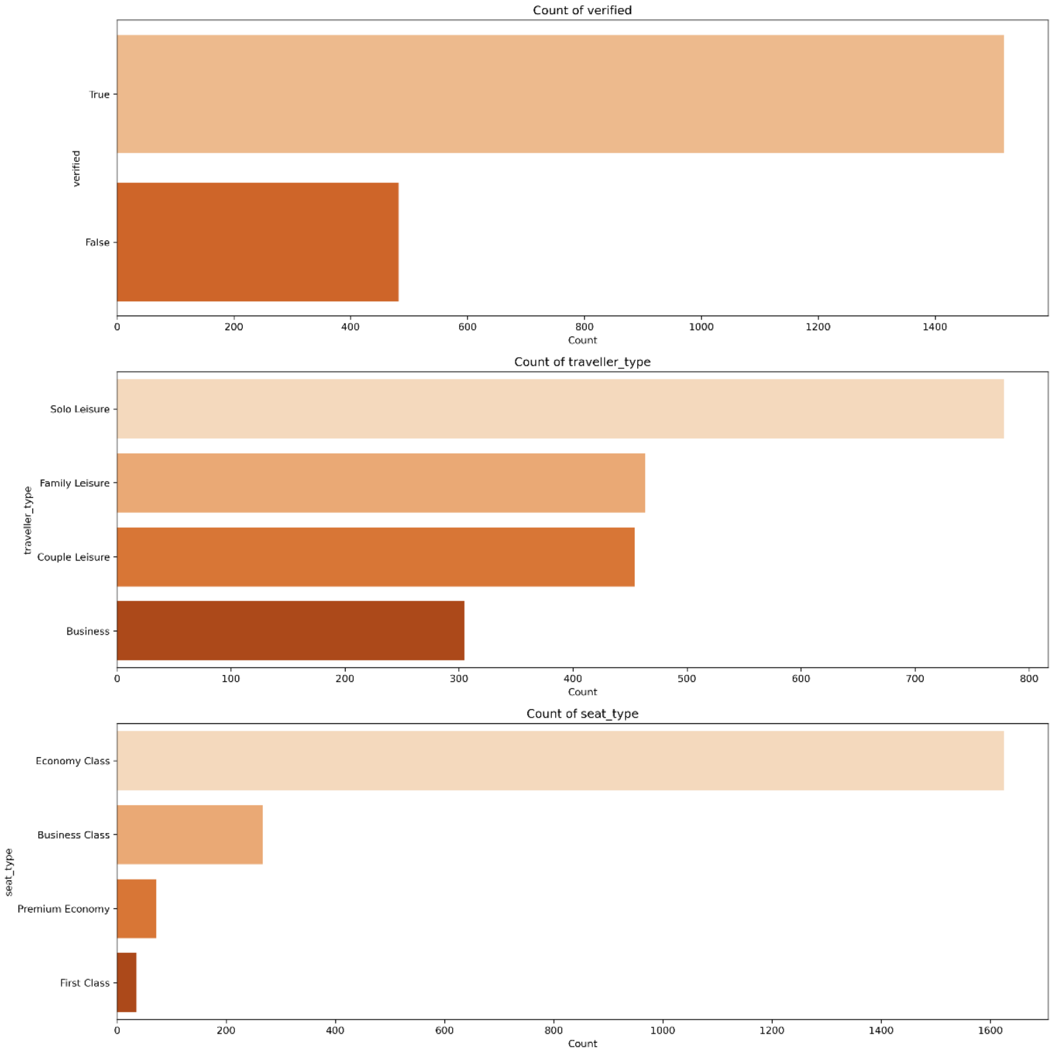

Figure 1.5

Quantitative analysis of the textual content of column review:

Airline	Average Chars per Word	Average Words per Sentence	Total Words	Vocabulary Size	Lexical Diversity
Norwegian	4	18	15200	2354	0.15
Finnair	4	20	17388	2640	0.15
Brussels Airlines	4	20	18034	2620	0.15
Go First	4	16	9478	1568	0.17
Garuda Indonesia	4	20	17622	2295	0.13
Frontier Airlines	4	18	16910	2305	0.14
FlySafair	4	20	12815	2102	0.16
flydubai	4	20	14822	2260	0.15
Flair Airlines	4	19	16979	2322	0.14
Fiji Airways	4	19	16012	2372	0.15
Hainan Airlines	4	17	11843	1914	0.16

EVA Air	4	17	17016	2714	0.16
Eurowings	4	18	15642	2567	0.16
Etihad Airways	4	20	18432	2683	0.15
Ethiopian Airlines	4	19	17245	2643	0.15
Emirates	4	19	17865	2743	0.15
El Al Israel Airlines	4	18	15122	2398	0.16
Egyptair	4	19	16406	2519	0.15
Gulf Air	4	19	14827	2361	0.16
Hawaiian Airlines	4	19	20050	2781	0.14

Table 1.2

Methods

For text preprocessing, we normalized (removed punctuation and stop words), performed tokenization, and applied vectorization.

In sentiment analysis, we used VADER (Valence Aware Dictionary and sentiment Reasoner), that is designed for informal text and a rule-based tool specifically attuned to social media. It could handle negations, intensity, and contextual polarity

To predict `overall_rating`, machine learning methods such as lasso, random forest, gradient boosting, XGBoost was used. These methods could handle complex relationships between features (e.g. review and sentiment_scores)

- Lasso – helps prevent overfitting, and only a subset of features is important for prediction
- Random Forest – ensemble method that builds multiple decision trees and combines their prediction. Robust to overfitting and can handle non-linear relationships
- Gradient boosting – builds trees sequentially, with each tree correcting the errors of the previous one. Highly effective for predictive accuracy. Good at handling imbalanced datasets
- XGBoost – optimized implementation of gradient boosting.

Evaluation metrics are MSE (Mean squared error) and adjusted r squared.

For topic modelling, we used RAG (Retrieval-Augmented Generation) and LDA (Latent Dirichlet Allocation). LDA automatically identifies topics in a corpus of text by clustering words that frequently occur together, while RAG combines retrieval-based and generation-based approaches for tasks like question answering and information retrieval. It incorporates external knowledge sources, hence the higher accuracy and more contextually relevant responses.

Analysis

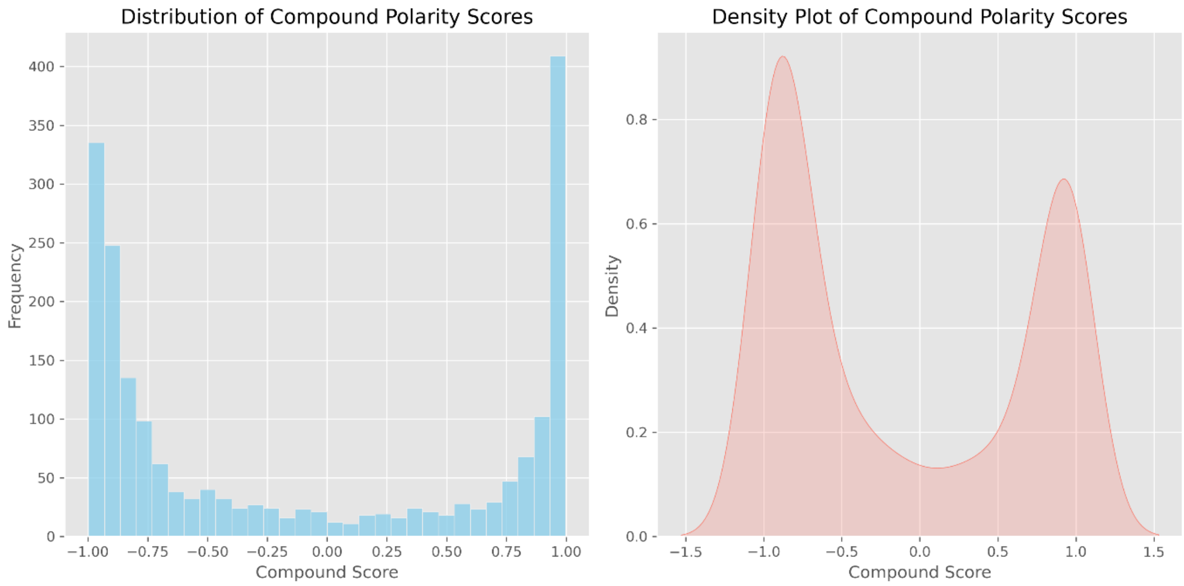

Figure 3.1

Bimodal distribution with peaks at the extreme ends of the compound score (-1 and 1) indicates that many reviews have strong negative or positive sentiments. Fewer reviews have neutral or mixed sentiments.

	MSE	**R^2**	**Adjusted R^2**
Lasso	0,1561	0	1,25
Random Forest	0,3992	0	1,25
Gradient Boosting	0.2812	0	1,25
XGBoost	0,7951	0	1,25

Table 3.1

For prediction, median scores of features are used

Lower MSE values suggest better model performance. In this regard, Lasso has the lowest, with an alpha of 0.4, followed by gradient boosting. However, having an R squared of 0 suggests none of the models explain any variance in the data. Optimally, the adjusted R squared should be within the range of 0 to 1. As the values are 1.25, there are potential issues within the data.

Based on feature importance analysis, the top five features that most significantly influence the prediction are::

1. `inflight_ent`
2. `seat_comfort`
3. `ground_service`
4. `positive`
5. `food_beverages`

In topic modelling through LDA, the following topics were inferred based on the identified words:

Topic #0: booking experience

Topic #1: business experience

Topic #2: flight delays

Topic #3: baggage, luggage

Topic #4: economy flight experience

Topic #5: boarding experience

In topic modelling through RAG, the following was extracted:

1. **Cabin Crew Service**
2. **In-Flight Entertainment**
3. **Food Quality**

4. **Check-In Experience**
5. **Baggage Handling**
6. **Flight Delays**
7. **Customer Service**
8. **Seat Comfort**
9. **Cleanliness**
10. **Boarding Process**

Conclusion

The analysis of airline reviews provided valuable insights into customer sentiment and highlighted key areas for improvement. The bimodal distribution of sentiment scores revealed that customers often express strong opinions, with most reviews being either highly positive or highly negative, and very few reflecting neutral or mixed sentiments. In the prediction analysis, Lasso regression outperformed other models, as indicated by its lower Mean Squared Error (MSE). However, the R^2 and Adjusted R^2 values were concerning, suggesting that the models did not explain the variance in the data effectively. This points to potential issues with the dataset or feature engineering, which may require further investigation and refinement to improve model performance.

The feature importance analysis identified the top factors influencing customer satisfaction, such as in-flight entertainment, seat comfort, ground service, positive sentiment, and food and beverages. These findings align closely with the results of topic modeling. Using Latent Dirichlet Allocation (LDA), key topics like booking experience, flight delays, and baggage handling were identified. Additionally, Retrieval-Augmented Generation (RAG) extracted further themes, including cabin crew service, check-in experience, and cleanliness. Together, these insights provide a comprehensive understanding of what matters most to passengers.

By leveraging computational tools like sentiment analysis, predictive modeling, and topic modeling, airlines can gain actionable insights on a larger scale. These insights can inform strategies to enhance customer satisfaction, address pain points, and improve overall service delivery. For instance, focusing on improving in-flight entertainment, seat comfort, and ground service could significantly boost passenger satisfaction.

In conclusion, while the predictive models require further refinement, the analysis offers valuable insights into customer preferences and areas for improvement. By prioritizing the key features and topics identified, airlines can take targeted actions to enhance the passenger experience, foster loyalty, and maintain competitiveness in the industry.

References

Bhojani, J. (n.d.). *Airline Reviews*. Www.kaggle.com. https://www.kaggle.com/datasets/juhibhojani/airline-reviews

Ignition GTM, Inc. (2024). *KPIs for Marketing: online reviews*. Haveignition.com. https://www.haveignition.com/kpis-for-marketing/kpis-for-marketing-online-reviews