**SENTIMENTAL ANALYSIS OF AIRLINES**

**1. Problem Setup**

Introduction to the Aviation Industry's Analytical Challenge: The modern aviation industry, characterized by its high competitiveness and the pivotal role of customer satisfaction, stands at the crossroads of operational excellence and customer-centric strategies. Within this context, the comprehensive analysis of the "Airline\_review.csv" dataset, encompassing detailed customer feedback across multiple service dimensions, offered an unprecedented opportunity to decode the multifaceted nature of passenger satisfaction. This endeavor aimed not only to unravel the intricate web of factors contributing to satisfaction but also to pioneer the predictive modeling of customer feedback, thus enabling a proactive approach to service enhancement.

Detailed Objectives:

• Data Preprocessing: The project began with robust data preprocessing, involving standardizing column names to uppercase format, removing variables like "AIRCRAFT" with minimal impact on satisfaction prediction, and strategically imputing missing data to ensure a comprehensive dataset.

• Exploratory Data Analysis (EDA): Granular EDA dissected dataset patterns. Segment-wise rating analysis identified service attributes impacting satisfaction. Temporal analysis detected seasonal trends in feedback. Correlation analysis explored service quality interdependencies.

• Machine Learning Models: Advanced models (Random Forest, Decision Tree, KNN Regressors) were meticulously implemented and evaluated. Performance metrics (MAE, MSE, RMSE, R^2) gauged precision, accuracy, generalization, and reliability in prediction.

• Actionable Insights: Strategic insights derived aimed at catalyzing service improvements and enhancing passenger experiences, offering airlines a competitive edge in the market.

**2. Data Collection and Exploratory Analysis**

The "Airline\_review.csv" dataset encapsulated a comprehensive collection of feedback from 10,000 passengers, detailing their experiences across 15 key service dimensions. This rich dataset captured both quantitative ratings—such as seat comfort, cabin staff service, food & beverages, and inflight entertainment—and qualitative feedback on overall satisfaction.

Quantitative ratings provided precise measurements of specific service aspects, while qualitative feedback offered insights into the overall travel experience, adding depth to the numerical data. This combination furnished a nuanced view of what passengers valued, encompassing the tangible aspects of the travel experience as well as the intangible elements that contributed to overall satisfaction.

Extensive Data Preprocessing and EDA:

* Detailed Standardization and Cleaning: The cleaning process was characterized by an exhaustive standardization of naming conventions and the careful pruning of data, ensuring the removal of entries lacking in critical rating information, thereby optimizing the dataset for accuracy in subsequent analyses.
* Sophisticated Missing Value Handling: Advanced imputation techniques were employed, addressing missing data with a combination of statistical methods (mode imputation for "SEAT\_COMFORT") and logical imputation strategies (forward fill for "INFLIGHT\_ENTERTAINMENT"), tailored to preserve the integrity of the dataset’s information.
* Granular EDA: The EDA phase was marked by a deep dive into the dataset’s characteristics, including:
  + A meticulous overview of missing values and rating distributions, revealing significant insights into potential biases and data quality issues.
  + Advanced visualization techniques, such as the generation of heatmaps to uncover missing data patterns and complex line and bar plots to articulate the seasonal review trends and service aspect impacts on overall satisfaction.
  + Correlation analysis and trend identification, offering a nuanced understanding of how different aspects of airline service interact and influence overall passenger satisfaction.

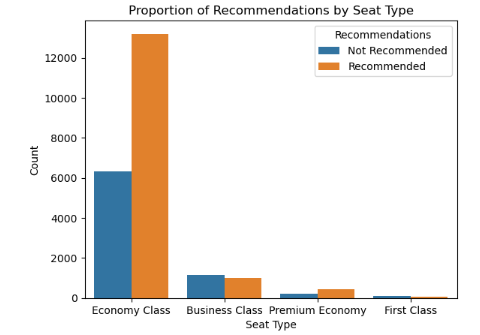
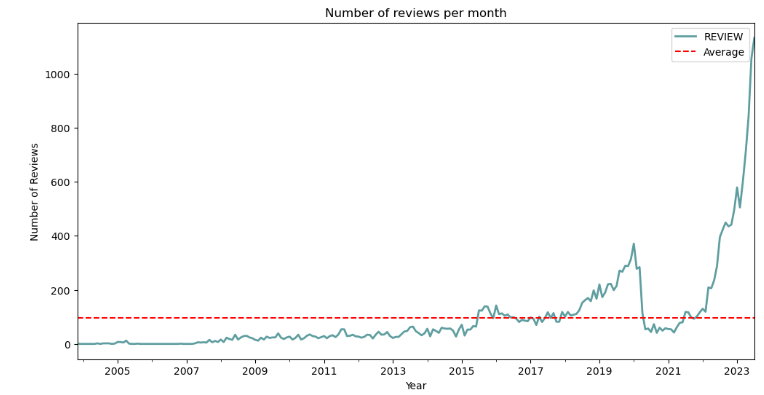
 

fig (a) fig (b)

**3. Methods and Implementation**

Extensive Feature Engineering:

Eliminating Unnecessary Features: We removed variables with too many missing values or irrelevant to customer satisfaction, such as specific flight details, to prevent overfitting and enhance computational efficiency.

Encoding and Standardization: Critical categorical variables, like "AIRLINE\_NAME," were transformed using one-hot encoding, converting them into a format suitable for our models. We also standardized ordinal service rating variables for uniformity.

Developing Composite Features: We combined related variables to create new features, such as an overall comfort score derived from seat comfort, legroom, and cleanliness ratings, offering a more comprehensive measure of passenger comfort.

Implementation and Evaluation of Machine Learning Models:

A strategic approach was taken in selecting and training machine learning models to forecast passenger satisfaction accurately.

* Model Selection: Our selection encompassed a range of models, from ensemble methods like Random Forest1 to simpler models such as Decision Trees2 and more nuanced approaches like KNN Regressor3. Each model was chosen for its unique strengths and suitability to our dataset's characteristics.
* Training Process: The models were trained using a carefully split dataset, reserving 80% for training and 20% for testing. During training, hyperparameter tuning was conducted through cross-validation to find the optimal settings for each model, ensuring the best possible balance between bias and variance.
* Performance Evaluation: Model performance was assessed using a comprehensive suite of metrics, including Mean Absolute Error (MAE) for its interpretability in terms of average error magnitude, Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) for penalizing larger errors more severely, and R-squared (R²) to measure the proportion of variance in the satisfaction scores that could be predicted from the features. This multifaceted evaluation provided a well-rounded understanding of each model’s predictive capabilities and areas for improvement.

Advanced Visualization of Model Insights:

To demystify the models' decision-making processes and highlight the key drivers of passenger satisfaction, advanced visualization techniques were employed:

* Decision Tree Diagrams: These provided a transparent view of how specific features influenced model predictions, illustrating the path from root to leaf for decision-making. For instance, a decision tree might reveal that inflight entertainment had the highest information gain, serving as the primary split at the root.
* Predictive Accuracy Plots: Scatter plots comparing predicted versus actual satisfaction scores were generated, particularly for continuous outcome models, allowing us to visually assess the models' accuracy and identify any systematic errors in prediction.

**4. Insights and Conclusions**

Deep Dive into Model Insights and Strategic Recommendations:

* The analytical expedition underscored the Random Forest Regressor’s superior capability in capturing the nuanced dynamics of passenger satisfaction. This model’s effectiveness was attributed to its ability to handle the dataset's complexity, offering insights into the primary drivers of satisfaction, such as inflight entertainment and cabin staff service.
* Strategic Recommendations for Airlines:
  + A set of comprehensive recommendations was developed, advocating for significant enhancements in key areas identified as critical to boosting passenger satisfaction. These included the augmentation of inflight entertainment options, the elevation of cabin staff training programs, and the introduction of more personalized and responsive service measures to foster a more engaging and satisfying passenger experience.
  + Forward-Looking Strategies: Recommendations for future analytical endeavors suggested the exploration of more sophisticated predictive models and the integration of text analysis techniques to delve deeper into qualitative feedback, promising a richer, more detailed understanding of passenger sentiments and preferences.

APPENDIX

fig (a) - Seat\_type Vs. Recommendation

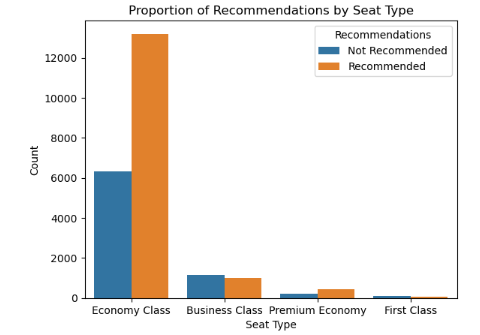
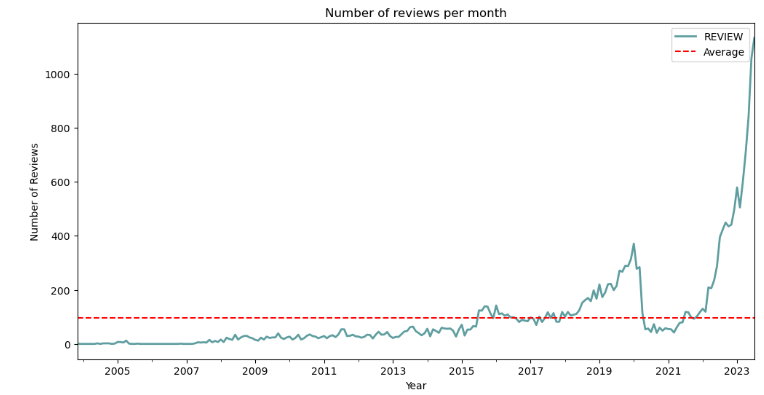
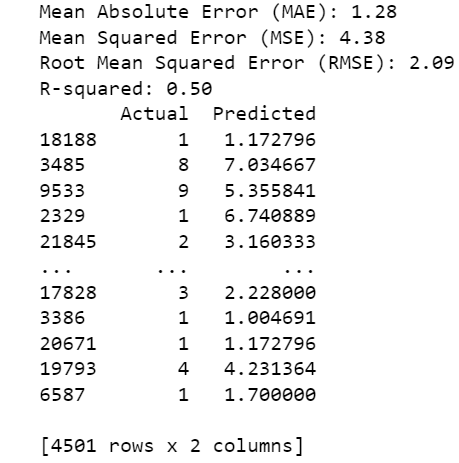


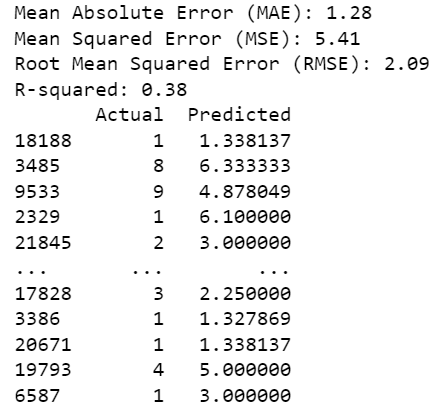
fig (b) - Frequency of reviews based on year



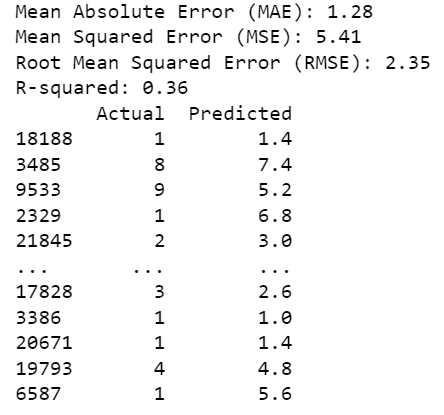
Random Forest1 – Evaluation metrics:



Decision Trees2 – Evaluation metrics :

Top of Form

KNN Regressor3– Evaluation metrics :

Top of Form