Prediction of Customer Satisfaction and Evaluating the Significance of Services and Performance

Berkay Dinç
Computer Engineering Department
Dokuz Eylül University
İzmir, Türkiye
berkay.dinc@ogr.deu.edu.tr

Güney Söğüt
Computer Engineering Department
Dokuz Eylül University
İzmir, Türkiye
guney.sogut@ogr.deu.edu.tr

Yusuf Gassaloğlu
Computer Engineering Department
Dokuz Eylül University
İzmir, Türkiye
yusuf.gassaloglu@ogr.deu.edu.tr

***Abstract*—This study explores the application of machine learning models to predict and evaluate customer satisfaction in the airline industry. Using a publicly available dataset and a custom-built survey application, the project examines key factors influencing passenger satisfaction, including service quality, flight punctuality, and customer support. Three machine learning models—Random Forest, Logistic Regression, and K-Nearest Neighbors—were tested under varying preprocessing scenarios, with Random Forest emerging as the most reliable and accurate. Rigorous evaluation metrics, including precision, recall, F1 score, and AUC, confirmed the model's robustness. Additionally, the incorporation of Importance-Performance Analysis (IPA) provided actionable insights for identifying critical areas for improvement. This study highlights the value of machine learning in enhancing service quality and meeting evolving customer expectations within the competitive airline industry.**

I. Introduction

The aviation industry is one of the most competitive sectors globally, characterized by dynamic customer expectations and the need for continuous service improvements. In this context, understanding and predicting customer satisfaction is not only a measure of operational success but also a critical driver of business growth and customer retention. With increasing access to data and advancements in machine learning, it has become feasible to develop predictive models that enable airlines to anticipate passenger satisfaction and take proactive measures to enhance the overall travel experience.

This study focuses on evaluating customer satisfaction in the airline industry by leveraging a publicly available dataset from Kaggle [1], which includes a wide range of features related to passenger demographics, service quality, and operational metrics. These features reflect various aspects of the airline experience, such as flight punctuality, in-flight services, customer support, and overall satisfaction. Understanding the significance of these features and their relationships with customer satisfaction is vital for improving service delivery.

To achieve these objectives, we implemented and compared three widely used machine learning models—K-Nearest Neighbors (KNN), Logistic Regression, and Random Forest. These models were chosen for their diverse strengths in handling classification and regression tasks, allowing for a comprehensive evaluation of their effectiveness in predicting satisfaction levels. The performance of each model was assessed using an extensive set of evaluation metrics, including Accuracy, Precision, Recall, F1-Score, Area Under the Curve (AUC), and Cohen's Kappa. Furthermore, confusion matrices and Receiver Operating Characteristic (ROC) curves were plotted using Matplotlib to provide a visual understanding of the models' predictive capabilities.

To ensure a robust analysis, the modeling process was conducted under three distinct scenarios:

With Feature Selection: To identify and use the most influential features, reducing noise and enhancing model efficiency. With Normalization: To standardize the dataset and ensure comparability between features with different scales. Without Preprocessing: To evaluate baseline performance and the impact of preprocessing techniques. In addition to analyzing historical data, this study incorporates real-time feedback through a custom-built survey application developed using Flask. This application was designed to collect data directly from passengers, focusing on their experiences and satisfaction with various aspects of their journey. By integrating survey responses with the existing dataset, the study provides a richer and more comprehensive perspective on customer satisfaction.

The research methodology was implemented using Python, with scikit-learn employed for machine learning and Matplotlib for data visualization. These tools facilitated efficient model development, evaluation, and interpretation. The integration of robust analytics with practical data collection highlights the study's dual focus on leveraging technology and understanding user experiences.

By combining advanced machine learning techniques, rigorous performance evaluation, and real-time data collection, this paper seeks to provide actionable insights for airline companies. The findings aim to support data-driven decision-making in areas such as service design, resource allocation, and customer relationship management. Ultimately, the study contributes to the broader goal of enhancing passenger satisfaction, fostering customer loyalty, and maintaining a competitive edge in the fast-paced airline industry.

II. Related Works

Yang [2] proposed the Importance-Satisfaction (I-S) model to assist firms in identifying critical quality attributes for improvement based on customer satisfaction surveys. Using a two-dimensional approach, the I-S model classifies attributes into four areas: excellent, to be improved, surplus, and careless. For optimal decision-making, Yang developed a satisfaction function, which helps evaluate the marginal satisfaction of each attribute. The study emphasized the need for firms to prioritize attributes with high importance but low satisfaction, located in the "to be improved" area. Through the I-S model and marginal satisfaction analysis, firms can achieve a balance in resource allocation and improve their customer satisfaction index (CSI).

Park et al. [3] proposed a deep learning methodology to analyze airline customer propensities, focusing on customer churn risk and satisfaction in South Korea. The study surveyed 340 adults who had flown at least once in the last five years, collecting data through 50 questions about physical and social servicescapes, brand experience, loyalty, and satisfaction. After preprocessing the data, machine learning models such as kNN, Decision Tree, RF, and XGBoost, along with deep learning models like CNN and CNN-LSTM, were evaluated. The CNN-LSTM model achieved the highest accuracy, with 94% for customer churn risk and 90% for satisfaction, outperforming traditional machine learning models by 11% and 7% on average. Incorporating social servicescape factors increased the accuracy by up to 6% for deep learning models and 5% for machine learning models

AlHabbal [4] conducted a study using machine learning to predict and optimize U.S. airline customer satisfaction. The research employed a dataset of 129,880 records containing 23 attributes, including customer demographics, trip details, and service evaluations. Seven classification models—Decision Tree (DTC), Random Forest (RF), Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naive Bayes (NB), and Artificial Neural Networks (ANN) —were evaluated. Random Forest demonstrated the highest accuracy of 93.32%, while ANN achieved 92.13%. The dataset's imbalance ratio between satisfied (53.59%) and dissatisfied (46.41%) customers was addressed through stratified sampling. Feature importance analysis identified 'Inflight Wi-Fi Service,' 'Ease of Online Booking,' and 'Leg Room Service' as the top predictors of satisfaction. The study concluded with a confusion matrix and radar chart to compare model performance across metrics such as precision, recall, and F1-score

Zhibin et al. [5] proposed the Importance-Performance-Impact Analysis (IPIA) framework to enhance customer satisfaction by addressing the shortcomings of the conventional Importance-Performance Analysis (IPA). IPIA incorporates advanced tools, including Back Propagation Neural Network (BPNN) and DEMATEL/Analytic Network Process (ANP). Empirical application in one of China's 'Big Four' airlines involved a web-based survey of 817 respondents, from which 298 valid responses were analyzed. The study identified ten key attributes grouped into basic, performance, and excitement factors. The IPIA Matrix revealed that punctuality (Importance: 0.16, Performance: 3.49, Impact: 0.26) and ticket price (Importance: 0.15, Performance: 3.28, Impact: 0.20) required resource prioritization. Airline reputation scored highest in all dimensions (Importance: 0.18, Performance: 3.83, Impact: 0.36). Results demonstrated IPIA's superior capability to guide resource allocation through visual tools like the IPIA Table and Matrix.

Cavalcante et al. [6] developed a methodology to predict customer satisfaction for electricity distribution companies using machine learning. The study analyzed four primary data sources: customer service, power outage, reliability indices, and satisfaction surveys, comprising over 12 million daily service records and 2.8 million power outage instances. The developed system used regression models trained on historical data from 20 operational areas, with 580 individual regressors generated to estimate satisfaction indicators. The validation process compared predictions against a 2017 satisfaction survey, achieving a Mean Absolute Percentage Error (MAPE) of 1.36% for the primary indicator, ISPQ. Processing time for the complete training exceeded 100 hours, involving 200,448 configurations per training session.

Aileen et al. [7] conducted a study to predict U.S. airline passenger satisfaction using machine learning, employing a dataset of 129,880 observations with 22 features. Six supervised classification models were implemented: Random Forest (RF), K-Nearest Neighbors (KNN), Decision Tree Classifier (DTC), Logistic Regression (LR), Naïve Bayes (NB), and AdaBoost. Feature selection reduced the variables to five key predictors: 'Customer Type,' 'Departure/Arrival time convenient,' 'Online boarding,' 'Inflight entertainment,' and 'Cleanliness.' The Random Forest model achieved the highest accuracy (89.2%), followed by KNN (87.2%), DTC (82.0%), AdaBoost (82.8%), LR (78.4%), and NB (76.8%). Performance evaluation metrics revealed RF as the best model with a precision of 93.04% and an F1-score of 88.80%, predicting 223 true values and 27 false values in the confusion matrix. Features such as 'Online boarding' and 'Inflight wifi service' showed the highest correlation with passenger satisfaction.

III. Design and Implementation

The project was designed to harness the power of data warehousing and machine learning techniques to analyze customer feedback and provide actionable insights for improving service quality. The system integrates predictive modeling on Fig. 1, Importance-Performance Analysis (IPA), and a user-friendly web interface to streamline the process of collecting, analyzing, and interpreting customer satisfaction data. By employing advanced machine learning models and robust evaluation metrics, the project offers a comprehensive approach to understanding and addressing the needs and expectations of customers.

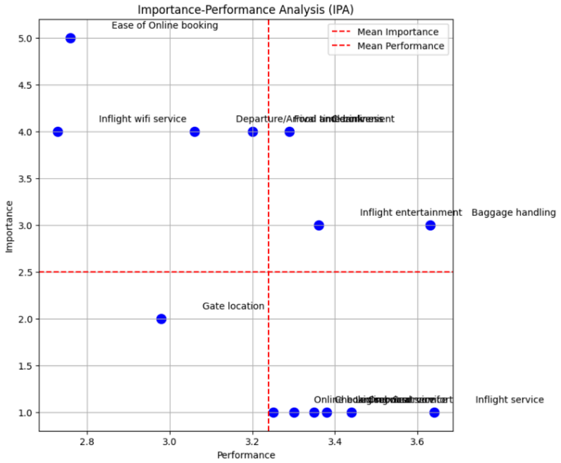

Fig. 1: IPA Example

The project design emphasizes a seamless flow of data, starting with a Flask-based web interface that enables customers to provide feedback through an intuitive survey interface hosted on the /survey route. When customers submit their feedback, the data is dynamically stored in a CSV file (test.csv), forming a centralized data warehouse for further analysis. This approach ensures real-time data capture and maintains a structured format that facilitates efficient preprocessing and model training. The survey interface is designed to maximize ease of use, encouraging widespread participation and generating a rich dataset for predictive modeling.

For the predictive analysis, three machine learning algorithms were employed: Random Forest, Logistic Regression, and K-Nearest Neighbors (KNN). These models were implemented using the scikit-learn library within a Jupyter Notebook environment. The dataset underwent rigorous preprocessing, including handling missing values, normalizing numerical features, and encoding categorical variables to ensure compatibility with the algorithms. Model evaluation was conducted using cross-validation to validate the performance on unseen data, with key metrics including accuracy, precision, recall, F1-score, Area Under the Curve (AUC), and Cohen's Kappa. These metrics provide a comprehensive view of each model's predictive capabilities, addressing both their ability to classify correctly and their agreement with true labels. Among the models, Random Forest demonstrated the highest performance across all metrics, making it the ideal candidate for deployment.

The selected Random Forest model was deployed within the Flask application by integrating it into the app.py file. This enabled real-time predictions on new survey data submitted by users. Administrators access these results through the /results route, which provides detailed insights, including predictions, evaluation metrics, and visualizations. Visual aids such as confusion matrices, ROC curves, and IPA charts help administrators understand the performance of the model and prioritize key service attributes. IPA specifically enables a clear distinction between attributes that are performing well and those requiring attention, providing actionable insights for strategic decision-making.

In terms of evaluation, the metrics used go beyond basic accuracy to include precision, recall, and F1-score, which reflect the balance between correctly identified positive cases and overall prediction performance. The AUC provides insight into the model's ability to distinguish between classes, while Cohen's Kappa evaluates the agreement between predicted and actual labels, accounting for the possibility of chance agreement. These metrics collectively ensure a robust assessment of the models' capabilities, addressing both technical and business objectives.

In summary, the project offers a comprehensive framework for analyzing customer satisfaction and service performance through predictive modeling and visualization. The integration of machine learning models like Random Forest, Logistic Regression, and KNN, combined with robust metrics and IPA, ensures that the system not only provides accurate predictions but also meaningful insights into the significance of different service attributes. By combining technical rigor with practical usability, the project empowers organizations to enhance customer satisfaction and strategically allocate resources to areas with the highest impact.

IV. Benchmarks and Tests

The testing phase for the project involved rigorous evaluation of three machine learning models—Random Forest, Logistic Regression, and K-Nearest Neighbors (KNN)—under different preprocessing conditions. These conditions included feature selection, normalization, and a scenario without any preprocessing. The models were assessed using several performance metrics: accuracy, precision, recall, F1-score, Area Under the Curve (AUC), and Cohen's Kappa. The goal was to determine the optimal model and preprocessing combination that would provide the most reliable results for customer satisfaction prediction.

Random Forest consistently demonstrated the highest level of performance across all scenarios, making it the standout model in this study. In the feature selection scenario, Random Forest achieved an impressive accuracy of 93%, with a precision of 93% and recall of 91%. The F1-Score, which balances precision and recall, was 92%, indicating a strong ability to correctly classify both positive and negative cases. Additionally, the model exhibited an AUC of 0.98, reflecting its excellent discrimination power between classes. Cohen's Kappa, a metric that accounts for agreement beyond random chance, reached 0.86, indicating substantial agreement between the model's predictions and the true labels.

When normalization was applied, Random Forest further improved its performance, achieving an accuracy of 96% and a precision of 97%. These results highlighted the model's capacity to perform well even when data normalization techniques were applied, improving its robustness against variations in input data. The recall was slightly lower at 94%, but the model's precision remained high, ensuring that it effectively minimized false positives. The F1-Score of 96% reinforced the model's ability to balance precision and recall under normalized conditions. AUC remained at 0.99, signifying near-perfect classification ability, and Cohen's Kappa increased to 0.92, demonstrating an even stronger agreement between predictions and actual values.

In the scenario without any preprocessing, Random Forest achieved near-identical results, maintaining an accuracy of 96%, precision of 97%, recall of 94%, and F1-Score of 96%. The AUC reached an exceptional 0.99, further solidifying the model's performance. Cohen's Kappa also remained high at 0.92, indicating that the model's predictions were highly reliable. In addition to the numerical performance metrics, the confusion matrices and ROC curves for Random Forest provided valuable insights into its predictive capabilities. The confusion matrix demonstrated a clear distinction between correctly and incorrectly classified instances, with a minimal number of false positives and false negatives on Fig. 2. The ROC curves also showcased the model's ability to discriminate between classes, with the curve staying close to the top-left corner, reflecting high sensitivity and specificity on Fig. 3.

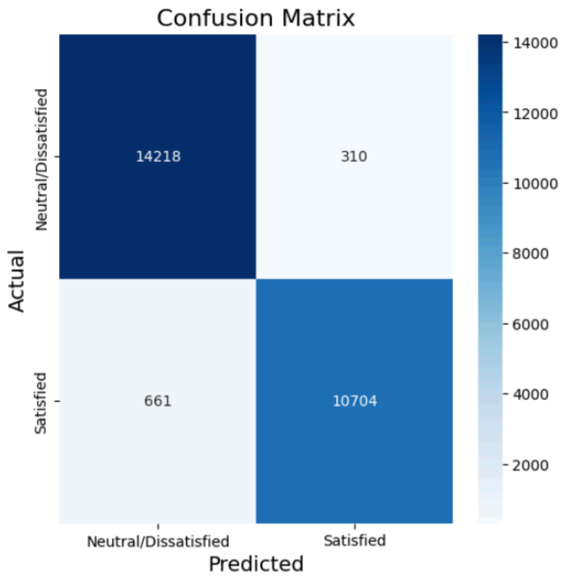

Fig. 2: Random Forest Confusion Matrix with normalization

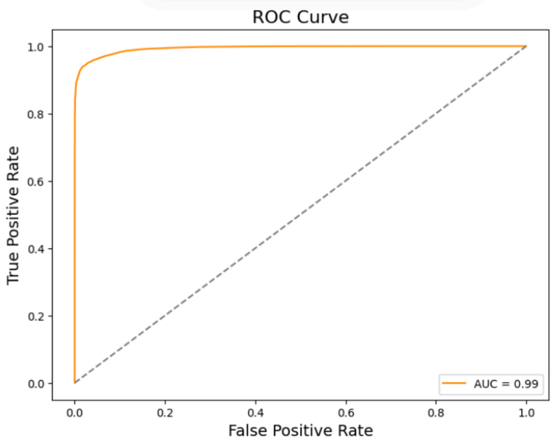

Fig. 3: Random Forest ROC Curve with normalization

Logistic Regression, while not as strong as Random Forest, still delivered reasonable performance across the different preprocessing scenarios. In the feature selection scenario, the model achieved an accuracy of 85%, with a precision of 84% and a recall of 82%. The F1-Score was 83%, which suggested that the model was moderately successful at identifying both positive and negative cases, but it was not as effective as Random Forest. The AUC was 0.91, indicating that Logistic Regression could distinguish between classes, although with lower confidence than Random Forest. Cohen's Kappa was 0.70, signifying moderate agreement between the predicted and actual labels, but it highlighted that the model still had room for improvement.

Normalization improved the performance of Logistic Regression, with accuracy rising to 87%, precision to 87%, and recall to 83%. The F1-Score also increased to 85%, and the AUC reached 0.93, showing a modest improvement in the model's ability to separate classes. Cohen's Kappa also improved to 0.74, suggesting better agreement with the actual labels. However, even with these improvements, Logistic Regression still lagged behind Random Forest in terms of overall predictive power and reliability.

In the scenario without preprocessing, Logistic Regression performed similarly to the normalized case, with accuracy, precision, recall, F1-Score, AUC, and Cohen's Kappa remaining identical to those achieved with normalization. This consistency suggested that the model's performance was not heavily influenced by preprocessing techniques, but it also indicated that Logistic Regression might not benefit as much from preprocessing as Random Forest.

KNN demonstrated varying results across the different preprocessing conditions, with its performance being highly sensitive to the preprocessing applied. In the feature selection scenario, KNN achieved a high accuracy of 93%, precision of 94%, and recall of 90%. The F1-Score was 92%, and the AUC was 0.97, indicating that KNN could effectively classify instances and was able to discriminate between the different classes with good accuracy. Cohen's Kappa was 0.86, suggesting that the model's predictions had a high level of agreement with the true labels. These results were competitive with Random Forest, indicating that KNN could be a viable model under the right conditions.

However, when normalization was applied, KNN's performance showed only a slight improvement, with accuracy and precision both remaining at 93%, and recall dropping slightly to 89%. The F1-Score stayed at 92%, and the AUC remained at 0.97, showing that the model still performed well but did not exhibit the same level of improvement seen in Random Forest. Cohen's Kappa decreased slightly to 0.85, indicating a marginal reduction in agreement.

The most significant decline in performance occurred when no preprocessing was applied. In this scenario, KNN's accuracy dropped dramatically to 75%, with precision falling to 73% and recall to 67%. The F1-Score also decreased to 70%, and the AUC dropped to 0.80, signaling poor classification performance. Cohen's Kappa also dropped to 0.48, reflecting weak agreement and indicating that the model's predictions were unreliable when raw, unprocessed data was used.

Random Forest emerged as the best-performing model across all preprocessing scenarios, especially in terms of confusion matrices and ROC curves, which clearly indicated its superior classification ability. The model's robust performance in all metrics—accuracy, precision, recall, F1-Score, AUC, and Cohen's Kappa—was evident, with AUC consistently approaching 1.0, demonstrating near-perfect class separation. The confusion matrices for Random Forest revealed a low incidence of misclassifications, and the ROC curves remained close to the top-left corner, indicating excellent sensitivity and specificity.

In contrast, Logistic Regression and KNN, while delivering solid results, were not as reliable or consistent as Random Forest. Logistic Regression showed moderate improvements with normalization but remained weaker in comparison to Random Forest, particularly in the AUC and Cohen's Kappa metrics. KNN demonstrated a good performance in feature selection but suffered a significant decline when preprocessing was omitted, underscoring its sensitivity to data preparation techniques.

In conclusion, Random Forest outperformed both Logistic Regression and KNN, making it the most suitable model for this application. Its ability to provide reliable and accurate predictions, coupled with clear and informative confusion matrices and ROC curves, makes it the optimal choice for predicting customer satisfaction and evaluating service performance in the context of this project.

V. Future Work

The findings and methodologies in this project provide a foundation for future research and development in customer satisfaction prediction within the airline industry. Expanding the scope of the study to include additional datasets, such as customer feedback from social media platforms, loyalty programs, and competitor benchmarking, could enhance the model's ability to capture a broader range of customer sentiment. Furthermore, advanced feature engineering techniques, such as natural language processing (NLP), could analyze qualitative data from open-ended survey responses, providing deeper insights into customer experiences. The integration of real-time analytics and streaming data capabilities would enable dynamic responses to emerging customer satisfaction trends. Exploring the performance of cutting-edge machine learning approaches, such as transformers and deep reinforcement learning, could further refine predictive accuracy and robustness. Moreover, addressing scalability concerns by testing the system with larger, multi-airline datasets and developing interactive visualization dashboards would enhance its practical applicability. Finally, tackling ethical considerations, such as bias in data collection and predictions, is crucial for fostering fairness and transparency in machine learning applications for customer satisfaction analysis.

VI. Conclusion

This project demonstrated the application of machine learning techniques in predicting and evaluating customer satisfaction in the airline industry. By implementing and rigorously testing models like Random Forest, Logistic Regression, and K-Nearest Neighbors under various preprocessing conditions, Random Forest emerged as the most reliable and accurate model. The integration of Importance-Performance Analysis

(IPA) and a user-friendly Flask-based interface highlighted the project's practical utility in delivering actionable insights for improving service quality. The results emphasize the importance of data-driven strategies in enhancing customer satisfaction, optimizing operational efficiency, and supporting strategic decision-making. The methodologies and insights presented in this study provide a significant contribution to the field and pave the way for further innovations in predictive analytics and customer relationship management.

References

[1] Teejmahal, "Airline Passenger Satisfaction." [Online]. Available: https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction

[2] C.-C. Yang, "Improvement actions based on the customers' satisfaction survey," *Total Quality Management & Business Excellence*, vol. 14, no. 8, pp. 919–930, 2003, doi: 10.1080/1478336032000090842.

[3] S.-H. Park, M.-Y. Kim, Y.-J. Kim, and Y.-H. Park, "A Deep Learning Approach to Analyze Airline Customer Propensities: The Case of South Korea," *Applied Sciences*, vol. 12, no. 4, 2022, doi: 10.3390/app12041916.

[4] M. R. AlHabbal, "Predicting & Optimizing Airlines Customer Satisfaction Using Classification," 2022. [Online]. Available: https://repository.rit.edu/theses/11383/

[5] Z. Lin and I. Vlachos, "An advanced analytical framework for improving customer satisfaction: A case of air passengers," *Transportation Research Part E: Logistics and Transportation Review*, vol. 114, pp. 185–195, 2018, doi: https://doi.org/10.1016/j.tre.2018.04.003.

[6] L. Cavalcante Siebert, J. F. Bianchi Filho, E. J. d. Silva Júnior, E. Kazumi Yamakawa, and A. Catapan, "Predicting customer satisfaction for distribution companies using machine learning," *International Journal of Energy Sector Management*, vol. 15, no. 4, pp. 743–764, Jan. 2021, doi: 10.1108/IJESM-10-2018-0007.

[7] A. C. Y. Hong, K. W. Khaw, X. CHEW, and W. C. YEONG, "Prediction of US airline passenger satisfaction using machine learning algorithms," *Data Analytics and Applied Mathematics (DAAM)*, 2023, [Online]. Available: https://api.semanticscholar.org/CorpusID:259672397