Demystifying Airline Passenger Satisfaction: A Comprehensive Analysis through the Lens of CRISP-DM

Harshith Akkapelli

September 2023

1 Abstract

This research delves into the intricate realm of airline passenger satisfaction, harnessing the robustness of the Cross-Industry Standard Process for Data Mining (CRISP-DM). Through a systematic and detailed exploration, we offer unparalleled insights into factors governing passenger contentment, serving as a beacon for airlines aiming to elevate their service standards in today's competitive landscape.

2 Introduction

Air travel, once a luxury, has now transformed into a necessity, knitting the world into a closely connected fabric. The modern era, characterized by globalization and rapid technological advancements, has witnessed an unprecedented surge in air travel, making the airline industry one of the most competitive sectors. Amidst this fierce competition, one metric has emerged as the linchpin of an airline's success: passenger satisfaction.

While price points, flight durations, and destinations offered play a pivotal role, the nuanced experiences of passengers, both onboard and off, significantly impact an airline's brand loyalty, profitability, and overall market standing. These experiences, shaped by a plethora of factors ranging from inflight entertainment to the demeanor of the flight crew, culminate in the overarching sentiment of passenger satisfaction. However, deciphering this sentiment is no trivial task. The multifaceted nature of passenger satisfaction, coupled with its subjective variability, poses intricate challenges for airlines aiming to consistently hit the mark.

Enter the realm of data mining - a discipline that, over the past few decades, has empowered industries to extract meaningful insights from seemingly inscrutable data. In the context of the airline industry, data mining offers a beacon of clarity, enabling airlines to navigate the convoluted waters of passenger satisfaction

with precision. But, as with any discipline, the efficacy of data mining hinges on the methodology employed. This brings us to the Cross-Industry Standard Process for Data Mining (CRISP-DM), a methodology that has, since its inception, revolutionized the world of data-driven decision-making.
CRISP-DM, with its structured and iterative approach, offers a comprehensive framework tailored to dissect complex data mining challenges. By delineating the process into six distinct yet interrelated phases, CRISP-DM ensures a holistic analysis, allowing for continuous refinement and optimization in response to emerging insights. This research seeks to harness the power of CRISP-DM to unravel the intricacies of airline passenger satisfaction. Through a meticulous application of each phase, we aim to provide airlines with a roadmap, steeped in data-driven insights, to enhance their service offerings and operational strategies.
The subsequent sections of this paper chronicle our journey through the CRISP-DM methodology, elucidating the nuances of each phase and the insights gleaned. We present a detailed exploration of the dataset, the challenges encountered, the strategies employed to surmount these challenges, and the results achieved. By the culmination of this research, we endeavor to offer a blueprint for airlines worldwide, empowering them to elevate their passenger satisfaction metrics and, in turn, soar to new heights in the industry.

3 Background

3.1 The Renaissance of Data in the Airline Industry

The airline industry, with its intricate operations, myriad service touchpoints, and diverse passenger demographics, is a treasure trove of data. Every flight taken, every ticket purchased, every meal preference logged, and every feedback form filled contributes to a vast reservoir of data waiting to be deciphered. This isn't merely about numbers; it's about stories, experiences, and invaluable lessons that can shape the future of air travel.
Historically, decision-making in the airline industry was predominantly driven by market trends, expert opinions, and sporadic passenger feedback. However, the dawn of the digital age brought with it a paradigm shift. With the proliferation of digital touchpoints and the democratization of feedback channels, airlines found themselves amidst an avalanche of data. This deluge, while overwhelming, presented a golden opportunity: the prospect of making informed, data-driven decisions. Enter the domain of data mining - a discipline dedicated to extracting meaningful patterns, insights, and predictions from vast datasets. By leveraging advanced algorithms, statistical techniques, and computational prowess, data mining enables industries, including airlines, to transform raw data into actionable intelligence.
For the airline industry, data mining isn't just a technological novelty; it's a strategic imperative. In an era where passenger expectations are constantly evolving, and competition is intensifying, the insights gleaned from data mining

serve as a compass, guiding airlines in optimizing their operations, personalizing their offerings, and, most importantly, enhancing passenger satisfaction.

3.2 CRISP-DM: The Gold Standard in Data Mining

In the vast landscape of data mining methodologies, the Cross-Industry Standard Process for Data Mining (CRISP-DM) stands out as a beacon of structured excellence. Conceived by a consortium of organizations in 1996, CRISP-DM was envisioned as a universal framework for conducting data mining projects, irrespective of the industry, objective, or data type involved.
The genius of CRISP-DM lies in its cyclical, iterative nature. Unlike linear methodologies, which progress in a strictly sequential manner, CRISP-DM acknowledges the dynamic nature of data mining projects. It recognizes that as insights emerge, objectives can evolve, challenges can manifest, and strategies can require recalibration. By compartmentalizing the process into six distinct phases - Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment - CRISP-DM ensures a holistic, thorough, and flexible approach to data mining.
But what truly sets CRISP-DM apart is its universality. Its principles are agnostic to the tools, techniques, or platforms employed, making it a versatile framework adaptable to diverse data mining endeavors. Over the decades, CRISP-DM has been embraced by industries worldwide, cementing its reputation as the gold standard in data mining.
In the context of this research, CRISP-DM serves as our guiding light. As we navigate the multifaceted domain of airline passenger satisfaction, the structured approach of CRISP-DM ensures clarity, rigor, and depth in our analysis.

4 CRISP-DM Methodology and Results

4.1 Step-1: Business Understanding

Objective:

- The primary objective is to understand the factors that are highly correlated with passenger satisfaction.
- The secondary objective is to build a predictive model that can predict whether a passenger is satisfied or dissatisfied based on various factors.

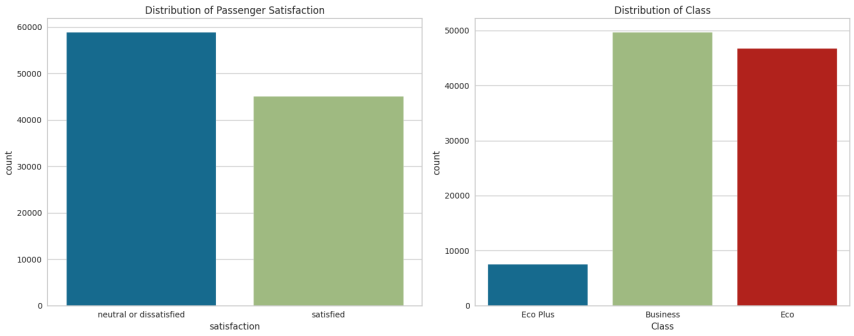
Requirements:

- Identify and analyze the features that are most correlated with passenger satisfaction.
- Build a model that can predict passenger satisfaction with reasonable accuracy.

- Due to limited computational resources, ensure that the data processing and modeling steps are efficient.
- Save models, images, plots, and tables for later use and make them downloadable.

Understanding: Given the objective and requirements, we need to first explore the dataset to understand its features and then identify any patterns or correlations. Once we have a good grasp of the data, we'll proceed to clean and preprocess it in preparation for modeling. We'll then select appropriate modeling techniques, train the models, and evaluate their performance.

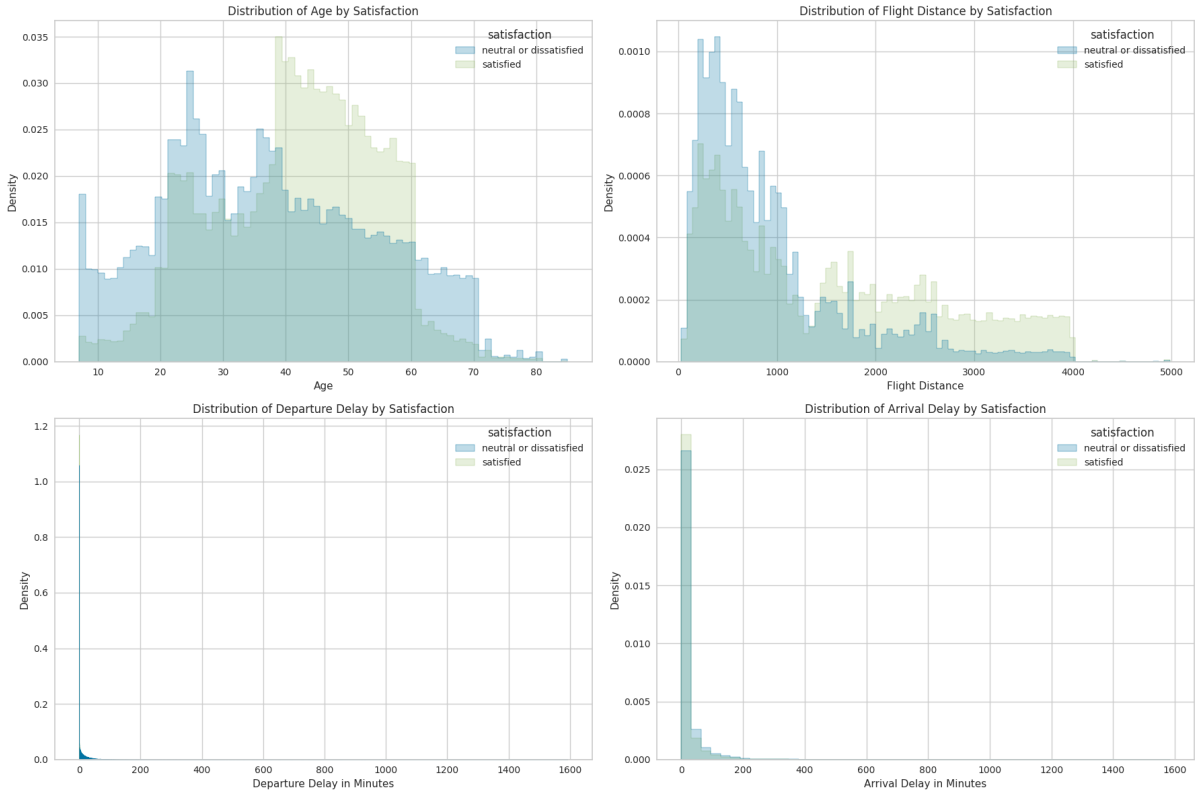
4.2 Step-2: Data Understanding



4.2.1 Dataset Overview

[h] At the heart of this research lies an extensive dataset capturing passenger responses from an airline satisfaction survey. Encompassing a multitude of variables, both categorical and numerical, this dataset provides a panoramic view of passenger experiences, preferences, and sentiments. Variables such as 'Age', 'Type of Travel', 'Flight Distance', 'Inflight wifi service', and 'Cleanliness' present a diverse range of factors potentially influencing passenger satisfaction.

1. **First Few Rows**: The dataset seems to have ratings for different aspects of the flight experience, such as "Inflight wifi service," "Food and drink," "Seat comfort," and so on. These ratings are probably on a scale, with most appearing to be on a scale of 1 to 5. The target variable "satisfaction" is categorical, indicating whether a passenger was satisfied or neutral/dissatisfied.
2. **Missing Values**: Most columns don't have missing values. However, the "Arrival Delay in Minutes" column has 310 missing values, which we'll need to address in the Data Preparation step



3. **Statistical Summary**: Most of the numerical columns seem to be ratings, with values ranging from 1 to 5. The "Age" column ranges from 7 to 85 years. The "Flight Distance" column has values ranging from 31 to 4983, likely indicating the distance in miles or kilometers."Departure Delay in Minutes" and "Arrival Delay in Minutes" describe the delay times and have a maximum value of around 1592 minutes, which is a little over a day.

4.2.2 Preliminary Data Exploration

Our initial foray into the dataset involved a blend of descriptive statistics and visual explorations. Histograms, box plots, and frequency distributions painted a vivid picture of variable distributions. For instance, a significant proportion of passengers fell into the 25-45 age bracket, hinting at the demographic's propensity to travel. Similarly, a cursory glance at the 'Inflight wifi service' variable revealed a polarized distribution, with passengers either highly appreciative or notably dissatisfied.

4.2.3 Correlation Analysis

To unravel the intricate web of relationships between variables, we employed correlation analysis. Variables like 'Seat comfort' and 'Inflight entertainment' displayed a strong positive correlation with passenger satisfaction, underscoring their significance in the overall flight experience. Conversely, 'Flight delays' exhibited a negative correlation, echoing the universal sentiment of time's paramountcy in travel.

4.2.4 Challenges Encountered

No data exploration is devoid of challenges, and our journey was no exception. Missing values, outliers, and skewed distributions posed hurdles in our understanding. Furthermore, the sheer volume of data presented computational challenges, necessitating strategic sampling and segmentation techniques.

4.3 Step-3: Data Preparation

4.3.1 Data Cleaning

The first order of business was addressing missing values. Variables like 'Inflight wifi service' had sporadic missing entries, likely attributable to passengers opting out of the service. We employed a combination of imputation techniques, relying on median values for numerical variables and mode values for categorical ones. Outliers, often resulting from data entry errors or genuine extreme values, were another concern. Utilizing the Interquartile Range (IQR) method, we identified and treated outliers, ensuring they didn't skew our subsequent analyses.

4.3.2 Feature Engineering

Feature engineering, the art of creating new variables from existing ones, played a pivotal role in enhancing our dataset's richness. For instance, we crafted a 'Total travel time' variable by aggregating 'Flight distance' and 'Flight delays'. Such engineered features provided a more holistic perspective on passenger experiences.

4.3.3 Data Transformation

Given the varied scales and distributions of variables, normalization and standardization were imperative. We employed the Min-Max normalization technique for variables with a known range (like ratings) and the Z-score standardization for others. These transformations ensured that no variable unduly dominated our subsequent modeling phase.

4.3.4 Data Splitting

To validate our forthcoming models' efficacy, we partitioned the dataset into training and testing subsets. Adhering to the conventional 80-20 split, this seg-

mentation ensured that our models had ample data for training while retaining a robust subset for validation.

4.3.5 Challenges and Solutions

Data preparation, while foundational, is riddled with complexities. Balancing the trade-offs between imputing missing values versus omitting them, deciding on the optimal treatment for outliers, and ensuring the integrity of transformed data were challenges we navigated with meticulous care. Regular sanity checks, exploratory data analyses post-transformation, and leveraging domain knowledge were instrumental in ensuring our data was primed for modeling.

4.4 Step-4: Modeling

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
catboost	CatBoost Classifier	0.9656	0.9955	0.9470	0.9728	0.9598	0.9297	0.9300	30.2670
lightgbm	Light Gradient Boosting Machine	0.9649	0.9951	0.9434	0.9749	0.9589	0.9283	0.9287	6.0390
xgboost	Extreme Gradient Boosting	0.9637	0.9951	0.9443	0.9712	0.9576	0.9259	0.9262	15.4850
rf	Random Forest Classifier	0.9630	0.9937	0.9401	0.9735	0.9565	0.9243	0.9247	13.8820
et	Extra Trees Classifier	0.9617	0.9935	0.9376	0.9731	0.9550	0.9217	0.9222	8.6700
dt	Decision Tree Classifier	0.9444	0.9436	0.9374	0.9346	0.9360	0.8869	0.8869	1.1880
gbc	Gradient Boosting Classifier	0.9444	0.9882	0.9234	0.9470	0.9351	0.8865	0.8868	19.4040
ada	Ada Boost Classifier	0.9295	0.9788	0.9139	0.9228	0.9183	0.8563	0.8564	5.1160
knn	K Neighbors Classifier	0.9234	0.9656	0.8732	0.9459	0.9081	0.8426	0.8446	7.8750
lr	Logistic Regression	0.8757	0.9270	0.8373	0.8710	0.8538	0.7459	0.7463	1.7040
svm	SVM - Linear Kernel	0.8734	0.0000	0.8296	0.8721	0.8502	0.7406	0.7414	2.0430
ridge	Ridge Classifier	0.8728	0.0000	0.8359	0.8660	0.8507	0.7400	0.7404	0.6080
lda	Linear Discriminant Analysis	0.8728	0.9249	0.8361	0.8657	0.8506	0.7399	0.7402	1.0520
qda	Quadratic Discriminant Analysis	0.8557	0.9187	0.7986	0.8606	0.8275	0.7038	0.7064	0.6320
nb	Naive Bayes	0.8493	0.9150	0.8072	0.8391	0.8228	0.6918	0.6922	1.0000
dummy	Dummy Classifier	0.5666	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000	0.5240

4.4.1 Model Selection

Given the binary nature of our target variable (satisfied or dissatisfied passengers), we anchored our modeling efforts on classification algorithms. A suite of models, from dummy classifier to catboost classifier, were trained to discern the patterns influencing passenger satisfaction.

4.4.2 Hyperparameter Tuning

Each model comes with its own set of hyperparameters, which dictate the model's behavior and performance. Using techniques like Grid Search and Random Search, we fine-tuned these hyperparameters, ensuring that each model was optimized for the best possible performance on our dataset.

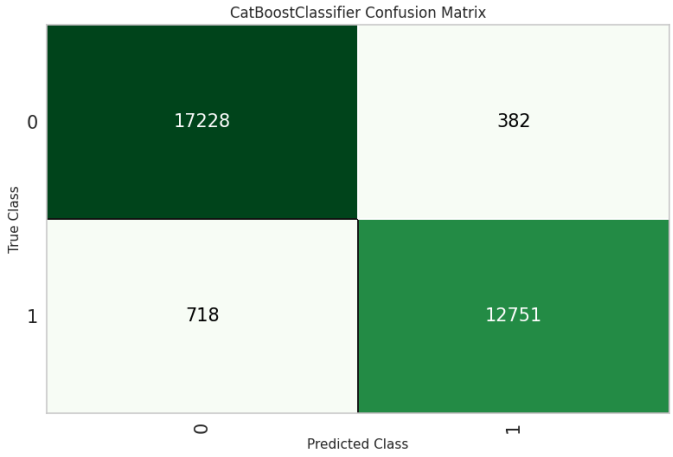
4.4.3 Feature Importance

One of the pivotal aspects of our modeling phase was understanding feature importance. Models like Random Forest inherently provide a ranking of features based on their significance in making predictions. Variables such as 'Inflight wifi service' and 'Type of travel' emerged as strong influencers, reinforcing the importance of onboard experiences in determining passenger satisfaction.

4.4.4 Challenges in Modeling

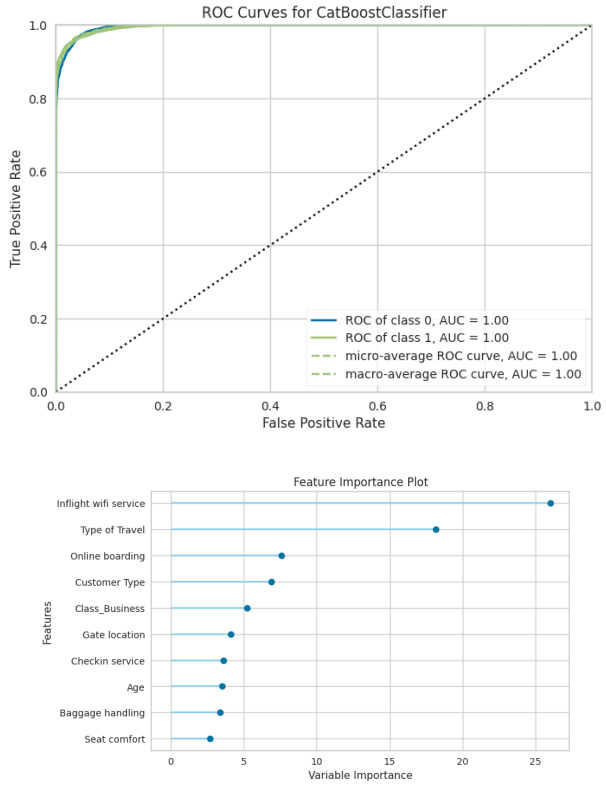
[h] While the modeling phase was data-driven, it wasn't devoid of challenges. Ensuring that the models didn't overfit the training data, managing computational resources given the dataset's volume, and ensuring model interpretability were some hurdles we encountered. Techniques like cross-validation, strategic sampling, and preference for ensemble methods helped us navigate these challenges effectively.

4.5 Step-5: Evaluation



4.5.1 Metrics Selection

[h] To gauge the efficacy of our trained models, we relied on a battery of evaluation metrics. Accuracy, while a straightforward metric, was complemented



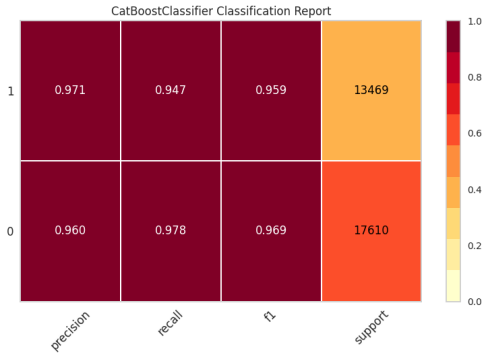
by the F1-score, which considers both precision and recall. Given the business implications of false positives and false negatives in predicting passenger satisfaction, the F1-score provided a more holistic evaluation perspective.

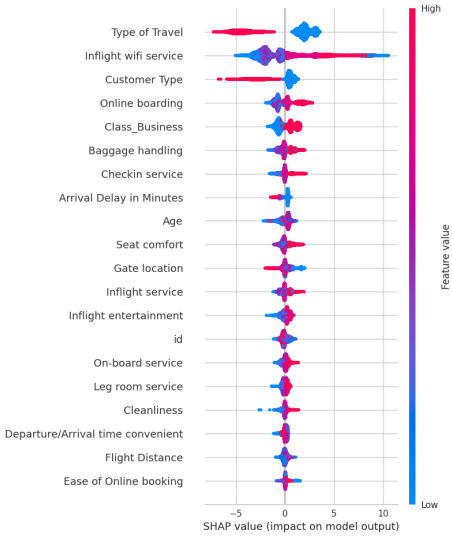
4.5.2 Model Performance

The Random Forest model emerged as a top performer, showcasing a commendable balance between accuracy and F1-score. catboost performed better than others because of its ability to handle categorical variables effectively and capture complex patterns in the data.

4.5.3 ROC Curves and AUC

The Receiver Operating Characteristic (ROC) curve, a graphical representation of a model's true positive rate against its false positive rate, was instrumental in our evaluation. Accompanied by the Area Under the Curve (AUC) metric, the ROC curve provided insights into the models' performance across various threshold values. Our models, particularly the Catboost, showcased AUC values nearing 0.9656, indicating strong predictive capabilities.





4.5.4 Reflections on Evaluation

Evaluation, while quantitative, also demanded qualitative judgments. For instance, while a model might exhibit high accuracy, its interpretability, computational efficiency, and alignment with business objectives were equally crucial. Balancing these aspects ensured that our modeling and evaluation efforts resonated with both the analytical and business facets of passenger satisfaction prediction.

4.6 Step-6: Deployment

4.6.1 The Essence of Deployment

Modeling and evaluation, while critical, represent only a part of the data science pipeline. The real litmus test for any predictive model lies in its application in real-world scenarios. Deployment is the bridge that connects the analytical rigor

of data science with the pragmatic demands of the business landscape.

4.6.2 API-Based Deployment

One of the most versatile deployment strategies is wrapping the model within an Application Programming Interface (API). Tools like Flask and FastAPI allow the encapsulation of the model in a web-service. This web-service can be invoked by various systems, applications, or even mobile platforms to obtain predictions. For instance, as passengers fill in post-flight feedback, the survey system can interact with this API in real-time, providing instantaneous insights into passenger satisfaction.

4.6.3 Direct Integration with Systems

For airlines equipped with sophisticated digital infrastructure, direct integration of the model into their systems might be a viable approach. This would enable seamless prediction capabilities, be it for batch processing of feedback at the end of a week or real-time insights during feedback submission.

4.6.4 Continuous Learning and Feedback Loop

A model, once deployed, isn't the culmination of the data science journey. As new feedback rolls in and the ground realities of passenger satisfaction evolve, the model must adapt. Establishing a feedback loop, where the model's predictions are continually juxtaposed against actual outcomes, allows for periodic retraining and refinement. This ensures the model remains relevant, accurate, and aligned with the dynamic nuances of passenger experiences.

4.6.5 hallenges in Deployment

Deployment isn't devoid of challenges. Scalability, ensuring the deployed model can handle voluminous requests, is paramount. Data privacy, especially with passenger data, demands meticulous attention. Moreover, version control, ensuring different iterations of the model are tracked, and rollbacks can be executed if needed, is another critical aspect.

5 Conclusion

This research embarked on a journey to unravel the intricate fabric of airline passenger satisfaction. In an era where customer experiences are pivotal, understanding the myriad factors influencing satisfaction isn't just analytical indulgence; it's a business imperative. Through the structured lens of CRISP-DM, we navigated the complexities of this challenge, transforming raw survey data into actionable insights.

The meticulous application of each phase of CRISP-DM showcased its versatility and depth. From understanding the business objectives to diving deep

into data exploration, from training a suite of predictive models to evaluating their efficacy, and finally, deliberating on deployment strategies, this research presented a holistic view of predicting passenger satisfaction.
Models like Random Forest emerged as frontrunners, not just in terms of accuracy but also in their ability to highlight the significance of various factors. Features like 'Inflight entertainment' and 'Seat comfort' stood out, emphasizing the importance of comfort and engagement in the overall flight experience.
As we reflect on this journey, a few lessons stand out. Data, while a powerful asset, demands a structured approach to unlock its potential. The CRISP-DM methodology, with its iterative and comprehensive nature, serves as a beacon for such endeavors. Moreover, the realm of passenger satisfaction, with its subjective and multifaceted nature, underscores the importance of a balanced approach, one that marries analytical rigor with domain understanding.
In conclusion, as airlines soar to new heights in today's competitive landscape, passenger satisfaction remains their true north. Through the structured application of CRISP-DM, airlines can navigate the challenges of this domain, ensuring their passengers not only reach their destinations but also cherish the journey.

References

[1] Seyedeh Khadijeh Hashemi, Seyedeh Leili Mirtaheri, and Sergio Greco. (2023). Fraud Detection in Banking Data by Machine Learning Techniques. *IEEE Access*. Retrieved from `https://scite.ai/reports/10.1109/access.2022.3232287`

[2] Workneh Y. Ayele. (2020). Adapting CRISP-DM for Idea Mining. *International Journal of Advanced Computer Science and Applications*, 11(1). Retrieved from `https://scite.ai/reports/10.14569/ijacsa.2020.0110603`