Business Data Science Project Report: Predicting and Understanding Airline Customer Satisfaction

I. Business Context & Dataset Insight

1.1 Introduction

In the airline industry, customer satisfaction is a key driver of loyalty, repeat purchase, and brand reputation. Because flight experiences are shaped by many touchpoints—booking, airport processes, onboard comfort, service interactions, and operational reliability—airlines need an evidence-based view of which factors most strongly influence satisfaction. Understanding these drivers supports prioritization of service improvements, better customer experience management, and targeted actions for different passenger segments. This project uses the OpenML dataset Customer Satisfaction in Airline , a large-scale customer survey dataset designed to study what explains whether a passenger reports being satisfied or dissatisfied with their flight experience.

Business question to answer: What are the main operational and service-quality drivers of satisfaction?

Methodology & Models used:

- predictive models that can estimate satisfaction from observed trip and service attributes (starting with a logistic regression baseline, then comparing with tree-based models).
- Segment customers into meaningful groups (clustering) to support differentiated experience strategies and more targeted improvement initiatives.

1.2 Dataset Description

The dataset contains roughly 130,000 observations and 22 features (including the target). The target variable is satisfaction, a binary label indicating whether the passenger was satisfied. The explanatory variables cover three main categories: Passenger and trip profile, Service quality ratings (ordinal features 0–5), Operational performance (flight delays). The target distribution is relatively balanced: about 55% satisfied vs 45% dissatisfied, making it well-suited for classification and for comparing model performance across approaches.

Originally, the dataset's features are in columns with these categories of data:

```
Data types:
CustomerType                      category
Age                                  uint8
TypeofTravel                      category
Class                             category
FlightDistance                       int64
Seatcomfort                       category
DepartureArrivaltimeconvenient    category
Foodanddrink                      category
Gatelocation                      category
Inflightwifiservice               category
Inflightentertainment             category
Onlinesupport                     category
EaseofOnlinebooking               category
Onboardservice                    category
Legroomservice                    category
Baggagehandling                   category
Checkinservice                    category
Cleanliness                       category
Onlineboarding                       uint8
DepartureDelayinMinutes              int64
ArrivalDelayinMinutes              float64
satisfaction                      category
dtype: object
```

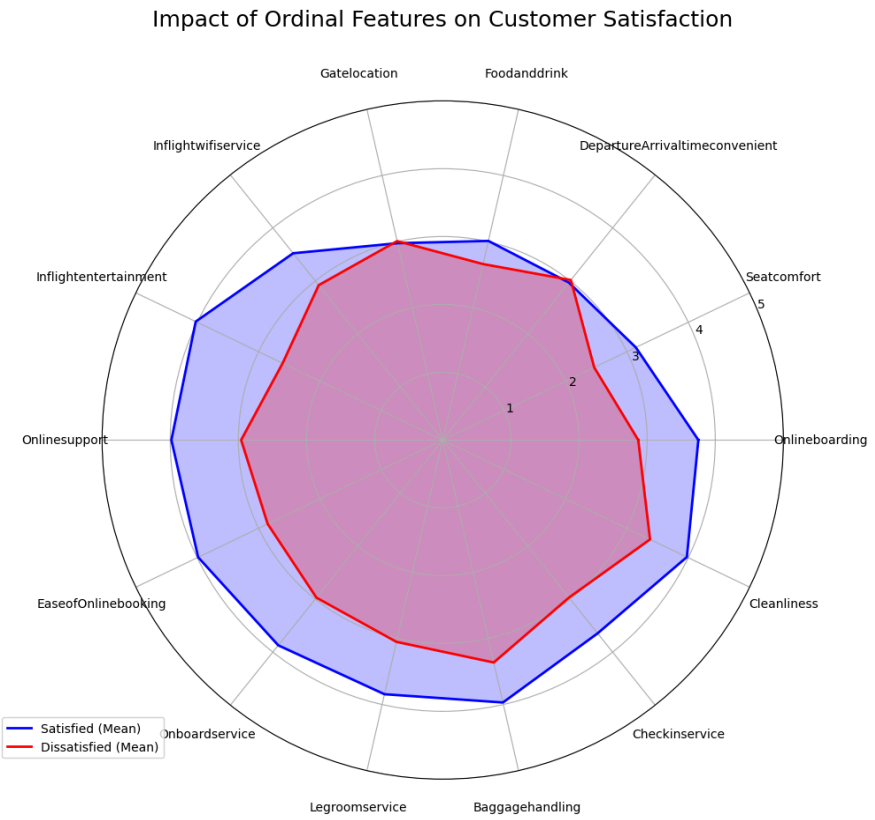
II. Comprehensive Exploratory Data Analysis

Passengers are, on average, about 39 years old (ranging from 7 to 85). Satisfaction tends to be higher among middle-aged travelers (around 40–60) and lower among younger passengers (around 20–40).

Satisfaction also differs clearly by cabin class: business-class passengers report higher satisfaction than economy travelers, with the biggest gaps showing up in areas like in-flight entertainment and online services (e.g., booking and boarding).

Flight distance does not appear to be a strong driver on its own. While the average trip is about 1,981 miles (up to roughly 6,951 miles), passengers in the 2,000–3,000 mile range look almost equally likely to be satisfied or dissatisfied.

Most service ratings are moderately positive, averaging roughly 3–3.7 out of 5, with baggage handling and cleanliness scoring slightly higher (around 3.7). The biggest gaps between satisfied and dissatisfied passengers show up in the most variable areas—online services and in-flight entertainment (Wi-Fi, entertainment, online support/booking).



Delay variables are strongly right-skewed: most flights have zero delay, but a small number have very large delays (over 1,500 minutes). Missing values in ArrivalDelayinMinutes are filled with the median to limit the influence of these outliers.

Finally, departure and arrival delays are almost perfectly correlated, meaning they carry overlapping information; To avoid redundancy, "DepartureDelayinMinutes" is dropped.

We decided to remove it because customers are more focused on their arrival time than their departure time, as the arrival time determines how late they'll be. While a delayed departure can often be made up for during the flight, a delayed arrival time means the plane has landed later than it was supposed to.

IV. Modeling Approach

Modeling goal & reason for model choice: We frame satisfaction as a binary prediction problem, so we can both quantify what drives satisfaction and build a tool that flags passengers at higher risk of dissatisfaction for targeted improvements.We chose this approach as logistic regression serves as a simple, interpretable baseline. Then, decision trees capture non-linear relationships in an easy-to-explain way. Random forests improve stability and accuracy by averaging many trees. XGBoost is included because boosted trees often perform best on tabular survey-style data.

Data preparation: satisfaction is the label so it's separated from the predictor. Categorical variables are encoded, and standardize features for logistic regression so coefficients are comparable. We use a stratified train/test split to keep the satisfied / dissatisfied balance consistent. Because departure and arrival delays are strongly correlated, we keep only one (dropping DepartureDelayinMinutes) to avoid redundant information.

4.1 Baseline Model: Logistic Regression

The logistic regression baseline delivers solid but clearly "baseline" performance on the held-out test set: accuracy = 0.8297, ROC-AUC = 0.8283, and F1 = 0.8443, which means the model benefits from stronger regularization to avoid overfitting while keeping a stable, interpretable set of effects. Below show features like: entertainment, service quality, comfort are positively important. Whereas customer type, travel type, and inconvenience can negatively impact the satisfaction.

In practical terms, this result shows that a simple linear model already captures a meaningful share of the satisfaction signal from service ratings and operational variables, making it a good reference point and an interpretable benchmark. However, it is also substantially outperformed by the tree-based models in the notebook, suggesting that satisfaction is influenced by non-linear patterns and interactions (e.g., different drivers depending on passenger type/class/age) that logistic regression cannot fully capture.

4.2 Tree-Based Models

The tree-based models outperform the logistic regression baseline by a wide margin, which suggests that satisfaction is driven by non-linear effects and interactions (for example, the same service rating can matter differently depending on passenger profile or travel context). Overall, XGBoost is the strongest production-style choice, while Random Forest is a very close second; the Decision Tree is useful mainly for simple explainable rules, not for peak performance.

The **decision tree** is the simplest of the three and trains very quickly (about 2.9 seconds). It already performs strongly (accuracy 0.9428, ROC-AUC 0.9433, F1 0.9472) and is easy to interpret because it produces clear "if–then" rules. The downside is that a single tree can be sensitive to the specific training sample, so its predictions may be less stable.

The **random forest** improves reliability by combining many trees. This gives very strong performance (accuracy 0.9564, ROC-AUC 0.9570, F1 0.9598) and tends to generalize well, but it takes much longer to train (about 91 seconds) because it builds many models.

XGBoost gives the best overall results in your comparison while staying relatively fast. It achieves the highest scores (accuracy 0.9570, ROC-AUC 0.9576, F1 0.9603) and trains in about 9 seconds. Its improvement over the random forest is small, but it is consistent, which makes it the best performance–speed trade-off in your results.

V. Model Evaluation

5.1 Evaluation Metrics

We evaluated the models using these essential metrics to have accurate and reliable information and avoid biases:

- **Accuracy**, to measure the global proportion of the correct predictions,
- **F1-score**, which is a balance between precision and recall,
- **ROC-AUC**, which allows us to measure the model capacity to differentiate satisfied clients to unsatisfied clients.

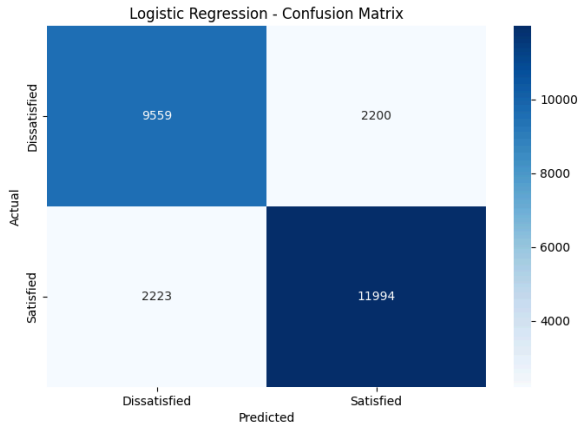
Figure 1 illustrates the ROC curves for the Logistic Regression and XGBoost models. The curve corresponding to XGBoost dominates the baseline model across almost all thresholds, confirming its superior discriminative power. This visual result is consistent with the higher ROC-AUC score obtained by tree-based models.

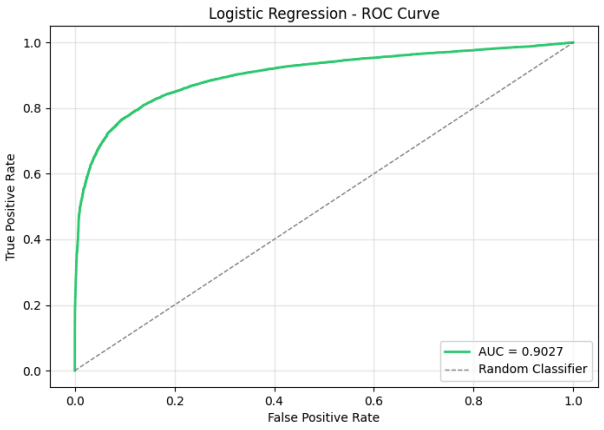
5.2 Results and Comparison

Results show that the logistic regression is a correct baseline but is limited. **XGBoost (Gradient Boosting)** has better global performances, with higher ROC-AUC and F1-score.

These models give a better prediction, even if they have an interpretability slightly reduced compared to logistic regression.

The ROC-AUC is 0.83, the F1-score is 0.84, and the accuracy of the logistic regression baseline is 82.97%. This demonstrates its shortcomings in capturing intricate and non-linear relationships while also confirming its applicability as a baseline model.





The baseline is greatly outperformed by tree-based models. With an F1-score of 0.95, the Decision Tree achieves an accuracy of 94.28%. The best outcomes are obtained through ensemble methods: While XGBoost marginally outperforms all other models with an accuracy of 95.70%, a ROC-AUC of 0.96, and an F1-score above 0.96, Random Forest achieves an accuracy of 95.64%, a ROC-AUC of 0.96, and an F1-score of 0.96.

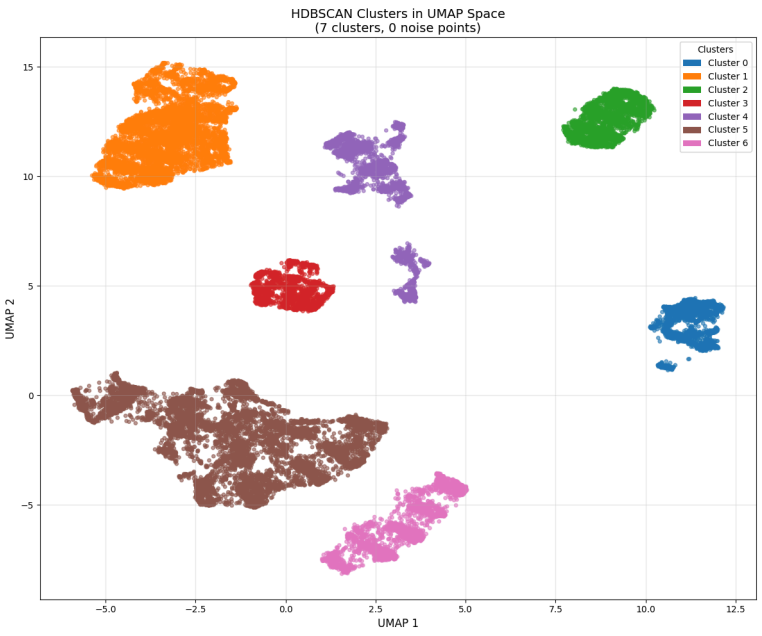
VII. Customer Segmentation (Clustering)

We evaluated multiple clustering approaches, some of them weren't very useful/conclusive (these were done with no Dimension Reduction techniques:

1. K-Means: Poor results with silhouette score <0.2, indicating weak cluster separation
2. Hierarchical Clustering: Dendrogram suggested 4 potential groups but uncertainty about cluster density

Final Approach: PCA + UMAP + HDBSCAN

- PCA: Retains 95% variance while reducing noise (17 components)
- UMAP: Preserves local and global structure for better cluster separation
- HDBSCAN: Density-based clustering that automatically determines optimal cluster count



The pipeline identified 7 distinct customer segments with 0 noise points, ordered by satisfaction:

Cluster 6: Size 2,843, Satisfaction 99%: Promoter

Cluster 5: Size 9,068, Satisfaction 68%: Promoter

Cluster 1: Size 7,407, Satisfaction 48%: Neutral

Cluster 4: Size 3,263, Satisfaction 46%: Neutral

Cluster 0: Size 2,157, Satisfaction 42%: Neutral

Cluster 3: Size 2,093, Satisfaction 39%: At-Risk

Cluster 2: Size 3,169, Satisfaction 15%:At-Risk

VIII. Limitations and Future Improvements

Not causal: The models show associations, so "important features" are not guaranteed to cause satisfaction changes. Since the results come from one dataset/context; performance and drivers may change across airlines, routes, seasons, or time.

Possible leakage / limited actionability: Many predictors are service ratings that closely reflect the satisfaction outcome, which can inflate performance and is less useful for early, operational decisions.

Validation depth: Using mainly a single split and a few metrics can miss instability across samples and hides business-relevant trade-offs (e.g., missing dissatisfied customers).

IX. Conclusion

Business insights: Drivers of Satisfaction

Across the analysis, customer satisfaction is driven primarily by the quality of the in‑flight experience—especially in‑flight entertainment, cleanliness, seat comfort, and related onboard services. Operational disruptions also matter: longer arrival delays are strongly associated with lower satisfaction. Together, these results suggest that the biggest gains come from improving high-impact service touchpoints and strengthening disruption management, rather than relying only on minor operational tweaks.

Target customers:

The segmentation points to three clear groups to act on. First, the "promoter" clusters (5 and 6) are highly satisfied (around 68–99%) and are mostly business-class and older travelers. The priority here is to keep them loyal through retention initiatives and premium offers. Second, the "at-risk" clusters (2 and 3) report very low satisfaction (around 15–39%) and are most likely to churn, so they should be addressed first with targeted improvements to the main pain points—especially the digital journey, onboard experience, and how delays are handled. Finally, the "neutral" clusters (0, 1 and 4) sit in the middle (around 42–48% satisfied) and can be moved upward with lighter interventions such as tailored upselling and loyalty nudges.

Key takeaways:

Premium segments are already performing well, while younger economy travelers are the main improvement area. Prioritizing upgrades to the onboard/digital experience and reducing the impact of delays is the most direct path to increasing satisfaction and retention.