Data Science Case Study

Customer Satisfaction Prediction for Rail Services

An End-to-End Data Science Case Study to Support Customer Retention

119K+ Records · EDA & ML · Customer Satisfaction Optimization

Business Context

Business Case

Challenge

A train company has collected customer feedback data on service ratings by post-service email request. The organization seeks to transform this raw data into actionable insights to drive strategic decision-making.

Objective

Leverage data analysis and machine learning classification models to predict customer satisfaction and identify the most influential service variables that drive positive customer experiences.

Strategic Goals

Improve Retention

Reduce main causes of dissatisfaction to increase repeat bookings

Targeted Promotions

Make promotions more efficient by targeting high-risk segments

Operational Insights

Help operations teams prioritize the right improvement actions

Caveat: a key business limitation intuition here is that service ratings are collected after the trip, and customers are not forced to submit them. As a result, rating-based features are informative but not consistently available for every customer.

Data Foundation

The Dataset

119,567
Total Records

25
Features

363
Missing Values

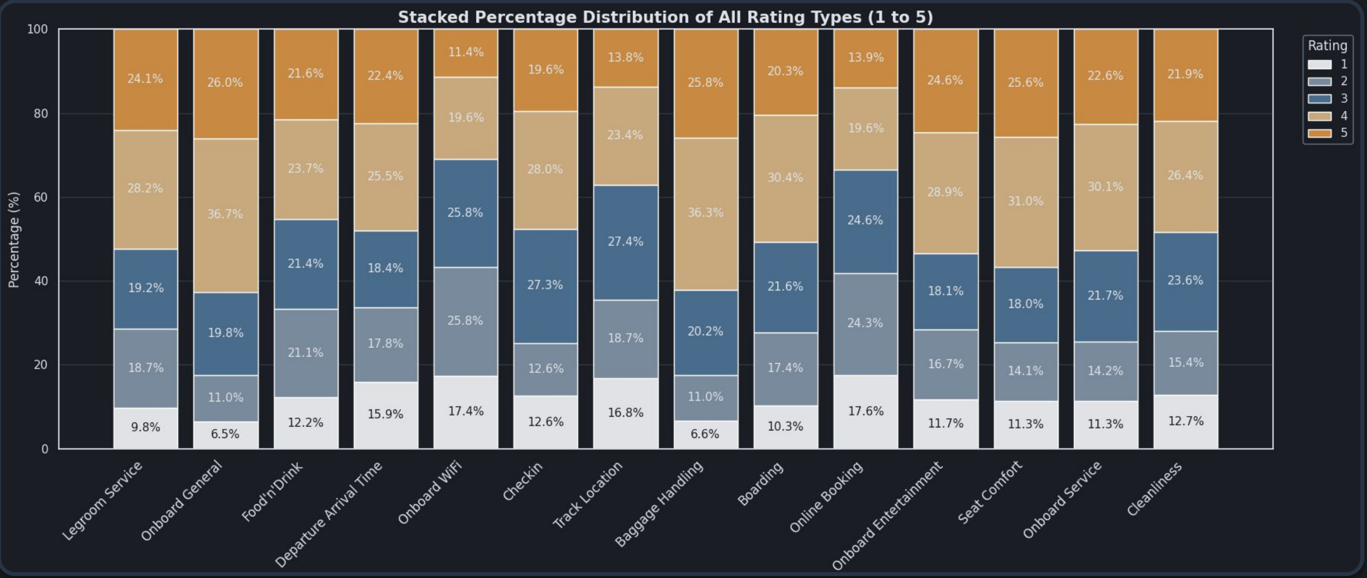
Data Quality Status

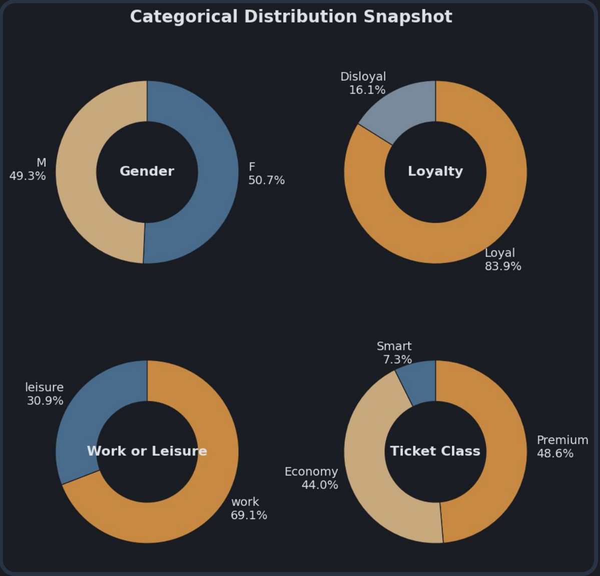
Completeness	99.7%
Missing Values	Arrival Delay in Minutes
Duplicate Records	0

✓ Data has to be treated to be ready for modelling

Feature Categories

- **Customer Info:** Age, Gender, Loyalty, Purpose
- **Travel Details:** Ticket Class, Distance, Delayes
- **Service Ratings (1-5 Stars):** 14 features
- **Target:** Satisfied (Y/N)





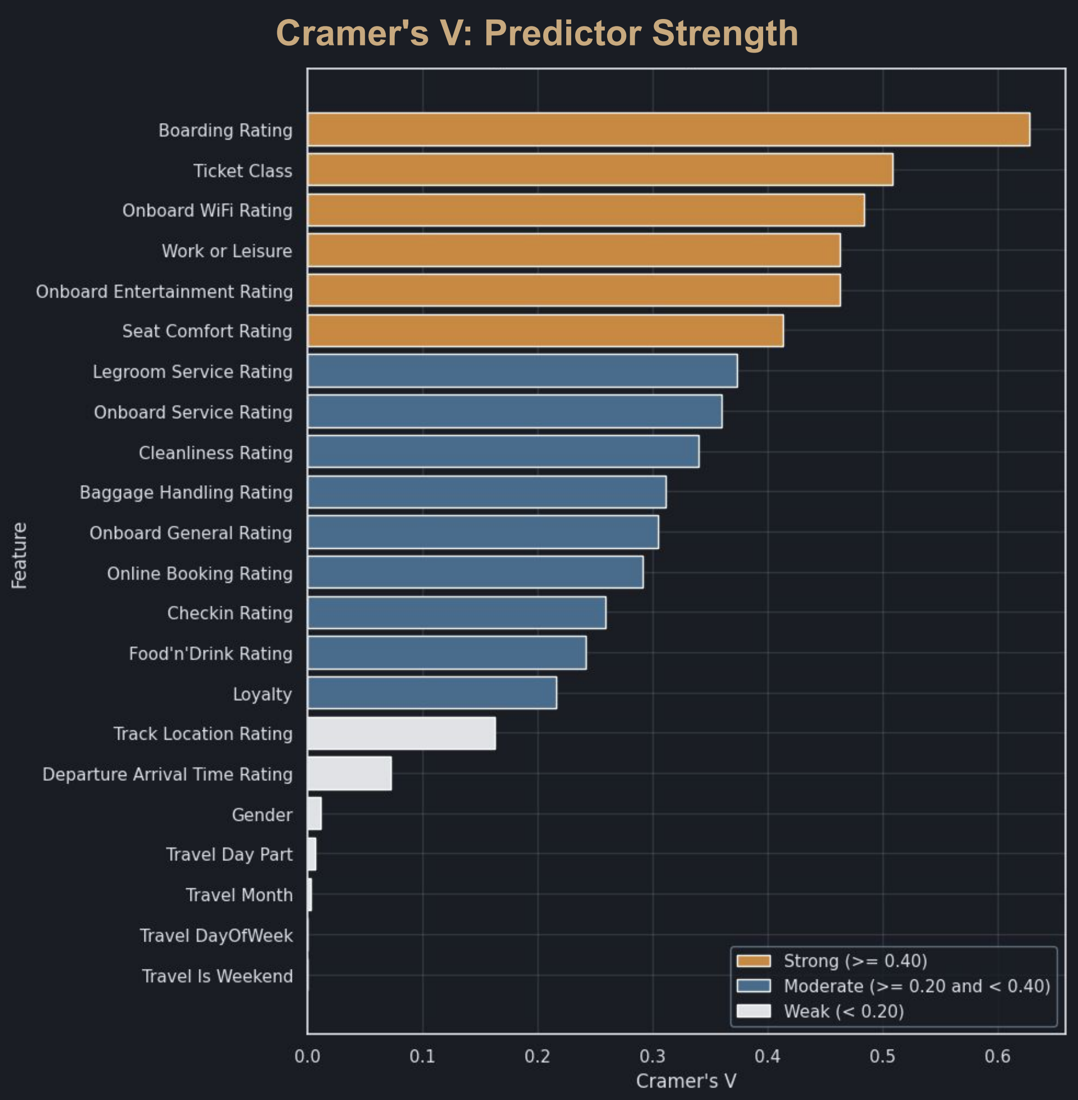
Satisfaction Distribution

56.7%
Satisfied (Y)

43.3%
Not Satisfied (N)

Exploratory Data Analysis

Feature Analysis



Satisfaction Rate Lowest 10 Categories

Purpose : Leisure	8%
Onboard Entert. : 1 ⭐	9%
Boarding : 2 ⭐	11%
Onboard Service : 1 ⭐	17%
Cleanliness : 1 ⭐	17%
Legroom Service : 1 ⭐	17%
Food and Drink : 1 ⭐	17%
Ticket Class : Economy	17%
Loyalty: Disloyal	18%
Seat Comfort : 1 ⭐	19%

Cramer's V Interpretation

Association strength with Satisfaction (0 none, 1 perfect). <0.10 negligible, 0.10–0.20 weak, 0.20–0.40 moderate, >0.40 strong. Not directional.

Near Zero Cramer's V Values

For Travel DayOfWeek / Travel Is Weekend, satisfaction rates are flat across days/weekends, so these engineered calendar features carry no meaningful signal for predicting satisfaction.

Interesting Finding

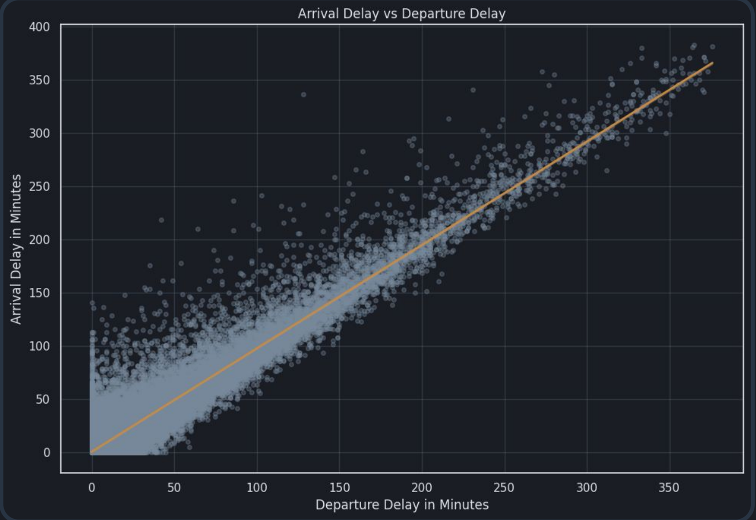
Customer that rated the WiFi with 5 stars are satisfied 99% of the times.

Note: Service-experience ratings dominate predictive power, especially Boarding and WiFi. Lowest satisfaction concentrates in Leisure + Economy + Disloyal segments with 1★–2★ service scores.

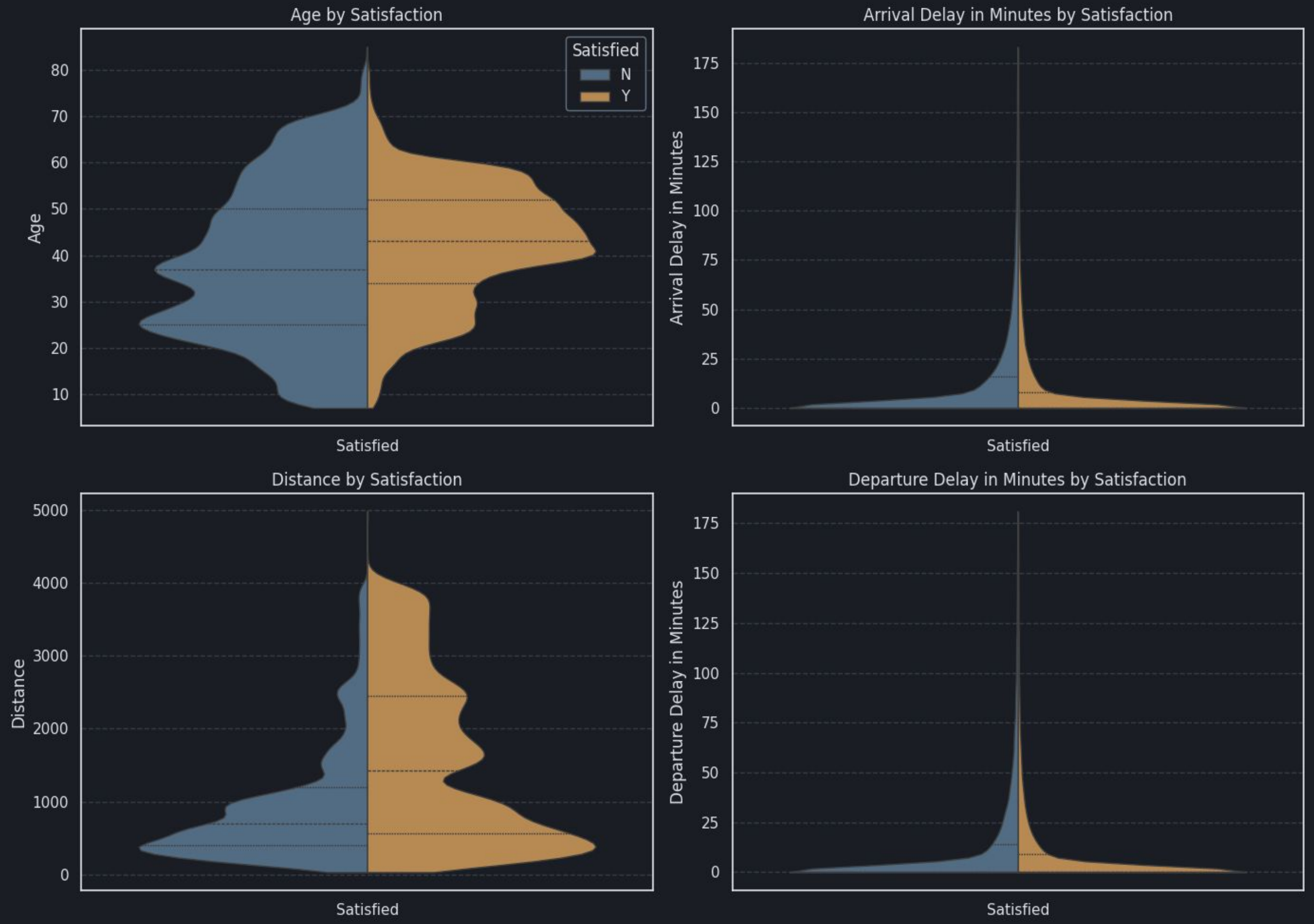
Exploratory Data Analysis

Linear Feature Analysis

Correlation Matrix: Numerical Features



Numerical Feature Satisfaction Distribution



Methodology

Modeling Process Structure

Preprocessing
Data cleaning, encoding, feature engineering

→

Stratified Split
65% Train / 15% Val / 20% Test

→

Models Pretesting
4 algorithms on validation set

↓

Evaluation Metrics
Accuracy, Precision, Recall, F1, AUC

→

Hyperparameter Tuning
RandomizedSearchCV optimization

→

Final Testing
Unbiased performance on test set

Models Evaluated

- **Random Forest** — Ensemble method
- **Decision Tree** — Interpretable
- **KNN** — Instance-based
- **Logistic Regression** — Linear baseline

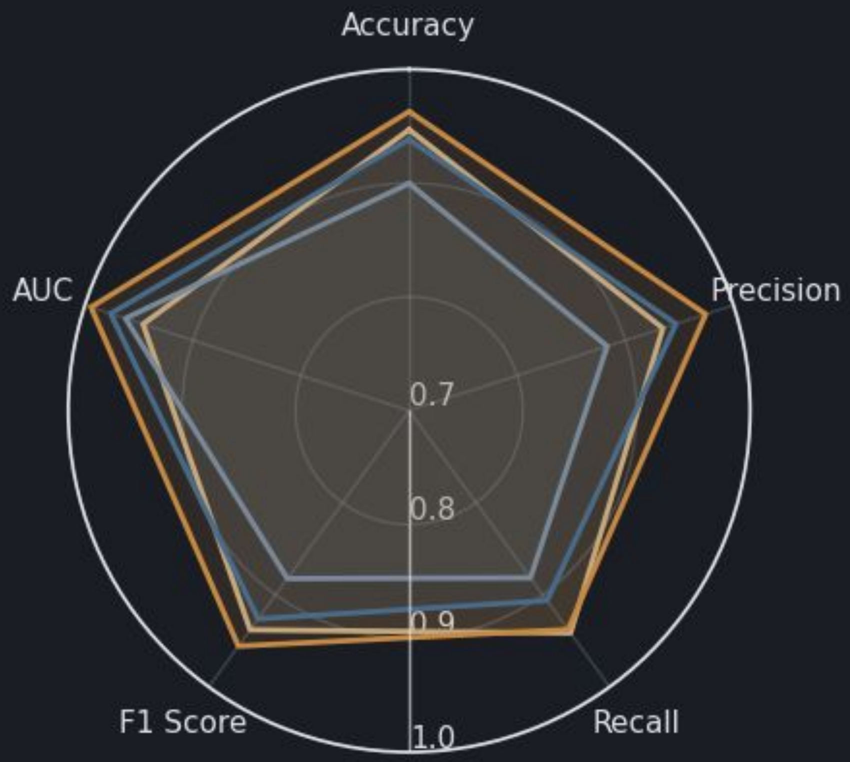
Key Methodology Decisions

- **Stratified split** ensures balanced class distribution across all sets
- **Cross-validation** for robust hyperparameter selection
- **Test set held out** until final evaluation to prevent data leakage
- **Multiple metrics** for comprehensive model assessment

Model Performance

Pre-Tuning Results & Comparison

Model Performance Radar (Zoomed)



Validation Metrics Summary

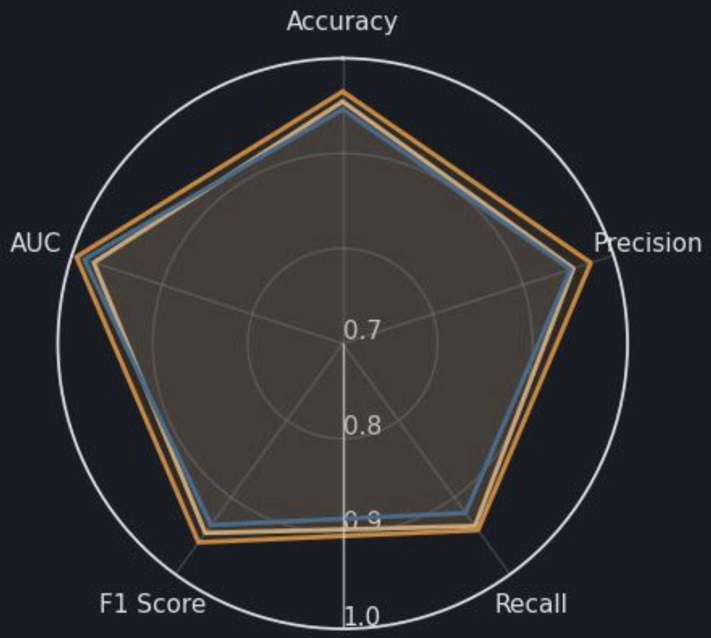
Model	P	R	F1	AUC	Accuracy
Random Forest	P: 97.1%	R: 93.7%	F1: 95.4%	AUC: 99.4%	96.1%
Decision Tree	P: 93.5%	R: 94.1%	F1: 93.8%	AUC: 94.4%	94.5%
KNN	P: 94.7%	R: 90.6%	F1: 92.6%	AUC: 97.5%	93.8%
Logistic Regression	P: 88.3%	R: 88.4%	F1: 88.3%	AUC: 96.3%	89.9%

Key Finding: Random Forest dominates across all metrics with 96.1% accuracy and AUC of 99.4%, indicating strong discriminative power and robust performance. On the other hand, in this pre-selection of model training, Logistic Regression gets outperformed from the other more complex models.

Final Evaluation

Tuned Test Results & Comparison

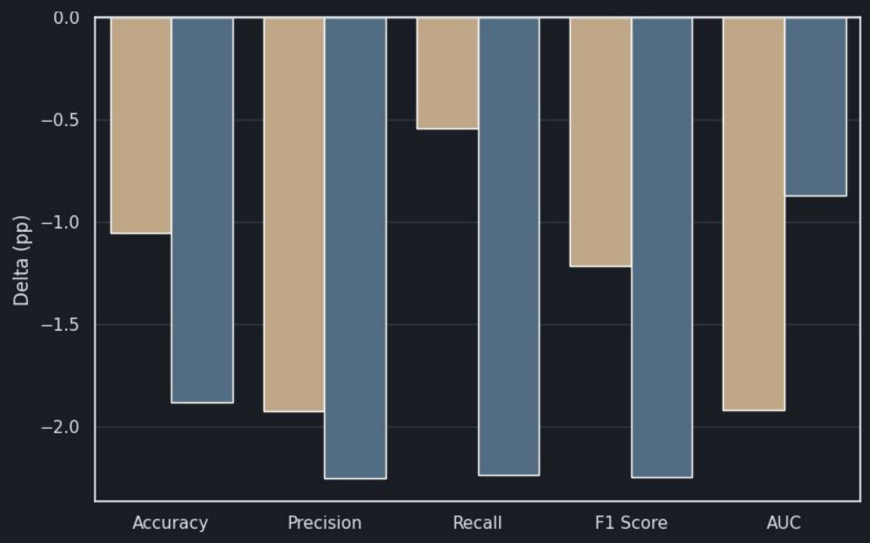
Test Performance Radar



Test Metrics Summary

	Accuracy
Random Forest P: 97.1% R: 93.1% F1: 95.6%	96.3% AUC: 99.4%
Decision Tree P: 95.6% R: 93.1% F1: 94.3%	95.2% AUC: 97.5%
K-Nearest Neighbors P: 94.7% R: 94.6% F1: 93.2%	94.3% AUC: 98.4%

Random Forest Point Percentage Delta



Performance Analysis

Validation-to-test deltas are very small, suggesting that the model probably didn't overfit.

Best Model

Random Forest maintains lead with 96.3% accuracy, though margin over the Decision Tree (95.2% -> -1.1%) is minimal.

Generalization Concern

Main risk is **real-world drift** (seasonality, policy/operations changes, survey-response bias) → monitor & retrain.

Extended Analysis

Limitations & Without Ratings Analysis

Limitations

Missing Values

Only "Arrival Delay in Minutes" has missing values (363 rows ≈ 0.30%). It was imputed using a linear regression from Departure Delay to preserve consistency and retain records.

Rating Dependency

Current models rely heavily on post-trip ratings, which may not be available in real-time prediction scenarios.

Distribution Shift

Train/val/test come from the same historical sample, but production may shift over time. Add drift monitoring + periodic retraining.

Without Ratings Scenario

To ensure deployability in real operations, a complementary analysis was conducted: **predicting satisfaction without using any post-trip rating features**.

- Uses only **demographics, travel details, and service class**
- Enables **pre-trip real-time predictions or when ratings are missing**
- More practical for **proactive customer retention**

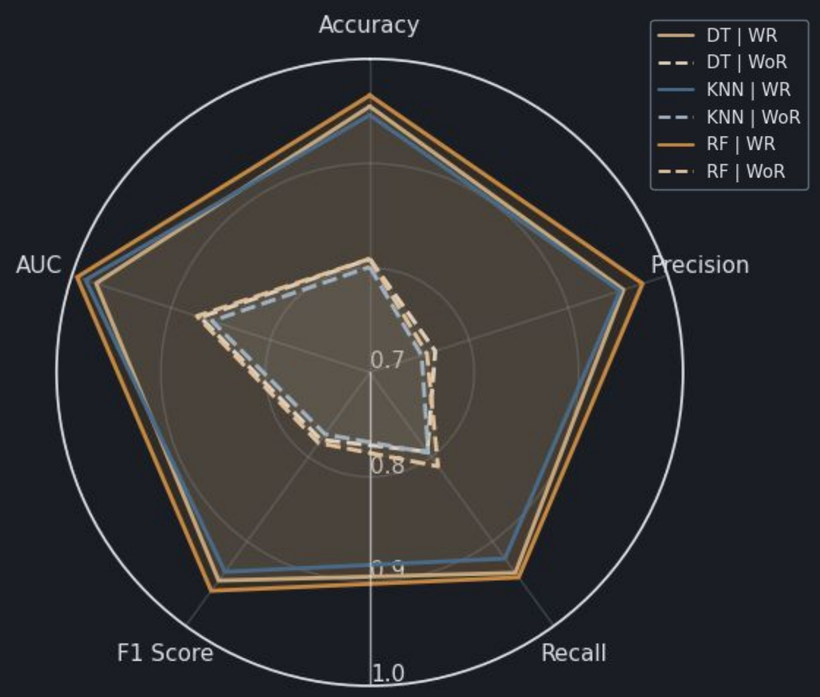
Research Question

Can we achieve a good enough performance **without rating features**?
This tests whether satisfaction is predictable from customer profile and trip characteristics alone.

Alternative Pipeline Results

Without Ratings - Performance Comparison

With vs Without Ratings



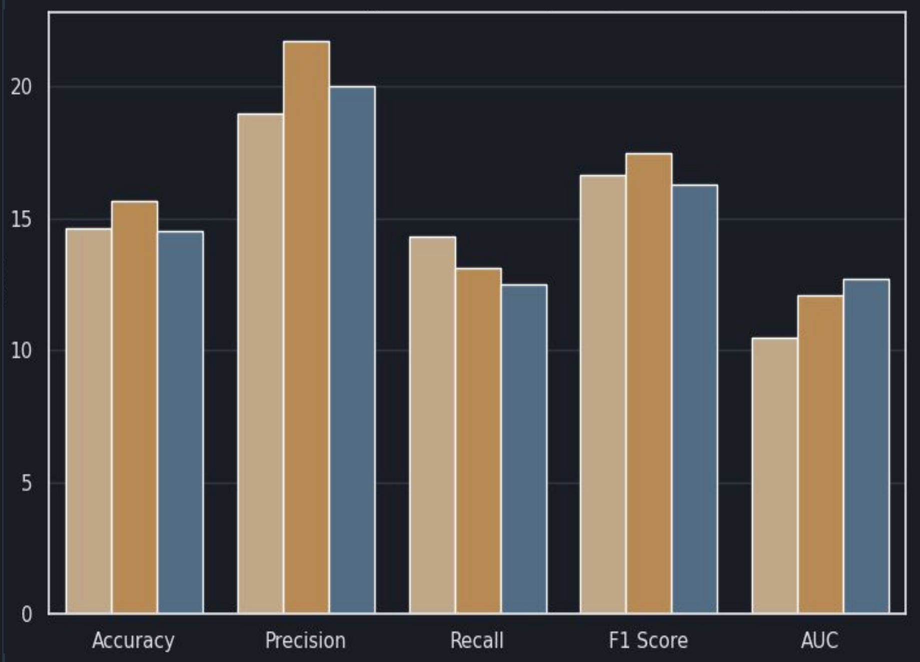
Accuracy Shift With vs Without Rating

Random Forest
96.3% → 81.1%

Decision Tree
95.2% → 80.6%

K-Nearest Neighbors
94.3% → 79.9%

Performance Delta: With - Without Ratings



Expected Metrics Drop

Removing rating features **lowered metrics by a considerable amount**. This is the **trade-off** for a model usable when post-trip surveys are missing.

Why This Happens

Ratings carry direct **service-experience feedback**, which is the **strongest possible signal**. Without ratings, the model relies on **weaker proxies**, so **predictive power drops**.

Deployment Insight

Use a two-pipeline setup:
With-ratings → post-trip diagnosis & target recovery.
Without-ratings → scoring when ratings are missing.

Final Assessment

Conclusions & Recommendations

Key Findings

1 **Random Forest Best Overall**

Achieved the best accuracy in both with and without ratings. Strongest and most consistent and versatile performer.

2 **Dissatisfaction is concentrated in clear service pain points**

Biggest risk groups: Leisure, Economy, Disloyal customers, plus low scores in Boarding, WiFi, Entertainment, Cleanliness, Onboard Service, Food & Drink, Seat Comfort.

3 **Ratings Drive Peak Accuracy But Operations Need Two Pipelines**

Models with ratings clearly outperform (96.3% vs 81.1%). Since many customers don't submit post-trip ratings, Deploy a two-model setup: with-ratings for post-trip diagnosis/recovery, and no-ratings for broad proactive coverage.

Business Call To Action

→ First focus on Boarding flow, WiFi reliability, Entertainment quality, Cleanliness, and Onboard service consistency.

→ Use the with-ratings model for high-precision post-trip recovery and root-cause diagnosis, and use the no-ratings model for broad proactive coverage when surveys are missing.

→ Build retention playbooks for Economy + Leisure + Disloyal passengers (service recovery vouchers, tailored offers, loyalty conversion campaigns)

Future Improvements

- Collect more diverse training data regarding customer behaviour (non-rating)
- Increase survey response rate to expand coverage of the higher-accuracy ratings pipeline
- Run A/B tests on model-driven interventions and measure retention uplift

Thanks for the attention!

Q&A