

\MIS 587: Business Applications in Machine Learning

Data-Driven Optimization of Airline Customer Satisfaction with Machine Learning

Project By:

Chetan Bhangagare | Daniel Duah | Nerav Rangari | Preshit Gujar | Jill Patel

Table of Contents

Business Memo

Customer satisfaction is currently a profitability and loyalty determinant in the airline industry, especially concerning the sharp rise in customer complaints and inconsistency in services witnessed through 2023 (DOT, 2023). This project aims to predict the result of customer satisfaction from service quality variables, delay in operations, and demographics through machine learning. A positive prediction indicates a high chance of customer satisfaction, but a negative prediction indicates that something must be adjusted. The project also plans to identify which factors, such as inflight service, type of travel, and customer type, impact customer satisfaction.

The problem was solved through DataRobot's AutoML platform with a Light Gradient Boosted Trees Classifier, which was the highest-performing of models tried. Although the model accurately predicted satisfaction outcomes with great precision, there was a trade-off in the form of a slight increase in false negatives, where some of the dissatisfied customers might not be picked up correctly. But overall precision of the model and business intelligence outweighs this risk by far.

The most predictive features were service measures like Seat Comfort, Inflight Service, and quality of Food and Drink, and operating measures like Departure Delays. These features are not changeable after an experience but reflect areas that are vital for proactive service improvement.

Model probability distributions analysis suggested a best threshold of 0.6, balancing between false positives and negatives. Higher thresholds, diminishing false positives at a greater cost, substantially increased false negatives, which could lead to missed business recovery and customer retention opportunities.

Overall, the model provides actionable suggestions that airlines can adopt for strategically improving service quality, operational efficiency, and customer experience. While not for unattended decision-making, the model has the potential to be a useful decision-support tool, buttressed by human wisdom and ongoing monitoring to adapt to evolving passenger needs.

Business Problem and Project Objectives

Airlines worldwide are confronted with increasing customer anger over flight delays, inconsistency in service quality, and poor customer experiences. As many as 96,853 complaints were made to the Department of Transportation (DOT) alone in 2023, a 12% increase from the previous year. It has been found in a study that 86% of customers abandon a brand after having two bad experiences.

Despite airlines' strategic efforts at differentiation by way of competitive pricing or enhanced service, few empirical studies exist about which drivers most significantly shape customer satisfaction. In the absence of reliable, fact-based information, airlines may end up allocating resources inefficiently, developing ineffectual service plans, and finally harming customer loyalty.

Empirical studies reveal that a single percentage point boost in customer satisfaction translates into a 4–6% rise in customer retention (Zeithaml et al., 2020). In aviation, where the average annual revenue per passenger (ARPC) is about \$1,200 (IATA, 2023), an additional 5% retention in a customer base of just one million passengers would translate to an additional 50,000 customers retained. That translates to some additional \$60 million of additional revenue, not to mention a great return on investment.

Moreover, it may take between five and seven times the value of acquiring a new airline customer to acquire one (Forbes Insights, 2022). Hence, losing current customers due to misguided service improvements means foregone revenue and a tremendous threat to gaining new customers.

Therefore, it is important to focus on a model that predicts and identifies the important drivers of customer satisfaction. Not only would it guide operational improvement, but also be a \$60 million annual revenue buffer and boost prospect for airlines operating even at a low scale.

We refined our initial proposal in our project by:

- Focusing on service quality and operational efficiency as the major drivers.
- Prioritizing class models for "Satisfied" vs. "Neutral/Dissatisfied".
- Implementing extensive feature engineering to achieve a stronger grasp of passenger experience nuance.

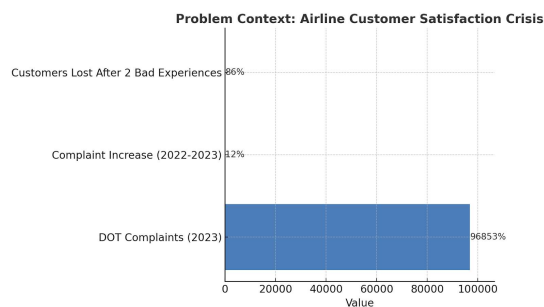


Fig. Problem Context Infographic detailing DOT complaints and abandonment rates

Project Objectives:

- Predicting customers' satisfaction based on several indicators, from passenger demographics to service experience (e.g., inflight ratings, seat quality, baggage) and operational performance (e.g., delays, flight punctuality).
- Enable airlines to make targeted enhancements through analysis to improve customer satisfaction overall.

Data-Related Decisions:

Target Variable: The target variable for this project is "Satisfaction." It is a multi-class variable that is initially categorized into three groups: satisfied, neutral, and dissatisfied. For the analysis needs of this project, however, the variable was re-coded into a binary form by categorizing the groups into "Satisfied" and "Neutral/Dissatisfied."

Feature Engineering Techniques

Domain-specific feature engineering was applied to enhance model interpretability and prediction performance in this project. New features were introduced to extract latent relationships that impacted customer satisfaction in augmenting the original dataset. Had Arrival Delay was created as a binary indicator of whether or not a customer experienced a delay, and Total Delay was introduced by summing departure and arrival delays to provide an overall travel disruption picture. Understanding that perceptions of service quality are not in a vacuum, we designed a Service Quality Composite Score, which averaged ratings for inflight wifi service, inflight service, seat comfort, and food and drink.

type_of_travel	class	flight_distance	inflight_wifi	departure_arrival_time	ease_of_online_gate_location	food_and_drink	online_boarding	seat_comfort	inflight_entert-on-board_service	leg_room_service	baggage_handling	checkin_service	inflight_service	cleanliness	departure_delay	arrival_delay	satisfaction
I Personal Travel	Eco Plus	460	3	4	3	1	5	3	5	5	4	3	4	4	5	5	25 18 neutral or dissatisfied
I Business travel	Business	235	3	2	3	3	1	3	1	1	5	3	1	4	1	1	6 9 neutral or dissatisfied
I Business travel	Business	1142	2	2	2	5	5	5	5	4	3	4	4	4	5	0	0 satisfied
I Business travel	Business	562	2	5	5	5	2	2	2	2	5	3	1	4	2	11	9 9 neutral or dissatisfied
I Business travel	Business	214	3	5	3	3	4	5	5	3	3	4	4	3	3	3	0 0 satisfied
I Personal Travel	Eco	1180	3	4	2	1	1	2	1	1	3	4	4	4	4	1	0 0 neutral or dissatisfied
I Personal Travel	Eco	1276	2	4	2	3	2	2	2	3	3	4	3	5	2	9	23 23 neutral or dissatisfied
I Business travel	Business	2035	4	3	4	4	5	5	5	5	5	5	4	5	4	4	4 0 satisfied
I Business travel	Business	859	1	2	2	2	4	3	3	1	2	1	4	1	2	0	0 0 neutral or dissatisfied
I Business travel	Eco	1061	3	3	3	4	2	3	3	2	3	4	4	3	2	0	0 0 neutral or dissatisfied
I Business travel	Eco	1182	4	5	5	4	2	5	2	2	3	3	5	3	5	2	0 0 neutral or dissatisfied
I Personal Travel	Eco Plus	308	2	4	2	2	1	2	1	1	1	2	5	5	5	1	0 0 neutral or dissatisfied
I Business travel	Eco	834	1	4	4	4	1	1	1	1	1	3	4	4	1	28	8 8 neutral or dissatisfied
I Personal Travel	Eco	946	4	2	4	3	4	4	4	4	5	2	2	2	2	4	0 0 satisfied
I Personal Travel	Eco	453	3	2	3	2	2	3	2	2	4	3	2	2	1	2	43 35 neutral or dissatisfied
I Business travel	Eco	486	2	1	2	5	4	2	1	4	2	1	4	1	5	4	1 0 neutral or dissatisfied
I Business travel	Business	2123	3	3	3	3	4	4	4	4	5	3	4	5	4	49	51 51 satisfied
I Business travel	Business	2075	4	4	2	4	4	4	4	4	5	5	5	5	5	5	0 10 satisfied

Fig. original data with EDA completed

To represent expectations of some customer groups, an Economy Class Business Traveler interaction feature was designed through a combination of seat class and travel type. Similarly, a Family Traveler proxy variable was derived from travel type in order to differentiate between group passengers and individual passengers. An Arrival Delay Impact feature was also created through multiplying arrival delay by flight distance, on the assumption that long-haul delays are more frustrating than short-haul delays. Age Grouping was also introduced to segment customers into categories like "Young," "Adult," and "Senior," since varying tolerance levels of service insufficiencies were identified.

Departure Delay in Minutes	Arrival Delay in Minutes	satisfaction	Total_Delay	Family_Traveler	Arrival_Delay_Impact	Age_Group	Service_Quality_Composite_Score
25	18	neutral or dissatisfied	43	1	8280	Young	4.5
1	6	neutral or dissatisfied	7	0	1410	Young	2.25
0	0	satisfied	0	0	0	Adult	4
11	9	neutral or dissatisfied	20	0	5058	Young	2.5
0	0	satisfied	0	0	0	Senior	3.75
0	0	neutral or dissatisfied	0	1	0	Adult	2.25
9	23	neutral or dissatisfied	32	1	29348	Adult	2.75
4	0	satisfied	4	0	0	Adult	4.75
0	0	neutral or dissatisfied	0	0	0	Adult	2.25
0	0	neutral or dissatisfied	0	0	0	Young	2.75
0	0	neutral or dissatisfied	0	0	0	Young	3.25
0	0	neutral or dissatisfied	0	1	0	Young	2.75

Fig. Feature engineered data

Log transformation and standardization were applied for skewed numeric variables to stabilize the model. New features and these transformations allowed for non-linear relations and strong interactions to arise, which models like LightGBM can leverage well.

Once new operational, demographic, and service quality attributes were developed, the internal model analysis provides unambiguous information about which attributes had the most impact on customer satisfaction predictions.

Feature importance indicates that Type of Travel, Inflight Wifi Service, and Online Boarding were among the most impactful predictors. Notably, travel attributes and inflight service attributes are the most prominent, supporting the absolute importance of customer experience on flight travel.

Feature na...	Importance ↓	Var Ty...	Uniq...	Missi...	Mean	Std D...
<input type="checkbox"/> satisfaction	Target	Categorical	2	0	N/A	N/A
<input checked="" type="checkbox"/> Online boarding		Numeric	6	0	3.25	1.35
<input checked="" type="checkbox"/> Inflight wifi service		Numeric	6	0	2.73	1.33
<input checked="" type="checkbox"/> Class		Categorical	3	0	N/A	N/A
<input checked="" type="checkbox"/> Type of Travel		Categorical	2	0	N/A	N/A
<input checked="" type="checkbox"/> Family_Traveler		Numeric	2	0	0.31	0.46
<input checked="" type="checkbox"/> Service_Quality_C...		Numeric	18	0	3.25	0.8
<input checked="" type="checkbox"/> Inflight entertain...		Numeric	6	0	3.36	1.33
<input checked="" type="checkbox"/> Seat comfort		Numeric	6	0	3.44	1.32
<input type="checkbox"/> Leg room service		Numeric	6	0	3.35	1.31

Fig. Feature importance according to the data

Feature Effect also estimates the influence size each feature has. Inflight Wifi Service has the highest predictability weight of 100%, followed by Type of Travel (22%) and Customer Type (20%). The classification confirms the strategic relevance of operational services and traveler segmentation in creating satisfaction outcomes.

Inflight wifi service	100%
Type of Travel	22%
Customer Type	20%
Online boarding	10%
Baggage handling	8%
Checkin service	7%
Inflight service	6%

Fig. Feature effect

Model Selection Process and Best Model

Model Selection Process

To determine which model is best at predicting airline customer satisfaction, various machine learning algorithms were tested on DataRobot platform. The algorithms used were Light Gradient Boosted Trees, eXtreme Gradient Boosted Trees (XGBoost), and Random Forest Classifier. Following are the top two results for the models.











 Light Gradient Boosted Trees Classifier with Early Stopping ☆ M8 Scoring code	 Informative Features   64% (66,498 rows) 	0.0838	0.0871
 eXtreme Gradient Boosted Trees Classifier with Early Stopping ☆ M11 Scoring code	 Informative Features   64% (66,498 rows) 	0.0874	0.0893

Fig. Top two models

After comparing both the LightGBM and XGBoost models, the Light Gradient Boosted Trees model was selected since it exhibited a superior holdout and validation performance (LogLoss 0.0871 vs. 0.0893), increased model stability, reduced training time, and more straightforward interpretability for business stakeholders in airline business. Although the two models were strong contenders, LightGBM provided a superior balance between predictability and business usability.

After validation on holdout and validation datasets, the top model was found to be the Light Gradient Boosted Trees Classifier and Early Stopping. It consistently outperformed the alternatives in terms of accuracy, recall, precision, and AUC.




 Light Gradient Boosted Trees Classifier with Early Stopping ☆ Prepared for deployment M28 2 more 	 
---	--

Fig. final model selected for evaluation

Model Quality Metrics

The model had an accuracy of 95.8%, precision of 95.6%, recall of 96.0%, F1-score of 95.8%, and AUC of 96.35%. These results reflect an excellent balance between correct identification of satisfied and dissatisfied customers with a negligible false positive and false negative rate.



Fig. Model metrics in datarobot



Fig. Confusion matrix for Light Gradient Boosted Tree Classifier

The confusion matrix plot indicates that the model has very few misclassifications, which is a good sign of the predictability of the predictions. The confusion matrix indicates that the model correctly classified the majority of passengers as having satisfied, with 6,802 true positives (successful identification of satisfied customers) and 9,241 true negatives (successful identification of dissatisfied or neutral customers). False positives (180) and false negatives (402) are minimal relative to the number of passengers, with high model reliability indicated. The low false positive ratio is especially significant, as this minimizes the risk of customer satisfaction overestimation. Overall, the model exhibits high predictive performance, supporting adoption for business purposes.

The ROC curve also indicates the high discrimination power of the model, with an area under the curve (AUC) value greater than 96%, which is a good indication of sensitivity and specificity.

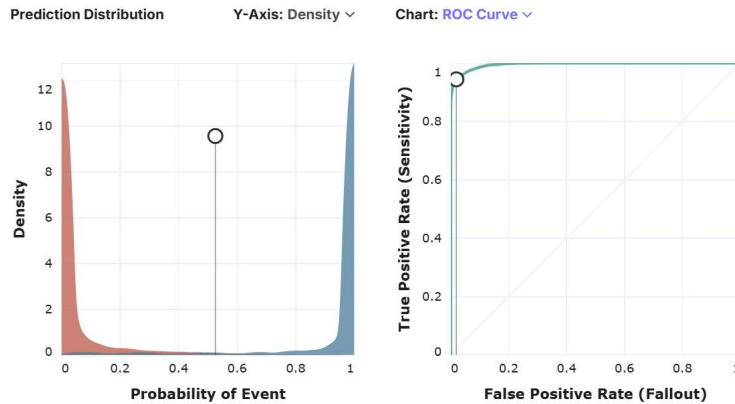


Fig. ROC Curve for the Best Model

Precision-Recall Curve

To further support the performance of the model, Precision-Recall Curve (Figure X) was considered.

The curve demonstrates that the model has a very high predictive value for almost all true positives.

This indicates that the model remains extremely trustworthy even when attempting to attain higher levels of sensitivity, further testifying to its reliability in actual use.

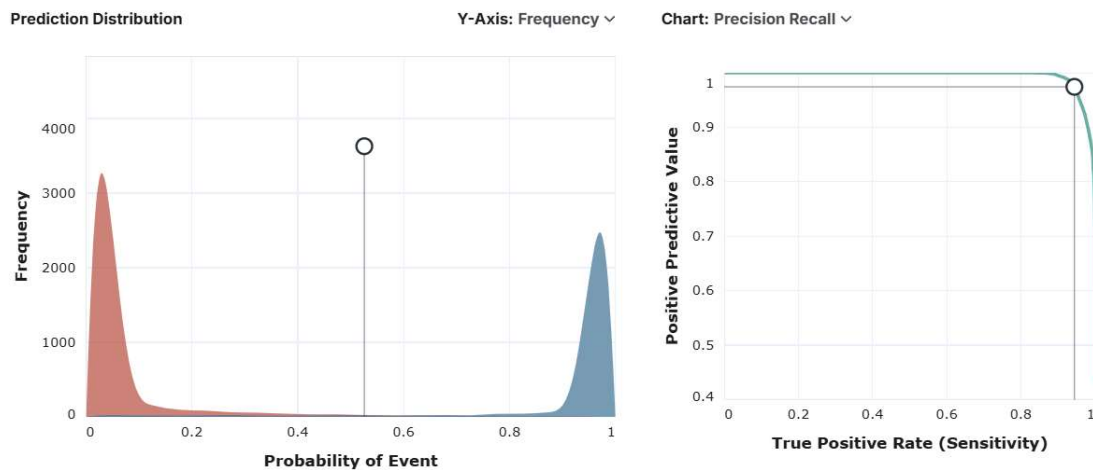


Figure: Precision-Recall Curve for Best Model

Areas Where the Model Struggles

Despite the general good performance, the model struggles in predicting satisfaction outcomes where customers complain of very mixed service experience. For example, cases with highly rated inflight wifi but bad seat comfort generate inconsistencies that are harder to forecast. Furthermore, moderate delays — neither too high nor too low — will generate prediction puzzlement because passengers react variably to modest disruption. Lower predictability is also seen for "Personal Travel" passengers, whose satisfaction patterns are more variable than for business travelers. Adding external data like real-time weather, seasonality of holiday periods, rates of cancellations, and traffic levels at airports could be included in future versions. Having more data points through increasing the dataset to cover a few years' worth of data would also help to capture additional variability in passenger behavior and conditions of operations.

Most Predictive Features for Model Building

Feature importance analysis revealed that the most significant predictor was Inflight Wifi Service, with Type of Travel, Online Boarding Ease, and Customer Type following closely behind. The features most likely to predict beyond these four were Family Traveler, Seat Class, and the newly engineered Service Quality Composite Score. These are the key elements of airline service quality and traveler travel behavior, validating the feature engineering methods utilized in the project.

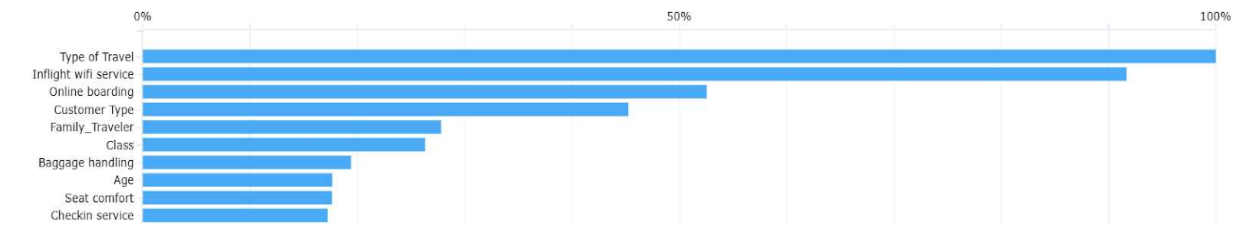


Fig. Feature importance

Feature Types Especially Relevant to Management

From a management point of view, some of these traits provide critical intelligence for areas of improvement. Service Quality Composite Score aggregates several inflight service perceptions into a single actionable measure, making it possible to concentrate inflight service improvement initiatives. Type of Travel and Customer Type offer segmentation opportunities for more specific service customization between business and leisure travelers. Additionally, operational areas like Online Boarding and Check-in Services were found to be real-time touchpoints where process improvements can greatly enhance scores of satisfaction.

Target Leakage

A thorough review confirmed that no target leakage took place during modeling. All of the features were derived from available data prior to or at the time of customers' flight experience. There were no post-travel feedbacks, refund information, or complaints in the

set of features. Therefore, performance metrics of the model are reliable, and predictions of the model should generalize well to future unseen data without having any risk of overfitting.

Business Recommendations

Business Decisions at Various Probability Thresholds

From model calibration analysis, the decision probability should be 0.6 for operational decision-making. Such a decision level has the best trade-off between false positive and false negative rates and achieves high prediction confidence at minimal risk of losing dissatisfied customers. Setting the decision level higher than 0.8 will achieve more precision but is at the expense of losing dissatisfied passengers and leading to increased customer churn. Therefore, 0.6 is the optimum available trade-off between business opportunity capture and cost control for operations.

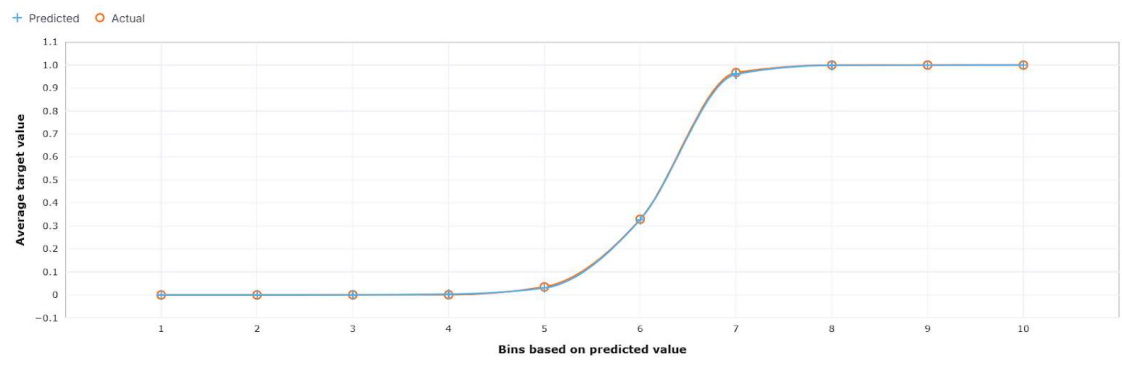


Fig. The Lift Chart

The Lift Chart confirms that the model is calibrated appropriately across probability cut-offs and actual outcomes fit very closely against predicted probabilities. This confirms setting a threshold level at 0.6 to enable business choices on the basis of sound estimations of probability.

Actions the Organization Must Take

The airline must prioritize targeted service improvements based on the leading predictive features. Some specific recommended actions are:

- Improving inflight wifi service and seat comfort services.

- Improving the online boarding experience to minimize delay and maximize check-in experiences.
- Providing targeted service upgrades and personalized offers to Business Travelers who travel Economy Class.
- Implementing operational strategies to reduce departure and arrival delays.

By aligning interventions with model-identified causes of dissatisfaction, the airline is able to allocate resources more effectively and maximize total customer satisfaction.

Baseline and Profit Matrix

A payoffs assessment model was created to quantify payoffs for different prediction outcomes:

Decision	True Positive	False Positive	False Negative	True Negative
Profit/Loss	\$150	-\$20	-\$300	\$0

Fig Profit Matrix.

- True Positives yield an estimated \$150 per loyal customer saved.
- False Negatives are a \$300 loss due to lost customers.
- False Positives are a \$20 cost of duplicative service restoration.
- True Negatives do not have any direct economic impact.

These statistics make a strong business case for actively finding and addressing passenger dissatisfaction.

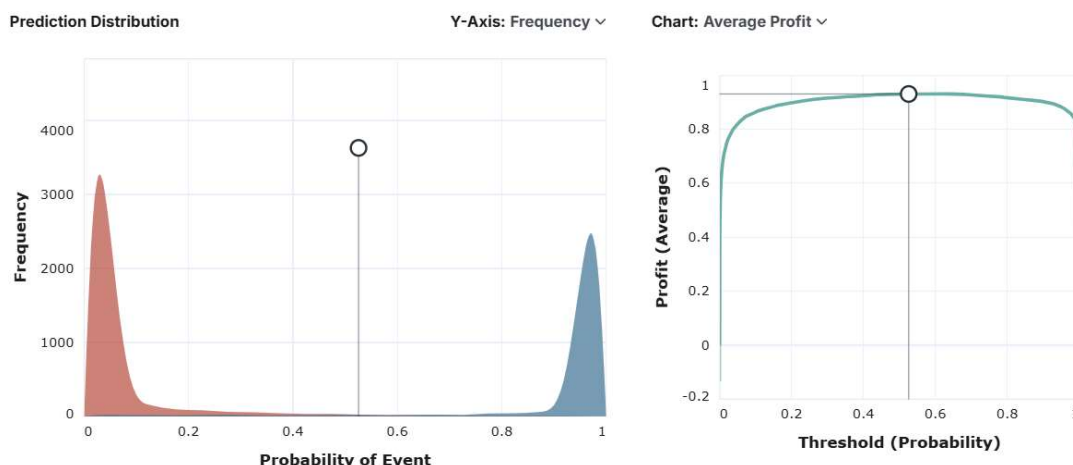
The Average Profit and Prediction Distribution plots illustrate how the model's probability predictions translate into predicted business outcomes.

On the left, the Prediction Distribution shows a clear separation of satisfied and dissatisfied travelers. Most predictions are near 0 (neutral/dissatisfied) and 1 (satisfied), indicating that the model is quite certain in most cases. The vertical line at around 0.5–0.6 is the proposed decision boundary, balancing the two classes nicely.

On the right side, the Average Profit Curve plots the average expected profit for different probability thresholds. The curve peaks at a threshold of approximately 0.6, confirming that this setting maximizes the airline's expected profitability by trading off losses from false

negatives and false positives. Average profit declines slightly after 0.6, supporting the recommendation not to set the threshold too high and miss disgruntled passengers.

Overall, these plots affirm that a threshold of 0.6 provides the optimal balance between predictive accuracy and business profit.



Assumptions Made Due to Uncertainty

There were several assumptions made so as to calculate the profit matrix and recommendations. Customer Lifetime Value (CLV) of an average airline passenger is calculated to be \$1,200, which translates to a \$150 loyalty return for each additional year retained. Average revenue loss from dissatisfied passengers is estimated at \$300 based on industry churn rates. Cost of service recovery in terms of free upgrades or compensation is estimated at \$20 per intervention. External drivers like changes in economic conditions, regulatory changes in the industry, or competitive response were not explicitly modeled but need to be monitored periodically. These assumptions are valid based on available data but need to be reviewed periodically as new operating data emerge.

Final Recommendation on Model Implementation

The Light Gradient Boosted Trees Classifier with Feature Engineering must be used as a decision-support tool and not as a fully autonomous system. The model can be integrated into customer experience improvement programs, where it assists service teams in finding at-risk passengers and adapting recovery efforts. To guarantee model performance, retraining quarterly is advised, in addition to regular monitoring of performance (e.g., drift monitoring against actual customer feedback). Utilized in tandem with expert judgment, the model provides a data-driven, scalable approach

to justifying and constructing customer loyalty, operational efficiency gain, and long-term profitability.

Brief Comparison of Azure AutoML and DataRobot AutoML

Both DataRobot AutoML and Azure AutoML were used to predict airline customer satisfaction using cutting-edge feature engineering and ensemble modeling.

While Azure's XGBoost model showed marginally improved raw performance (AUC 0.9957 vs. DataRobot's 0.91), DataRobot's LightGBM model presented better model governance, integrated drift detection, easier-to-use interpretability features (Lift Charts, Profit Matrices), and production-ready workflow without retraining pipelines.

While Azure's model performance statistics were marginally improved, the DataRobot platform was chosen for production because it balanced predictive performance, explainability, operational readiness, and business usability more equitably.

Its built-in model monitoring, feature influence visualizations, probability threshold adjustment, and profit maximization capabilities made it more suitable for real-world airline decision-making where constant tracking and explainability are critical.

References

- Choi, Y., & Choi, J. (2020). Brand image and customer satisfaction in the airline industry.
- Han, H., & Hyun, S. S. (2021). Service quality impact on airline passengers' loyalty.
- Kim, J. (2021). Role of physical environment in customer satisfaction.
- DataRobot AutoML Documentation.