Machine Learning and Web Design Project

Airplane Flight Satisfaction

By Marta Baker, Sarah Chauvin, Elizabeth Viramontes, and Gabriela Zarate

October 3, 2024

Introduction and Purpose

For this project, we wanted to explore the trends that make airplane flights seem enjoyable to passengers. We used a dataset containing airplane passenger survey responses from 2018 to create a supervised machine learning model that airlines could use to predict customer satisfaction and better the user experiences they offer. We compared this data with a second dataset containing information about individual airlines to create multiple visualizations that show how the manipulation of various features can improve the passenger's perception of satisfaction with their experience.

Research Questions and Design Concepts

Our data prompted us to seek out which features in the survey were the most significant in determining customer satisfaction. We also sought to discover which type of machine learning model might provide the most robust predicting tool. With the addition of visualizations from a second dataset, we wanted to know if there were trends in customer responses to specific airlines, and if the airline itself or certain features were more important in determining satisfaction.

The aesthetic of our website and presentation slide deck reflects common colors seen in many airline logos and plane designs: blue, red, and black. Since our study was not a deep analysis of any specific airline, we included mostly stock-type photos in our website, along with descriptive text for each page. To create the website, our datasets were manipulated in Jupyter notebooks to create the machine learning models, one of which was chosen and saved for import into the JavaScript and HTML files in our app folder. The templates and static folders also include embedded Tableau dashboards with the visualizations we created from both datasets, which are interactive along with the predictor tool. The app.py file can be used to run the website locally via the Flask application. The website is also hosted publicly and can be viewed at this link.

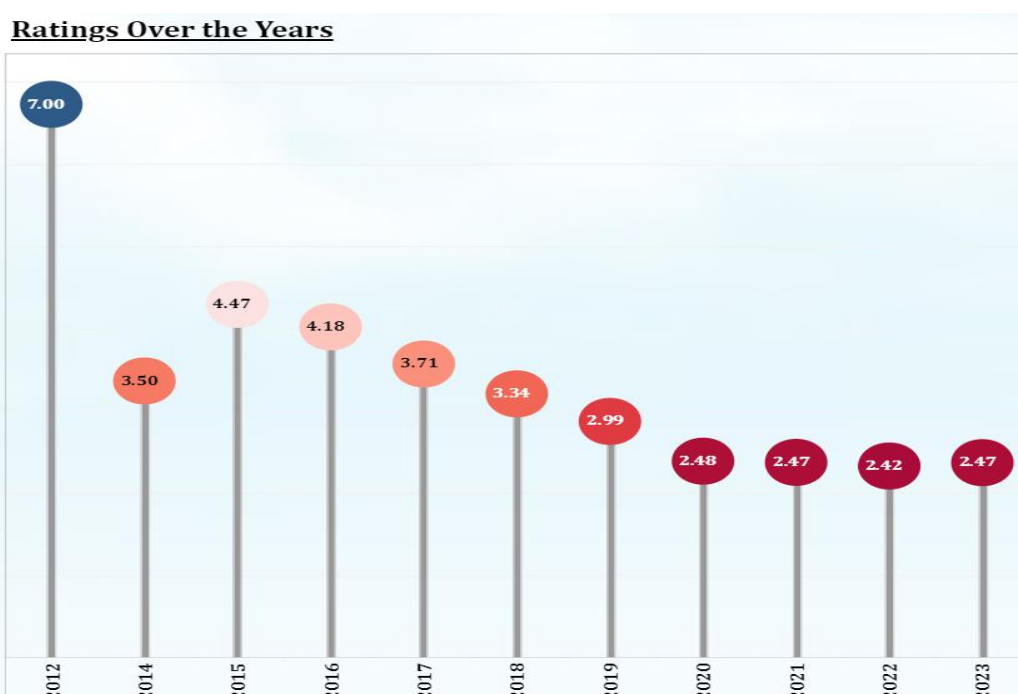
Home page and Visualization Dashboards

The home page includes a short summary of the function of our website and a description of the predictor tool. Both the Airline Satisfaction and Demographic Satisfaction pages show embedded Tableau dashboards, each with multiple visualizations to tell our passenger satisfaction story. The first, Airline Satisfaction, uses the SkyRatings dataset (Pitroda), and the second, Demographics Satisfaction, uses the Passenger Satisfaction dataset (D). Each page contains a brief description explaining how the customer ratings are considered and how to use the filters to obtain specific views.

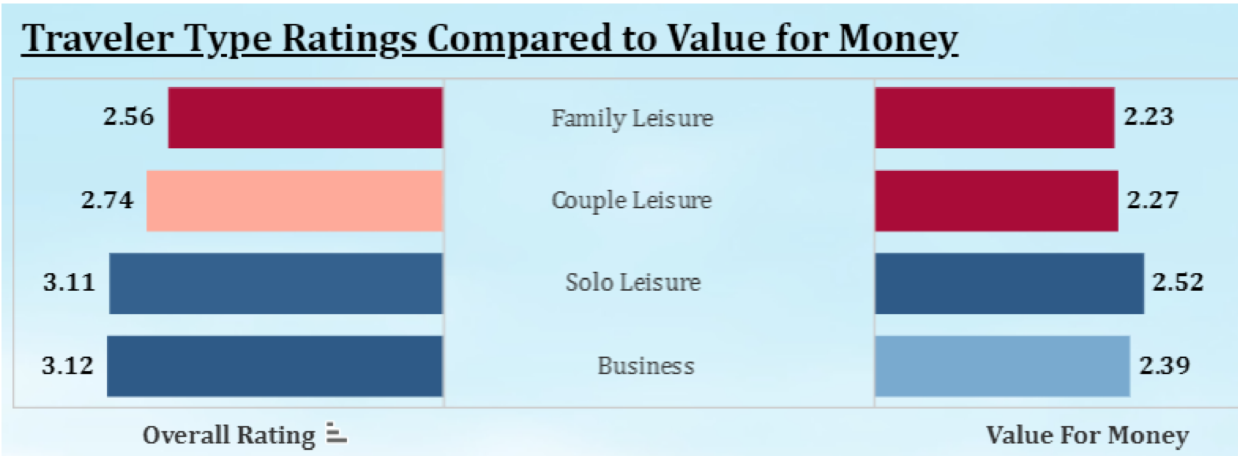
Embedded Tableau pages: Airline Satisfaction

The General Airline Satisfaction Rating dashboard reviews the relationships between passenger ratings of services provided and flight classes and changes in those ratings over time. The second page of this storyboard, Most Popular Airlines in the US, identifies which airlines offered the most highly rated services and whether passengers would recommend them or not. Both pages of the storyboard contain global filters that interact with the graphs, allowing for analysis of specific airline, year, and type of traveler.

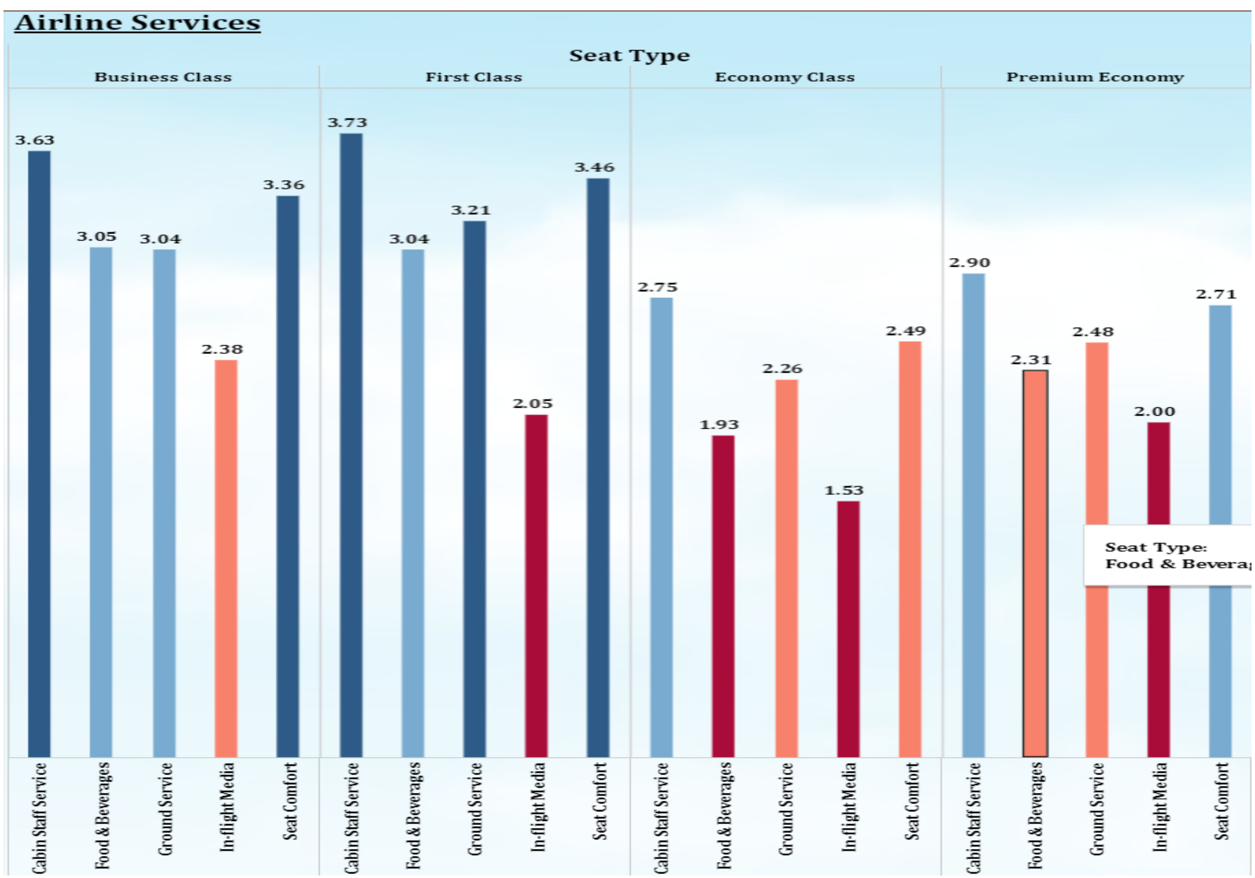
On the General Customer Satisfaction page, the lollipop chart shows the changes in overall customer ratings, ranging from 1 at the lowest to 10 as the highest rating, from 2012 to 2023. It appears that the ratings have worsened over time, likely because the overall number of travelers have increased and, consequently, their expectations of airlines have increased.



In all cases, customers believe that the type of travel is expensive compared to the services each offers. However, upon closer review of the horizontal bar chart, we see that the Solo Leisure and Business traveler types rate the cost significantly lower than the overall service rating, meaning that they expect more from the airline's service.

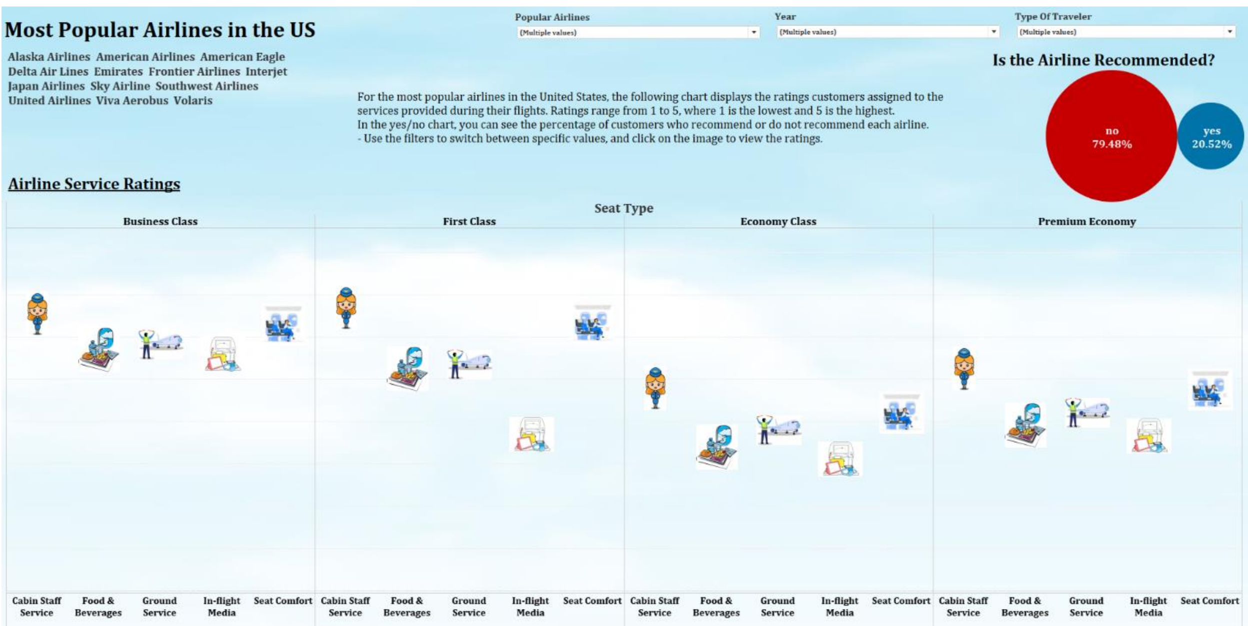


The vertical bar chart describes the overall ratings of specific services more in-depth, organized by seat type. Similarly to the above image, this graph shows how different travel types affect the passenger's satisfaction with their flight.



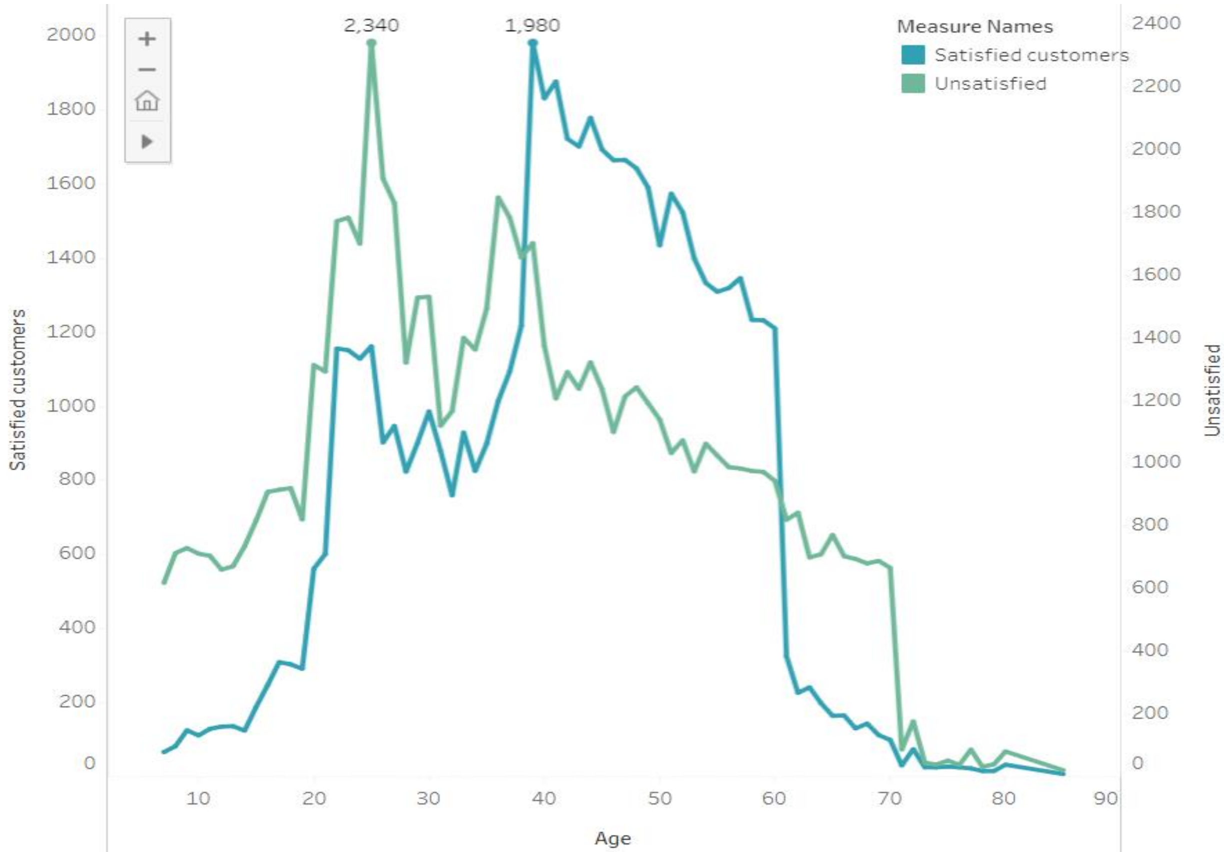
To emphasize the airlines that many people are familiar with, which are prevalent across the internet, in street advertisements, and on television, The Popular Airlines in US page contains a dashboard specifically selecting 13 of the most popular airlines in the United States. This provides an opportunity to see whether customers recommend these airlines.

The description on this page explains how the rating information for this data was obtained, indicating that scores range from 1 at the lowest to 5 at the highest. To make the visualization more intuitive, the points on the graph are depicted by icons of the service they represent. Users can click on each icon to view the average ratings given for that service by passengers with that seat type. This dashboard also features a bubble plot at the top of the page displaying whether the airline is recommended along with the percentage of customers who rated it as such.

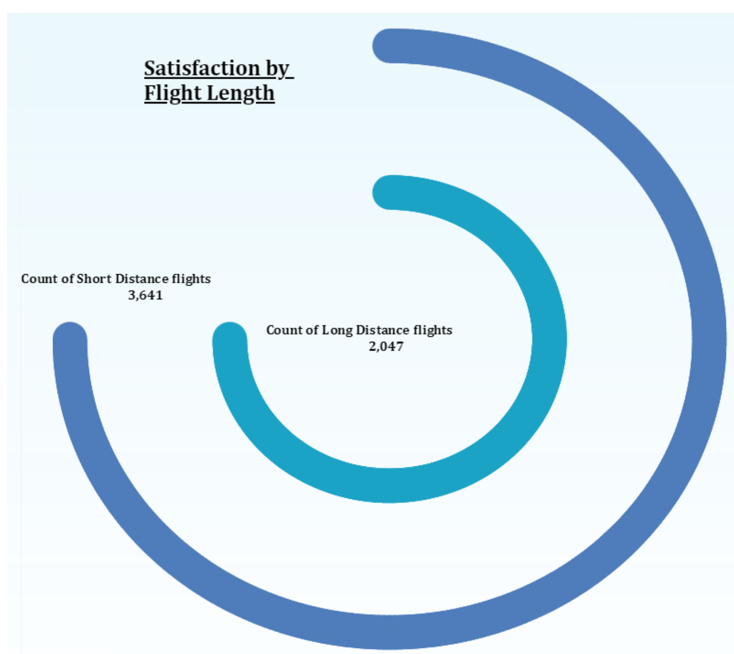


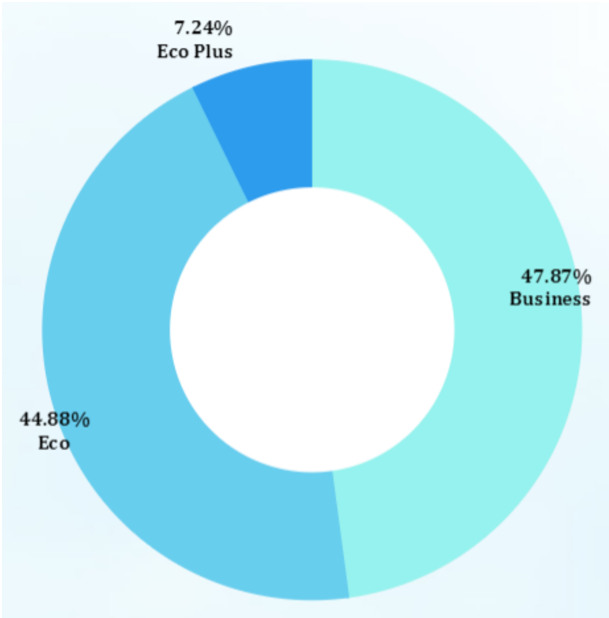
Embedded Tableau pages: Demographic Satisfaction

The analysis of customer satisfaction by demographic dashboard depicts some features from our original dataset (D) that we considered pertinent to our predictive model. Based on age, we noticed that most customer satisfaction was present among people 39 – 60 years old, while most dissatisfaction was present among younger groups, especially 23 – 26 years old. This could be because the people in the younger group are more likely to fill out a survey when they are dissatisfied with the service they received. The highest number of passengers for each rating is noted on the graph below, with 2340 customers age 25 who voted that they were unsatisfied with their service, and 1980 customers age 39 who felt satisfied with their service.



We also analyzed the correlation between flight distance and passenger satisfaction. We classified flight length as being short distance if under 844 miles and long distance if 844 or more miles. The circle graph shows that 3641 passengers were satisfied with short-distance flights and 2047 were satisfied with long-distance flights. This indicates a positive correlation between flight length and customer satisfaction, possibly due to increased services and amenities for long-distance flights, while short distance flying experiences could be more stressful.





The donut chart on this dashboard shows one of the features that we expected to have high impact on passenger satisfaction: travel class. Results are fairly consistent between the business and economy classes, though economy plus had a much lower satisfaction rate at 7.24%. We concluded that economy plus passengers may have higher expectations for their class, and overall satisfaction should not be drastically impacted by travel class if the passenger is traveling economy or business class.

The last visualization on the demographics dashboard shows that the overall satisfaction rate between genders is balanced, with both males and females generating an even distribution of the results. This indicates that there is no correlation between gender and satisfaction among flight passengers.

	Gender	
Satisfaction	Female	Male
neutral or dissatisfied	51%	49%
satisfied	50%	50%

Passenger Satisfaction Prediction Model

This page of the website contains the predictive tool for airlines to use in determining where to focus improvements on the services they offer to their passengers. Each filter offers options to select specific categorical values or numeric ratings for each feature, and the Check Passenger Satisfaction button runs the model and returns both text and a pie chart depicting the predicted passenger satisfaction rate.

The method we used to create and choose the model for our predictive tool involved many steps, which can be seen in more detail in the various Python files in the Notebooks folder of our GitHub repository. For both modeling notebooks, the preprocessing involved dropping the ID column and casting some or all of the ratings features to strings so they could be one-hot encoded. To create a preprocessing pipeline, the features were listed as numeric (age, flight distance, departure delay (minutes), and arrival delay (minutes)), binary (gender, customer type, and type of travel), and categorical (all of the ratings features plus travel class). Five separate preprocessing pipelines were created and tested to determine if label encoding or ordinal encoding the categorical features was better, if the numerical columns should all be scaled or some binned and the rest scaled, and to determine how the ratings features impacted the models.

We compared the Logistic Regression and multiple classifier models, including Random Forest, Extra Trees, K-Nearest Neighbors, Adaptive Boost, Gradient Boosting (GB), Extra Gradient Boosting (XGBoost), and Light Gradient Boosting (LGBM), to determine the best metrics. For accuracy and Receiver Operating Characteristic area under the curve (ROC AUC), the best models for all the datasets including all of the features appeared to be Gradient Boosting, XGBoost, and LGBM. For datasets not including the ratings features, the LGBM for the one-hot encoded categorical features was the best because the feature importances showed a fairly even significance on all of the features. For the Gradient Boosting and XGBoost, one or two features usually had an importance over 10% while the next highest feature was about half as important, and the LGBM placed the highest importance at 8%.

After determining the best model options, we ran grid searches to tune the hyperparameters for optimization of performance. Datasets involved in grid searches included various amounts of one-hot encoding and ordinal encoding, with some or none of the features removed. We performed several rounds of grid searches on GB, XGBoost, and LGBM models to determine the best n_estimators and learning_rate, refitting on f1 score because for our data, we determined that precision and recall were equally important. F1 scores for datasets with full feature inclusion with both one-hot and ordinal encoding were essentially identical; however, the LGBM for one-hot encoding and binning some numerical features was slightly better, so it was used for cross validation.

We performed cross validation on several of the closely-scoring datasets and pipelines to determine which had the best overall metrics. Our ideal model used binning and one-hot encoding for categorical features, a standard scaler for numeric features, ordinal encoding for binary features, and a simple imputer for any missing values upon productionalization. The cross-validation results of our ideal LGBM model are as follows: highest f1 score = 0.9654, mean accuracy = 0.9650, mean precision = 0.9752, mean recall = 0.9435, and mean AUC = 0.9953.

While ideally we would use our LGBM model for the predictive tool in our website, we were limited by the constraints of PythonAnywhere, so we chose the best model we had that the deployment site could support. Our LGBM and XGBoost models returned the best metrics, followed by GB, and since our top two choices are unsupported at this time, we set up our predictive tool with our GB model having the same dataset and preprocessing as our top LGBM and XGBoost models. The metrics produced by cross validation on this model are as follows: highest f1 score = 0.9594, mean accuracy = 0.9589, mean precision = 0.9673, mean recall = 0.9372, and mean AUC = 0.9933. While these metrics are very similar to the set mentioned above, they do indicate that the GB model will not perform quite as well as the LGBM. We can also observe this conclusion in the imbalance in the model's top feature importances, as seen in the image to the right.

	Feature	Importance
6	inflight_wifi_service	0.343154
1	customer_type	0.243293
17	checkin_service	0.150171
18	inflight_service	0.110571
8	ease_of_online_booking	0.046481
16	baggage_handling	0.030555
10	food_and_drink	0.020350
12	seat_comfort	0.019795
14	on_board_service	0.009166
7	departure_arrival_time_convenient	0.005958

Report, About Us, and Resources pages

The last three pages of our website include this report, so users can fully understand the context, and the contributors to our project, including each of our team members and our data sources. The About Us page also includes links to each team member's GitHub and LinkedIn pages, where users are able to contact us and learn more about our work.

Conclusions, Limitations, and Future Work

The visualizations created in our Tableau dashboard effectively illustrate the evolving landscape of customer satisfaction in the airline industry. By analyzing passenger ratings from 2012 to 2023, we can see a clear trend indicating a decline in overall satisfaction, likely influenced by increasing traveler expectations. The interactive dashboards allow viewers to explore specific airlines and their services, providing valuable insights into customer perceptions. Importantly, the focus on well-known airlines highlights the broader trends in customer sentiment, while also shedding light on smaller, lesser-known carriers.

The use of filters and engaging visual elements on the dashboard pages of our website enhances the user experience, enabling a deeper understanding of how different traveler types perceive value and quality. Additionally, using a dataset for creating our machine learning model that doesn't list specific airlines keeps our predictive tool from becoming biased against smaller airlines. Overall, these visualizations not only tell a compelling story about airline customer satisfaction but also serve as a useful tool for both consumers and industry stakeholders seeking to improve service quality and meet evolving customer demands.

Unfortunately, we discovered several limitations with our data. While we appreciated the lack of bias in the exclusion of specific airlines in our original dataset (D), the fact that the data did not include airport locations could mean that data on boarding and ticket purchase experience will vary for singular airlines. We do not know which countries or airports any of these flights are departing from or arriving to, and there could be some bias in those ratings based on individual airport function and layout. For the dataset that did include specific airlines (Pitroda), the maximum number of customers surveyed per airline was 100, and we do not know how these customers were chosen.

The minimum age surveyed in our data was seven years old, so some of the responses may not be reliable. Since we discovered that more young people appeared to be unsatisfied with their flights, we concluded that young people are more likely to fill out surveys when they are dissatisfied, again potentially skewing our data. We also do not have a source for our original dataset, so we cannot verify the quality of the survey process.

The dataset we used to create our machine learning model (D), did not include the dates for the survey responses, so we do not know if any of the data would present differently before or after the covid pandemic. While experimenting with our models, we discovered some imbalance in the feature importances, the top 15 of which are demonstrated in the image below, even when the models were producing desirable metrics. This is likely due to some skewing in the data. With loyalty, the data included five times as many loyal customers as disloyal customers, so it is possible that there could have been an incentive for loyalty program customers to fill out positive surveys.

Gradient Boosting

	Feature	Importance
12	seat_comfort	0.152148
6	inflight_wifi_service	0.115398
7	departure_arrival_time_convenient	0.067965
11	online_boarding	0.056960
5	flight_distance	0.056910
9	gate_location	0.003973
0	gender	0.002995
3	type_of_travel	0.002172
10	food_and_drink	0.000870
8	ease_of_online_booking	0.000803
18	inflight_service	0.000246
1	customer_type	0.000142
4	travel_class	0.000000
2	age	0.000000
13	inflight_entertainment	0.000000

LightGBM

	Feature	Importance
0	gender	0.086667
5	flight_distance	0.061667
6	inflight_wifi_service	0.054333
1	customer_type	0.053667
11	online_boarding	0.053000
12	seat_comfort	0.041000
7	departure_arrival_time_convenient	0.037000
3	type_of_travel	0.029333
8	ease_of_online_booking	0.016000
2	age	0.015333
10	food_and_drink	0.013000
9	gate_location	0.010333
19	cleanliness	0.009000
13	inflight_entertainment	0.004000
18	inflight_service	0.004000

XGBoost

	Feature	Importance
12	seat_comfort	0.117386
7	departure_arrival_time_convenient	0.063524
11	online_boarding	0.043181
6	inflight_wifi_service	0.026692
5	flight_distance	0.021610
9	gate_location	0.004885
8	ease_of_online_booking	0.004018
10	food_and_drink	0.003550
0	gender	0.002668
18	inflight_service	0.002201
16	baggage_handling	0.001891
14	on_board_service	0.001858
15	leg_room_service	0.001811
3	type_of_travel	0.001774
13	inflight_entertainment	0.001302

We also saw responses from twice as many business travelers compared with personal travelers, and our model does seem to predict higher satisfaction when the filters for business travel or business class are selected. We concluded that passengers traveling for business are more likely to be satisfied with their experience when their flights are subsidized by their employers. Business travelers are also more likely to be traveling alone or with other adults, which could lead them to perceive a more positive experience than a personal traveler who might be stressed from traveling with children or elderly companions. We also concluded that business travelers are more likely to fill out a survey after their flight, which could be leading to some customer bias in the way our model performs.

Lastly, as mentioned previously, we ran many different model types with multiple preprocessing pipelines to choose the most optimized one, but we were unable to use our top performer for the model on this website because it is not currently supported by PythonAnywhere. If we deployed our website elsewhere, we might update the model to our ideal LGBM version, which may perform more reliably. For additional expansion of our data analysis, more survey results including dates, airport locations, and specific airlines could provide a more robust machine learning model.

Resources and Works Cited

D, John. “Passenger Satisfaction,” Kaggle datasets. Updated 2018. https://www.kaggle.com/datasets/johndddddd/customer-satisfaction

Klein, TJ. “Airline Passenger Satisfaction,” Kaggle datasets. Updated 2019. https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction

Nadya, Gissella. “Predicting Airlines Passengers Satisfaction using Logistic Regression and K-Nearest Neighbor,” website. Updated 2021. https://rpubs.com/gissellanadya/LBB-Airlines

Pitroda, Khushi. “SkyRatings: Unleashing 23K+ Airline Reviews!” Kaggle datasets. Updated 2023. https://www.kaggle.com/datasets/khushipitroda/airline-reviews

Slave et al. “Flight Satisfaction Prediction,” Kaggle datasets. 13 September, 2024. https://www.kaggle.com/code/bejondbirthday/flight-satisfaction-prediction

Project deployment: “Airlines Passenger Satisfaction” web application. https://sarahchauvin.pythonanywhere.com/