Can an Airline Company predict customer satisfaction?

Robert Pearson, rpearson2@bellarmine.edu

ABSTRACT

This project aims to study the relationship between a customer's satisfaction with their flight and the variables that could influence these results. By using logistic regression, the project seeks to provide an accurate model that airline companies could use to predict if a customer will be satisfied or dissatisfied with the service they provide. It will use a wide range of variables that the consumer has input to create models using all or a number of these variables. From this, airlines can use the model to gain valuable insight into what drives customer satisfaction and adjust their service accordingly.

I. INTRODUCTION

The dataset that I am working with is called *Customer Satisfaction in Airline*. It is from a survey taken from an undisclosed Airline company that focuses on measuring airline passengers' satisfaction through a wide variety of variables that the customer either ranked or responded to. The classification target for this project is customer satisfaction, predicting if the customer is satisfied with or not with the service they received from the airline.

II. BACKGROUND

Measuring customer satisfaction is a key component of any business as it can contribute to the effort of improving service quality. The airline could have collected the data originally for several reasons, but the main reason is likely due to wanting to them wanting to improve the customer's experience. In an industry, like the airline industry, customer satisfaction plays a big role in customer retention. Airlines want to make customers into lifetime customers and that happens from customers having a good experience on their flight. Questions the author could be trying to answer is if it's possible to predict passenger satisfaction and what factors influence a satisfied or dissatisfied passenger.

III. EXPLORATORY ANALYSIS

The dataset contains 129,880 samples of an airline's customer ratings. There are 22 columns of various data types. **Table 1** lists each variable in the dataset along with the data type (nominal, ordinal, continuous, or discrete). The dependent variable in this study is *satisfaction*, this variable's output is either "satisfied" or "dissatisfied." There were some missing values for the variable "*Arrival Delay in Minutes*", however this should not be significant as it was only missing 393 values out of 129,880. **Figure 1** shows the distribution of satisfied and dissatisfied customers and from that it shows that they are evenly distributed.

Table 1: Data Types

Variable Name	*Data Type*
Satisfaction	Nominal – Categorical
Customer Type	Nominal – Categorical
Age	Continuous – Numerical
Type of Travel	Nominal – Categorical
Class	Nominal – Categorical
Flight Distance	Continuous – Numerical
Seat Comfort	Ordinal – Numerical
Departure/Arrival Time Convenient	Ordinal – Numerical
Food and drink	Ordinal – Numerical
Gate location	Ordinal – Numerical
Inflight wifi service	Ordinal – Numerical
Inflight entertainment	Ordinal – Numerical
Online support	Ordinal – Numerical

Ease of Online booking	Ordinal – Numerical
On-board service	Ordinal - Numerical
Leg room service	Ordinal – Numerical
Baggage handling	Ordinal – Numerical
Checking service	Ordinal – Numerical
Cleanliness	Ordinal – Numerical
Online boarding	Ordinal – Numerical
Departure Dealy in Minutes	Continuous – Numerical
Arrival Delay in Minutes	Continuous – Numerical

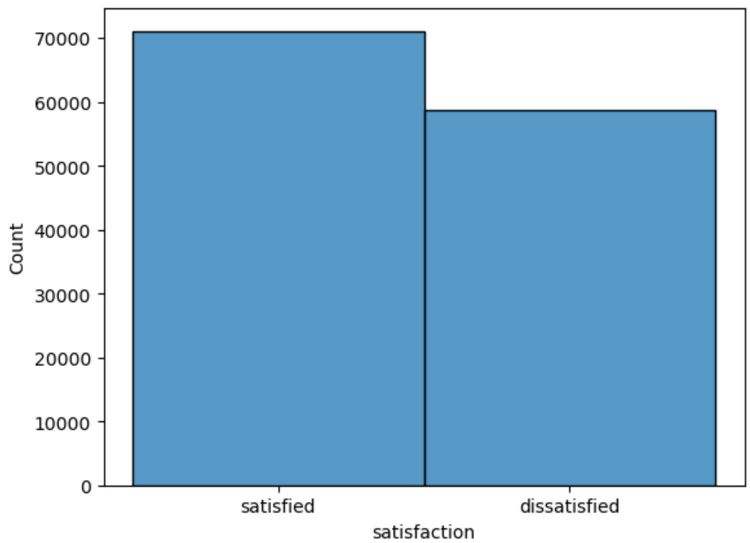


Figure 1: Histogram of the distribution of satisfied or dissatisfied customers

IV. METHODS

A. Data Preparation

Overall, there was not much work that went into preparing the data for the logistic regression model. As stated previously, the only missing values were in "*Arrival Delay in Minutes*", the rest of data did not require any cleaning procedures in preparing for analysis. The data for "*Arrival Delay in Minutes*" was well distributed and is a continuous variable. Because of this, I was able to compute the missing with the mean to fill in the missing data before proceeding with model. I did not need to normalize any data or drop/create any other variables. The categorical data within the dataset required encoding to be used in the model. This was done with one-hot encoding and was applied to all the categorical variables because the data within these variables was not ordinal and is instead nominal.

B. Experimental Design

The selected variables for experiment 3 and 4 are "*Seat comfort*", "*Departure/Arrival time convenient*", "*Food and drink*", "*Gate location*", "*Inflight wifi service*", "*Inflight entertainment*", "*Online support*", "*Ease of Online booking*", "*On-board service*", "*Leg room service*", "*Baggage handling*", "*Check-in service*", "*Cleanliness*", and "*Online boarding*". These variables were selected due to them all being ordinal and based on consumer ranking of a service provided by the airline.

Table 2: Experiment Parameters

Experiment Number	**Parameters**
1	All twenty-one (21) variables with 80/20 split for train, and test

2	All twenty-one (21) variables with 70/30 split for train, and test
3	Selected variables with 80/20 split for train, and test
4	Selected variables with 70/30 split for train, and test

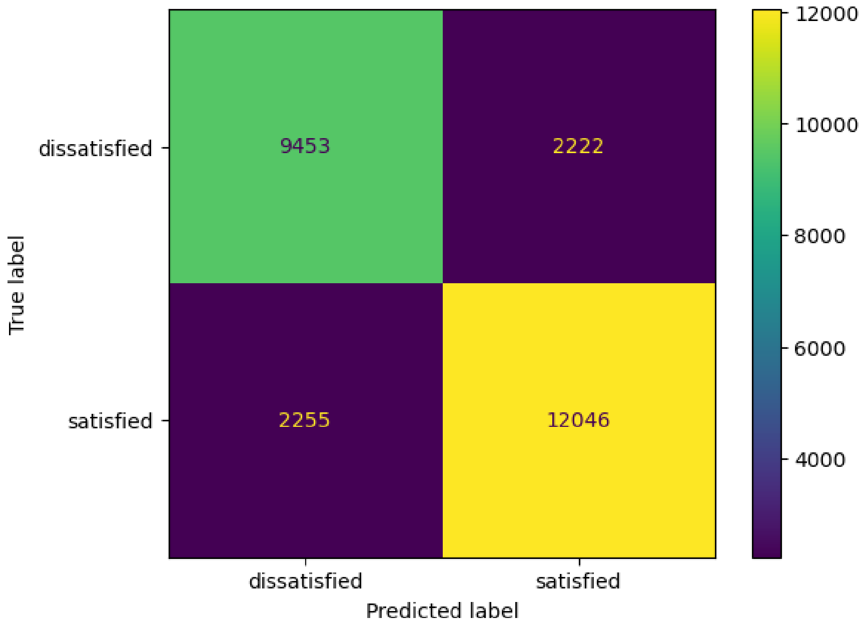
C. Tools Used

The following tools were used for this analysis: Python v3.11.4 running the Anaconda 1.10.0 environment for Windows 11 computer was used for all analysis and implementation. In addition to base Python, the following libraries were also used: Pandas 1.5.3, Numpy 1.24.3, Matplotlib 3.7.1, Seaborn 0.12.2, and SKLearn 1.3.0.

These tools were for their efficiency and versatility in data analysis and machine learning. Python serves as the foundation for the Anaconda environment simplifying the process with pre-installed libraries. Pandas and NumPy provide data manipulation, Matplotlib and Seaborn aid in the visualizations, and scikit-learn offers a machine learning toolkit.

V. RESULTS

A. Classification Measures/ Accuracy measure

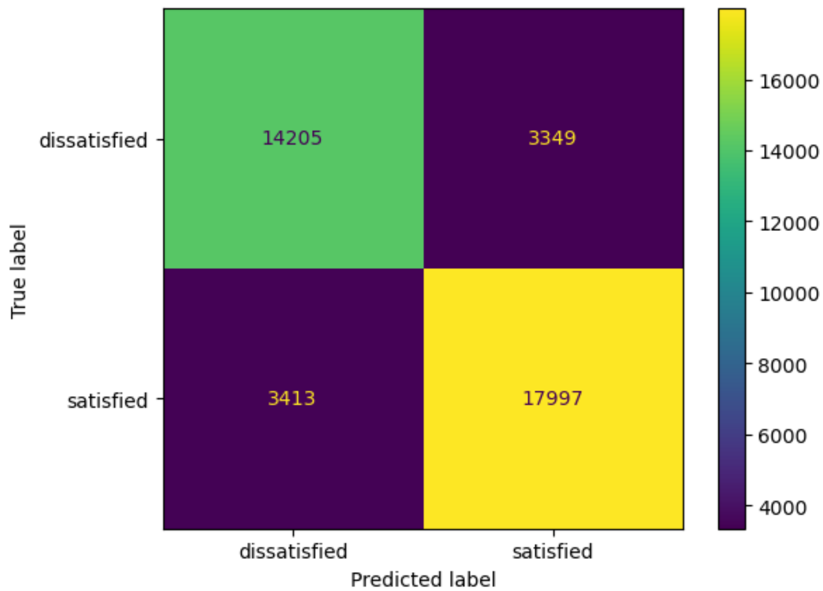


a)

	precision	recall	f1-score	support
dissatisfied	0.81	0.81	0.81	11675
satisfied	0.84	0.84	0.84	14301
accuracy			0.83	25976
macro avg	0.83	0.83	0.83	25976
weighted avg	0.83	0.83	0.83	25976

b)

Figure 2: All 21 variables with 80/20 split for train and test a) Confusion Matrix b) Classification Report

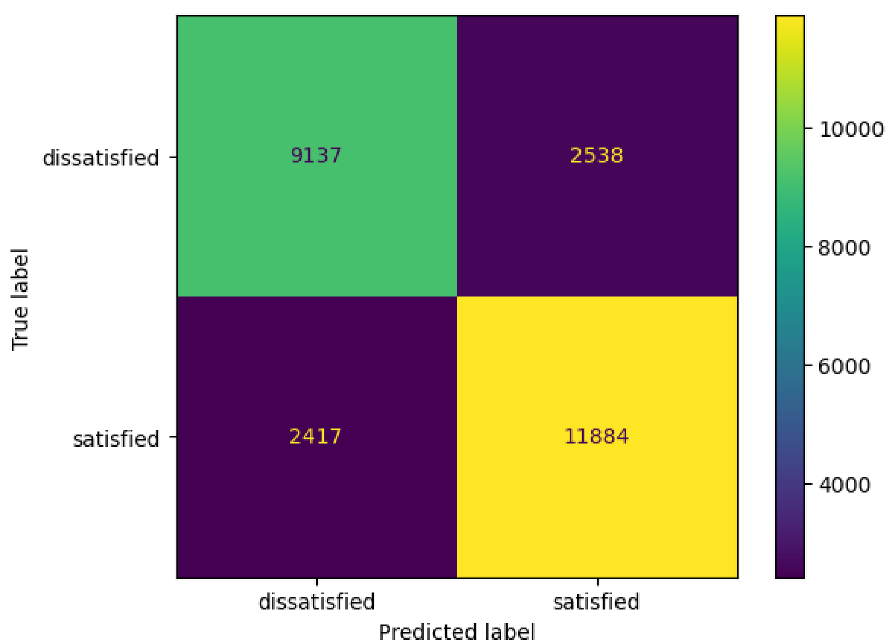


a)

	precision	recall	f1-score	support
dissatisfied	0.81	0.81	0.81	17554
satisfied	0.84	0.84	0.84	21410
accuracy			0.83	38964
macro avg	0.82	0.82	0.82	38964
weighted avg	0.83	0.83	0.83	38964

b)

Figure 3: All 21 variables with 70/30 split for train and test a) Confusion Matrix b) Classification Report

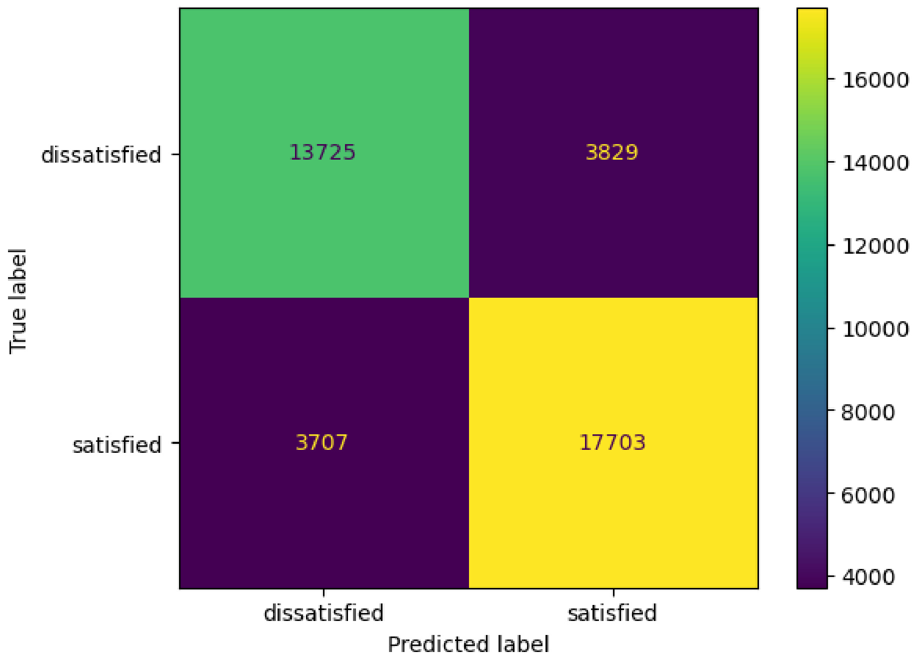


a)

	precision	recall	f1-score	support
dissatisfied	0.79	0.78	0.79	11675
satisfied	0.82	0.83	0.83	14301
accuracy			0.81	25976
macro avg	0.81	0.81	0.81	25976
weighted avg	0.81	0.81	0.81	25976

b)

Figure 4: Selected variables with 80/20 split for train and test a) Confusion Matrix b) Classification Report



a)

	precision	recall	f1-score	support
dissatisfied	0.79	0.78	0.78	17554
satisfied	0.82	0.83	0.82	21410
accuracy			0.81	38964
macro avg	0.80	0.80	0.80	38964
weighted avg	0.81	0.81	0.81	38964

b)

Figure 5: Selected variables with 70/30 split for train and test a) Confusion Matrix b) Classification Report

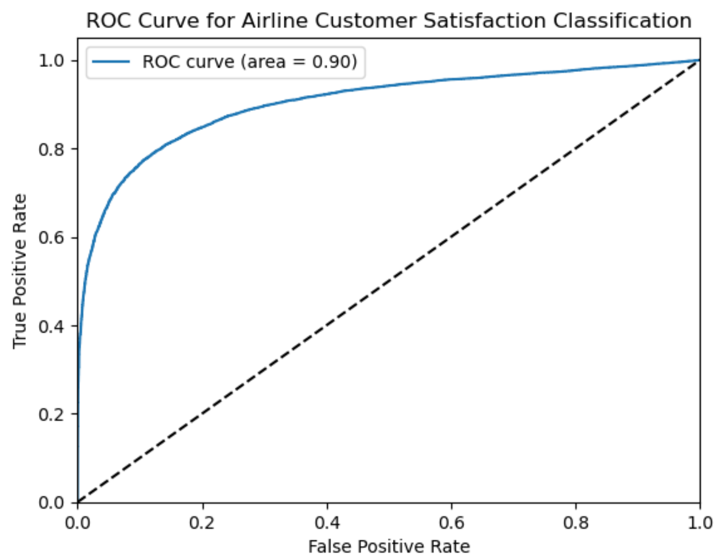


Figure 6: ROC Curve and AUC for All twenty-one (21) variables with 80/20 split for train, and test

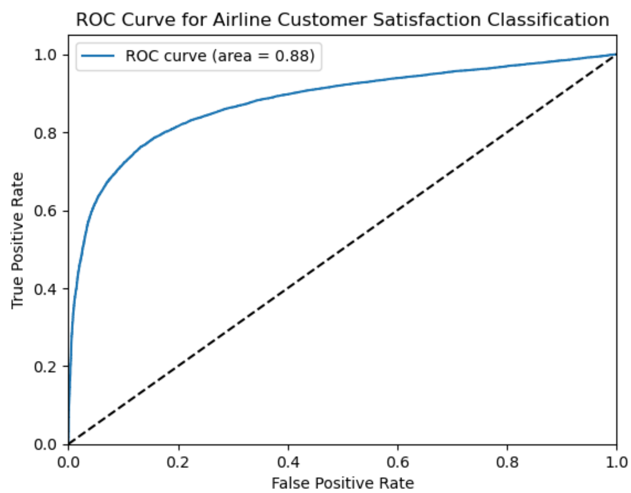


Figure 7: ROC Curve and AUC for All selected (14) variables with 80/20 split for train, and test

Split	Model	True Positives (TP)	False Positives (FP)	True Negatives (TN)	False Negatives (FN)
80/20	21 features	12046	2222	9453	2255
70/30	21 features	17997	3349	14205	3414
80/20	Selected	11884	2538	9137	2417
70/30	Selected	11703	3829	13725	3707

a)

True Positives (TP): These are instances where the model correctly predicts that a customer is satisfied, and in reality, they are satisfied.
False Positives (FP): These are instances where the model incorrectly predicts that a customer is satisfied, but in reality, they are dissatisfied.
True Negatives (TN): These are instances where the model correctly predicts that a customer is dissatisfied, and in reality, they are dissatisfied.
False Negatives (FN): These are instances where the model incorrectly predicts that a customer is dissatisfied, but in reality, they are satisfied.

b)

Table 3: a) Model Prediction Results b) Reference for Confusion Table

B. Discussion of Results

Based on the results from the classification tables, the models built with all twenty-one features consistently outperformed the models with the selected features by a relatively small margin. This indicates that the additional seven features—"Customer Type", "Age", "Type of Travel", "Class", "Flight Distance", "Departure Delay in Minutes", and "Arrival Delay in Minutes"—contribute positively to the model's performance.

Among the models with all features, both the 80/20 and 70/30 splits yielded similar performance. This suggests that the model's performance is strong across the different train/test sizes. A high accuracy score of 83% for both test and train splits indicate the proportion of correctly classified instances. Precision, recall, and F1-score for both satisfied and dissatisfied customers had values between 0.81 to 0.84. The ROC curves for both models had an AUC score of 0.9. This shows that the model can achieve a high true positive rate while keeping the false positive rate relatively low. Therefore, the models with all features provide a good balance between correctly recognizing satisfied and dissatisfied customers while keeping a high level of accuracy.

It is important to note that the AUC for the models with selected variables was 0.88. This shows that those models can also accurately classify instances. This suggests that all models are effective at distinguishing between positive and negative cases.

Looking at Table 3, the model with twenty-one features and a 70/30 split for train/test has the highest true positives, showing that the model correctly identifies the most satisfied customers. While the model with twenty-one features and an 80/20 split for train/test has the lowest false negatives, showing that it incorrectly predicts the fewest dissatisfied customers as satisfied. Considering these factors, if we were to prioritize correctly identifying satisfied customers (true positives), the model with a 70/30 split is better. But, if we prioritize minimizing the number of dissatisfied customers that are incorrectly reported as satisfied (false negatives), the model with an 80/20 split is better.

Among the other models, the one with all fourteen features and an 80/20 split for train/test appears to be the worst performer. This model had a decently high false positive count of 2,538 and false negative count of 2,417. It misclassified satisfied customers as dissatisfied and vice versa more often compared to the other models. Additionally, it had a lower accuracy score of .81, as well as lower precision, recall, and F1-score for both satisfied and dissatisfied customers, with values around 0.78 to 0.83 compared to the models with all features.

C. Problems Encountered

Problems I encountered during this process included obtaining the data and evaluating the model. The first problem I had was with obtaining the data was picking what topic I wanted to make a regression model on. This was challenging because I had multiple topics that sounded interesting before ultimately deciding on doing one with airline customer satisfaction. The second problem was with evaluating the model and creating a ROC curve (receiver operating characteristic curve) for the model that best fit based off the classification report. We had not done this in class, so it took a while for me to understand the basic concept of how it worked and what information it displayed.

D. Limitations of Implementation

Some limitations with my model could include the number of variables being imputed. If important variables are excluded or irrelevant ones included, it could through off the whole prediction. It could also lead to overfitting, where the model focuses more on random stuff instead of recognizing patterns.

E. Improvements/Future Work

For improving my model in the future, I would want to run more experiments with different feature selections, or sampling techniques. Having 21 different variables to predict if a customer will be satisfied or not with their flight, some of these might play a bigger role in improving customer satisfaction. I would like to do a more in-depth analysis to find which ones impact it the most. It would also be more useful for the airlines to know what variables play the biggest role. Another thing I would like to look into for future work is looking at the possibility of using different models, if that's even possible.

VI. CONCLUSION

In conclusion, the analysis of airline customer satisfaction reveals that the models with all twenty-one features only slightly outperformed those with fourteen features; indicating that the additional variables such as customer type, age, and flight details may play some importance. Both 80/20 and 70/30 splits exhibited strong performance with high accuracy, precision, recall, and F1-scores. The ROC curves had AUC scores of 0.9 and 0.88, displaying each model's ability to distinguish between satisfied and dissatisfied customers. While the models with twenty-one variables, 70/30 split model excelled in identifying satisfied customers, the 80/20 split model minimized misclassification of dissatisfied customers. However, the model with all fourteen features and an 80/20 split was the least favorable performer due to higher false positive and false negative counts. Future work on this project could focus on refining the feature selection and exploring alternative modeling techniques.

REFERENCES

https://www.kaggle.com/datasets/yakhyojon/customer-satisfaction-in-airline