Using Metrics to Predict Airline Customer Satisfaction

Joseph Seiba [13] Chanse Bhakta [12] Minh Vu [1]

[1]Department of Computer Science [2]Department of Management Science and Engineering [3]Department of Electrical Engineering

Introduction

For our project, we aimed to answer the question: "What factors contribute to passenger's satisfaction when flying?"

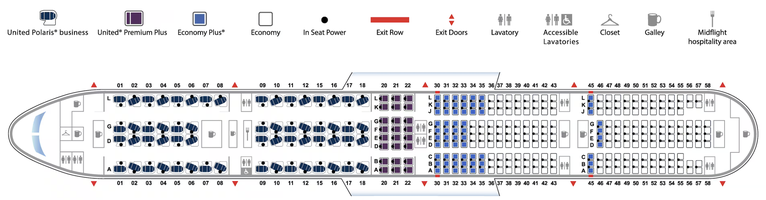

Figure 1. United Airlines Boeing 777 Seat Map.

Air travel is a staple in the transportation industry and customer satisfaction plays a monumental role in an airline's reputation, loyalty, and ultimately, success. Factors impacting customer satisfaction may be linked to aircraft type, location of seat, frequency of delays, etc. Aside from the previously listed customer-dependent factors, airline-dependent factors strongly correlate with customer satisfaction.

Dataset and Features

Our project utilized the "Airline Passenger Satisfaction" dataset from Kaggle, consisting of approximately 130,000 entries from real U.S. airline customer satisfaction surveys. This dataset covers around 24 parameters, providing a comprehensive overview of demographic information (such as gender, age, customer type, and class) and detailed customer experiences (including inflight wifi, seat comfort, and cleanliness). This allowed us to analyze how demographic factors and service details influence satisfaction. For preprocessing, we removed the customer ID's and any entries with missing data-points, and converted the final decision column to a binary format, indicating customers as either "satisfied" or "neutral or dissatisfied."

Results

Customer loyalty is one of the most highly correlated aspects to customer satisfaction. Customers, across industries, will provide higher satisfaction rates to the companies which they are loyal to, and will be more likely to scrutinize companies they have not done business with.

Type of Travel also is correlated to customer satisfaction. This makes sense because often, flyers travel for business. If the business pays for a ticket, then the flyer might be indifferent towards their experience because they have have no financial investment towards their experience.

Discussion/Future Work

The performance of both datasets on this model was generally great. However, given that the results are so high, it could indicate an overfit model. In the future, we would like to introduce more noise in the dataset to potentially alleviate these concerns. Additionally, investigating other machine learning algorithms and comparing their performance could provide insights into the best approaches for this specific classification task. Improving feature engineering and selection processes could also lead to better model performance.

References

[1] S. Chowdhury & M. Schoen, "Research Paper Classification using Supervised Machine Learning Techniques," in *Proc. IETC*, 2020, pp. 1-6, doi: 10.1109/IETC47856.2020.9249211.

[2] P. Rotella and S. Chulani, "Analysis of customer satisfaction survey data," in *Proc. MSR*, IEEE, 2012.

[3] Y.-H. Hsieh, C. J. Lin, & J. C. Chen, "Customer Satisfaction Measurement with Neural Network," 2007, pp. 47-52.

[4] C. Lawson & D. C. Montgomery, "Logistic regression analysis of customer satisfaction data," *Qual. Reliab. Eng. Int.*, vol. 22, no. 8, pp. 971-984, 2006.

[5] Dataset: Airline Passenger Satisfaction, Available at: `https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction`

[6] Code: numpy, sklearn, matplotlib. See code readme for details.

Model 1: Logistic Regression Model

The logistic regression algorithm is known for being an efficient and robust algorithm for binary classification tasks. Due to these reasons, and primarily because of this efficiency, we wanted to employ this model to see if we can achieve a high accuracy score.

1. **Sigmoid Function:** We applied a sigmoid function to which we employed a threshold of 0.5. In other words, of the probability of a customer being satisfied was 0.5 or greater, we predicted it as such.
2. **Interpretability:** Logistic regression offers interpretability that is often lacking in more complex models. We are able to extract (normalized) parameter values for the model. This allows us to gauge which parameters contribute to a negative or positive experience when flying.
3. **Implementation:** We employed sklearn's logistic regression model

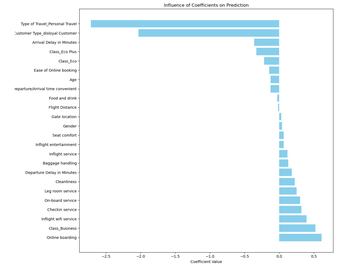

Figure 2. Logisitic Regression Coefficients

Model Configuration

Hyperparameter - Max Iterations: Set to 1000 to ensure thorough optimization, particularly vital for complex datasets where default iterations might not suffice for optimal performance.

Logistic Regression Model Performance

- **Accuracy:** Achieved 87.16%, indicating a high level of correct predictions.
- **Precision:** Approximately 86.92%, showcasing the model's ability to accurately identify 'satisfied' customers.
- **Recall:** 83.28%, reflects the model's strength in minimizing missed 'satisfied' classifications.
- **F1 Score:** 85.06%, demonstrates a balanced performance between precision and recall.

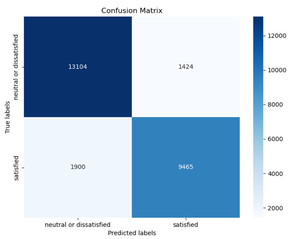

Figure 3. Logistic Regression Confusion Matrix

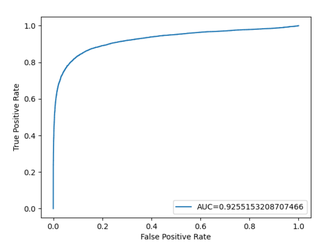

Figure 4. Logistic Regression ROC Curve

Model 2: Neural Network

Neural networks are able to cope and better represent a dataset's complex relationships. As such, we looked at our previous usage examples in class. We leveraged the coursera neural networks module. As such, we used a tri-layer neural network model, each with sigmoid activation.
The binary cross entropy loss function is an ideal choice binary classification task because it effectively measures the discrepancy between actual binary outcomes and the probabilities predicted by the model, thereby facilitating the enhancement of classification accuracy directly.
We also chose the Adam optimizer for its effectiveness in managing sparse gradients and its flexible nature, both of which contribute to a marked decrease in the convergence time to reach the optimal solution.

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{1}$$

$$L(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)] \tag{2}$$

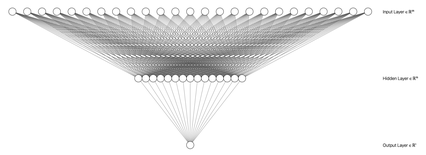

Figure 5. Neural Network Layers

Model Configuration

Hyperparameter - Learning rate: Set to 0.0001 to mitigate overfitting concerns, and overshoot the optimal parameters.

Neural Network Model Performance

- **Accuracy:** Achieved 92.33%, indicating a high level of correct predictions.
- **Precision:** Approximately 93.13%, showcasing the model's ability to accurately identify 'satisfied' customers.
- **Recall:** 88.99%, reflects the model's strength in minimizing missed 'satisfied' classifications.
- **F1 Score:** 91.01%, demonstrates a balanced performance between precision and recall.

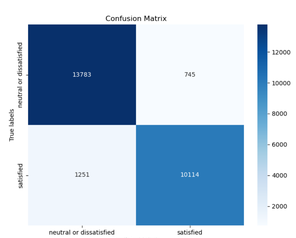

Figure 6. Neural Network Confusion Matrix

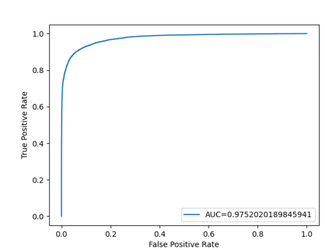

Figure 7. Neural Network ROC Curve