BA305 Airline Passenger Satisfaction



**Analyzing Airline Passenger Satisfaction Levels to Inform Investment Strategies in the Airlines Industry**

Team A9: Qianqi Dong, Shimon Dasgupta, Ramsey Shaheen, Dima Kazlouski, Sarah Broderick

**Introduction**

**1.1 Background and the Problem**

Similar to all service-oriented industries, airline’s success is largely attributed to their customer service. When you take a flight, what are the factors that make your experience satisfactory and keep you coming back to your airline? If we can identify the key factors that contribute to customer dissatisfaction, airlines can focus on optimizing these areas, reducing unnecessary costs and increasing overall customer satisfaction.   
  
We are going to use three different models to analyze different features impacts on a customers overall satisfaction; KNN, Decision Trees, and Logistic Regression. KNN will help us understand the impact of similar review profiles, Decision Trees will allow us to visually and hierarchically explore satisfaction determinants, and Logistic Regression will enable us to quantify the influence of different service features on overall satisfaction.   
  
We want to develop a predictive model that can not only help airlines identify the most crucial areas of flight services that need improvements, but also gives overall insights into their customers' expectations. Additionally, by understanding these dynamics, consumers can make more informed choices, and airlines can tailor their services to meet customer expectations more effectively.

**1.2 Selecting a Dataset**

When looking for a dataset for our user story, we wanted to make sure it included qualitative and quantitative data regarding customer demographics as well as flight details and service specifics. We found a lot of data sets that had general flight details but wanted actual quantitative customer data on services and flight details. We eventually found a dataset that had good feature data and our predicted variable.

**1.3 The Air Travel Dataset**

Data Description: The dataset includes a comprehensive collection of customer reviews on flight services, aimed at quantifying various aspects of the travel experience that influence overall satisfaction. It contains a wide array of variables, from demographic information about travelers to detailed evaluations of different service attributes.

Population/Sampling: Our population is people traveling on commercial flights in the U.S. We are using 20% of the whole dataset as our sample which consists of 25976 flights.

Data Collection: The dataset was collected from kaggle and then preprocessed by our team.  
  
Variables: The original dataset includes 23 variables, covering demographic data, flight details, service ratings, and textual reviews. This set of variables allows for analysis into different areas of what affects overall passenger satisfaction.  
[**https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction**](https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction)

**2 Data Preprocessing**

**2.1 Cleaning Data Entries and Removing Unnecessary Features**

We first checked for any rows with missing or NAN values, and found that there were 83 records missing Arrival Delay in Minutes. We removed these 83 records and kept the 25893 non duplicated records with no missing values. Then we moved on to dropping unnecessary columns from the dataset. We dropped the ID which is irrelevant to our analysis. Since there is a high correlation between delay in departure and arrival, we also decided to drop the Arrival\_Delay\_in\_Minutes to keep the data concise.

**2.2 Reformatting and Scaling Our Data**

We converted the values of certain categorical features into dummies to enhance the interpretability of the data. The original dataset comprises two classes: 'satisfied' and 'neutral or unsatisfied'. For more efficient model processing, we assigned the code 0 to 'satisfied' and 1 to 'neutral or unsatisfied'. Additionally, we assigned numeric codes to various categorical variables as follows: For the 'Class' variable, 'Business' is coded as 2, 'Eco Plus' as 1, and 'Eco' as 0. For 'Gender', 'Male' is coded as 0 and 'Female' as 1. For 'Customer Type', 'loyal customer' is coded as 0 and 'disloyal customer' as 1. Finally, for 'Type of Travel', 'personal travel' is coded as 0 and 'business travel' as 1.

**2.3 Checking Correlations**

In addition to the basic customer information variables, there are 14 other variables that measure customer satisfaction on a scale from 1 to 5 across various aspects.

Since we label satisfaction as 0, we can see from the correlation diagram that most variables have an inverse relationship with satisfaction, the lower the rate leads to higher dissatisfaction.

We observed significant correlations among some of these variables, suggesting redundancy, and we may not need these 14 dimensionalities. Consequently, we employed Principal Component Analysis (PCA) to extract key components from these 14 variables. This approach allowed us to identify and eliminate unnecessary features, streamlining our dataset for more effective analysis.

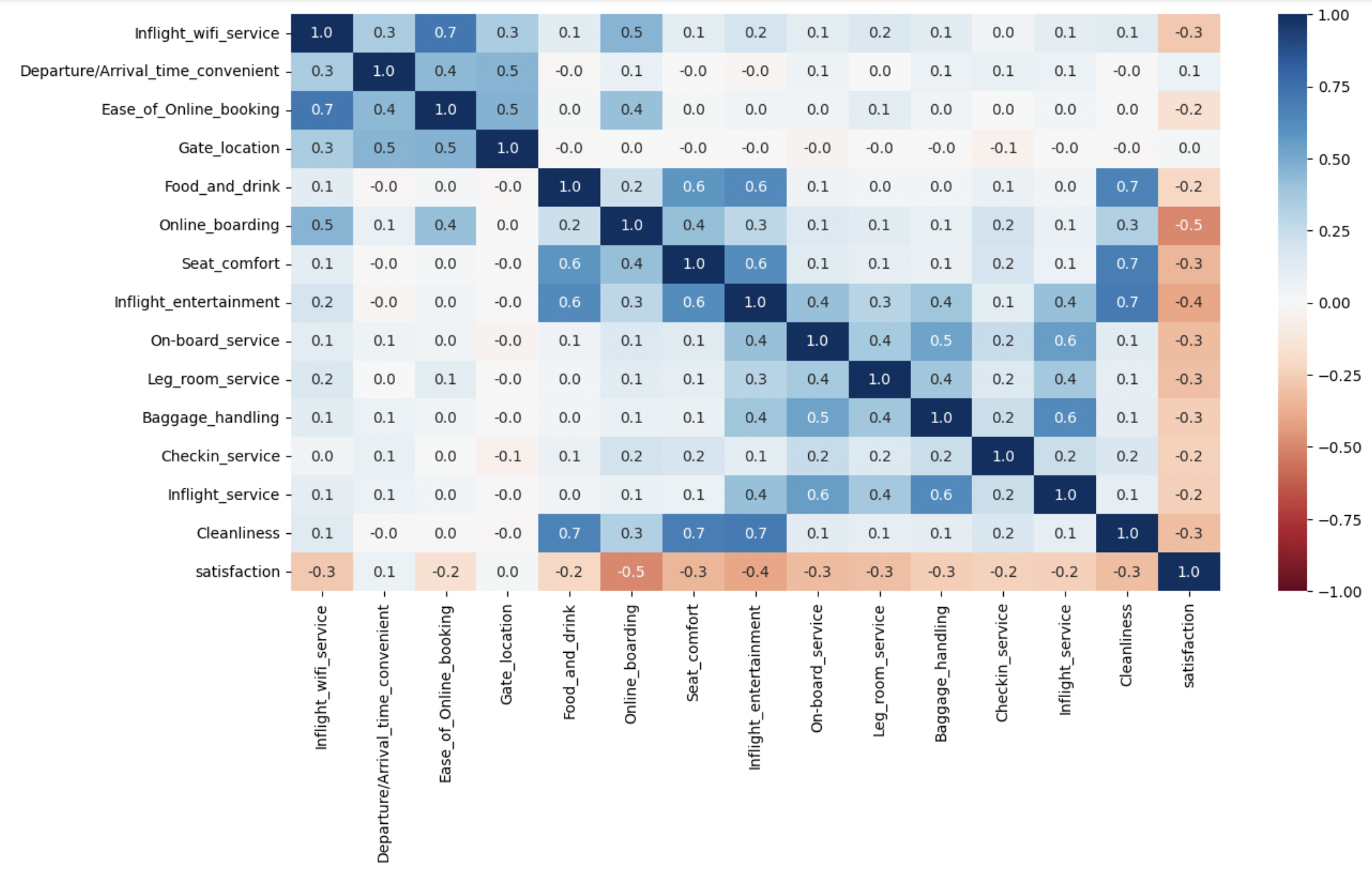


Figure 1: Correlation matrix of 14 feature

**2.4 Removing Unnecessary Features and Combining Variables**

We ran PCA on these 14 variables. Based on the eigenvalues table, we decided to use Latent Root Criterion to keep 4 components that can explain around 67% variance. Then we draw the correlation of components with original variables and try to interpret and come up with names for these factors.

* The first component is aptly named 'Inflight Discomfort Factor', reflecting its relatively high negative correlation with inflight entertainment, cleanliness, and seat comfort.
* The second component, 'Digital Convenience and Logistics Discontent Factor', captures the elements of customer experience related to logistical ease and technological facilitation, indicated by its strong negative correlation with inflight wifi service, departure/arrival timing, ease of online booking, and gate location.
* The third component, which we have labeled 'Service Quality Discontent Factor', is intricately linked to the dissatisfaction of direct customer service interactions, as evidenced by its high negative correlation with inflight service, baggage handling, and on-board service.
* Lastly, the fourth component, 'Online Boarding and Check-in Inefficiency', is characterized by its negative correlation with online boarding and check-in service.

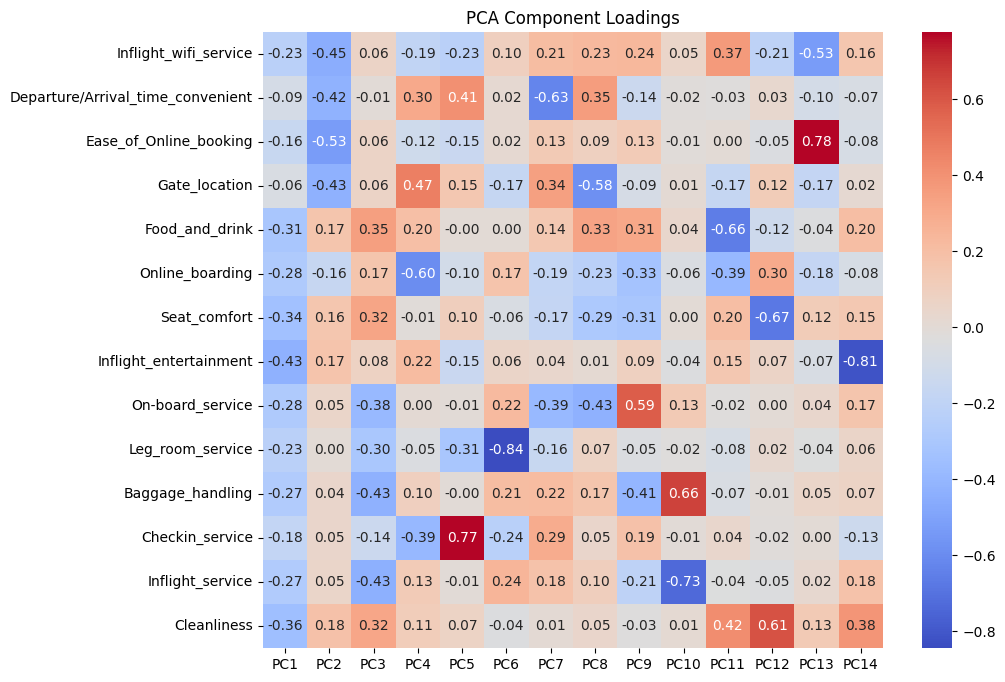


Figure 2: Components correlation with variables

Finally, we have successfully consolidated the original set of 24 variables into a more manageable set of 11 variables.



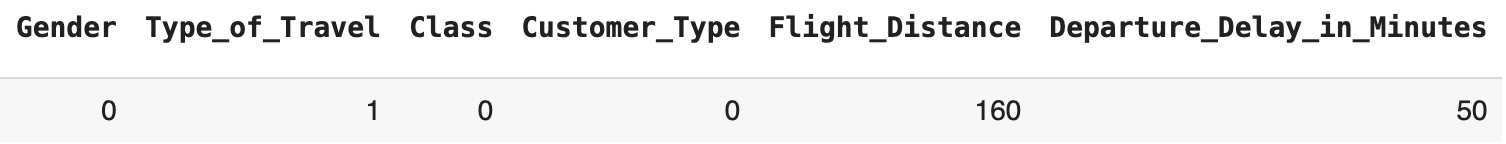


Figure 3: Final Dataset Sample (X only)

**3 Methodology**

**3.1 KNN**

To begin our exploration of predictive models, we implemented the K-Nearest Neighbors (KNN) model as our foundational approach to predict customer dissatisfaction levels. We normalized our feature set using the “StandardScaler” class from the “sklearn.preprocessing” module to ensure that each feature contributes equally to the distance calculations. We initially configured the KNN model with 5 neighbors to establish a performance baseline, achieving an accuracy of 80%. Recognizing the potential for improvement, we performed hyperparameter tuning for the number of neighbors, k. We executed a loop from 1 to 20, training a new KNN model at each increment and evaluating its accuracy on the test data.

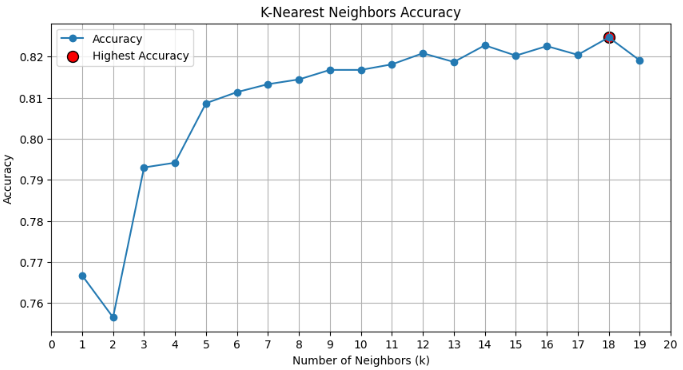


Figure 4: K-Nearest Neighbors Accuracy

This process revealed an optimal k value of 18, enhancing our model's accuracy to 82%. Following the optimization we evaluated the model’s effectiveness by analyzing its confusion matrix and computing both the true positive rate (TPR) and false positive rate (FPR). Our optimized KNN model demonstrated relatively robust performance, distinguishing satisfied from dissatisfied customers with an accuracy of 82%, a TPR of 83%, and an FPR of 21%, across a test dataset of 5,179 passengers which included 2,890 actual positives and 2,289 actual negatives. The model accurately identified 2,419 true positives and 1,815 true negatives, illustrating its reasonable efficacy in predicting dissatisfaction.

**3.2 Logistic Regression**

For our second model, we implemented Logistic Regression, initializing it with no regularization penalty and utilizing the 'lbfgs' solver for optimization. We first fit the model on the training data, achieving an initial accuracy of 79.7%. By introducing a cutoff threshold of 0.45 for classifying predicted probabilities, we refined our accuracy to 80.1%. Unlike the KNN model, this model provided a set of insightful coefficients, illustrating the influence of each feature on the likelihood of dissatisfaction.

Notably, 'Inflight Discomfort Factor' and 'Online Boarding and Check-in Inefficiency' showed positive coefficients, significantly increasing the probability of dissatisfaction, whereas 'Digital Convenience and Logistics Discontent Factor' had a minor negative impact. Taking this analysis further, by taking the exponential of a given coefficient we were able to determine the impact of one unit increase on the odds of being dissatisfied. As an example, the model suggests that increasing “Inflight Discomfort Factor” by one unit increases the odds of being dissatisfied by approximately 2.21 times. Additionally, we employed “statsmodels.api” for a deeper validation of the model, which involved assessing p-values and confidence intervals for the coefficients. This detailed analysis substantiated the importance of our key factors, with a Pseudo R-squared of 0.309, signifying a moderate level of explanatory power of the model.

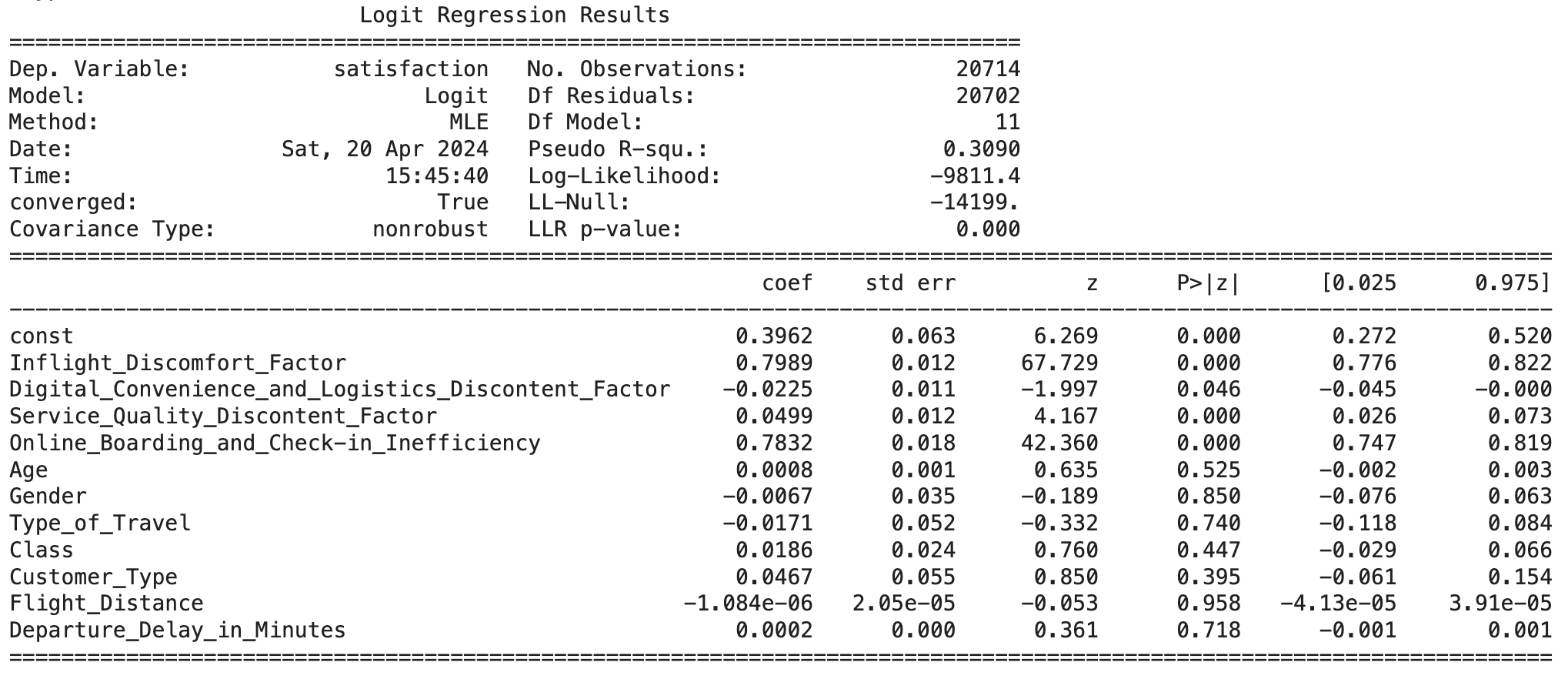


Figure 5: Logistic Regression Summary

**3.3 Decision Tree**

With an aim to delve deeper into the factors influencing customer dissatisfaction, we employed a Decision Tree classifier. Starting with an unpruned Decision Tree, we achieved a 100% accuracy on the training data and a 79.2% accuracy on the test data, indicating overfitting. To address this, we tuned the tree using hyperparameters such as ‘max\_depth’, ‘min\_samples\_split, min\_samples\_leaf’, and ‘min\_impurity\_decrease’, which brought the training accuracy to 80.5% and test accuracy to 79.3%, suggesting a better generalization. To gain actual insights, this new tree was visualized and we saw that significant splits were made according to ‘Inflight Discomfort Factor’ and ‘Digital Convenience and Logistics Discontent Factor’ features.

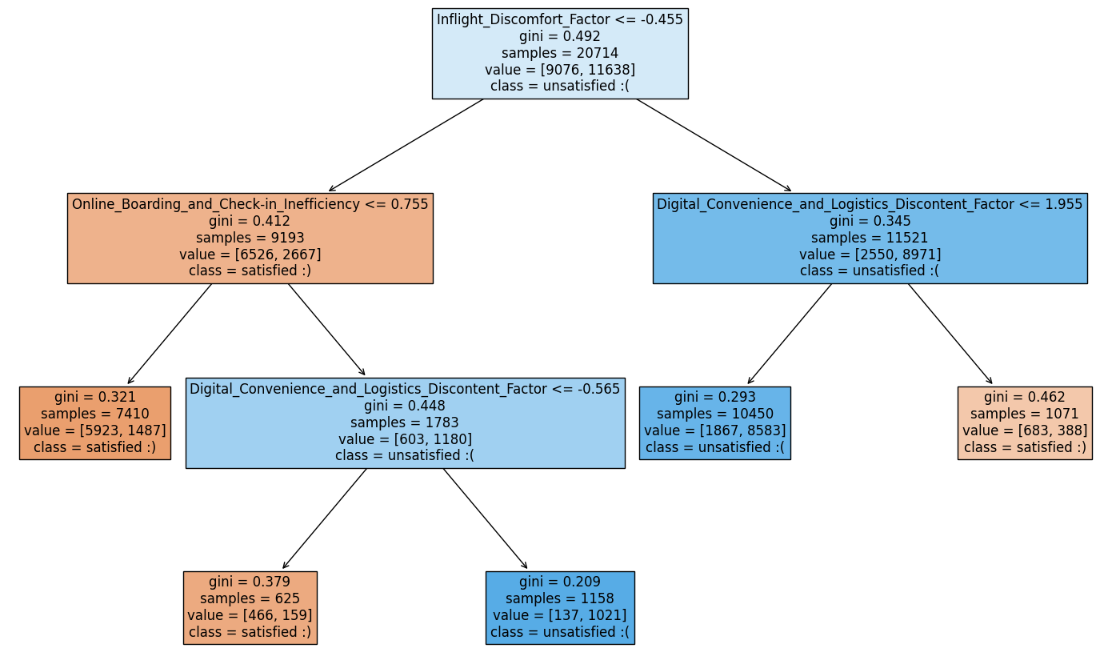


Figure 6: Decision Tree

Looking to improve on the 79.3% accuracy, we decided to apply cost complexity pruning on the full tree. By plotting accuracy against different values of the complexity parameter alpha we obtained an alpha score of 0.00019 which improved our test accuracy to 83.5%.

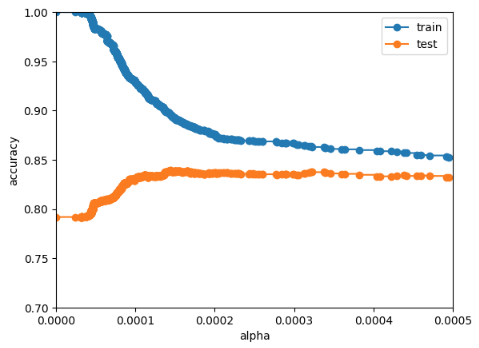


Figure 6: Accuracy of training and testing set under different ɑ

Next, we conducted feature importance analysis on the pruned Decision Tree and discovered that ‘Inflight Discomfort Factor’ was the most significant predictor of customer dissatisfaction, followed by the ‘Digital Convenience and Logistics Discontent Factor’, which corroborates the results found during Logistic Regression modeling, indicating areas for potential airline service improvement.

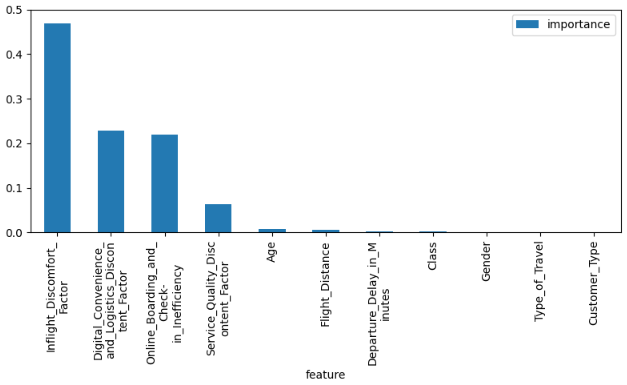


Figure 7: Feature Importance

Lastly, a final effort to finetune the model was made by utilizing GridSearchCV, which identified the best hyperparameter combination as a ‘max\_depth’ of 10, ‘min\_impurity\_decrease’ of 0, and ‘min\_samples\_leaf’ of 20, achieving an improved accuracy of 84.3%. Not only was this our best accuracy out of the 3 implemented models, Decision Tree analysis also provided us with interpretable insights into the key drivers of customer dissatisfaction.

**4 Model Comparison**

**4.1 Cost Benefit Model**

To effectively measure the impact of predictive models on participant outcomes, we simulated and assigned a cost-benefit weight to each potential result, facilitating a comprehensive evaluation. The assigned weights are as follows

* Benefit of True Positive (100): The potential value in correctly identifying dissatisfied customers, as this can lead to direct intervention, improve customer service, and potentially turn an unhappy customer into a loyal one.
* Cost of False Positive (-30): Incorrectly targeting a satisfied customer as dissatisfied can potentially lead to some unnecessary costs
* Benefit of True Negative (50): Correctly identifying a satisfied customer can help avoid unnecessary interventions, preserving resources and customer goodwill.
* Cost of False Negative (-70): Failing to identify a dissatisfied customer may lead to churn or negative word-of-mouth, potentially damaging the brand and causing loss of revenue.

|  | Predicted Negative | Predicted Positive |
| --- | --- | --- |
| Actual Negative | Benefit: 50 | Cost: -70 |
| Actual Positive | Cost: -30 | Benefit: 100 |

Table 1: Cost Benefit Model

**4.2 Model Comparison**

To comprehensively assess the performance of each predictive model, we consider accuracy and the total benefits calculated from the cost-benefit model.

|  | KNN | Logistic Regression | Decision Tree | Random Forest | Neural Network |
| --- | --- | --- | --- | --- | --- |
| Accuracy | 82% | 80% | Full Tree: 79.2%  Alpha: 83%  Gridsearch: 84% | 85% | 82% |
| Total Benefits | $294,720 | $283,630 | $299,400 | $311,740 | $283,710 |

Table 2: Accuracy and Benefits Comparison

The K-Nearest Neighbors (KNN) model demonstrates optimal accuracy of 82% and a total benefits of $294,720 when k is set to 18, which is comparatively large and could potentially lead to overfitting. Despite its simplicity, the model's outcomes lack interpretability. Moreover, the original dataset we first tired comprised 100,000 records, which significantly hampered the model's computational speed. Consequently, to enhance time efficiency, we were compelled to utilize the 20% dataset for the analysis.

The logistic regression model, despite yielding the lowest accuracy (80%) among the tested models, offers substantial interpretative value. It provides clear insights into the significance, directionality, and magnitude of each variable's influence on the outcome, allowing for targeted improvements in these areas. Additionally, the model's performance remains efficient when processing large datasets, maintaining swift operational speeds.

The Decision Tree model leads to an accuracy comparable to logistic regression while offering additional interpretive insights with the highest total benefits. It enables us to understand the combinations of factors that contribute to customer dissatisfaction or satisfaction. Moreover, through the strategic pruning of the tree, we are able to refine the model further. The decision tree runs reasonably fast but may slow considerably when looping through alphas on large datasets.

Finally, even though random forest and neural networks are powerful tools for predicting satisfaction, they do not provide insights into how these predictions are made, hindering our ability to understand or interpret the underlying factors driving the outcomes. Thus, we decided to not move forward with these two models.

**5 Findings and Implications**

KNN

Our team was able to achieve a high accuracy KKN model using our data set meaning that there are certain features that contribute to customer satisfaction. This allowed us to pivot to other models that would give us more specific insights towards individual features that impact customer satisfaction.

Logistics Regression

The logistic regression model provided the highest positive coefficient for 'Inflight Discomfort Factor' and 'Online Boarding and Check-in Inefficiency' when examining impact on customer dissatisfaction. Logistic regression indicates that airlines should prioritize customer inflight comfort and their pre-boarding process to prevent customer dissatisfaction. Investment in these facets of the experience will be relatively more effective and impactful than investments in other areas such as food and beverage.

Decision Trees

From our decision tree model, ‘Inflight Discomfort Factor**’** was the most significant predictor of customer dissatisfaction, followed by the ‘Digital Convenience and Logistics Discontent Factor’ and 'Online Boarding and Check-in Inefficiency'. These components are highlighted in the “splits” in the branches of the decision tree indicating which factors and branches lead to satisfied rather than dissatisfied customers.

**6 Conclusion and Our Recommendation**

Our team settled on primarily using our decision tree and logistic regression models to inform our recommendations. These two models provided the highest accuracy score with our data set. Additionally, we felt they were best suited to give insights in this situation. For example, the decision tree model is not a “black box” and allows for additional information given the structure of the branches unlike a model such as a neural network. By focusing on these two models analyzing customer satisfaction, we determined that ‘Inflight Discomfort Factor’ and 'Online Boarding and Check-in Inefficiency' are the most important factors for customer dissatisfaction. The decision tree model output shows ‘Inflight Discomfort Factor’, ‘Digital Convenience and Logistics Discontent Factor’, and 'Online Boarding and Check-in Inefficiency' as the three most significant components that predict customer dissatisfaction. The logistics regression model matches regarding ‘Inflight Discomfort Factor’ and ‘Digital Convenience and Logistics Discontent Factor’ as these are the two components with the highest absolute coefficients. However, ‘Digital Convenience and Logistics Discontent Factor’ is of relatively low significance according to this model.

‘Inflight Discomfort Factor’ includes the inflight entertainment, cleanliness, and seat comfort variables in the data set while ‘Online Boarding and Check-in Inefficiency’ includes the online boarding and check-in service variables. These variables are thus considered the most important for airlines to focus on to prevent customer dissatisfaction. Our team recommends airlines should make investments in these aspects of the flight customer experience as they will have an outsized impact on customer satisfaction when compared with other variables in the data set. For example, airlines prioritizing overall customer satisfaction could make improvements in seat cushions or seat reclining to improve satisfaction related to the ‘Inflight Discomfort Factor’. Additionally, they could make investments in their online check-in user interface and reliability to improve customer satisfaction related to the ‘Online Boarding and Check-in Inefficiency’. These insights paired with potential further testing provide critical insights for airlines to prioritize investment in areas that truly matter to the experiences of their customers.

**7 Potential Areas for Improvement**

While we generally believe that our results provide enough information to put together a comprehensive list of recommendations for airline companies to improve their in-flight service, there are still several actions we could take to improve our models and dataset and uncover even more insights.

First, instead of having two outputs labeled “satisfied” and “dissatisfied”, we could add an additional category labeled “neutral”. This distinction would prevent the ambiguity currently present when trying to determine whether customers marked “unsatisfied” are all genuinely unhappy with their in-flight experience. Since it is likely that there are many people in the dataset who do not have a strong opinion regarding their flight experience, it would improve the accuracy of our findings to take these customers into consideration as well.

Secondly, handling large datasets has proven to be slow, particularly when testing various combinations and experimenting with different parameters, such as the number of K clusters and diverse alpha values. This limitation somewhat limits our ability to optimize our models efficiently. The extensive time required for each iteration restricts our experimentation with larger training datasets which could potentially enhance model accuracy. If not constrained by time, using a larger dataset might allow us to significantly refine our model's accuracy by providing a richer foundation for training and validation. This approach would enable more thorough testing and potentially lead to more robust and precise outcomes. A larger dataset would give us a broader scope of customers to analyze and could improve the overall accuracy of our findings.

Additionally, it might be useful to look deeper into our data collection and make a larger and more variably comprehensive dataset. As of now, the dataset is focused only on U.S. commercial flights; expanding our collection to include international flights could uncover valuable information about how different cultural norms influence expectations that customers have regarding their flight experience. We have also not looked into the income class differences of people in the dataset; creating an additional variable set that separates people into classes such as “working class” and “upper class” would give us an idea of how expectations change based on a customer's income, which would help airlines target particular customer segments.

Finally, in our current model comparison approach, the costs and benefits are subjectively estimated by our team. Unfortunately, we lack concrete data to accurately quantify the actual benefits and costs associated with correctly or falsely identifying dissatisfied passengers. This limitation underscores the need for more rigorous evidence-based research. Creating a more objective costs and benefits analysis as a result of having more data and more specific models would provide more accurate information as to the significance of each individual outcome.

**8 Appendix**

**Variable Description:**

Gender: Gender of the passengers (Female, Male)

Customer Type: The customer type (Loyal customer, disloyal customer)

Age: The actual age of the passengers

Type of Travel: Purpose of the flight of the passengers (Personal Travel, Business Travel)

Class: Travel class in the plane of the passengers (Business, Eco, Eco Plus)

Flight distance: The flight distance of this journey

Inflight wifi service: Satisfaction level of the inflight wifi service (0:Not Applicable;1-5)

Departure/Arrival time convenient: Satisfaction level of Departure/Arrival time convenient

Ease of Online booking: Satisfaction level of online booking

Gate location: Satisfaction level of Gate location

Food and drink: Satisfaction level of Food and drink

Online boarding: Satisfaction level of online boarding

Seat comfort: Satisfaction level of Seat comfort

Inflight entertainment: Satisfaction level of inflight entertainment

On-board service: Satisfaction level of On-board service

Leg room service: Satisfaction level of Leg room service

Baggage handling: Satisfaction level of baggage handling

Check-in service: Satisfaction level of Check-in service

Inflight service: Satisfaction level of inflight service

Cleanliness: Satisfaction level of Cleanliness

Departure Delay in Minutes: Minutes delayed when departure

Arrival Delay in Minutes: Minutes delayed when Arrival

Satisfaction: Airline satisfaction level (Satisfaction, neutral or dissatisfaction)