ML FINAL PROJECT REPORT

Introduction and Motivation

The airline industry is a highly competitive one in which profit margins are low. A study by Kim and Lee (2011) showed that satisfied customers are more likely to recommend an airline and repurchase tickets with the same airline. This contributes to an increase in an airline's profitability as well as its market share, according to studies by Dagger et al (2007) and Buttle (1996). A study by Mattila (2001) concluded that loyal customers are more likely to forgive service failure and more likely to purchase tickets even when prices rise. Data from the 2019 Airline International Destination Satisfaction study demonstrated that positive customer satisfaction is more important than cheaper prices in a customer's decision to rebook with an airline. Furthermore, even if a passenger had not traveled with a particular airline in the past, they would be willing to pay more to travel with it if it has a good reputation. A good airline customer satisfaction score can help create a competitive advantage. For example, while United Airlines ranked higher for fewer canceled flights, Southwest ranked higher than United for customer satisfaction and ended up outranking United on the list of top airlines for overall performance.

On the other hand, low customer satisfaction scores could result in passengers switching to competitors and sharing their negative experiences about the airline with others, discouraging potential customers from flying with the airline. If a significant proportion of the airline's user base switches to competitors, this can substantially decrease the airline's profitability and can hinder the airline from expanding into new markets.

In response to the discussions in the above two paragraphs, we aim to utilize machine learning techniques to identify key factors driving airline customer satisfaction. Machine learning can help by extracting key insights and revealing patterns amongst large datasets. Using machine learning, we can answer the following question: Amongst various factors, which factors are most important in predicting airline customer satisfaction? Furthermore, machine learning models can predict how much airline customer satisfaction scores are likely to change in the future if investments into various features are made. For example, it is feasible that improving the in-flight wifi service could result in significant improvements in airline customer satisfaction scores. By utilizing machine learning, we aim to encourage airline companies, especially those with relatively low satisfaction scores, to prioritize and enhance these factors to boost overall customer satisfaction.

Delta Airlines, which had the lowest customer satisfaction score in 2023 according to Business Insider out of the four major American airline carriers (American, Southwest, Delta, and United), is poised to benefit substantially from utilizing machine learning models to enhance passenger satisfaction. Delta is America's oldest and second-largest airline company, founded in 1925. Delta has major hubs in Atlanta and Detroit and currently operates 5,400 daily flights. Delta primarily targets the business traveler and seeks to boost profitability by utilizing the hub and

spoke model, purchasing used aircraft to minimize fixed costs, and owning their fuel refineries. By employing machine learning to identify the most important factors driving airline customer satisfaction, Delta can work to improve these factors and hence improve the experience of their passengers, likely boosting their profitability.

Dataset and Data Cleaning

For this project, we utilized the passenger satisfaction dataset from Kaggle. The dataset contains customer satisfaction scores, an assessment of various factors (such as seat comfort and cleanliness), and additional information (gender, age, etc) from over 120,000+ airline passengers. Before data cleaning, the dataset contains 129,880 rows and 24 columns.

We first checked whether the dataset contains any missing values. The only column that contained missing values was “Arrival Delay”, a column that measures flight arrival delay in minutes. After checking for missing values, we dropped all the rows containing missing values and replaced all spaces and hyphens in column names with underscores. We then dropped the “ID” column as we believed that it wouldn't be significant in predicting airline customer satisfaction scores. We then mapped all the non-numeric columns (such as gender and type of travel) to numeric ones and converted the type of these columns to int64.

```
#convert satisfaction to numbers
#gender
#customer type
#type of travel
#class
sat = {'Neutral or Dissatisfied': 0, 'Satisfied': 1}
df.loc[:,'Satisfaction'] = df['Satisfaction'].map(sat)

gd = {'Female':0, 'Male':1}
df.loc[:,'Gender'] = df['Gender'].map(gd)

ct = {'First-time':0, 'Returning':1}
df.loc[:,'Customer_Type'] = df['Customer_Type'].map(ct)

tt = {'Business':0, 'Personal':1}
df.loc[:,'Type_of_Travel'] = df['Type_of_Travel'].map(tt)

cl = {'Business':0, 'Economy Plus': 1, 'Economy' : 2}
df.loc[:,'Class'] = df['Class'].map(cl)
```

```
df['Satisfaction'] = df['Satisfaction'].astype('int64')
df['Gender'] = df['Gender'].astype('int64')
df['Customer_Type'] = df['Customer_Type'].astype('int64')
df['Type_of_Travel'] = df['Type_of_Travel'].astype('int64')
df['Class'] = df['Class'].astype('int64')
```

The dataset contained 129,487 rows and 23 columns after undergoing data cleaning. Furthermore, each of these 23 columns was now in numeric format, ensuring that the dataset was ready for machine learning models to be applied.

Model 1: Logistic Regression

```
from sklearn.model_selection import train_test_split
df_train, df_test = train_test_split(df, test_size=0.3, stratify=df['Satisfaction'], random_state=88)
df_train.shape, df_test.shape
```

To prepare to fit a logistic regression model on the dataset, we split the dataset into a training dataset (df_train) and a testing dataset (df_test). The training dataset contained 70% of the original data, and the testing dataset contained 30% of the original data.

We then fit the logistic regression model on the training data. Our logistic regression model predicted the overall satisfaction level based on all the other variables in the dataset.

```
Optimization terminated successfully.
         Current function value: 0.335322
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:           Satisfaction   No. Observations:                90640
Model:                          Logit   Df Residuals:                    90617
Method:                           MLE   Df Model:                           22
Date:                Mon, 06 May 2024   Pseudo R-squ.:                  0.5102
Time:                        18:51:45   Log-Likelihood:                -30394.
converged:                       True   LL-Null:                       -62047.
Covariance Type:            nonrobust   LLR p-value:                     0.000
=============================================================================================================
                                            coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------------------------------
Intercept                                -7.8654      0.083    -94.223      0.000      -8.029      -7.702
Gender                                    0.0870      0.021      4.186      0.000       0.046       0.128
Age                                      -0.0080      0.001    -10.490      0.000      -0.009      -0.006
Customer_Type                             1.9964      0.032     63.100      0.000       1.934       2.058
Type_of_Travel                           -2.7312      0.033    -81.644      0.000      -2.797      -2.666
Class                                    -0.3625      0.014    -26.560      0.000      -0.389      -0.336
Flight_Distance                        2.651e-06   1.19e-05      0.223      0.824   -2.07e-05     2.6e-05
Departure_Delay                           0.0039      0.001      3.747      0.000       0.002       0.006
Arrival_Delay                            -0.0089      0.001     -8.600      0.000      -0.011      -0.007
Departure_and_Arrival_Time_Convenience   -0.1314      0.009    -15.032      0.000      -0.149      -0.114
Ease_of_Online_Booking                   -0.1573      0.012    -13.026      0.000      -0.181      -0.134
Check_in_Service                          0.3323      0.009     36.437      0.000       0.314       0.350
Online_Boarding                           0.6148      0.011     56.247      0.000       0.593       0.636
Gate_Location                             0.0280      0.010      2.867      0.004       0.009       0.047
On_board_Service                          0.3088      0.011     28.444      0.000       0.287       0.330
Seat_Comfort                              0.0633      0.012      5.291      0.000       0.040       0.087
Leg_Room_Service                          0.2483      0.009     27.340      0.000       0.230       0.266
Cleanliness                               0.2256      0.013     17.506      0.000       0.200       0.251
Food_and_Drink                           -0.0347      0.011     -3.044      0.002      -0.057      -0.012
In_flight_Service                         0.1095      0.013      8.533      0.000       0.084       0.135
In_flight_Wifi_Service                    0.3913      0.012     32.032      0.000       0.367       0.415
In_flight_Entertainment                   0.0702      0.015      4.626      0.000       0.040       0.100
Baggage_Handling                          0.1394      0.012     11.426      0.000       0.115       0.163
=============================================================================================================
```

The above picture shows the output of the logistic regression model. We interpret the logistic regression model as follows: Holding all other variables constant, each 1 (unit/point/dollar/etc) increase in (variable of interest) yields $e^{\text{Coefficient for the variable of interest}}$ times higher estimated odds of a customer being satisfied with the airline.

As an example, the coefficient for Seat_Comfort, the satisfaction level with the comfort of the airplane seat, is 0.0633. $e^{0.0633} = 1.0653$. So, the interpretation for the Seat_Comfort variable would be as follows: Holding all other variables constant, each 1-point increase in the satisfaction level with the comfort of the airplane seat yields a 1.0653 times higher estimated odds of a customer being satisfied with the airline.

After fitting the logistic regression model on the training data, we then applied the model to the test set, rounding the result to either 0 or 1 based on a threshold of 0.5 (numbers over 0.5 were rounded to 1, and numbers below 0.5 were rounded to 0), and evaluated its performance by calculating the following metrics: confusion matrix, accuracy, true positive rate, and false positive rate, and AUC. The results of the model evaluation are discussed in the next paragraph.

The number of true negatives was 19832, the number of false negatives was 2790, the number of false positives was 2136, and the number of true positives was 14089. The accuracy of the model in predicting customer satisfaction with an airline was 87.32%. The true positive rate was 83.47%, and the false positive rate was 9.72%. The AUC, which evaluates the overall quality of a logistic regression model, was 0.93 out of 1, suggesting that the logistic regression model performed well in predicting customer satisfaction level with an airline.

Model 2: Decision Tree

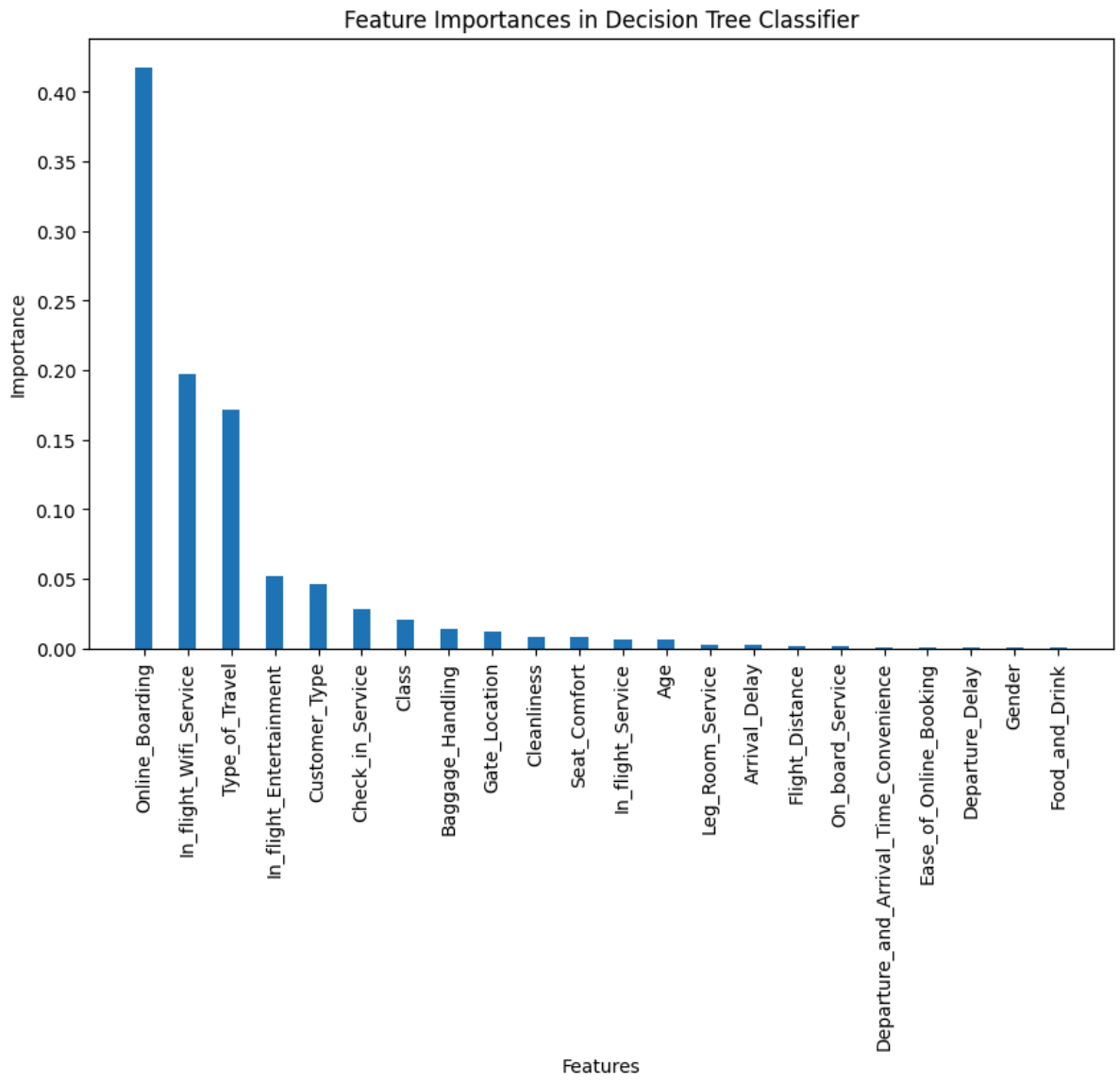
```
y = df['Satisfaction']
X = df.drop('Satisfaction', axis=1)

X_train, X_test, y_train, y_test = train_test_split(X, y,
                                                    test_size=0.3,
                                                    stratify=df['Satisfaction'],
                                                    random_state=88)

X_train.shape, X_test.shape
```

To prepare to fit a decision tree model on the dataset, we split the dataset into a training dataset and a testing dataset in the same way that we split the dataset in the logistic regression model, except we explicitly isolated the dependent variable (Satisfaction) from the independent variables, as we split the data into X_train, X_test, y_train, and y_test instead of df_train and df_test. The training dataset contained 70% of the original data, and the testing dataset contained 30% of the original data. Note that by stratifying the split on the dependent variable and setting the random_state, a parameter that enables reproducibility, parameter equal to 88 (the random_state parameter was 88 when we split the logistic regression model as well), the

distribution of the training dataset is identical to that of the distribution of the training dataset in the logistic regression model.

We used cross-validation, exploring depths from 1 to 10, to select the best depth for the decision tree model. A depth of 10 turned out to be the best depth for the decision tree model using cross-validation, with an AUC of about 0.986. We then used this depth as a parameter in applying the decision tree model to the testing dataset and then evaluated its performance by calculating the following metrics: confusion matrix, accuracy, true positive rate, and false positive rate, training AUC, testing AUC. Finally, we plotted the most important features of the decision tree classifier in predicting whether a customer is satisfied with the airline. The next couple of paragraphs discuss the results of the decision tree model evaluation and the most and least important predictive features of the model.

The number of true negatives was 21182, the number of false negatives was 1378, the number of false positives was 786, and the number of true positives was 15501. The accuracy of the model in predicting customer satisfaction with an airline was 94.43%. The true positive rate was 91.84%, and the false positive rate was 3.58%. The training AUC was 0.9906 and the testing AUC was 0.9854, suggesting that the decision tree model performed well in predicting customer satisfaction with an airline. The training AUC being close to the testing AUC suggests that overfitting is not an issue with the decision tree model.



As can be observed in the plot above, the three most important features (in descending order) of the decision tree model in predicting airline customer satisfaction were satisfaction with the online boarding experience (corresponding to the Online_Boarding variable), satisfaction with the in-flight wifi service (corresponding to In_flight_Wifi_Service), and the type of travel (corresponding to Type_of_Travel), whether business or personal.

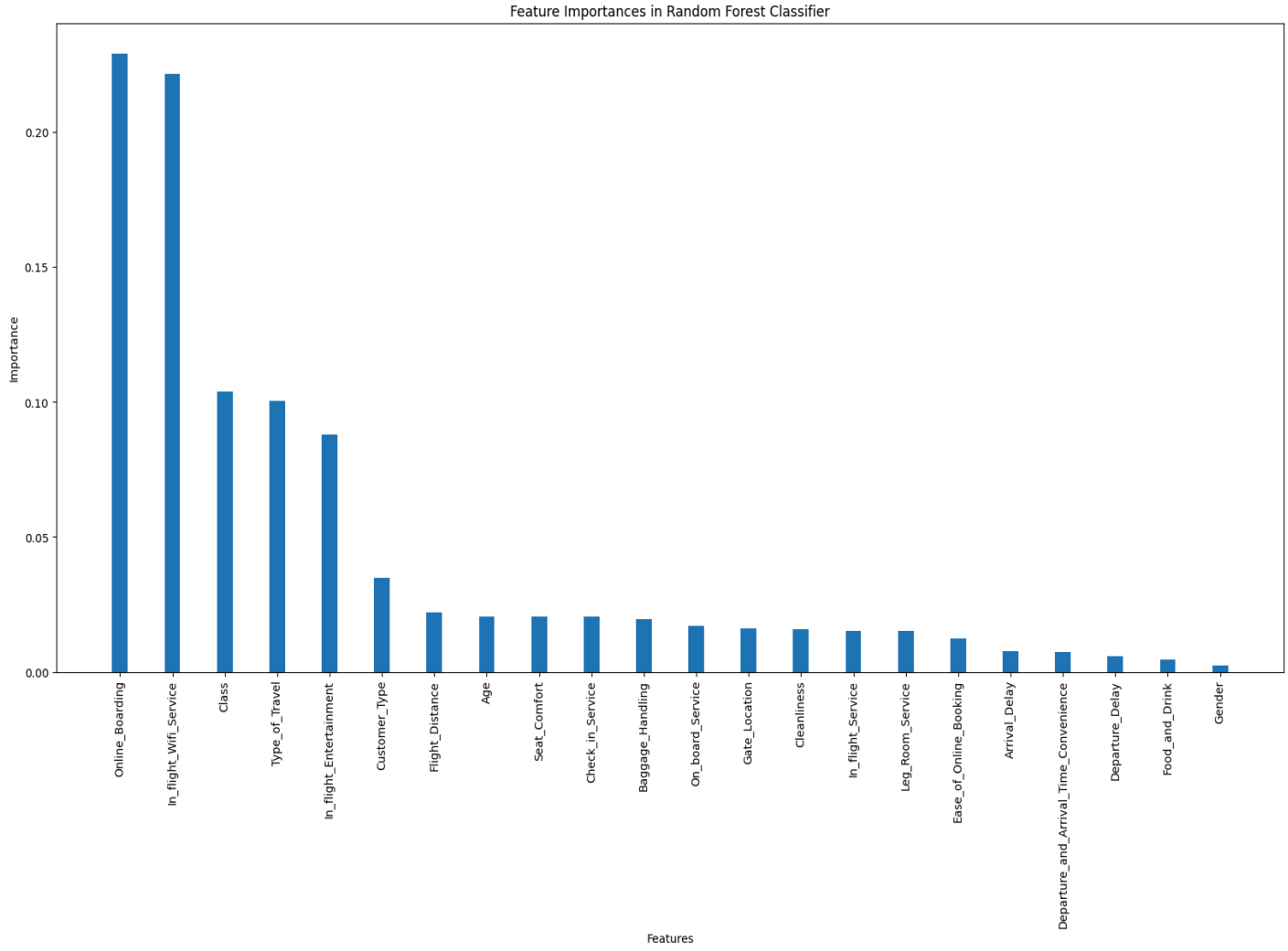
The three least important features of the decision tree model in predicting airline customer satisfaction, starting with the least significant feature and going in ascending order, were satisfaction with the food and drinks on the airplane (corresponding to Food_and_Drink), the gender of the passenger (corresponding to Gender), and flight departure delay in minutes (corresponding to Departure_Delay).

Model 3: Random Forest

We reused the data split from the decision tree model for the random forest model, so we did not have to split the dataset into the training and testing datasets again.

We used cross-validation to determine the max_features (number of features considered when looking for the best split at each node of a tree) parameter for the random forest model. Using cross-validation, 13 features turned out to be the optimal choice for the random forest's max_features parameter. We set the n_estimators (number of trees in the random forest) parameter to 5 and the min_samples_leaf parameter to 3 (minimum number of leaf node samples per tree). We then used this depth as a parameter in applying the random forest model to the testing dataset and then evaluated its performance by calculating the following metrics: confusion matrix, accuracy, true positive rate, and false positive rate, training AUC, testing AUC. Finally, we plotted the most important features of the random forest classifier in predicting whether a customer is satisfied with the airline. The next couple of paragraphs discuss the results of the random forest model evaluation and the most and least important predictive features of the model.

The number of true negatives was 21317, the number of false negatives was 1033, the number of false positives was 651, and the number of true positives was 15846. The accuracy of the model in predicting customer satisfaction with an airline was 95.67%. The true positive rate was 93.88%, and the false positive rate was 2.96%. The training AUC was 0.9992 and the testing AUC was 0.9899, suggesting that the random forest model performed well in predicting customer satisfaction with an airline. The training AUC being close to the testing AUC suggests that overfitting is not an issue with the random forest model.



As can be observed in the plot above, the three most important features (in descending order) of the random forest model in predicting airline customer satisfaction were satisfaction with the online boarding experience (corresponding to Online_Boarding), satisfaction with the in-flight wifi service (corresponding to In_flight_Wifi_Service), and the travel class (corresponding to Class), which includes business, economy, and economy plus.

The three least important features of the random forest model in predicting airline customer satisfaction, starting with the least significant feature and going in ascending order, of the random forest model were the gender of the passenger (corresponding to Gender), satisfaction with the food and drinks on the airplane (corresponding to Food_and_Drink), and flight departure delay in minutes (corresponding to Departure_Delay).

Discussion of Benefits and Risks

All three classification models that we applied (logistic regression, decision tree, and random forest) achieved testing AUCs of over 0.9, indicating that the overall quality of each of the three models is high. The good testing AUCs also indicate that overfitting is not an issue with the

three models. Furthermore, all three models had accuracies of over 85%, true positive rates of over 80%, and false positive rates of below 10%.

Assuming the underlying distribution of airline customer satisfaction data is the same, Delta can continue to use the random forest model to identify the most important factors driving airline customer satisfaction and work to improve them if needed. This could result in an improved experience for their passengers, likely boosting their profitability. However, consumer preferences and market conditions may change over time, so machine learning models may not be able to accurately predict airline customer satisfaction in the future without fine-tuning at regular time intervals. For example, the quality of social distancing measures temporarily became a factor in predicting airline customer satisfaction during the COVID lockdowns, which this dataset (and hence the models) did not account for.

We do not believe that the models are oversimplified. Over 120,000 rows and 20 columns were used in predicting airline customer satisfaction. Furthermore, the models were rigorously trained and tested. Each of the three models that we applied was trained on over 85,000 rows of data and tested on over 35,000 rows of data. In the case of the decision tree model, we explored depths from 1-10 to find out the depth that gives the best cross-validation AUC, one of the best methods to check the overall quality of a model and applied this depth to the testing dataset. In the case of the random forest model, we evaluated various options of the max_features parameter ranging from 1-18 to find out the best number of features to consider for the split at each node of a tree that gives the best cross-validation AUC and applied this parameter to the testing dataset.

We do not see serious ethical concerns about applying machine learning models to this specific dataset for improving airline customer satisfaction scores. The only personally identifying information in this dataset is a person's gender and age. Given only a person's gender and age, it is very unlikely that an individual can be uniquely identified. However, if machine learning models are applied to a dataset in the same setting that includes full names and additional personal information, then it could be the case that a person is uniquely identified, potentially leading to misuse of their data. Airline companies may use the satisfaction scores to target specific customers with personalized recommendations without explicit consent. For example, customers who rated the online boarding experience poorly might receive emails advertising new features of the airline's online boarding experience. If customers believe that their information is being used in ways they did not consent to, they may switch to traveling with a different airline and discourage others from traveling with the airline, likely decreasing the airline's profitability.

Conclusion

Out of the three classification models (logistic regression, decision tree, and random forest) that we fit on the dataset, random forest performed the best in predicting airline customer satisfaction. Random forest had the highest AUC among the three models. Hence, it is no surprise that the random forest model also had the highest True Positive Rate and the lowest

False Positive Rate among the three models. Furthermore, the random forest model also had the highest accuracy among the three models. So, random forest performed the best amongst the models for all the above metrics. Given the robust performance of the Random Forest Model, it can be used to provide insights for airline companies with relatively low airline satisfaction scores like Delta Airlines. As the online boarding experience is the most important factor in predicting airline customer satisfaction according to the Random Forest model, Delta can target improvements in this area if needed.

Future work could incorporate additional datasets from multiple sources (not just Kaggle) related to airline customer satisfaction to refine predictive accuracy. It is important to continue monitoring the performance of all three models (the best-performing model may change over time) and updating them as necessary to account for potential changes in customer preferences over time. This will help ensure that the models remain accurate and up-to-date, helping airlines continuously optimize passenger experiences and likely boost their profitability.

WORKS CITED

- Bhat, Mysar Ahmad. “Airline Passenger Satisfaction.” Kaggle, 19 May 2022, www.kaggle.com/datasets/mysarahmadbhat/airline-passenger-satisfaction/data.
- C, Sam. “Delta Airlines: Flying High in a Competitive Industry.” Harvard Business School, 8 Dec. 2015, d3.harvard.edu/platform-rctom/submission/delta-airlines-flying-high-in-a-competitive-industry/.
- Giacobone, Bianca. “Alaska Airlines Is Passenger’s New Favorite Airline According to a New Customer Satisfaction Survey. See How US Airlines Ranked, from Worst to Best.” Business Insider, Business Insider, 20 Apr. 2023, www.businessinsider.com/us-airlines-ranking-customer-satisfaction-index-travel-alaska-spirit-2023-4.
- Kochański, Rafał. “5 Ways to Improve Airline Customer Service.” Hubtype, 24 Jan. 2024, www.hubtype.com/blog/ways-to-improve-airline-customer-service#:~:text=It%20increases%20customer%20loyalty,industry%20it%20is%20particularly%20notable.
- Leon, Steven, and Sonoma Dixon. “Airline satisfaction and loyalty: Assessing the influence of personality, trust and Service Quality.” Journal of Air Transport Management, vol. 113, Oct. 2023, https://doi.org/10.1016/j.jairtraman.2023.102487.