Predicting Airline Passnger Satisfaction

Team Number (or name):

9

Team Members:

Boris Topalov (bnt4yb),

Iain Muir (iam9ez),

Jingxuan Liu (jl2hv),

Ethan Chen (ec5rdn)

Course: COMM 4522 - Business Analytics with Python

Instructor: Dr. Mousavi

Date Submitted: April 28th, 2022

1 Executive Summary

1.1 Business Problem Statement

Recently, the largest U.S. airlines disclosed that the vast majority of their profits are generated by their frequent flyer (i.e. customer loyalty) programs. As such, the biggest priority for each airline is to attract users to their programs, even if it means operating the actual flights at a loss, as they currently do. With such fierce competition among frequent flyer programs, airlines need to figure out how to attract customers to their company over a rival. This is especially critical in the airline industry where customer loyalty and satisfaction is generally low.

1.2 Business Goal

The goal of our project is to figure out which factors effectively increase or decrease customer satisfaction on flights. By improving on these factors, airlines will be able to deliver more satisfying flight experiences to their customers, giving them a competitive edge over their rivals and attracting customers that are up for grabs to their programs. In addition, by improving the customer experience, airlines can improve their brands and the reputation of flying itself, making the experience seem less of a hassle and attracting more customers to flights in general. Thus, the findings of this project will also help airlines pull new customers into the market as well. Ultimately, the goal of the project is to help airlines attract more customers into their customer loyalty programs by figuring out how to improve their flight experience.

1.3 Data Profile

The data we used for investigation and analysis was the ‘Airline Passenger Satisfaction’ dataset published in 2020 on Kaggle. It was divided into separate training and testing sets, where the training set has 103,904 records and the testing set has 25,976 records. There are 24 columns in total, including 4 continuous variables, 5 categorical variables, 14 discrete rating variables, and the target variable which has two categories. There are 8,488 records that have missing or N/A values, which were either dropped from the dataset or replaced with median value depending on the column type. Our target attribute are in ones and zeros, where 1 represents satisfied customers and 0 represents neutral or dissatisfied customers. The target attribute is of roughly equal balance with 56.7% being neutral or dissatisfied and 43.3% being satisfied.

1.4 Results

Our best-performing model on our training dataset was a 4-layer Neural Network model, which scored a 94.95% accuracy. However, the best performing model on our test dataset was an XGBoost model with hyperparameters tuned using a cross-validated grid search. This model scored a 67.79% accuracy and a ROC-AUC score of 80.444%. It had a recall score for the positive class of 49.64%, and recall score of 82.0% for the negative class. Based on this model, we found that the most important predictors of customer satisfaction for airlines are passenger class, inflight service, and the arrival delay of a plane in minutes. Airlines should use the findings from our analysis to know what areas of customer satisfaction to focus on.

2 Project Report

2.1 Introduction

Recently, the largest U.S. airlines disclosed that the vast majority of their profits are generated by their frequent flyer (i.e. customer loyalty) programs. As such, the biggest priority for each airline is to attract users to their programs, even if it means operating the actual flights at a loss, as they currently do. With such fierce competition among frequent flyer programs, airlines need to figure out how to attract customers to their flights and programs over a rival's. This is especially critical in the airline industry where customer loyalty and satisfaction is generally low.

The goal of our project is to figure out which factors effectively increase or decrease customer satisfaction for airlines. In 2020, the number of US airline complaints soared to a new record, revealing a 568% increase from the previous year. However, according to J.D. Power, airline customer satisfaction reached a new high in 2021, marking a quick turn-around as corporations adapted to the COVID-19 pandemic. By improving on these factors, airlines will be able to deliver more satisfying flight experiences to their customers, giving them a competitive edge over their rivals and attracting customers that are up for grabs to their programs.

A survey published in the Harvard Business Review states that 9 out of 10 customers expect businesses to anticipate their needs ahead of time. With such high expectations and pressure on businesses, airlines must do everything they can to ensure high customer satisfaction. Identification of the underlying factors behind the dissatisfaction will help airline companies have a better idea of passenger's focus and concerns, set improvement targets accordingly, and allocate resources efficiently to increase customer satisfaction.

Increasing customer satisfaction is important both for competing with competitors, as well as making customers more likely to frequently use air travel. If customer satisfaction is too low, not only may customers turn to other competitors in the airline industry, they may even turn to non-airline competitors like trains or long-distance buses, or may even reduce their traveling altogether. As a business, airlines seek to increase satisfaction in the most efficient ways possible. Thus, they cannot afford to simply increase travel quality across the board - thus, this project will help them pick and choose factors so that they can balance improving customer satisfaction with cost effectiveness.

2.2 Background

Using supervised and unsupervised classification techniques for predicting customer satisfaction has been a common data science task over the past decade or so. Whether a company is struggling with their customer satisfaction reputation or they want to ensure their competitive advantage in that aspect, all companies strive to know what makes their customers happy. Prominent examples of these use cases include Banco Santander and Brazilian E-Commerce Marketplace Olist.

Banco Santander is a Spanish multinational financial services company based in Madrid and Santander in Spain, and in the late 2000s, it was quickly building the reputation of being one of the worst institutions in customer satisfaction. Specifically, Santander came bottom of JD Power's survey in 2007, 2008, and 2009. Six years ago, the bank took to Kaggle to ask freelance Data Scientists to help them identify dissatisfied customers early in their relationship, with a prize of $60,000 for the winner. Although there are no publications from the winners of the competition, many exist from the competition that had over 5,000 teams.

One of the best publications we found was from Ashish Thomas on Towards Data Science. In their article, Thomas details the business problem, the use of machine learning for solving the problem, the key evaluation metric, and then dives into extensive EDA and model building. Specifically, Thomas used various forms of feature engineering and data transformation to create six separate datasets: Normal, Normal with One Hot Encoding, Normal with Response Encoding, Log Transformed, Log Transformed with Response encoding, and Log Transformed with One Hot Encoding. Thomas then proceeded to test five separate models on each of the datasets: Logistic Regression, Decision Trees, Random Forest, XGBoost and LightGBM. All of this goes to show how many data science techniques can be applied to solve this business problem.

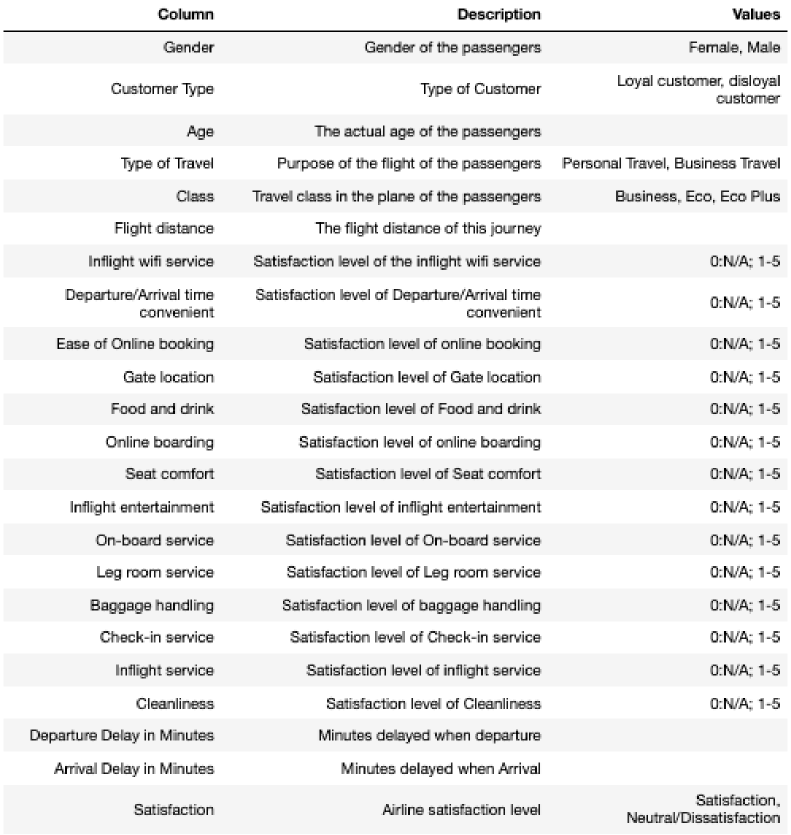
For further exploration into this problem, Thomas also lists two existing solutions he thought were equally informative.

Olist is the largest department store/e-commerce marketplace in Brazil, which connects small businesses from all over Brazil to channels without hassle and with a single contract. Although not in the form of a competition, Olist generously released their real and anonymised commercial data, containing over 100,000 records. The dataset has gotten over 2,000 “up-votes” on Kaggle, while over 200 people have submitted their code solutions. Potential ideas listed as inspiration include Natural Language Processing, Clustering, Sales Prediction, Delivery Performance, Product Quality, and Feature Engineering.

One of the best publications we found was from Praveen Hegde on Towards Data Science. Similar to Thomas, Hegde provides an extensive background to the problem and uses a wide variety of machine learning models. Specifically, Hegde implements a few more basic models—Logistic Regression, KNN, SVM, and Decision Tree—as well as a few more complex models—Random Forest, LightGBM, XGBoost, and RUSBoost. Within the realm of predicting customer satisfaction and binary classification in general, there are clearly a vast amount of data science techniques that can deliver business results.

2.3 Data

Our dataset comes from Kaggle, a subsidiary of Google LLC that allows users to download and publish datasets. This dataset “Airline Passenger Satisfaction” was published in 2020, and divided into a training set and a testing set. There are 103,903 and 25,975 records, respectively, and 25 columns within the datasets, and the data types include integer, string, and float. Below is a description of the attributes of our dataset:

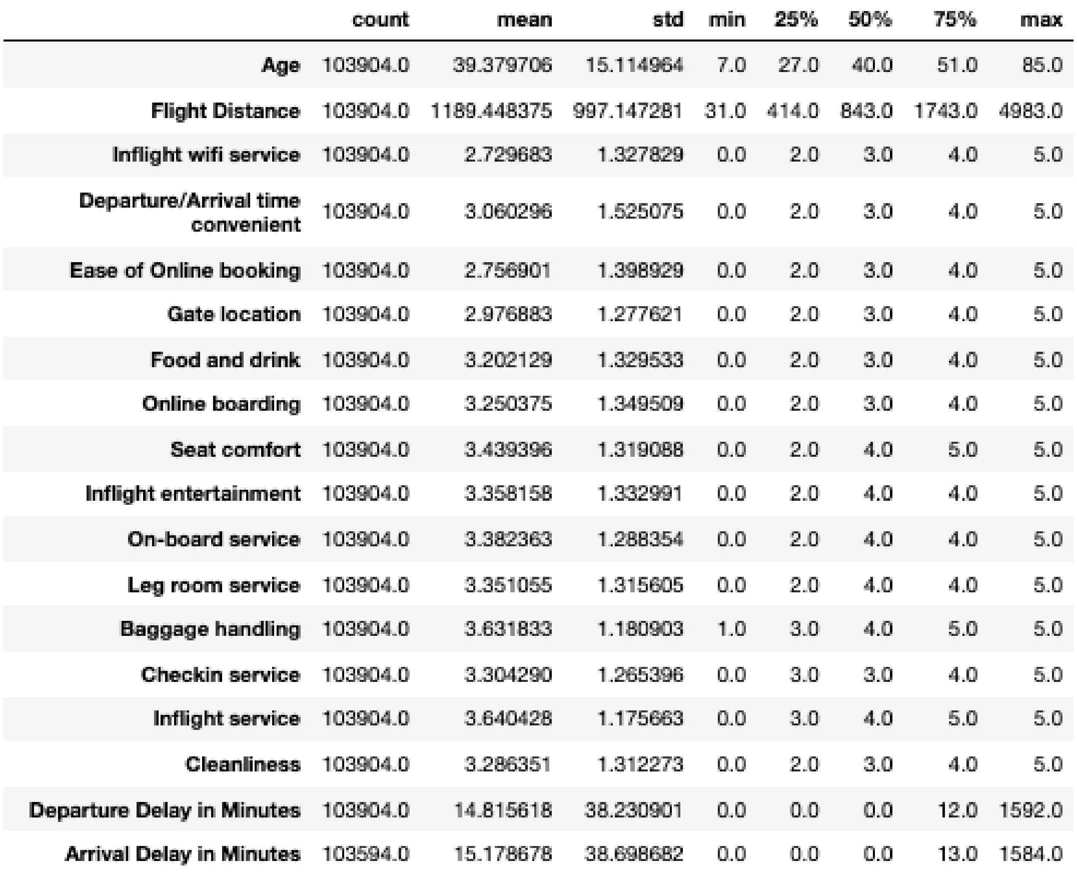


Column	Description	Values
Gender	Gender of the passengers	Female, Male
Customer Type	Type of Customer	Loyal customer, disloyal customer
Age	The actual age of the passengers	
Type of Travel	Purpose of the flight of the passengers	Personal Travel, Business Travel
Class	Travel class in the plane of the passengers	Business, Eco, Eco Plus
Flight distance	The flight distance of this journey	
Inflight wifi service	Satisfaction level of the inflight wifi service	0:N/A; 1-5
Departure/Arrival time convenient	Satisfaction level of Departure/Arrival time convenient	0:N/A; 1-5
Ease of Online booking	Satisfaction level of online booking	0:N/A; 1-5
Gate location	Satisfaction level of Gate location	0:N/A; 1-5
Food and drink	Satisfaction level of Food and drink	0:N/A; 1-5
Online boarding	Satisfaction level of online boarding	0:N/A; 1-5
Seat comfort	Satisfaction level of Seat comfort	0:N/A; 1-5
Inflight entertainment	Satisfaction level of inflight entertainment	0:N/A; 1-5
On-board service	Satisfaction level of On-board service	0:N/A; 1-5
Leg room service	Satisfaction level of Leg room service	0:N/A; 1-5
Baggage handling	Satisfaction level of baggage handling	0:N/A; 1-5
Check-in service	Satisfaction level of Check-in service	0:N/A; 1-5
Inflight service	Satisfaction level of inflight service	0:N/A; 1-5
Cleanliness	Satisfaction level of Cleanliness	0:N/A; 1-5
Departure Delay in Minutes	Minutes delayed when departure	
Arrival Delay in Minutes	Minutes delayed when Arrival	
Satisfaction	Airline satisfaction level	Satisfaction, Neutral/Dissatisfaction

Figure 1

The model we aim to build will predict customer satisfaction (*Satisfaction*) with the airline, which becomes our dependent variable. We will refer to the training set as our dataset in the discussion on data characteristics that follows. In the dataset, Arrival Delay in Minutes is the only attribute that has missing values, where 310 values are missing. We set missing values as null values and removed all rows with null values from our dataset. For all the satisfaction level attributes where customers rate each type of service they receive, a 0 represents that the service was not available to the customer (e.g. if the flight the customer takes does not provide inflight wifi, satisfaction level for inflight wifi service will be an NA). We substituted all zeros in these columns with the median of the column to make sure they don't bias our result. We also removed two columns that are unnecessary for the analysis, row numbers and customer ids. There aren't any outliers in this dataset.

After rows with null values and unnecessary columns are removed, below is a table summarizing key statistics of the final set of variables we will use in the model:



	count	mean	std	min	25%	50%	75%	max
Age	103904.0	39.379706	15.114964	7.0	27.0	40.0	51.0	85.0
Flight Distance	103904.0	1189.448375	997.147281	31.0	414.0	843.0	1743.0	4983.0
Inflight wifi service	103904.0	2.729683	1.327829	0.0	2.0	3.0	4.0	5.0
Departure/Arrival time convenient	103904.0	3.060296	1.525075	0.0	2.0	3.0	4.0	5.0
Ease of Online booking	103904.0	2.756901	1.398929	0.0	2.0	3.0	4.0	5.0
Gate location	103904.0	2.976883	1.277621	0.0	2.0	3.0	4.0	5.0
Food and drink	103904.0	3.202129	1.329533	0.0	2.0	3.0	4.0	5.0
Online boarding	103904.0	3.250375	1.349509	0.0	2.0	3.0	4.0	5.0
Seat comfort	103904.0	3.439396	1.319088	0.0	2.0	4.0	5.0	5.0
Inflight entertainment	103904.0	3.358158	1.332991	0.0	2.0	4.0	4.0	5.0
On-board service	103904.0	3.382363	1.288354	0.0	2.0	4.0	4.0	5.0
Leg room service	103904.0	3.351055	1.315605	0.0	2.0	4.0	4.0	5.0
Baggage handling	103904.0	3.631833	1.180903	1.0	3.0	4.0	5.0	5.0
Checkin service	103904.0	3.304290	1.265396	0.0	3.0	3.0	4.0	5.0
Inflight service	103904.0	3.640428	1.175663	0.0	3.0	4.0	5.0	5.0
Cleanliness	103904.0	3.286351	1.312273	0.0	2.0	3.0	4.0	5.0
Departure Delay in Minutes	103904.0	14.815618	38.230901	0.0	0.0	0.0	12.0	1592.0
Arrival Delay in Minutes	103594.0	15.178678	38.698682	0.0	0.0	0.0	13.0	1584.0

Figure 2

From our preliminary analysis of the attributes, we found that across all services, inflight wifi service is the least satisfactory area, followed by ease of online booking and gate location, all of which have an average satisfaction level of below 3. Furthermore, overall satisfaction tends to be lower among adolescents and elder people.

2.4 Methods

In order to classify airline passengers as satisfied or unsatisfied, we tested a variety of Classification models. We tested five models: a Decision Tree Classifier as our baseline, a Random Forest Classifier, an XGBoost Classifier, a Gradient Boosting Classifier, and a Neural Net Model. Generally, we kept the architecture for our models fairly simple. We wanted the models to be as accurate as possible but also avoid overfitting on the training data. For all five of our models, we evaluated performance based on accuracy score.

For data preprocessing, we created two pipelines using sklearn's Pipeline class to ensure consistency in how new data would be handled. One pipeline was for numerical columns, and the other for categorical columns. We first made sure to eliminate any missing values. We used sklearn's SimpleImputer class to achieve this, and replaced all missing values with the median value for each column. Additionally, we treated all numerical 0 values as missing, since each numerical value was on a 1-5 scale. Thus, each 0 value was replaced with the computed Imputer value. We then applied a StandardScaler, which 'standardizes features by removing the mean and scaling to unit variance' (taken from skLearn's documentation). For our categorical columns, our pipeline encodes each feature into binary values using a OneHotEncoder.

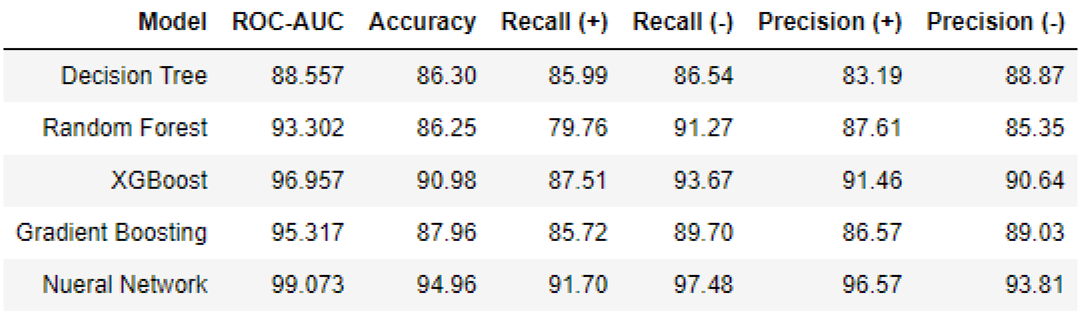
Next, we engineered new features that we believed could be statistically significant in our classification. For example, we created an *Average Rating* feature that averages all numerical columns given in the dataset (15 to be exact). We attempted to engineer additional features but were unsuccessful in finding statistical significance with any of them. Our models either did not use our engineered features at all, or the weights of these features were too insignificant to have a substantial impact on overall model performance. If we were to continue working on this analysis in the future, feature engineering would be a priority, as there are many possibilities beyond the features we created.

After data preprocessing and feature engineering, We ran initial tests on our models to keep track of baseline performances. After preprocessing and initial model testing, we began the hyperparameter tuning phase. We used a Five-Fold Cross-Validated Grid Search to determine the optimal hyperparameters for each model. We first tuned our Random Forest model. The hyperparameters we tweaked were the number of estimators, the max depth, and the criterion. After tuning, our final model had 30 estimators, a depth of 10, and used entropy as its criterion. Interestingly, 30 estimators and a depth of 10 were both the maximum possible in our grid search. For our XGBoost Model and Gradient Boosting Model, we tuned the number of estimators, max depth, and the learning rate (alpha). Our final XGBoost hyperparameters included a Learning Rate of 0.3, with a max depth of 3 and 30 estimators. Again, all of these values were the highest possible in our grid search. This was a concern as we thought our models were going to overfit on training data– in section 2.5 (results), we discuss model performance and issues with overfitting.

Experimentation in this project revolved around hyperparameter tuning and data cleaning. We tested the performance of our models with various hyperparameters such as learning rate. One issue that we ran into was how to tweak the architecture of our Neural Net. It was very easy to overfit this model by using too many parameters. We attempted to use dropout layers and normalization layers to improve the model's performance but struggled to achieve strong results. Ultimately, we concluded that a neural network may not be the best-suited architecture for the type of analysis we were conducting.

2.5 Results & Discussions

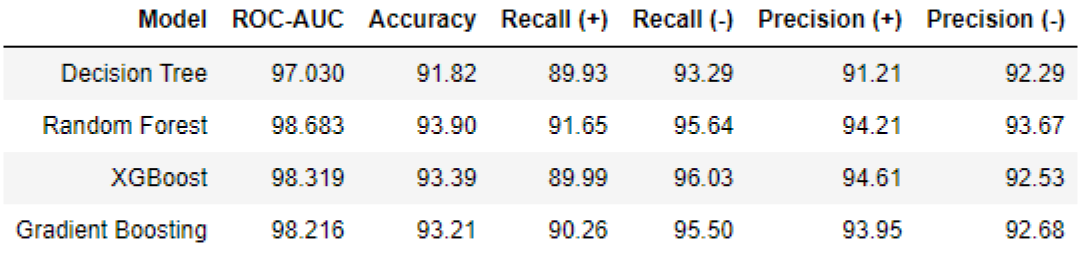
Given that our project is a binary classification problem, we are evaluating our machine learning algorithms using a standard confusion matrix, ROC curve, and scikit-learn's classification report function. Specifically, we have paid close attention to the accuracy of the model and the precision/recall trade-off. In our initial model testing, our baseline Decision Tree Classifier scored an accuracy of 86.30%, which we thought was quite high given the lower level of complexity of a Decision Tree model compared to our other models. The 4-layer Neural Network model performed the best with an accuracy of 94.95%. Figure 1 below shows the various performance metrics for each model in our initial testing.



Model	ROC-AUC	Accuracy	Recall (+)	Recall (-)	Precision (+)	Precision (-)
Decision Tree	88.557	86.30	85.99	86.54	83.19	88.87
Random Forest	93.302	86.25	79.76	91.27	87.61	85.35
XGBoost	96.957	90.98	87.51	93.67	91.46	90.64
Gradient Boosting	95.317	87.96	85.72	89.70	86.57	89.03
Nueral Network	99.073	94.96	91.70	97.48	96.57	93.81

Figure 3

After hyperparameter tuning using a grid search, our Random Forest model scored the highest, with an accuracy of 93.90%. Figure 2 below shows the various performance metrics for each model after hyperparameter tuning.



Model	ROC-AUC	Accuracy	Recall (+)	Recall (-)	Precision (+)	Precision (-)
Decision Tree	97.030	91.82	89.93	93.29	91.21	92.29
Random Forest	98.683	93.90	91.65	95.64	94.21	93.67
XGBoost	98.319	93.39	89.99	96.03	94.61	92.53
Gradient Boosting	98.216	93.21	90.26	95.50	93.95	92.68

Figure 4

Our models did not perform as well on test data. Despite the Random Forest model performing slightly better than the XGBoost model on training data, the XGBoost model turned out to be the strongest performer when classifying test data. Our tuned XGBoost model had an accuracy of 67.79% when classifying the test data. Furthermore, the model's recall score for the positive class was just 49.64%, while the recall score for the negative class was 82.0%. There were multiple possibilities for why this drop in model performance occurred. We hypothesized it was due to either our model overfitting on training data, or there being an imbalance between positive and negative classes in the dataset. We studied the dataset for differences and found that there was no significant imbalance between negatively and positively classified data (0 or 1). As mentioned in section 2.4 (Methods), we suspected our tuned models of overfitting on training data. The low recall score for the positive class reinforced this idea, and we believe that our models did indeed overfit on the training data.

We also analyzed feature importances for our tuned XGBoost model. We found that the class a passenger is flying in (first class, business class, economy, etc) and inflight service were the two most heavily weighted features when determining if a customer is satisfied or unsatisfied, with class having more than twice the weight of inflight service. Intuitively, this makes sense, as first class and business class passengers have preferential treatment and are much more likely to be satisfied with their flight experience.

Other models on Kaggle tended to score higher than ours. Many other studies resulted in test accuracies of >95%. Furthermore, some of these models did not use any feature engineering and lacked hyperparameter tuning, which led us to believe that our tuning may have resulted in worse performance than if we had not changed hyperparameters. This is entirely possible if we selected poor values for our hyperparameters, or if we tweaked the wrong hyperparameters.

In phase 1 of our project, we found that inflight service and online booking were the two features that had the lowest average satisfaction scores. A conclusion we made is that airlines should prioritize Inflight Service, as it not only scores low on customer satisfaction surveys, but it is also one of the best predictors for customer satisfaction. We believe that these business implications are the most important take-aways from our model as it provides airlines actionable insights into how they can improve their customer experience.

2.6 Conclusion

The goal of this project is to figure out which factors most affect customer satisfaction on flights. By doing so, airlines will be able to effectively improve their flight experiences. This will allow them to gain a competitive edge over rivals, as well as attract more customers into the market by making flying a more appealing mode of transportation. To this end, the project has succeeded, finding that passenger class, the quality of inflight service, the length of arrival delays, the quality of inflight entertainment, and the quality of inflight Wi-Fi are the factors that most affect flight experience.

Our analysis involved applying 5 different classification models to try to predict customer satisfaction for airlines. Our 5 models included a Decision Tree Classifier, a Random Forest Classifier, a Gradient Boosting Classifier, an XGBoost classifier, and a 4-layer Neural Network. After tuning hyperparameters of each model, we found that a Random Forest Classifier can best predict a customer's satisfaction. Using this model, we were able to predict customer satisfaction with 68% accuracy and 80.4% AUC. More importantly, we discovered that the most important factors contributing to customer satisfaction were passenger class, inflight service, arrival delays, inflight entertainment, and inflight Wi-Fi.

There exist very actionable results from this project. While it is difficult for airlines to act upon the most important factor, passenger class, the next four factors can be translated into clear priorities for airlines to improve satisfaction. For instance, airlines can try improving the quality of inflight entertainment and Wi-Fi to see if the increased customer satisfaction is worth the cost. Airlines can also experiment with modifying their training programs for airplane staff, to see if they can effectively increase the quality of inflight service. They can also look for ways to mitigate the impact of flight delays on customer satisfaction, given how much of an impact they make. Finally, airlines could conduct further analysis on customer acquisition costs and customer lifetime value for different passenger classes. This will provide insight into if more satisfied passengers in higher classes provide more revenue. Overall, the project is successful in figuring out which improvements in flight experience should be prioritized to improve customer satisfaction.

3 Appendix

No other analytical tools were used for the project beyond Python. For the coding portion of our project, see *COMM4522_FinalProject.ipynb*.