Predicting Airline Passenger Satisfaction

Brandon Habschied and Geordy Aponte

Responsibilities:
Brandon Habschied: Classifier, Regressor, Clustering Algorithm, Transformation, Outlier, and k-fold
Geordy Aponte: Classifier, Clustering Algorithm, Statistical Test, Outlier, and Advance Method

Abstract:

This study investigates factors that contribute to customer satisfaction in the airline industry. The dataset includes responses from customers on 20 different features of the service provided, as well as their overall satisfaction rating. We used a variety of machine learning algorithms, like Random Forest Classifier, XGBoost, and K-Means to achieve our goal of creating a model which can be used to predict customer satisfaction and a secondary goal of determining which features have the highest contribution to customer satisfaction. Using k-fold, we were able to create a model which predicted customer satisfaction with an accuracy of 96% and from this model, it was found that the most important features for predicting customer satisfaction were Customer Type, Online Boarding, In-flight Wifi Service Class, and Check-in Service. These findings have practical implications for service providers, as they can use the results to tailor their services to meet the specific needs and preferences of their customers.

Table of Contents

Introduction:

Customers' satisfaction and loyalty are essential factors in the airline industry's success, and the COVID19 pandemic has reinforced this fact. It is essential for airlines to understand what customers expect and value, identify their pain points, and tailor their services to satisfy their needs. This study aims to do as such, to create a model that can best predict customer satisfaction. Individual consumers have been surveyed and asked to rate their experience across multiple travel factors, with a final declaration of satisfaction.. With a model created from this dataset, airlines can focus their efforts on specific areas of improvement to ensure satisfactory customer satisfaction, thus creating brand loyalty and future success.

We will be using a dataset found on Kaggle which has a sample size of almost 130,000 with 23 different features with the individuals rating of the features and providing a satisfactory or unsatisfactory mark. Features include gender, ease of check-in, departure and arrival delay, flight time, and much more which can be used to determine whether each passenger was satisfied. Because the dataset is so large, k-fold methods can be implemented here to split the training data many times to create a very accurate training set.

With these interests in mind, we predict that the Class of the ticket will affect customer satisfaction e.g., premium tickets like business class will provide higher customer satisfaction as opposed to economy. In addition to ticket class, we predict that in-flight Wi-Fi service, ease of check in, seat comfort, and cleanliness will be the most important features for determining satisfaction while age will have little effect.

Methodology:

Classifier: Because our target variable “Satisfaction” is binary, we had a number of classifiers to choose from before we decided to use the Random Forest Classifier (RFC). The first reason we chose RFC is because it is an ensemble of decision trees and for our binary target variable, a decision tree is suitable but can be inaccurate. RFC ensembles multiple decision trees to increase accuracy by making it less prone to overfitting. Also, with the vast size of this dataset, we needed something which would be efficient and RFC could provide this because it has the ability to deal with large and high dimensional datasets.

Regressor: With our target variable "Average Score" being numerical, we will be using a regressor instead of a classifier. We chose to use XGBoost to create a model that uses decision trees to train our model. XGBoost is a powerful algorithm that is known for its accuracy and speed, and it can handle both small and large datasets with high dimensionality. Additionally, XGBoost includes regularization techniques to prevent overfitting and improve generalization.Using XGBoost provided us fast and effective results compared to other regressors that we tried such as KNN, SVM, and Gradient Boosting Regressor.

Clustering: We will use this method to create clusters of passengers with similar satisfaction levels and other attributes. This will allow us to gain insights into which features are most closely related to passenger satisfaction, as well as identifying groups of passengers with similar characteristics. As such, for easier visualization, we used a dendrogram from hierarchical clustering to cluster our data. This method provides easy visualization of the clusters and allows us to see which samples are grouping together to help provide us insight as to which instances have similar experiences.

Feature Engineering: This dataset from kaggle is strictly categorical, with a mix of both numbers and words; because there was a mix, we needed to do a large amount of preprocessing in order to clean the data. Much of what we did required manual conversions in order to have the data match, and then be able to run it in our models. The specifics will be discussed later in the paper.

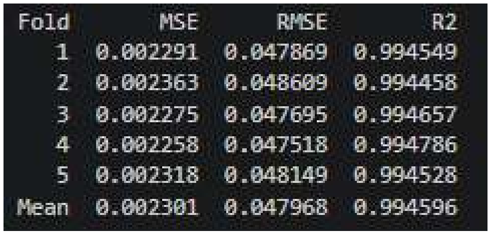
Anomaly Detection: For our anomaly detection, we decided to use Local Outlier Factor (LOF). The reason for this is because we want to find passengers who have significantly different ratings compared to others in their local neighborhood. For example, if a passenger gives very low ratings for all aspects of their flight experience, while other passengers with similar characteristics rate their experience much higher, then that passenger may be considered an outlier.

Advanced Modeling: We decided to use XGBoost as well for a regressor for our target variable of customer satisfaction. The reason for this is actually very similar to our reasoning for choosing RFC. XGBoost also uses an ensemble of decision trees in order to create a more accurate model and it is also efficient with large amounts of data.

Results:

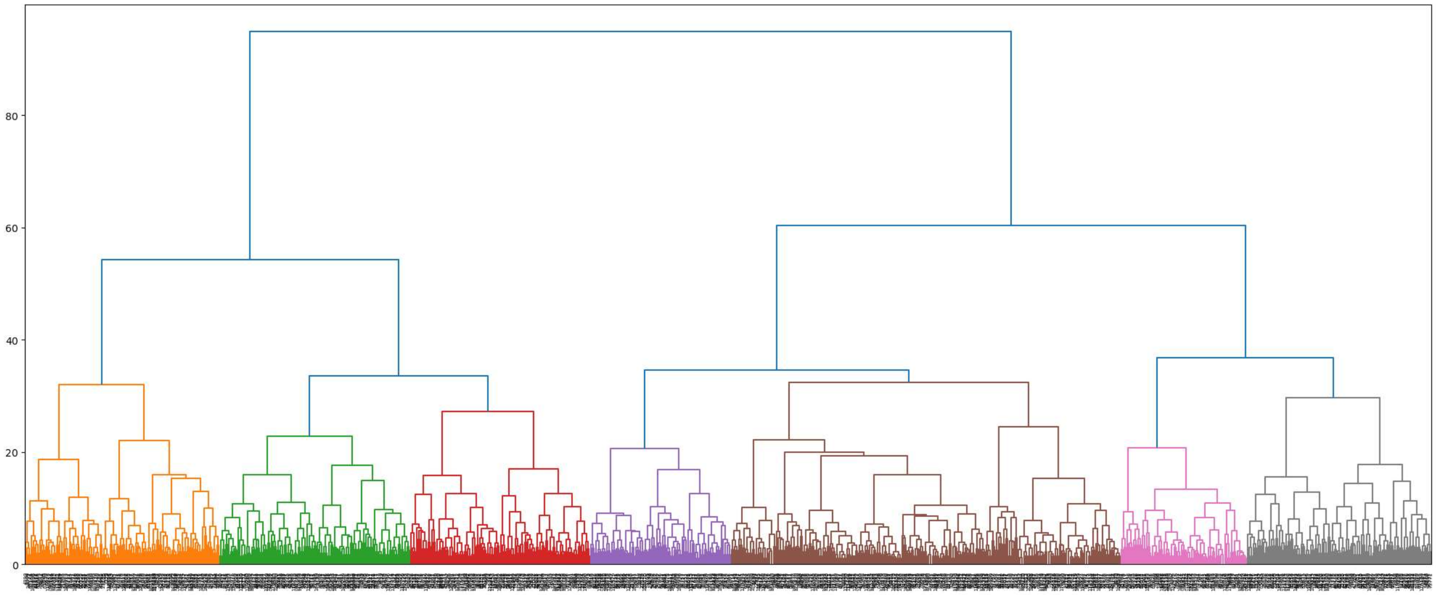
Classifier: After performing a grid search for the best parameters, we were able to tune our RFC to produce a very accurate predictive model that could predict customer satisfaction. With the original data set, we found that the RCF created a model that was 96% accurate, after implementing k-fold. From this model, we also determined feature importance and as such, we found that the top five most important features for this model were Online Boarding, In-flight Wifi Service, Type of Travel, Class, and Average Rating in that order.

Regressor: To begin our XGBoost Regressor, we separated the target variable "Average Rating" from the dataframe. After splitting the data with k-fold, we were able to determine the Mean Standard Error (MSE), Root Mean Standard Error (RSME), and Coefficient of Determination (R2) for each of the folds in the k-fold. With each fold, we then took an average of each measurement for each fold and achieved a desirable low average MSE of 0.0023 and RSME of 0.047 while achieving a high average R2 of 0.9945 which indicates that our regression model was a very good fit to the data, potentially overfit to our data.

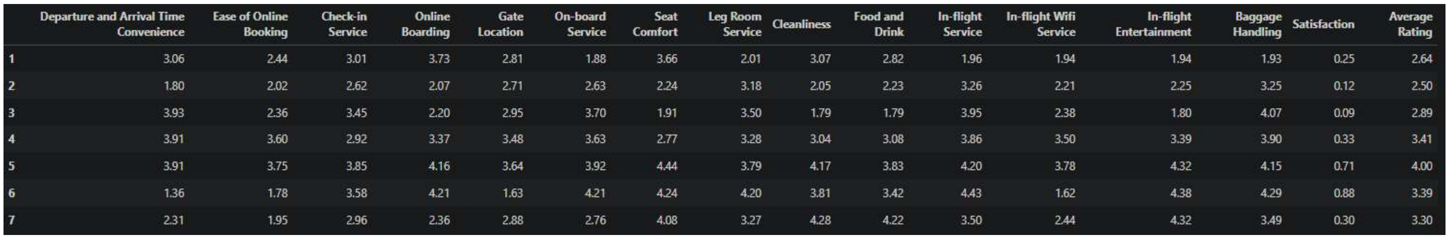


```
Fold       MSE      RMSE        R2
   1  0.002291  0.047869  0.994549
   2  0.002363  0.048609  0.994458
   3  0.002275  0.047695  0.994657
   4  0.002258  0.047518  0.994786
   5  0.002318  0.048149  0.994528
Mean  0.002301  0.047968  0.994596
```

Clustering Algorithm: In an attempt to discover patterns, we implemented hierarchical clustering using linkage on a random sample of 1000 instances. The reason we chose to reduce our clustering to only 1000 samples was due to the fact that our computers would crash during computation of samples of significant size. After plotting our dendrogram, we visually identified approximately 7 clusters and set our color threshold to 33. Doing so gave us the following dendrogram.



After plotting our dendrogram, we found the sizes of our 7 clusters and their associated centroids by taking an average of each instance within that cluster.



	Departure and Arrival Time Convenience	Ease of Online Booking	Check-in Service	Online Boarding	Gate Location	On-board Service	Seat Comfort	Leg Room Service	Cleanliness	Food and Drink	In-flight Service	In-flight Wifi Service	In-flight Entertainment	Baggage Handling	Satisfaction	Average Rating
1	3.06	2.44	3.01	3.73	2.81	1.88	3.66	2.01	3.07	2.82	1.96	1.94	1.94	1.93	0.25	2.64
2	1.80	2.02	2.62	2.07	2.71	2.63	2.24	3.18	2.05	2.23	3.26	2.21	2.25	3.25	0.12	2.50
3	3.93	2.36	3.45	2.20	2.95	3.70	1.91	3.50	1.79	1.79	3.95	2.38	1.80	4.07	0.09	2.89
4	3.91	3.60	2.92	3.37	3.48	3.63	2.77	3.28	3.04	3.08	3.86	3.50	3.39	3.90	0.33	3.41
5	3.91	3.75	3.85	4.16	3.64	3.92	4.44	3.79	4.17	3.83	4.20	3.78	4.32	4.15	0.71	4.00
6	1.36	1.78	3.58	4.21	1.63	4.21	4.24	4.20	3.81	3.42	4.43	1.62	4.38	4.29	0.88	3.39
7	2.31	1.95	2.96	2.36	2.88	2.76	4.08	3.27	4.28	4.22	3.50	2.44	4.32	3.49	0.30	3.30

Statistical Test: Chi-Squared statistical testing was used to determine the independence of satisfaction compared to the other features collected in the survey. Recall that if a p-value of 0.05 or less is found, then this would indicate that there is a dependence for these variables. What we found was that all of the available features had a p-value less than 0.05.. What this indicates is that customer satisfaction is dependent on all of the features customers' were surveyed on and as such these features all serve as a good predictor when creating a predictive model, which is what we've done with our predictive models.

Scaling/Transformation: The dataset from kaggle is strictly categorical, with a mix of both numbers and words; because there was a mix, we decided to clean the data before beginning any methodology selection. To begin with, we convert our data that were words to numerical representations: gender , customer type, type of travel, and satisfaction was turned into a binary variable (0 for male and 1 for female, 0 for first time flyers and 1 for returning flyers, 0 for business travel and 1 for personal travel, and 0 for not satisfied and 1 for satisfied) and the class variable determine what class type the passenger purchase and we changed this to 0 for economy class, 1 for economy plus, and 2 for business class. Missing data was only found in the arrival variable and this was filled in with zero as the number of missing data was small so zero would not have been significant. We also debated the idea of splitting gender into two different columns but because we had a large number of features, we did not see it necessary to split this variable. As such, we were able to transform the dataset into usable data that our models could read.

Outlier/Anomaly Detection: Using LOF, we found that 2% of the sample was an outlier. The presence of outliers in a dataset can have a significant impact on the analysis and modeling results, potentially leading to biased estimates, incorrect predictions, and reduced model performance. As such, these data points were removed and then RCF was run again. Luckily for us, the results of our model remained the same at 96% and feature importance remained in the same order.

k-fold: k-fold was used in both the classifier and the regressor. This dataset benefits from k-fold in many ways as the dataset is so large. With k-fold, we get more robust data as we are able to train and test over a large number of sub-datasets. As a result, we have less overfitting and this is why we are able to achieve such high accuracy scores for RFC and XGBoost.

Advance Method: While XGBoost was used as our regressor, it was also used as our advanced method as well. After using RFC, we used XGBoost in an attempt to create a more accurate model, which was already a challenge because of how accurate it was as RFC is an ensembled classifier. XGBoost created a model that was as accurate as our RCF model, again with a 96% accuracy. However, feature importance was significantly different from what was found with RCF. With XGBoost, we found that the top five most important features were Online Boarding, Type of, Travel, In-flight Wifi Service, Customer Type, and Class in that order.

Actionable Insights:

We found that the features with the most impact to customer satisfaction were Online Boarding, In-flight Wifi and Type of Travel. As such, airliners should focus on these areas, specifically Online Boarding and In-flight Wifi. We exclude Type of Travel because the airline cannot control if a customer is using the airline for business purposes or for personal reasons. As such, we will focus on Online Boarding and In-flight Wifi.

For Online Boarding, it would be best to make this as seamless and intuitive as possible. This can be done by creating simple screens for the customer with milestones when they arrive at the airport. For example, when the customer arrives at the airport, they can open an app on their phone to start the process by letting the app know they've arrived at the airport. The app can then guide the customer to the proper gate that they will be departing from. Afterwards, they will see their boarding pass and know quickly where their seat is. Removing ambiguity from the check in process can ensure a positive response from the customer.

For In-flight Wifi, we have a few more options. The performance of In-flight Wifi can be improved, as some technology is advancing and ensuring both fast and reliable internet connection can dramatically improve customer experience with this service. At the moment, there is a cost to use In-flight Wifi but if we give it to the customer as a free courtesy, they will be happy they get to experience this "luxury" in the air. The only issue with making this service free is that with more users, there will be more network congestion, and thus poorer performance so a premium wifi network could be offered to customers in addition to the free option but would still require performance upgrades.

Discussion and Conclusion:

We believe our biggest obstacle was the actual dataset itself. As you may have noticed, we aimed at using methods which were designed for large datasets in order to try to reduce runtime. However, the dataset we used has over 3.1 million data points and as a result, we experienced time issues with running some methods. For example, we wanted to create a dendrogram using all of the data points but this was causing the kernel to die as we were running out of RAM on our computers. To resolve this, we had to use a fraction of the data in order to be able to run this. Additionally, we decided to change our regressor because it also had issues running as well. We've changed it a few times but we wanted to use SVM but we were not able to complete running the model; after about 60 minutes, we quit and ran it again and then after 90 minutes, we realized that this was probably not going to work for us. We then began to look at other regressors which aligned with the results we wanted and decided to use XGBoost. We saw that large dataset as an advantage just because the data should have been normal based on its sample size but didn't take into account the stress it would cause our machines as we've been familiar working with datasets that are magnitudes smaller in size.

Dataset:
https://www.kaggle.com/datasets/mysarahmadbhat/airline-passenger-satisfaction?select=airline_passenger_satisfaction.csv

Dataset Dictionary:
https://www.kaggle.com/datasets/mysarahmadbhat/airline-passenger-satisfaction?select=data_dictionary.csv