Project Report - “Happier Passengers“

Problem Statement and Goal

Competition on passenger satisfaction is one of the many dimensions in which airlines compete. Based on this, how could an airline approach the question of increasing passenger satisfaction given its finite resources?

In order to achieve this, one approach is to:

(a) quantify and rank the impact of various factors of flight quality on passenger satisfaction, and

(b) based on this ranking, analyze how the flight process could be tailored to maximize passenger satisfaction in a cost-effective manner.

The scope of this project is to implement only the first part of the above approach: statistically quantify and rank the factors that impact passenger satisfaction and forward the analysis to relevant parties with suggestions.[1]

In terms of implementation, given a dataset of airline passenger surveys, I will first create a model of customer satisfaction [target variable] via decision trees and then perform a sensitivity analysis on factors of flight quality.

The rest of the report will follow the analysis charted above. It will continue with an exposition of the data and dataset, followed by feature extraction, modelling and analysis. The report will end with a set of recommendations and directions for future research.

[1] Taking a wider view, the analysis in this project is intended to serve as a foundation to:

(a) [In the Short-Term] Understand the types of flights that are vulnerable to passenger dissatisfaction and prioritize their further analysis

(b) [In the Long-Term] Deploy large-scale policy changes to increase passenger satisfaction given a budget constraint or impose cost-cutting measures on factors that do not impact customer satisfaction. Some potential steps to this outcome could include:

(i) Deploying small-scale experiments to validate the project’s quantitative results in the field.

(ii) If validated, perform a more extensive cost-benefit analysis of the proposed improvement.

Dataset Description

The dataset is an airline passenger survey and is taken from Kaggle.[2] While the original data source is not mentioned in the link, the author of the dataset points to another dataset in Kaggle which claims that the data is from a US airline passenger satisfaction survey.[3]

The dataset contains 129,880 surveys in total, out of which 103,904 (80%) were separated out in Kaggle as training and the rest, 25,976 (or 20%), as test. [If I did this research again, I would randomize these myself or check if the distributions of both sets were similar]

For passenger surveys, following are the variable names, a brief note (description, contents, etc) and their classification in my analysis:
Continuous Variables: Age, Flight Distance, Arrival Delay in Minutes, Departure Delay in Minutes.
Categorical Variables (numerical): Inflight wifi service [1 (worst) to 5 (best), 0 is N/A: also applies to following], Departure/Arrival time convenient, Ease of Online booking, Gate location, Food and drink, Online boarding, Seat comfort, Inflight entertainment, On-board service, Leg room service, Baggage handling, Checkin service, Inflight service, Cleanliness.
Categorical Variables (string): Gender [Female, Male], Customer Type [Loyal, disloyal], Type of Travel [Personal, Business], Class [Eco, Eco Plus, Business]. satisfaction [neutral or dissatisfied, satisfied]

Exploratory Data Analysis

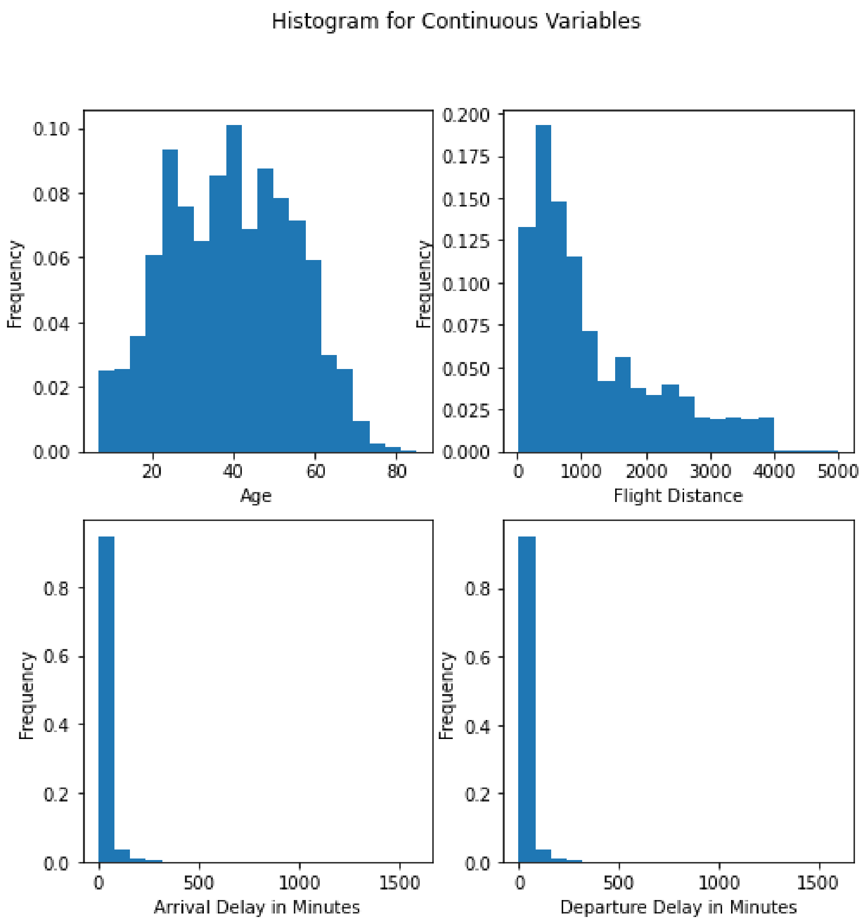
Please note that all exploratory analysis is done on the training data. You could also read this section accompanied by its Jupyter Notebook: 'Happier Passengers - Part 1 - Objective - Data - EDA.ipynb'.

---

[2] https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction
[3] https://www.kaggle.com/johnddddd/customer-satisfaction

Explore Continuous Variables: Distribution and Notes

Distribution



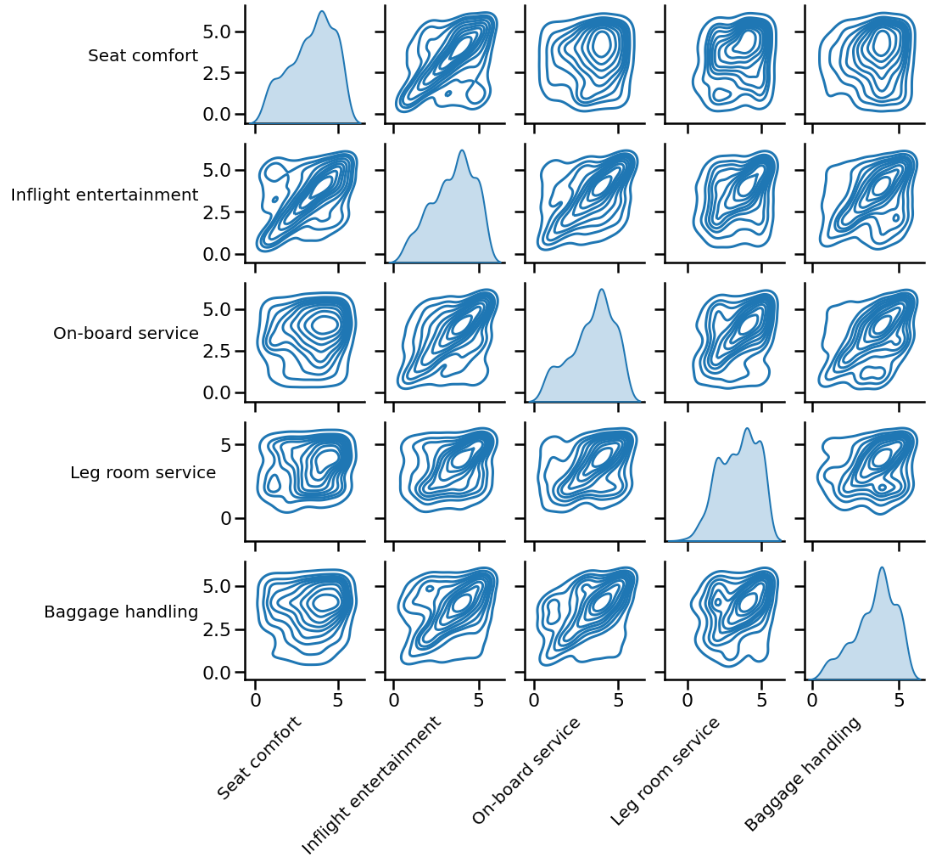
Notes

1) Departure and Arrival distributions seem to have most of their mass at or around zero, however tend to also have large outliers. We might want to keep this in mind in the future. We could try to come up with a feature that normalizes them.
2) Flight distance could indicate generally two types of flights - short distance (<900 miles?) and long-distance (≥900?). This could be important since it might indicate different types of flights and different behaviors by their respective passengers.

Explore Categorical Numerical Variables

Distribution

Below is a pairwise kde-plot of a subset of the 14 categorical numerical variables in the dataset. While only 5 variables are presented here due to limited space, a pair plot of all variables can be found in the "Happier-Passengers - Part 1…" notebook.

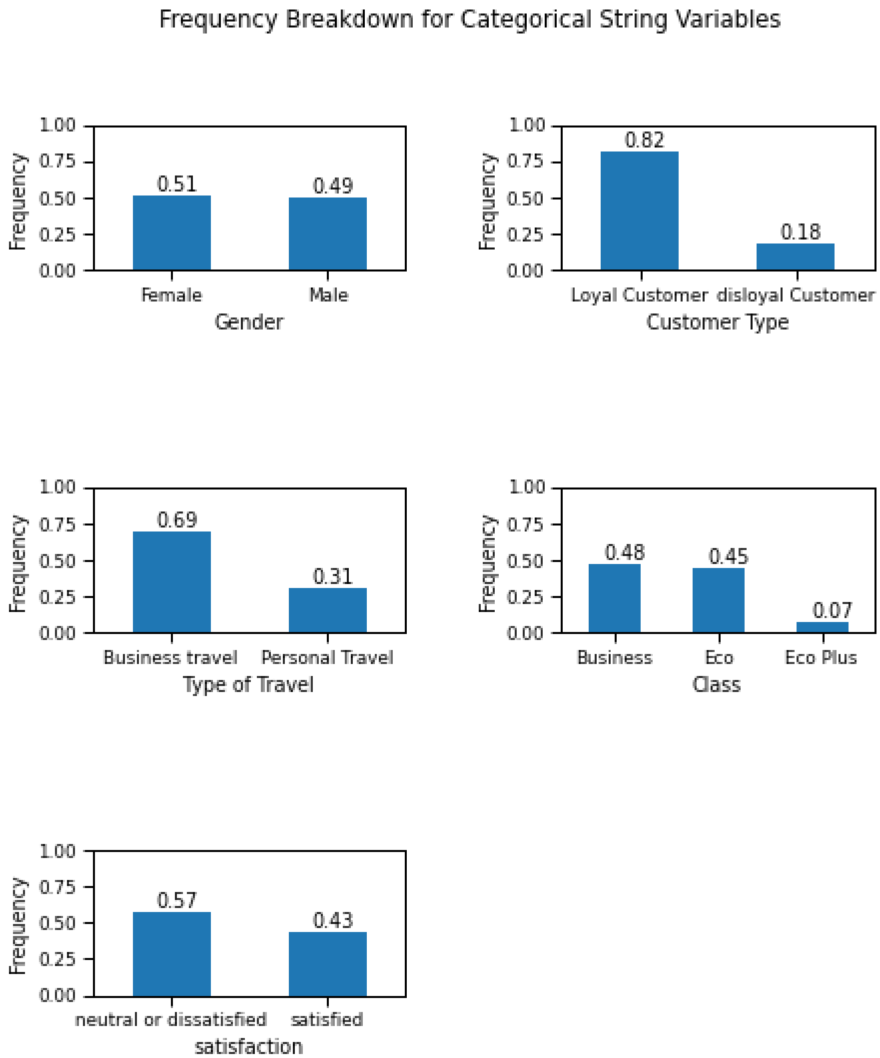


Notes

1) Most distributions look unimodal with a mean above 3 (the mean of uniform excluding 0's.) This may bring up questions as to whether or not there might be enough room to be able to measure improvements in services.
2) Bimodal distributions generally indicate there are two different populations (examples here: Food & Drink, Ease of Online Booking, Leg room service) and could disappear when looked from a different population conditioning (e.g. Class, Age) These in turn can help indicate pockets of demographics where impacts of an improvement could be higher.
3) The total number of samples with any zeros (pointing to either non-answered questions or unavailable services) is around 8%. (An analysis could be done to see whether this data was relatively similar to the remaining data)

Explore Categorical String Variables

Distribution



Notes

1) One important thing to note is that the data seems to be balanced with respect to ‘satisfaction’, the target variable.
2) Among the remaining variables the gender ratio looks normal while the Business class vs Eco class travel ratio seems not to be. Given the seating in most passenger planes, I would expect the Business/Eco ratio to be much lower. This alone could indicate a sampling bias in the survey - however we have no information regarding the sampling methodology.

Data Wrangling

In this section, I will prepare the data for the decision tree analysis. (As before the section also has a companion file: “Happier Passengers - Part 2 - Data Wrangling.ipynb”.)

As a first step, I assigned numerical values to categorical string variables. All except one of such variables are binary and have been assigned 0’s and 1’s. For the only categorical variable with three values, we tried to make it such that the outcome would be more or less ordinal (i.e. in general prices paid Eco is less than Eco Plus and price paid for Eco Plus is less than Business) Given that the plan is to use an ensemble of tree, it made more sense for the splits to result in more easily interpretable sets.

Next, I dropped the surveys that contained any NAs anywhere and 0s in the categorical numeric variables. Justification here was mostly based on the observation that 0’s would introduce a non-ordinal element into the 1-to-5 mix possibly hindering a sensible function of the information gain step of the decision tree approach.

After the drops, resulting data sizes are as follows: (Note that sample loss is 8.2% for the Training Set and 8.4% for the Test Set - relatively similar figures, another data point in favor of the initial Train/Test split being random.

	Training Size	Test Size
Before Drop	103,904	25,976
After NA Drop	103,594	25,893
After 0 Drop	95,415	23,789

The next step is feature engineering where I will make an attempt to combine some features in the data to generate a more informative feature.

Feature Engineering

The main goal here is to generate a feature that is more informative than its constituents and able to be precisely generated by a decision tree approach. [You can follow the text with the companion notebook: “Happier Passengers - Part 3 - Feature Engineering.ipynb”

In the data, although the delays in arrival and departure times could be useful features that impact passenger satisfaction, another idea for a feature would be to normalize them by the distance traveled. The core idea here is that the same amount of delay could be more tolerable in a long flight than a short one.

Following the above idea and assuming a more or less constant flight velocity, normalizing the delays with distance traveled would let us calculate how long the delay was with respect to the total expected flight time.

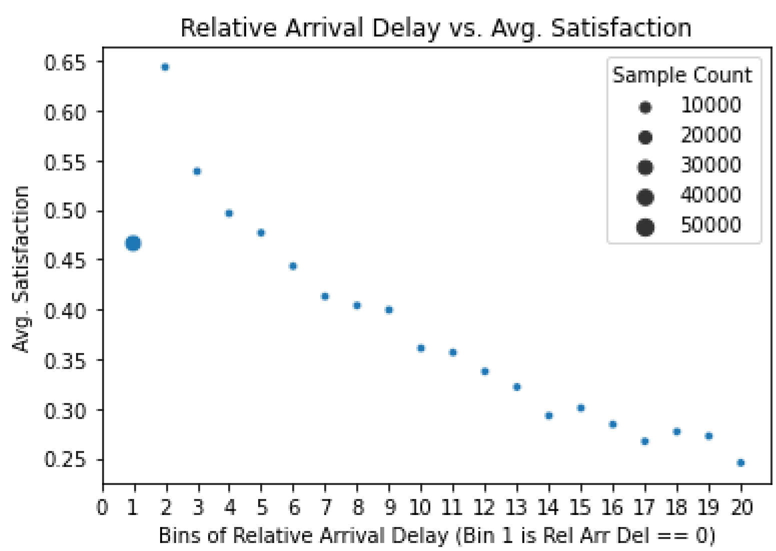
Relative Arrival Delay = *Arrival Delay in Minutes* / *Flight Distance*
Relative Departure Delay = *Departure Delay in Minutes* / *Flight Distance*

Once done, we find that the correlations of these “relative” variables to ‘satisfaction’ are higher in absolute terms than the simple delay variables. This could indicate a justification to include the relative delay variables.

Correlation of ‘satisfaction’ vs	Arrival	Departure
Relative Delay	-0.088	-0.084
Delay in Minutes	-0.058	-0.051

At this point, I choose to include ‘Relative Arrival Delay’ (corr: -0.088) in the model over the ‘Relative Departure Delay’. I argue that the two variables are pretty similar and in a tie, passengers might care more about arrival delays than departure delays.

Finally, this new variable itself could use some exploratory data analysis:



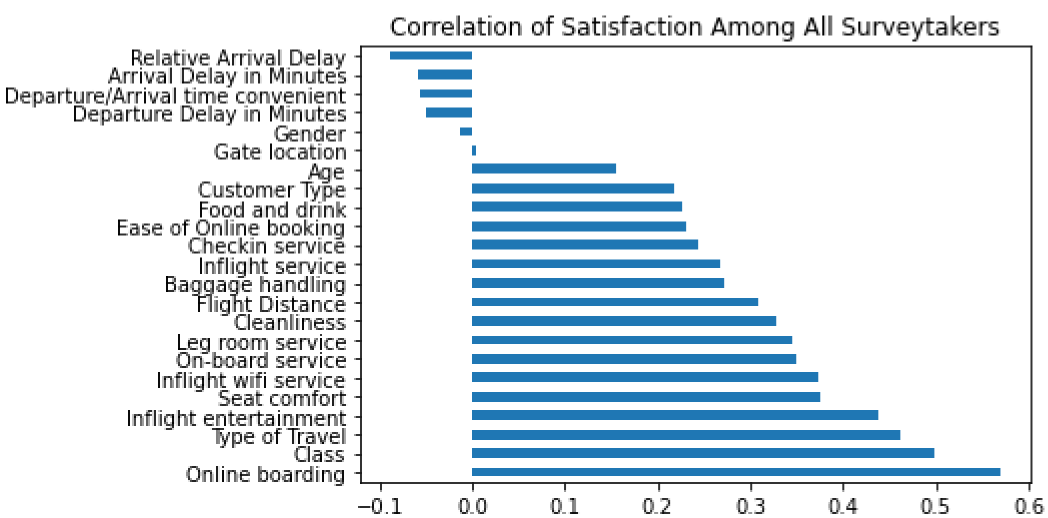
From the above graph, we note that for flights with delays, average ‘satisfaction’ tends to drop with ‘Relative Arrival Delay’. One interesting point to note is that the average ‘satisfaction’ of “less-delayed” flights seems to be higher than the general average ‘satisfaction’. It might be interesting to see if there might be a mechanism behind that. Nevertheless, given the correlation and this structure, ‘Relative Arrival Delay’ could be an interesting addition to the model - and one that might be worth working on the minimize [yet, out of the purview of this decision tree approach at this time]

Modeling

This section will be concerned with choosing a decision tree model that best represents the survey responses. The accompanying code is in “Happier Passengers - Part 4 - Modeling.ipynb”.

Investigate Input Correlations with ‘satisfaction’

Let’s first take a look at the correlation of ‘satisfaction’ with the model inputs to have a general idea about the relationships and see, to what extent, how the final model may reflect these.



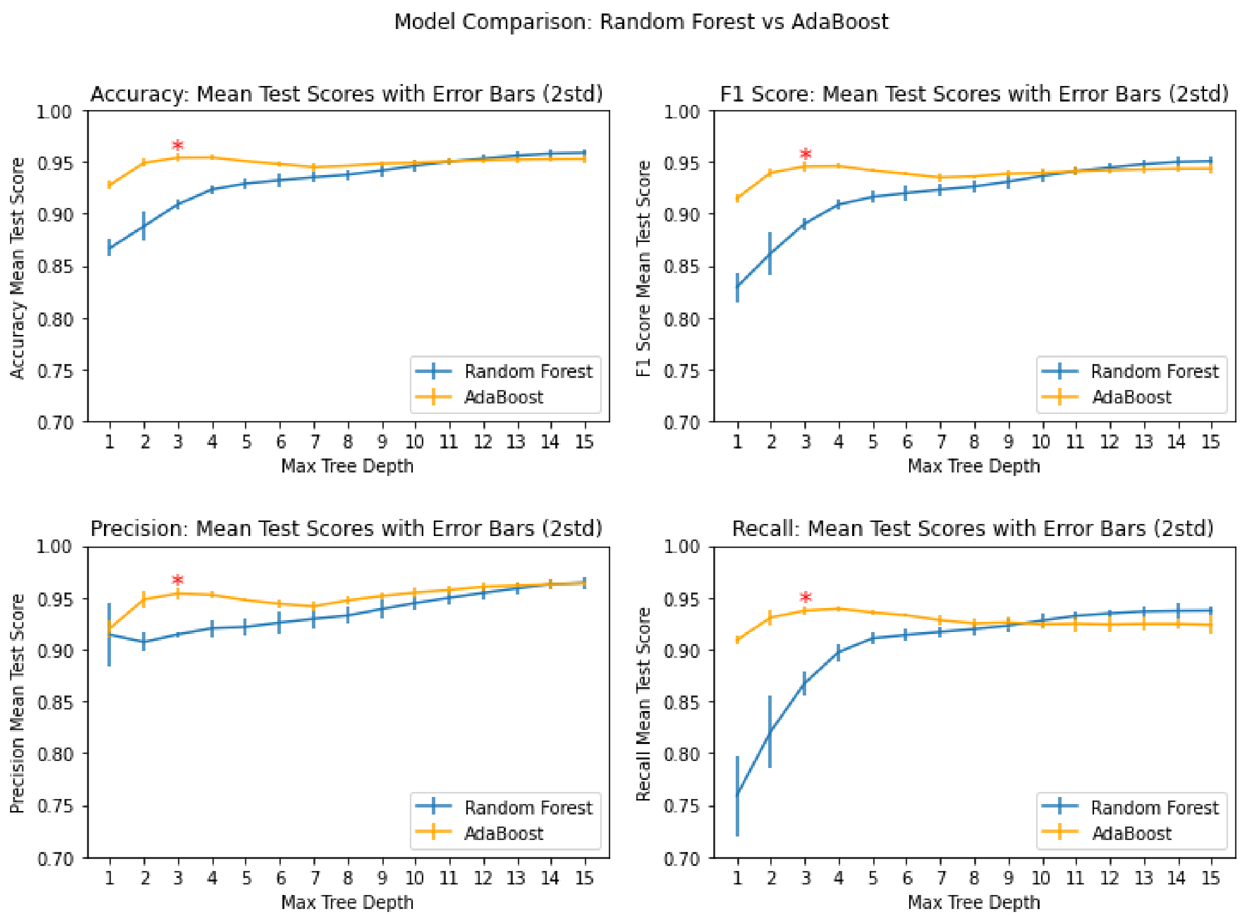
In the above, it’s interesting to note ‘Class’ and ‘Type of Travel’ as having high correlation with ‘satisfaction’. Given how these categorical variables are mapped in an ordinal-like sense, we can see that higher satisfaction is related to Business Travel and Business Class travelers. Among other factors, ‘Gender’ and ‘Gate location’ seem to have no correlation with ‘satisfaction’ - which makes sense a-priori.

Model Selection

As for the candidates to model passenger satisfaction , the competing approaches will be: Random Forest and AdaBoost. For both of these models, different hyperparametrizations will be trained over a 5-fold cross-validation on the training set and tested on the test set.

For the Random Forest model, the hyperparametrization will be over the maximum tree depth. Range will be from 1 to 15.
The hyperparametrization of the AdaBoost model will also take a similar approach. The base estimator will be a decision tree and the maximum tree depth here will be from 1 to 15.
For each model, the number of trees in the ensemble will be 100.

Below is the model comparison:



Although the models seem to behave similarly in terms of scores at higher Max Tree Depth values, we see that the AdaBoost with lower Max Tree Depth values also shows a similar performance to the higher end models. Given the parsimony of these low Max Tree Depth AdaBoost models, I choose an AdaBoost model - particularly the one with depth 3 - noted by the star in the above figure.

Final Model

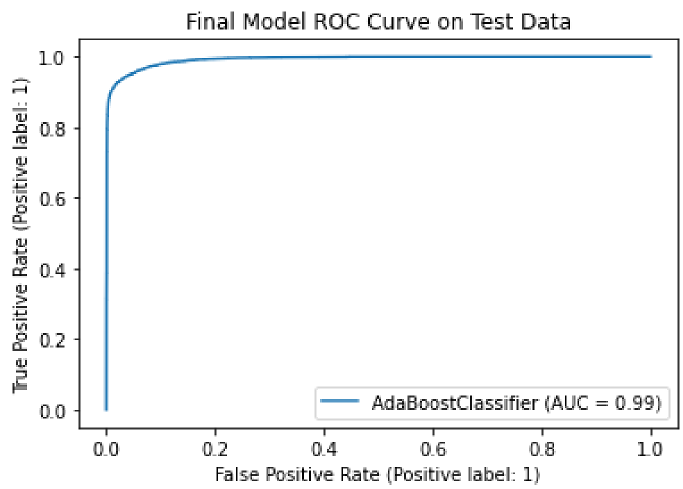
Once again, the final model is the following:

AdaBoost model based on a decision tree classifier with a max tree depth of 3 and 100 estimators.

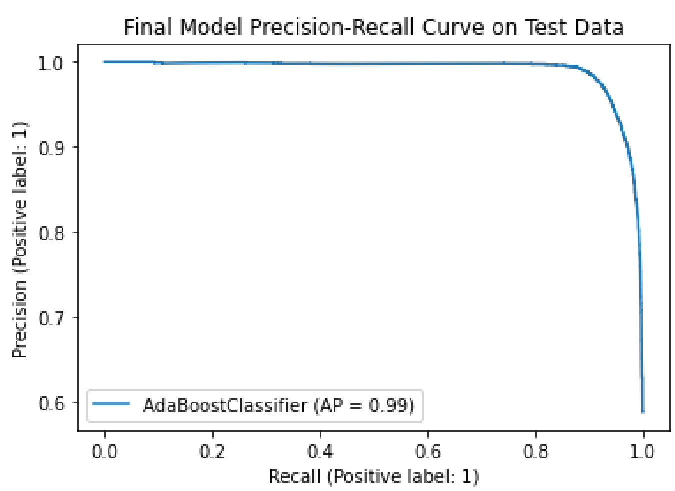
The model trained on all of the training dataset and applied on the test set has the following scores,

Final Model (AdaBoost) Scores on Test Data	
Accuracy	0.956
F1 Score	0.949
Precision	0.956
Recall	0.941

ROC Curve,

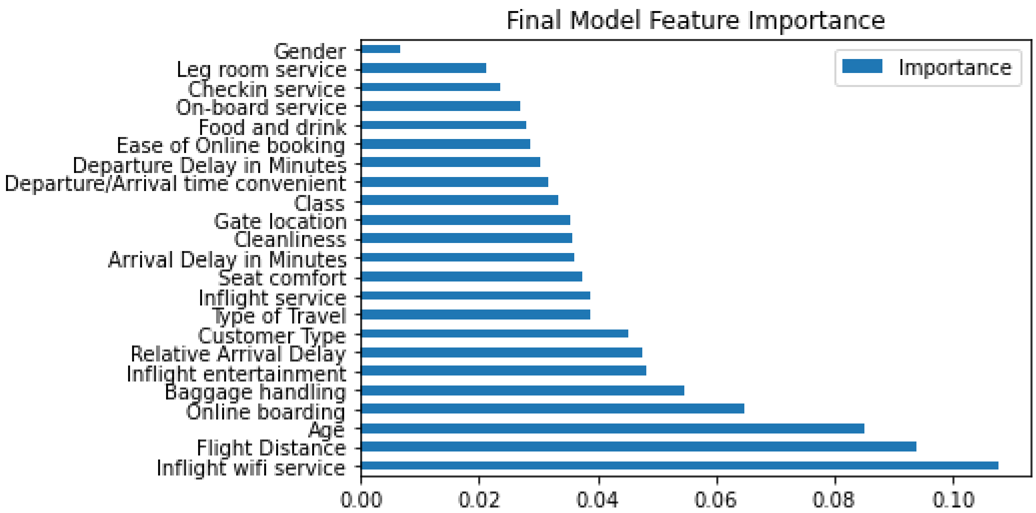


and the Precision-Recall Curve:



Both curves above show a great fit. [Perhaps, too good]

Finally, I close the model section with the feature importance graph.



Regarding the above figure, observe that among the top five "most important" variables, Inflight wifi service, 'Online boarding' and 'Baggage handling' are survey questions. This is good news as the airline has some direct control over these flight quality factors.

On the other hand, while the remaining variables in top five ('Flight Distance' and 'Age') are not directly controllable, observe that they moved up in ranking when compared to the prior correlation figure. This gives one hope that there might be a more interesting story there. Namely, it might mean that decision tree structure has found some structure in the data that it mined and that this model could be split into sub-models perhaps based on these variables, in particular 'Flight Distance'.

Sensitivity Analysis

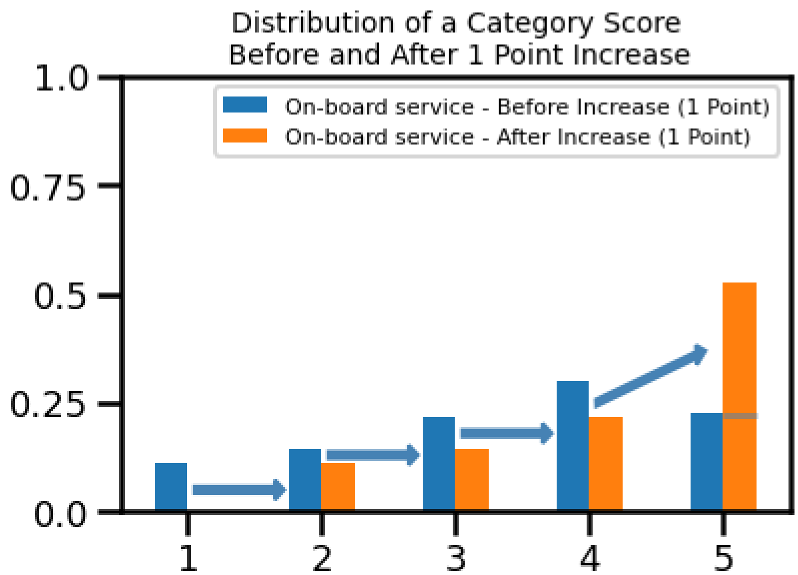
This section will perform a sensitivity analysis on factors of flight quality on passenger satisfaction. The code is in "Happier Passengers - Part 4 - Modeling.ipynb" starting with the "Quantify and Rank Factors…" section.

The sensitivity analysis will be performed by perturbing the responses of passengers to survey questions and noting the new satisfaction rates.

The exercise can be thought as asking varying versions of the following question:

If all the passengers who responded to, say, 'On-board service' with a 3 (out of 5) were to increase their scores to 4, how would this group's average satisfaction change?

Given the responses to these questions across all survey categories and responses 1 through 4, one can estimate the change in average satisfaction for each category for a variety of scenarios. In this exercise, the focus will be on the impact of each category score (1...4) increasing by one. An example is below - note that the distribution is shifting right by one point (average score increases from 3.39 to 4.16)



Baseline average satisfaction for passengers in each category and score are as follows.

```
Average Satisfaction per Category and Score
```

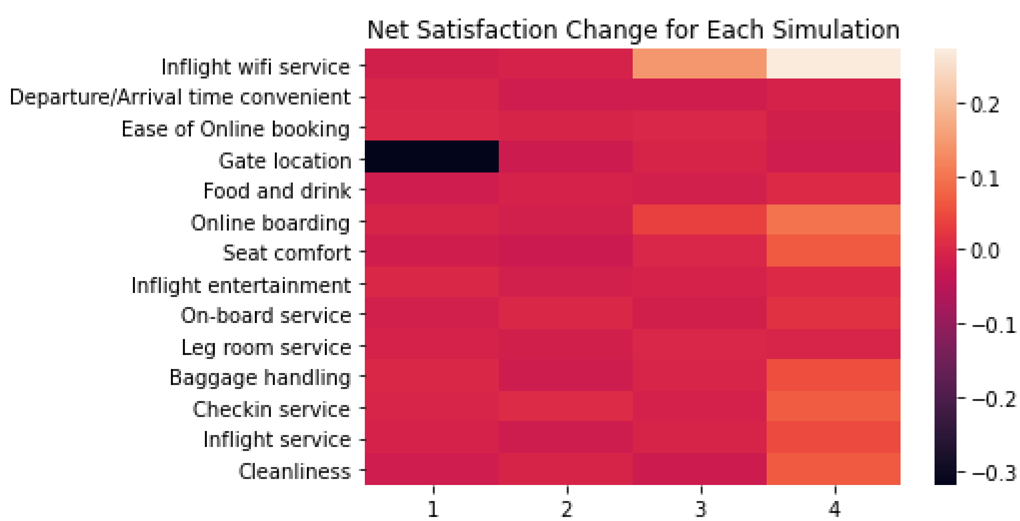
	1	2	3	4	5
Inflight wifi service	36.4%	25.3%	26.8%	61.1%	98.7%
Departure/Arrival time convenient	50.2%	44.2%	44.2%	38.3%	41.3%
Ease of Online booking	39.5%	31.2%	32.6%	53.3%	71.0%
Gate location	50.6%	46.5%	33.4%	38.9%	55.1%
Food and drink	16.8%	37.9%	39.7%	52.7%	55.4%
Online boarding	12.5%	10.6%	14.2%	61.9%	86.7%
Seat comfort	19.5%	20.2%	20.7%	55.6%	65.8%
Inflight entertainment	10.7%	17.5%	26.4%	61.4%	66.6%
On-board service	17.9%	23.8%	30.2%	52.4%	67.2%
Leg room service	17.1%	25.6%	26.5%	58.6%	62.0%
Baggage handling	27.7%	26.8%	23.0%	47.6%	63.3%
Checkin service	22.3%	23.2%	45.2%	45.3%	62.1%
Inflight service	27.1%	27.5%	22.7%	47.6%	62.1%
Cleanliness	17.2%	18.0%	43.1%	53.4%	63.1%

Below are the results of the experiment where each score is increased by one.

Average Satisfaction after Scores increase by 1

	1->2	2->3	3->4	4->5	N/A
Inflight wifi service	35.0%	24.8%	41.5%	88.3%	NaN
Departure/Arrival time convenient	50.1%	42.4%	42.7%	37.7%	NaN
Ease of Online booking	39.5%	30.8%	32.6%	51.9%	NaN
Gate location	18.6%	44.2%	33.0%	37.0%	NaN
Food and drink	15.1%	37.4%	38.7%	53.1%	NaN
Online boarding	12.1%	9.6%	17.9%	71.6%	NaN
Seat comfort	18.0%	17.7%	20.8%	62.3%	NaN
Inflight entertainment	11.0%	16.4%	25.9%	61.9%	NaN
On-board service	16.8%	24.1%	29.0%	54.2%	NaN
Leg room service	16.4%	24.3%	26.6%	58.4%	NaN
Baggage handling	27.9%	24.8%	22.9%	53.0%	NaN
Checkin service	22.3%	24.0%	44.4%	52.3%	NaN
Inflight service	26.6%	25.5%	22.3%	52.5%	NaN
Cleanliness	15.4%	17.5%	40.8%	60.2%	NaN

Net satisfaction change resulting from subtracting simulation results from the baseline.



In the above figure, lighter colors correspond to increases in customer satisfaction over and above the baseline satisfaction and therefore contribute to increases in ‘satisfaction’. Some examples are Inflight wifi service, ‘Online boarding’ and ‘Checkin service’.

On the other hand, for ‘Gate location’, an increase of score from 1 to 2 in ‘Gate location’ seems to significantly detract from average satisfaction. Given the direction of change, the direction of this large impact does not make sense however I do not have an explanation as to why this might happen. This could be studied further.

Combining the above changes with score distributions and existing satisfaction percentages, the (sorted) outcome of the shift is as follows. Note the increase in ‘satisfaction’ for Inflight wifi service’ and ‘Online boarding.’

Impact of 1 Point Increase on Average Satisfaction (Sorted)

	Impact of 1 Point Increase	Baseline Satisfaction	Satisfaction After 1 Point Increase
Inflight wifi service	8.8%	43.0%	51.8%
Online boarding	3.5%	43.0%	46.5%
Checkin service	1.8%	43.0%	44.8%
Baggage handling	1.8%	43.0%	44.8%
Seat comfort	1.6%	43.0%	44.6%
Inflight service	1.5%	43.0%	44.5%
Cleanliness	0.9%	43.0%	44.0%
On-board service	0.2%	43.0%	43.2%
Inflight entertainment	-0.1%	43.0%	42.9%
Leg room service	-0.3%	43.0%	42.7%
Ease of Online booking	-0.4%	43.0%	42.7%
Food and drink	-0.4%	43.0%	42.6%
Departure/Arrival time convenient	-0.8%	43.0%	42.3%
Gate location	-6.3%	43.0%	36.7%

Recommendations

While the above data exploration and methodology can be used in a variety of ways, a first look at the results above indicate that the following can benefit passenger satisfaction the most:

1) Improve ‘inflight wifi service’
2) Improve ‘online boarding’

All of the results however should consider the cost of implementing those improvements and perhaps run small-scale real-life experiments to test the validity of the above results before committing fully.

In addition to the result and recommendations, data exploration has brought to focus some factors that could help better model passenger satisfaction. Two cases in point are:

1) 'Relative Arrival Delay': As seen in the Feature Engineering section, satisfaction drops strongly with this factor as well as its high place in the feature importance list. An investigation into causes could shed light on prescriptive measures and increase satisfaction.
2) 'Flight Distance': The distribution suggests a sample split could be fruitful while the variable's high ranking in the feature importance list may hint at more insights to uncover.

Taken together, I believe that the two sets of recommendations above would be a good starting point to start to understand passenger preferences and drive large-scale changes across the company.

Further research for this project could include

1) Designing field experiments to check the validity of the improvement recommendations
2) Investigating causes of 'Relative Arrival Delay'
3) Creating smaller models for short- and long-distance flights
4) Coding the infrastructure that enables analysts to simulate impacts of response changes in a more granular manner: e.g. ability to create customized scenarios with only a percent/demographic of users responding positively.

Conclusion

This project attempted to initiate an analysis regarding how an airline could approach the question of increasing passenger satisfaction given its finite resources.

While the work done in this project generated recommendations about possible highest impact flight quality factors, it is also meant as a starting point to a more data-based decision-making.