Modeling the Characteristics of Airport Travelers: Final Report

Melanie Malinas
Springboard Data Science Career Track
Capstone Project 1

Introduction and Problem Statement:

Travel and tourism is a huge industry. According to the World Travel and Tourism Council, the world travel and tourism sector grew 3.9 percent in 2018, outpacing the global GDP growth of 3.2 percent, contributing $8.8 trillion and 319 million jobs to the global economy. In 2018, it generated 10.4 percent of all global economic activity. There is therefore a huge opportunity to use data science to advance the travel and tourism industry.

According to the same report, 21.5 percent of travel and tourism spending in 2018 was on business. If you look at just airlines, however, business travelers represent a huge portion of an airline's profitability. Business travelers account for 12% of an airline's passengers, but are usually twice as profitable - in fact, on some flights, business travelers make up 75% of an airline's profits. This is because businesses these days are willing to pay more to book last-minute flights for their employees, or book non-stop flights or first-class/business class seats. It would therefore be useful to use data science be able to distinguish between business and non-business travelers for the purpose of better targeting these different types of travelers.

For this project, I will be using the San Francisco International Airport (SFO) Customer Survey from 2017 to build a model of the profiles of business vs. non-business travelers. This survey asked over 2,500 SFO travelers, mostly in person at their gates, about their experience with SFO as well as information about their trip and demographic information. Using this information, I will statistically analyze the characteristics of SFO travelers as well as use machine learning to predict who is a business traveler and who is a non-business traveler.

Data Cleaning:

I acquired this survey data freely from DataSF (San Francisco Open Data), a project of the city of San Francisco. DataSF has many different datasets from different sectors of San Francisco. The survey data was downloadable in CSV format, which made it simple to open and work with. Each row in the survey was a different survey respondent, and each column was a question on the survey. A data dictionary described the different questions and the codes for each answer. Almost all of the data was categorical data, with the answers coded as 1, 2, 3, etc. with 0 for blank/non-response.

The data also had weights for each respondent that were given to try and make the survey more representative of the population. For various different questions, I created dictionaries with weighted counts for each answer. For example, when the data was weighted, the number of pleasure travelers went down and the number of business travelers went up.

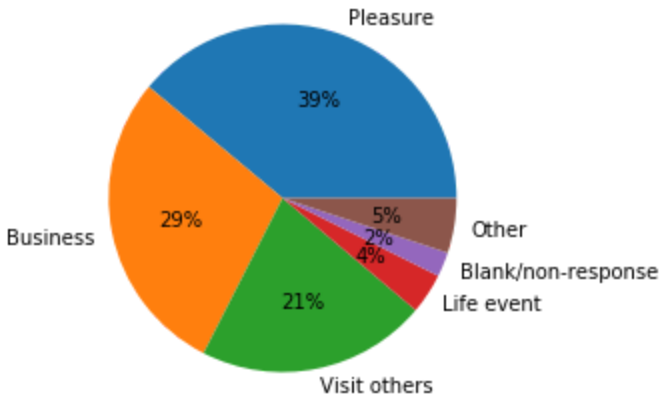
Another step I took to clean the data was to rename several of the columns. I renamed 'Q2PURP1' to 'Purpose', 'Q3GETTO1' to 'Transportation', etc. This made the column names better formatted and easier to read. I also mapped the actual descriptions of the answers from the data dictionary onto their codes to make the data more readable.

Other than that, this data was pretty clean, well-organized, and easy to work with.

Exploratory Analysis:

For the first part of my project, I did exploratory data visualization and statistical analysis. I found several interesting trends in the data that could help distinguish between business and non-business travelers.

Looking at the weighted dictionary of counts, I found that about 29% of travelers were business travelers. I created a new column classifying all travelers as either business or non-business travelers.



I then looked at correlations between variables. Because the data are entirely categorical,it is not possible to directly compute correlations. Therefore, I used Cramer's V, which is able to approximate the correlations for categorical data. Cramer's V is based on the chi-squared statistic:

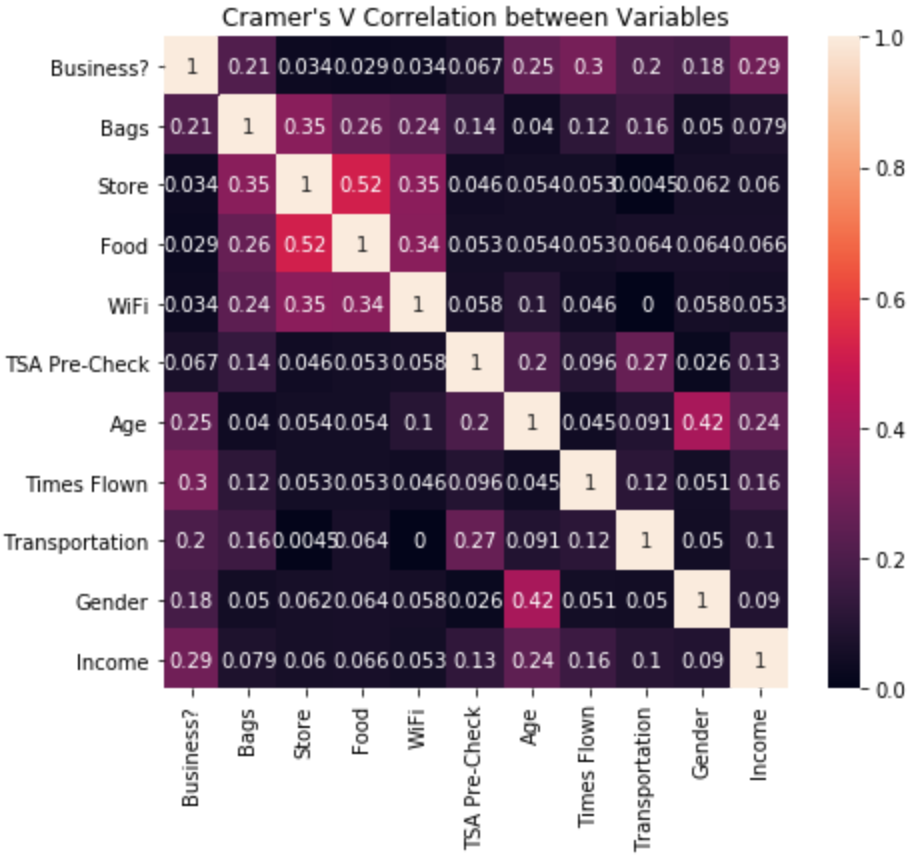
$$V = \sqrt{\frac{\varphi^2}{\min(k-1, r-1)}} = \sqrt{\frac{\chi^2/n}{\min(k-1, r-1)}}$$

Here, φ is the phi coefficient, χ^2 is the chi-squared statistic, n is the number of observations, and k and r are the number of columns and rows, respectively.

The columns of interest that I looked at were business/non-business (Business?), whether bags were checked (Bags), whether someone went to a store (Store), whether someone bought

food (Food), whether someone used the free WiFi (WiFi), whether someone went through TSA Pre-Check (TSA Pre-Check), the traveler's age (Age), the number of times flown out of SFO in the past year (Times Flown), the type of transportation used to get to the airport (Transportation), the gender of the traveler (Gender), and the traveler's household income (Income).

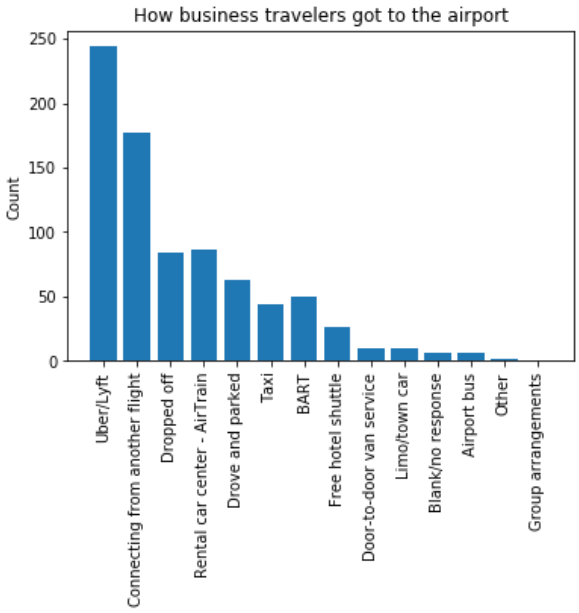
I then created a heatmap of the Cramer's V values:

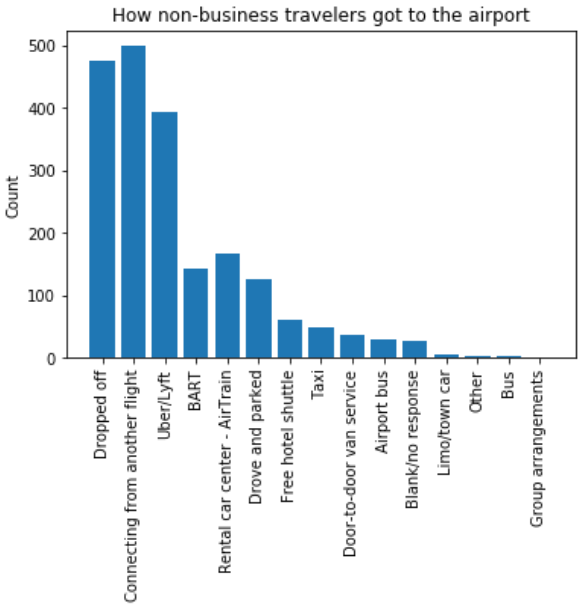


The highest correlation was between whether someone went to a store and whether they bought food, with a Cramer's V value of 0.52. The next highest correlation was between gender and age. Examining the gender and age breakdown more closely, it appears that there are more female travelers between the ages of 18 and 34, whereas there are more male travelers between the ages of 35 and 54. At ages 55 and over it appears that the gender breakdown of travelers is about equal.

In terms of correlations between whether someone was a business traveler and the other categories, I saw small correlations between that and Bags, Age, Times Flown, Transportation, Gender, and Income - with Income being the largest. These are therefore good categories to explore more in depth.

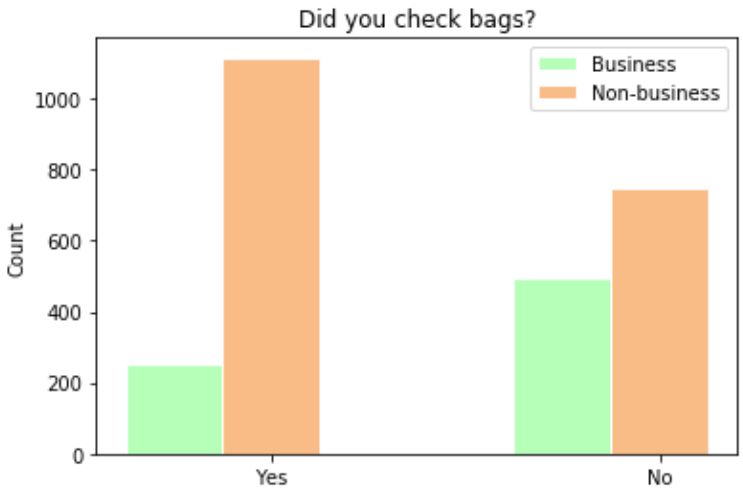
My next step in my exploratory analysis was to compare business travelers and non-business travelers in my columns of interest. I made graphs comparing how business and non-business travelers got to the airport:





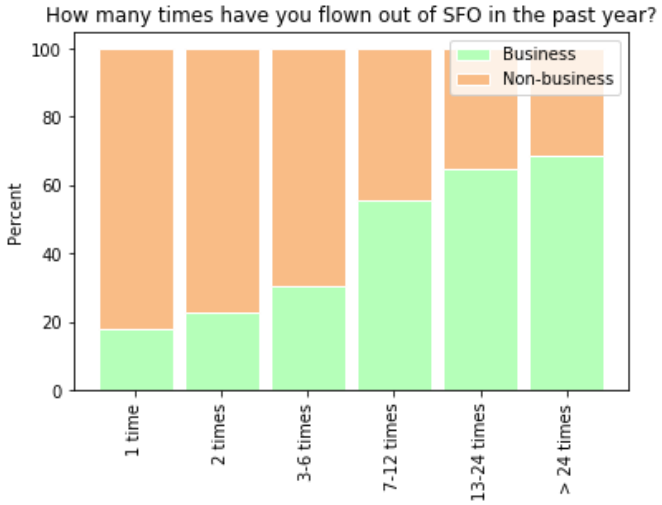
As you can see, the most common way that business travelers got to the airport was using a ridesharing service such as Uber or Lyft, whereas the most common way that a non-business traveler got to the airport was being dropped off. The proportion of business travelers using Uber or Lyft was statistically significantly higher than the proportion of non-business travelers using Uber or Lyft ($p<0.00001$), with about 30% of business travelers using Uber or Lyft compared to about 20% of non-business travelers. I therefore believe that this may be a good category to use in predicting whether or not someone is a business traveler.

I also examined the breakdowns between business and non-business travelers on whether the traveler checked a bag, and found that most business travelers did not check a bag, while most non-business travelers checked a bag. This difference was statistically significant, with about 31% of business travelers checking a bag compared to about 55% of non-business travelers checking a bag.



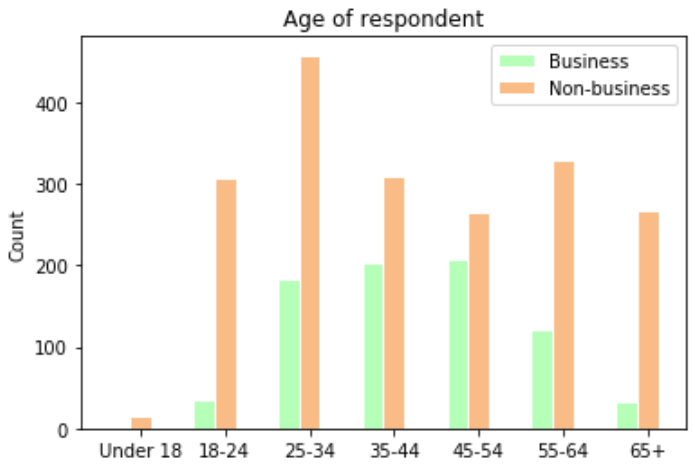
I also examined the variables of whether someone had visited a store at the airport, bought food, or used the free WiFi, and I found that there were no significant differences between business and non-business travelers in these categories.

I then looked at how many times a traveler had flown out of SFO in the past year. For this I created a stacked bar chart, where each bar shows the percentage of each category that are business travelers vs. non-business travelers:



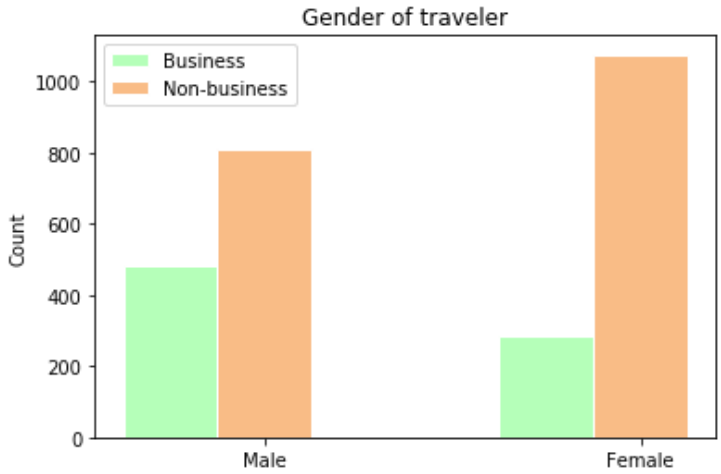
As the number of times flown out of SFO increases, the proportion of travelers who are business travelers increases. The proportion of travelers flying more than 13 times per year is significantly different between business and non-business travelers.

I also looked at the distribution of ages of the travelers:



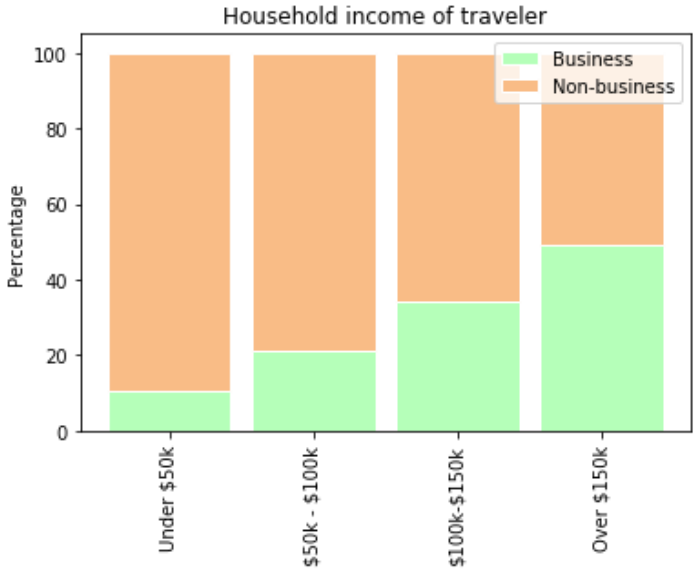
Business travelers appear to be mostly in the range of 25-64, with very few being age 18-24 or 65 and over. By contrast, the non-business travelers have large cohorts of travelers ages 18 to 24 and 65 and over. However, the average age of the travelers, around 43 years, was not significantly different between business and non-business travelers.

I then looked at the gender of the travelers:



Business travelers were majority male, and non-business travelers were majority female, and this difference was statistically significant. It is perhaps not surprising that the business travelers were majority male, as it seems more likely to have male businessmen that travel for work. However, it was surprising to me that the non-business travelers were majority female. I am not sure why this is or if it is actually representative of SFO travelers. I would think that the weighting would have corrected for this, but it did not.

Finally, I analyzed household income levels for business and non-business travelers using a stacked bar plot:



Perhaps unsurprisingly, I found that as the household income level went up, the proportion of business travelers also went up. The average household income for a business traveler was about $134,000, which was statistically significantly different from the average household income for a non-business traveler, $97,000.

The categories of Bags, Age, Times Flown, Transportation, Gender, and Income are therefore all good variables for predicting whether someone is a business traveler, which is what I will do in the next section using machine learning.

m

Summary of Differences Between Business and Non-Business Travelers:

	Business Travelers	**Non-business travelers**
Whether they checked bags (**Bags**)	More likely to not check a bag (31% checked a bag)	More likely to check a bag (55% checked a bag)
Number of times flown out of SFO in the past year (**Times Flown)**	Those who flew high number of times more likely to be business travelers	
Age of traveler (**Age**)	Mostly between the ages of 25-64	Large cohorts under 25 and over 64
Method of transportation to the airport (**Transportation**)	More likely to take Uber or Lyft	Less likely to take Uber or Lyft
Gender of traveler (**Gender**)	More likely to be male	More likely to be female
Household income of traveler (**Income**)	Higher average income	Lower average income

Machine Learning and In-Depth Analysis:

The goal of the machine learning aspect of this project was to create a model that would successfully predict whether someone was a business or non-business traveler. Based on my exploratory data analysis, I knew that there were several features of the dataset that differed between business and non-business travelers: namely, whether they checked bags (Bags), the number of times flown out of SFO in the past year (Times Flown), the age of the traveler (Age), the method of transportation to the airport (Transportation), the gender of the traveler (Gender), and the household income of the traveler (Income). However, in thinking about the business aspect of this project, it is important to consider which of these features a business would actually have access to. For example, if a business marketing to business travelers set up a booth outside the airport where it was possible to see where travelers were coming in from, then the business might have access to the Transportation feature. However, it is unlikely that any business would have access to the traveler's household income.

For my business case, I decided to take the position of an airline marketing an airline miles rewards credit card. I have seen booths for airline rewards credit cards inside the terminal before, but I believe that it might be possible to market the credit card to the traveler when the traveler checks in at the desk, where the airline would have clear information about whether the traveler checked a bag (Bags), the age and gender of the traveler (Age and Gender), as well as

information about that traveler's past flights with the airline (approximated by Times Flown). I thus decided to build my model using the Bags, Age, Gender, and Times Flown features.

For this project, it is important to strike a balance between precision and recall. Higher precision means fewer false positives (non-business travelers classified as business travelers), whereas higher recall means fewer false negatives (business travelers classified as non-business travelers). In general, if you are advertising a special deal for business travelers where you could possibly lose money if the deal is taken by non-business travelers, then you want to minimize the number of false positives. However, you could also lose money if you have a large number of false negatives and fail to give the deal to enough business travelers. Whether you weight precision or recall more heavily would therefore depend on the specific deal you were offering and the projected cost-benefit analysis. For this project, I will weight precision and recall equally and attempt to maximize the F1 score, which is the harmonic mean of the precision and recall.

Building the Model:

I tried three different algorithms for this project: logistic regression, random forests, and support vector machines. Because my data was somewhat imbalanced (only 29% of travelers were business travelers) I also tested whether it was necessary to compensate for this in some way, because most machine learning algorithms work best when you have the same number of data points for each class. One method of compensating in scikit-learn is a parameter called class_weight, which allows you to specify how you want to weight the different classes. If you specify that the class_weight parameter should be "balanced", the software will use the values of your dependent variable column to automatically adjust weights inversely proportional to the class frequencies. I also used GridSearchCV to perform cross-validation and optimize my hyperparameters. I then defined a function to calculate the maximum F1 score over all of the predicted probabilities.

The results of my six different models (logistic regression, random forests, and SVM with and without class balancing) are as follows, where the F1 score is the maximum over all predicted probabilities and the precision and recall are for that maximum F1 score:

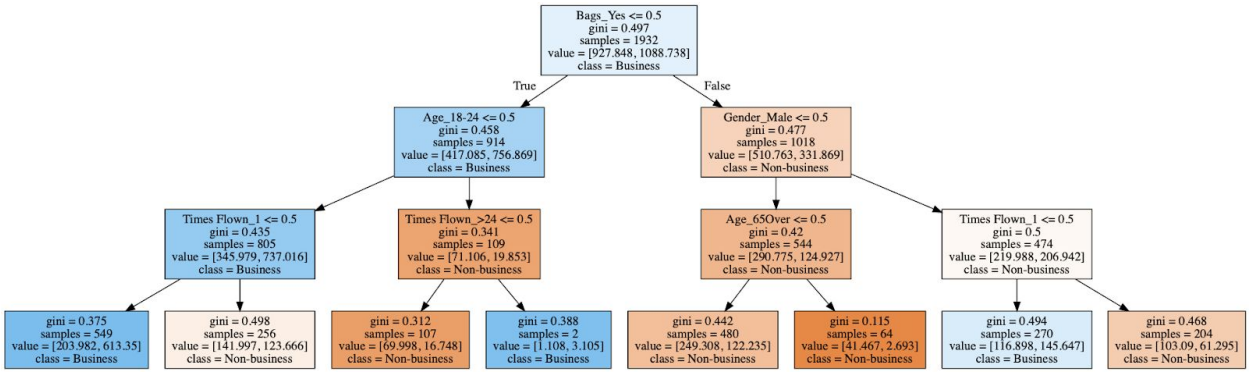
Method	Precision	Recall	F1 Score
Logistic regression - no class balance	0.54	0.62	**0.576**
Logistic regression - class balance	0.56	0.59	**0.577**
Random forests - no class balance	0.59	0.52	**0.551**
Random forests - class balance	0.60	0.51	**0.552**
SVM - no class balance	0.50	0.67	**0.571**
SVM - class balance	0.50	0.66	**0.573**

It appears that the best method is logistic regression, followed by SVM, followed by random forests. Class balancing makes only a small difference in the F1 score, 0.001 or 0.002 at the most. I therefore selected logistic regression as the best method for my data.

Using logistic regression, I then tried using a different method for dealing with the imbalanced data: upsampling. With upsampling, you resample, with replacement, the data in the minority class so that you end up with equal numbers of majority and minority class members. Doing logistic regression with upsampling and choosing the maximum F1 score, I got an F1 score of **0.591**, with a precision of 0.58 and a recall of 0.60. This is better than any of my previous results. I thus conclude that the best model for my data is logistic regression with upsampling.

Decision Tree:

To better visualize the contributions of the different features, I also created a single decision tree and visualized it using export_graphviz. From the decision tree, it appears that all of the variables of interest - Gender, Age, Bags, and Times Flown - contribute to whether a person is considered a business traveler or a non-business traveler. This is in agreement with my initial statistical analysis that showed differences between business and non-business travelers in all of these features.



At the bottom of the tree are the nodes. The darkness of a leaf represents the purity of the node. As you can see, the features that are most useful are whether someone checked a bag (those who checked bags are more likely to be non-business travelers), their gender (male travelers are more likely to be business travelers), whether they were age 18-24 or 65 and over (these age groups were more likely to be non-business travelers) as well as whether they flew more than one time or 24 times or more out of SFO (those who flew out more than one time are more likely to be business travelers). As you can see, the largest node of business travelers, with a gini impurity of 0.375, is those who did not check a bag, are not ages 18-24, and flew out more than one time.

Conclusion and Next Steps:

In this project I demonstrated that it is possible to predict whether someone is a business or non-business traveler based on just their gender, their age, whether they checked a bag, and how many times they have flown out of the airport. This project also showed the usefulness of upsampling when dealing with imbalanced classes. I was a little surprised that logistic regression was the best model and that random forests did less well, but I suppose that it is not always possible to predict which model will perform the best.

This project involved a lot of assumptions that might not hold up in a business scenario. For one, it may be possible for an airline to figure out whether someone was a business traveler based on whether they used a business credit card to book a flight or had a business-related email address, or if they used a travel agency that caters to business travelers. This was information that we did not have access to with this data. Therefore in a real-world scenario, it's likely we would be training this model on only a subset of the data and could build it into part of a larger ensemble model to distinguish the non-obvious business travelers from non-business travelers.