Modeling the Characteristics of Airport Travelers

Melanie Malinas
Springboard Data Science Career Track

Capstone 1

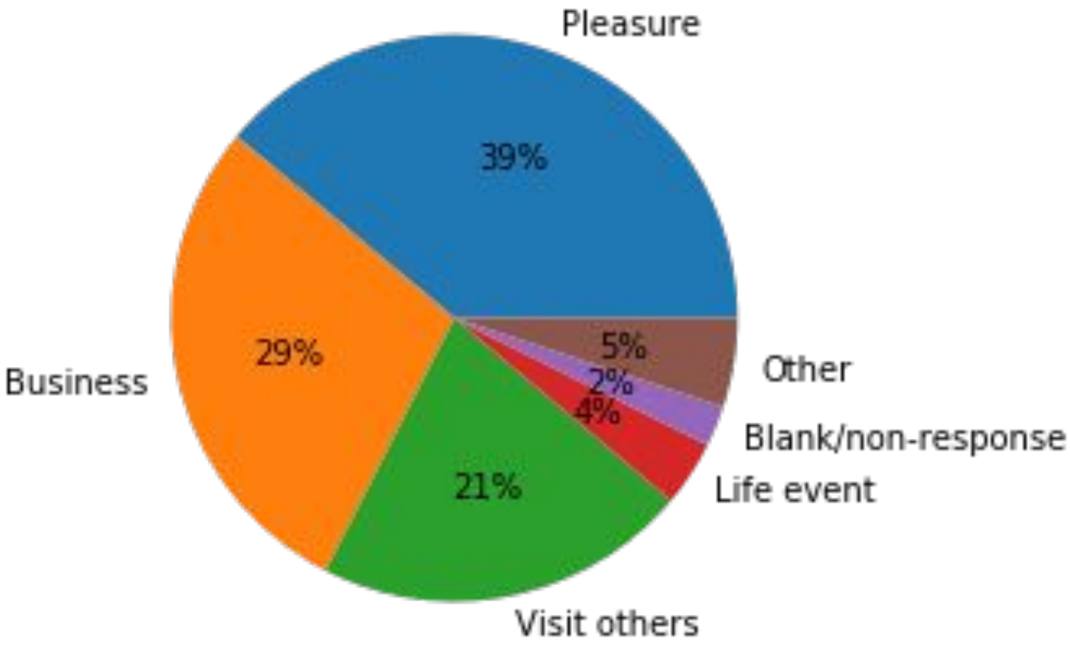
Problem Statement

- Analyzing the characteristics of business vs. non-business travelers
    - Business travelers are twice as profitable for an airline as non-business travelers
- Analysis of San Francisco International Airport (SFO) survey data from 2017
- Survey asked over 2,500 SFO travelers about their experience at SFO as well as information about their trips and demographic information

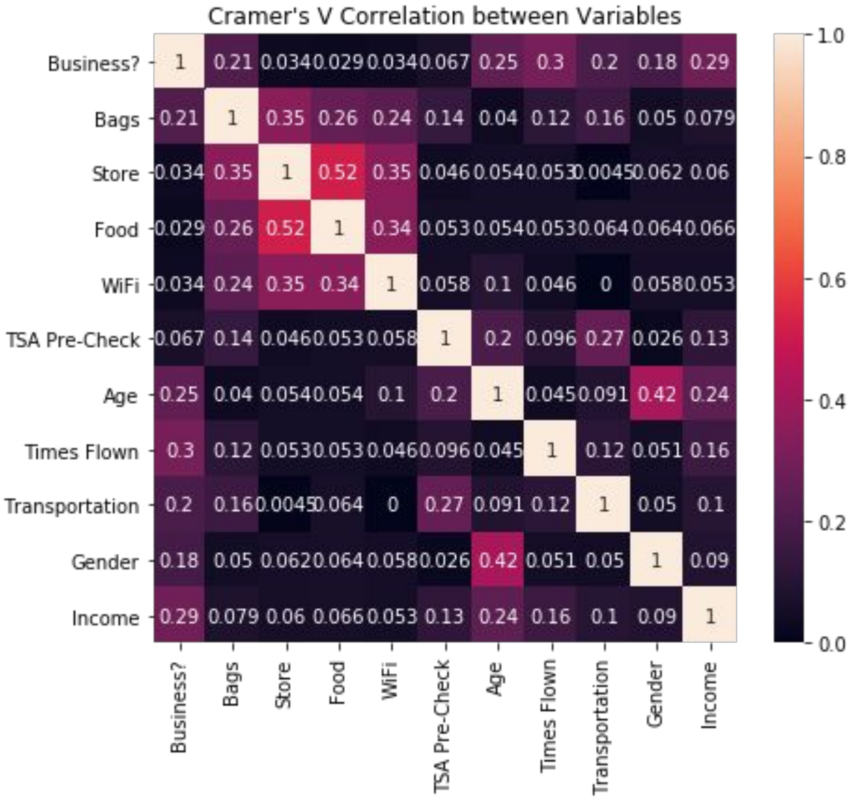
Data Cleaning

- Data was acquired in a CSV file from SF Open Data
- Almost all the data was categorical data
- Data was weighted, so 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
- Renamed columns
- Mapped descriptions from data dictionary onto data
- Created a new column for whether someone was a business traveler or a non-business traveler

Types of SFO travelers



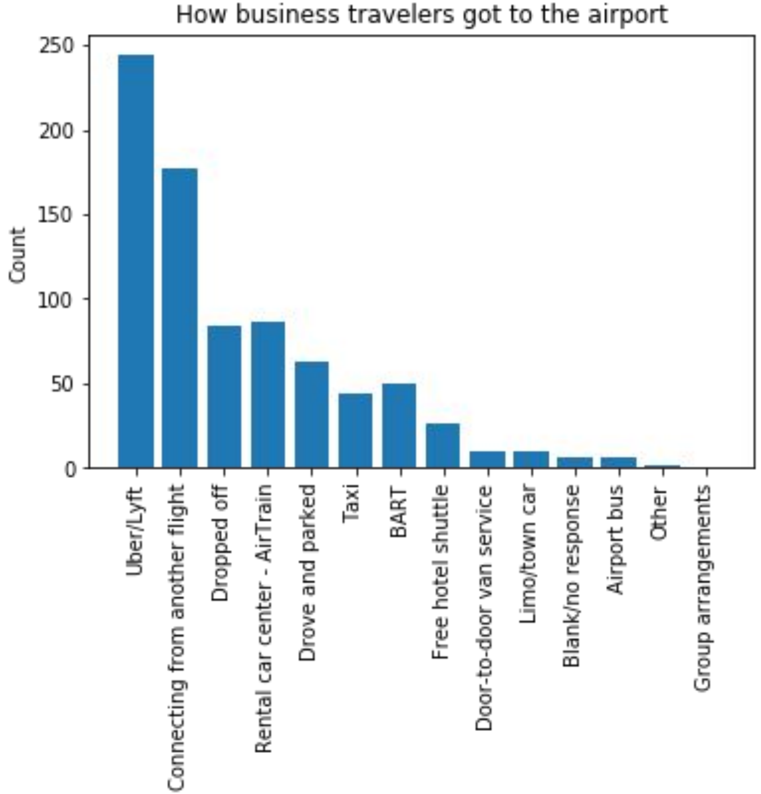
Categorical correlations between variables

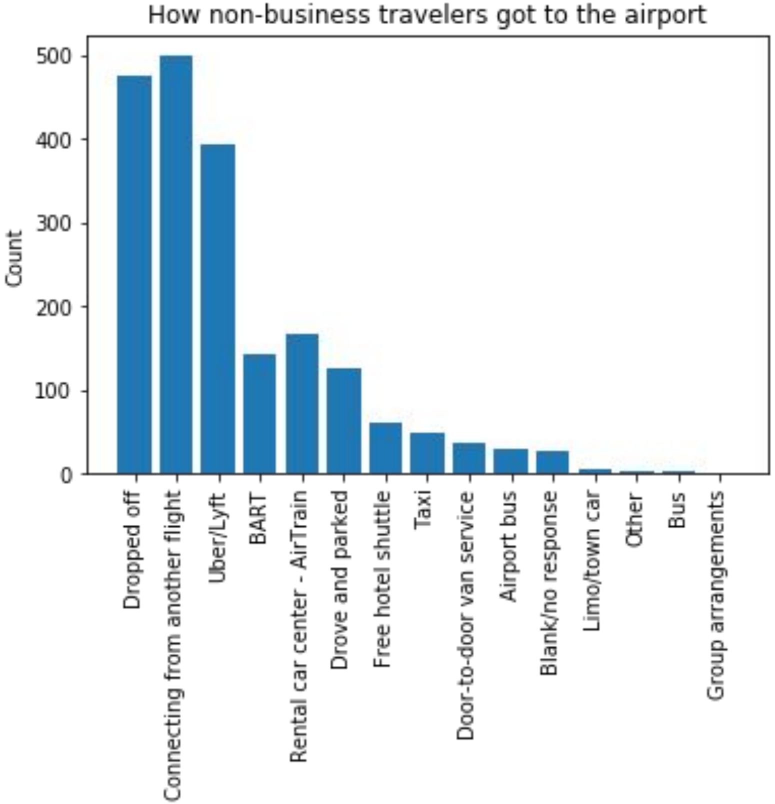


Cramer's V Correlation between Variables

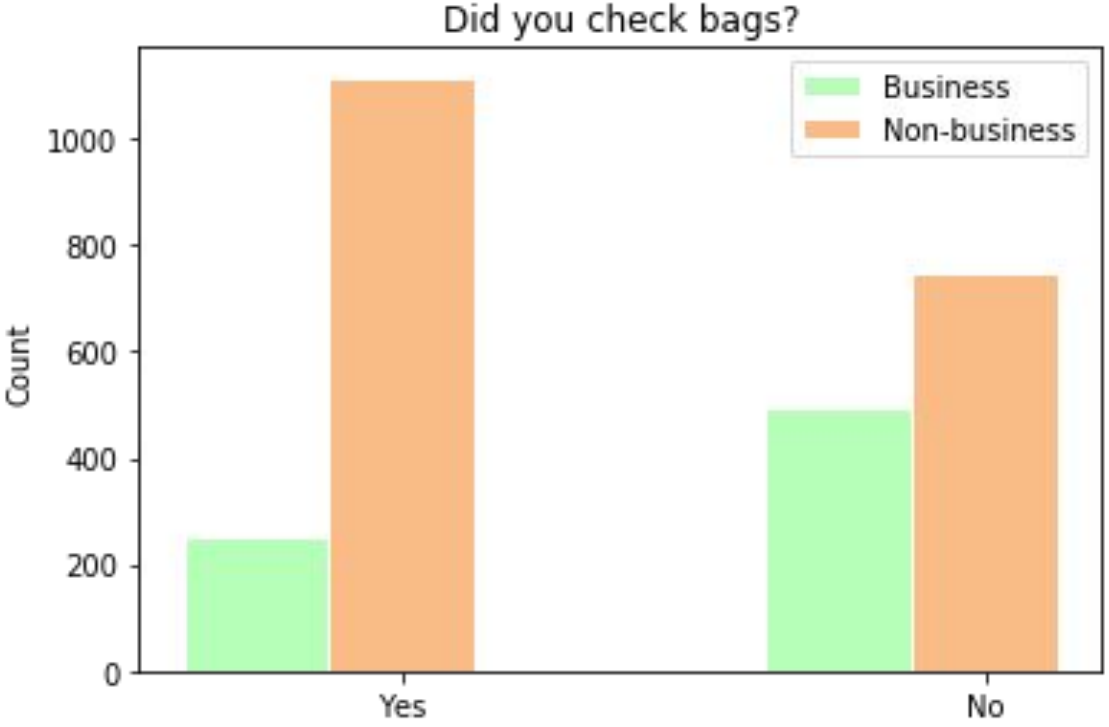
	Business?	Bags	Store	Food	WiFi	TSA Pre-Check	Age	Times Flown	Transportation	Gender	Income
Business?	1	0.21	0.034	0.029	0.034	0.067	0.25	0.3	0.2	0.18	0.29
Bags	0.21	1	0.35	0.26	0.24	0.14	0.04	0.12	0.16	0.05	0.079
Store	0.034	0.35	1	0.52	0.35	0.046	0.054	0.053	0.0045	0.062	0.06
Food	0.029	0.26	0.52	1	0.34	0.053	0.054	0.053	0.064	0.064	0.066
WiFi	0.034	0.24	0.35	0.34	1	0.058	0.1	0.046	0	0.058	0.053
TSA Pre-Check	0.067	0.14	0.046	0.053	0.058	1	0.2	0.096	0.27	0.026	0.13
Age	0.25	0.04	0.054	0.054	0.1	0.2	1	0.045	0.091	0.42	0.24
Times Flown	0.3	0.12	0.053	0.053	0.046	0.096	0.045	1	0.12	0.051	0.16
Transportation	0.2	0.16	0.0045	0.064	0	0.27	0.091	0.12	1	0.05	0.1
Gender	0.18	0.05	0.062	0.064	0.058	0.026	0.42	0.051	0.05	1	0.09
Income	0.29	0.079	0.06	0.066	0.053	0.13	0.24	0.16	0.1	0.09	1

How people got to the airport

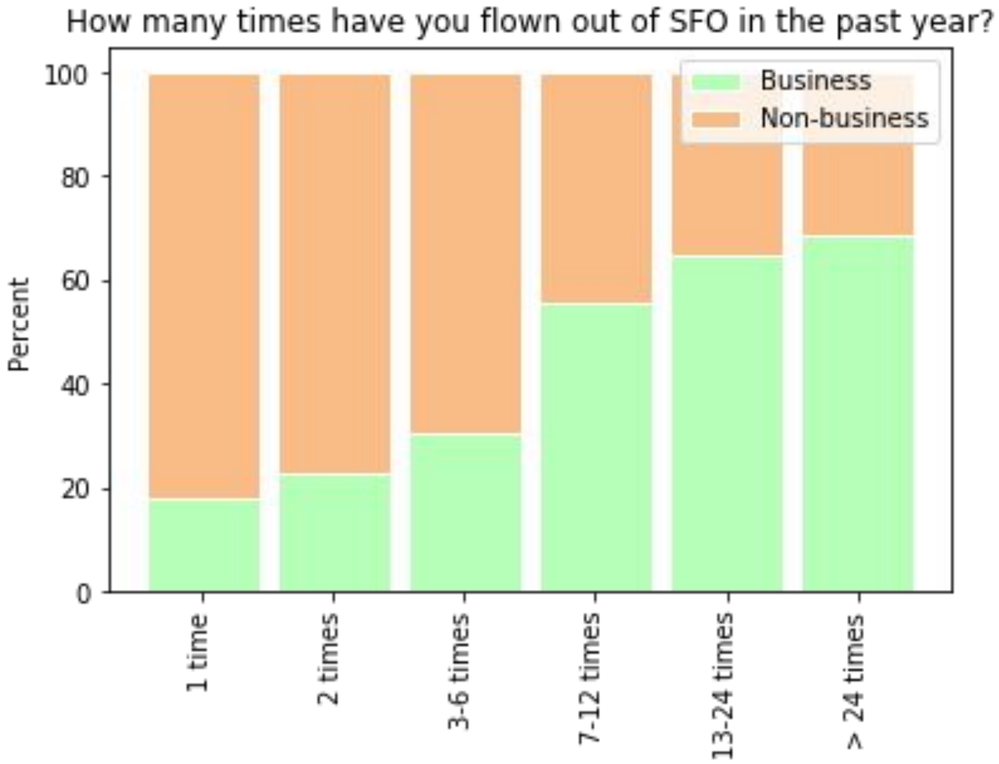




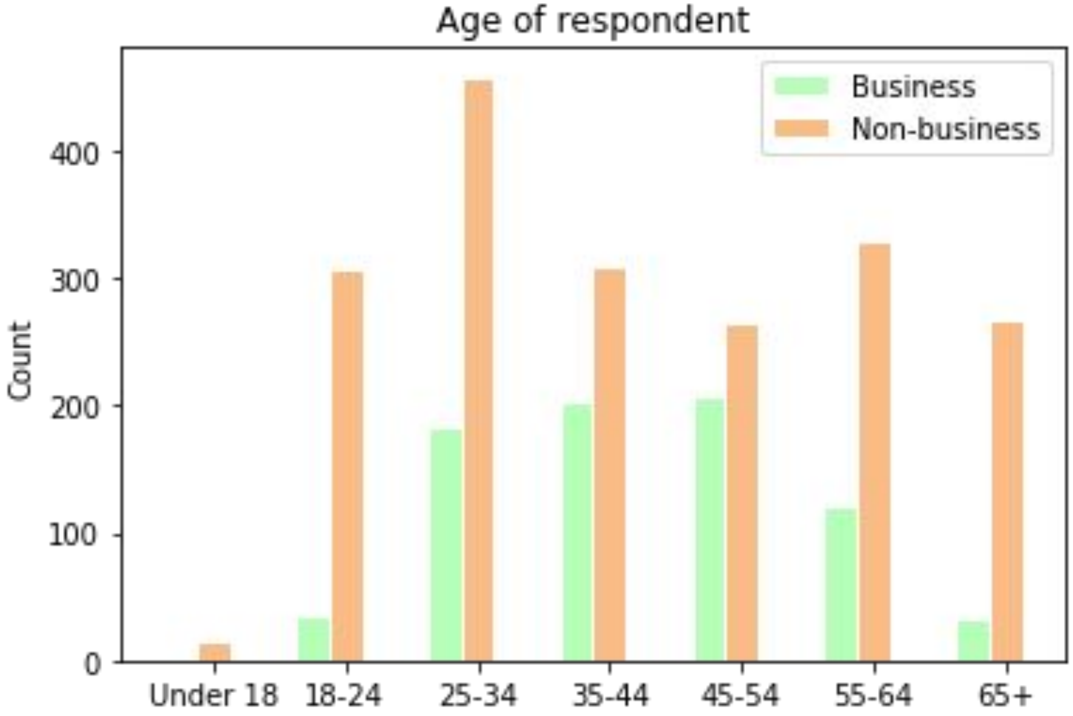
Checking bags



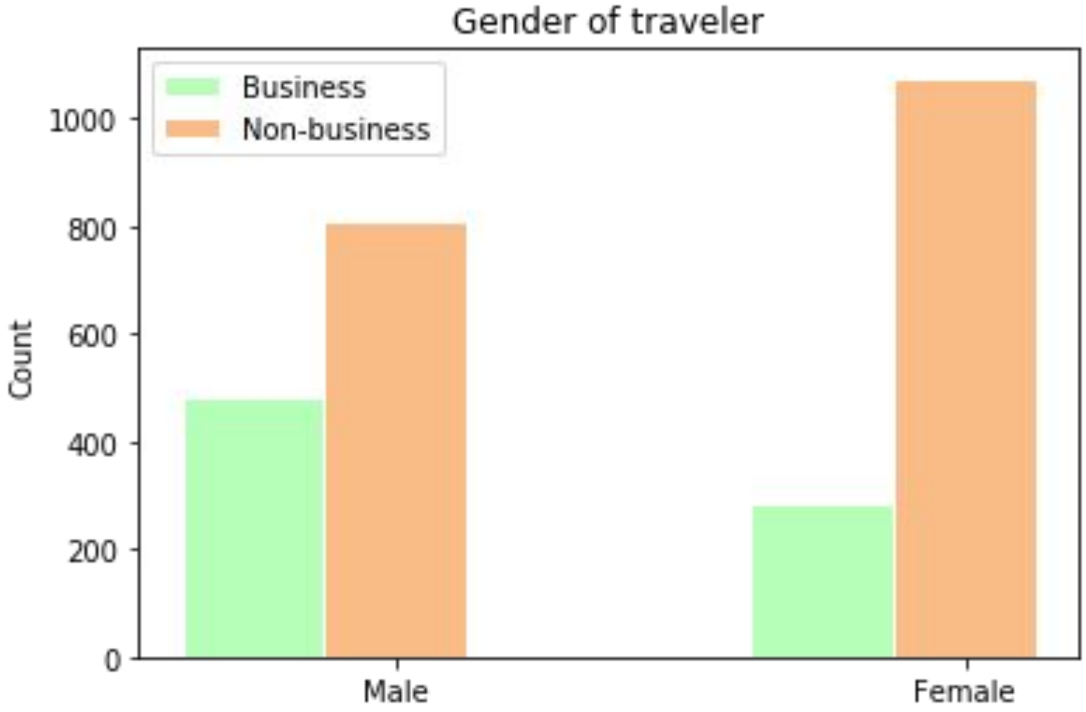
Times flown out of SFO



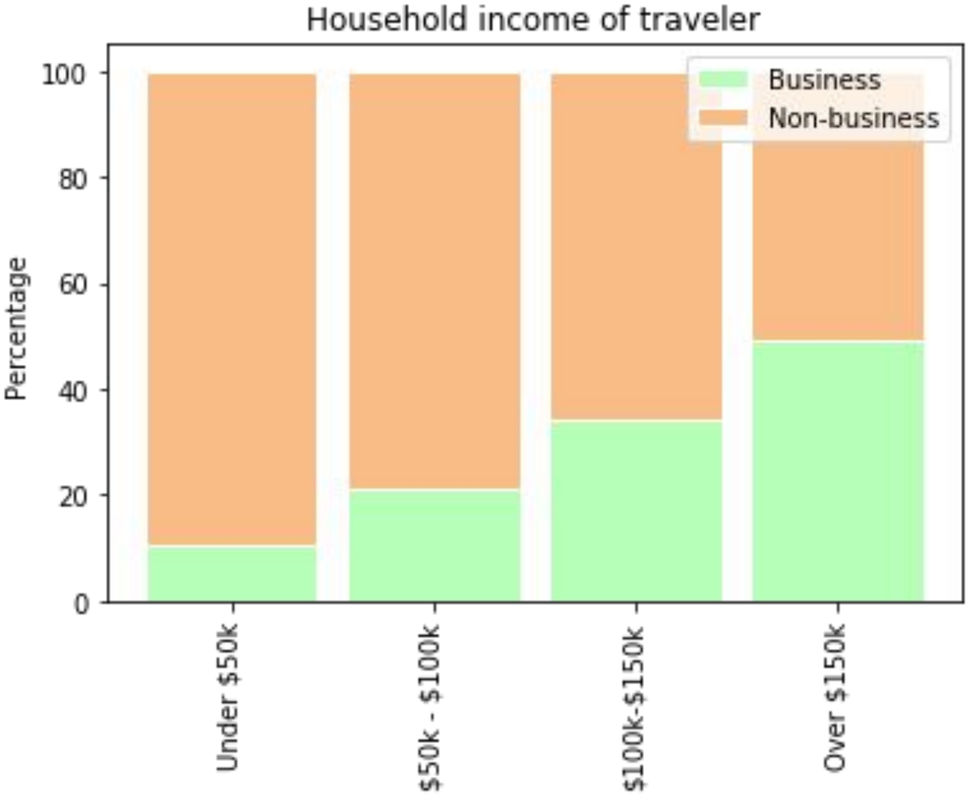
Age of traveler



Gender of traveler



Household income of traveler



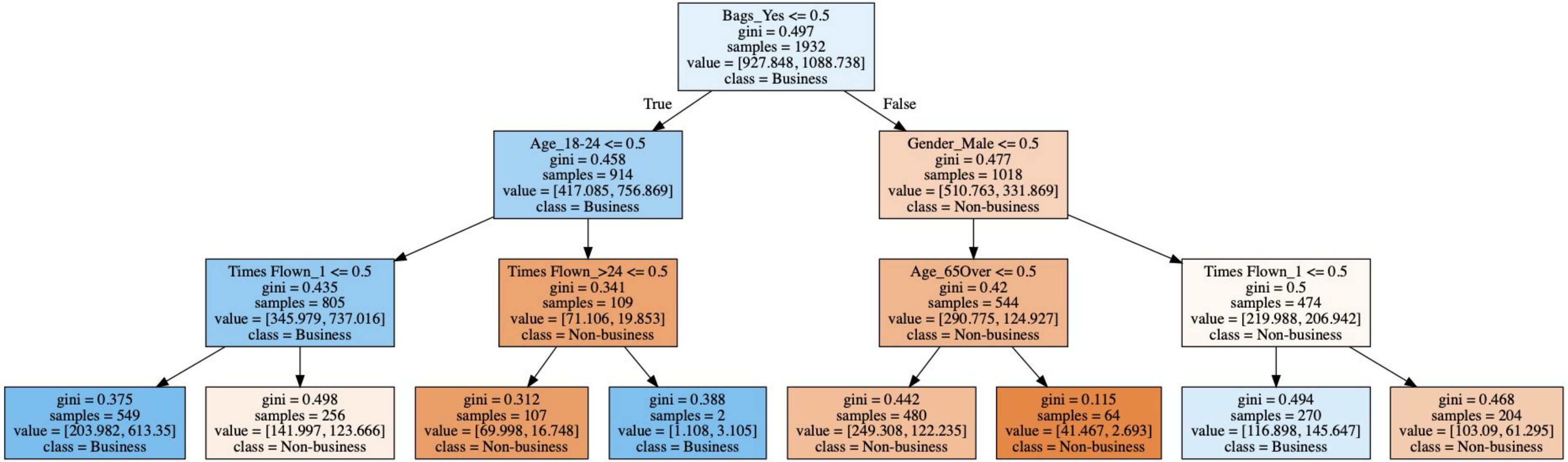
Business case

- Airline marketing a travel rewards credit card
- Giving deal when checking in at the desk of the airport
  - Features to use: Bags, Age, Gender, Times Flown
- Optimizing F1 score
  - Striking a balance between precision and recall

Machine learning methods

Method	Precision	Recall	F1 Score
Logistic regression - no class balance	0.54	0.63	**0.581**
Logistic regression - class balance	0.55	0.61	**0.578**
Random forests - no class balance	0.59	0.52	**0.550**
Random forests - class balance	0.61	0.51	**0.553**
SVM - no class balance	0.66	0.42	**0.511**
SVM - class balance	0.49	0.64	**0.553**
Logistic regression - upsampling	0.58	0.60	**0.591**

Decision tree visualization



Conclusions

- It is possible to predict whether someone is a business traveler based only on their gender, age, number of times flown out of SFO, and whether they checked a bag.
- Surprising that logistic regression was the best model and random forests did not do as well
- Usefulness of upsampling
- Assumptions of this project:
  - We did not have info on whether someone used a business credit card to book flight or whether they used a business-related email
- Work may be useful for distinguishing non-obvious business travelers from non-business travelers