What Determines Airline Demands: Evidence From 1993-2024 US Airline Dataset

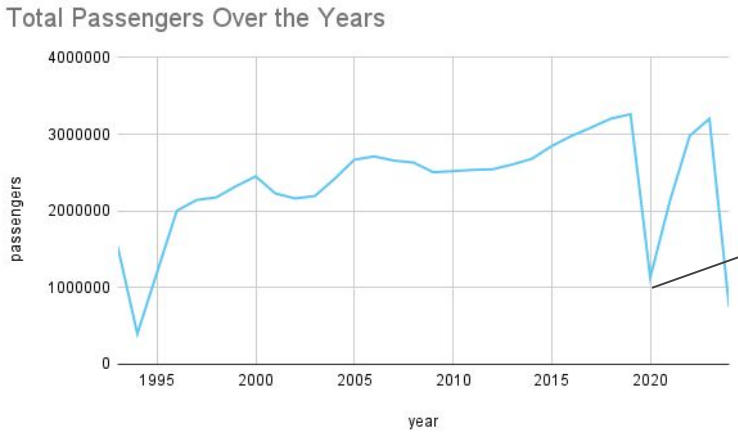
BY: Feng, Michelle, Dora

Introduction

Air travel is important for connecting people and economies and, as a result of this, it's crucial for researchers and even companies to understand air routes, flights, and passenger volumes. Using a dataset that consists of US airline demands data from 1993 to 2024, we first constructed a multivariate regression model to grasp the potential relationship between variables, followed by a Random Forest algorithm to determine the variables that are most important in predicting airline demands – all of which provide valuable insights for decision-makers and airlines companies to prepare for strategies relevant to the industry.

Raw data

The dataset includes a variety of variables such as identifiers for tables and routes, details about the origin and destination cities and airports, the year (from 1993 - 2024) and quarter of each record, and flight-related metrics. These metrics include distance between airports in miles, the number of passengers, average fares, market shares, and carrier information.



total_passeng~r	Coefficient
time	59604.54
post_covid	-447019.3
post_covid_time	-30343.44
_cons	1636415

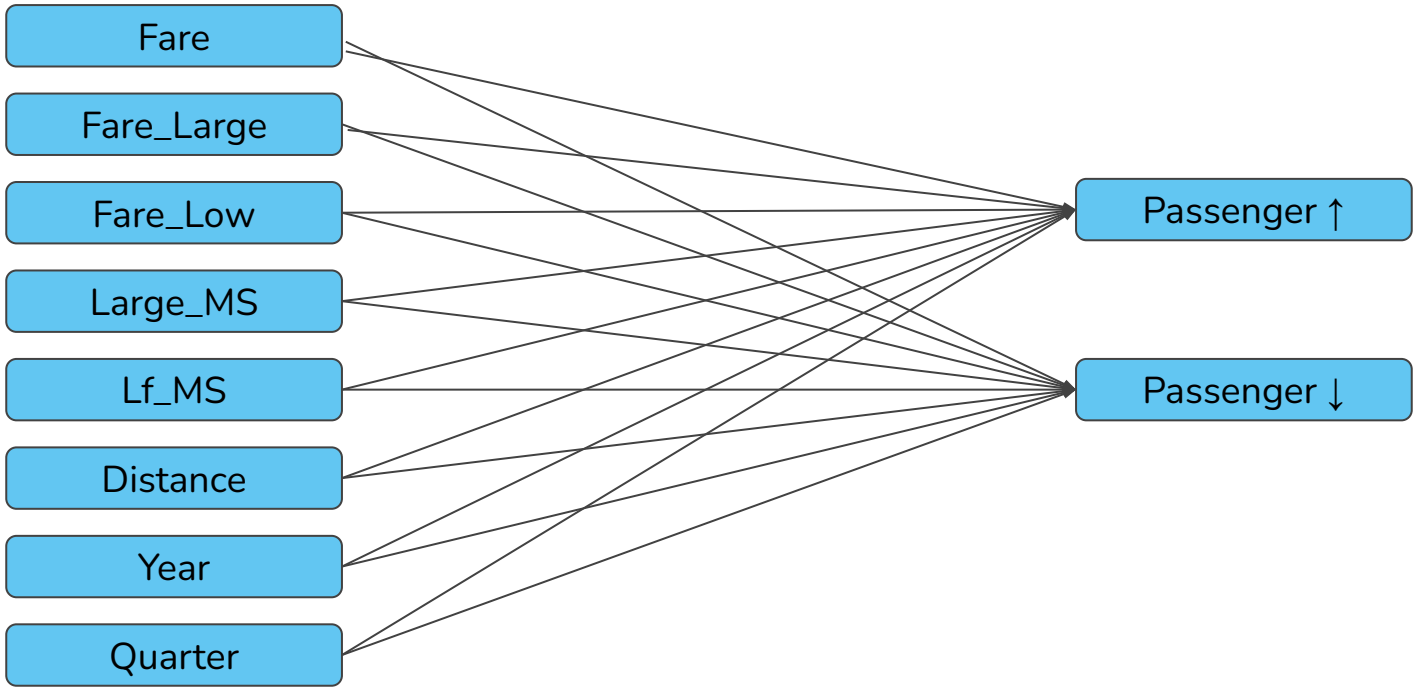
Multivariate Regression Model

$$\text{Passengers}_{i,t} = \beta_0 + \sum_{k=1}^{8} \beta_k X_{k,i,t} + \epsilon_{i,t}$$

Where:

- β_0: Intercept (Constant Term)
- β_k: Coefficients for the Independent Variables (for k = 1, ..., 8)
- $X_{k,i,t}$: Independent Variables for Individual *i* at time *t*
- Independent Variables Include *Distance*, *Fare*, *Fare_Large*, *Fare_Low*, *Lf_MS*, *Large_MS*, *Year*, *Quarter*.

Theoretical Framework



Baseline Results

Table 1. Baseline Regression Results

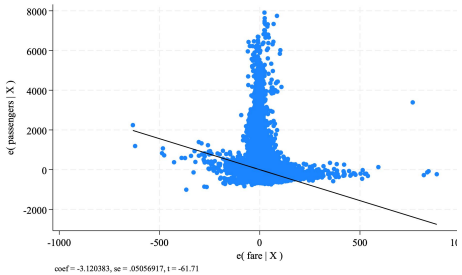
	Passengers
Fare	−3.120***
	(0.051)
Fare (Large)	2.357***
	(0.041)
Fare (Low)	−0.819***
	(0.032)
Market Share (Large)	−192.26***
	(5.58)
Market Share (Low Fare)	−143.93***
	(4.06)
Distance (Miles)	−0.022***
	(0.002)
Year	8.362***
	(0.117)
Quarter	4.698***
	(0.881)
Constant	−15964.32***
	(233.80)
Observations	244,343
R-Squared	0.0916
Adjusted R-Squared	0.0916
F-statistic	3079.45
Root MSE	488.47

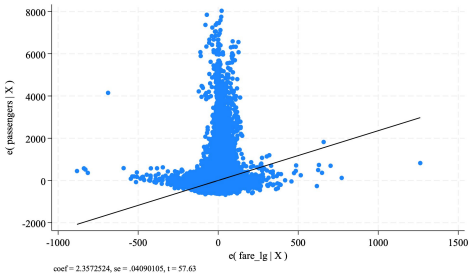
Note: Coefficients are presented as raw values, with standard errors in parentheses. $***p < 0.01, **p < 0.05, *p < 0.1$

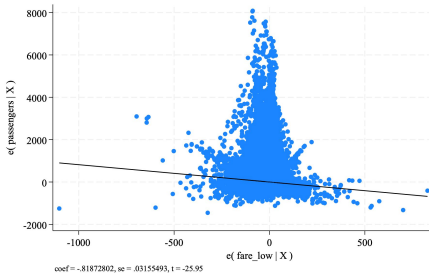
*Only About **9.16%** of variations in **Passengers** can be explained by this multivariate regression model

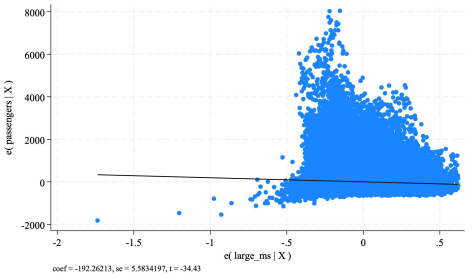
*On average, the difference between the predicted and actual number of passengers is around **488 passengers**

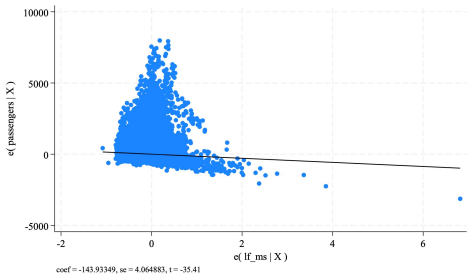
Partial Regression Plots

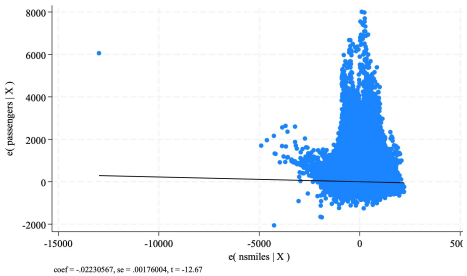


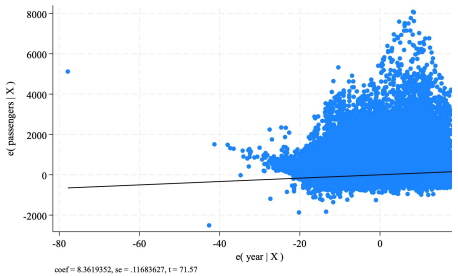


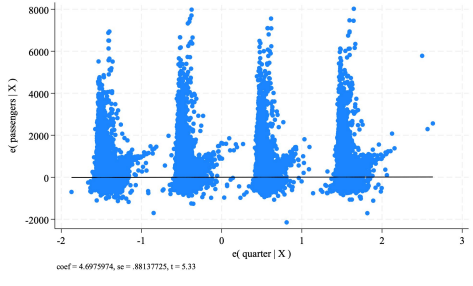






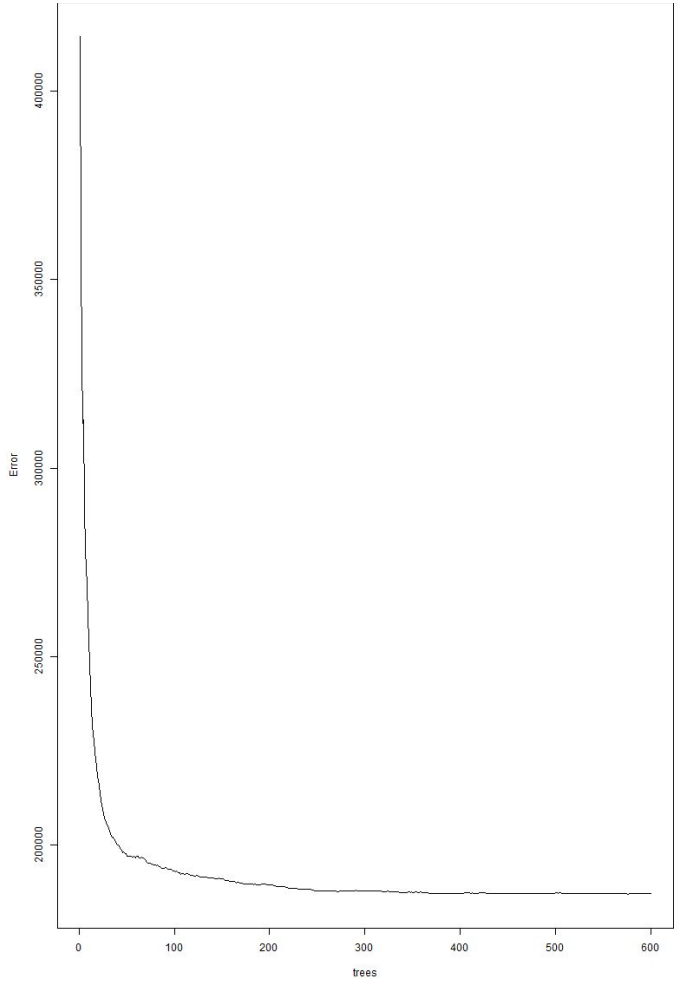




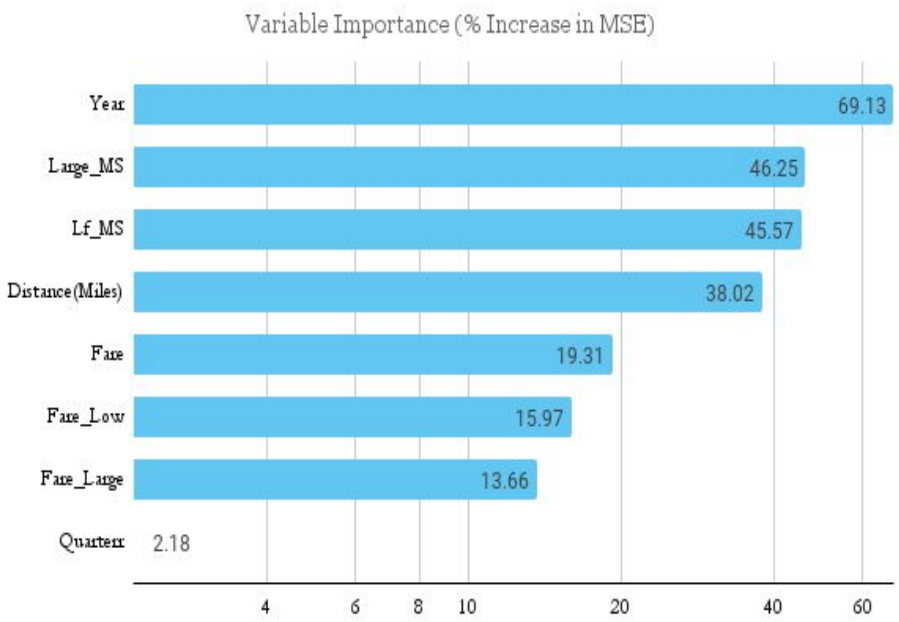


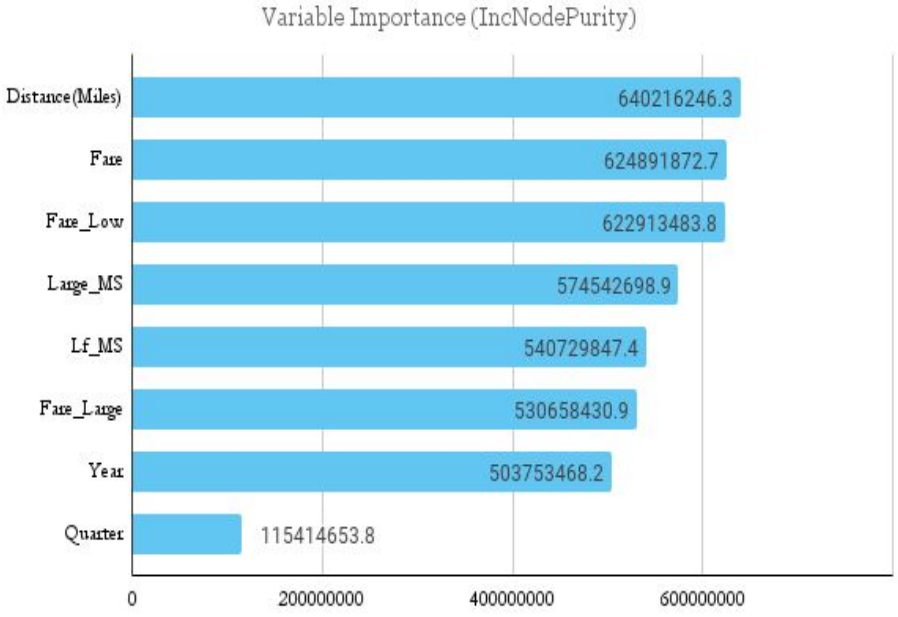
Random Forest

- Randomly selected 20,000 observations from the original dataset of 245,955 observations due to its large size.
- I focused on 9 key variables – 8 variables being the predictors and 1 being the target variable – and applied the random forest algorithm in R studio.
- The model was built with **ntree** set to 600 to ensure for sufficient tree depth/number for accurate predictors, and **mtry** set to 3, which controlled how many predictor variables are considered during each split



Variable Importance





Model(s) Performance

	Training Data	Testing Data
RMSE	199.16	415.05
R-Squared	0.9276	0.3075
MAE	113.32	249.82

Discussions

- The multivariate regression model provides valuable baseline for understanding the relationship between the chosen independent variables (predictors) and the dependent variables (target variables).
- From IncNodePurity, distance, fare, and lowest fare were the top three important variables in shaping or predicting airline demands.
- When considering % Increase in MSE, year, market shares (both largest and lowest fare carrier), and distance were the most important — **Fare-related variables showed lower influence here.**
- Model perform way better on training dataset than on testing dataset, suggesting potential concern of overfitting.
- **For Industry:** Distance, fare, and even market share highlights the importance of route management and fare optimization, as well as competition (as this would bring fare down).

Limitations

- Overfitting in Random Forest
- Unobserved Confounding Variables
- Single-country focus
- Limited explanatory power of the regression model (based on R-squared)

thank YOU