

Forecasting in the U.S. Domestic Airline Industry

Dennis Johnson Arapurayil, Jacob Tassos, Tony Diehl-Verdugo
University of San Diego AAI 501: Introduction to Artificial Intelligence
Professor Andrew Van Benschoten
August 8, 2025
YouTube Link: https://youtu.be/z\$V9K6icnEM



Overview

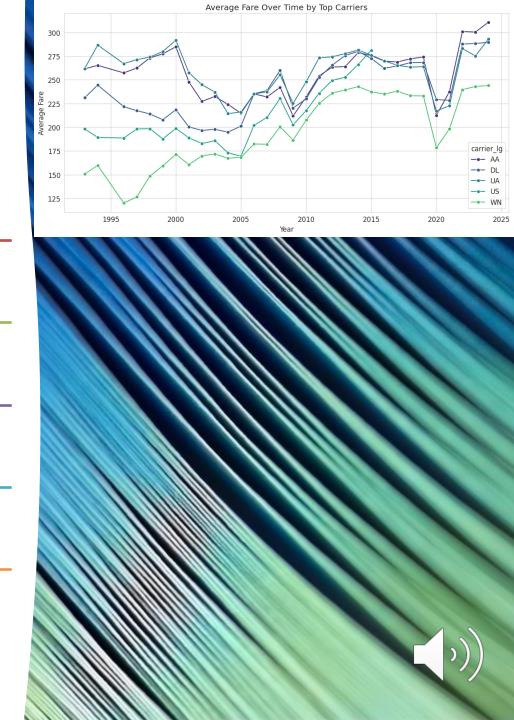
STUDY APPLIES AI FORECASTING TO U.S. DOMESTIC AIRLINE MARKET.

DATASET: 1993–2024 WITH FARES, ROUTES, CARRIERS, PASSENGER VOLUMES.

TESTED: XGBOOST, RANDOM FOREST, ARIMA/SARIMA.

XGBOOST: 90% DIRECTIONAL ACCURACY, MAPE 1.3%.

SUPPORTS STRATEGIC DECISION-MAKING FOR STARTUPS.



Introduction

- Airline industry shaped by fuel prices, consumer preferences, macroeconomics, seasonality.
- Startups lack historical insight; need predictive tools.
- Traditional models miss nonlinear and seasonal patterns.
- Machine learning enables adaptive, data-driven forecasts.



Data Cleaning & Preparation

- Removed >50% missing value columns, duplicates, standardized formats.
- Mapped cities to airports, linked to likely carriers.
- Applied IQR filtering; retained major event shifts like COVID-19.

1.2

Total Revenue

0.4

0.2

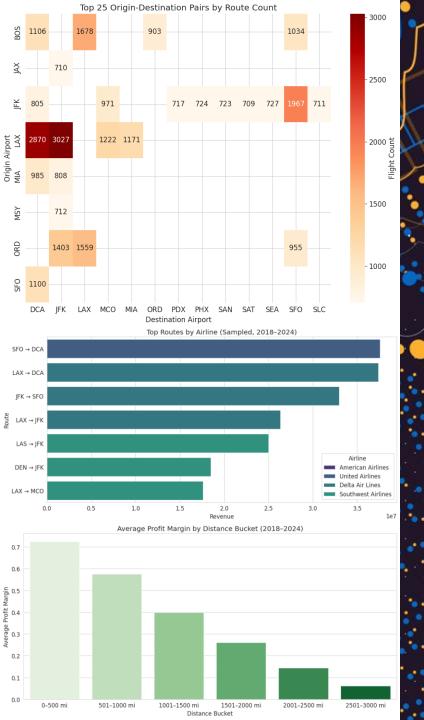
Result: structured quarterly panel dataset.



Large Ms by Year

0.675 0.650

Passengers by Year



Exploratory Data Analysis

- Fare trends by year/region; hubs LAX,
 JFK, ATL dominate.
- Seasonality: leisure destinations spike in Q2/Q3.
- Price variability linked to distance, carrier type.
- Identified route behavior patterns.



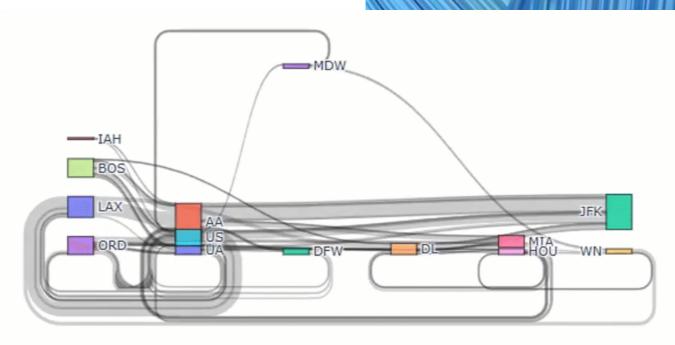
Correlation Heatmap Year 1.00 0.00 0.00 -0.01-0.01-0.03 0.00 0.13 0.16 0.14 0.15 0.12 0.18 0.13 0.14 -0.06 0.14 1.00 0.00 quarter 0.00 1.00 0.00 0.00 0.00 0.00 0.00 -0.00 0.02 -0.02-0.00-0.02-0.00-0.01-0.010.02 -0.010.01 0.00 1.00 - 0.8 citymarketid 1 0.00 0.00 1.00-0.01 0.53 0.03 0.12-0.05 0.06-0.15 0.06-0.07 0.05-0.07-0.09 0.00 -0.04 0.00 0.00 citymarketid 2 -0.010.00 -0.01 1.00 0.04 0.52 0.09 -0.080.01 -0.04 0.00 0.01 0.02 -0.11-0.08 0.01 -0.08-0.010.00 airportid 1 -0.01 0.00 0.53 0.04 1.00 0.05 0.12 -0.04 0.06 -0.12 0.05 -0.06 0.05 -0.08-0.08-0.01-0.03-0.010.00 0.6 airportid 2 -0.03 0.00 0.03 0.52 0.05 1.00 0.19 -0.02 0.07 -0.09 0.06 -0.03 0.07 -0.17-0.08 0.03 -0.01-0.03 0.00 nsmiles 0.00-0.00 0.12 0.09 0.12 0.19 1.00 0.06 0.54-0.37 0.51-0.22 0.45 0.71-0.29-0.02 0.15 0.00 -0.00 0.4 passengers 0.13 0.02 -0.05-0.08-0.04-0.02 0.06 1.00 -0.05-0.03 0.00 -0.16-0.11-0.13 0.81 -0.11 0.94 0.13 0.02 fare 0.16-0.020.06 0.01 0.06 0.07 0.54-0.05 1.00-0.20 0.95-0.20 0.86 0.03-0.24 0.14 0.12 0.16-0.02 -0.2large ms 0.14-0.00-0.15-0.04-0.12-0.09-0.37-0.03-0.20 1.00-0.19 0.52-0.100.28 0.14 0.13-0.060.14-0.00 fare lg 0.15-0.02 0.06 0.00 0.05 0.06 0.51 0.00 0.95-0.19 1.00-0.25 0.81 0.03-0.19 0.10 0.17 0.15-0.02 -0.0If ms 0.12-0.00-0.070.01-0.06-0.03-0.22-0.16-0.200.52-0.25 1.00 0.05 0.12-0.040.18-0.200.12-0.00 fare low 0.18-0.01 0.05 0.02 0.05 0.07 0.45-0.11 0.86-0.10 0.81 0.05 1.00 0.05-0.26 0.15 0.03 0.18-0.01 fare_per_mile 0.13-0.01-0.07-0.11-0.08-0.17-0.71-0.130.03 0.28 0.03 0.12 0.05 1.00 0.17 0.10-0.100.13-0.01 -0.2sengers_per_mile 0.14 0.02 -0.09-0.08-0.08-0.08-0.29 0.81 -0.240.14 -0.19-0.04-0.260.17 1.00 -0.11 0.67 0.14 0.02 e per passenger -0.06-0.01 0.00 0.01 -0.010.03 -0.02-0.110.14 0.13 0.10 0.18 0.15 0.10 -0.11 1.00 -0.09-0.06-0.01 revenue 0.14 0.01 -0.04-0.08-0.03-0.010.15 0.94 0.12 -0.06 0.17 -0.20 0.03 -0.10 0.67 -0.09 1.00 0.14 0.01 year num 1.00 0.00 0.00 -0.01-0.01-0.03 0.00 0.13 0.16 0.14 0.15 0.12 0.18 0.13 0.14 -0.06 0.14 1.00 0.00 -0.6quarter num 0.00 1.00 0.00 0.00 0.00 0.00 -0.00 0.02 -0.02-0.00-0.02-0.00-0.01-0.010.02 -0.010.01 0.00 1.00 nsmiles airportid_1 revenue **passengers** passengers per mile are_per_passenger year_num citymarketid_2 citymarketid 1

Pearson correlation analysis guided feature selection

Strong revenue-passenger correlation

Fare per mile tied to route length

Visualization of Patterns

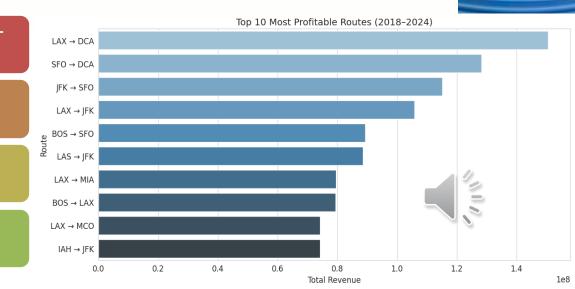


Sankey: top city pairs (LAX-LAS, ATL-JFK).

Heatmaps: market saturation and gaps.

Seasonality plots: recurring peaks.

Segmented routes into business, leisure, premium categories.



Model Selection Process (1)

```
Testing 222 configurations for target variable: log_revenue...

Test 50: Best model for log_revenue: XGBoostRegression[lag_4+lag_16], MAPE: 1.7%, R²: 0.011

50/222 configs tested...

Test 100: Best model for log_revenue: RandomForest[lag_1+lag_2+lag_3+lag_4], MAPE: 1.5%, R²: 0.355

100/222 configs tested...

Test 150: Best model for log_revenue: XGBoostRegression[quarter+lag_1+lag_4+rolling_mean_4], MAPE: 1.6%, R²: 0.168

150/222 configs tested...

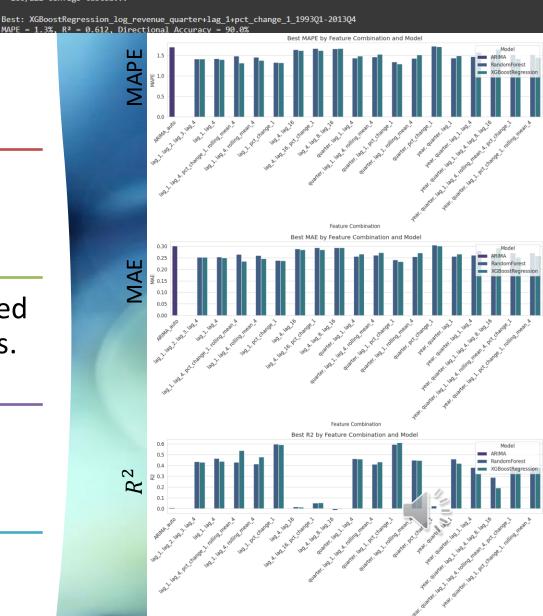
Test 200: Best model for log_revenue: XGBoostRegression[year+quarter+lag_1+lag_4], MAPE: 1.7%, R²: 0.002

200/222 configs tested...
```

Models: ARIMA/SARIMA, Random Forest, XGBoost.

222 configurations with varied features and training periods.

Metrics(1): MAPE, MAE, and R²



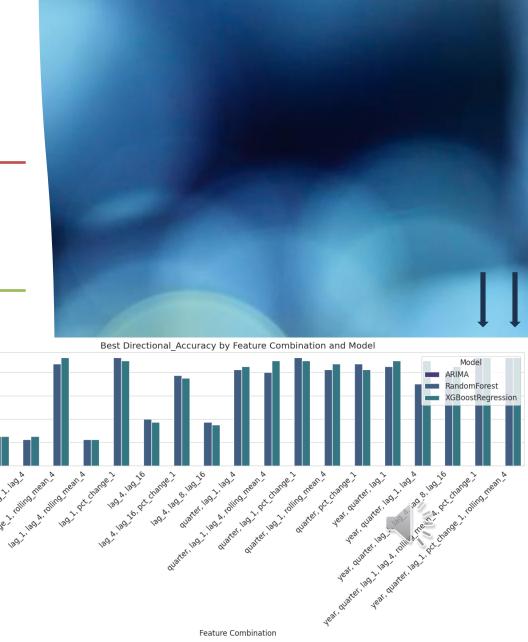
Feature Combination

Model Selection Process (2)

Metrics (2): Directional Accuracy

Time Series Forecasting Features

DA prioritized for final selection.



Why XGBoost Excelled

TOP 10 BY MAPE (Ascending)

SEQUENTIAL TREE-BUILDING WITH REGULARIZATION.

CAPTURES NONLINEAR
PATTERNS WITH
LAGGED/SEASONAL FEATURES.

LOWEST MAPE, HIGH DA PERFORMANCE.

ADAPTABLE TO MARKET VOLATILITY.

```
XGBoost_log_revenue_quarter+lag_1+pct_change_1_1993Q1-2013Q4
                                                                                                   1.29 0.61
                  XGBoost_log_revenue_lag_1+lag_4+pct_change_1+rolling_mean_4_1993Q1-2013Q4
                                        XGBoost log revenue lag 1+pct change 1 1993Q1-2013Q4
                                        RandomF log revenue lag 1+pct change 1 1993Q1-2013Q4
                               RandomF log revenue quarter+lag 1+pct change 1 1993Q1-2013Q4
                               XGBoost log revenue lag 1+lag 4+rolling mean 4 1993Q1-2013Q4
                                               XGBoost log revenue lag 1+lag 4 1993Q1-2013Q4
                                   RandomF log revenue lag 1+lag 2+lag 3+lag 4 1993Q1-2013Q4
                                   XGBoost_log_revenue_lag_1+lag_2+lag_3+lag_4_1993Q1-2013Q4
                                                                                                   1.41 0.43
                                                                                                                      17.5
                                                                                                   1.42 0.38
     XGBoost log revenue year+quarter+lag 1+lag 4+rolling mean 4+pct change 1 1993Q1-2013Q4
  OP 10 BY R2 (Descending)
                                                                  Configuration
                                                                                   R<sup>2</sup> MAPE (%) Dir Acc (%)
                   XGBoost_log_revenue_quarter+lag_1+pct_change_1_1993Q1-2013Q4 0.61
                                                                                            1.29
                          RandomF log revenue lag 1+pct change 1 1993Q1-2013Q4
                                                                                                         92.5
                   RandomF log revenue quarter+lag 1+pct change 1 1993Q1-2013Q4
                                                                                                         90.0
                          XGBoost log revenue lag 1+pct change 1 1993Q1-2013Q4
                                                                                            1.32
                                                                                                         90.0
     XGBoost log revenue lag 1+lag 4+pct change 1+rolling mean 4 1993Q1-2013Q4
                                                                                            1.31
                                                                                                         90.0
                                                                                                         22.5
                  XGBoost log revenue lag 1+lag 4+rolling mean 4 1993Q1-2013Q4 0.48
                                                                                            1.38
                                                                                                         22.5
                                  RandomF_log_revenue_lag_1+lag_4_1993Q1-2013Q4 0.47
                                                                                            1.42
                          RandomF_log_revenue_quarter+lag_1+lag_4_1993Q1-2013Q4 0.46
                                                                                            1.43
                                                                                                         80.0
                          RandomF_log_revenue_year+quarter+lag_1_1993Q1-2013Q4 0.46
                                                                                            1.43
                                                                                                         67.5
                          XGBoost log revenue quarter+lag 1+lag 4 1993Q1-2013Q4 0.46
                                                                                                         80.0
                                        Best MAPE by Target Variable and Model Type
                                                                                                         Model Type
  1.6
                                                                                                        RandomForest
  1.4
                                                                                                       XGBoostRegression
  1.2
Best MAPE
  1.0
  0.6
  0.4
  0.2
```

Target Variable

Model Performance

XGBoost: MAPE 1.3%, R² 0.612, DA 90%.

Random Forest: similar MAPE, lower DA.

ARIMA/SARIMA: higher error, weaker trends.

Top features: lagged revenue, rolling averages, seasonal flags.



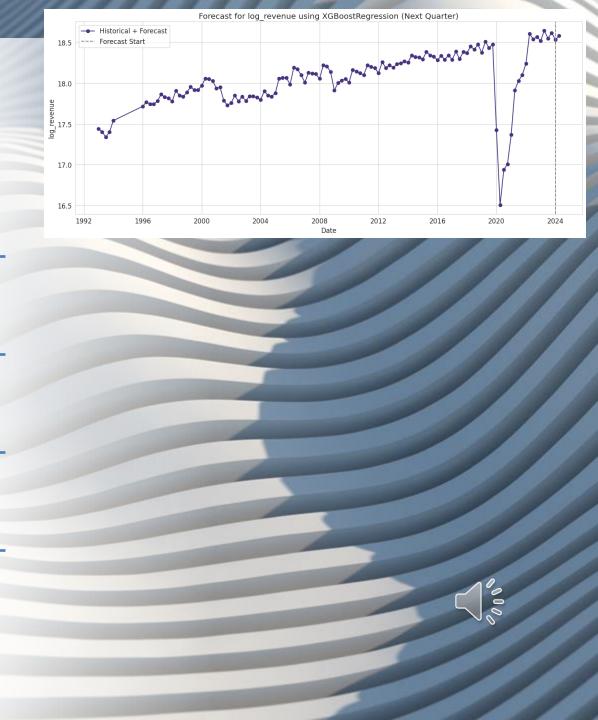
Model Analysis & Forecast Example

Training: 2018Q1–2023Q4.

Forecasted next quarter without look-ahead bias.

DA drives proactive route planning.

Identified high-profit routes: LAX-DCA, SFO-DCA, JFK-SFO.



Conclusion

XGBoost: best performer balancing error and DA.

Handles nonlinear, complex relationships.

Directional accuracy key for strategic planning.

Supports segmentation-aware pricing and route optimization.



Recommendations



ENHANCE SEASONAL FEATURES: HOLIDAYS, WEATHER, SCHOOL CALENDARS.



INTEGRATE MACROECONOMIC DATA: FUEL, INFLATION, CONSUMER SPENDING.



References

Breiman, L. Random Forests. Machine Learning 45, 5-32 (2001).

https://doi.org/10.1023/A:1010933404324

Chen, T., & Guestrin, C. (2016). XGBoost. Proceedings of the 22nd ACM SIGKDD

International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785

denjcodes, diehl64, & jtassos. (2025). AAA-501-Group-3 [Source code].

GitHub. https://github.com/denjcodes/AAA-501-Group-3

Geeks for Geeks. (2025, July 23). What is lag in time series forecasting.

https://www.geeksforgeeks.org/machine-learning/what-is-lag-in-time-series-forecasting/

Hyndman, R. J., Athanas opoulos, G., Garza, A., Challu, C., Mergenthaler, M., & Olivares,

K. G. (2025, July 27). Forecasting: Principles and Practice, the Pythonic Way. https://otexts.com/fpppy/

Jikadara, B. (2024, August 4). US airline flight routes and fares 1993-2024. Kaggle.

https://www.kaggle.com/datasets/bhavikjikadara/us-airline-flight-routes-and-fares-1993-2024

Lemke, C., & Gabrys, B. (n.d.). Forecasting and Forecast Combination in Airline Revenue Management Applications.

Surakhi, O., Zaidan, M. A., Fung, P. L., Hossein Motlagh, N., Serhan, S., AlKhanafseh, M.,

Ghoniem, R. M., & Hussein, T. (2021). Time-lag selection for time-series forecasting using neural network and heuristic algorithm. *Electronics*, 10(20), 2518. https://doi.org/10.3390/electronics10202518

Arapurayil, D. J., Tassos, J. A., & Diehl Verdugo, T. A. (n.d.). Forecasting in the U.S. domestic airline industry [Video]. YouTube. https://youtu.be/zSV9K6icnEM

