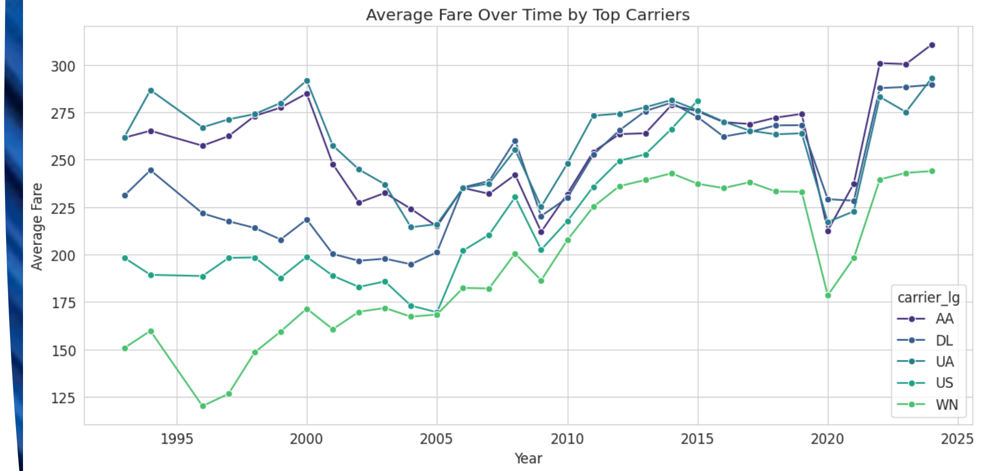
Forecasting in the U.S. Domestic Airline Industry

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/zSV9K6icnEM**

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%.

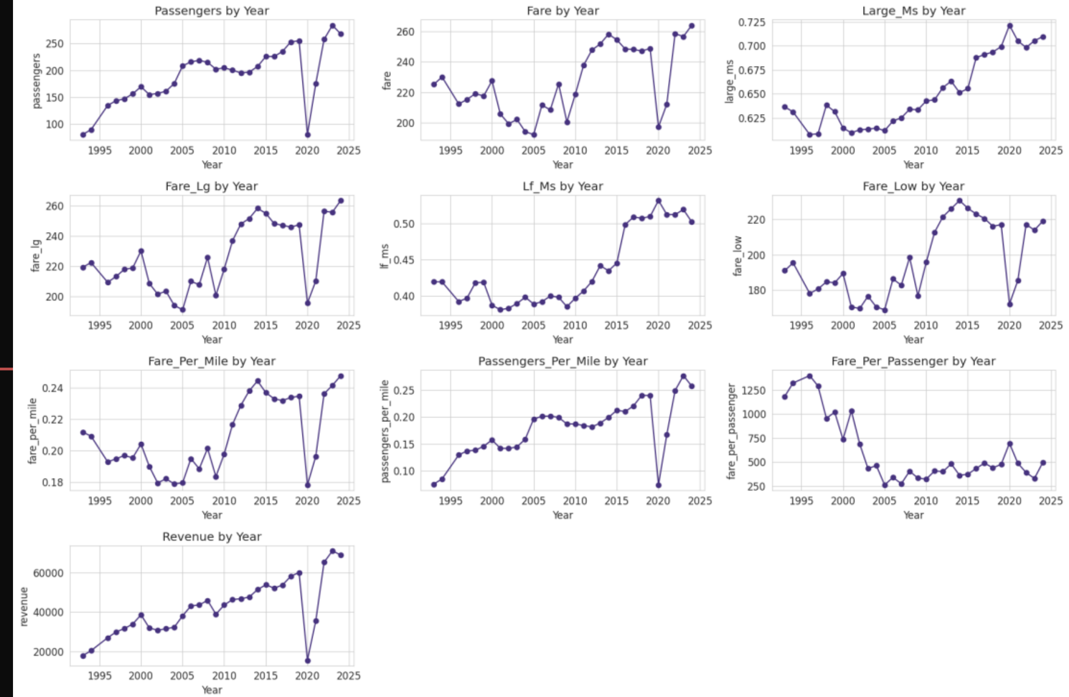
SUPPORTS STRATEGIC DECISION-MAKING FOR STARTUPS.

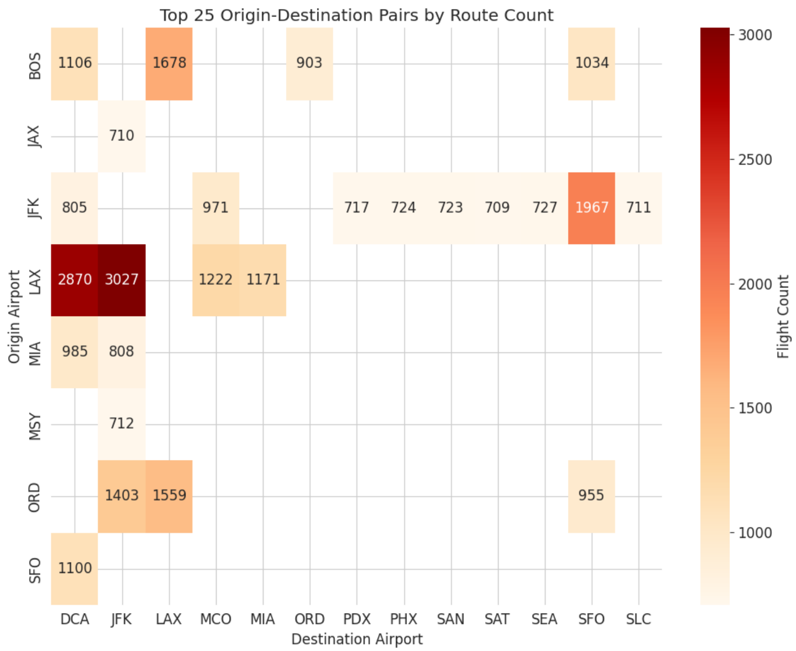
Introduction

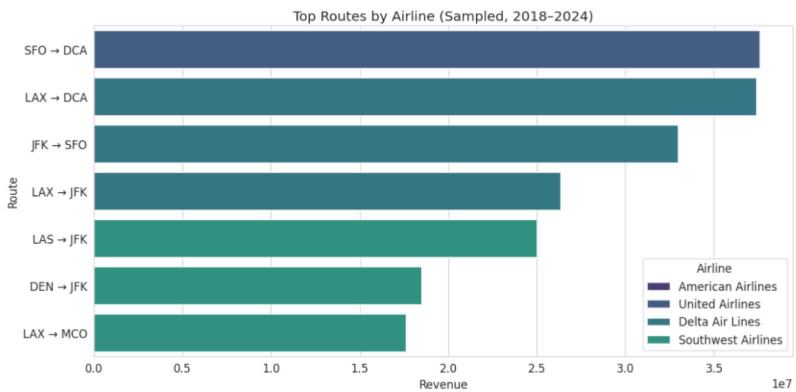
- Airline industry shaped by fuel prices, consumer preferences, macroeconomics, seasonality.
- Startups lack historical insight; need predictive tools.
- Traditional models miss nonlinear and seasonal patterns.
- Machine learning enables adaptive, data-driven forecasts.

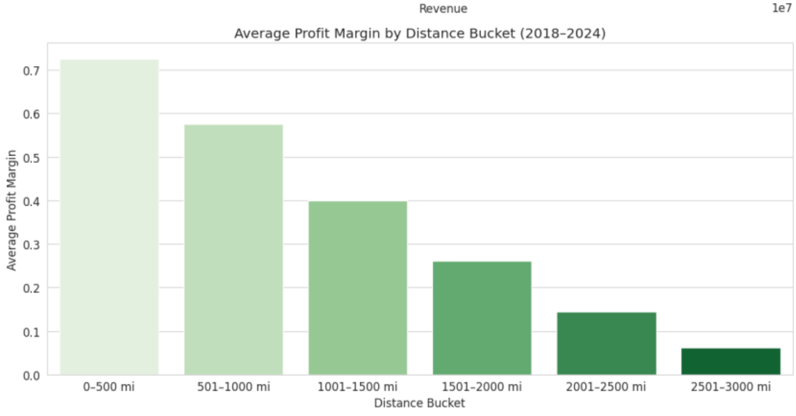
Data Cleaning & Preparation

- 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.
- Result: structured quarterly panel dataset.





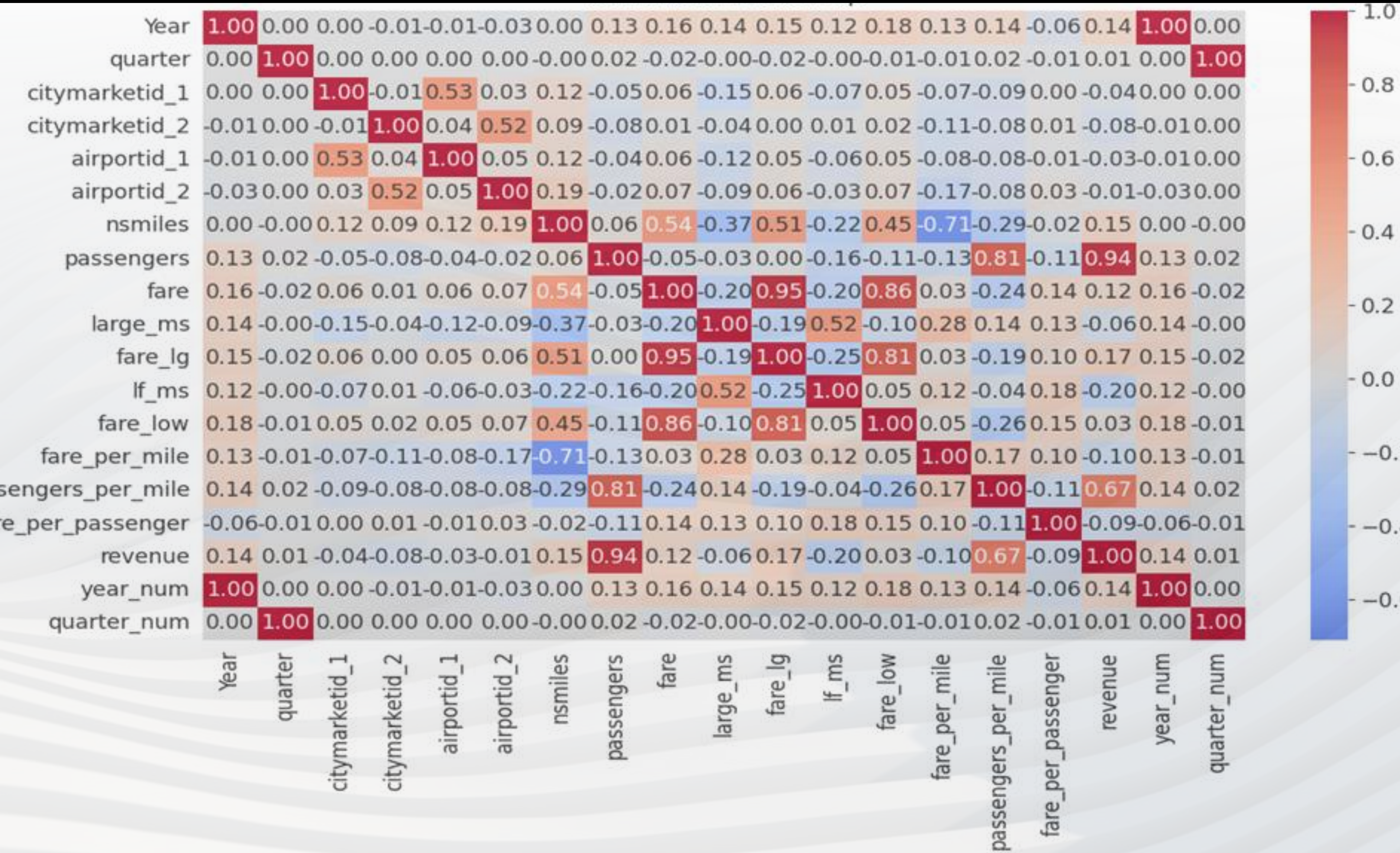




Exploratory Data Analysis

- 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

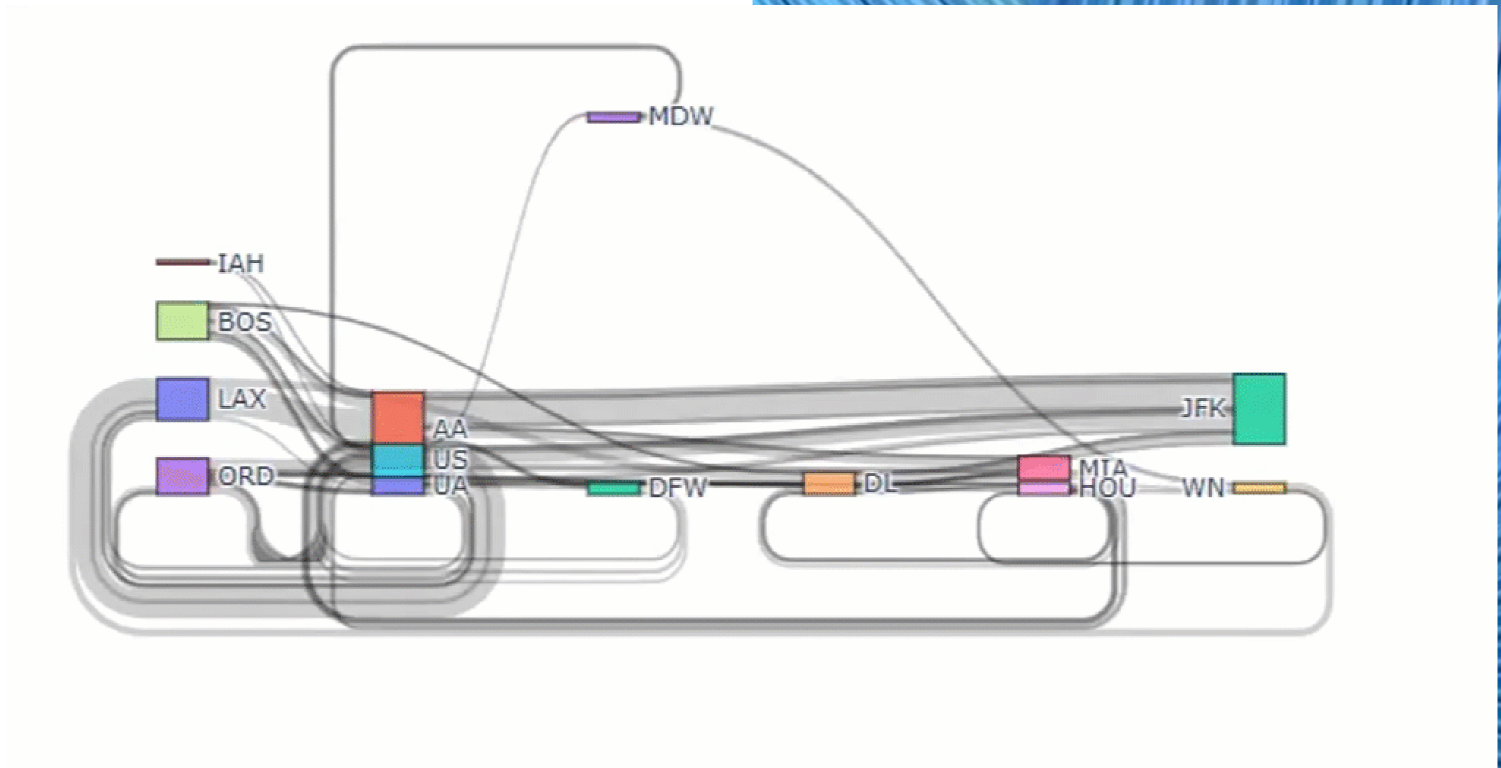


Pearson correlation analysis guided feature selection

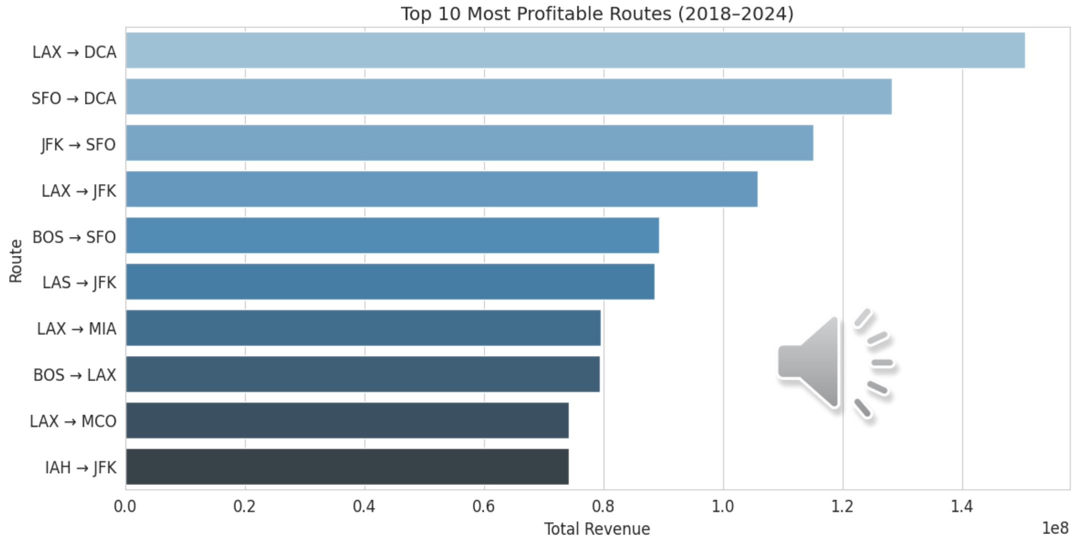
Strong revenue-passenger correlation

Fare per mile tied to route length

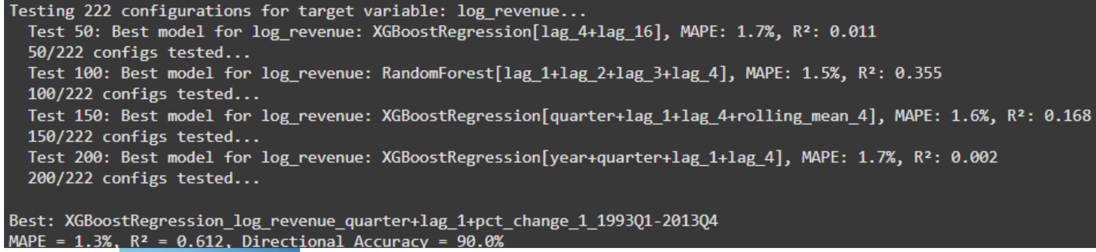
Visualization of Patterns



- Sankey: top city pairs (LAX–LAS, ATL–JFK).
- Heatmaps: market saturation and gaps.
- Seasonality plots: recurring peaks.
- Segmented routes into business, leisure, premium categories.



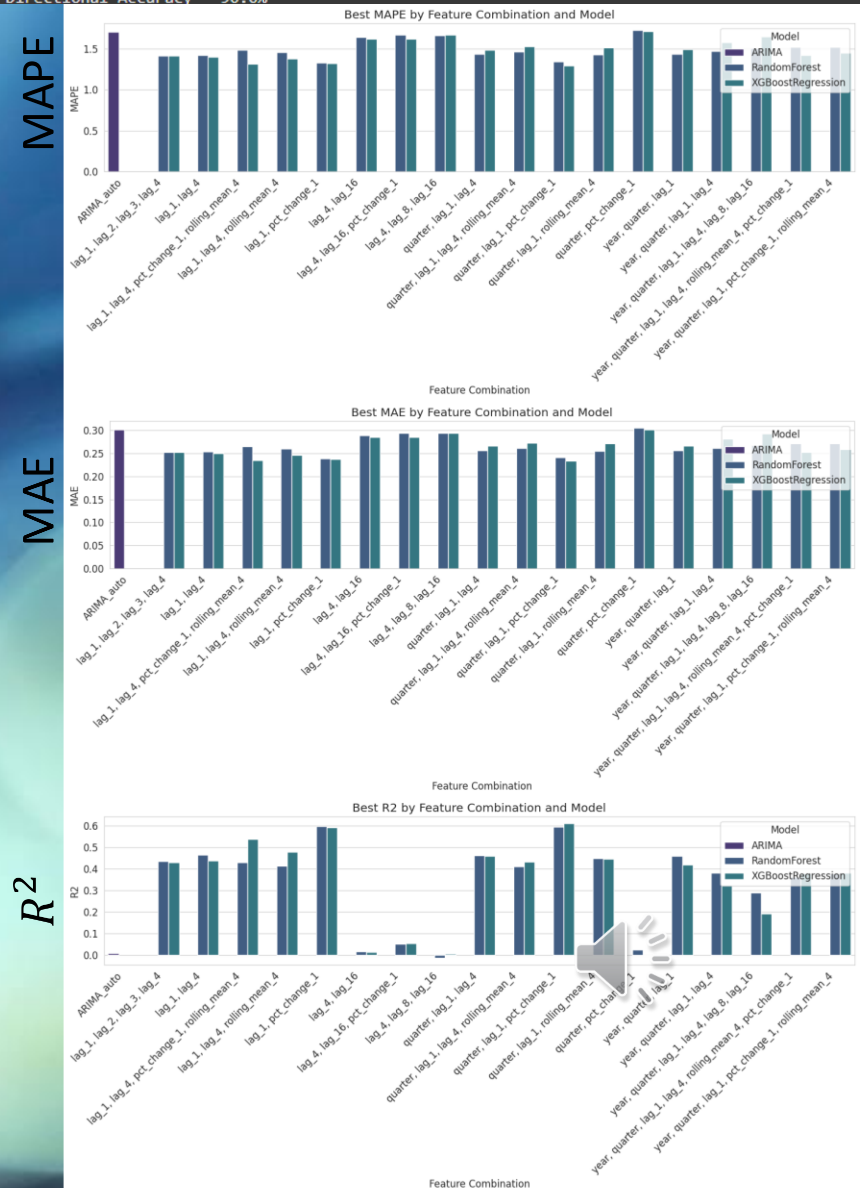
Model Selection Process (1)



```
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...

Best: XGBoostRegression_log_revenue_quarter+lag_1+pct_change_1_1993Q1-2013Q4
MAPE = 1.3%, R² = 0.612, Directional Accuracy = 90.0%
```

- Models: ARIMA/SARIMA, Random Forest, XGBoost.
- 222 configurations with varied features and training periods.
- Metrics(1): MAPE, MAE, and R^2

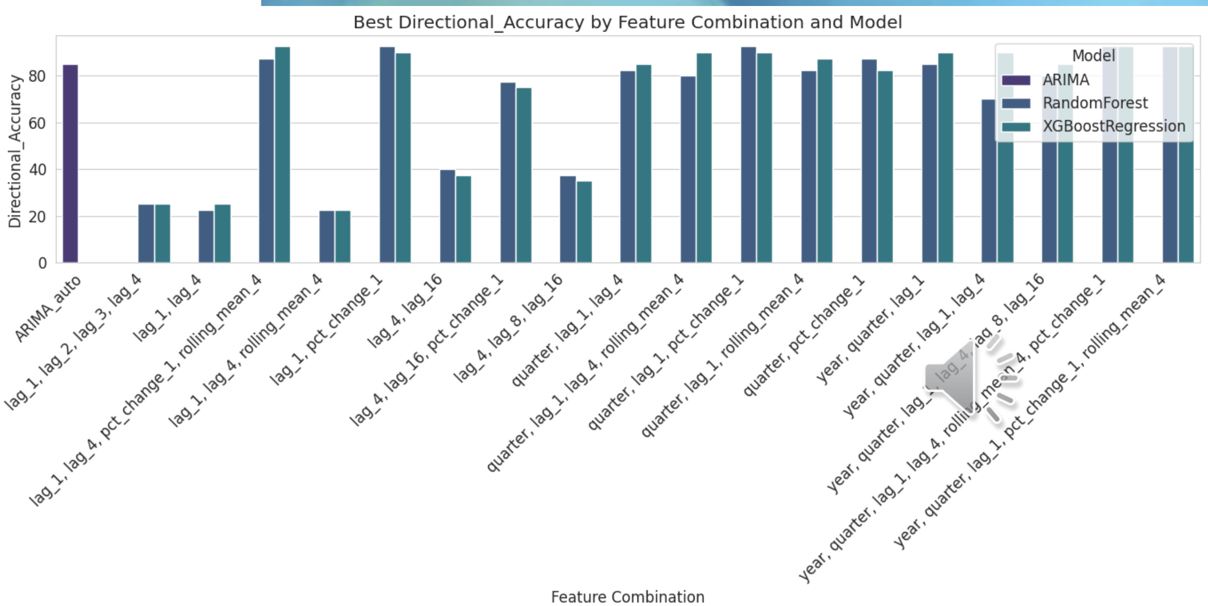


Model Selection Process (2)

Metrics (2): Directional Accuracy

Time Series Forecasting Features

DA prioritized for final selection.



Why XGBoost Excelled

SEQUENTIAL TREE-BUILDING WITH REGULARIZATION.

CAPTURES NONLINEAR PATTERNS WITH LAGGED/SEASONAL FEATURES.

LOWEST MAPE, HIGH DA PERFORMANCE.

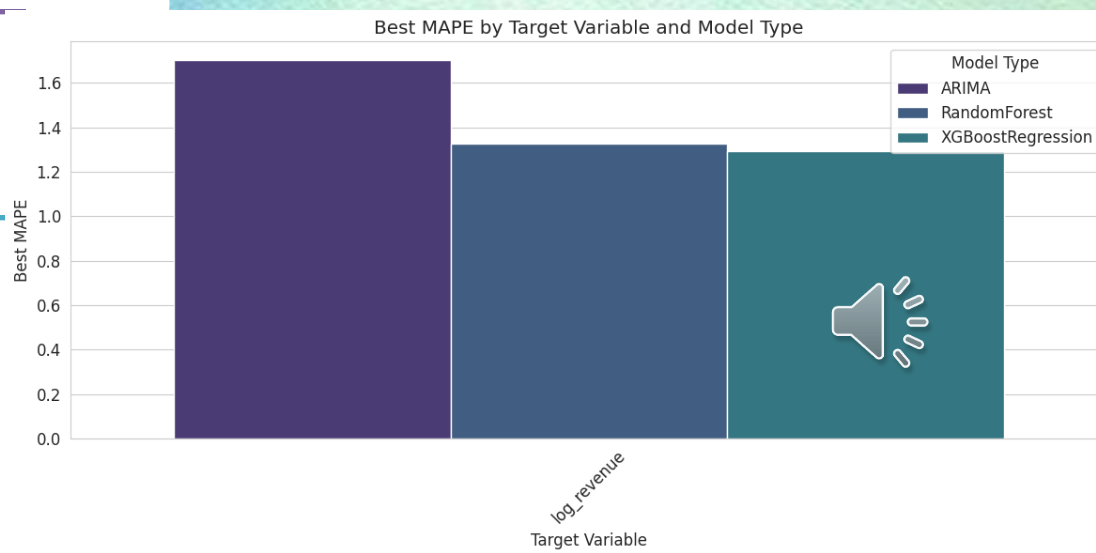
ADAPTABLE TO MARKET VOLATILITY.

TOP 10 BY MAPE (Ascending)

	Configuration	MAPE (%)	R^2	Dir_Acc (%)
1	XGBoost_log_revenue_quarter+lag_1+pct_change_1_1993Q1-2013Q4	1.29	0.61	90.0
2	XGBoost_log_revenue_lag_1+lag_4+pct_change_1+rolling_mean_4_1993Q1-2013Q4	1.31	0.54	90.0
3	XGBoost_log_revenue_lag_1+pct_change_1_1993Q1-2013Q4	1.32	0.59	90.0
4	RandomF_log_revenue_lag_1+pct_change_1_1993Q1-2013Q4	1.33	0.60	92.5
5	RandomF_log_revenue_quarter+lag_1+pct_change_1_1993Q1-2013Q4	1.34	0.60	90.0
6	XGBoost_log_revenue_lag_1+lag_4+rolling_mean_4_1993Q1-2013Q4	1.38	0.48	22.5
7	XGBoost_log_revenue_lag_1+lag_4_1993Q1-2013Q4	1.40	0.44	25.0
8	RandomF_log_revenue_lag_1+lag_2+lag_3+lag_4_1993Q1-2013Q4	1.41	0.44	22.5
9	XGBoost_log_revenue_lag_1+lag_2+lag_3+lag_4_1993Q1-2013Q4	1.41	0.43	17.5
10	XGBoost_log_revenue_year+quarter+lag_1+lag_4+rolling_mean_4+pct_change_1_1993Q1-2013Q4	1.42	0.38	87.5

TOP 10 BY R2 (Descending)

	Configuration	R^2	MAPE (%)	Dir_Acc (%)
1	XGBoost_log_revenue_quarter+lag_1+pct_change_1_1993Q1-2013Q4	0.61	1.29	90.0
2	RandomF_log_revenue_lag_1+pct_change_1_1993Q1-2013Q4	0.60	1.33	92.5
3	RandomF_log_revenue_quarter+lag_1+pct_change_1_1993Q1-2013Q4	0.60	1.34	90.0
4	XGBoost_log_revenue_lag_1+pct_change_1_1993Q1-2013Q4	0.59	1.32	90.0
5	XGBoost_log_revenue_lag_1+lag_4+pct_change_1+rolling_mean_4_1993Q1-2013Q4	0.54	1.31	90.0
6	XGBoost_log_revenue_lag_1+lag_4+rolling_mean_4_1993Q1-2013Q4	0.48	1.38	22.5
7	RandomF_log_revenue_lag_4_1993Q1-2013Q4	0.47	1.42	22.5
8	RandomF_log_revenue_quarter+lag_1+lag_4_1993Q1-2013Q4	0.46	1.43	80.0
9	RandomF_log_revenue_year+quarter+lag_1_1993Q1-2013Q4	0.46	1.43	67.5
10	XGBoost_log_revenue_quarter+lag_1+lag_4_1993Q1-2013Q4	0.46	1.48	80.0



Model Performance

```
XGBoostRegression:
  Best MAPE: 1.3%
  Best R²: 0.612
  Best Directional Accuracy: 90.0%
```

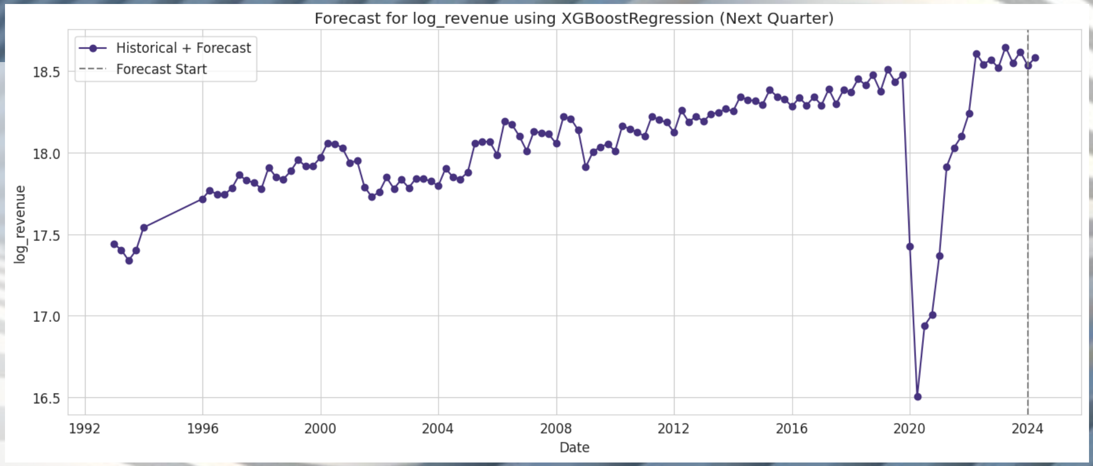
XGBoost: MAPE 1.3%, R^2 0.612, DA 90%.

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

GeeksforGeeks. (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., Athanasopoulos, 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