Forecasting U.S. Domestic Flight Prices:

A Predictive Model
Using Historical Data and Seasonal Trends

Group 5

Thao Nguyen, Maralmaa Batnasan, Hazel Foo
Arial Huang, Hana Kalinova

PROJECT SUMMARY

Executive Summary

Goal: ***Develop a robust and accurate model for flight price prediction based on historical data to support strategic pricing decisions.***

Findings:

- Data Preprocessing and Feature Engineering: Introduce dummy variables and lagged features to capture temporal patterns, seasonality, and route-specific aspects.
- Modeling and Model Selection: Experiment with time-series and tree-based models and select XGBoost as final model
- Understanding Feature Importance:
  - Short-haul flight impact pricing strategies
  - Market shares of low fare carriers and popular carriers shape pricing decisions
  - Specific regional markets have stronger impacts on price

Executive Summary
Business Implications

Strategic Pricing
Prioritize pricing strategies for short haul flights and focus on route-specific adjustments/expansions for popular regional markets

Market Share Insights:
highlight the need to monitor competitive dynamics to inform pricing strategies

Time-Dependency:
highlight the need to incorporate quarterly and yearly trends into pricing decisions

Key Analytical Problems

Predictive Modeling

Seasonal and Year-Over-Year Analysis

Stakeholder Business Value

DATASET

Data Description

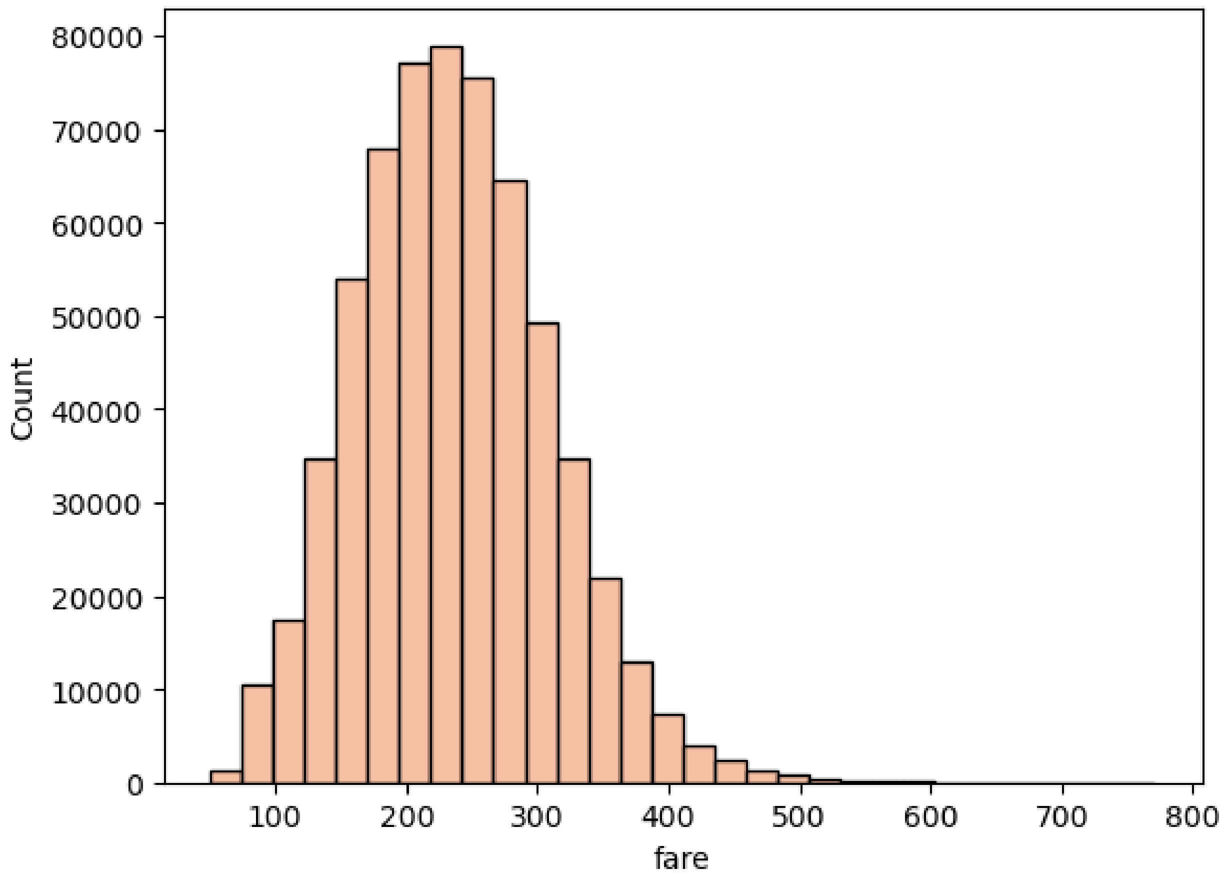
- **Data source:** U.S. Department of Transportation - Domestic Airline Consumer Airfare Report
- **Target variable:** Average fare
- **Potential predictors:**
  Year, quarter, city1 and city 2 (directionless), non-stop market miles, passenger per day, Carrier with the Largest Market Share/Percentage of Share/Average fare, Carrier with the Lowest Average Fare/Percentage of Share/Average fare

Data Description

Year/Quarter	Airfare prices fluctuate seasonally and over time due to economic conditions, holidays, and other factors like oil prices and demand.
Global Price of Brent Crude	Fuel is a significant operational cost for airlines, and fluctuations in crude oil prices directly affect ticket prices.
City pair	The specific cities being connected affect pricing due to demand, competition, and market size.
Non-stop Market Miles	Distance is a critical factor in airfare pricing, with longer routes generally costing more.
Passenger per Day	Higher demand typically leads to higher prices.
Overall Average Fare	Baseline fare used to compare with specific carriers and identify trends.
Carrier with the Largest Market Share/Market Share/Average fare	The dominant carrier can influence prices significantly due to market control.
Carrier with the Lowest Fare/Market Share/Average fare	Price competition is key, and low-cost carriers often drive overall fare reductions.
Fare	Average fare of the indicated year, quarter, and city pair

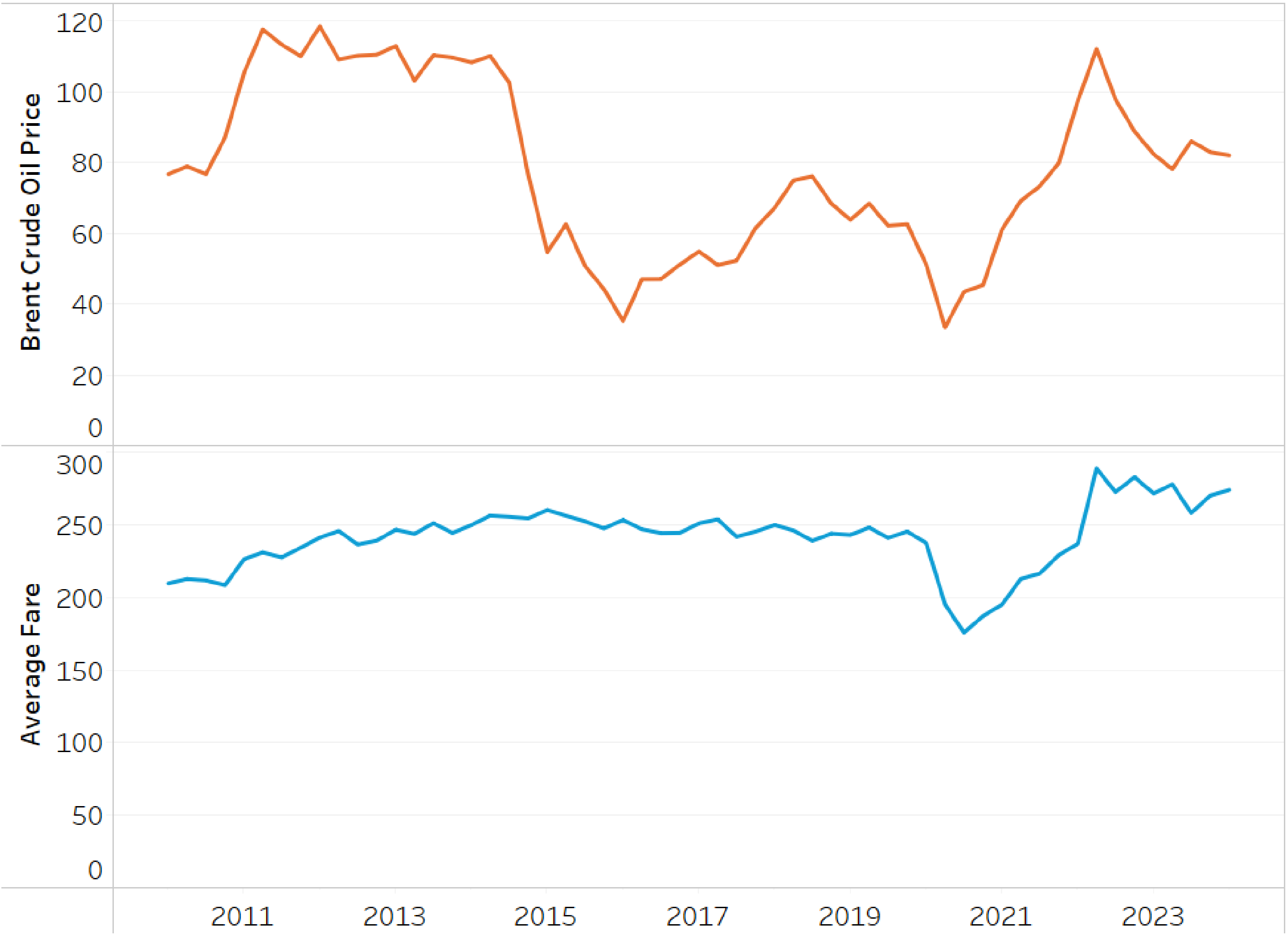
Data Description

Dataset	
Row count	617,305
Unique route (city pair)	9,694
Target Variable: Fare	
mean	236.3
min	50.45
25%	184.05
50%	232.54
75%	283.11
max	770.65



Pre-processing & Feature Engineering

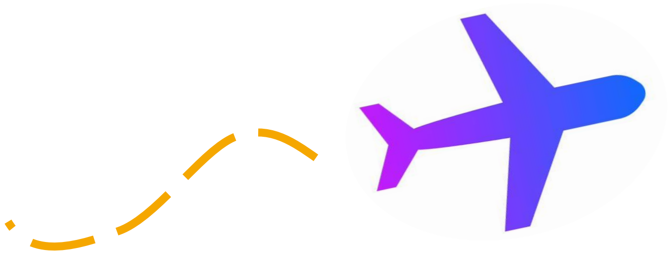
- Add external data – [brent crude oil price]
- Remove routes lacking complete data from 2010Q1 to 2024Q1
- Dummify year, quarter, and city pairs



Pre-processing & Feature Engineering

Map nsmiles to haul category	
~900 miles	Short haul
901~2200 miles	Medium haul
2201~ miles	Long haul

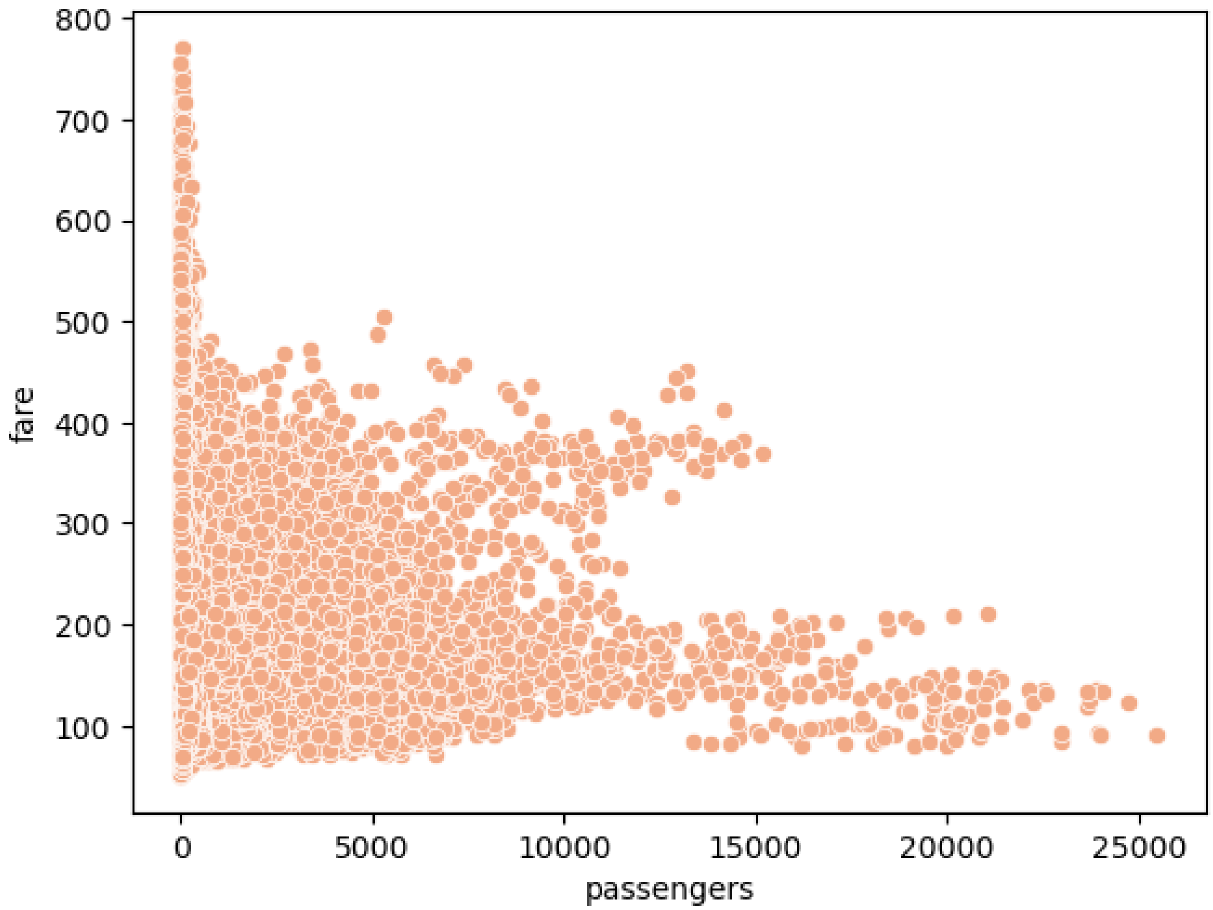
Map carriers to service category	
DL, AA, AS, UA, NW, US, CO	Full service
FL, B6, VX, WN, YX, U5	lcc
MX, G4, XP, SY, NK, F9	Ultra lcc



Pre-processing & Feature Engineering

- Remove outliers

Average passengers per day	
mean	181
min	10
25%	17
50%	33
75%	100
max	25,471

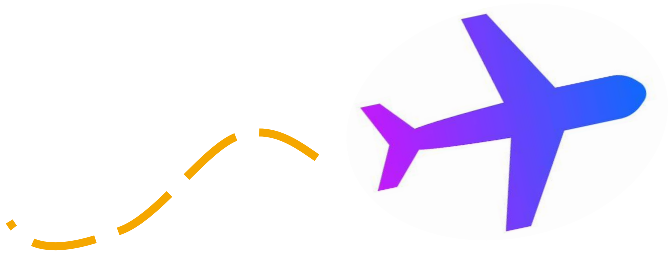


Pre-processing & Feature Engineering

- Fare difference

  (carrier with largest market share / carrier with lowest avg fare)

Fare	Average fare of carrier with largest market share	Fare difference
100	110	10



Pre-processing & Feature Engineering

- Lagged features: using historical data to predict future fare

Year, quarter	Fare	Lagged 1 fare	pctchange	Lagged 1 pctchange
2022 Q1	100	-	-	-
2022 Q2	120	100	0.2	-
2022 Q3	150	120	0.25	0.2
2022 Q4	150	150	0	0.25
2023 Q1	120	150	-0.2	0
2023 Q2	180	120	0.5	-0.2
2023 Q3	200	180	0.11	0.5
2023 Q4	200	200	0	0.11

- Lagged 1 fare (2022 Q3)

  = Fare (2022 Q2)
- pctchange (2023 Q3)

  $= \frac{Fare\ (2023\ Q3)}{Fare\ (2023\ Q2)} - 1 = 0.25$
- Lagged 1 pctchange (2022 Q3)

  = Lagged 1 pctchange (2022 Q2)

Pre-processing & Feature Engineering

- Lagged features: using historical data to predict future fare

Year, quarter	Fare	Lagged 4 fare	pctchange	Lagged 4 pctchange
2021 Q1	80	-	-	-
2021 Q2	100	-	-	-
2021 Q3	100	-	-	-
2021 Q4	120	-	-	-
2022 Q1	100	80	0.25	-
2022 Q2	120	100	0.2	-
2022 Q3	150	100	0.5	-
2022 Q4	150	120	0.25	-
2023 Q1	120	100	0.2	0.25
2023 Q2	180	120	0.5	0.2
2023 Q3	200	150	0.33	0.5
2023 Q4	200	150	0.33	0.25

- Lagged 4 fare (2023 Q1)

  = Fare (2022 Q1)

- pctchange (2023 Q1)

  $$= \frac{Fare\ (2023\ Q1)}{Fare\ (2022\ Q1)} - 1 = 0.2$$

- Lagged 4 pctchange (2023 Q1)

  = Lagged 4 pctchange (2022 Q1)

Feature Selection

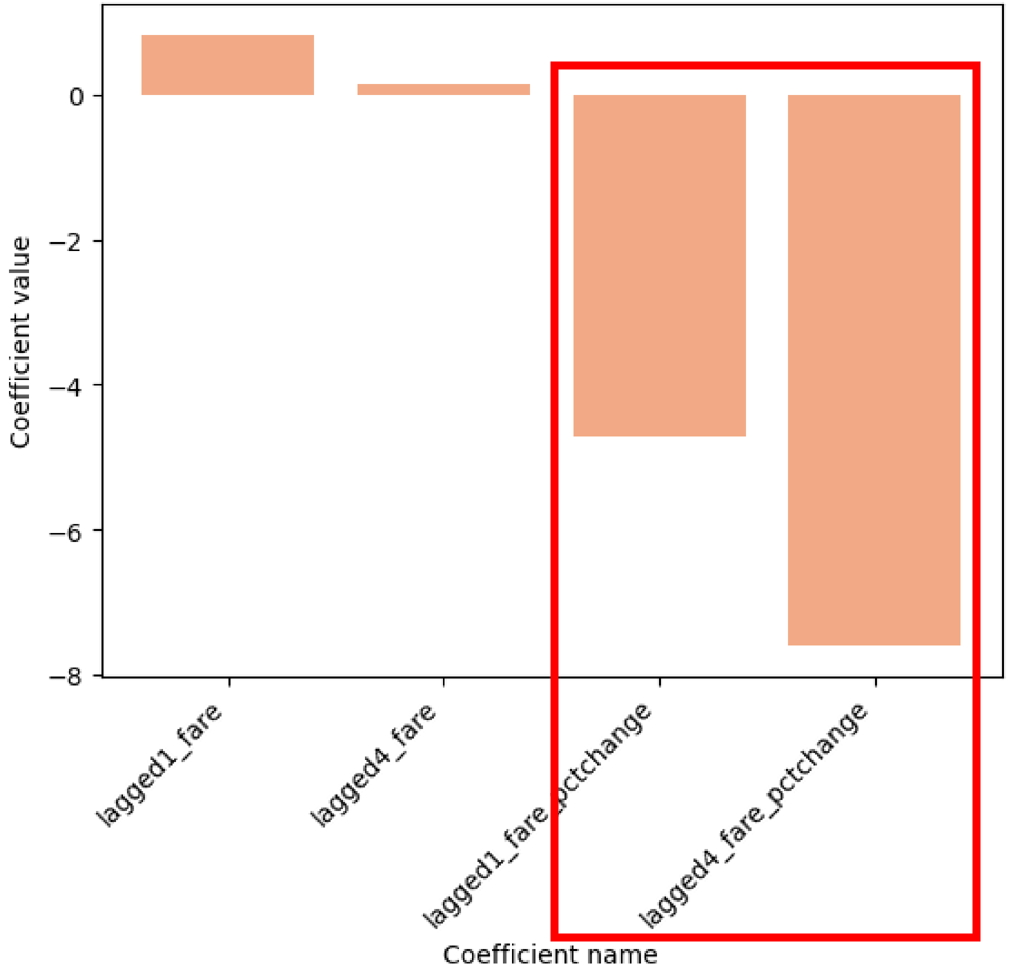
```
X = df_sample[['lagged1_fare', 'lagged4_fare',
'lagged1_fare_pctchange', 'lagged4_fare_pctchange']]
y = df_sample['fare']

model = LinearRegression()
model.fit(X, y)

fig, ax = plt.subplots()
ax.bar(X.columns, model.coef_)
ax.set(xlabel='Coefficient name', ylabel='Coefficient value')

plt.setp(ax.get_xticklabels(), rotation=45,
horizontalalignment='right')
plt.show()
```

Larger absolute values of coefficients mean that a given feature has a large impact on the output variable



Feature Selection

Dummies

year, quarter, city pairs, haul category

Lagged 1

carrier service category, fare difference, market share of carrier with lowest fare

Lagged 1 pctchange

fare, passengers, brent crude

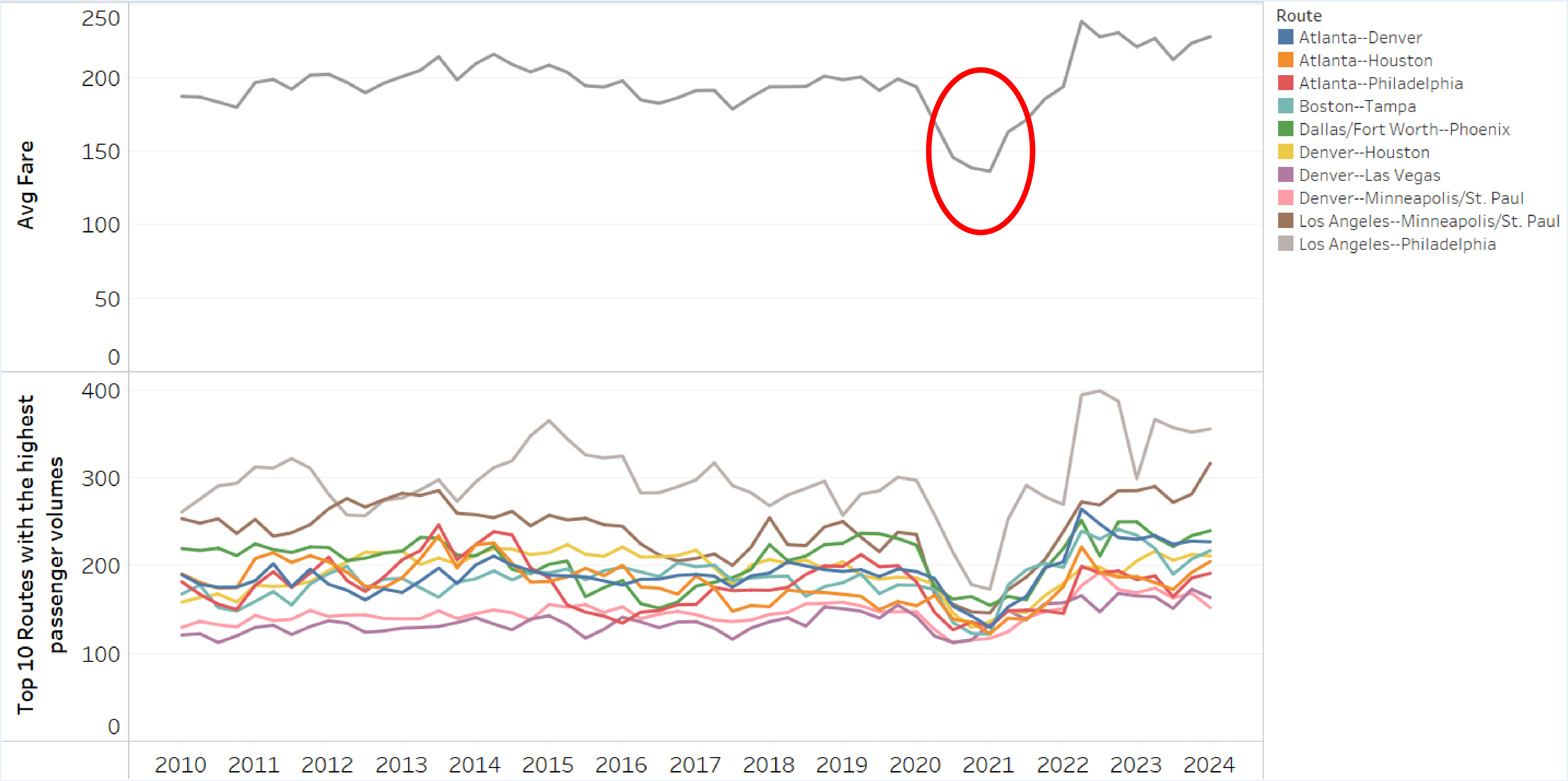
Lagged 4

market share of carrier with largest market share

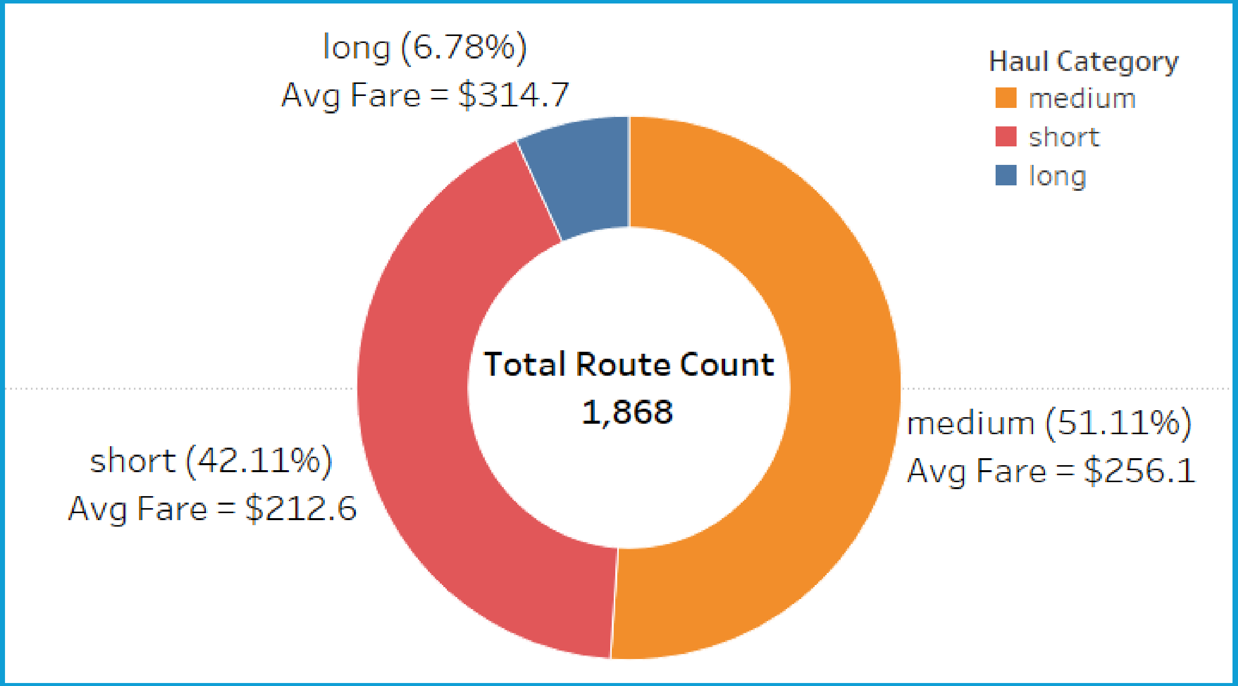
Lagged 4 pctchange

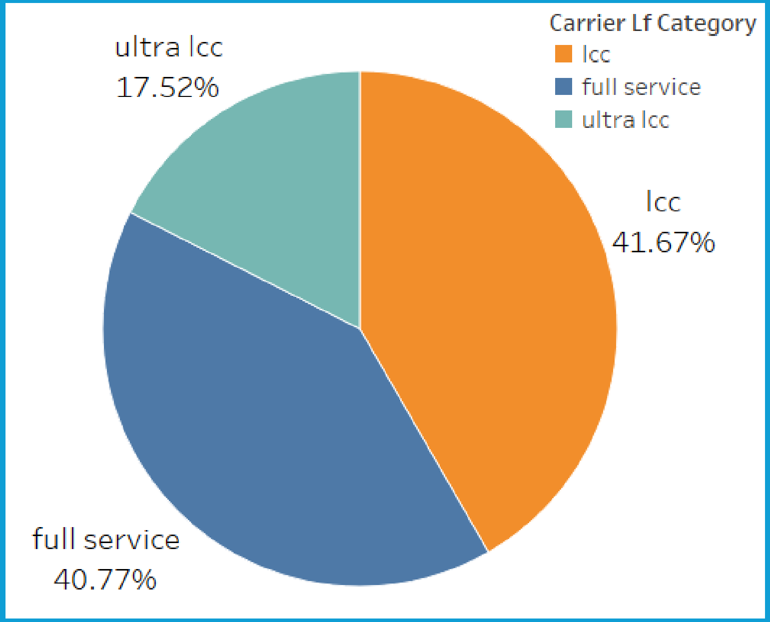
fare, passengers

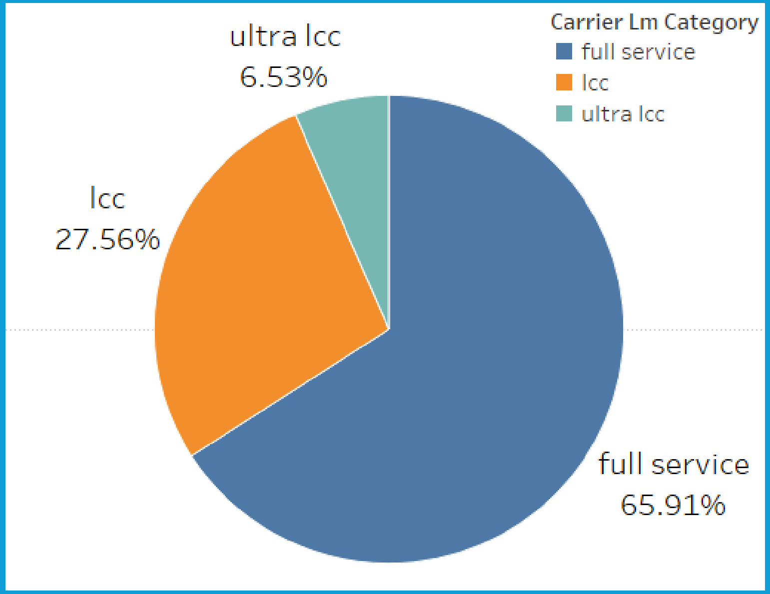
Exploratory Results



Exploratory Results





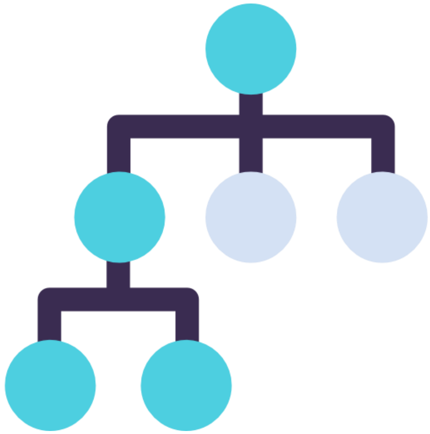


MODEL SELECTION

Model Options

Random Forest

SARIMA

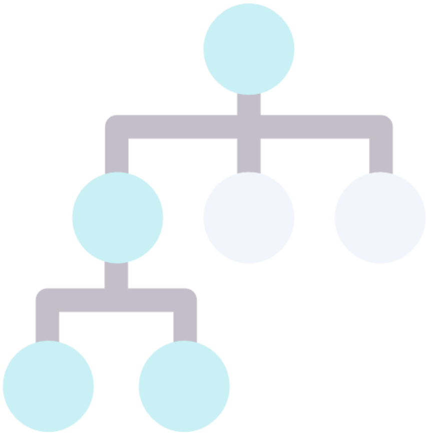


XGBoost

Model Options

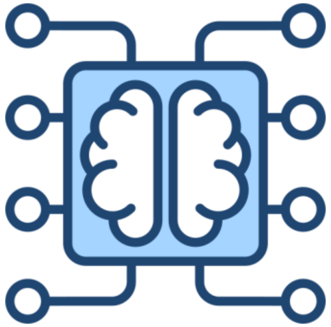
Random Forest

SARIMA

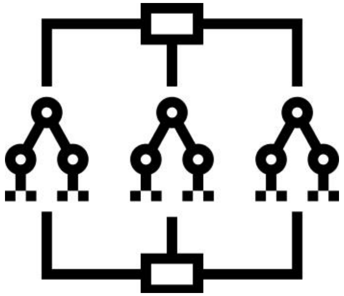


XGBoost

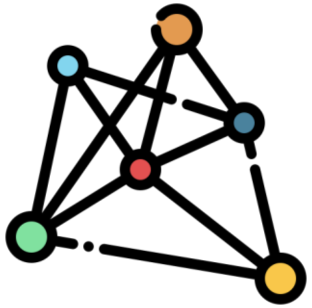
Random Forest – Characteristics



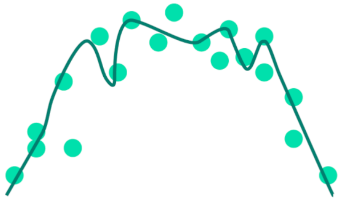
Ensemble model: Combines multiple trees for robust predictions



Randomness with Bagging technique



Capture non-linear relationship well



Less prone to overfitting thanks to averaging across multiple trees

Random Forest – Rationale

- Good at handling large datasets with high dimensions (~190 variables)
- Good at handling heterogeneous data (numerical and categorical)
- Capture non-linear relationships well
- Less sensitive to outliers
- More time efficient: require less parameter tuning (XGBoosts) and computational power(ANN)

Random Forest – Limitations

LIMITATIONS	MITIGATIONS
Computational Costs: increase number of n_estimators (number of trees) can be computationally expensive	Identify diminishing returns for additional trees and fine tune n_estimators as performance starts to plateau.
Time Series Trends: Random Forest does not capture sequential trends	Incorporate lagged variables to capture time dependencies. Use TimeSeriesSplit for CV evaluation to make sure that CV follows temporal order.

Random Forest – Results

Best Parameters				
max_depth: 30	max_features: None	min_samples_leaf: 1	min_samples_split: 2	n_estimators: 500

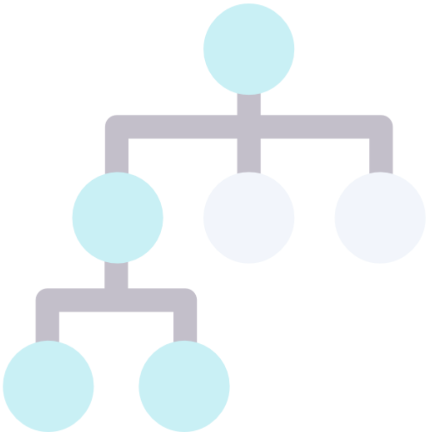
Results and Interpretations

MSE = 1761.03	R2 = 0.55	MAPE = 0.14	MAE = 31.71
• Average of squared differences between predicted and actual values • Used to compare performance between different models	• 55% of variability in flight prices (fare) can be explained by the model	• On average, the model predictions deviate from the actual values by approximately 14%	• On average, the model predictions deviate from the actual fare by approximately $31.71

Model Options

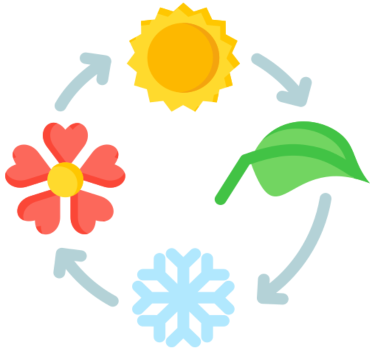
Random Forest

SARIMA



XGBoost

SARIMA – Characteristics



Seasonality (S): identifies and models repeated patterns over time

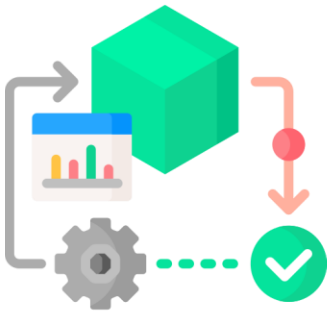
Integrated (I): transforms non-stationary data into stationary by differencing

Exogenous (X): incorporates external predictors not intrinsic to historical trends/ patterns

Autoregressive (AR): captures relationship between current and past data

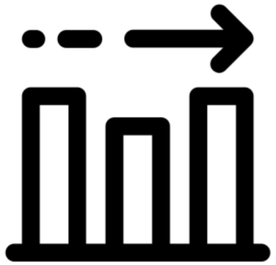
Moving average (MA): models dependency between current value and prediction errors

SARIMA – Limitations



Univariate model

Assumes linearity



Assumes stationarity

Computationally intensive

SARIMA – Results

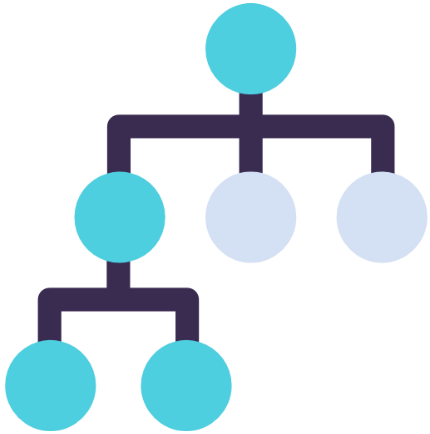
Best Parameters and Interpretations			
AR order (p): 0	differencing (d): 0	MA order (q): 1	seasonality (s): 4
No autoregressive component created through the model because we've already included lagged features	Data is assumed to be stationary because the features are in lagged state.	The model includes one lagged error term (a moving average component of order 1).	s=4 when data is recorded quarterly, capturing the seasonality that occurs every year.

Results
MSE = 1640.53

Model Options

Random Forest

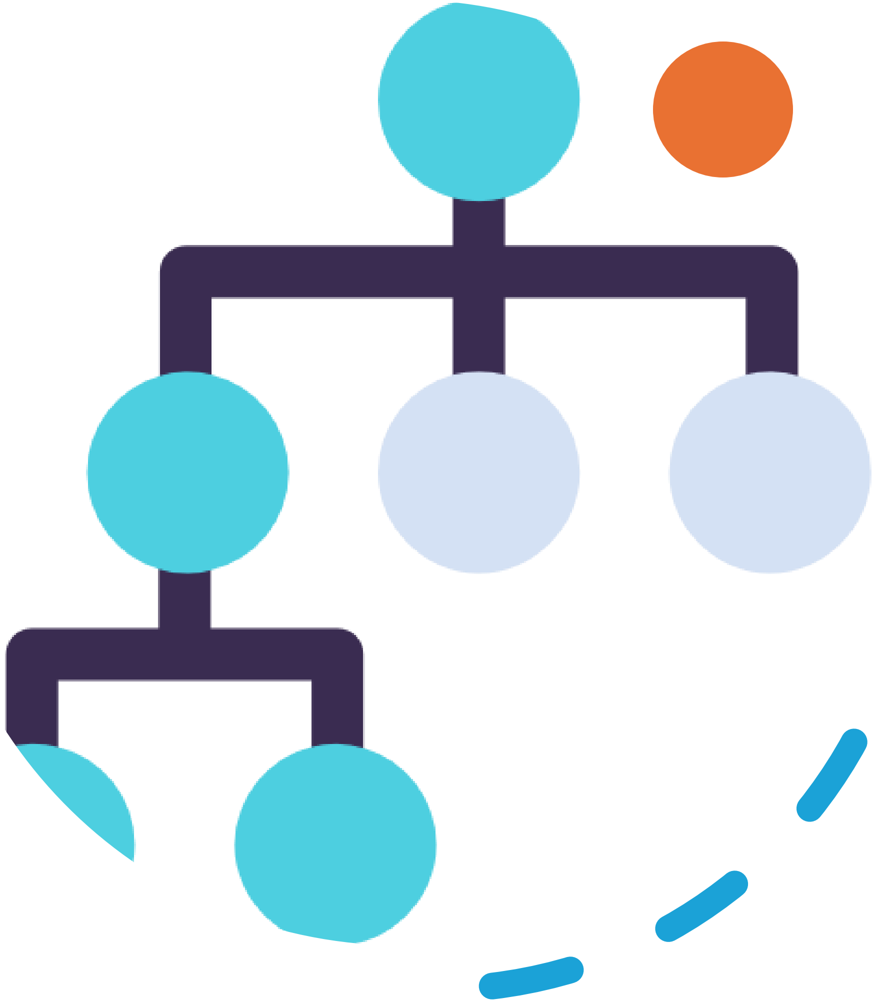
SARIMA



XGBoost

XGBoost – Characteristics

- Type of gradient boosted tree algorithm
- Combines predictions from multiple decision trees to build a strong, robust model
- Optimises the model iteratively by minimizing the difference between predicted and actual values
- Builds decision trees sequentially, with each one prioritising the errors from previous trees
- Special features
  - In-built regularisation to prevent overfitting
  - Parallelisation to build trees faster



XGBoost – Rationale

High predictive accuracy

Straightforward & interpretable

Ability to incorporate temporal dynamics and early stopping

Scalable & more computationally efficient

XGBoost – Parameters

Tree specific parameters: define how trees are constructed	max_depth, min_child_weight, colsample_bytree, subsample
Boosting parameters: control how boosting is performed	learning_rate, n_estimators
Regularisation parameters	lambda, alpha, gamma

XGBoost – Parameters

Tree specific parameters: define how trees are constructed	max_depth, min_child_weight, colsample_bytree, subsample
Boosting parameters: control how boosting is performed	learning_rate, n_estimators
Regularisation parameters	lambda, alpha, gamma

XGBoost – Parameters

Hyperparameter testing using GridSearchCV

n_estimators, learning_rate, max_depth, subsample, min_child_weight, colsample_bytree

Cross-validation approach: TimeSeriesSplit

Cross-validation strategy for time series data

Avoids data leakage

Retains seasonality and time-based features

Performance evaluation

MSE, MAPE, MAE, R squared

XGBoost – Results

Best Parameters

n_estimators	max_ depth	learning_rate	Subsample	colsample_bytree	min_child_weight
2000	8	0.04	0.7	0.8	5

Results and Interpretations

MSE = 932.22

- Lowest MSE across 3 different models, suggesting XGBoost model provides more accurate and consistent forecasts

R2 = 0.67

- 32.6% of the price variability is unexplained by the model
- Note: High variability in flight prices makes achieving a perfect R^2 difficult

MAPE = 12.59

- On average, the predictions are 12.59% off from the actual prices relative to their magnitude

MAE = 28.30

- On average, predictions deviate from the actual prices by $28.30

Top 10 features

Feature	Importance
haul_category_short	0.057432
city_Aspen, CO	0.055768
city_Atlantic City, NJ	0.054937
lagged1_ms_lf	0.051260
city_Tampa, FL (Metropolitan Area)	0.030351
city_Huntsville, AL	0.027704
city_Fayetteville, AR	0.027576
city_Las Vegas, NV	0.026694
city_Orlando, FL	0.023969
lagged4_ms_lm	0.017254

RESULTS & IMPLICATION

Interpretation of results

Observation 1:

Short-haul category is the most significant factor in determining flight prices

- Implement dynamic pricing strategies to maximize revenue and provide differentiated offerings to attract price-sensitive travelers
- Optimize fleet utilization by assigning the right aircraft (e.g., smaller, more fuel-efficient planes) to minimize operational costs

Interpretation of results

Observation 2:

Flight prices from the previous quarter for the airline with the lowest fare emerged as the one of the most influential factors

- Monitor the price adjustments in lowest-fare airline to set competitive prices
- Airlines can design campaigns to counteract the influence of the lowest-fare carrier
- Partner with the lowest-fare airline on specific routes or code-sharing agreements by pooling resources and reducing operational costs

Interpretation of results

Observation 3:

Flight price of the carrier with largest market share from the previous year showed a relatively significant influence on prediction

- Develop dynamic pricing models that respond quickly to changes in the market leader’s pricing
- Refine revenue management practices by tracking the pricing history of the market leader and aligning promotions or fare structures to maintain competitiveness in key markets
- Use the market leader's historical pricing to competitively price new routes while ensuring profitability

Interpretation of results

Observation 4:

Specific cities are influential in predicting flight prices

2 main categories:

- Emerging/niche markets
    - Huntsville, Fayetteville
- Major markets/ travel hubs
    - Aspen, Las Vegas, Tampa, Orlando, Atlantic City

Observation 4

Emerging/ Niche Markets

Represent untapped markets where lower prices might stimulate greater demand

- Could position them as an affordable entry point to expand market share, offering competitive fares to attract customers
- Tailor offerings to cater to the specific needs of businesses or institutions like NASA (Huntsville) or University of Arkansas (Fayetteville)
- Can serve as testing grounds for innovative pricing strategies, loyalty programs, or route expansions
- Ensure connectivity with major hubs (Dallas, Atlanta, Chicago, etc.) to meet unmet demand

Observation 4

Major Markets/ Travel hubs

- High-volume, high-demand destinations with robust competition between carriers
- Flight pricing is closely tied to external factors like seasonality, local events, or economic activity
    - ➢ Airlines could capitalize on premium pricing during peak seasons while ensuring adequate capacity to meet demand
    - ➢ Partnerships with local attractions can be established to offer bundles to stimulate demand
    - ➢ Tracking local events to dynamically adjust prices

Implication of Results in Business

Travel Platforms (Expedia, Kayak, Booking)	• Offer personalized price alerts and predictive tools to improve user engagement and booking decisions. • Leverage seasonal and regional trends to optimize marketing strategies and promotions.
Airlines (Delta, American Airlines, Southwest, United Airlines, JetBlue, etc.)	• Use forecasts to adjust fares dynamically and optimize route planning and demand management • Build customer trust by offering consistent, transparent, and competitive pricing.
Travel Agencies (AAA, Flight Centre, Travel Leaders, Carlson Wagonlit, TUI Group)	• Align travel package pricing with seasonal trends to maximize value for customers. • Prepare for peak travel periods by optimizing resources and tailoring deals effectively.
Consumers	• Access cost-saving tools and make informed travel decisions based on reliable price predictions. • Build loyalty to platforms offering transparency and clear pricing insights.

Feasibility and Practicality of Implications

Travel Platforms

- Implementation: Integrate predictive models via APIs or plugins into existing platforms.
- Feasibility: Low cost, high user engagement, and adaptable for various destinations.

Airlines

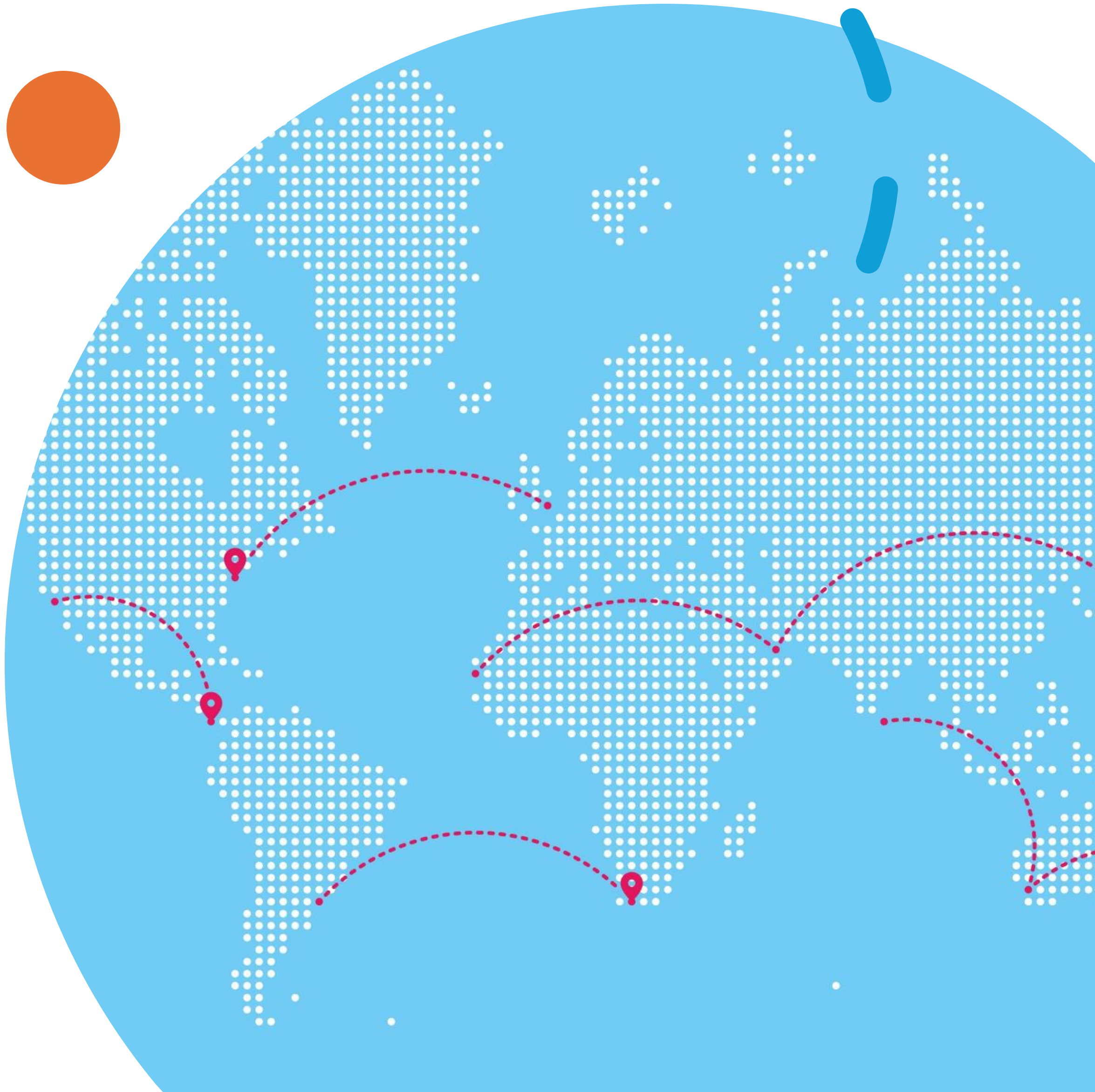
- Implementation: Leverage existing data systems for deploying forecasting tools.
- Feasibility: High ROI with improved pricing strategies and efficient demand management.

Travel Agencies

- Implementation: Use predictive insights for creating dynamic travel packages.
- Feasibility: Cost-effective adoption with tailored offerings to maximize revenue.

Consumers

- Implementation: Deliver insights via user-friendly apps and platforms.
- Feasibility: Easy adoption requiring no technical expertise, fostering trust and transparency.



References

Resources:

- https://www.datacamp.com/tutorial/tutorial-time-series-forecasting
- https://www.geeksforgeeks.org/time-series-analysis-and-forecasting/
- https://www.statista.com/markets/424/topic/488/airlines/
- https://www.geeksforgeeks.org/sarima-seasonal-autoregressive-integrated-moving-average/#:~:text=SARIMA%2C%20which%20stands%20for%20Seasonal,handle%20data%20with%20seasonal%20patterns.

Dataset:

- https://www.transportation.gov/policy/aviation-policy/domestic-airline-consumer-airfare-report
- https://fred.stlouisfed.org/series/POILBREUSDQ

THANK YOU