Week 9 Milestone: Final Presentation Outline

Flight Price Predictor

Business case and problem statement:

Travelers often struggle with knowing when to book flights to secure the lowest prices. Airline pricing algorithms are complex, and last-minute purchases often result in higher costs. By using time series forecasting techniques and machine learning models, we aim to:

- Identify price trends over time
- Predict future ticket prices based on historical data
- Provide insights for consumers to make cost-effective purchasing decisions

Modeling hypothesis and assumptions:

- Model Hypotheses:
  - Flight prices generally increase as departure date approaches
  - Price fluctuations follow discernible patterns based on time to departure
  - Historical price data can predict future price trends
- Key Assumptions:
  - **Non-stationary data:** Means and variance change over time.
  - **Seasonal trends:** Holidays, weekends, and major events affect prices.
  - **Consistent relationship:** Time-to-departure influences price predictability.
  - **External factors:** Events and demand fluctuations impact pricing but can be modeled.
  - **Data availability:** Sufficient historical data exists for reliable predictions.

Data description and properties:

Dataset: 82 million flight itineraries with 46 features related to the flight itineraries.
Key Variables:

Variables	Description
totalFare (Target)	The final ticket price for a given flight departure.
search_date	The date the price is being checked.
departure_date	Date and time of departure.
seatsRemaining	Number of available seats at time of booking.
isRefundable	Indicates whether the ticket is refundable.
totalTravelDistance	Distance traveled (some missing values).

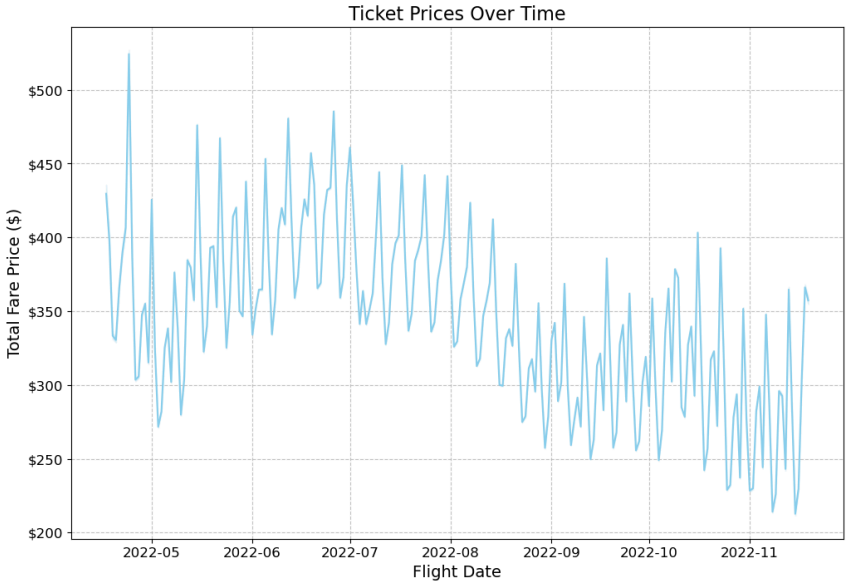
Data Challenges & Processing:

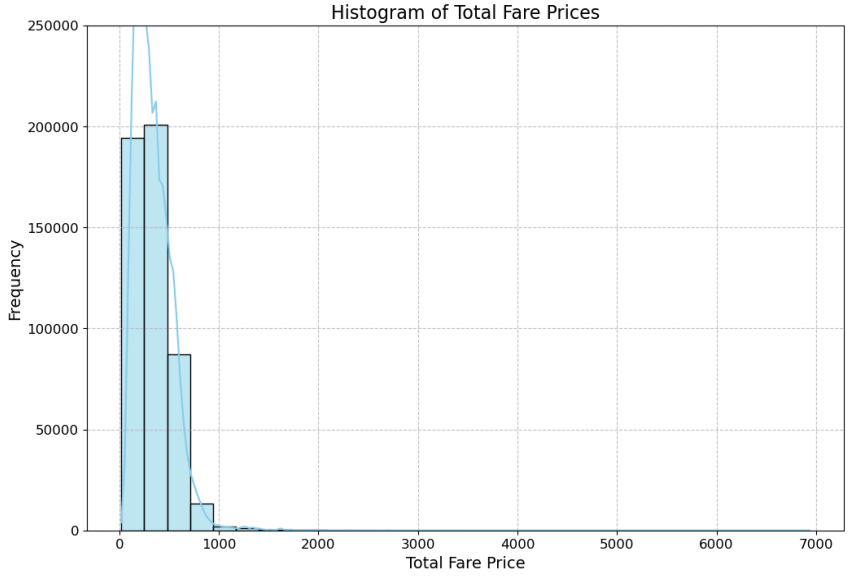
- Missing values in totalTravelDistance → Imputed using median-based techniques.
- Categorical variables (e.g., airport codes, airlines) → Encoded using one-hot encoding.
- Non-stationary trends in ticket prices → Feature transformations were applied.

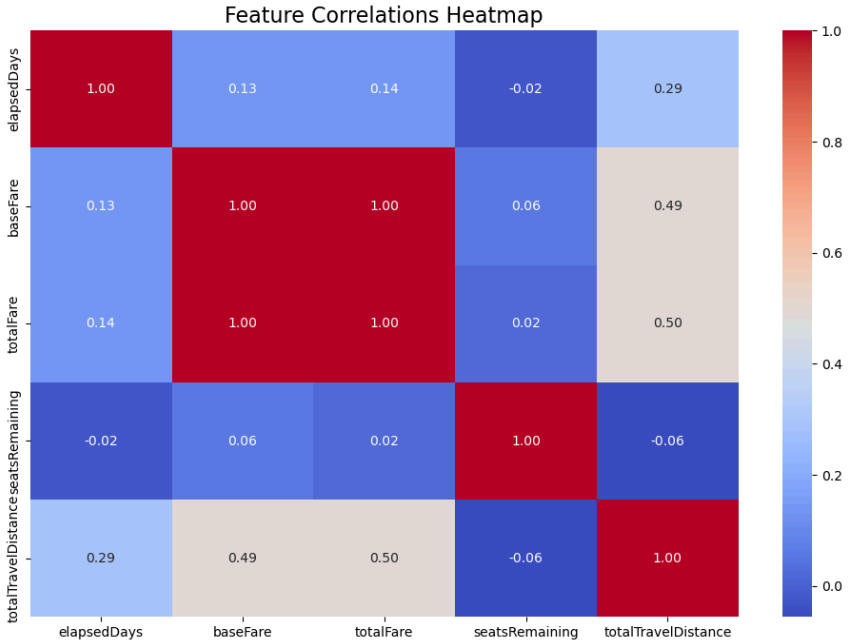
Key Data Insights

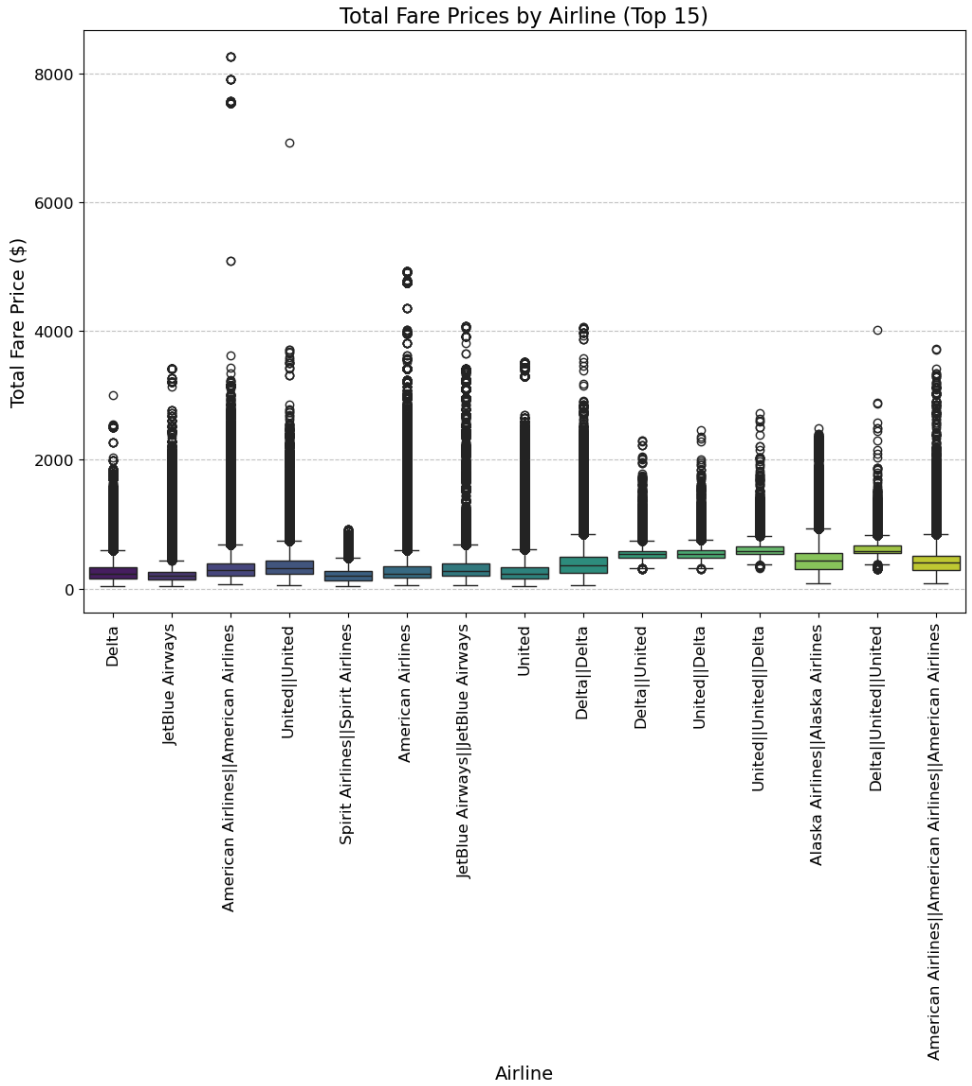
- Flight prices increase closer to departure, but the rate varies based on seasonality.
- Some airlines exhibit more volatile pricing trends than others.
- The number of available seats influences price, lower availability leads to higher fares.

Data processing and Exploratory Data Analysis:






Feature Correlations Heatmap
elapsedDays
baseFare
totalFare
seatsRemaining
totalTravelDistance
1.00
0.13
0.14
-0.02
0.29
0.13
1.00
1.00
0.06
0.49
0.14
1.00
1.00
0.02
0.50
-0.02
0.06
0.02
1.00
-0.06
0.29
0.49
0.50
-0.06
1.00
1.0
0.8
0.6
0.4
0.2
0.0

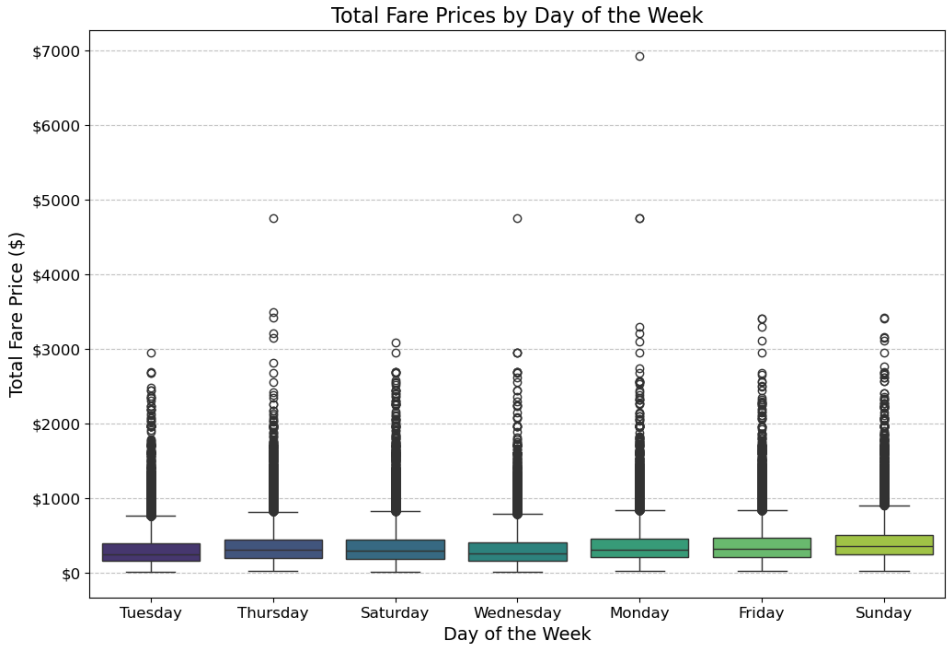

Total Fare Prices by Airline (Top 15)
Total Fare Price ($)
8000
6000
4000
2000
0
Delta
JetBlue Airways
American Airlines||American Airlines
United||United
Spirit Airlines||Spirit Airlines
American Airlines
JetBlue Airways||JetBlue Airways
United
Delta||Delta
Delta||United
United||Delta
United||United||Delta
Alaska Airlines||Alaska Airlines
Delta||United||United
American Airlines||American Airlines||American Airlines
Airline

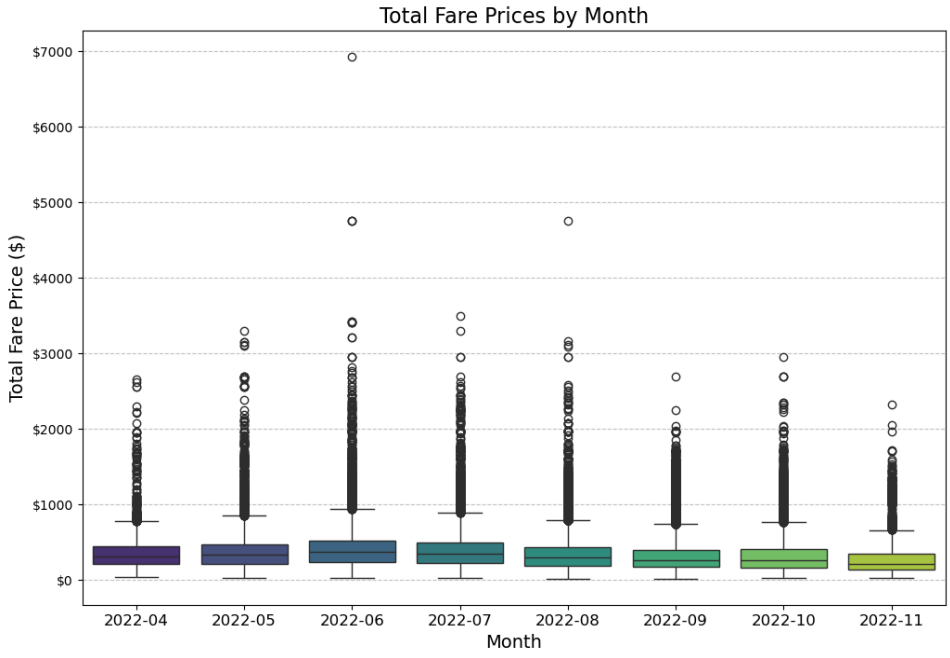
Feature Engineering:

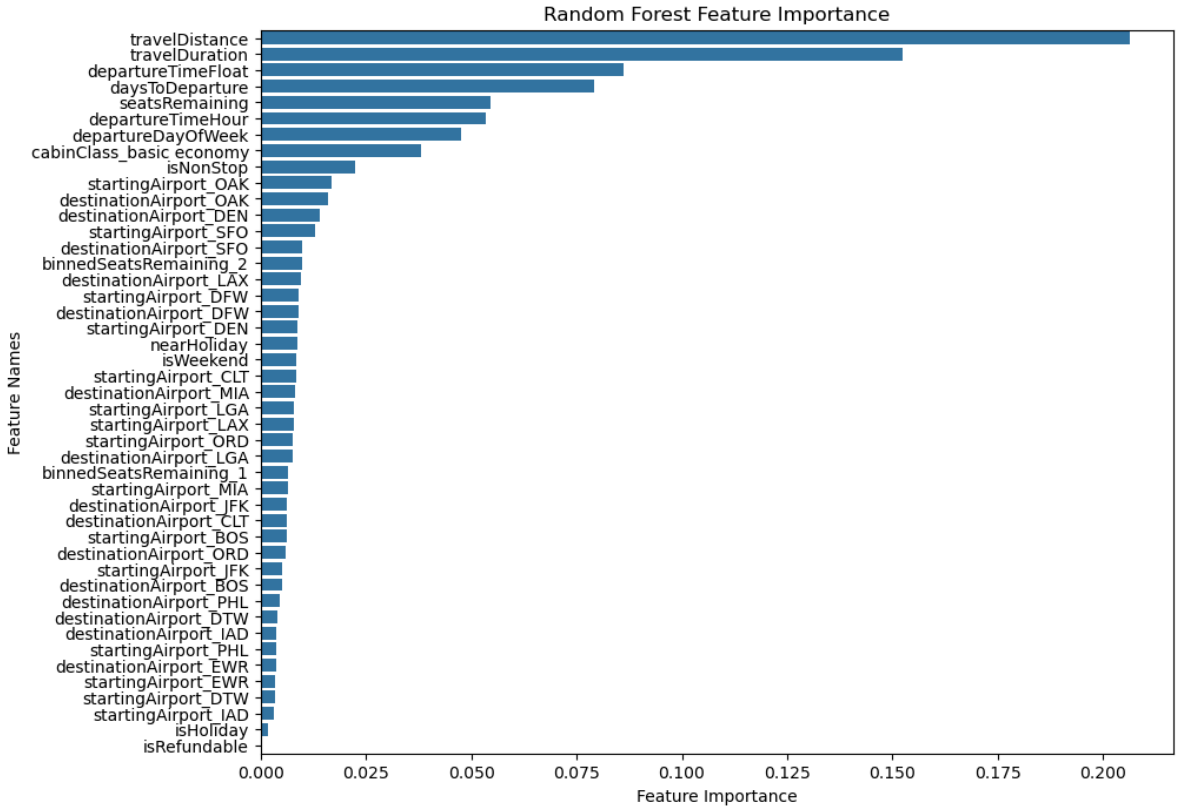
1. *Derived Features:*
   - Binned seatsRemaining into categories (low, medium, high availability).
   - Extracted temporal features (e.g., month, weekday, hour of departure).
   - Created a holiday indicator → Flights near holidays flagged for price impact.

2. *Encoding Categorical Variables:*
   - One-hot encoding for airports and airlines.

3. *Missing Data Handling:*
   - Imputation of missing totalTravelDistance values.







Proposed modeling approaches:

Models Tested:

Traditional Time Series Models

- ARIMA
- SARIMA
- Prophet (did not generalize well)

Machine Learning Models

- Decision Tree (Baseline)
- Random Forest (Best-performing model)
- XGBoost (Close second in performance)

Deep Learning Model

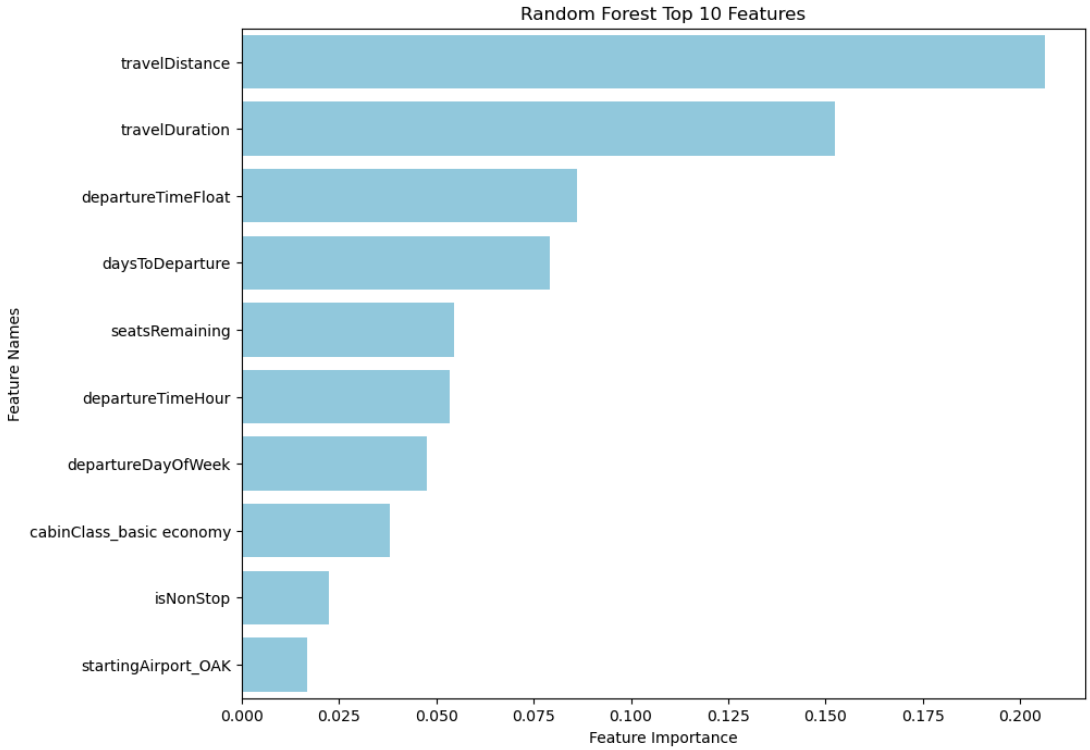
- RNN LSTM (Struggled due to price variance)

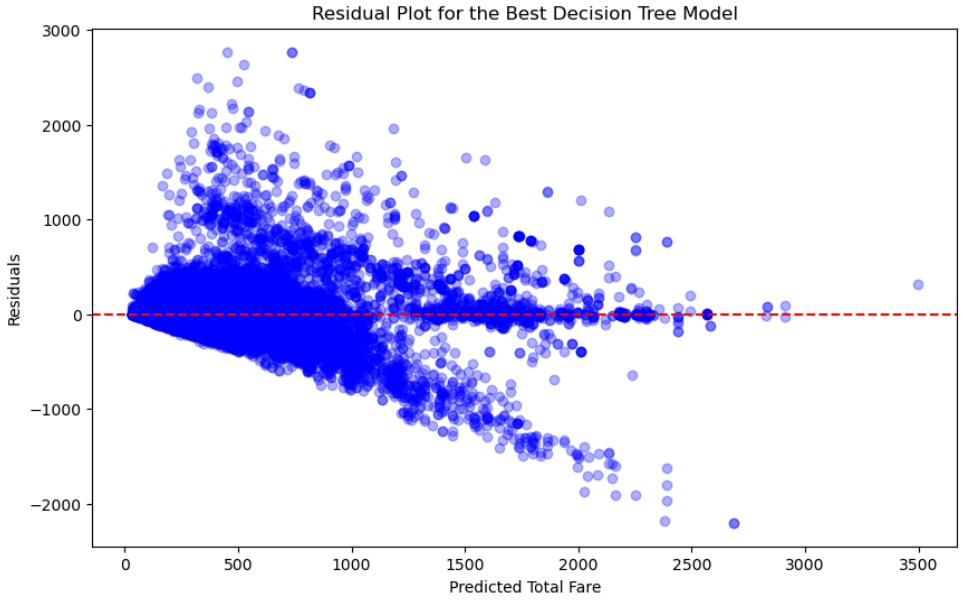
Selected model results with justifications and tradeoffs:

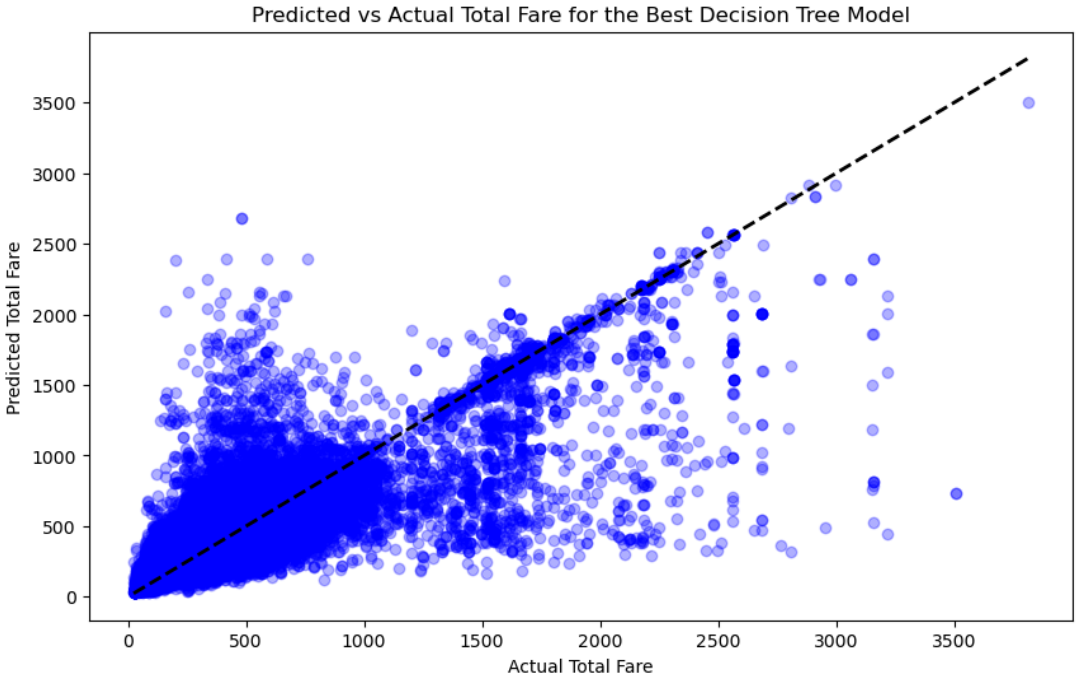
Model	BIC
ARIMA	55,701.33
SARIMA	52,994.49

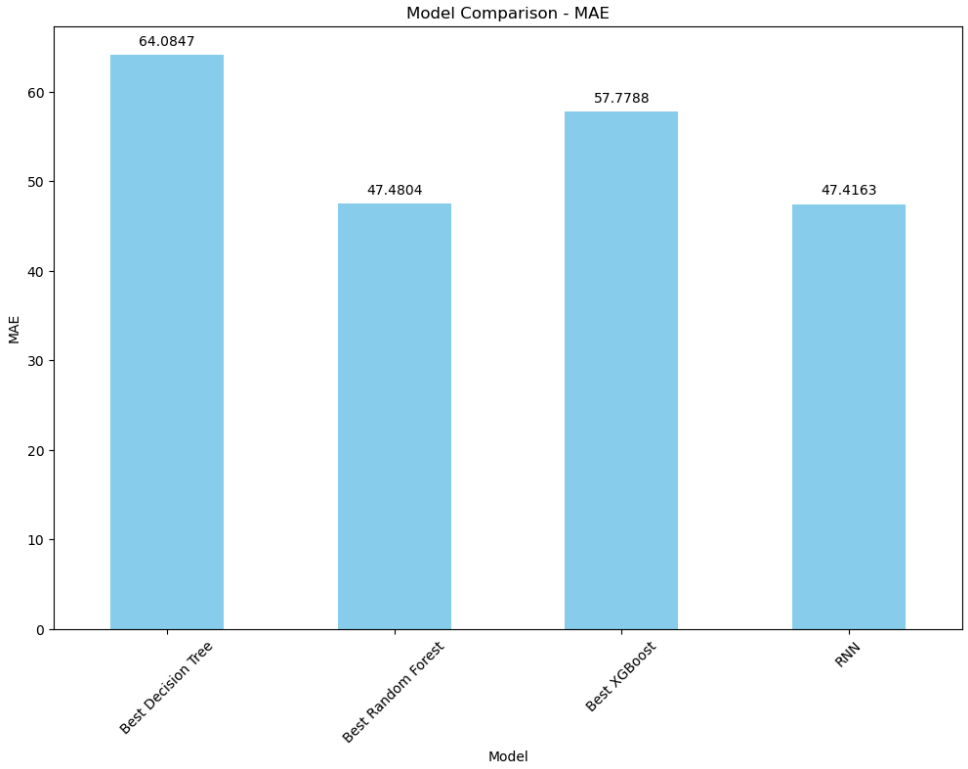
Model	MAE	RMSE	R^2 Score
Prophet	23,447.19	149.89	
XG Boost	56.60	99.70	0.82
Decision Tree	63.81	124.56	0.72
Random Forest	44.87	93.76	0.84
RNN LSTM	85.46	166.38	0.42

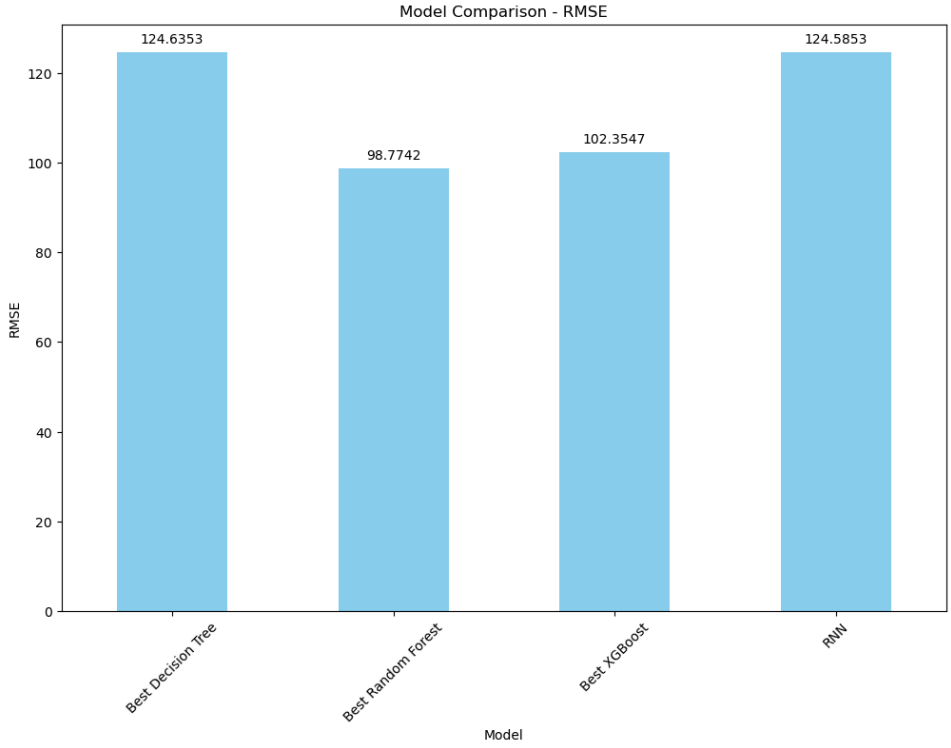
- Random Forest was the best-performing model, with an R^2 score of 0.84 and the lowest RMSE.
- XGBoost was a strong contender, showing slightly lower accuracy than Random Forest.
- Decision Trees had higher error rates due to overfitting on training data.
- RNN LSTM struggled with high variance, likely due to limited sequence dependencies in flight pricing data.

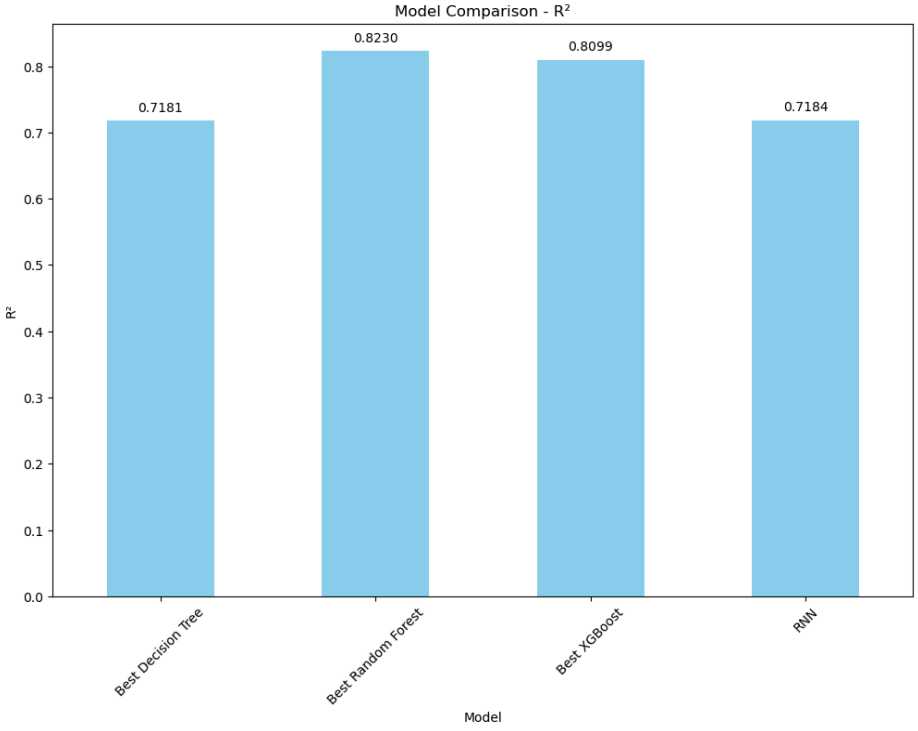

Random Forest Top 10 Features
travelDistance
travelDuration
departureTimeFloat
daysToDeparture
seatsRemaining
departureTimeHour
departureDayOfWeek
cabinClass_basic economy
isNonStop
startingAirport_OAK
Feature Names
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
Feature Importance


Residual Plot for the Best Decision Tree Model
3000
2000
1000
0
−1000
−2000
Residuals
0
500
1000
1500
2000
2500
3000
3500
Predicted Total Fare


Predicted vs Actual Total Fare for the Best Decision Tree Model
Predicted Total Fare
Actual Total Fare
0
500
1000
1500
2000
2500
3000
3500


Model Comparison - MAE
MAE
Model
64.0847
47.4804
57.7788
47.4163
Best Decision Tree
Best Random Forest
Best XGBoost
RNN
0
10
20
30
40
50
60


Model Comparison - RMSE
124.6353
98.7742
102.3547
124.5853
RMSE
0
20
40
60
80
100
120
Best Decision Tree
Best Random Forest
Best XGBoost
RNN
Model


Model Comparison - R²
0.7181
0.8230
0.8099
0.7184
R²
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Best Decision Tree
Best Random Forest
Best XGBoost
RNN
Model

Insights/Recommendations & Future work

Insights from the Analysis:

- Tree-based models (Random Forest, XGBoost) outperform deep learning models for flight price prediction.
- Feature engineering significantly improved model accuracy, particularly holiday effects and seat availability binning.
- Flight prices are highly non-stationary, requiring advanced techniques for accurate forecasting.

Recommendations:

- Real-time prediction system: Deploy a model API that continuously updates with new flight data.
- External data sources: Incorporate weather conditions, real-time seat availability, and airline promotions to improve accuracy.
- Hyperparameter tuning: Further refine model parameters for even lower error rates.
- Transformer-based models: Test newer deep learning approaches (e.g., Time Series Transformers) to improve sequence-based learning.

Future Research Directions:

- Expanding the dataset to include international flights and budget airline fares.
- Exploring ensemble methods that combine Random Forest and XGBoost for improved predictions.
- Investigating reinforcement learning for dynamic pricing recommendations.