Flight Price Predictor

Group 8
Sam Ding, Lulu Wang, Neal Xu, Richard Yang

Team Members

Sam Ding

Lulu Wang

Neal Xu

Richard Yang

01
Problem Statement

02
More about Data

03
Methodology

04
Summary

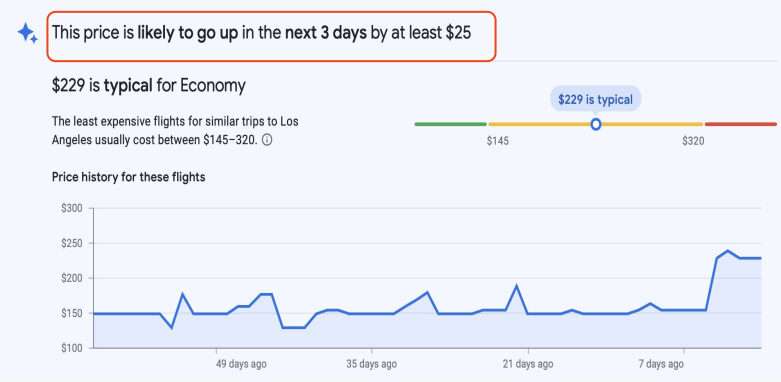
Problem Statement

- Predict prices interval of airline tickets
- Traditional ML and Regression methods aim to predict the mean of a dependent variable
- Predict multiple quantiles of the ticket price distribution
- Find the optimal interval for client

Business Value

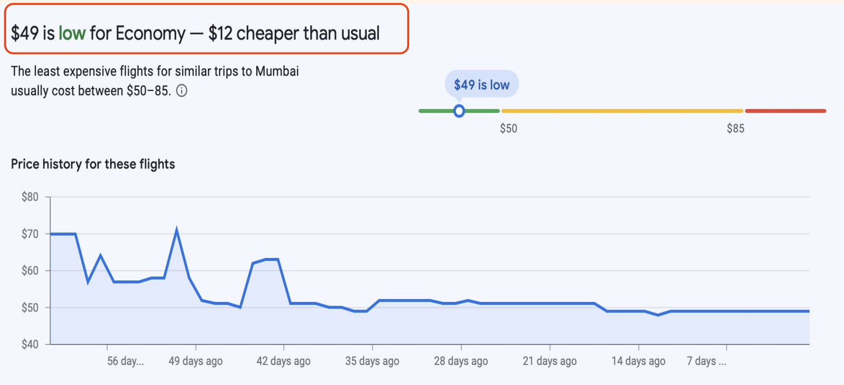
On a given day...

Google Flights UI for
New York to Los Angeles



How consumer will benefit?

Google Flights UI for
Delhi to Mumbai



Data Description

Source Data

Data was distinct flight booking options for travel between India's top 6 metro, collected for 50 days in 2022.

Predictor

Flight information (airline, time of departure, origin & destination city, duration), days before departure

Target Variable

Flight price in Indian Rupee (₹)

Feature Engineering

City

Origin & destination city GDP, population

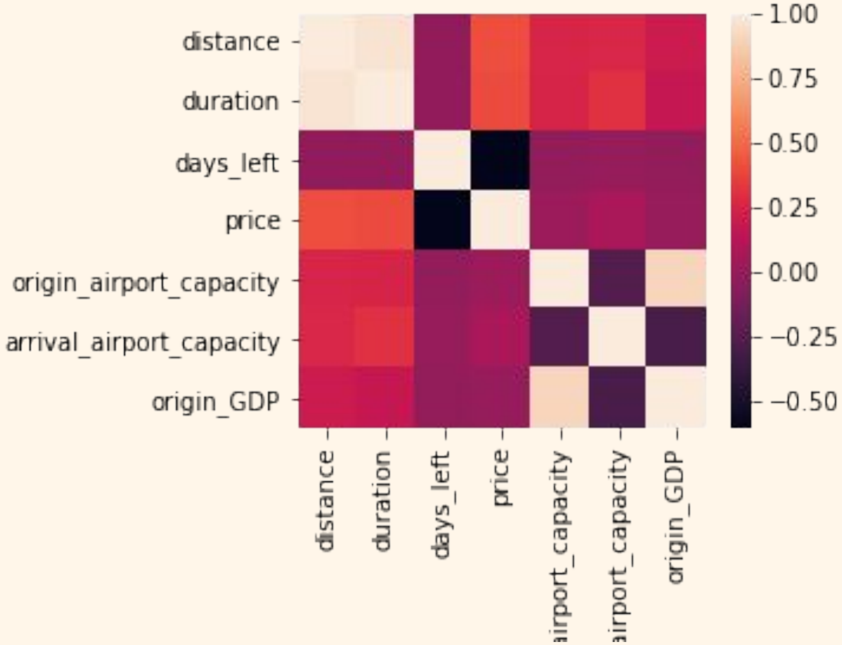
Airport

Origin & destination airport capacity

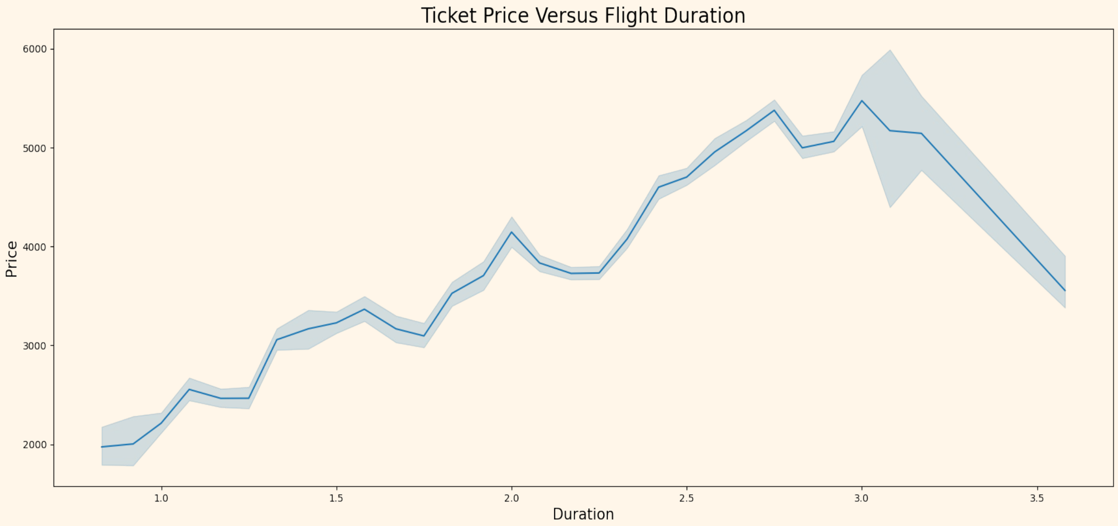
Flight

Carrier Category, Distance

Data Exploration - Correlation

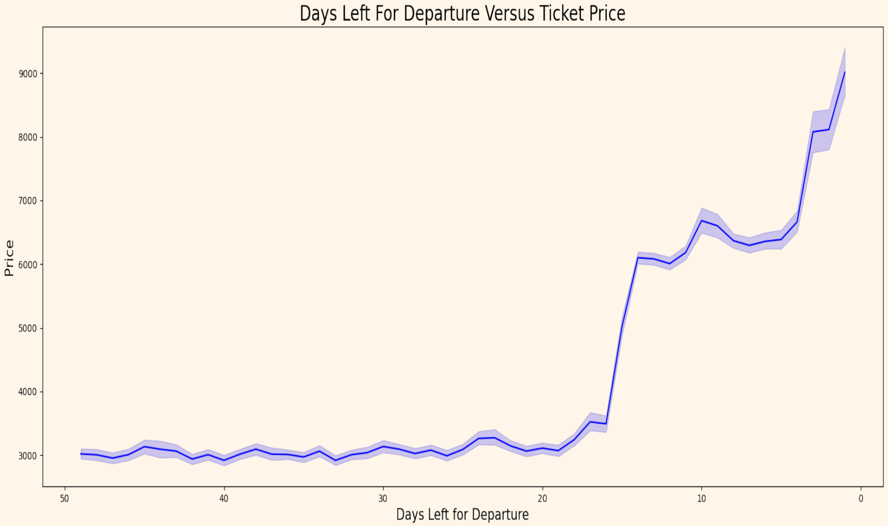


Data Exploration - Duration



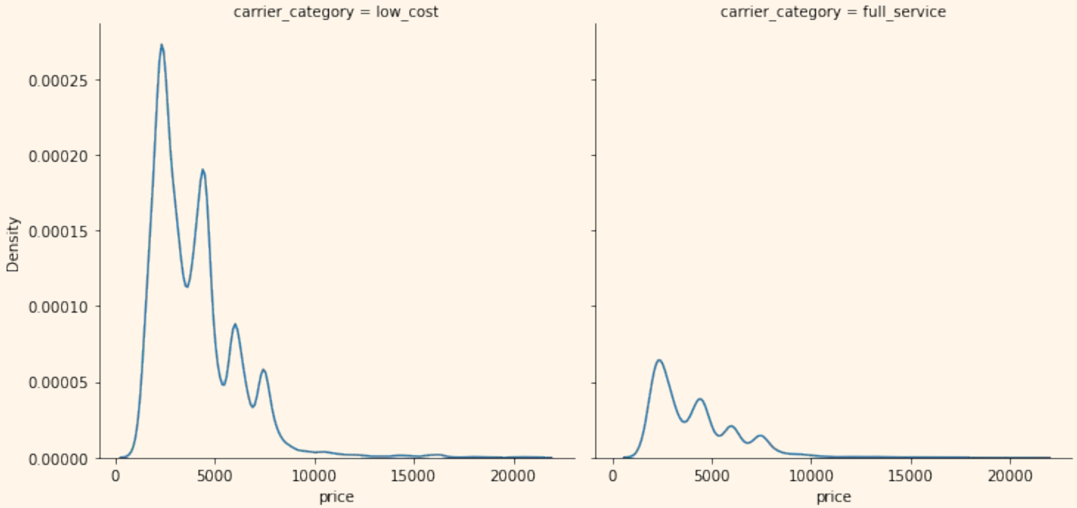
As flight duration increases, the price gradually increases.
But for duration larger than 3h, the price declines steeply.

Data Exploration - Days Left



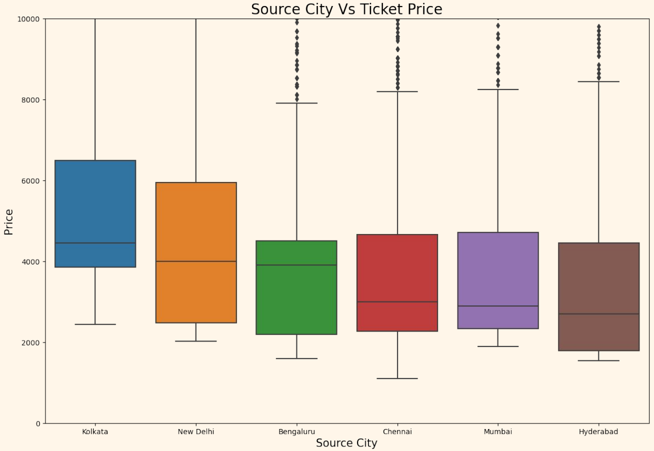
As days left for departure decreases, the price gradually increases

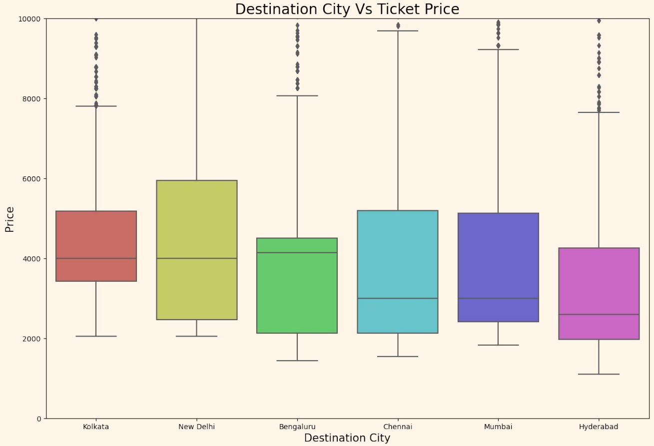
Data Exploration - Airline



Two airlines are full-service airlines, while four are low-cost carriers.

Data Exploration- City





Flight involving Kolkata and New Delhi have higher price, and flight involving Mumbai and Hyderabad have lower price

Feature Selection

We used **Recursive Feature Elimination (RFE)** method to fit a **Random Forest Regressor**, which generates the optimal number of features to select, as well as the features to select given that optimal number.

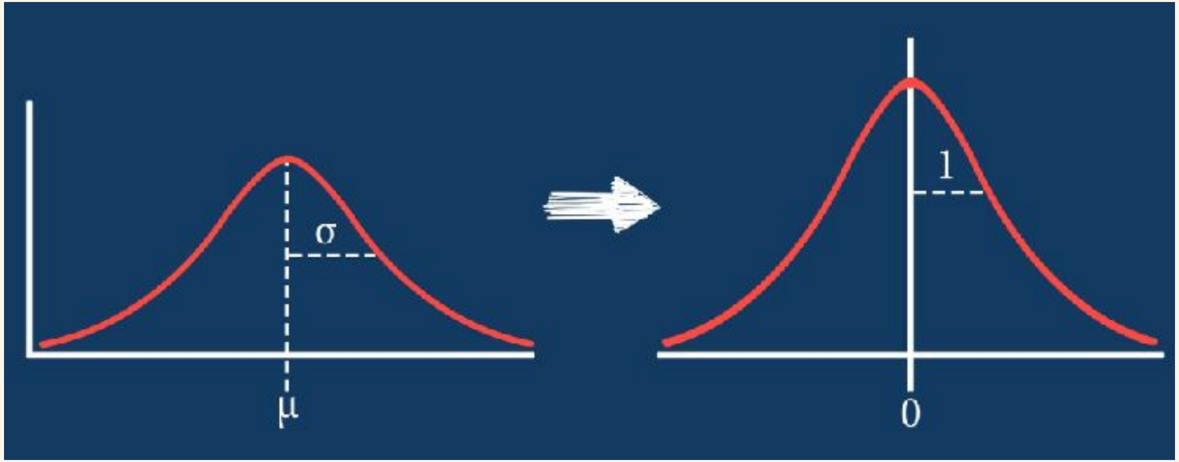
RFE calculates that **10 features** demonstrates the best metrics.

The 10 variables selected by RFE are:

```
Departure Time,
Arrival Time,
Distance,
Duration,
Days Left for Departure,
Origin Airport Capacity,
Destination Airport Capacity,
Origin City GDP,
Destination City GDP,
Carrier Category
```

Data Transformation

We used **Preprocessing** method from *Sklearn* Package to transform our dataset to a standard scale, which is mean=0, standard deviation=1



Model Selection

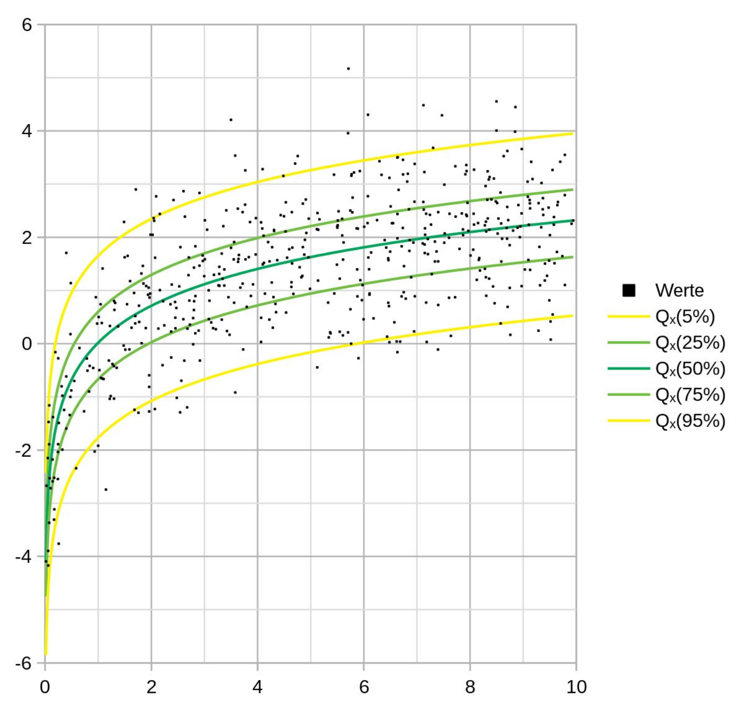
	Model_Name	R2_score	MAPE
0	RandomForestRegressor	0.879689	62.120630
1	XGBRegressor	0.877535	61.906102
2	BaggingRegressor	0.859311	62.252076
3	ExtraTreesRegressor	0.855853	62.399239
4	KNeighborsRegressor	0.827792	7.848034
5	GradientBoostingRegressor	0.815205	60.714184
6	DecisionTreeRegressor	0.791894	62.705725
7	Lasso Regression	0.528483	59.827220
8	Ridge Regression	0.528476	30.142381
9	LinearRegression	0.528475	30.142488

We first tried to predict the airline ticket prices by applying 10 Models including ML and Regression.

Among these models,
Random Forest Regressor has the best R2 score, and **KNN** has the best MAPE.

We can also see that there is a gap in R Squared between **Lasso Regression**, **Ridge Regression** and **Linear Regression**.

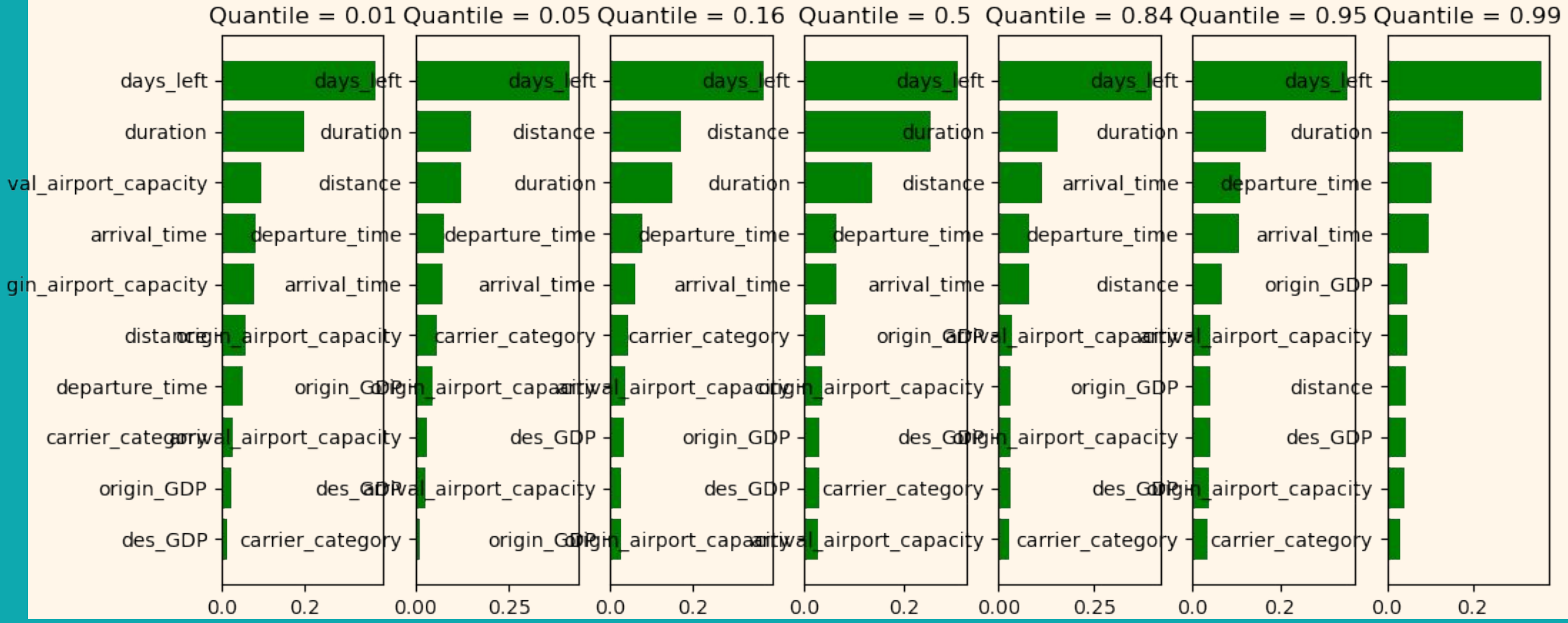
What is Quantile Regression?



	0.01	0.05	0.5	0.95	0.99	actual	interval
0	2409.87	2410.24	1813.06	2789.67	2825.24	2791	415.37
1	2409.47	2409.94	1709.26	2783.37	2888.82	2700	479.35
2	2409.92	2409.76	1983.08	2789.79	2946.36	2791	536.44
3	2409.58	2409.63	1876.19	2875.51	2956.26	2700	546.68
4	2409.68	2410.07	1808.96	2786.45	2981.99	2410	572.31
...	...	...	...	...	...	...	...
8366	2333.58	2754.39	8103.92	12417.39	18049.43	12392	15715.85
8367	2334.01	2755.71	8029.32	12515.85	18057.88	16383	15723.87
8368	2333.99	2754.84	8246.83	12514.81	18146.05	9243	15812.06
8369	2334.01	2755.55	8101.08	12515.56	18166.99	11027	15832.98
8370	2334.52	2756.77	8100.21	12615.14	18205.92	20268	15871.40

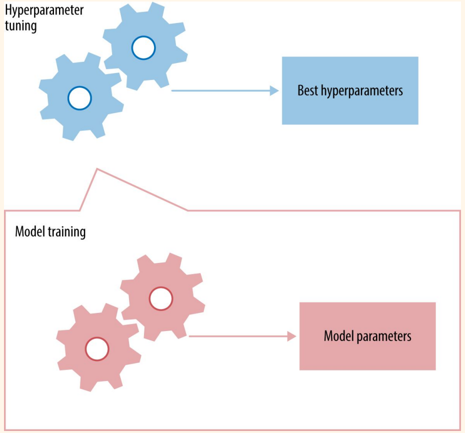
8371 rows × 7 columns

Feature Importance



Model Engineering
(Hyper-parameter Tuning)

```
grid['n_estimators'] = [100, 200, 300, 400, 500]
grid['max_depth'] = [2, 4, 6, 8, 10]
grid['learning_rate'] = [0.001, 0.01, 0.1, 0.2, 0.3]
grid['min_samples_leaf'] = [1, 2, 4, 6, 8]
grid['min_samples_split'] = [2, 4, 6, 8, 10]
```



Run time: ~10 hours

✓ 585m 38.8s

Tuning Result:

```
Best: -266.941791 using {'learning_rate': 0.1, 'max_depth': 10, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 500}
```

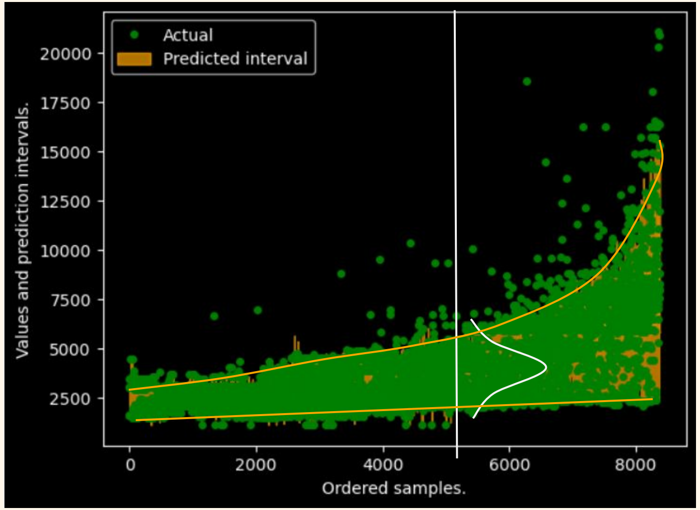
Quantile Regression

Quantile Regression (Linear)

R-Squared	MAPE
0.29	34.71%

Quantile Boosting Regression (Non-Linear)

R-Squared	MAPE
0.85 √	**5.34%** √



Interval: 0.16 to 0.84 Quantile (1σ)

Further Exploration–Days Left

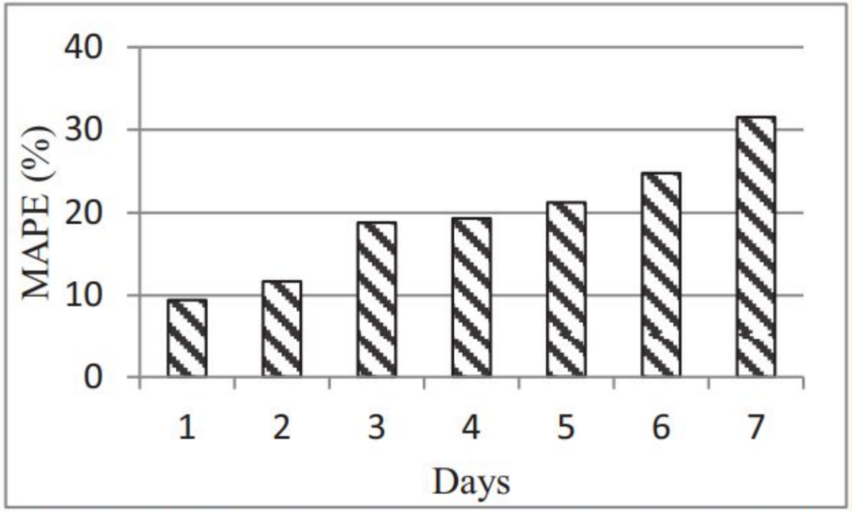


Figure 2: Spot price forecasting error for one week ahead

From the Amazon Spot Prices Prediction Literature Review (Murphy & Brockman, 2000), we can observe an obvious pattern that the **regression model tends to perform better when the days left is smaller.**

Results -Quantile Regression

Days Left	R - Squared	MAPE (Mean Absolute Percentage Error)
1-10	-0.12	18%
11-20	0.11	37%
21-30	0.44	19%
31-40	0.48	18%
41-49	0.36	19%

Quantile Boosting Regression

Days Left	R - Squared	MAPE (Mean Absolute Percentage Error)
1-10	0.52	8%
11-20	0.84	7%
21-30	0.76	6%
31-40	0.78	5%
41-49	0.61	6%

Conclusions

1. The **Quantile Boosting Regression** performs better based on R-Squared and MAPE score.

2. The Most suitable Use Case: The model has the best prediction accuracy when days left is between **11-20 days**.

3. Feature Importance: Days Left, Duration and Departure Time are the three most significant features in our model.

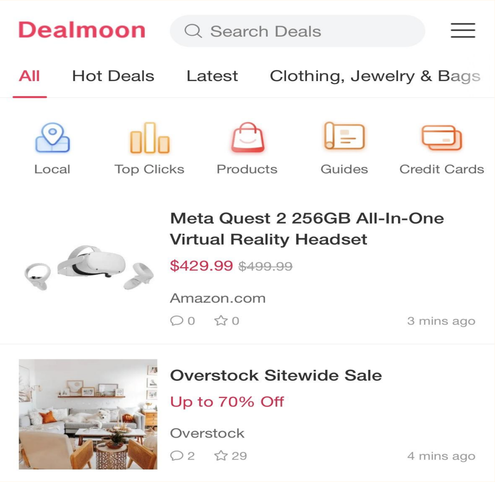
Additional Implementation

More Convinced Interval (For Consumer)

Interval prediction is better than point estimate. The point estimate may be overvalued or undervalued which may lead to higher risks of missing the best purchasing price

Partnership (For Airline)

Collaborate with third party platforms, airlines, and put the purchasing link directly in our app so that user can buy the ticket more conveniently with lower price



Q&A

Thanks!