Estimating flight prices from booking data

Group 7: Amy Merkelz, Jordan Stump, Laila Sultan, Nayoung Lim, Vani Kalra

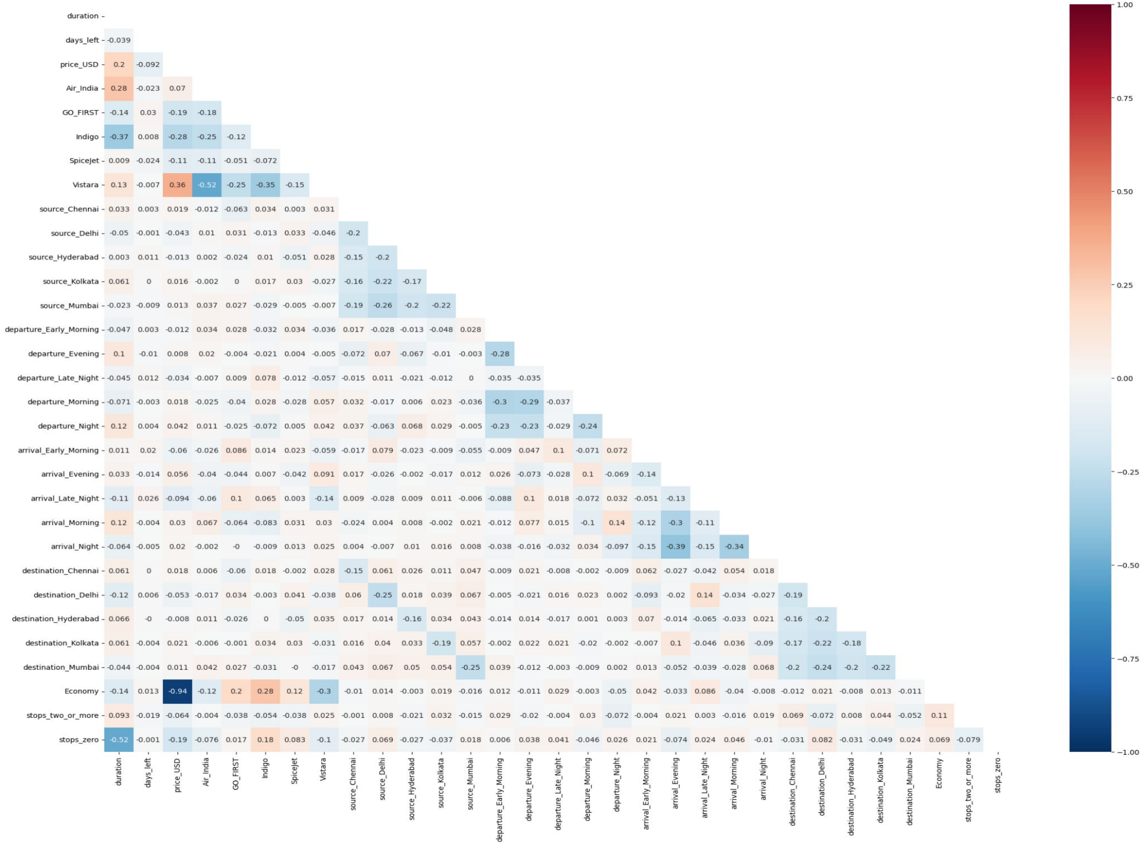
Background

- Dataset is booking data from EaseMyTrip for 6 major metropolitan areas of India
- Data was collected from February 11th, 2022 to March 31, 2022
- Metadata:
  - 11 features, including airline, departure/arrival times and cities, days before departure that the flight was booked
  - Many categorical features that were one-hot encoded
  - 300,261 rows
- Our goal was to build a regression model to predict flight prices

Correlation Between Price and Other Features

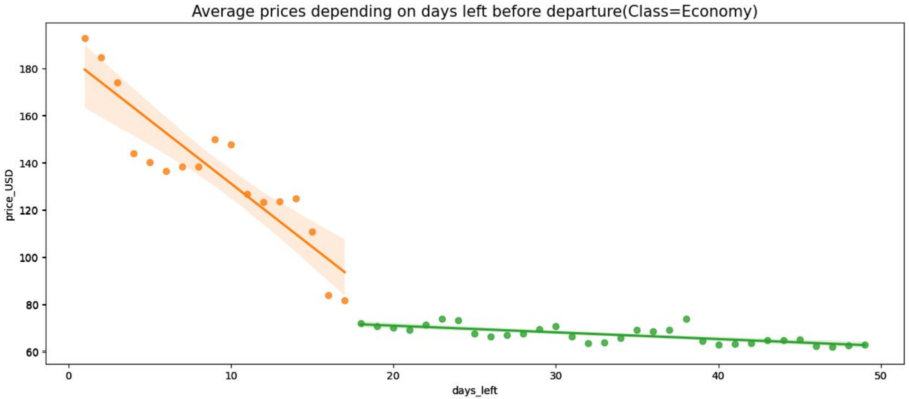
Features :

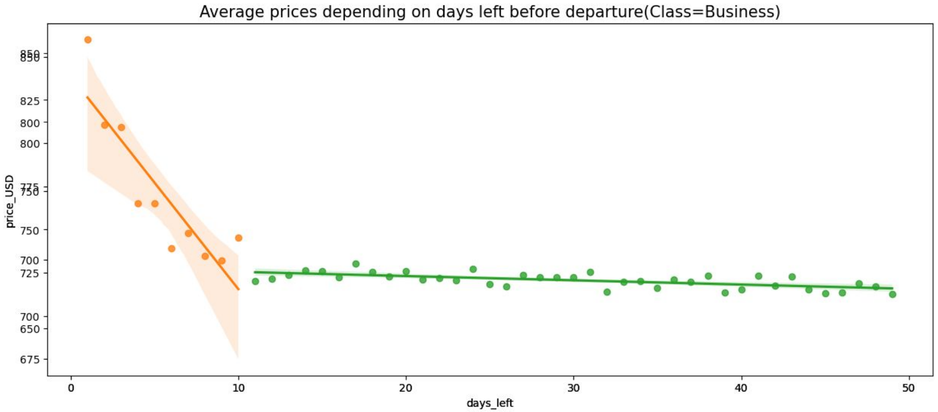
- Class(Economy, Business)
- Airline
- Number of stops between source city and destination city
- Days left before departure
- Flight duration
- Departure time & Arrival time
- Duration



	duration	days_left	price_USD	Air_India	GO_FIRST	Indigo	SpiceJet	Vistara	source_Chennai	source_Delhi	source_Hyderabad	source_Kolkata	source_Mumbai	departure_Early_Morning	departure_Evening	departure_Late_Night	departure_Morning	departure_Night	arrival_Early_Morning	arrival_Evening	arrival_Late_Night	arrival_Morning	arrival_Night	destination_Chennai	destination_Delhi	destination_Hyderabad	destination_Kolkata	destination_Mumbai	Economy	stops_two_or_more	stops_zero
duration																															
days_left	-0.039																														
price_USD	0.2	-0.092																													
Air_India	0.28	-0.023	0.07																												
GO_FIRST	-0.14	0.03	-0.19	-0.18																											
Indigo	-0.37	0.008	-0.28	-0.25	-0.12																										
SpiceJet	0.009	-0.024	-0.11	-0.11	-0.051	-0.072																									
Vistara	0.13	-0.007	0.36	-0.52	-0.25	-0.35	-0.15																								
source_Chennai	0.033	0.003	0.019	-0.012	-0.063	0.034	0.003	0.031																							
source_Delhi	-0.05	-0.001	-0.043	0.01	0.031	-0.013	0.033	-0.046	-0.2																						
source_Hyderabad	0.003	0.011	-0.013	0.002	-0.024	0.01	-0.051	0.028	-0.15	-0.2																					
source_Kolkata	0.061	0	0.016	-0.002	0	0.017	0.03	-0.027	-0.16	-0.22	-0.17																				
source_Mumbai	-0.023	-0.009	0.013	0.037	0.027	-0.029	-0.005	-0.007	-0.19	-0.26	-0.2	-0.22																			
departure_Early_Morning	-0.047	0.003	-0.012	0.034	0.028	-0.032	0.034	-0.036	0.017	-0.028	-0.013	-0.048	0.028																		
departure_Evening	0.1	-0.01	0.008	0.02	-0.004	-0.021	0.004	-0.005	-0.072	0.07	-0.067	-0.01	-0.003	-0.28																	
departure_Late_Night	-0.045	0.012	-0.034	-0.007	0.009	0.078	-0.012	-0.057	-0.015	0.011	-0.021	-0.012	0	-0.035	-0.035																
departure_Morning	-0.071	-0.003	0.018	-0.025	-0.04	0.028	-0.028	0.057	0.032	-0.017	0.006	0.023	-0.036	-0.3	-0.29	-0.037															
departure_Night	0.12	0.004	0.042	0.011	-0.025	-0.072	0.005	0.042	0.037	-0.063	0.068	0.029	-0.005	-0.23	-0.23	-0.029	-0.24														
arrival_Early_Morning	0.011	0.02	-0.06	-0.026	0.086	0.014	0.023	-0.059	-0.017	0.079	-0.023	-0.009	-0.055	-0.009	0.047	0.1	-0.071	0.072													
arrival_Evening	0.033	-0.014	0.056	-0.04	-0.044	0.007	-0.042	0.091	0.017	-0.026	-0.002	-0.017	0.012	0.026	-0.073	-0.028	0.1	-0.069	-0.14												
arrival_Late_Night	-0.11	0.026	-0.094	-0.06	0.1	0.065	0.003	-0.14	0.009	-0.028	0.009	0.011	-0.006	-0.088	0.1	0.018	-0.072	0.032	-0.051	-0.13											
arrival_Morning	0.12	-0.004	0.03	0.067	-0.064	-0.083	0.031	0.03	-0.024	0.004	0.008	-0.002	0.021	-0.012	0.077	0.015	-0.1	0.14	-0.12	-0.3	-0.11										
arrival_Night	-0.064	-0.005	0.02	-0.002	-0	-0.009	0.013	0.025	0.004	-0.007	0.01	0.016	0.008	-0.038	-0.016	-0.032	0.034	-0.097	-0.15	-0.39	-0.15	-0.34									
destination_Chennai	0.061	0	0.018	0.006	-0.06	0.018	-0.002	0.028	-0.15	0.061	0.026	0.011	0.047	-0.009	0.021	-0.008	-0.002	-0.009	0.062	-0.027	-0.042	0.054	0.018								
destination_Delhi	-0.12	0.006	-0.053	-0.017	0.034	-0.003	0.041	-0.038	0.06	-0.25	0.018	0.039	0.067	-0.005	-0.021	0.016	0.023	0.002	-0.093	-0.02	0.14	-0.034	-0.027	-0.19							
destination_Hyderabad	0.066	-0	-0.008	0.011	-0.026	0	-0.05	0.035	0.017	0.014	-0.16	0.034	0.043	-0.014	0.014	-0.017	0.001	0.003	0.07	-0.014	-0.065	-0.033	0.021	-0.16	-0.2						
destination_Kolkata	0.061	-0.004	0.021	-0.006	-0.001	0.034	0.03	-0.031	0.016	0.04	0.033	-0.19	0.057	-0.002	0.022	0.021	-0.02	-0.002	-0.007	0.1	-0.046	0.036	-0.09	-0.17	-0.22	-0.18					
destination_Mumbai	-0.044	-0.004	0.011	0.042	0.027	-0.031	-0	-0.017	0.043	0.067	0.05	0.054	-0.25	0.039	-0.012	-0.003	-0.009	0.002	0.013	-0.052	-0.039	-0.028	0.068	-0.2	-0.24	-0.2	-0.22				
Economy	-0.14	0.013	-0.94	-0.12	0.2	0.28	0.12	-0.3	-0.01	0.014	-0.003	0.019	-0.016	0.012	-0.011	0.029	-0.003	-0.05	0.042	-0.033	0.086	-0.04	-0.008	-0.012	0.021	-0.008	0.013	-0.011			
stops_two_or_more	0.093	-0.019	-0.064	-0.004	-0.038	-0.054	-0.038	0.025	-0.001	0.008	-0.021	0.032	-0.015	0.029	-0.02	-0.004	0.03	-0.072	-0.004	0.021	0.003	-0.016	0.019	0.069	-0.072	0.008	0.044	-0.052	0.11		
stops_zero	-0.52	-0.001	-0.19	-0.076	0.017	0.18	0.083	-0.1	-0.027	0.069	-0.027	-0.037	0.018	0.006	0.038	0.041	-0.046	0.026	0.021	-0.074	0.024	0.046	-0.01	-0.031	0.082	-0.031	-0.049	0.024	0.069	-0.079	

Average price depending on the number of days left before departure





Class = **Economy**

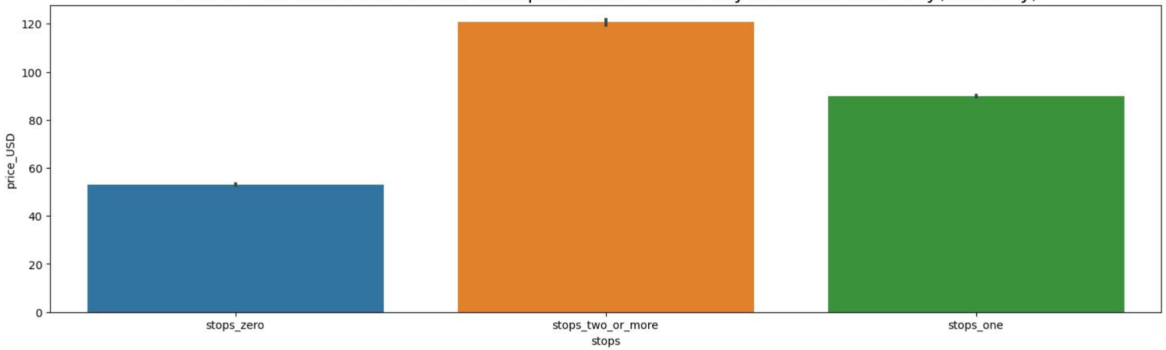
- 18 to 50 days left before departure (green): Prices remain stable during this period.
- 1 to 18 days left before departure (orange): Prices rise starting from 18 days before departure and continue to rise.

Class = **Business**

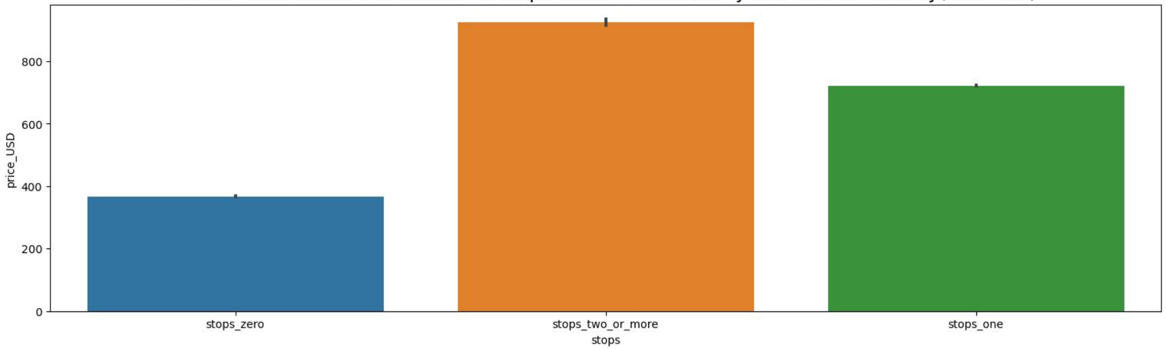
- 10 to 50 days left before departure (green): Prices remain stable during this period.
- 1 to 10 days left before departure (orange): Prices rise starting from 10 days before departure and continue to rise.

Average price depending on number of stops between origin city and destination city

Ticket Prices based the number of stops between source city and destination city(Economy)



Ticket Prices based the number of stops between source city and destination city(Business)

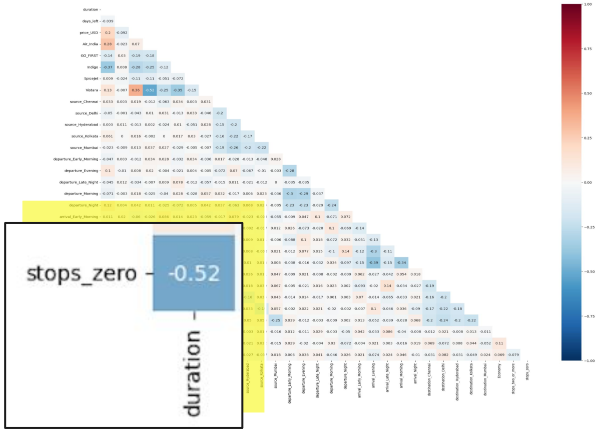


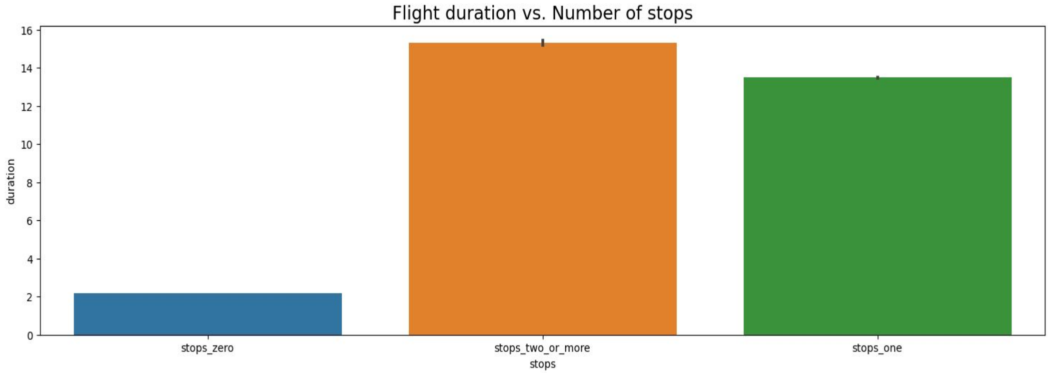
“For both Economy and Business class, as the number of stops increases, the prices also rise.”

Correlation between Covariates

High correlation between covariates (abs>=0.5) : Flight duration vs. stops_zero

	Correlation Value
Flight duration vs. stops_zero	-0.52





“As the number of stops increases, the flight duration(time) increases."

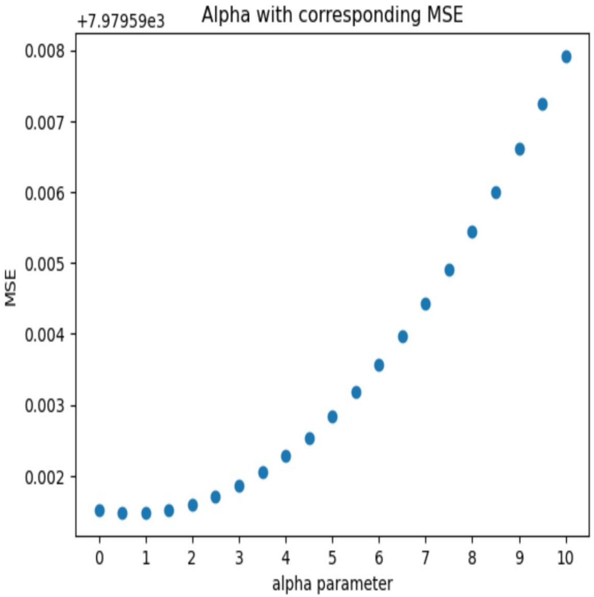
Ridge Regression

Grid search (hyperparameter tuning for alpha parameter)

- Tested values from 0 to 10 incremented by 0.5.
- Alpha of .5 provided the lowest MSE.
- Tested large alpha values, 950-1050 incremented by 5.
- Provided a slightly higher MSE and slightly lower R^squared value.
- Overall the tuning seemed to very slightly increase accuracy.

Standard Scaler (for “distance” and “days_left” features)

- Decreased runtime of fit for grid search by ~40%
- Decrease in runtime of score by ~70%
- No significant impact observed on MSE or R^squared values



	MSE	R^squared	Adj R^squared
Train	7.98e+03	.911	.911
Validation	7.72e+03	.913	.913
Test	7.94e+03	.911	.911

LASSO Regression

- Hyperparameter Tuning - generates 50 alpha values from {0.0001, 10} to tune model
- Best alpha - 0.000828642772854684
- Use grid search with 5-fold cross validation to find best alpha value, - using negative mean squared as the scoring metric
- Fit model on training data and uses validation and test sets for prediction
- Arrival time in the evening are the cheapest - largest coefficient of -549.7239506107213

	MSE	R^2
Validation	7721.137484452601	0.9129648213744022
Test	7943.1293464585	0.9112205609361346

Decision Tree Regression: all data & random samples

Sample Size	R^2 - Test	MSE - Test Data	R^2 - Validation	MSE - Validation
100	0.862	12.3 x 10^3	0.869	11.7 x 10^3
1,000	0.896	9.30 x 10^3	0.898	9.06 x 10^3
10,000	0.934	5.87 x 10^3	0.937	5.60 x 10^3
100,000	0.979	1.91 x 10^3	0.978	1.95 x 10^3
All data	0.977	2.06 x 10^3	0.976	2.10 x 10^3

Kernel Regression

- Hyperparameter Tuning:
  - Performed grid search for hyperparameters `alpha` and `gamma`.
  - Explored `alpha` values of 0.01 and 0.001, with `gamma` values spanning `np.logspace(-2, 0, 3)`.
  - Identified best parameters as `{'alpha': 0.01, 'gamma': 0.01}`.
- Model Evaluation:
  - On the validation set, the model achieved an MSE of 80,250.18 and an R^2 of 0.105.
  - On the test set, the model had an MSE of 84,477.87 and an R^2 of 0.098.
- Computational Details:
  - The grid search involved 3-fold cross-validation for each of 6 parameter combinations, totaling 18 fits.
- Observations:
  - The increase in sample size to 10,000 data points led to higher MSE and significantly lower R^2 values compared to the initial results with a smaller sample size. This indicates a decrease in model performance and suggests overfitting with the smaller sample or that the model may not generalize well to larger datasets.

Kernel Regression

Set	Mse	R^2 Value
Validation	80,250.18	0.105
Test	84,477.87	0.098

Comparison of Regression Models

Method	R^2 Value - Test	Mean Squared Error - Test
Ordinary Linear Regression	0.911	7.94×10^3
LASSO Regression	0.911	7.94×10^3
Ridge Regression	0.911	7.94×10^3
Decision Tree Regression	0.977	2.07×10^3
Kernel Regression	0.098	84.5×10^3

References

Dataset: https://www.kaggle.com/datasets/shubhambathwal/flight-price-prediction/data

INR to USD Conversion: https://www.exchangerates.org.uk/

Questions?