Flight Fare Prediction Project

Group M: Ricatti-Tavano-Valeri

Flight Fare Prediction - Introduction

The objective of the study is to analyse the flight booking dataset obtained from “Ease My Trip” website and to conduct various statistical methods in order to get meaningful information from it. 'Easemytrip' is an internet platform for booking flight tickets, and hence a platform that potential passengers use to buy tickets.
Data was collected in two parts: one for economy class tickets and another for business class tickets. A total of 300261 distinct flight booking options was extracted from the site. Data was collected for 50 days, from February 11th to March 31st, 2022.

The aim of our study is to answer the below research questions:

- Does price vary with Airlines?
- How is the price affected when tickets are bought in just 1 or 2 days before departure?
- How does the ticket price vary between Economy and Business class?
- How the price changes with respect to the flight duration?

Exploratory Data Analysis

Index	airline	flight	source_city	departure_time	stops	arrival_time	destination_city	class	duration	days_left	price
0	SpiceJet	SG-8709	Delhi	Evening	zero	Night	Mumbai	Economy	2.17	1	5953
1	SpiceJet	SG-8157	Delhi	Early_Morning	zero	Morning	Mumbai	Economy	2.33	1	5953
2	AirAsia	I5-764	Delhi	Early_Morning	zero	Early_Morning	Mumbai	Economy	2.17	1	5956
3	Vistara	UK-995	Delhi	Morning	zero	Afternoon	Mumbai	Economy	2.25	1	5955
4	Vistara	UK-963	Delhi	Morning	zero	Morning	Mumbai	Economy	2.33	1	5955

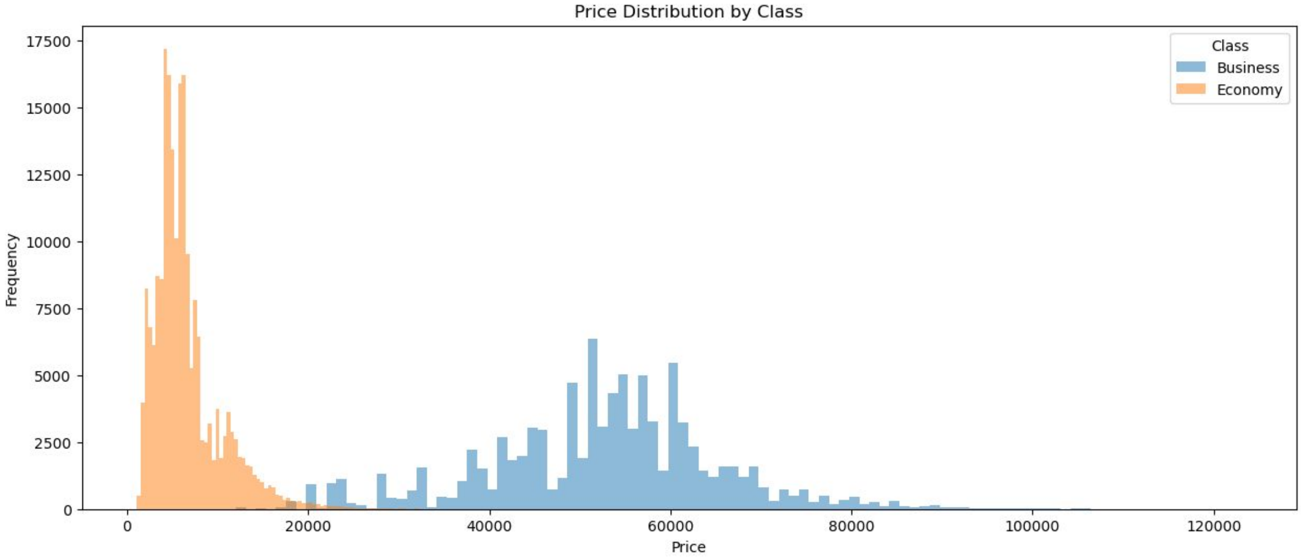
Dataset contains information about flight booking options from the website Easemytrip for flight travel between India's top 6 metro cities. There are 300261 data points and 11 features in the cleaned dataset.

Exploratory Data Analysis

- **Airline**: The name of the airline company, it is a categorical feature having 6 different airlines.
- **Flight**: Plane's flight code. It's a categorical feature.
- **Source City**: City from which the plane take off. It's a categorical feature with 6 cities.
- **Departure Time**: Categorical feature obtained by grouping time periods into bins. It has 6 time tables.
- **Stops**: Stores the number of stops between the source and destination cities, it's a categorical feature with 3 distinct values.
- **Arrival Time:** Categorical feature similar to Departure Time.
- **Destination City**: Categorical City similar to Source City.
- **Class**: It has two different values, Business and Economy.
- **Duration**: Continuous variable that report the time of the flight in hour.
- **Days Left**: Continuous characteristic derived by subtracting the trip day by the booking date.
- **Price**: Continuous variable that stores the ticket price.

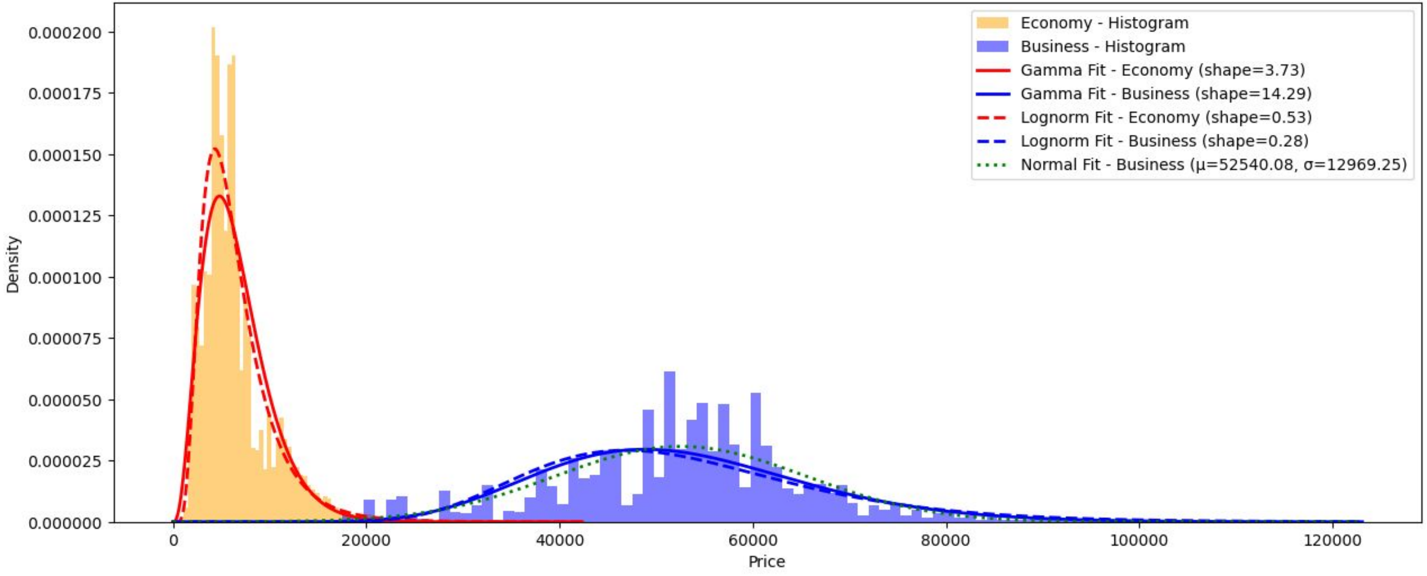
Price Analysis Distribution

	count	mean	std	min	25%	50%	75%	max
class								
Business	93487	52540	12969	12000	45185	53164	60396	123071
Economy	206666	6572	3743	1105	4173	5772	7746	42349



Is easy to note that the distribution of the prices has two distinct peaks, one for the economy class eand the other for the business class, so it can be defined as a **bimodal** distribution.

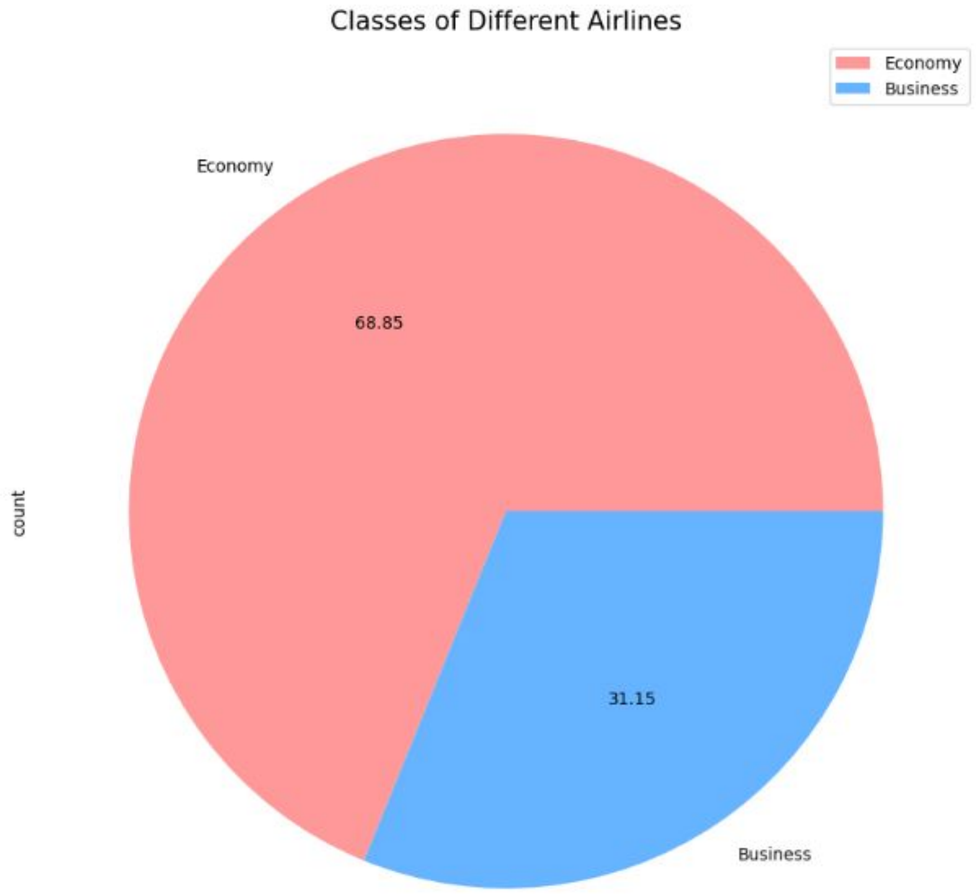
Price Analysis Distribution



The first peak correspond to the **economy class** that is characterized by an highly right-skewed distribution. Both Gamma and Lognormal distribution aligns with the peak of the histograms, but the lognormal seems to follow better the distributions. While the second peak corresponds to the **business class**, we can see that the gamma distribution fit better the central part of the data than the lognormal.

Class Distribution

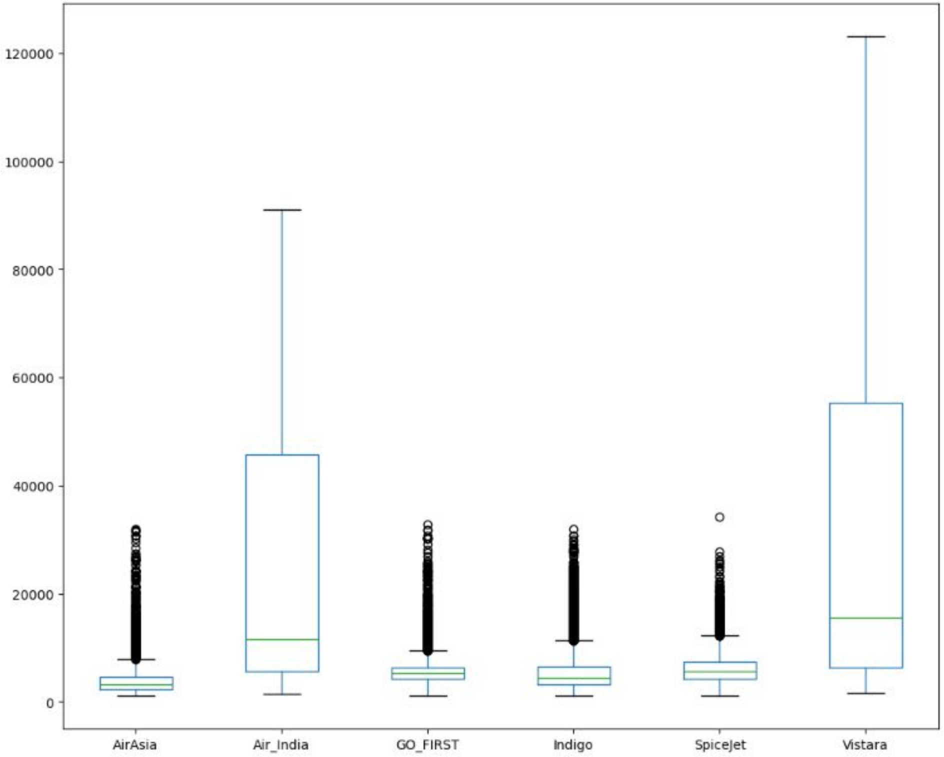
Pie chart shows that the majority of the flights belongs to Economy class with almost 70% of flight. In a future model training we can expect that the model will predict lower prices, corresponding to economy class, better than higher prices. So the class feature will be crucial in price prediction. So it is reasonably to try split the dataset in two and prepare separate model for Economy and Business class.



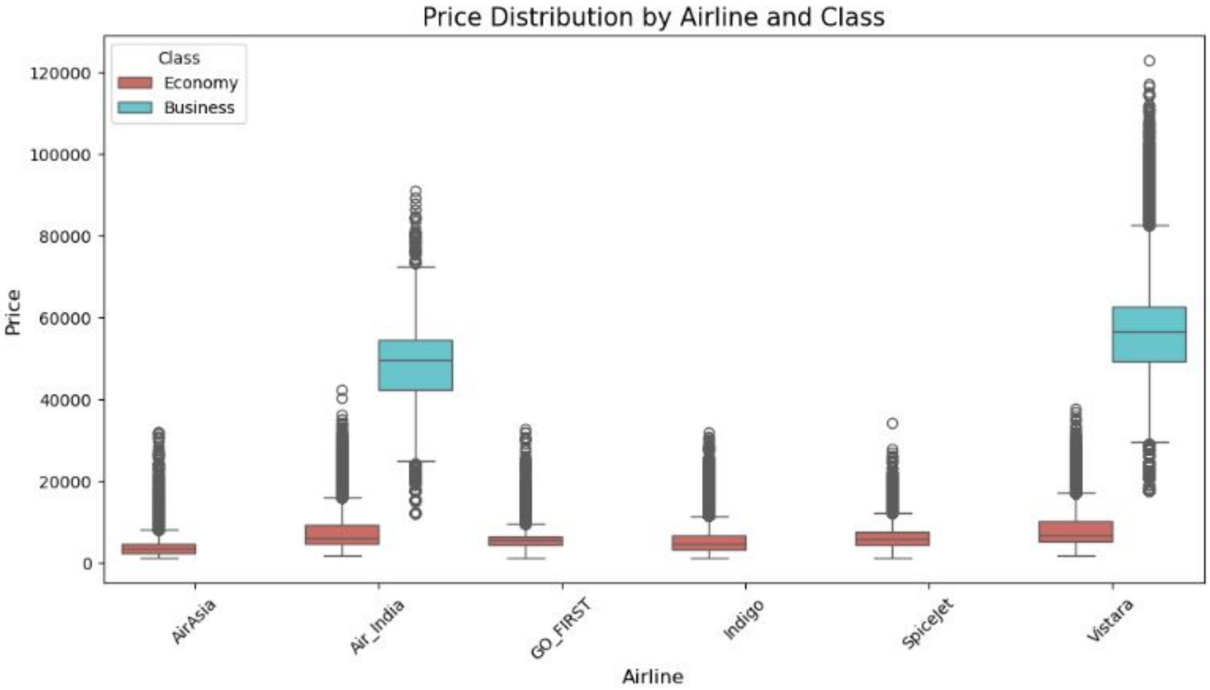
Price vs Airline

It is observed that *Vistara* and *Air India* exhibit the highest price distributions compared to other airline companies. Specifically, their median, maximum values, and interquartile range (IQR) are notably higher. In particolar a wider interquartile range indicates a greater variability in ticket prices. Both the median values and the maximum values of *Vistara* and *Air India* suggest the fact that these two company offers business-class ticket.

airline	AirAsia	Air_India	GO_FIRST	Indigo	SpiceJet	Vistara
count	16098.00	80892.00	23173.00	43120.00	9011.00	127859.00
mean	4091.07	23507.02	5652.01	5324.22	6179.28	30396.54
std	2824.06	20905.12	2513.87	3268.89	2999.63	25637.16
min	1105.00	1526.00	1105.00	1105.00	1106.00	1714.00
25%	2361.00	5623.00	4205.00	3219.00	4197.00	6412.00
50%	3276.00	11520.00	5336.00	4453.00	5654.00	15543.00
75%	4589.00	45693.00	6324.00	6489.00	7412.00	55377.00
max	31917.00	90970.00	32803.00	31952.00	34158.00	123071.00

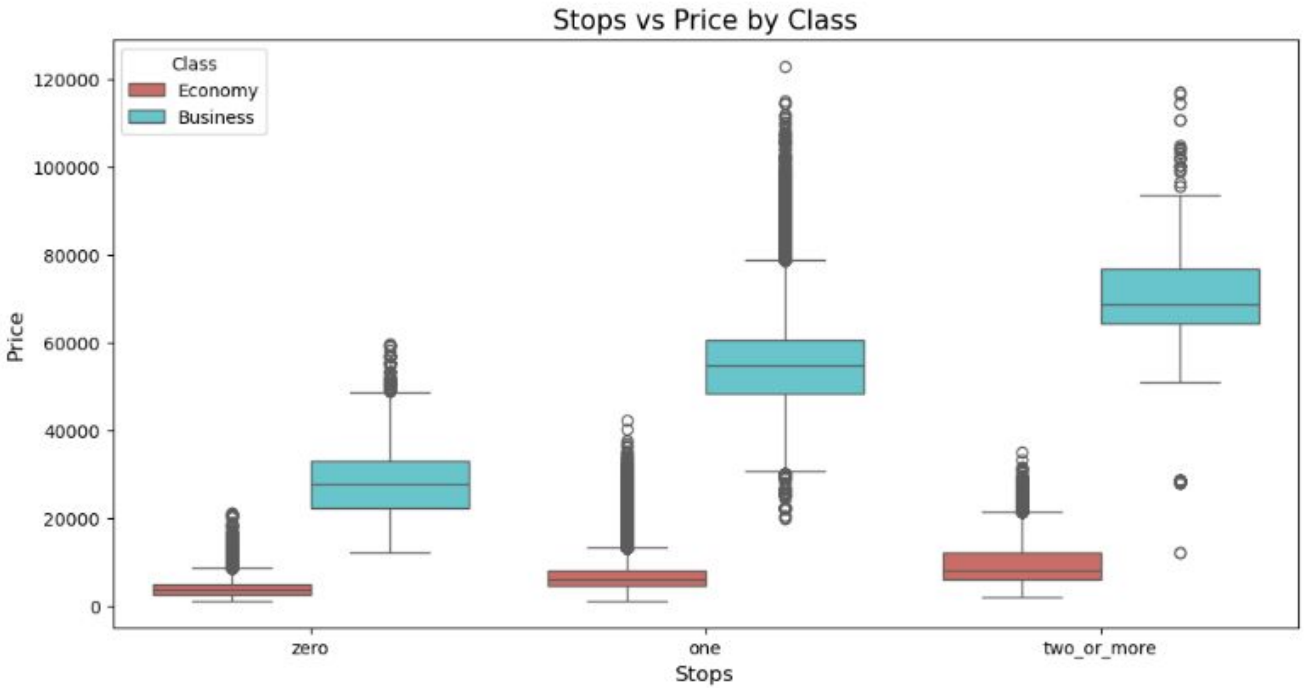


Price vs Airline by Class



As before anticipated only Air India and Visitara offers business class. It is notable that Economy Class have lower and more stable ticket prices, with similar median values and IQR. Only Air India and Vistara show a higher median value for the economy class. For the Business class is important so note that both of the airlines have a large interquartile range indicating high price variability. Also numerous outlier are visible for the *Vistara* and *Air India* while the budget airlines show a more stable fare structures. We can expect that the *Vistara* and *Air India* airlines will be more influential in the prediction phase.

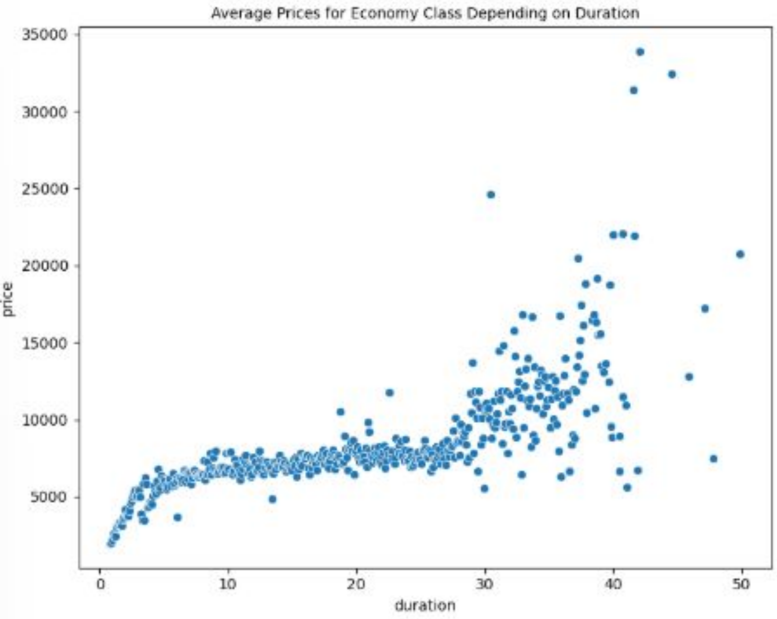
Price vs Stops



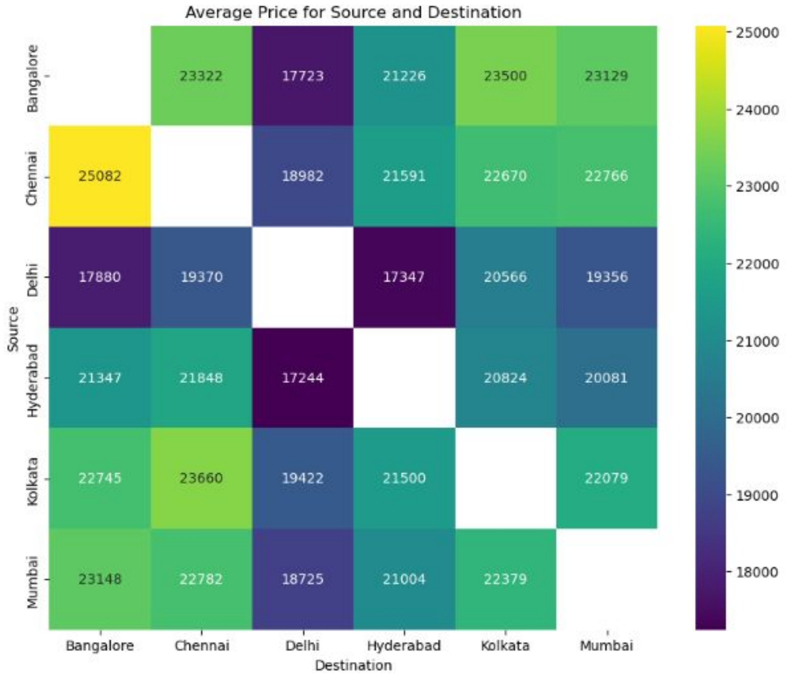
Business class fares are consistently higher than Economy class fares across all stop categories. Prices tend to increase with the number of stops, especially for Business class, where flights with two or more stops show the highest median price and greater variability. Outliers are present in both classes, particularly in Business class.

Price vs Flight Duration

These scatter plot visualizes the relationship between flight duration and ticket price, distinguishing between economy and business classes. It is clear that the flight price increases with increasing flight duration in both classes. Business class prices increase steeply compared to Economy. There is a positive correlation between flight duration and the price also because the duration is strictly related to the number of stop of the flight

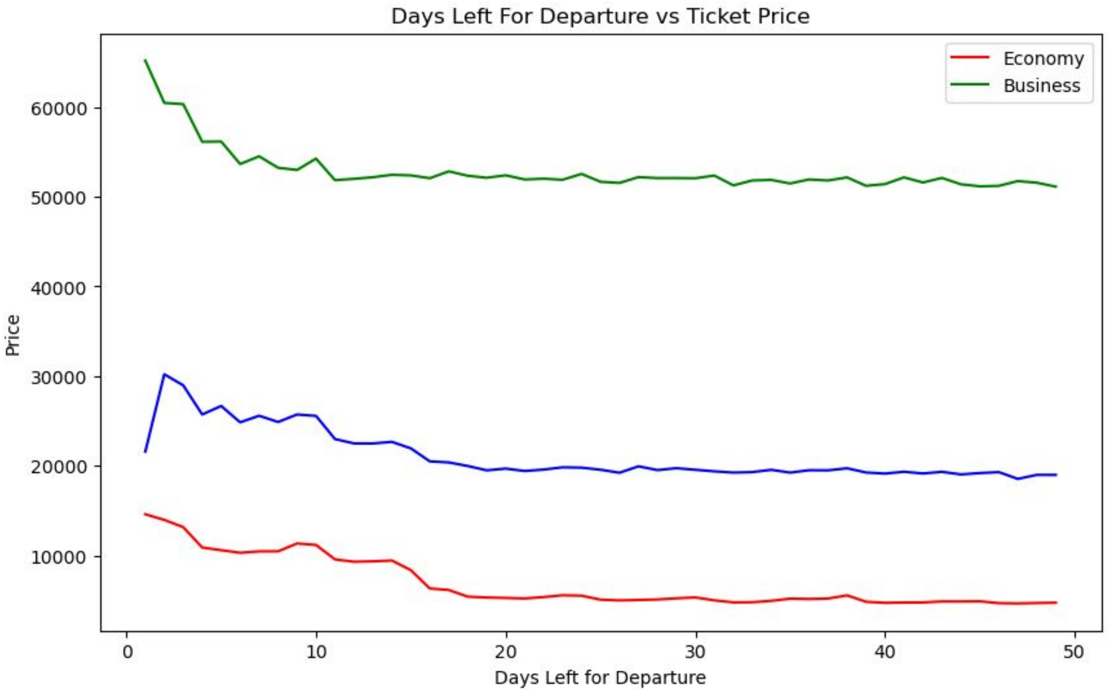


Price vs Source and Destination



This heat map visualizes the average ticket price between different source and destination cities, the highest average price can be justified by the presence of business tickets or by the absence of direct flights. While the lower average price can be explained by route with high competition and multiple flight options

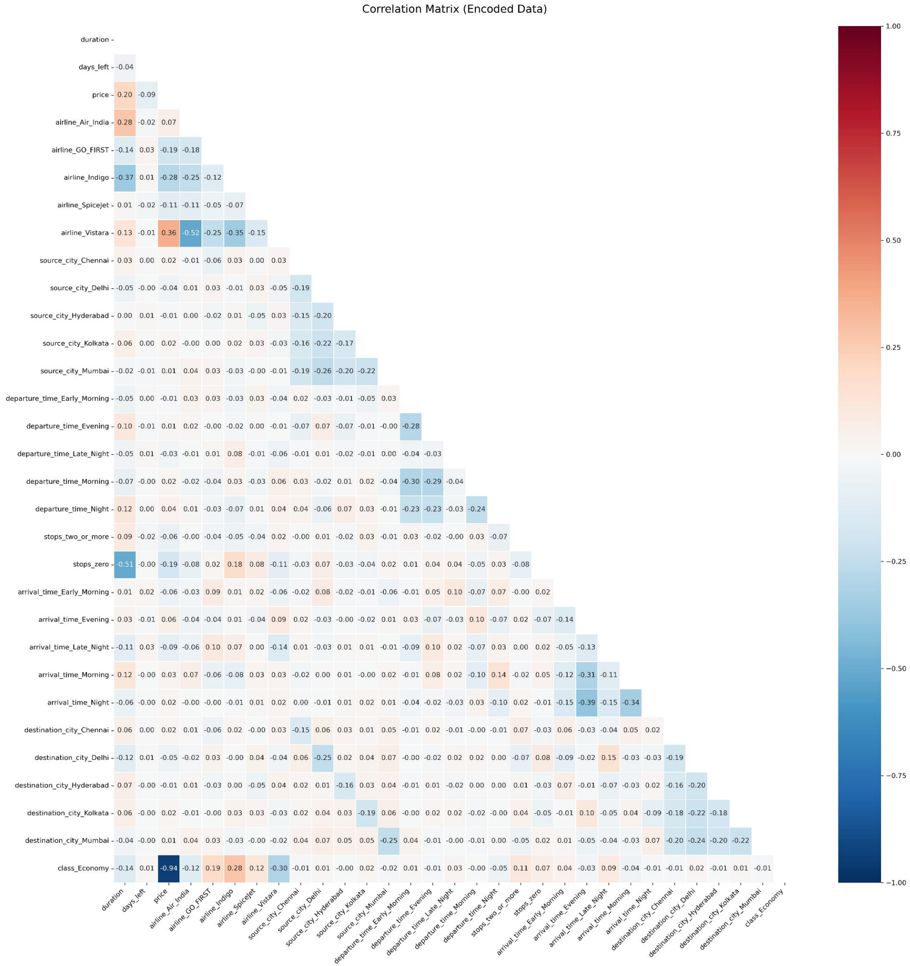
Price vs Days Left



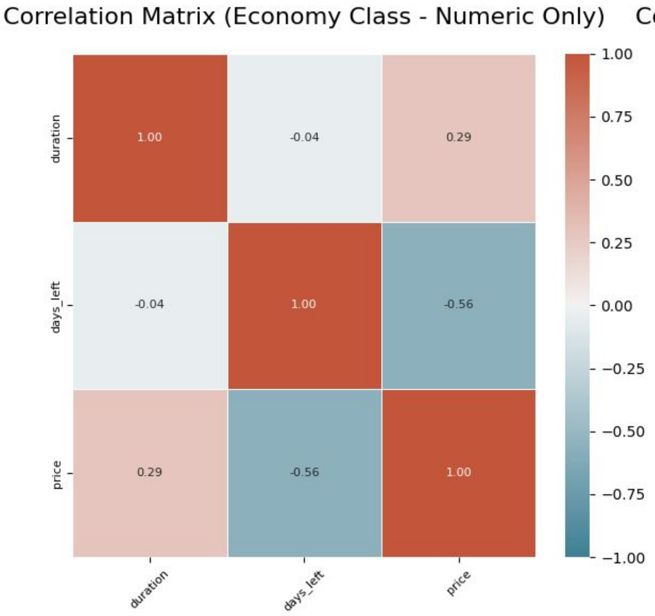
This line plot illustrates the relationship between the number of days left before departure and ticket price. Ticket prices spikes in the 0-15 days left interval. After 15 days left the price stabilizes. This graph suggest a negative correlation between price and days left and also underline a non linear relationship.

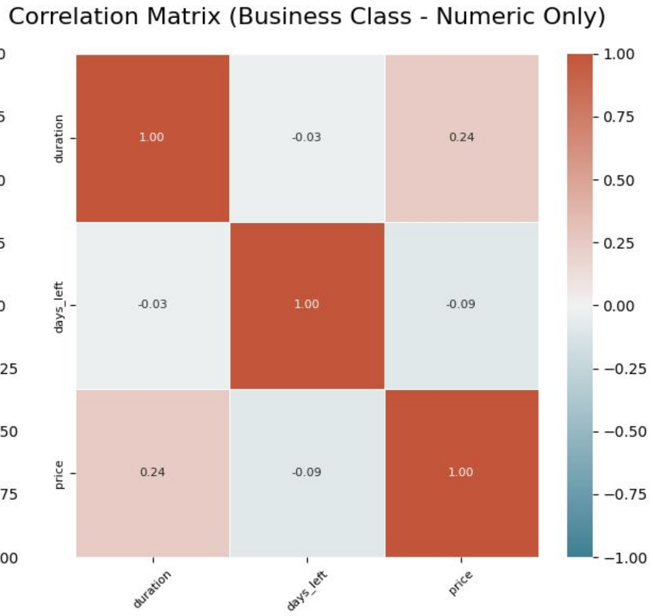
Correlation Matrix

The correlation matrix reveals notable relationships between variables within the dataset. There is a strong negative correlation (-0.94) between class and ticket price. A moderate positive correlation (0.20) is noticed between duration and price as already noted also by the negative correlation with the zero stop. Other correlations regards the different airlines. Variables related to *departure* and *arrival* times have minimal correlations with *price* indicating that these factors do not significantly influence flight costs.



Correlation Matrix





In these plots we can observe the correlation between the numeric variables divided by class. We can observe a strong negative correlation (-0.56) between price and days left in the economy class suggesting that budget-conscious travelers book economy ticket early to get lower fares. Another moderate correlation (0.29) there is between duration and price for the economy class underlining the fact that a longer flight has an higher flares.

Evaluation Metrics

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

$$AIC = 2k - 2\ln(L)$$

$$BIC = \ln(n)k - 2\ln(L)$$

Linear Regression and Variants

We will present:

- Complete Model: all features
- Partial Model: log Price, removed
- Polynomial Model: quadratic and interaction for duration and days_left
- Ridge Regression CV
- Splitted models - Economy
- Splitted models - Economy Quad: price log and quadratic terms
- Splitted models - Business Quad: price log and quadratic terms

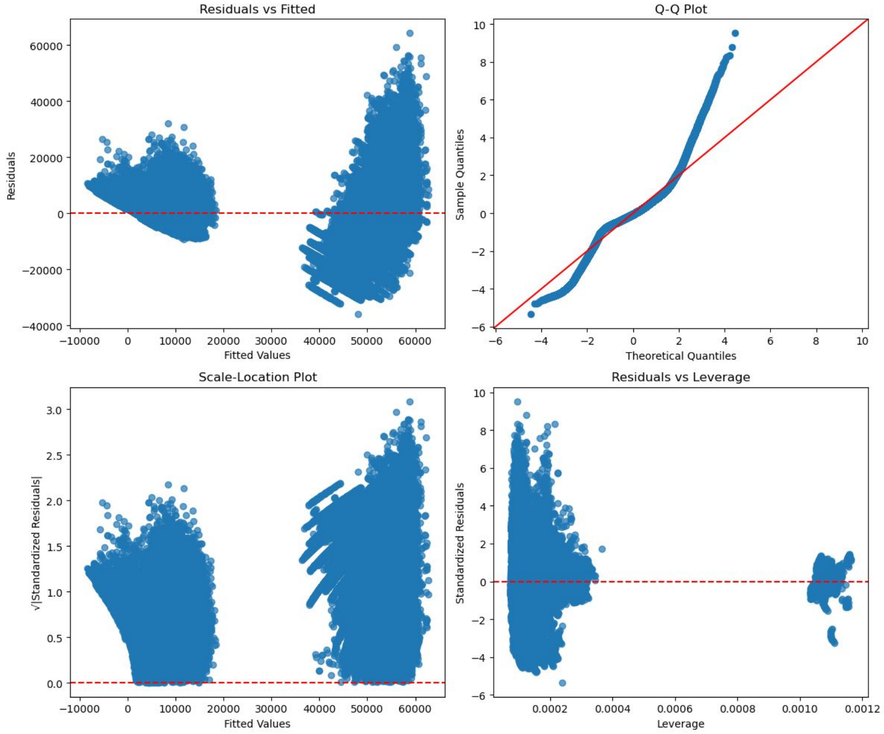
Complete Model

Results:	
R^2:	0.912
R^2 test set:	0.911
MAE:	4584
RMSE:	6792
MAPE:	46.49%

Coef:	
const:	52,550
class Economy:	44,930
stops zero:	-7,597
airline Vistara:	4,097
destination Delhi:	-1,559

- All Classes
- All features
- only 1st degree
- No interaction
- One hot encoding
- All VIF < 10
- One p > 0.5 (Mumbay destination)

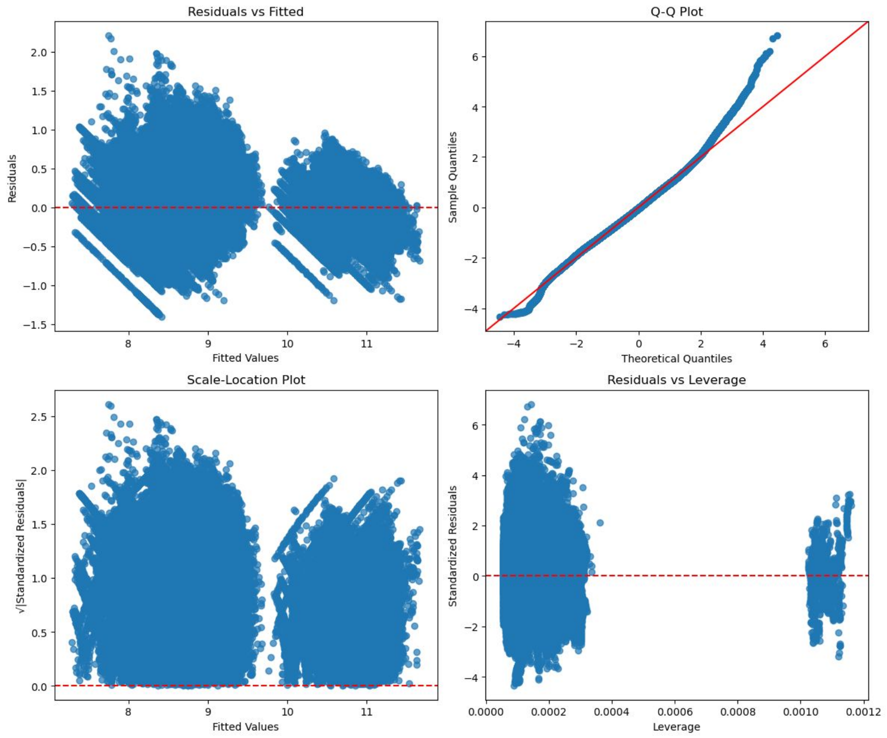
Heteroscedasticity: variance increases with fitted values in Residuals vs Fitted Plot, confirmed by trend in Scale-Location Plot. Presence of clusters.
Non normality, especially in the tails, emerge in Q-Q Plot.
High leverage points indicates the presence of outliers



Partial Model

Results:		**Coef:**	
R^2:	0.915	const:	10.558
R^2 test set:	0.881	class Economy:	-2.026
MAE:	4605	stops zero:	-0.451
RMSE:	7833	airline Vistara:	0.647
MAPE:	26.19%	airline Air India:	0.521

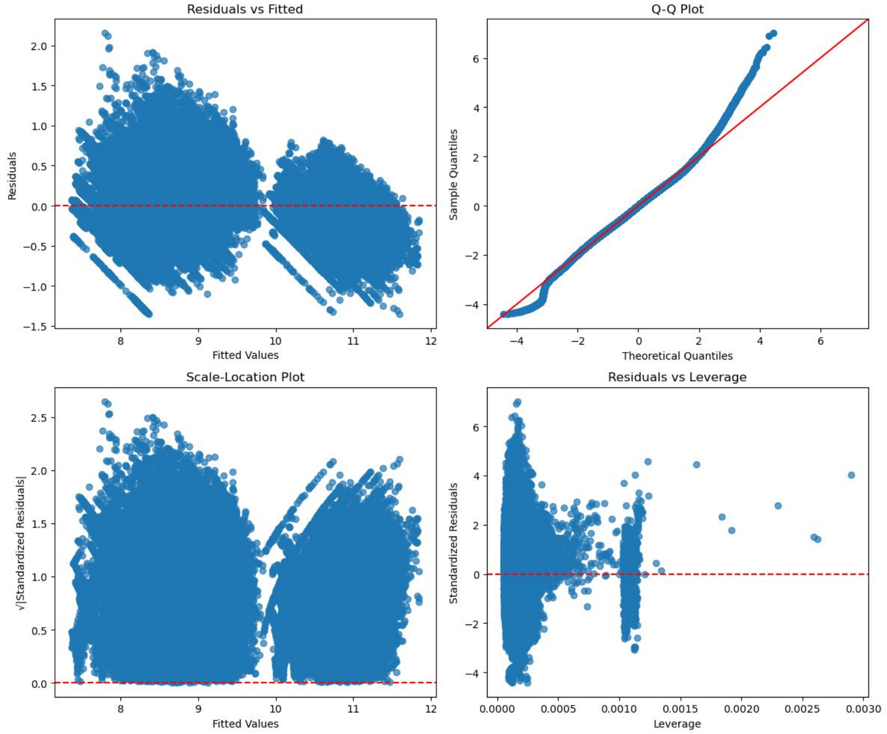
- Log Price
- improvements in R^2, AIC, BIC
- Coefficients for squared features not influent (-0.0004, 0.0006)
- Normality:Q-Q plot slightly improved



Polynomial Model

Results:		**Coef:**	
R^2:	0.924	const:	10.843
R^2 test set:	0.860	class Economy:	-2.025
MAE:	4716	airline Vistara:	0.637
RMSE:	8495	airline Air India:	0.514
MAPE:	24.88%	airline SpiceJet:	0.458

- Log Price
- slightly improvements in R^2, AIC, BIC
- Coefficients for squared terms and interaction are not influent (-0.0004, 0.0006, 0.0002) but high multicollinearity



Polynomial - Ridge CV

Results:

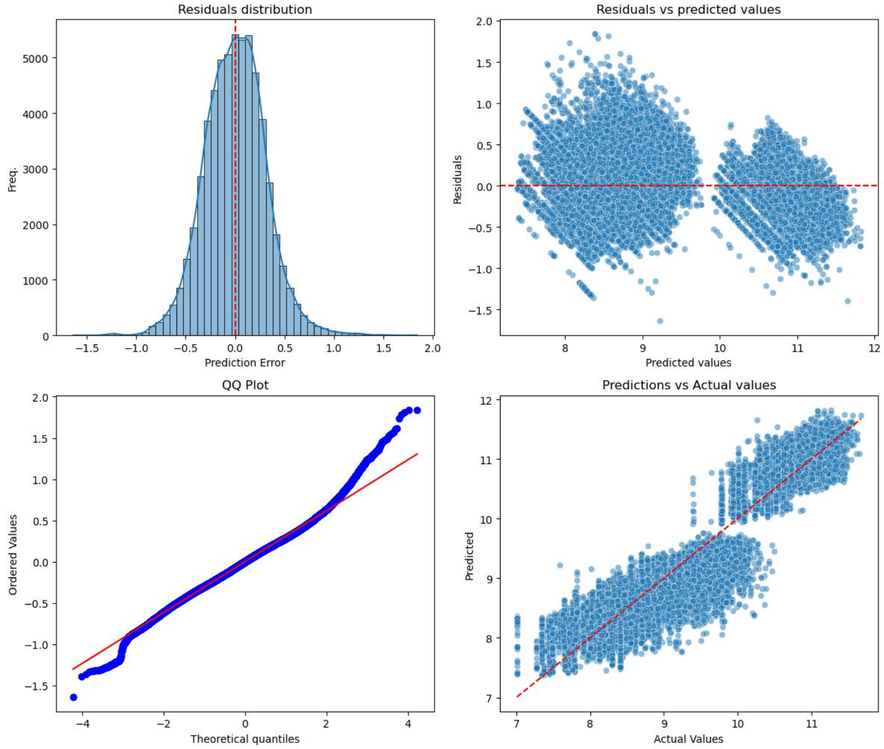
R^2:	0.923
R^2 test set:	0.923
MAE:	4716
RMSE:	8492
MAPE:	24.88%

Coef:

const:	0.0
class Economy:	-0.937
days left:	-0.645
days left^2:	0.418
airline Vistara:	0.315

- Log Price
- Scaled values
- Best alpha 1.0

Normality: no improvements



Splitted - Economy

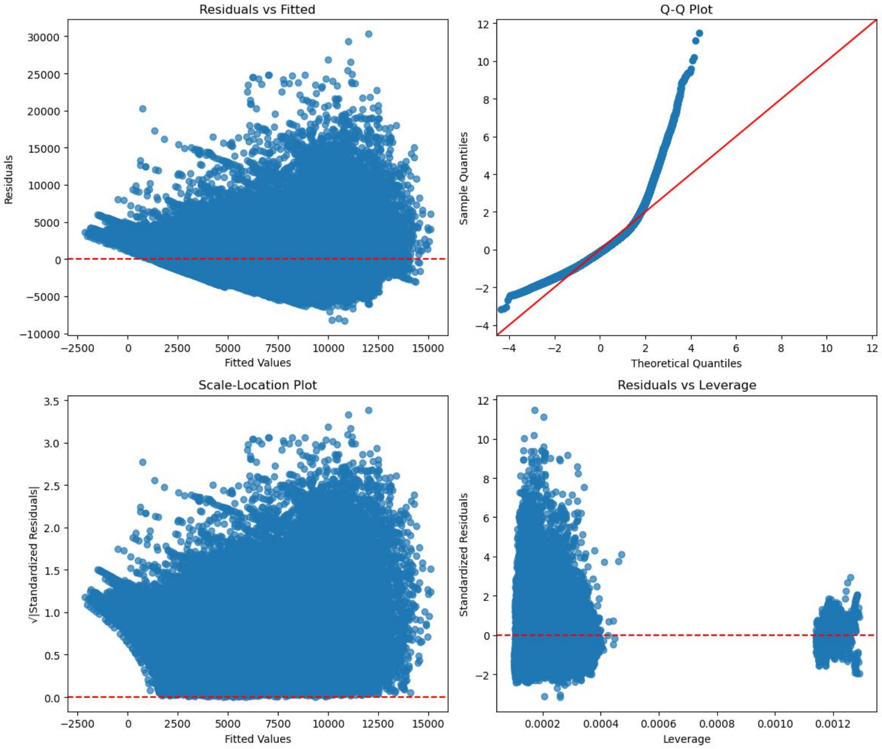
Results:

R^2:	0.502
MAE:	1893
RMSE:	2612
MAPE:	33.8%

Coef:

const:	8,034
airline Vistara:	3,304
airline Air India:	2,749
stops zero:	-1,886
airline SpiceJet:	1,832

- All features
- Non normality
- Heteroscedasticity



Splitted - Economy Quad

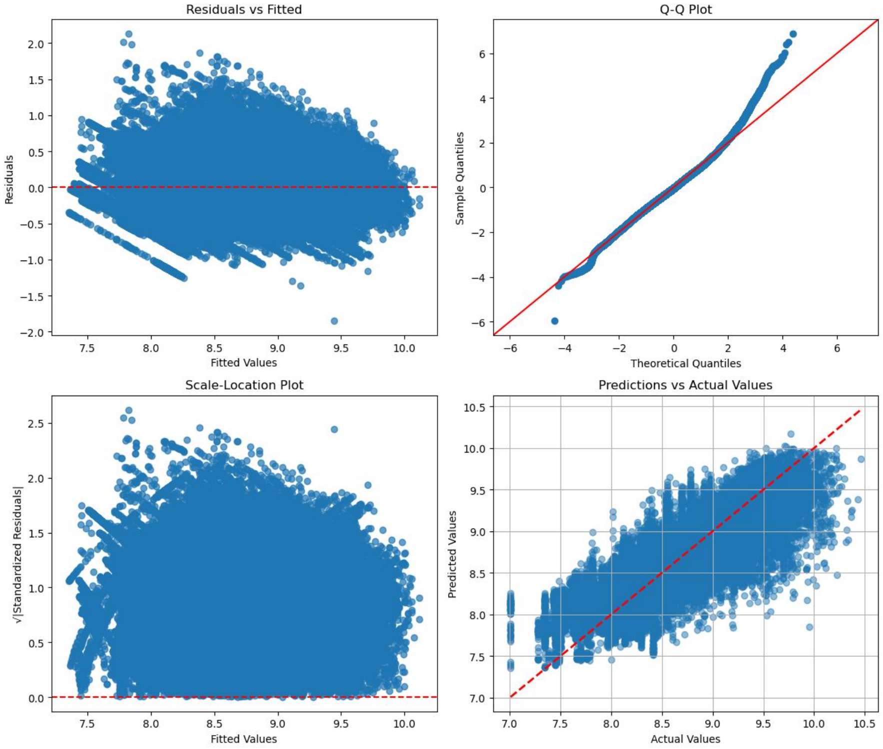
Results:

R^2:	0.654
MAE:	1540
RMSE:	2329
MAPE:	24.53%

Coef:

const:	9,067
airline Vistara:	0,615
airline Air India:	0,526
airline Go First:	0,429
airline SpiceJet:	0,429

- Log Price
- Quadratic terms
- All terms included
- Non normality
- Heteroscedasticity

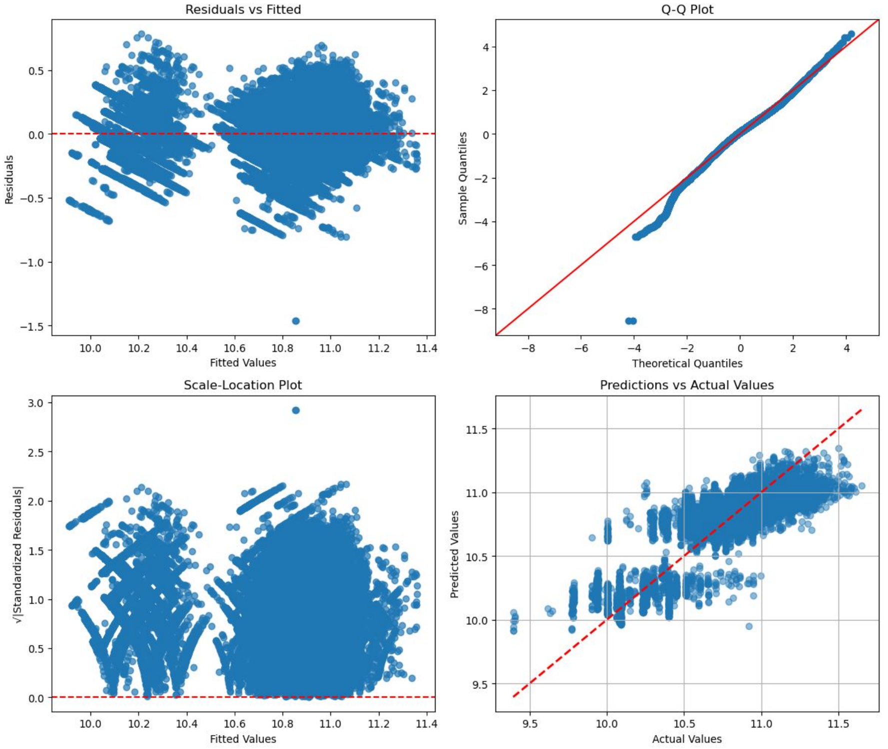


Splitted - Business Quad

Results:	
R^2:	0.630
MAE:	6647
RMSE:	8700
MAPE:	13.27%

Coef:	
const:	10.675
stop Zero:	-0.552
stop two or more:	0.203
airline Visitara:	0.151
source Kolkata:	0.768

- Log Price
- Quadratic terms (low coef -0.0009, 9.806e-05)
- All terms
- Non normality
- Heteroscedasticity



Non-parametric regression

We will present:

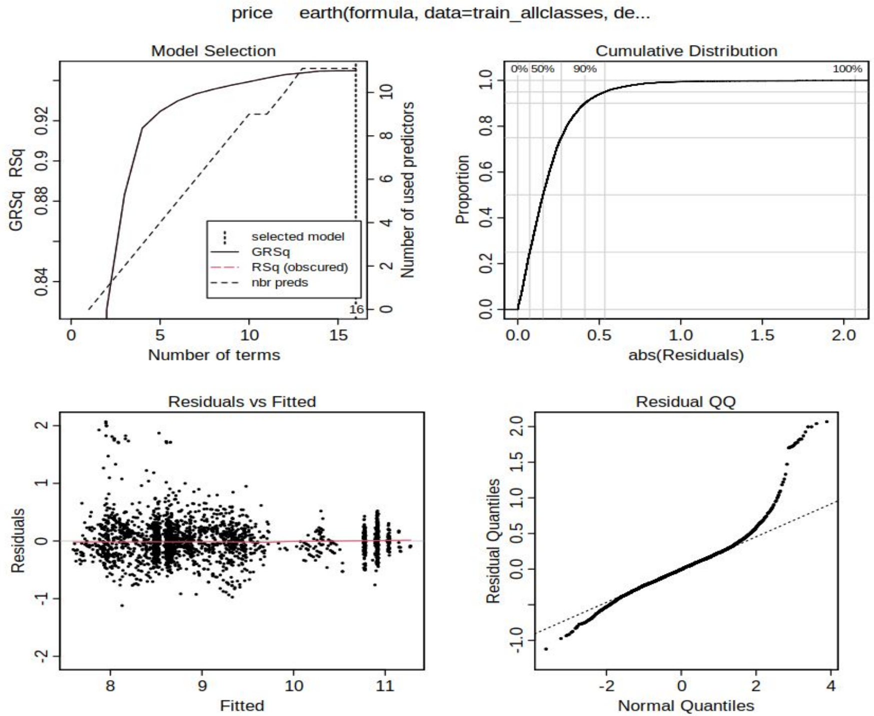
- MARS: Multivariate Adaptive Regression Spline
- Splitted models - Mars Economy
- Splitted models - Mars Business

MARS

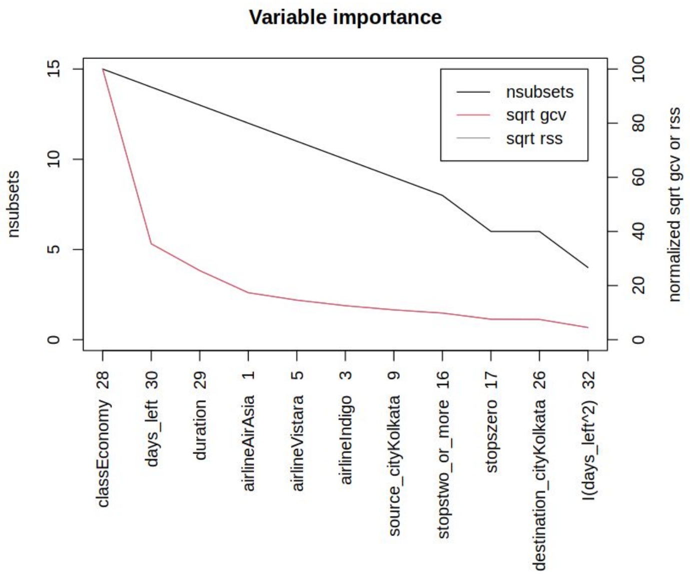
Results:	
GR^2:	0.945
MAE:	3269
RMSE:	5697
MAPE:	20.04%

Coef:	
intercept:	10.764
Economy * h(-0.37-days_left)…:	-2.764
class Economy:	-2.240
Economy*h(-0.37-days_left):	0.867
Asia*stopszero:	0.593

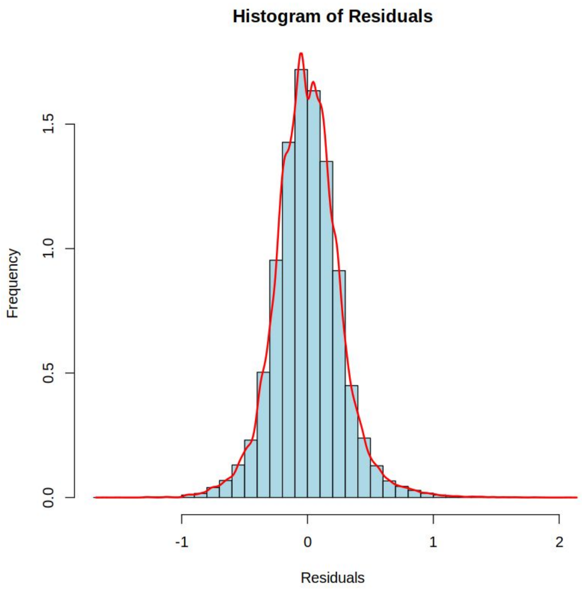
- Log Price
- Quadratic terms
- degree 4 for interactions



MARS



Variable importance
nsubsets
sqrt gcv
sqrt rss
nsubsets
normalized sqrt gcv or rss
0
5
10
15
0
20
40
60
80
100
classEconomy 28
days_left 30
duration 29
airlineAirAsia 1
airlineVistara 5
airlineIndigo 3
source_cityKolkata 9
stopstwo_or_more 16
stopszero 17
destination_cityKolkata 26
I(days_left^2) 32



Histogram of Residuals
Frequency
Residuals
0.0
0.5
1.0
1.5
-1
0
1
2

MARS - Economy

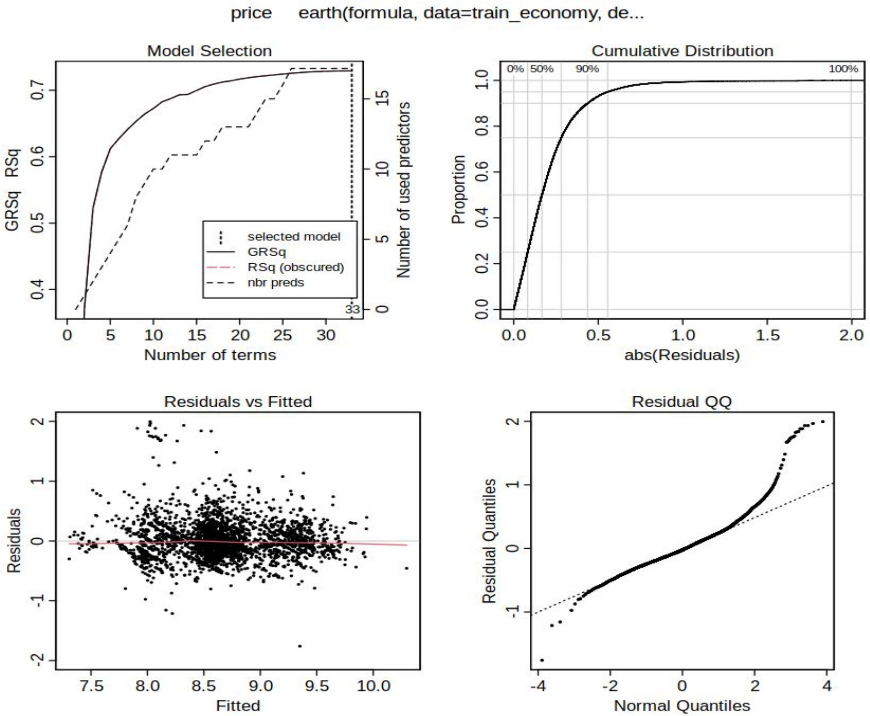
Results:

GR^2:	0.729
MAE:	1347
RMSE:	2119
MAPE:	20.94%

Coef:

h(-0.973592-duration)*h(1.21141-I(duration^2)):	-13.665
intercept:	5.806
h(-0.378701-days_left)*h(0.675968-I(days_left^2)):	-13.665

- Log Price
- Quadratic terms
- degree 3 for interactions

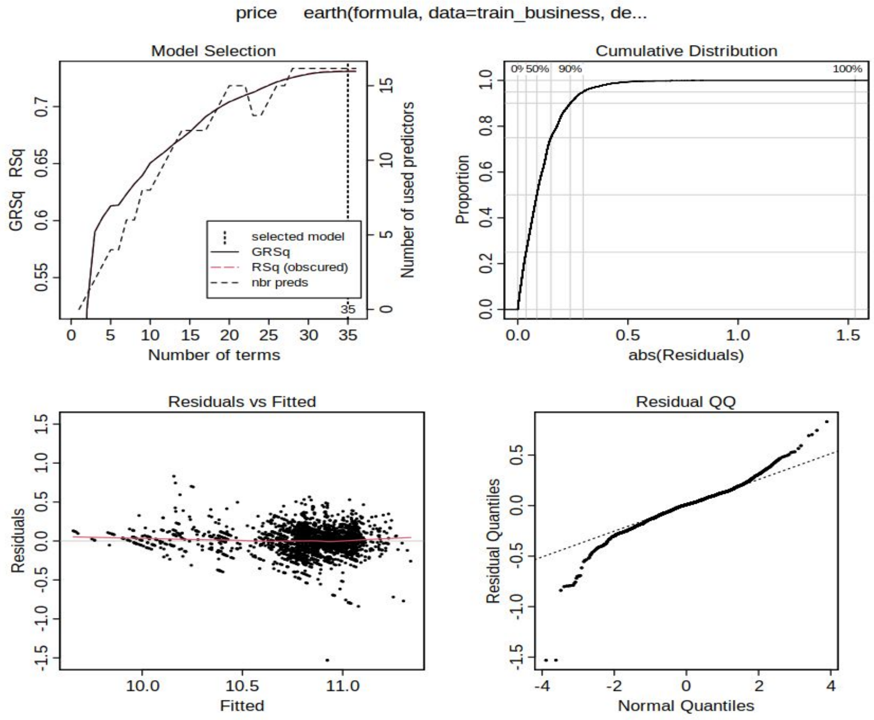


MARS - Business

Results:	
GR^2:	0.731
MAE:	5744
RMSE:	7867
MAPE:	11.16%

Coef:	
Intercept:	10.742
h(-1.13597-duration):	-2.764
airlineVistara * h(-1.68321-duration)	-2.518

- Log Price
- Quadratic terms
- degree 4 for interactions



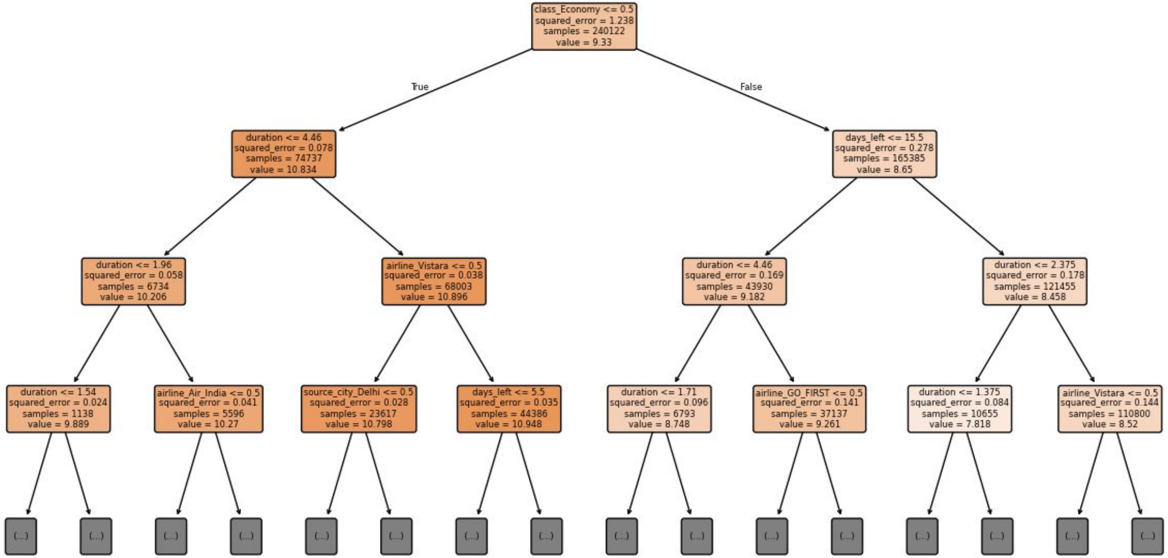
Regression Trees, XGBoost and Random Forest

Regression Tree - Complete Model

- As first approach we use a regression Tree, a simple and effective regression model for predicting price. The latter variable has already been logarithmically transformed.
- DecisionTreeRegressor - sklearn.tree library.
- **parameters**: criterion (mse) , splitter (best/random), max_depth, random_state.

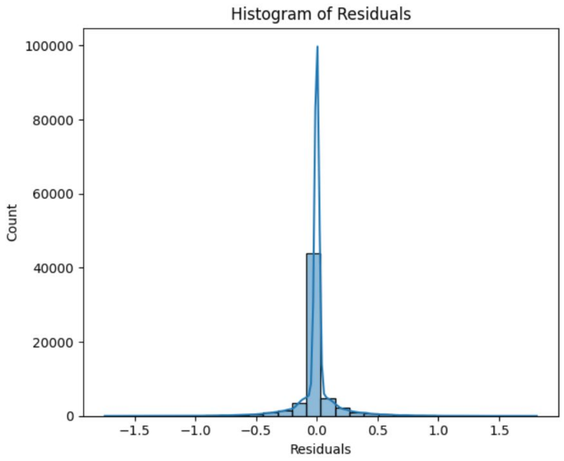
```
▾ DecisionTreeRegressor
DecisionTreeRegressor(random_state=30)
```

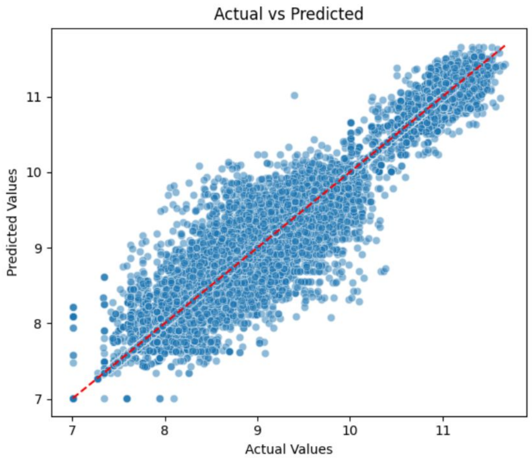
First 3 levels of the regression tree.



- The **regression tree shows** how various factors influence airline ticket prices, with the most significant feature being flight class.
- The root node splits based on whether the flight is in **Economy** class, showing that **Business** tickets have significantly higher prices.
- Flight **duration** is another key determinant, as longer flights generally lead to higher fares, with multiple splits refining this effect.
- Airlines also play a crucial role, particularly **Air India** and **Vistara**, indicating that different carriers have distinct pricing structures.
- The number of **days left** until departure is an important factor, as last-minute bookings tend to be more expensive.
- Additionally, the departure city, especially whether the flight originates from **Delhi**, affects the price. These features collectively determine ticket pricing, with each split in the tree reducing variance and refining predictions.

Some diagnostic plots on the Regression Tree.

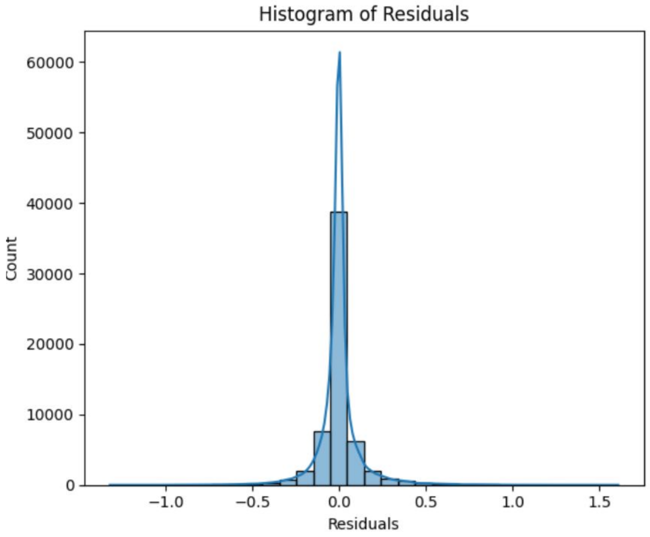


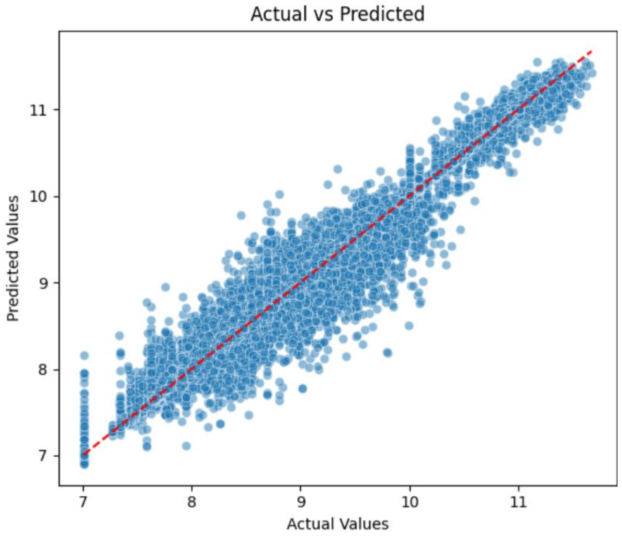


XGBoost - Complete Model

- The results of the performance metrics related to the Regression Tree shows a good fit on the test set.
- Anyway, we try to improve our model. For this purpose, we will use **Xgboost** to find the best alpha and train the model with it.
- cpp_alphas has been retrieved from the previous **regression tree,** using np.linspace(0, max(cpp_alpha), num=10).
- We will set some hyperparameters to use this library, such as **max_depth**, **learning_rate**, **n_estimators** and **num_boost_round**. This part of the notebook has been implemented using a GPU to speed up the training phase.

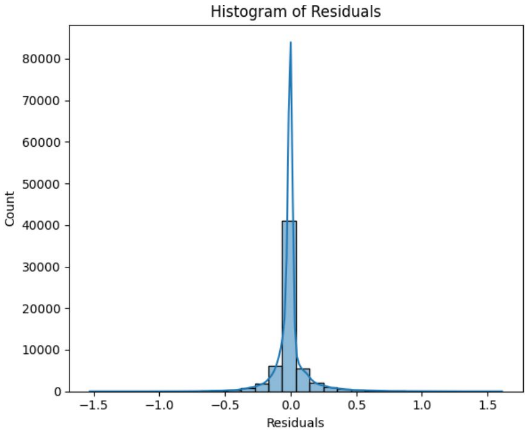
Some diagnostic plots - XGBoost.

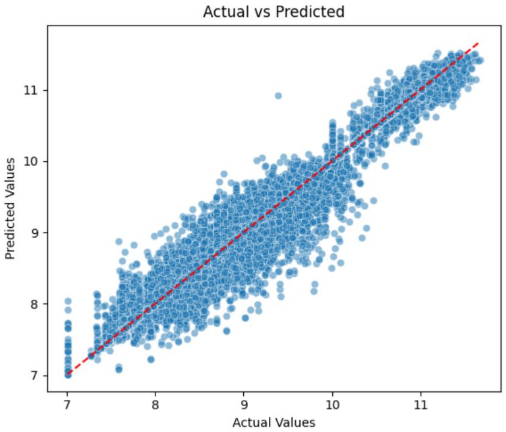




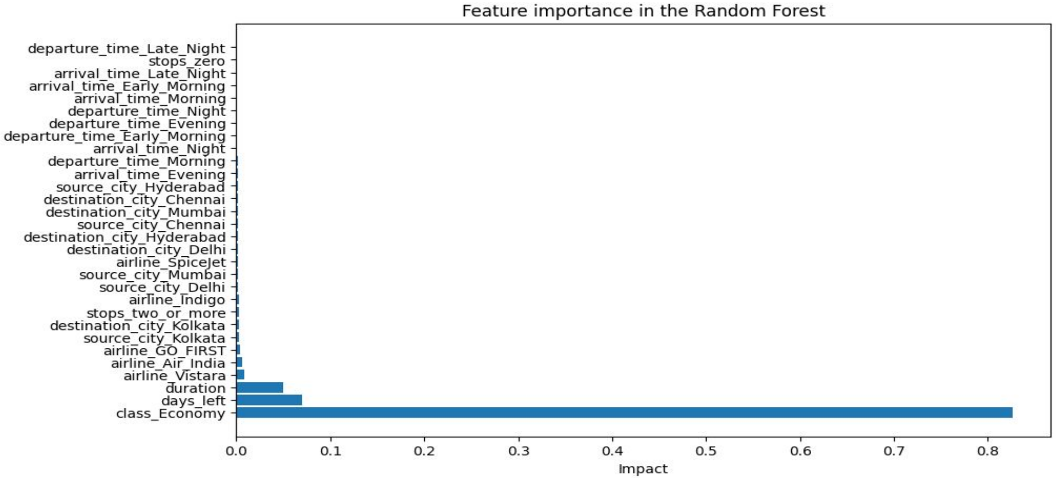
Random Forest - Complete Model

- Let us now try to set up a **Random Forest** model, using *RandomForesRegressor* from the sklearn library.
- The result we expect, compared to a Regression Tree, offers several **advantages**, mainly in terms of *accuracy*, *robustness* and *generalization* ability. A single Regression Tree tends to suffer from *overfitting*, especially if it is very deep, fitting too well to the training data and having difficulty generalizing to new data.
- A Random Forest, on the other hand, is an **ensemble** of many regression trees built on different subparts of the dataset and with random choices in the variables considered at each split.



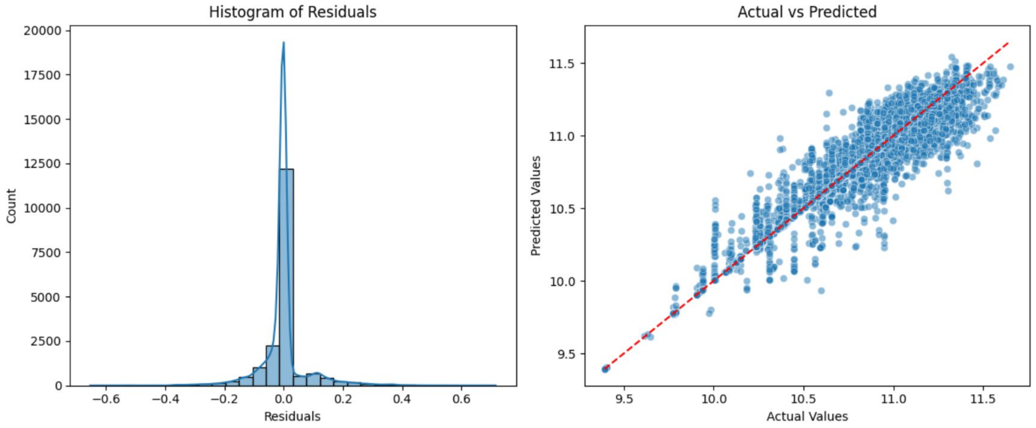


We notice a very strong impact of **class** on the model we set up.
At this point, we will use the two datasets for the distinct classes we derived earlier to conduct a more in-depth analysis.

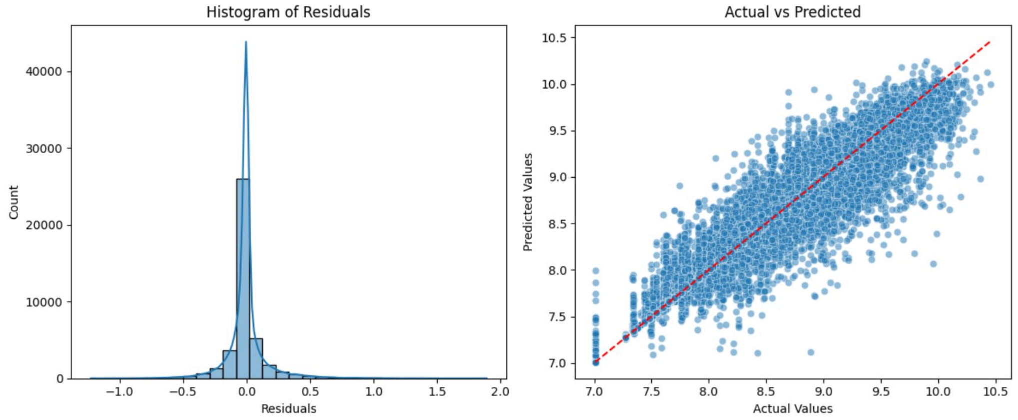


Regression Trees, XGBoost and Random Forest for Business and Economy

- Once the previous approach is completed, we replicate the results on the separate datasets (Business and Economy).
- As in the linear regression phase, we expect different results in terms of explained variance, diagnostic plots, and residuals.
- Next, we will examine some diagnostic plots based on the **Random Forest** approach for both datasets.



Diagnostic Plots - Random Forest - Business Dataset



Diagnostic Plots - Random Forest Economy Dataset

Results

Results for Linear Models and Variants

Model name	R^2 train	R^2 test	MAE	RMSE	MAPE	AIC	BIC
result_lr_complete	0.912	0.911	4583.92	6791.88	46.49%	4.91e06	4.91e06
result_lr_partial1	0.912	0.911	4584.58	6791.87	46.50%	4.91e06	4.91e06
result_lr_partial2	0.915	0.881	4605.23	7833.93	26.19%	1.401e05	1.401e05
result_pr_partial1	0.924	0.860	4716.45	8493.33	24.88%	1.152e05	1.152e05
ridge_regression	0.923	0.860	4716.35	8492.98	24.88%	-1.41e05	-1.41e05

Results for MARS

Model name	R^2 train	R^2 test	MAE	RMSE	MAPE	AIC	BIC
mars_model_allclasses	0.945	0.945	3269.51	5697.62	20.04%	9152.89	9296.93
mars_model_economy	0.729	0.737	1347.96	2119.80	20.94%	9439.12	9723.90
mars_model_business	0.731	0.721	5744.14	7867.81	11.16%	-18591.92	-18317.66

Results for Regression Trees & XGBoost and Random Forest

Model name	R^2	MAE	RMSE	MAPE
tree_reg	0.9760	1159.61	3518.97	7.25%
XGBoost	0.9880	1044.99	2493.25	6.65%
random_forest	0.9849	1057.18	2788.90	6.46%

Model name	R^2	MAE	RMSE	MAPE
tree_reg_eco	0.7548	684.08	1847.83	9.10%
XGBoost_eco	0.8613	601.85	1369.66	8.18%
random_forest_eco	0.8653	587.41	1369.50	7.77%

Model name	R^2	MAE	RMSE	MAPE
tree_reg_bus	0.8084	2264.54	5641.60	3.90%
XGBoost_bus	0.8743	2130.04	4569.41	3.74%
random_forest_bus	0.8779	2084.50	4503.21	3.60%

Conclusions

- In conclusion, our analysis of the dataset and the subsequent modeling efforts have led several **key insights**. The dataset, characterized by its division into two distinct **classes** influencing ticket prices, presented challenges such as numerous outliers and a predominance of categorical variables, with only **duration** and **days_left** being numeric.
- In terms of model quality, we explored various approaches to predict ticket **prices** across the two datasets. *Linear* and *Polynomial* Models, including those with *interactions*, were initially employed but were eventually supplemented by *MARS* and *Ridge Regression* to better manage **outliers**, **VIF**, **heteroscedasticity**, and **prediction accuracy**.
- These models effectively captured variance with relatively low complexity as indicated by **AIC** and **BIC** metrics. Ultimately, *Regression Trees* and *XGBoost* emerged as the top performers for this dataset.

- Throughout the analysis, we addressed several critical aspects, including **heteroscedasticity**, **normality** of **residuals**, **VIF** and **multicollinearity**, **non-linear features**, and the application of **logarithmic transformation** on price. Splitting the dataset into two **classes** mitigated **clustering** issues but at the cost of reduced prediction **accuracy**. **Overfitting** was controlled through **penalization** techniques and validation on test sets.
- Splitting the datasets reduced **clustering** problems but decreased prediction accuracy.
- **Overfitting** was managed using penalization and test set verification.
- The *economy* dataset had more observations, lower **prices**, and more outliers.
- The *business* dataset had higher **prices** but fewer observations.
- This **comprehensive** approach allowed us to develop robust **models** capable of handling the complexities inherent in the data, ultimately leading to more accurate and reliable price predictions.

Thanks for listening!