AIR FARE PREDICTION

1 Abstract

Flight price prediction is pivotal in providing customers with information on booking flight tickets. Flight aggregators such as Kayak or Hopper collect information from multiple websites and offer guidance on fares based on historical data. Depending on the circumstances, it can advise customers to buy a ticket immediately or wait for the right time to get the cheapest price. By using historical data, the travel sector can provide better customer service to foster royalty. This project aims at predicting flight prices based on historical data obtained from the Office of Airline Information of the Bureau of Transportation and Statistics from the reporting carriers. The project uses several machine learning techniques to predict flight prices with an R-squared score of 0.47.

2 Introduction

Airfares vary on a daily basis based on demand and supply. During holiday seasons, ticket prices tend to be high as people tend to travel more for vacation. Airfares consist of multiple components such as fuel charge, tax, operator fees, service fees, convenience fees and other factors. Additionally, some of these components also change with time, for example, based on crude oil price, the fuel price tends to change. Another example could be the days left to travel, as the travel date nears, the price increases. Tickets prices are increasingly driven by economic, marketing, and societal factors as well. Not only do airfare predictions help customers decide when to book tickets, but it can also help airline companies make critical business decisions and strategies in the competitive market. This project focuses primarily on using linear statistical techniques to predict airfares assuming a linear relationship is sufficient for the best prediction. It also uses other non-linear modelling techniques to compare the results of the different models.

3 Dataset

3.1 Data Origin

The DB1B dataset maintained by the Office of Airline Information of the Bureau of Transportation and Statistics from the Airline Origin and Destination Survey is a 10% random sample of airline tickets from reporting carriers which consist of domestic carriers. It has 25 years of data but for this project we will be using data only for one quarter, i.e., 2022 Q2. A single quarter has almost 3 million observations. The dataset has 3 tables Ticket, Coupon and Market. The Ticket table gives basic information on an itinerary such as reporting carrier, distance flown, origin airport, origin city. The Coupon table gives coupon specific information for each itinerary such as destination airport, coupon type, fare class and number of passengers. The market data holds summary characteristics of each itinerary. Most of the information are repeated in the 3 tables and all 3 tables have the 'ITIN_ID' (Itinerary ID) as the primary index.

We will use the Market table as our main source as it holds information at the itinerary level with price per passenger. We will obtain some relevant features from the Coupon and Ticket tables. In total, we have 22 features with 2.1M data points. Below is a summary of the features obtained from the 3 tables:

Table	Features
Market	Origin airport, origin city, destination airport, ticket carrier, operating carrier, passengers, distance and airfare
Ticket	Origin state, roundtrip flag, online flag, price credibility
Coupon	Coupon type, fare class and trip break

3.2 Data Cleaning

Null values are present in variables coming from the coupon table. Fare class is an important variable in determining the price of the ticket and data points with null Fare class was dropped. For coupon type and trip break, the null values were given a separate label instead of using imputing techniques. Some classes in 'Fare Class' variable were combined. Data points with price credibility = 0 are not reliable and hence were removed. 'ITIN_ID' was removed as it is a unique identifier. Other features such as 'Distance Group', 'Bulk Fare', 'Destination City', 'Origin City' were removed because there was already a similar field. The size of the dataset reduced to 1.3M data points and 13 features.

4 Feature Engineering

4.1 Treating Outliers

Some data points contained ticket prices less than $50. These data points were removed as it most likely resulted from a discount. Prices greater than 99th percentile were capped to the 99th percentile value. For distance and number of passengers, 1st percentile and 99th percentile was used to replace values below and above the respective points. Other numerical variables were normalized using the Min Max Scaler.

4.2 Handling categorical variables

Variables like 'Origin Airport', 'Origin State' and 'Operating Carrier' had very high cardinality. Quantile encoding and manual grouping were tested. For 'Operating Carrier', manual grouping into 3 buckets followed by quantile encoding helped in reducing model's loss. Grouping was done based on the mean airfares of each category. For 'Origin Airport' and 'Origin state' just quantile encoding was performed. For variables with binary categories such as 'Coupon Type', 'Trip break' and 'Market geography type', one hot encoding was performed.

4.3 Multicollinearity

Pearson's correlation was used to build a correlation matrix among all the predictor variables. The response variable was also included to check which of the predictor variables would have a good correlation and hence a good predictor. The correlation matrix is shown in the

Appendix. The 'Origin State' was removed because it was correlated with the 'Origin Airport'. Also, 'Distance' and 'Fare Class' seemed to be correlated well with 'Airfare' when compared to other variables.

5 Experiment and Analysis

The aim here is to predict the 'Airfare' or 'Ticket Prices' and this is a regression problem. The datapoints were split into training (80%) and test sets (20%). An unbiased linear model (OLS) is the main focus of this project. Other non-linear models such as Random forests, XGBoost and Neural Networks were also tested. The dataset contained 11 features and 1.2M data points. The metric used to compare different models would be adjusted R-squared and Root Mean Squared Error (RMSE).

5.1 Ordinary Least Squares (OLS)

The model assumes the below linear relationship between the response and predictor variables

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

The model yielded an adjusted R-squared value of 0.37. The RMSE of the training set was 153.03 and that of the test set was 153.04. All the variables turned out to be significant at 95% confidence interval. This is possible in large datasets where $n >> p$ and small effects can be found 'significant'. The Durbin-Watson test produced a value of 2. A value greater than 3 would indicate negative autocorrelation in the residuals.

In order to reliably use our results, we need to check for normality and skewness of the residuals. The kurtosis value between +3 and -3 are acceptable and the model produced a value of 5.5. This could indicate possible outliers in the data that we have not treated.

5.1.1 Cook's Distance

We can use the Cook's distance to identify outlying predictor observations in the data. The Cook's distance is given by:

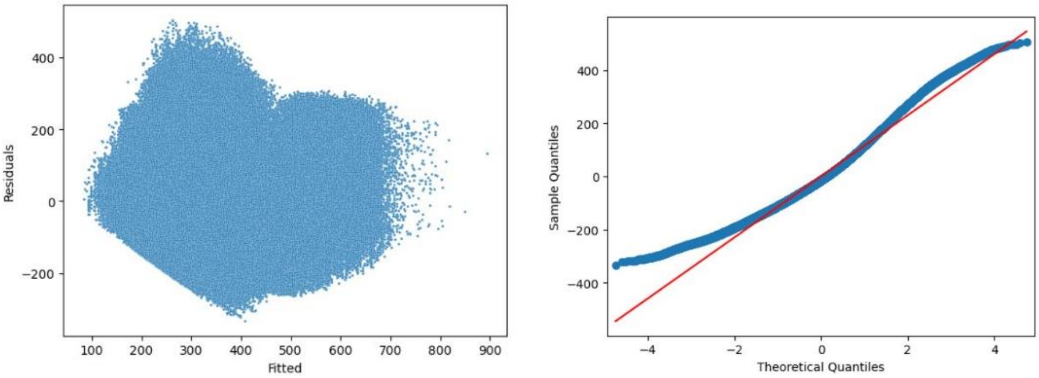
$$D_i = \frac{e_i^2}{pMSE}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right]$$

Where D_i is the Cook's distance for the i^{th} data point. A value greater than 3 times the mean value of the Cook's distances indicates that it is an outlying observation. 91K outlying observations were removed using this method. The kurtosis value reduced to 3.2 which a little past the borderline. The model was fit again and the adjusted R-squared value improved to 0.4. The RMSE of the train and test sets reduced to 115.06 and 114.98 respectively.

5.1.2 Studentized Deleted Residuals

When outlying observations influence the regression, the model can be pulled towards the potential outlier and this could pose a problem. In order to identify outlying response observations, studentized deleted residuals are used. The idea is to delete observations one at a time and refit the regression model on the remaining data points. Then the observed and fitted values are compared to get the studentized deleted residuals. If the value is greater than 3, it can be identified as an outlying observation. 6321 outlying observations were found and were removed before fitting the model again. No noticeable improvements in the model were observed.

5.1.3 Residuals vs Fitted



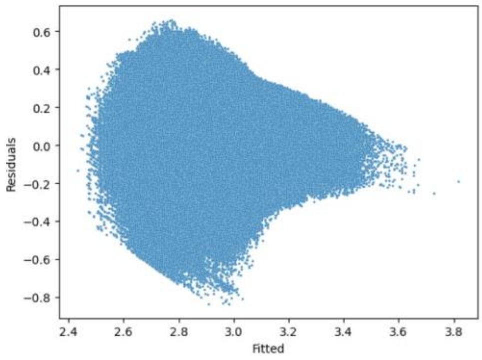
The residuals seem to bounce randomly around 0 for fitted values >400. For fitted values <400, there seems to be some decrease in variance which suggests some heteroscedasticity. Box-cox transformation or weighted least squares approach are a couple of techniques that can be used to counter this. The QQ plot shown in the figure suggests that there is a positive skew in the residuals.

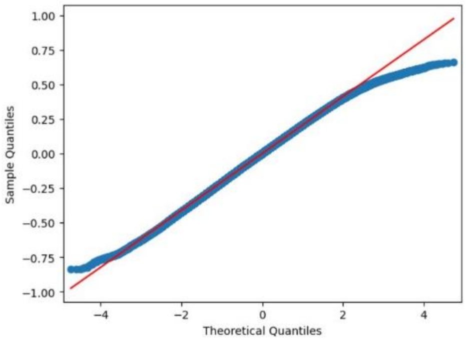
5.1.4 Box Cox Transformation

The constant variance assumption is important in OLS as it ensures that the model has minimum variance. If not, the standard errors of the estimators would not be accurate. The Box-cox transformation can be used to reduce the variance in the error terms. It is used on the response variable to transform it based on the below equation:

$$y(\lambda) = \begin{cases} y^{\lambda} - 1, & if\ \lambda \neq 0 \\ \log(y), & if\ \lambda = 0 \end{cases}$$

The residuals vs fitted plot is shown below after the box-cox transformation. The change in variance of the residuals is more pronounced. This indicates that we might have to look for other alternatives. The QQ plot shown in the figure suggests that there is a negative skew in the residuals.





5.1.4 VIF

The variable inflation factor measures the collinearity between the predictor variables. The correlation matrix can help in detecting collinearity between 2 predictors but VIF can detect multicollinearity between multiple variables. When collinearity exists between predictors, it becomes difficult for the model to decide the right coefficients for the correlated variables. The VIF table is shown in appendix. High VIF was detected for the column 'Coupon Type' and hence it was removed. Removing the variable did not change the results but it reduced the VIF for other correlated variables.

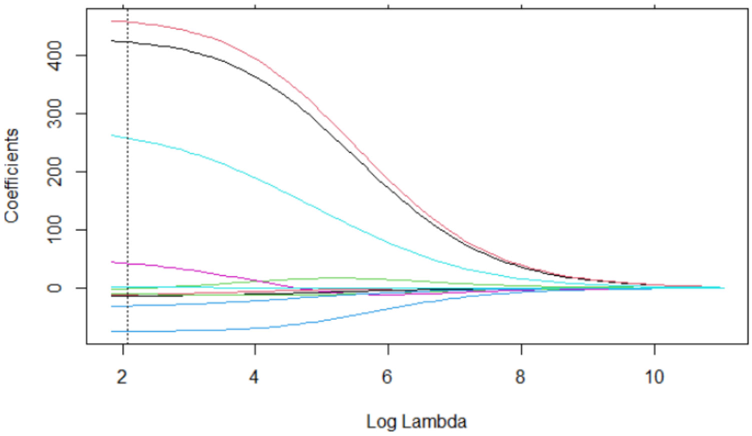
5.1.5 F Test for regression relation

The F-test was used to test the regression relation between the response variable and the predictor variables. The F-test value is 55309 and it is greater than the F-critical value of 1.78. This indicates regression relation is adequate.

5.1.6 Ridge Regression

Ridge regression model produces predictions with lesser variance and can help with multicollinearity in the variables. The parameters of the Ridge regression are estimated by optimizing the below equation:

$$\hat{\beta}^{ridge} = argmin_{\beta\in R}\left|\left|Y - X\beta\right|\right|_2^2 + \lambda\left|\left|\beta\right|\right|_2^2$$



The ridge regression model was tested for multiple values of lambda and the best value of lambda produced was 0 which is equivalent to OLS. When we have n>>p, and the underlying relationship is actually linear, OLS performs well. Even though VIF values are a little high for some variables (shown in appendix), Ridge regression is not helping here. Lasso regression was also tested but it did not help in improving the results.

5.2 Random Forest

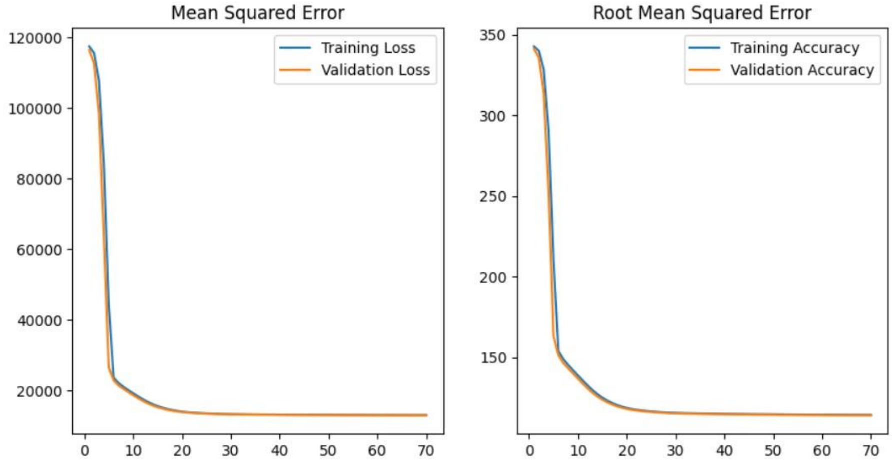
Random Forests is an ensemble learning method that uses multiple decision trees. It can handle all types of features and requires very little pre-processing. Hyperparameter tuning was performed to get optimal number of trees, criteria for information gain, depth and minimum samples in the leaf node. The final model had a training RMSE of 111.584 and test RMSE of 111.587 The adjusted R-squared value of the model was 0.43. It produced slightly better results when compared to OLS.

5.3 XGBoost

Extreme Gradient Boosting or XGBoost is a boosting algorithm where several optimization techniques are combined to get results within a short span of time. The objective is to minimize the loss function of the model by adding weak learners using gradient descent. Hyperparameter tuning was performed to get optimal gradient boosted trees, learning rate and depth. The model had a training RMSE of 107.94 and test RMSE of 109.22. The adjusted R-squared value of the model was 0.47. It produced even better results than Random Forests.

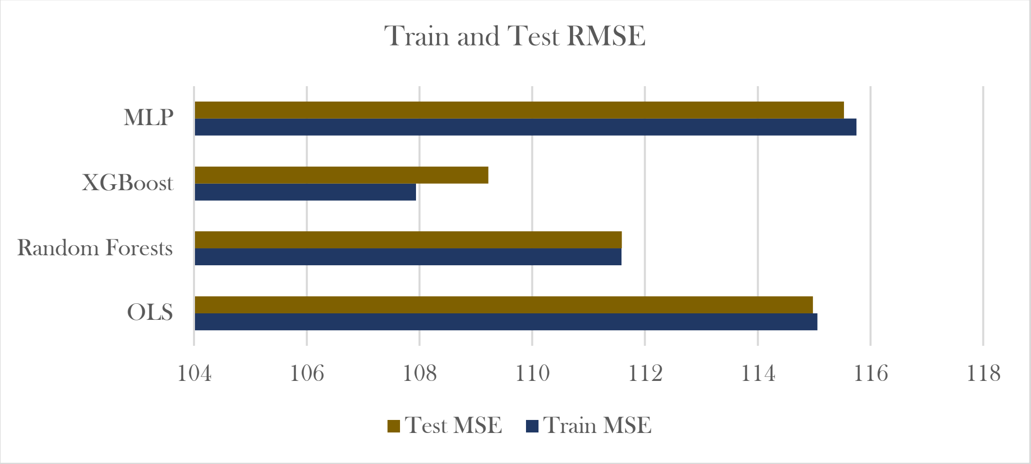
5.4 Multi-layer Perceptron

A Multilayer Perceptron model was implemented with 4 layers. Kernel L2 regularization with penalty of 0.01 was used. The activation function used was 'ReLU' for intermediate layers. The model was compiled using 'Mean Squared Error' loss function and 'Adam' optimiser with a learning rate of 0.001. 20 Epochs was ideal for the model to reach a minimum loss value obtained from the loss graph below. The model had a training RMSE of 118.17 and testing RMSE of 117.96. The adjusted R-squared value of the model was 0.36.



6 Conclusion

The RMSE values are the lowest for XGBoost followed by Random Forests.



The adjusted R-squared value measures the proportion of explained variance in the response variable using the predictor variables. The best adjusted R-squared value achieved was 0.47 by XGBoost. The adjusted R-squared value is not adequate because there might be some other factors that could be affecting the variation in prices. Airfares are typically cheaper when booked in advance and they increase a lot when the departure date nears. This 'days until journey when booking' information is not available in the dataset. Some macroeconomic factors also influence flight ticket prices like crude oil prices for which we also need the departure date for each itinerary (not available in the dataset). These crucial pieces of information could make up for some of the missing explained variance in the models.

7 References

Kutner et al. M. C. K. N. (2022). Applied Linear Statistical Models 5th Edition (5th Edition).

Wang, Tianyi & Pouyanfar, Samira & Tian, Haiman & Tao, Yudong & Alonso, Miguel & Luis, Steven & Chen, Shu-Ching. (2019). A Framework for Airfare Price Prediction: A Machine Learning Approach. 200-207. 10.1109/IRI.2019.00041

R. R. Subramanian, M. S. Murali, B. Deepak, P. Deepak, H. N. Reddy and R. R. Sudharsan, "Airline Fare Prediction Using Machine Learning Algorithms," 2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT), 2022, pp. 877-884, doi: 10.1109/ICSSIT53264.2022.9716563

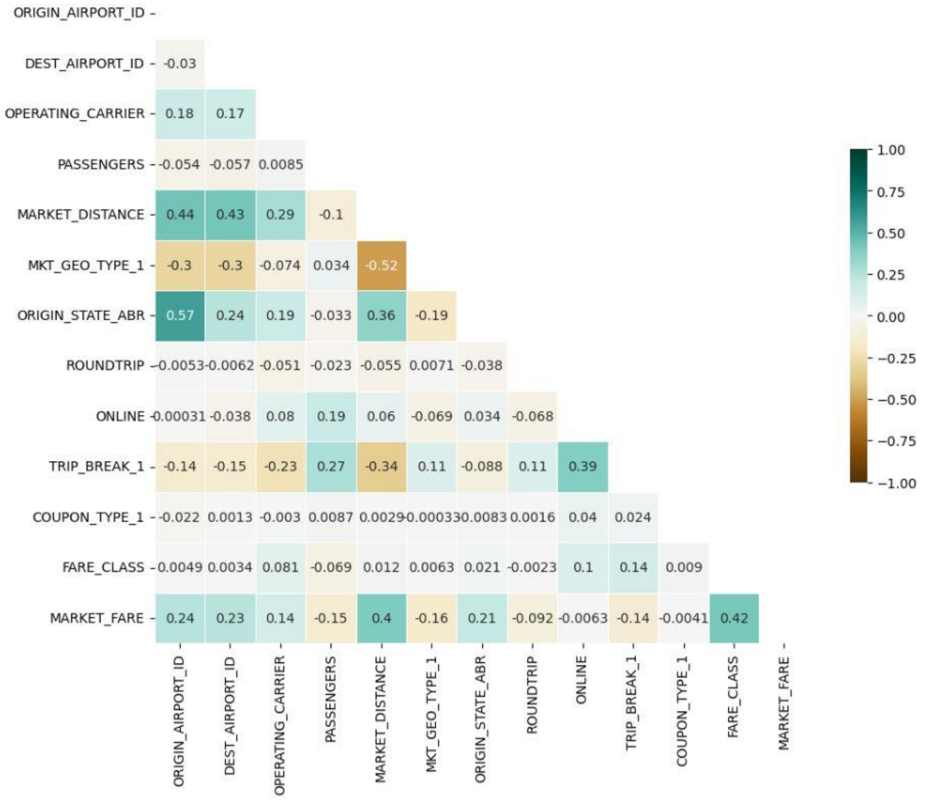
Hale, J. (2020, October 5). Smarter Ways to Encode Categorical Data for Machine Learning. Medium. https://towardsdatascience.com/smarter-ways-to-encode-categorical-data-for-machine-learning-part-1-of-3-6dca2f71b159

Kézdi, G. B. A. G. (n.d.). Datasets summary. Gabors Data Analysis. https://gabors-data-analysis.com/datasets/

Taylor, J. (2009). statsmodels Package in R. https://www.statsmodels.org

8 Appendix

8.1 Correlation matrix

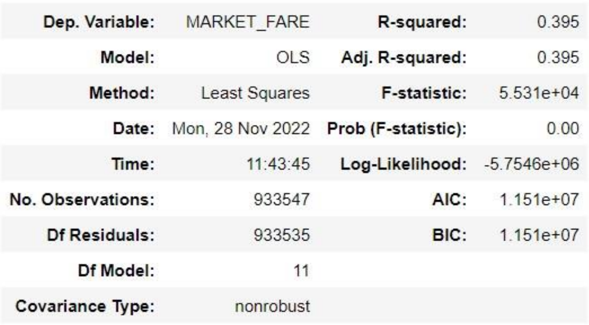


	ORIGIN_AIRPORT_ID	DEST_AIRPORT_ID	OPERATING_CARRIER	PASSENGERS	MARKET_DISTANCE	MKT_GEO_TYPE_1	ORIGIN_STATE_ABR	ROUNDTRIP	ONLINE	TRIP_BREAK_1	COUPON_TYPE_1	FARE_CLASS	MARKET_FARE
ORIGIN_AIRPORT_ID													
DEST_AIRPORT_ID	-0.03												
OPERATING_CARRIER	0.18	0.17											
PASSENGERS	-0.054	-0.057	0.0085										
MARKET_DISTANCE	0.44	0.43	0.29	-0.1									
MKT_GEO_TYPE_1	-0.3	-0.3	-0.074	0.034	-0.52								
ORIGIN_STATE_ABR	0.57	0.24	0.19	-0.033	0.36	-0.19							
ROUNDTRIP	-0.0053	-0.0062	-0.051	-0.023	-0.055	0.0071	-0.038						
ONLINE	-0.00031	-0.038	0.08	0.19	0.06	-0.069	0.034	-0.068					
TRIP_BREAK_1	-0.14	-0.15	-0.23	0.27	-0.34	0.11	-0.088	0.11	0.39				
COUPON_TYPE_1	-0.022	0.0013	-0.003	0.0087	0.0029	-0.00033	0.0083	0.0016	0.04	0.024			
FARE_CLASS	0.0049	0.0034	0.081	-0.069	0.012	0.0063	0.021	-0.0023	0.1	0.14	0.009		
MARKET_FARE	0.24	0.23	0.14	-0.15	0.4	-0.16	0.21	-0.092	-0.0063	-0.14	-0.0041	0.42	

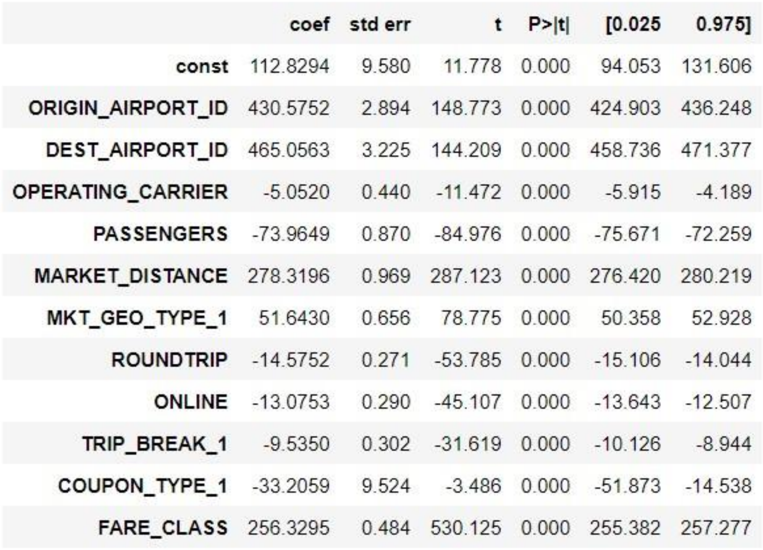
8.2 Variance Inflation Factor (VIF)

Feature	Before Removing Coupon Type	After removing Coupon Type
COUPON_TYPE_1	76.13	
DEST_AIRPORT_ID	14.40	11.36
FARE_CLASS	1.12	1.12
MARKET_DISTANCE	6.35	6.31
MKT_GEO_TYPE_1	28.79	12.93
ONLINE	3.95	3.88
OPERATING_CARRIER	12.35	11.79
ORIGIN_AIRPORT_ID	13.19	10.72
PASSENGERS	1.24	1.24
ROUNDTRIP	3.75	3.63
TRIP_BREAK_1	3.02	2.95

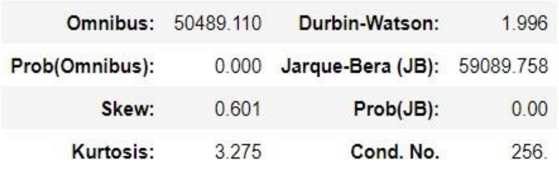
8.3 OLS Summary:



Dep. Variable:	MARKET_FARE	**R-squared:**	0.395
Model:	OLS	**Adj. R-squared:**	0.395
Method:	Least Squares	**F-statistic:**	5.531e+04
Date:	Mon, 28 Nov 2022	**Prob (F-statistic):**	0.00
Time:	11:43:45	**Log-Likelihood:**	-5.7546e+06
No. Observations:	933547	**AIC:**	1.151e+07
Df Residuals:	933535	**BIC:**	1.151e+07
Df Model:	11		
Covariance Type:	nonrobust		

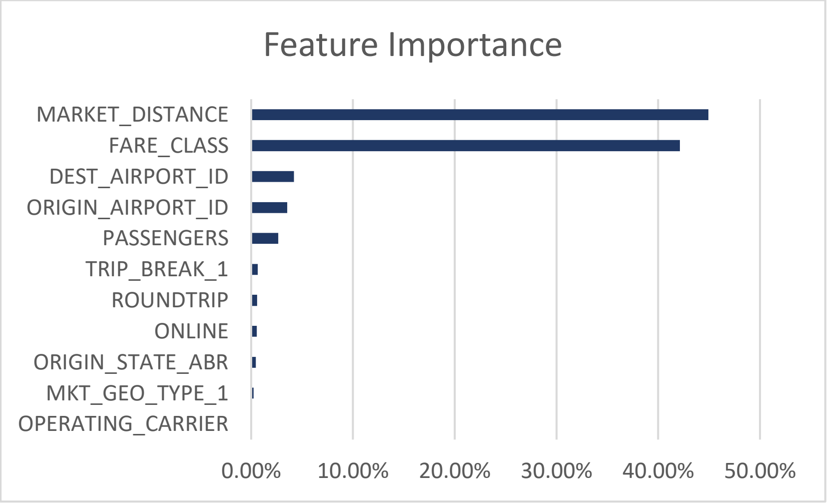


	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	112.8294	9.580	11.778	0.000	94.053	131.606
ORIGIN_AIRPORT_ID	430.5752	2.894	148.773	0.000	424.903	436.248
DEST_AIRPORT_ID	465.0563	3.225	144.209	0.000	458.736	471.377
OPERATING_CARRIER	-5.0520	0.440	-11.472	0.000	-5.915	-4.189
PASSENGERS	-73.9649	0.870	-84.976	0.000	-75.671	-72.259
MARKET_DISTANCE	278.3196	0.969	287.123	0.000	276.420	280.219
MKT_GEO_TYPE_1	51.6430	0.656	78.775	0.000	50.358	52.928
ROUNDTRIP	-14.5752	0.271	-53.785	0.000	-15.106	-14.044
ONLINE	-13.0753	0.290	-45.107	0.000	-13.643	-12.507
TRIP_BREAK_1	-9.5350	0.302	-31.619	0.000	-10.126	-8.944
COUPON_TYPE_1	-33.2059	9.524	-3.486	0.000	-51.873	-14.538
FARE_CLASS	256.3295	0.484	530.125	0.000	255.382	257.277



Omnibus:	50489.110	**Durbin-Watson:**	1.996
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	59089.758
Skew:	0.601	**Prob(JB):**	0.00
Kurtosis:	3.275	**Cond. No.**	256.

Variable importance from Random Forests:


Feature Importance
MARKET_DISTANCE
FARE_CLASS
DEST_AIRPORT_ID
ORIGIN_AIRPORT_ID
PASSENGERS
TRIP_BREAK_1
ROUNDTRIP
ONLINE
ORIGIN_STATE_ABR
MKT_GEO_TYPE_1
OPERATING_CARRIER