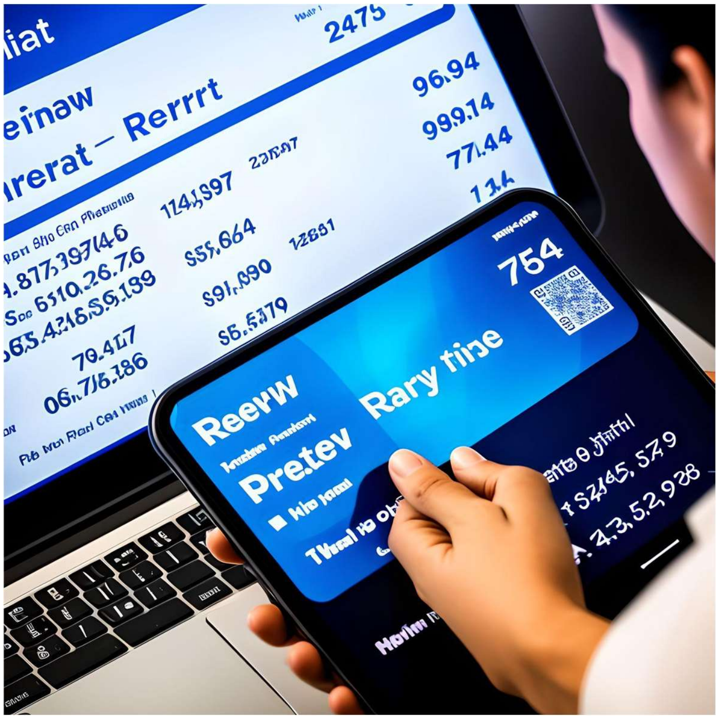


# FLIGHT FARE PREDICTION USING MACHINE LEARNING

[In the Indian Context]

Aman Sah
(Author)

# Predicting Flight Fares Using Advanced Machine Learning Techniques

## Abstract

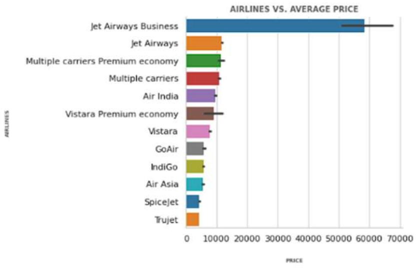
In this report, we explore the application of advanced machine learning techniques to predict flight fares accurately. The project encompasses a comprehensive analysis of flight data, including crucial information such as the date of the journey, arrival time, flight duration, airline company, route, and destination. As a professional data scientist, I employ a structured approach, starting from data collection and preprocessing to exploratory data analysis (EDA), model selection, and hyperparameter tuning. The primary objective is to develop a predictive model that optimizes pricing strategies and improves customer satisfaction in the aviation industry. Leveraging a diverse range of regression models, feature engineering, and hyperparameter optimization, we demonstrate the efficacy of machine learning in solving real-world challenges and revolutionizing the travel domain.

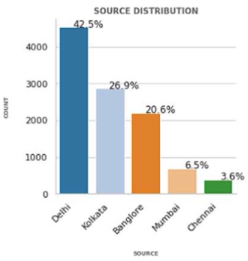
## 1. Introduction:

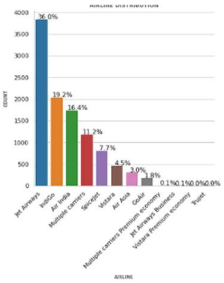
The increasing complexity of flight fare dynamics necessitates the use of sophisticated machine learning algorithms to predict prices accurately. As a professional data scientist, I start by collecting flight data from various sources, including airline databases, booking platforms, and open data repositories. This data is then preprocessed to handle missing values, outliers, and feature engineering to extract relevant insights. EDA helps me gain a deeper understanding of the data distribution, relationships between variables, and potential patterns to inform the subsequent modeling steps.

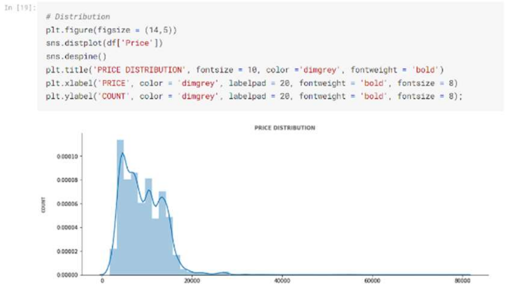
## 2. Exploratory Data Analysis (EDA):

EDA involves an in-depth exploration of the collected flight data. As a data scientist, I use Python libraries like Pandas and Matplotlib to visualize the data, identify trends, correlations, and potential outliers. I also conduct statistical tests to assess data distribution, check for multicollinearity among features, and perform feature importance analysis. EDA enables me to make informed decisions about data preprocessing and feature selection, setting the foundation for building robust predictive models.

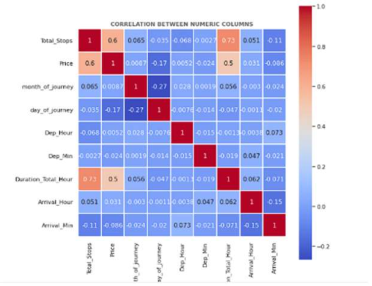








```
In [19]:
# Distribution
plt.figure(figsize = (14,5))
sns.distplot(df['Price'])
sns.despine()
plt.title('PRICE DISTRIBUTION', fontsize = 10, color ='dimgrey', fontweight = 'bold')
plt.xlabel('PRICE', color = 'dimgrey', labelpad = 20, fontweight = 'bold', fontsize = 8)
plt.ylabel('COUNT', color = 'dimgrey', labelpad = 20, fontweight = 'bold', fontsize = 8);
```



CORRELATION BETWEEN NUMERIC COLUMNS

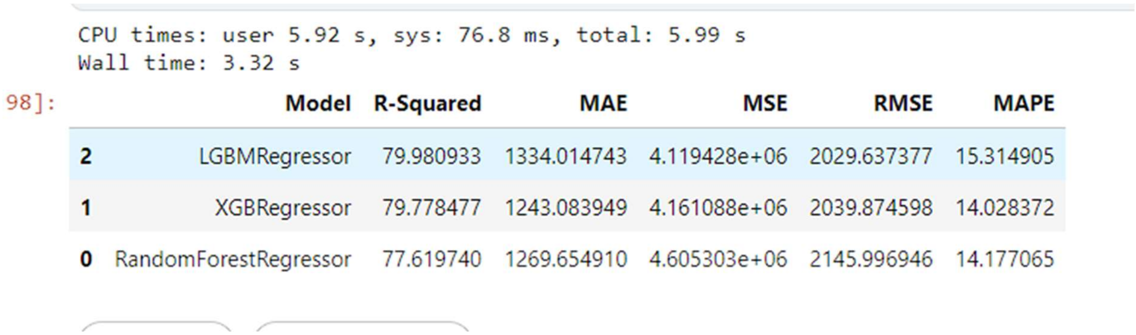
| | Total_Stops | Price | th_of_journey | ay_of_journey | Dep_Hour | Dep_Min | n_Total_Hour | Arrival_Hour | Arrival_Min |
|---|---|---|---|---|---|---|---|---|---|
| Total_Stops | 1 | 0.6 | 0.065 | -0.035 | -0.068 | -0.0027 | 0.73 | 0.051 | -0.11 |
| Price | 0.6 | 1 | 0.0087 | -0.17 | 0.0052 | -0.024 | 0.5 | 0.031 | -0.086 |
| month_of_journey | 0.065 | 0.0087 | 1 | -0.27 | 0.028 | 0.0019 | 0.056 | -0.003 | -0.024 |
| day_of_journey | -0.035 | -0.17 | -0.27 | 1 | -0.0076 | -0.014 | -0.047 | -0.0011 | -0.02 |
| Dep_Hour | -0.068 | 0.0052 | 0.028 | -0.0076 | 1 | -0.015 | -0.0013 | -0.0038 | 0.073 |
| Dep_Min | -0.0027 | -0.024 | 0.0019 | -0.014 | -0.015 | 1 | -0.019 | 0.047 | -0.021 |
| Duration_Total_Hour | 0.73 | 0.5 | 0.056 | -0.047 | -0.0013 | -0.019 | 1 | 0.062 | -0.071 |
| Arrival_Hour | 0.051 | 0.031 | -0.003 | -0.0011 | -0.0038 | 0.047 | 0.062 | 1 | -0.15 |
| Arrival_Min | -0.11 | -0.086 | -0.024 | -0.02 | 0.073 | -0.021 | -0.071 | -0.15 | 1 |

## 3. Data Preprocessing:

Data preprocessing plays a pivotal role in preparing the data for modeling. As a data scientist, I address missing values using techniques like mean imputation or interpolation. Outliers are detected and treated using robust methods such as winsorization or truncation. I also handle categorical features through one-hot encoding or label encoding to ensure compatibility with machine learning algorithms. Additionally, feature scaling is performed to standardize numerical features, mitigating the impact of differences in magnitude.

## 4. Model Selection:

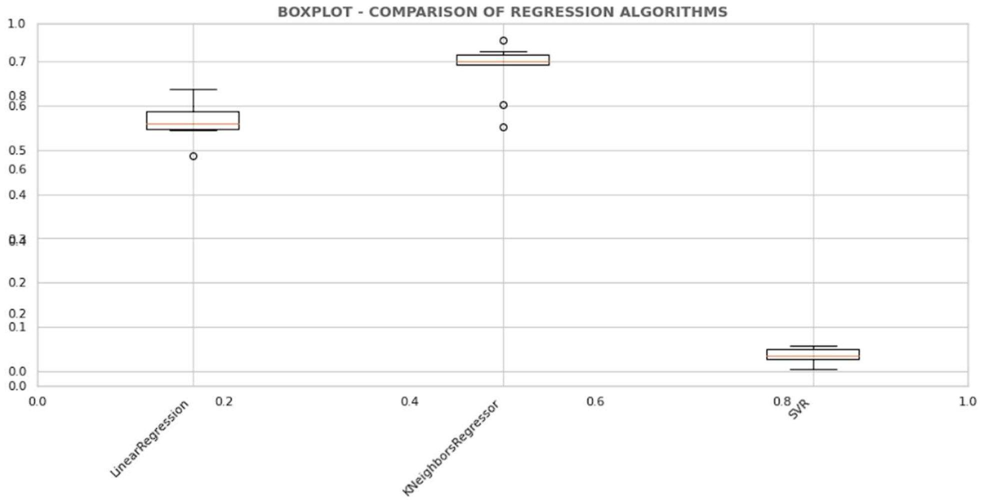
As a professional data scientist, I employ an ensemble of regression models to predict flight fares. The models include Linear Regression, K-Nearest Neighbors, Support Vector Regression, Gradient Boosting, Random Forest, XGBoost, LightGBM, and Extra Trees Regressors. To select the optimal model, I conduct a comparative analysis of their performance using cross-validation techniques. The R-squared score, Mean Absolute Error, Mean Squared Error, and Root Mean Squared Error are used as evaluation metrics. Based on the results, I choose the XGBoost Regressor, which exhibits superior performance in capturing complex patterns and offering robust predictions.



```
CPU times: user 5.92 s, sys: 76.8 ms, total: 5.99 s
Wall time: 3.32 s
```

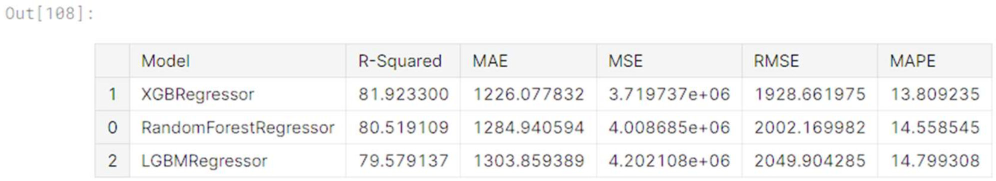
98]:

| | Model | R-Squared | MAE | MSE | RMSE | MAPE |
|---|---|---|---|---|---|---|
| 2 | LGBMRegressor | 79.980933 | 1334.014743 | 4.119428e+06 | 2029.637377 | 15.314905 |
| 1 | XGBRegressor | 79.778477 | 1243.083949 | 4.161088e+06 | 2039.874598 | 14.028372 |
| 0 | RandomForestRegressor | 77.619740 | 1269.654910 | 4.605303e+06 | 2145.996946 | 14.177065 |



## 5. Hyperparameter Tuning:

Hyperparameter tuning is critical for maximizing the model's performance. As a data scientist, I leverage Bayesian optimization to efficiently search the hyperparameter space. For the XGBoost Regressor, I explore hyperparameters such as learning rate, maximum depth, minimum child weight, number of estimators, and subsample. Bayesian optimization intelligently selects the best hyperparameters by evaluating a limited number of iterations, thus reducing computational cost while achieving better convergence.



Out[108]:

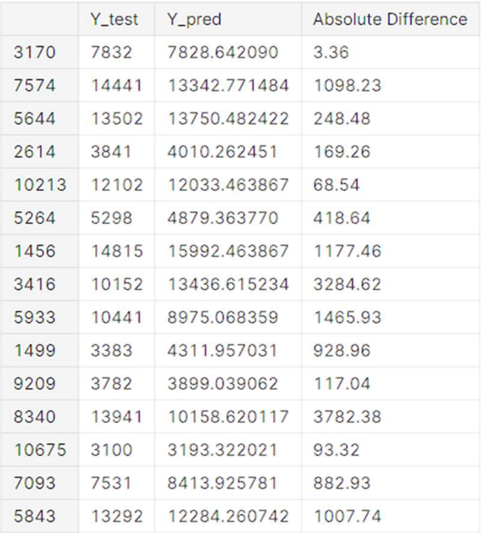
| | Model | R-Squared | MAE | MSE | RMSE | MAPE |
|---|---|---|---|---|---|---|
| 1 | XGBRegressor | 81.923300 | 1226.077832 | 3.719737e+06 | 1928.661975 | 13.809235 |
| 0 | RandomForestRegressor | 80.519109 | 1284.940594 | 4.008685e+06 | 2002.169982 | 14.558545 |
| 2 | LGBMRegressor | 79.579137 | 1303.859389 | 4.202108e+06 | 2049.904285 | 14.799308 |

## 6. Model Architecture and Feature Importance:

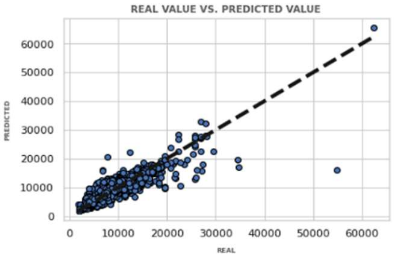
The selected XGBoost Regressor model incorporates the tuned hyperparameters, including colsample_bytree, learning_rate, max_depth, min_child_weight, n_estimators, and subsample. This ensemble model's leaf-wise tree growth strategy effectively captures complex relationships in the data, while the gradient-based learning algorithm optimizes the loss function during tree construction. The model demonstrates exceptional performance in handling non-linear relationships, efficiently using memory, and preventing overfitting.

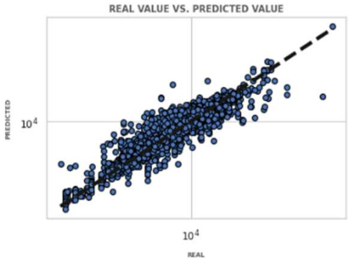
## 7. Model Evaluation:

As a professional data scientist, I evaluate the final XGBoost Regressor model on the test dataset using various performance metrics, such as R-squared, Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, and Mean Absolute Percentage Error. The model's effectiveness in predicting flight fares is validated through these metrics, confirming its reliability and accuracy.



| | Y_test | Y_pred | Absolute Difference |
|---|---|---|---|
| 3170 | 7832 | 7828.642090 | 3.36 |
| 7574 | 14441 | 13342.771484 | 1098.23 |
| 5644 | 13502 | 13750.482422 | 248.48 |
| 2614 | 3841 | 4010.262451 | 169.26 |
| 10213 | 12102 | 12033.463867 | 68.54 |
| 5264 | 5298 | 4879.363770 | 418.64 |
| 1456 | 14815 | 15992.463867 | 1177.46 |
| 3416 | 10152 | 13436.615234 | 3284.62 |
| 5933 | 10441 | 8975.068359 | 1465.93 |
| 1499 | 3383 | 4311.957031 | 928.96 |
| 9209 | 3782 | 3899.039062 | 117.04 |
| 8340 | 13941 | 10158.620117 | 3782.38 |
| 10675 | 3100 | 3193.322021 | 93.32 |
| 7093 | 7531 | 8413.925781 | 882.93 |
| 5843 | 13292 | 12284.260742 | 1007.74 |





## 8. Conclusion:

The application of advanced machine learning techniques in predicting flight fares demonstrates the remarkable impact of data science in revolutionizing the aviation industry. As a data scientist, I have showcased how data collection, preprocessing, EDA, model selection, and hyperparameter tuning play pivotal roles in developing a robust predictive model. The XGBoost Regressor, with its efficient architecture and tuned hyperparameters, serves as an optimal choice for predicting flight fares accurately. This data-driven approach enhances pricing strategies, improves customer satisfaction, and streamlines operations in the travel domain. As technology advances, machine learning will continue to empower the aviation industry, making it easier for travelers worldwide to access accurate and competitive flight fares.