# ✈️ Flight Price Prediction Using AWS SageMaker

## Overview

This project demonstrates an end-to-end machine learning workflow for predicting flight prices using Amazon SageMaker. It encompasses data preprocessing, exploratory data analysis (EDA), feature engineering, model training, and deployment.

---

## 1. Why Use Amazon SageMaker?

Amazon SageMaker is a fully managed service that provides every developer and data scientist with the ability to build, train, and deploy machine learning models quickly. The reasons for choosing SageMaker in this project include:-

- **Integrated Jupyter Notebooks:** Facilitates easy data exploration and preprocessing without managing servers.

- **Built-in Algorithms and Frameworks:** Supports popular machine learning algorithms and frameworks, streamlining the model development process.

- **Scalable Model Training:** Automatically manages the infrastructure required for training models, allowing for efficient scaling.

- **Model Deployment:** Simplifies the deployment of trained models into a production-ready hosted environment.

- **Cost-Effective:** Offers pay-as-you-go pricing, ensuring cost efficiency during experimentation and deployment phases.

---

## 2. Libraries Used

The project utilizes several Python libraries for data manipulation, visualization, and model development:

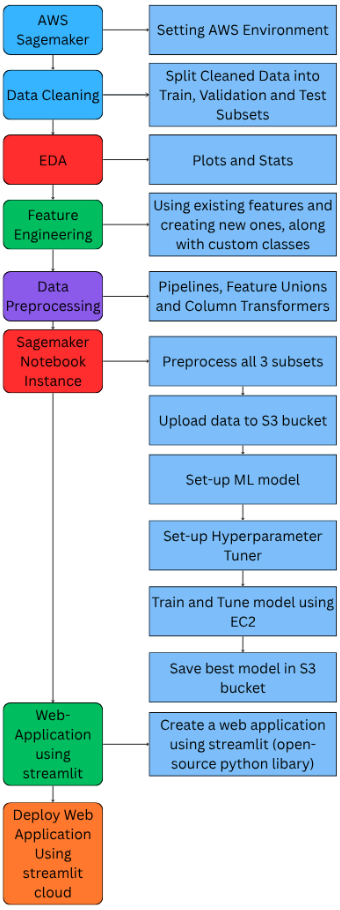
- **Pandas:** For data manipulation and analysis.

- **NumPy:** For numerical computations.
- **Matplotlib & Seaborn:** For data visualization.
- **Scikit-learn:** For machine learning algorithms and evaluation metrics.
- **XGBoost:** For implementing the XGBoost regression model.
- **Joblib:** For model serialization.
- **Streamlit:** For building the interactive web application interface.
- **Boto3:** AWS SDK for Python to interact with AWS services.

---

# 3. Project Workflow

The project follows a structured workflow divided as follows:-


AWS Sagemaker
Setting AWS Environment
Data Cleaning
Split Cleaned Data into Train, Validation and Test Subsets
EDA
Plots and Stats
Feature Engineering
Using existing features and creating new ones, along with custom classes
Data Preprocessing
Pipelines, Feature Unions and Column Transformers
Sagemaker Notebook Instance
Preprocess all 3 subsets
Upload data to S3 bucket
Set-up ML model
Set-up Hyperparameter Tuner
Train and Tune model using EC2
Save best model in S3 bucket
Web-Application using streamlit
Create a web application using streamlit (open-source python libary)
Deploy Web Application Using streamlit cloud

### 1) Introduction to AWS SageMaker

- Setting up the AWS environment and SageMaker instance.

### 2) GitHub Setup

- Setting up local and remote repositories using GitHub.

### 3) Data Cleaning

- Data cleaning using Pandas and NumPy.

### 4) Exploratory Data Analysis (EDA)

- Analyzing datasets to understand underlying patterns.
- Utilizing various plots and statistical measures for data analysis.
- Deep dive into data visualization and hypothesis testing.
- Identifying correlations and significant features affecting flight prices.

### 5) Feature Engineering

- Transforming raw data into meaningful features.
- Handling categorical variables, date-time features, and encoding techniques.

### 6) Model Training and Deployment

- Training the XGBoost regression model using SageMaker.
- Evaluating model performance using appropriate metrics.
- Deploying the trained model and building a Streamlit web application for Predictions

---

## 4. Additional Details

### Data Source

The dataset used contains information about various flight details, including airline, source, destination, departure and arrival times, duration, total stops, and price.

### Model Selection

The XGBoost regression model was chosen due to its efficiency and superior performance in handling structured data.

### Model Evaluation

The model's performance was evaluated using metrics such as Root Mean Squared Error (RMSE) and R-squared ($R^2$) to ensure accuracy and reliability.

### Web Application

A user-friendly web application was developed using Streamlit, allowing users to input flight details and receive predicted prices in real-time.

---

## 5. Conclusion

This project serves as a comprehensive guide for implementing a machine learning model for flight price prediction using AWS SageMaker. It covers the entire pipeline from data preprocessing to model deployment, providing valuable insights and practical experience in building scalable machine learning solutions on the cloud.