A logo of a company

Description automatically generated A close-up of a logo

Description automatically generated

**Joint Tech Internship Community Program**

### **Assignment: Predicting Flight Ticket Prices**

**Problem Statement:**

An online travel agency wants to predict the price of flight tickets based on various factors such as departure date, distance, airline, and class of service. Your task is to develop a machine learning model that predicts the price of a flight ticket given these attributes.

**Objective:**

Build a predictive model to estimate the price of flight tickets. Evaluate the model using appropriate metrics and provide insights into the factors that most influence ticket pricing.

**Dataset:**

You are provided with a dataset containing the following columns:

1. **FlightID**: Unique identifier for each flight.
2. **Airline**: The airline operating the flight (e.g., Delta, United, Southwest).
3. **DepartureAirport**: The airport from which the flight departs.
4. **ArrivalAirport**: The airport at which the flight arrives.
5. **DepartureDate**: The date of departure.
6. **Distance**: The distance between the departure and arrival airports (in miles).
7. **FlightDuration**: The duration of the flight (in hours).
8. **Class**: The class of service (e.g., Economy, Business, First).
9. **NumberOfStops**: The number of stops during the flight (e.g., Non-stop, 1 stop, 2+ stops).
10. **DaysUntilDeparture**: The number of days between booking and departure.
11. **DayOfWeek**: The day of the week the flight departs (e.g., Monday, Tuesday).
12. **TicketPrice**: The target variable representing the price of the flight ticket (in USD).

**Tasks:**

1. **Data Exploration and Preprocessing:**
   * Load the dataset and perform initial exploration to understand the data.
   * Identify and handle any missing values appropriately.
   * Convert categorical variables into numerical ones using techniques such as One-Hot Encoding.
2. **Feature Engineering:**
   * Perform feature scaling (e.g., Standardization or Normalization) on continuous variables.
   * Create new features if relevant, such as price per mile or duration per mile.
   * Use dimensionality reduction techniques (e.g., PCA) to reduce the feature space if necessary.
3. **Model Building:**
   * Split the dataset into training and testing sets (e.g., 80-20 split).
   * Train different regression models (e.g., Linear Regression, Decision Trees, Random Forest, Gradient Boosting).
   * Perform hyperparameter tuning using techniques like Grid Search or Random Search to optimize model performance.
4. **Model Evaluation:**
   * Evaluate your models using appropriate metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared.
   * Compare the performance of different models and select the best one.
   * Analyze feature importance to understand the most significant factors contributing to ticket pricing.
5. **Insights and Recommendations:**
   * Provide insights based on your model analysis regarding the factors that influence flight ticket prices.
   * Suggest actionable strategies for the travel agency to optimize pricing strategies and improve customer satisfaction.
6. **Documentation:**
   * Document your process, including data exploration, preprocessing steps, model selection, and evaluation.
   * Include visualizations where necessary to support your findings.

**Deliverables:**

* A Jupyter notebook (or Python script) with the entire workflow.
* A report summarizing your findings, including the model's performance and recommendations for optimizing flight ticket pricing.

This assignment problem focuses on predicting flight ticket prices, a linear regression problem in the travel and aviation domain. The problem involves data preprocessing, feature engineering, model training, and evaluation, with a focus on understanding the factors that influence ticket prices in the airline industry.