High-Level Design Document (HLD)

Flight Fare Prediction Project

**Project Name:**

**Flight Fare Prediction**

**Objective:**

The primary objective of the Flight Fare Prediction project is to develop a model that accurately predicts flight fares based on various features using Random Forest Regression. This model will help customers to plan their travel budget more efficiently and allow businesses to optimize pricing strategies.

**Scope**

The scope of this project includes data collection, preprocessing, model development, evaluation, and deployment of the Flight Fare Prediction system. This document outlines the overall flow, the data to be used, the steps involved in the project, and the tools and resources required.

**Data Description**

**Data Sources:**

Flight fare data from online travel agencies and airlines.

Historical flight pricing data.

Additional datasets such as airport information, holiday schedules, and weather data.

**Key Data Features:**

Flight Details: Airline, Date of Journey, Source, Destination, Route, Departure Time, Arrival Time, Duration, Total Stops.

Pricing Information: Fare.

**High-Level Architecture**

**Data Collection:** Aggregating flight fare data from various sources.

**Data Preprocessing:** Cleaning the data, handling missing values, feature engineering, and transforming categorical data.

**Model Development**: Training a Random Forest Regression model using the processed data.

**Model Evaluation:** Assessing the model's performance using appropriate metrics.

**Deployment:** Deploying the model for real-time fare prediction.

**Steps Involved**

**Requirement Analysis:** Understanding the requirements and defining the scope of the project.

**Data Collection and Exploration:**

Collect data from link: https://www.kaggle.com/datasets/nikhilmittal/flight-fare-prediction-mh

Explore and understand the data structure and features.

**Data Preprocessing:**

**Data Cleaning:** Remove duplicates, handle missing values, and correct erroneous data.

**Feature Engineering:** Create new features such as duration in minutes, departure hour, etc.

**Encoding:** Transform categorical variables into numerical formats using techniques like one-hot encoding.

**Model Development:**

Split data into training and testing sets.

Train the Random Forest Regression model using the training data.

Tune hyperparameters for optimal performance.

**Model Evaluation:**

Evaluate the model using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared.

Validate the model using cross-validation techniques.

**Model Deployment:**

Integrate the model into a web application or API for real-time predictions.

Set up monitoring and logging for model performance and usage.

**Documentation and Reporting:**

Document the entire process, including data sources, preprocessing steps, model parameters, and evaluation results.

Generate reports and visualizations to communicate findings to stakeholders.

Tools and Resources

**Data Collection and Preprocessing:**

**Python:** Pandas, NumPy

**Model Development and Evaluation:**

**Machine Learning Libraries:** Scikit-learn,

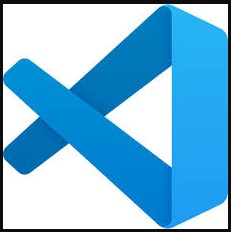
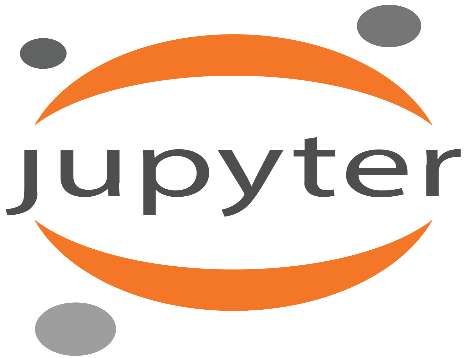
**Data Visualization:** Matplotlib, Seaborn.

**Evaluation Metrics:** MAE, MSE, R-squared.

**Deployment:**

Web Frameworks: Flask.

**Tool Used:**

**    **

**Visual Studio Code (VS Code)**

* Integrated Development Environment (IDE) for coding.
* Supports multiple languages and extensions.
* Provides debugging, version control, and task running.

**Pandas**

* Powerful data manipulation and analysis library.
* Provides data structures like DataFrame for handling tabular data.
* Essential for data cleaning, transformation, and aggregation.

**NumPy**

* Fundamental package for numerical computing in Python.
* Supports large, multi-dimensional arrays and matrices.
* Offers a wide range of mathematical functions.

**Seaborn**

* Data visualization library based on Matplotlib.
* Provides a high-level interface for drawing attractive statistical graphics.
* Useful for exploring and understanding data distributions and relationships.

**Matplotlib**

* Comprehensive library for creating static, animated, and interactive visualizations in Python.
* Supports a variety of plots such as line, bar, scatter, and histogram.
* Highly customizable for detailed visual analysis.

**Scikit-learn (sklearn)**

* Machine learning library for Python.
* Provides simple and efficient tools for data mining and data analysis.
* Includes algorithms for classification, regression, clustering, and dimensionality reduction.

**Flask**

* Lightweight web framework for Python.
* Used for developing web applications and APIs.
* Ideal for deploying machine learning models for real-time predictions.

**HTML**

* Standard markup language for creating web pages.
* Structures content on the web.
* Essential for developing front-end interfaces of web applications.

**GitHub**

* Platform for version control and collaboration.
* Hosts repositories and facilitates code sharing and collaboration.
* Provides tools for issue tracking, code review, and project management.

**Jupyter**

* Interactive computing environment for creating notebooks.
* Supports live code, equations, visualizations, and narrative text.
* Excellent for data exploration, visualization, and sharing results with stakeholders.

**Project Timeline**

Week 1-2: Requirement Analysis and Data Collection.

Week 3-4: Data Preprocessing and Feature Engineering.

Week 5-6: Model Development and Initial Training.

Week 7: Model Evaluation and Tuning.

Week 8: Deployment and Testing.

Week 9: Documentation and Final Reporting.

**Risk Management**

**Data Quality:** Ensuring the collected data is accurate and up-to-date.

**Model Performance:** Regularly tuning the model to adapt to new data trends.

**Deployment Challenges:** Ensuring the deployment environment is secure and scalable.

**Conclusion**

This High-Level Design document provides an overview of the Flight Fare Prediction project using Random Forest Regression. The detailed steps, tools, and resources mentioned here will guide the development team through the project, ensuring the successful implementation and deployment of the flight fare prediction model. This project aims to offer significant value to customers and businesses by providing accurate fare predictions and facilitating informed decision-making.