Capstone Project 2: Flight Price Prediction

1. Executive Summary:

This capstone project focuses on predicting flight prices using a machine learning model. The goal is to develop a predictive model that can accurately forecast flight prices based on various factors such as departure time, arrival time, flight duration, and more. The project utilizes Python for data processing and modelling, and Excel for initial data exploration and cleaning.

2. Problem Statement:

Background:
Airline ticket prices are highly variable and can fluctuate based on numerous factors including time of booking, flight duration, and seasonal demand. An accurate predictive model can help traveller's, airlines, and travel agencies make informed decisions and optimize costs.

Objective:
The objective of this project is to build a machine learning model that can predict flight prices with high accuracy using historical data. The model will leverage key features such as airline, departure time, arrival time, duration, number of stops, and other relevant variables.

Scope:
The project will involve data preprocessing, feature engineering, model training, evaluation, and optimization. The final output will be a Python-based solution that predicts flight prices using a trained machine learning model.

3. Data Sources:

Primary Data:
The primary dataset for this project is the Excel file Data_Train.xlsx, which contains historical flight price data including various features such as date of journey, airline, source, destination, route, duration, and additional price-related information.

4. Methodology:

Data Preprocessing:

- Load data from Data_Train.xlsx using Python libraries such as Pandas.
- Perform data cleaning (handling missing values, removing duplicates).
- Feature engineering to create new variables that may improve the model's predictive power.

Model Development:

- Split the data into training and test sets.
- Train multiple machine learning models (e.g., Linear Regression, Decision Trees, Random Forest, etc.).
- Evaluate model performance using metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared.

Model Optimization:

- Hyperparameter tuning using Grid Search or Random Search to improve model accuracy.
- Implement cross-validation to ensure model robustness.

5. Expected Outcomes:

- A machine learning model capable of predicting flight prices with reasonable accuracy.
- A set of insights regarding the most important features influencing flight prices.
- A Python script that can be used for future predictions or integrated into a larger application.

6. Tools and Technologies:

- Python (Pandas, NumPy, Scikit-Learn, Matplotlib, Seaborn)
- Jupyter Notebook for development and experimentation
- Excel for data storage and initial exploration

7. Risks and Challenges:

- Handling missing or incomplete data in the Excel dataset.
- Ensuring model generalizability to unseen data.
- Managing potential overfitting due to limited data size or high feature dimensionality.

8. Conclusion:

This capstone project aims to develop a robust predictive model for flight prices, leveraging historical data and machine learning techniques. By accurately forecasting flight prices, the model can provide valuable insights for stakeholders in the travel and aviation industry.