Task Performed: Data Tuning and Optimization

To achieve accurate flight fare prediction using machine learning, data tuning and optimization are crucial steps. These steps involve selecting the most important features, preprocessing the data, and selecting the best machine learning model and hyperparameters to achieve optimal performance.

The first step in data tuning and optimization is selecting the most important features. The features used in the model should be relevant to the problem at hand and have a significant impact on the target variable, which is the flight fare in this case. As mentioned in the reference article, several features such as departure date, departure airport, arrival airport, flight duration, number of stops, and airline have a moderate to significant impact on air ticket prices. Therefore, these features should be considered when selecting the features for the model.

The second step is preprocessing the data. This involves dealing with missing values, scaling the features, and encoding categorical variables. Missing values can be imputed using various techniques such as mean imputation or regression imputation. Scaling the features is necessary to ensure that each feature is on the same scale and does not dominate the model. Encoding categorical variables is also essential to ensure that the model can handle categorical data. One-hot encoding or label encoding are common techniques for encoding categorical variables.

The third step is selecting the best machine learning model and hyperparameters. As mentioned in the reference article, various regression models such as linear regression, decision tree, random forest, and support vector regression (SVR) can be used for flight fare prediction. However, selecting the best model and hyperparameters for the problem requires experimentation and evaluation. Cross-validation and grid search are common techniques for selecting the best model and hyperparameters. Cross-validation helps in evaluating the performance of the model on unseen data, while grid search helps in selecting the best hyperparameters for the model.

In addition to these steps, it is essential to ensure that the model is robust to outliers and overfitting. Outliers can be handled by removing them or transforming the data, while overfitting can be prevented by using regularization techniques or ensemble methods such as bagging or boosting.