

Task Performed: Data Exploration and Visualization

Data exploration and visualization are critical steps in any machine learning project, including predicting flight fares. These steps involve understanding the data, identifying any patterns, relationships or outliers, and presenting the data in a visually appealing way to gain insights and communicate findings to stakeholders.

In the case of flight fare prediction, the data set may include features such as departure date, arrival date, departure airport, arrival airport, flight duration, number of stops, and airline. The first step in data exploration is to get an overview of the data, which can be done by examining basic statistics such as the mean, median, standard deviation, minimum, and maximum values of each feature. This information can help identify any outliers or extreme values that need to be dealt with during preprocessing.

Next, data visualization techniques such as scatter plots, histograms, and box plots can be used to visualize the relationships between different features and the target variable (fare). Scatter plots can show the relationship between two continuous variables, such as flight duration and fare, while histograms can show the distribution of a single continuous variable, such as departure date. Box plots can be used to show the distribution of a continuous variable across different categories, such as the distribution of fares for different airlines.

In addition to these basic techniques, more advanced visualization techniques such as heat maps and parallel coordinates can be used to explore more complex relationships between multiple features. For example, a heat map can show the correlation between different features, while parallel coordinates can be used to visualize the relationship between multiple features and the target variable.

Data exploration and visualization can also help identify any missing values or inconsistencies in the data that need to be dealt with during preprocessing. For example, if there are missing values for a feature such as departure airport, this information may need to be imputed using methods such as mean or median imputation, or more advanced techniques such as regression imputation.