Machine Learning Project Report

Flight Fare Prediction

Presented by Manvendra Singh

Project Report

Abstract

Travelling through flights has become an integral part of today's lifestyle as more and more people are opting for faster travelling options. The flight ticket prices increase or decrease every now and then depending on several factors like timing of the flights, destination, and duration of flights on various occasions such as vacations or festive season. Therefore, having some basic idea of the flight fares before planning the trip will surely help many people save money and time. This System is designed to predict flight fares using machine learning algorithms based on various input parameters such as airline, source, destination, departure time, arrival time, duration, and total stops. This system integrates a web application for user interaction, allowing users to input their travel details and receive fare predictions.

Presented by Manvendra Singh

Project Report

Problem

To create a Machine Learning model that will predict the fares of the flights based on several factors available in the provided dataset.

Presented by Manvendra Singh

Project Report

Objectives

The classical machine learning tasks like Data Exploration, Data Cleaning, Feature Engineering, Model Building and Model Testing. Try out different machine learning algorithms that are best fit for the above case. To create a Machine Learning model that will predict the fares of the flights based on several factors available in the provided dataset.

Presented by Manvendra Singh

Project Report

Implementation

Data Gathering

The dataset for this project was sourced from Kaggle. It can be accessed and downloaded from the following link: Flight Fare Prediction Dataset.

Data Preprocessing

Cleansing and preparing the data for training, including handling missing values, encoding categorical variables, and normalizing the data.

Feature Selection

Selecting the most relevant features that influence flight prices.

Presented by Manvendra Singh

Project Report

Implementation

Model Training

Using processed data to train a machine learning model. Various algorithms like Random Forest, Gradient Boosting, and Linear Regression can be explored to find the best performer.

Model Evaluation

Evaluating the model's performance using metrics like MAE (Mean Absolute Error) and R^2 score.

Prediction

Using the trained model to predict flight fares based on user input.

Presented by Manvendra Singh

Project Report

Detailed Implemention

DATA PREPROCESSING

- Converting date columns to datetime format to extract useful features such as day and month.
- Handling missing values by dropping rows with missing data, as the dataset is large enough to afford losing a small fraction of data.
- Encoding categorical variables like Airline, Source, and Destination using one-hot encoding to convert them into a format that can be provided to the model.
- Feature engineering to create new features that might be relevant for the prediction, such as extracting the day of the week from the date of journey.

FEATURE SELECTION

The project selects features that are deemed relevant for predicting flight fares. This includes:

- Numerical features like Duration of the flight.
- Categorical features that have been one-hot encoded.
- Newly engineered features such as DAY and MONTH extracted from the DATE_OF_JOURNEY column.

MODEL SELECTION AND HYPERPARAMETER TUNING

- The project employs RandomizedSearchCV for hyperparameter tuning, which is a more efficient approach than GridSearchCV when dealing with a large number of hyperparameters and data.
- Multiple regression models are evaluated, including potentially Linear regression, Decision trees, Random forests, and Gradient boosting machines, though the exact models used are not specified in the provided inputs.

Project Report

Detailed Implemention

MODEL TRAINING AND EVALUATION

- The selected model is trained on the preprocessed training dataset. This involves learning the relationship between the input features and the target variable (PRICE).
- Cross-validation is used during hyperparameter tuning to ensure the model's generalizability and to prevent overfitting.
- The model's performance is evaluated using the R2 score, which measures the proportion of variance in the dependent variable that is predictable from the independent variables. This is a common metric for regression tasks.
- The evaluation is done on a separate test set that the model has not seen during training to assess its performance on unseen data.

System Design Flow Chart