Prediction of Price of Flight Tickets

ISM 6136- Data Mining Project

TEAM MEMBERS

ADITYA SAI MADDILA -U77227975

LEELA MAHESH CHAKRAVARTHY KODI -U28275523

VASUDEVA REDDY BOLLEDDULA - U44373196

INTRODUCTION:

In today's digital age, where convenience and efficiency are paramount, online platforms like "Ease My Trip" have revolutionized the way people book their flights. This research explores a wide range of flight reservation data from the well-known online travel agency "Ease My Trip." This dataset will be carefully examined, and relevant information will be extracted through the use of several data mining techniques. This study attempts to find important insights that could be very helpful to travelers as well as the airline business through thorough analysis and the use of cutting-edge machine-learning models.

One of the main characteristics of this work is the careful, methodical inspection of the dataset. The data is subjected to a thorough examination via a range of data mining techniques, allowing relevant information to be extracted. These are effective tools that help researchers find patterns, verify hypotheses, and reach relevant conclusions. The goal is apparent: to enable educated decision-making by analyzing the subtleties and underlying patterns that characterize the flight reservation industry.

The goal of this project is to unearth crucial insights that will increase their usefulness for travelers and the airline sector. These insights may provide real benefits for passengers, such as knowing when to make reservations of flights at lower prices, and which popular routes and destinations are available. Equipped with this understanding, tourists may make informed decisions, guaranteeing affordable yet fulfilling travel experiences.

In addition, the results of this study have the potential to have a big influence on the aviation industry. Airlines may adjust their services, marketing plans, and pricing structures by understanding the nuances of passenger preferences and behaviors. This customized strategy increases client happiness, encourages repeat business, and eventually supports the expansion and sustainability of the sector. Furthermore, this study's insights enable airlines to maximize their operational efficiency, guaranteeing a smooth and pleasurable travel experience for customers.

This study project is, at its core, an endeavor to gain information that has the capacity to change lives rather than just analyze data. This study provides evidence of the ability of analytical understanding and state-of-the-art machine-learning methods to shape traveler and airline industry futures by providing light on the way forward. In a data-driven age, this study project emphasizes how critical it is to use this abundance of data for the benefit of all travelers and the aviation industry as a whole, guaranteeing that the skies are not only crossed but also experienced with unmatched comfort and pleasure.

BACKGROUND:

Accurately anticipating flight ticket pricing is crucial in the dynamic world of air travel. Both travelers and airlines are always looking for the greatest offers. Researchers are exploring large datasets that include variables like as travel dates, routes, and passenger preferences by utilizing modern data analytics and machine learning techniques. The goal of these initiatives is to create predictive algorithms that can accurately predict changes in pricing.

These models provide travelers with crucial data that help them decide when to buy flight tickets and guarantee big discounts. Conversely, airlines maximize profitability through efficient revenue management—that is, by adjusting pricing in accordance with market trends and demand. Travellers' journey planning is being revolutionized by this convergence of technology in the aviation industry, which also gives airlines the ability to adjust to the ever-changing environment of price, guaranteeing everyone a smooth and financially feasible travel experience.

PROBLEM STATEMENT:

Travelers can make well-informed decisions regarding their air travel by using accurate and timely information that is provided by flight price prediction, which is an essential tool. Travelers may save money, arrange their budgets wisely, and improve their entire trip experience by predicting the fluctuations in ticket prices. In addition, this predictive capacity helps travel agencies and airlines promote competitive prices, optimize pricing strategies, and guarantee dynamic pricing adjustments in response to demand and market conditions. In the end, it promotes a more effective and customer-friendly aviation sector, which is advantageous to both passengers and service providers.

Our main problem statement would be -

How can we accurately predict flight ticket prices to help travelers save on costs and make more informed travel decisions?

SOLUTION METHODOLOGY

We have used the dataset from the Kaggle repository. There are 11 features in the dataset and we have filtered to 8 features that we will be used. We would need to calculate the R2 score and Mean Absolute Error, to assess the better fit of the model and to measure how far, on average, the predictions are from the true values. Finally, we will compare the R2 and MAE scores of different models. We have used Decision tree Regression, Random Forest Regression, Linear Regression, Lasso Regression, and Ridge Regression. We have divided the data set completely into training and testing datasets in 70 to 30 percent ratio and then we have calculated the R2 and MAE score which enable us to determine which model performs better.

DESCRIPTION OF DATASET:

The dataset consists of 10683 instances and 11 features, out of which 10 are independent variables and one Dependent variable which is our target variable.

Details of Independent Variables-

Airline: Airline company that provides air transport services.

Date_of_Journey: The specific Date on which the Journey is carried out.

Source: Airport from which the flight originates.

Destination: The airport where the flight is scheduled to arrive.

Route: The specific path or sequence of locations between Source and Destination airports.

Dep_Time: The specific time at which the flight is scheduled to begin.

Arrival_Time: The specific time at which the flight is scheduled to reach its destination.

Duration: The total time taken for the journey to be completed.

Total_Stops: The number of intermediate stops or layovers that occur during a journey.

Additional_Info: Specific details about services, and amenities related to the flight.

Details of Dependent Variable:

Price: The amount of money for the journey which is our target variable

MODEL COMPARISON:

The target variable, which after dataset analysis indicates is of continuous numerical variable type and predicts the fair prices of the Flight, suggests that the type of data mining problem we are dealing with is one of the regression models. We will thus take into consideration the most popular and efficient regression methods, which are listed below, in order to predict the target variable with less MAE score and more R2 score.

Decision Tree Regression:

Decision Tree Regression is a machine learning algorithm used for predicting continuous values. It works by recursively splitting the dataset into subsets based on features, creating a tree-like structure. Each leaf node represents a continuous output value. Decision trees can capture complex patterns but may overfit without constraints like limiting tree depth. They are valuable for regression tasks due to their ability to handle nonlinear relationships in data.

Random Forest Regression:

Random Forest Regression is a machine learning algorithm that uses multiple decision trees to predict numerical values. To increase accuracy and manage intricate relationships in data, it aggregates the predictions of these trees. In several domains, including environmental science and economics, Random Forest Regression provides reliable outcomes for jobs requiring exact numerical predictions.

Linear Regression:

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. The line that minimizes the discrepancy between the observed and anticipated values is determined to be the best fit. Because of its simplicity and efficacy, this method is extensively used in a variety of industries and is useful for predicting outcomes based on input factors.

Lasso Regression:

Lasso Regression is a linear regression technique that adds a penalty to the objective function, encouraging simpler models by setting some feature coefficients to zero. It is useful for feature selection, especially in datasets with many features, preventing overfitting and promoting sparsity.

Ridge Regression:

Ridge Regression is a regularization technique in linear regression that prevents overfitting by adding a penalty term to the objective function. This penalty discourages large coefficients, making the model more robust and stable, especially in the presence of correlated predictors. It helps address multicollinearity and maintains a balance between model complexity and accuracy.

DATA EXPLORATION AND DATA CLEANING:

After importing the dataset, we removed the unnecessary features that had no impact on the target variable were dropped and we used the summarize data feature to check for any missing values in any feature column and auto-filled them with the most frequent value, and checked the data type of each feature. We have converted the Duration feature value from hours to minutes and checked for any anomalies in the data. We have removed some of the instances from Total_Stops which have no effect on the price variable. We have scaled all the features to bring all the numerical values of all features on the same scale for better prediction. We then visualized the data by plotting various graphs like between Price variation of flights over Departure cities, the relationship between flight duration and prices and prices with the total_stops. Finally, we have Encoding the categorical data in each column to numerical labels. Now, the preprocessed and cleaned data is given as input to the split data feature.

Split data:

It will divide the dataset into a train dataset and a test dataset and here we have taken 70 percent as train data and 30 percent as test data.

Train and Test Model:

Training and testing a machine learning model involves using a subset of your data to train or teach the model and then using another subset to evaluate its performance. We have trained the data with five different models which are Decision Tree Regression, Random Forest Regression, Linear Regression, Lasso Regression, and Ridge Regression.

Model Evaluation:

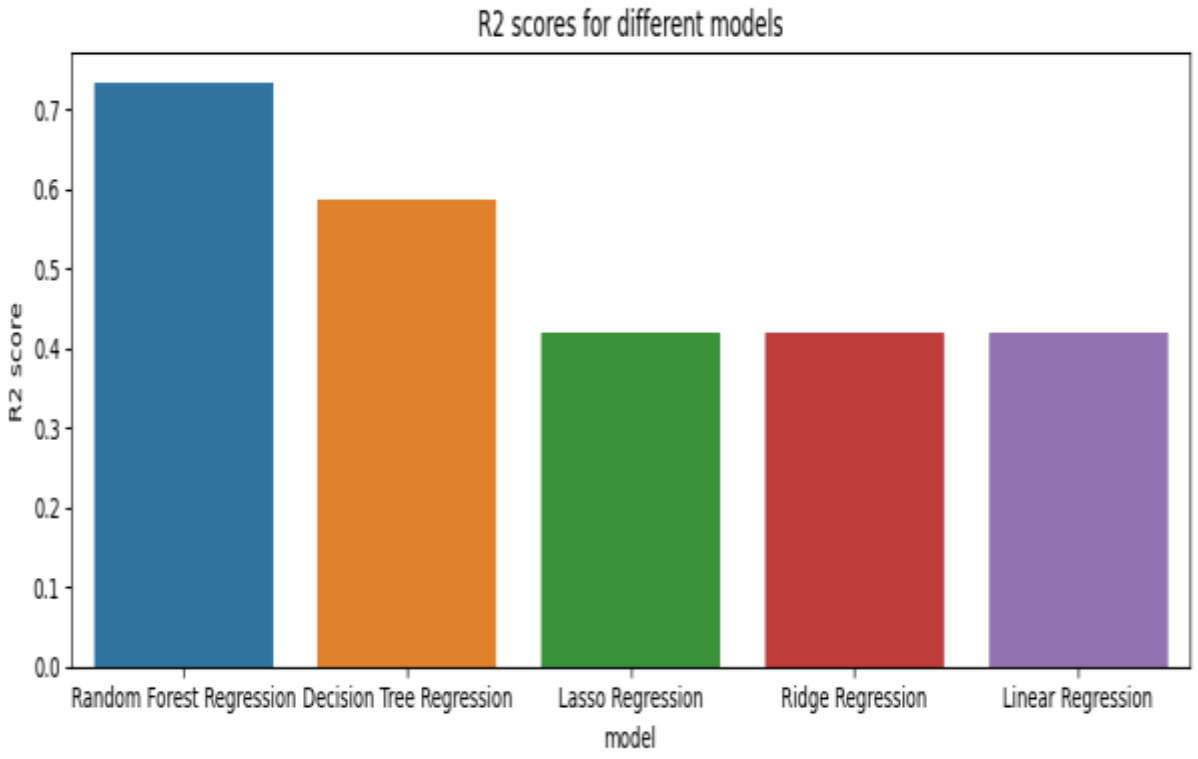
We have calculated the R2 score and MAE score of each model for evaluating the model performance

RESULTS:

All the R2 scores, and MAE scores were calculated and stored as a list for comparison.

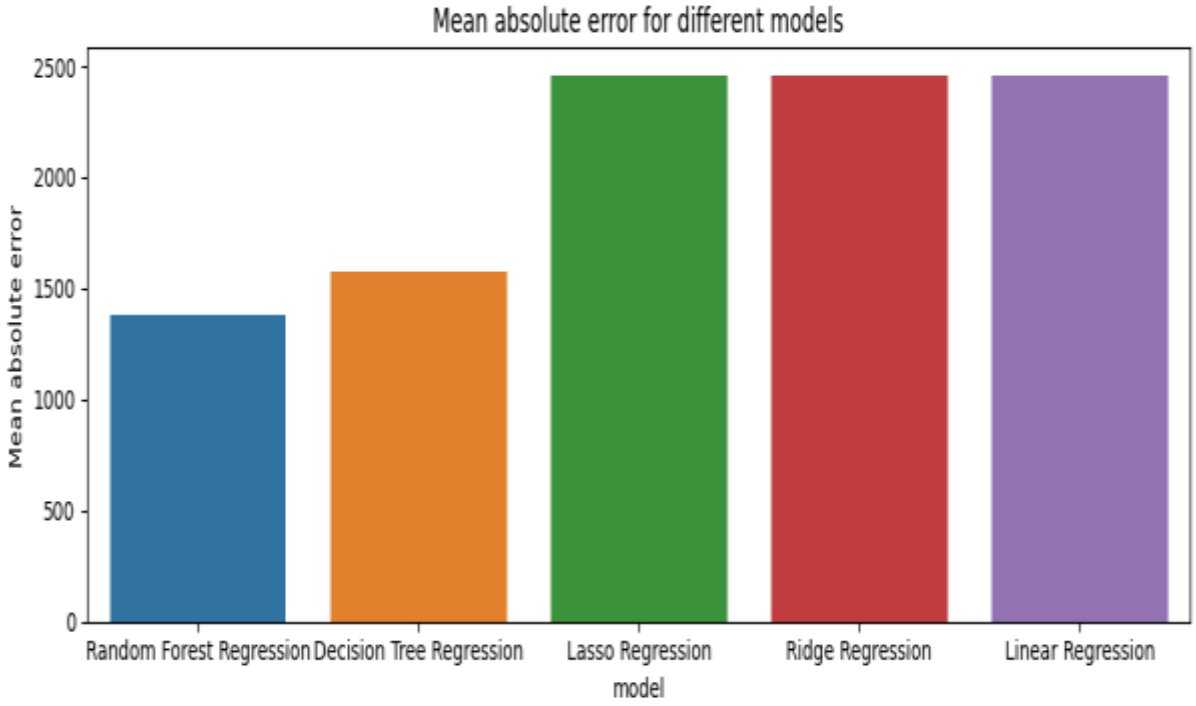
COMPARISON OF R2 SCORES OF ALL MODELS:

As R2 score is a metric used to evaluate the goodness of fit of a regression model and it indicates the proportion of the variance in the dependent variable that is predictable from the independent variables. On looking at the graph below of comparison of r2 scores of all models, the Random Forest regression has better score



COMPARISON OF MAE SCORES OF ALL MODELS:

The Mean Absolute Error (MAE) is a metric used to evaluate the accuracy of a regression model. It measures the average absolute difference between the predicted values and the true values, providing an indication of the model's accuracy. By looking at the comparison graph below Random Forest regression has a better MAE score and Decision Tree Regression has somewhat better in comparison to other remaining three models



CONCLUSION:

In our flight price detection project, Random Forest Regression emerged as the leading model, boasting an impressive 73.42% accuracy without fine-tuning. This robust performance underscores the model's innate ability to navigate the intricate landscape of airfare fluctuations. The findings provide a solid foundation for practical applications in the travel industry, offering stakeholders a reliable tool for anticipating and adapting to pricing dynamics.

Looking forward, the project opens avenues for optimization and further exploration. While the current accuracy is promising, fine-tuning the model could potentially unlock additional performance gains. Additionally, delving into advanced algorithms may present opportunities to refine the predictive capabilities of the model. This project not only advances our understanding of flight price prediction but also sets the stage for future research, encouraging a deeper dive into optimization strategies to enhance the effectiveness of machine learning applications in the domain of dynamic pricing.

Contributions:

Name	Contribution
Aditya Sai Maddila	Data Cleaning, visualization, and ppt
Leelamahesh Chakravarthy Kodi	ML model evaluation and Report preparation
Vasudeva reddy Bolleddula	Report preparation and Identification of data set

References:

https://www.kaggle.com/datasets/jillanisofttech/flight-price-prediction-dataset/data