iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VI **Month of publication:** June 2022

DOI: https://doi.org/10.22214/ijraset.2022.43666

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VI June 2022- Available at www.ijraset.com

Flight Price Prediction: A Case Study

Prithviraj Biswas, Rohan Chakraborty, Tathagata Mallik, Rohan Chakraborty, Sk Imran Uddin, Shreya Saha, Pallabi Das, Sourish Mitra
Computer Science & Engineering, Guru Nanak Institute of Technology, Kolkata

***Abstract*—A lot of factors that affect the overall price of airline tickets, including the airline, the date of travel, source, destination, route, duration, and so on. Each provider seems to have its own unique set regulations and methods for determining pricing. Recent breakthroughs in Artificial Intelligence (AI) and Machine Learning (ML) allow for the inference of such principles as well as the modelling of price volatility. This article is a study conducted on predicting flight prices. Utilizing two datasets for testing and training, this study analyses various machine learning methods for predicting flight prices.**

***Keywords*—artificial Intelligence, machine learning, air price, prediction model, dataset**

I. INTRODUCTION

Since the airline company's privatization, the airfare pricing scheme has evolved into a complex framework of sophisticated regulations including numerical simulations that determine airfare marketing strategies [1] [2] [3]. Even though these principles are still mostly unknown, research has revealed that they are influenced by a range of circumstances [4] [5]. Conventional characteristics such as distance, while still important, are no longer the only determinants of pricing structure. Economic, marketing, and sociological factors have all played a growing influence in determining flight pricing.

The majority of research on airfare forecasting has concentrated either on state scale or a single market. However, analysis at the business segment level is still relatively scarce. The marketing strategy is defined as the market/airport pairing in between aircraft sender and receiver.

II. RELATED WORK

It is critical for airlines to be capable of predicting airfare trends at the market segment level in order to alter strategies and resources for a given route. Scientific literatures on business segment price prediction, on the other hand, use biased conventional predictive methods, including such linear regression [6] [7], and thus are founded on the supposition that the selected variables have a linear relationship, that might not be true in most cases. Proposed research [8] Prediction of airfare prices utilizing machine learning approach, A dataset of 1814 Aegean Airways data flights was gathered and utilized to develop the machine learning technique for the study effort. Various figures of variables have been used to train the classifiers to demonstrate how feature extraction might affect validity of the model.

A study by William Groves' [9] shows that an operator can be introduced who can maximize purchase time on behalf of consumers. A model is constructed using the partial least squares approach.

Supriya Rajankar's survey report [10] on aircraft fare forecasting using machine learning models employs a tiny dataset consisting of flights between Delhi and Bombay. K-nearest neighbours (KNN), linear regression, and support vector machine (SVM) algorithms are used.

Over the course of several months, Santos[11] conducts research on airline routes between Madrid via London, Frankfurt, New York, as well as Paris. The figure depicts the acceptable number of days before purchasing an airline ticket.

Tianyi Wang[12] suggested a system in which two databases with macroeconomic data are integrated and machine learning methods including such support vector machines and XGBoost are being used to estimate the average price of tickets based on input and output pairings. With the updated R squared performance indicators, the framework obtains a high accuracy of 0.869.

In [13], a desired model is developed utilizing a Linear Quantile Blended Regression approach for the San Francisco–New York path, where daily airfares are provided via an internet domain. The model takes into account two factors: the number of days for travel and how much the trip is on a weekend or even a weekday.

III.DATA SET AND METHOD

A. Dataset

To begin, we need information on aircraft business and mass transit in order to develop the airline ticket pricing model just at market level. As a result, we have two datasets: training and testing. The training dataset contains 10,684 items with parameters such as airline, date of travel, source, destination, route, time of departure, estimated time of arrival, length, maximum stop, extra data, and price. The testing data set contains 2,672 items with the following attributes: Airlines, Date of Travel, Origin, End, Path, Time Of departure, Arrival Rate, Length, Maximum Stop, and Other Info.

©IJRASET: All Rights are Reserved

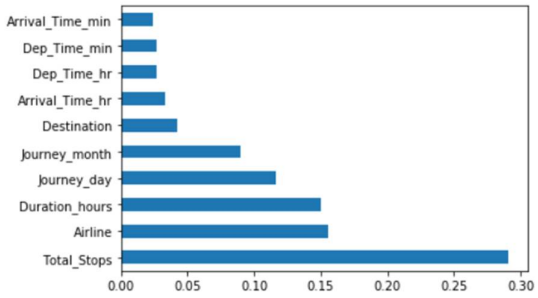


Fig. 1. Features of Dataset

TABLE I. DATASET FEATURE DESCRIPTION

Feature Name	Description
Airlines	As a result, this article will include all sorts of airlines such as Indigo, Jet Airways, Air India, among others.
Date of Journey	This column will inform us of the date the passenger's travel will begin.
Source	This column includes the names of the location from which the guest's journey will begin.
Destination	This column contains the name of the location where the passenger's journey will begin.
Route	This column includes the names of the location from where the customer's journey would begin.
Departure Time	The duration of a flight is the amount of time it takes to go from point A to point B.
Arrival Time	It will indicate how many spots the flight will stop over its journey.
Duration	The flight's endurance in hours..
Total Stops	The total number of breaks in the voyage.
Additional Information	It will indicate whether a meal is included with the journey or not.

B. Proposed Framework

Our suggested approach makes use of datasets to forecast airfare at the business segment levels. Fig 2 depicts a high-level view of the project framework's primary components. During the data pretreatment stage, all databases are cleaned to remove any potentially erroneous examples, then converted and integrated depending on market group. The feature extractor extracts and generates handmade attributes that are intended to describe the segment of the market. The goal of adaptive filtering modules is to improve accurate channels by assessing the utility of the characteristics and removing any unnecessary characteristics. Finally, we use the selected criteria to build our forecasting techniques, that result in the finished product of the projected airline cost of the ticket.

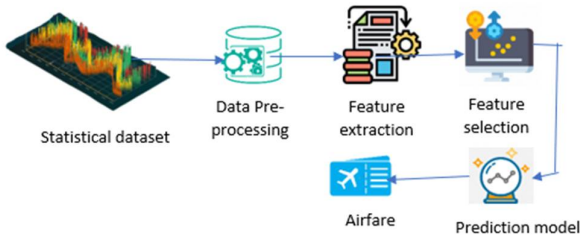


Fig. 2. Proposed framework for airfare price prediction.

1) *Data Pre-processing:* Many characteristics in the datasets have the same data. Furthermore, the statistics given by airlines can contain inaccurate figures due to human mistake, payment processing error, and so on. As a result, a well-designed data pre-treatment pipeline is critical for generating reliable input data for the machine learning algorithms. We discovered that the variables 'Route' and 'Total Stops' contain extremely few missing values in the data. We have one date form variable named 'Date of Journey,' as well as time variables called 'Dep Time' and 'Arrival Time.' The 'Journey Day' and 'Journey Month' variables may be extracted from the 'Date of Journey' field. 'Voyage day' indicates the month in which the journey began. Similarly, we may extract 'Departure Hour' as well as 'Departure Minute' from the 'Dep Time' as well as 'Arrival Time' variables, as well as 'Arrival Hour' and 'Arrival Minute' from the 'Arrival Time' variables. This 'Duration' field also contains duration information. This variable combines time hours & minutes data. 'Duration hours' as well as 'Duration minutes' can be extracted individually from the 'Duration' variable.
2) *Feature Engineering*: Here, we divide the characteristics and names first, then convert the hours to minutes. We are arranging the form of such a date of travel in our information in Date of Journey for easier preprocessing inside the model phase. Dep Time converts the departing date into clock time. Arrival Time is translated to minutes and hours by Arrival Time.
3) *Feature Selection*: To enhance model performance, a features extraction approach is used to investigate the level of influence of each information on the prediction outcome. The price section has been removed since it is no longer useful.

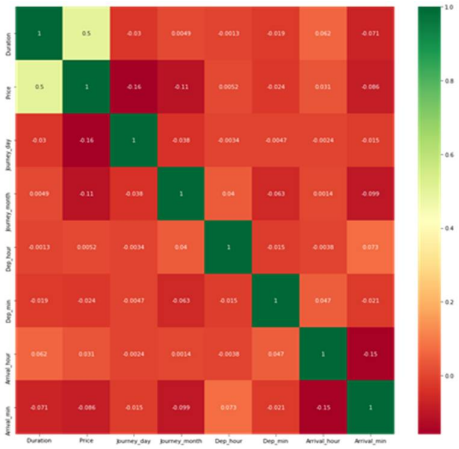


Fig. 3. Correlation between all Features

4) *Modelling*: We will now fit the image and forecast the results. We fit given data into different regression analysis in order to compare their efficiency and choose the right model.

IV. ALGORITHMS AND ANALYSIS

While we go through the algorithms we employed (XGBoost, Random Forest, and Decision Tree) and also how they operate in our models, please read the discussion below.

A. Decision Tree

The decision tree appears to be the most well-known and commonly employed categorization technique. A decision tree is a collection of nodes that resembles a diagram, for each junction indicating a test on the a characteristic and each branch indicating a test outcome, such that each node in a decision tree (terminal node) has a class label. A tree can be "trained" by dividing the resources collection into subgroups depending on a characteristic values test. This procedure is known as partitioning the data because it is performed iteratively on each derived subset. The recursion ends when all subgroups at a node have the same posterior probability, or when the split no longer adds additional value to the predictions. A decision tree is appropriate for experimental extracting knowledge since it does not need subject matter expertise or parameters configuration.
Assume S is a collection of cases, A is a property, Sv is the subgroup of S with Such a = v, as well as Value (A) is the collection of all number of values of A, then

$$\boldsymbol{Gain(S,A) = Entropy(S) - \Sigma_{ve} Values(A) \frac{|S_v|}{|S|} Entropy(S_v)}$$

B. Random Forest

A Random Forest is an ensemble approach that can handle simultaneous regression and classification problems by combining many decision trees using a technique known as Bootstrap as well as Aggregation, or bagging. The core idea is to use numerous decision trees to determine the final result instead of depending on personal decision trees. Random Forest's foundation learning methods are numerous decision trees. We arbitrarily choose rows and characteristics from the dataset to create sample datasets for each model. This section is known as Bootstrap. We simply have to understand the purity in our dataset, and we'll use that characteristic as the root of the tree which has the smallest impurity or, in other words, the smallest Gini index. Mathematically Gini index can be written as:

$$Gini\ Index = 1 - \sum_{i=1}^{n} (P_i)^2$$
$$= 1 - [(P_+)^2 + (P_-)^2]$$

C. XGBoost

XGBoost is an effective method for developing supervised regression models. Knowing as to its (XGBoost) goal function and baseline learners can help determine the truth of this proposition. This optimization problem has both a loss function and a regularization component. It makes a distinction between real and theoretical predictions, i.e. how far the model outputs deviate from the real amounts. In XGBoost, the most used standard error in regression problems is quarantine, whereas reg:logistics is used for classifications.
The formula may be used to compute the output value of each model.

$$Output\ value = \frac{\Sigma Residual}{no.of\ Residual\ +\ \lambda}$$

V. EXPERIMENTAL RESULTS

We drew the graph using data visualization to highlight the significance of each attribute for predicting flight prices.

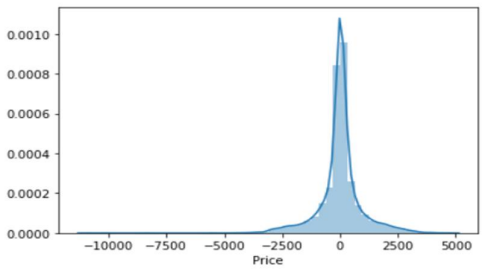


Fig. 4. Normal Distribution Curve between the difference of x-axis and y-axis of training dataset and price.

Meanwhile, depending on our studies, the highest accuracy in the training data for the decision tree method is 97 percent, and the greatest efficiency in Random Forest using testing data is roughly 78 percent.

TABLE I. RESULTS

Algorithm	Training Accuracy	Testing accuracy
XGBoost	0.92	0.77
Random Forest	0.95	0.78
Decision Tree	0.97	0.67

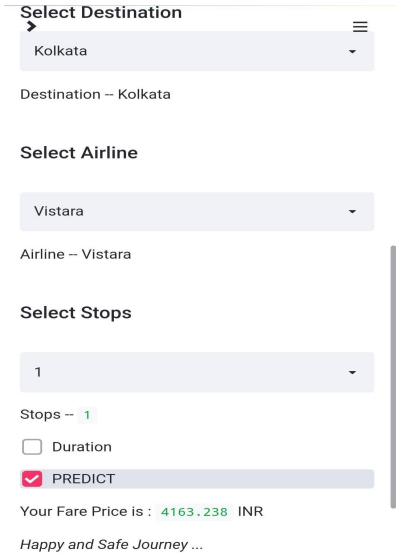


Fig. 5. Prediction of flight price

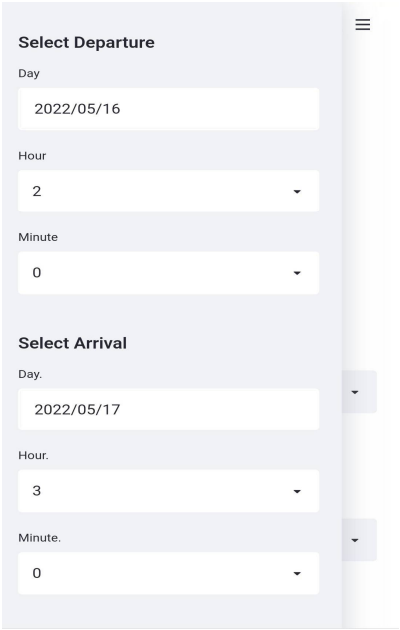


Fig. 6. Taking user input of Day, hour and minute to predict price



Fig. 6. Deployment architecture

VI. CONCLUSION AND FUTURE WORK

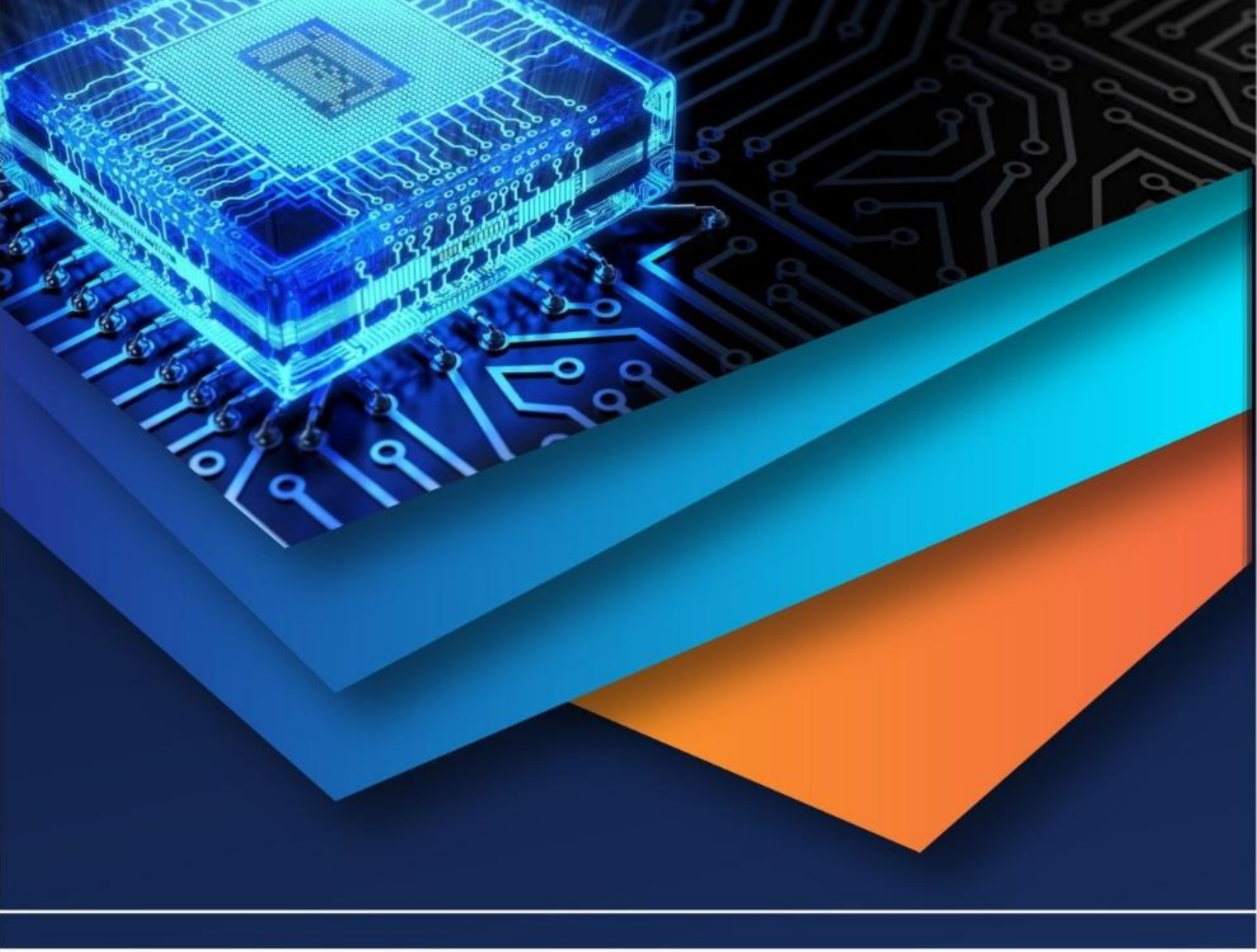
Three machine learning models were examined in this case study to forecast the average flight price at the business segment level. We used training data to train the training data and test data to test it. These records were used to extract a number of characteristics. Our suggested model can estimate the quarterly average flight price using attribute selection strategies.To the highest possible standard, much prior studies into flight price prediction using the large dataset depended on standard statistical approaches, which have their own limitations in terms of underlying issue estimates and hypotheses. To our knowledge, no other research have included statistics from holidays, celebrations, stock market price fluctuations, depression, fuel price, and socioeconomic information to estimate the air transport market sector; nonetheless, there are numerous restrictions.As example, neither of the databases provide precise information about ticket revenue, including such departing and arrival times and days of the week. This framework may be expanded in the future to also include airline tickets payment details, that can offer more detail about each area, such as timestamp of entry and exit, seat placement, covered auxiliary items, and so on. By merging such data, it is feasible to create a more robust and complete daily and even daily flight price forecast model. Furthermore, a huge surge of big commuters triggered by some unique events might alter flight costs in a market sector. Thus, incident data will be gathered from a variety of sources, including social media sites and media organizations, to supplement our forecasting models. We will also examine specific technological Models, such as Deeper Learning methods, meanwhile striving to enhance existing models by modifying their hyper-parameters to get the optimum design for airline price prediction.

VII. ACKNOWLEDGMENT

The authors would like to thank Pallabi Das (Assistant Prof. of Guru Nanak Institute of Technology, Kolkata), Dr. Santanu Kr. Sen (Principal of Guru Nanak Institute of Technology, Kolkata) for providing input and support.

REFERENCES

[1] J. Stavins, “Price discrimination in the airline market: The effect of market concentration,” Review of Economics and Statistics, vol. 83, no. 1, pp. 200–202, 2001.
[2] B. Mantin and B. Koo, “Dynamic price dispersion in airline markets,” Transportation Research Part E: Logistics andTransportation Review, vol. 45, no. 6, pp. 1020–1029, 2009.
[3] P. Malighetti, S. Paleari, and R. Redondi, “Has Ryanair's pricing strategy changed over time? an empirical analysis of its 2006–2007 flights,” Tourism Management, vol. 31, no. 1, pp. 36–44, 2010.
[4] T. H. Oum, A. Zhang, and Y. Zhang, “Inter-firm rivalry and firm-specific price elasticities in deregulated airline markets,” Journal of Transport Economics and Policy, vol. 7, no. 2, pp. 171–192, 1993.
[5] B. Burger and M. Fuchs, “Dynamic pricing – A future airline business model,” Journal of Revenue and Pricing Management, vol. 4, no. 1, pp. 39–53, 2005.
[6] T. M. Vowles, “Airfare pricing determinants in hub-to-hub markets,” Journal of Transport Geography, vol. 14, no. 1, pp. 15–22, 2006.
[7] K. Rama-Murthy, “Modeling of United States airline fares–using the official airline guide (OAG) and airline origin and destination survey (DB1B),” Ph.D. dissertation, Virginia Tech,2006.
[8] K. Tziridis T. Kalampokas G.Papakostas and K. Diamantaras ``Airfare price prediction using machine learning techniques" in European Signal Processing Conference (EUSIPCO), DOI: 10.23919/EUSIPCO .2017.8081365L. Li Y. Chen and Z. Li” Yawning detection for monitoring driver fatigue based on two cameras” Proc. 12th Int. IEEE Conf. Intell. Transp. Syst. pp. 1-6 Oct. 2009.
[9] William Groves and Maria Gini "An agent for optimizing airline ticket purchasing" in proceedings of the 2013 international conference on autonomous agents and multi-agent systems.
[10] J. Santos Dominguez-Menchero, Javier Rivera and Emilio TorresManzanera "Optimal purchase timing in the airline market".
[11] Supriya Rajankar, Neha sakhrakar and Omprakash rajankar “Flight fare prediction using machine learning algorithms” International journal of Engineering Research and Technology (IJERT) June 2019.
[12] Tianyi wang, samira Pouyanfar, haiman Tian and Yudong Tao "A Framework for airline price prediction: A machine learning approach"
[13] T. Janssen "A linear quantile mixed regression model for prediction of airline ticket prices"



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)