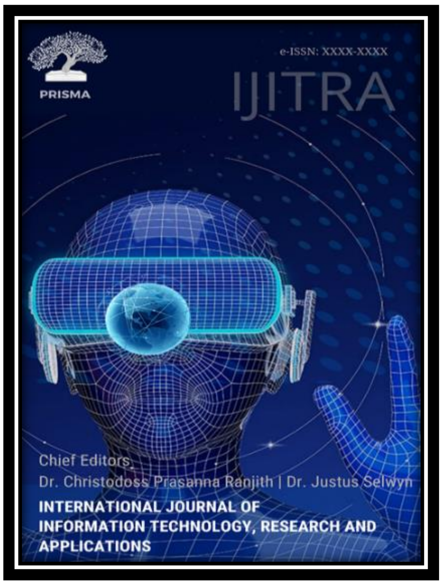
PRISMA
e-ISSN: XXXX-XXXX
IJITRA
Chief Editors
Dr. Christodoss Prasanna Ranjith | Dr. Justus Selwyn
INTERNATIONAL JOURNAL OF INFORMATION TECHNOLOGY, RESEARCH AND APPLICATIONS

International Journal of Information Technology, Research and Applications (IJITRA)

Ms Sarwath Unnisa, Rupam Kumari, (2024). Airline Price Prediction Using Extra Trees Regressor, 3(3), 07-14.

ISSN: 2583 5343

DOI: 10.59461/ijitra.v3i3.97

The online version of this article can be found at:
https://www.ijitra.com/index.php/ijitra/issue/archive

Published by:
PRISMA Publications

International Journal of Information Technology, Research and Applications (IJITRA) is a journal that publishes articles which contribute new theoretical results in all the areas of Computer Science, Communication Network and Information Technology. Research paper and articles on Big Data, Machine Learning, IOT, Blockchain, Network Security, Optical Integrated Circuits, and Artificial Intelligence are in prime position.

PRISMA
PUBLICATIONS

https://www.prismapublications.com/

Airline Price Prediction Using Extra Trees Regressor

Ms Sarwath Unnisa[1], Rupam Kumari[2]
[1,2]Dept. of Computer Science, Mount Carmel College Autonomous, Bangalore, Karnataka, India

Article Info

Article history:

Received March 31, 2024
Revised July 18, 2024
Accepted August 22, 2024

Keywords:

Airlines price Prediction,
Machine learning
Random Forest
XG Boost Regressor
Extra Trees Regressor

ABSTRACT

Airlines employ dynamic pricing strategies, which are based on demand estimation models, to set ticket prices. Airlines often set the price of their seats based on the necessary demand because they can only sell a certain number of seats on each journey. Airlines typically raise travel costs during times of strong demand, which slows down the rate at which seats are filled. When demand declines and seats become unsold, airlines typically lower their ticket prices to draw in new passengers. It will be more beneficial to sell those seats for more than the cost of service per passenger because unsold seats result in a loss of revenue. In order to forecast airline ticket prices, the primary objective of this study was to determine the factors influencing airline ticket costs and look into the relationships between them. Consequently, a model is created to predict the cost of plane tickets, allowing consumers to make better-informed decisions regarding their purchases. The cost of airline tickets in India is then predicted by this study using Four different machine learning algorithms: Extra Trees Regressor, XG Boost Regressor, Random Forest, and Decision Tree. Furthermore, hyperparameter optimization is done to obtain the most accurate and ideal prediction results. The Extrat Trees regressor yielded the greatest results, with the lowest RMSE of about 1807.59 and the highest accuracy of nearly 88%.

Corresponding Author:

Rupam Kumari
Dept. of Computer Science,
Mount Carmel College Autonomous,
Bangalore, Karnataka, India
Email: rupam12san@gmail.com

1. INTRODUCTION

Businesses that have standing inventory frequently employ intricate and dynamic pricing strategies in an effort to increase sales. To address this, airline corporations employ dynamic pricing tactics based on hidden variables and proprietary algorithms. The cost of the ticket. Customers find is more challenging to forecast the price of plane tickets going forward as a result . Customers who have access to flight pricing information are better able to forecast future price fluctuations because they can track price changes over a predetermined period of time. However, one cannot accurately forecast the cost of the flight based only on observations. There are various buckets of seats on a flight, and each bucket has a different cost. To make more money, the airlines reorganize these seats across buckets, raising the cost of the flight tickets for passengers. Consequently, different clients pay varying costs for the same flight. A ticket for an airline is priced based on a number of variables, including the cost of gasoline on the international market, the length of the journey, the time the consumer buys the ticket, etc. Each airline sets its own pricing for tickets based on its own set of guidelines and algorithms. Predicting when plane tickets will be the lowest can be a very profitable business because ticket costs fluctuate over time. Numerous studies on buy-wait methods and ticket price prediction have been published . Yield management, in which airlines adjust rates in response to past demand and aircraft capacity, is a significant method by which they control their standing inventory. Yield management was once done entirely by hand, but YMS (Yield Management Systems) has fully automated the process presently.

Therefore, in order to optimize income, YMS strives to sell the right seat to the right consumer at the right price at the right time. Yield management is relevant not only to airline reservations but also to hotel reservations, cruise ship reservations, and vehicle rentals. Using a multi-strategy data mining technique called HAMLETT, a YMS applies airfare data that has been crawled from the internet to estimate demand and modify prices accordingly. They developed a model that could forecast flight prices using a number of data mining approaches, enabling the clients to save a significant sum of money. Over the course of 41 days, they used a total of 12,000 observations in their investigation. The authors also state that the quantity of available seats on a flight has a significant impact on ticket prices. employed regression and clustering approaches to assist clients in making informed judgments. Initially, the k-means clustering technique was employed to group similar airlines whose costs were almost same. Later, Random Forest was utilized to ascertain the feature importance for predicting airline fares. A other strategy, proposed by, which is based on the random tree forest algorithm and the theory of (marked) point processes, ought to be less computationally demanding than the HAMLETT approach. The outcome demonstrated that while they perform almost as well as HAMLETT, their prediction is more valuable due to the specific situation and potential interpretation. This article predicts aircraft ticket prices fairly effectively, using three distinct machine learning methods, many days in advance of the travel time. Therefore, applying this strategy could help future travelers decide whether to buy a ticket at a particular cost. The study's objectives..[1],[2]

The following are this paper's primary goals.

To determine the key elements that affect the cost of an airline ticket.

To use three distinct machine learning (ML) methods to estimate airline pricing.

To assess how accurate each machine learning model was throughout this investigation.

1.1. Organization of paper

This is how the remainder of the paper is structured. The review of the literature is covered in Section 2. The dataset's exploratory data analysis and the study's methodology are covered in Sections 3 and 4, respectively. The outcomes and an accuracy comparison of all the models are presented in Section 5. Sect. 6 finally wraps up and explains the results.

2. LITERATURE REVIEW

The modeling of airline pricing through the use of data mining and machine learning algorithms has grown significantly during the last ten years. Regression models such as support vector machines (SVM), random forests (RF), decision trees (DT), and linear regression (LR) are frequently employed in order to forecast flight prices with accuracy. In order to forecast airline ticket prices, eight machine learning models—including ANN, RF, SVM, and LR—are also used. In their testing, Bagging Regression tree outperformed all other machine learning techniques with the greatest R2 of over 88% accuracy. used data with more than a lakh observations and four statistical regression models to predict the price of the flight ticket. Their program suggests to the passenger waiting till the price drops or purchasing a ticket at a specific price. The fact that this approach could only be used to economy tickets—and only to flights from San Francisco to Kennedy Airport—was a constraint on their efforts. They claimed that the best method for predicting ticket prices several days ahead of departure was to use mixed linear models with linear quantile. attempted to use support vector machines (SVM), linear regression, naive bayes, and softmax regression to forecast the cost of airline tickets. A model was constructed using almost 9,000 data and six distinct attributes, such as the total number of stops, the time from the booking, the departure date, etc. The linear regression model outperformed the other models with the lowest error rate of about 22.9%, while the SVM performed the worst in their analysis. Nevertheless, SVM was used to separate the airfares into two groups according to whether the price was "lower" or "higher" than the average. created a model that predicted airline tickets according to the kind of flight, such as direct, one-stop, or non-stop. proposed a mathematical approach to predict the best deals on airfares for particular flights . With the use of a two-month dataset and the isotonic regression technique—which is essentially a non-parametric technique—the researchers created a model that suggested users buy plane tickets at a particular time. The writers of looked into how much airline tickets cost over a specific length of time using variables like the number of stops made, the number of days until departure, etc. utilized support vector machines to forecast future airline ticket prices, with a 69.4% accuracy rate. a partial least squares regression model with a 75.3% accuracy rate to maximize the purchase of airline tickets. employed Artificial Neural Networks (ANN) and Genetic Algorithms (GA) to forecast airline ticket revenue. The weighted index of the Taiwan stock market, the worldwide oil price, and Taiwan's monthly unemployment rate were added as input features. By choosing the best input features, the GA enhances the ANN's performance. With a mean absolute percentage error of 9.11%, the model demonstrated strong performance. To create a program that could anticipate airline ticket prices more correctly, sophisticated machine learning algorithms are being used. suggested using Deep Regressor Stacking to produce forecasts that are more accurate. To increase the model's accuracy, they employed ensemble models such as SVM and Random forest[1], [3], [4].

3. EXPLORATORY DATA ANALYSIS

The dataset for the study consists of eleven distinct attributes, which are shown in Table 1: airline, date of travel, source, destination, route, duration, total stops, additional information, and cost. The dataset was made available for download on the Kaggle website. 10,683 observations and 11 columns were present at first. We examined each null and duplicate value in the dataset. Since the characteristics of Date of Journey, Arrival Time, and Dep Time were object types, data processing was done for them. In order to make an accurate prediction, they were translated into date and time. The hours and minutes of duration were split from the overall duration because the duration column had undergone pre-processing. Because of a few outliers, the Price element was changed to the median value. To handle category data and convert it into numerical format, two main encoding techniques were used. The original approach of nominal data without any kind of order was one hot encoding. Using the ordinal data, which were organized in a specific order, the label encoder was the second approach used

Table 1. Dataset description

Sl. No.	Variable name	Description	Data type
1	Airline	Name of the Airline	Object
2	Date of Journey	Date on which the journey starts	Object
3	Source	Boarding place	Object
4	Destination	The landing place	Object
5	Route	Route from which the flight travels	Object
6	Dep Time	The departure time of the flight	Object
7	Arrival Time	The arrival time of the flight	Object
8	Duration	Total duration of journey	Float
9	Total Stops	Total number of stops	Object
10	Additional Info	Any additional information	Object
11	Price (dependent variable)	Price of the airline ticket in Indian rupees (INR)	Int

The relationship between the average price and the airlines is shown in Figure 1. Figure 1 shows that while the average cost of the other carriers is roughly equal, the average cost of Jet Airways Business class is the highest. With the lowest average cost is SpiceJet. Figure 2 shows the relationship between the price and the source. While the average price among the other sources is around the same, Bangalore has a slightly larger percentage of outliers. Figure 3 shows the link between price and month. It's clear that April has somewhat cheaper flight costs than other months of the year.

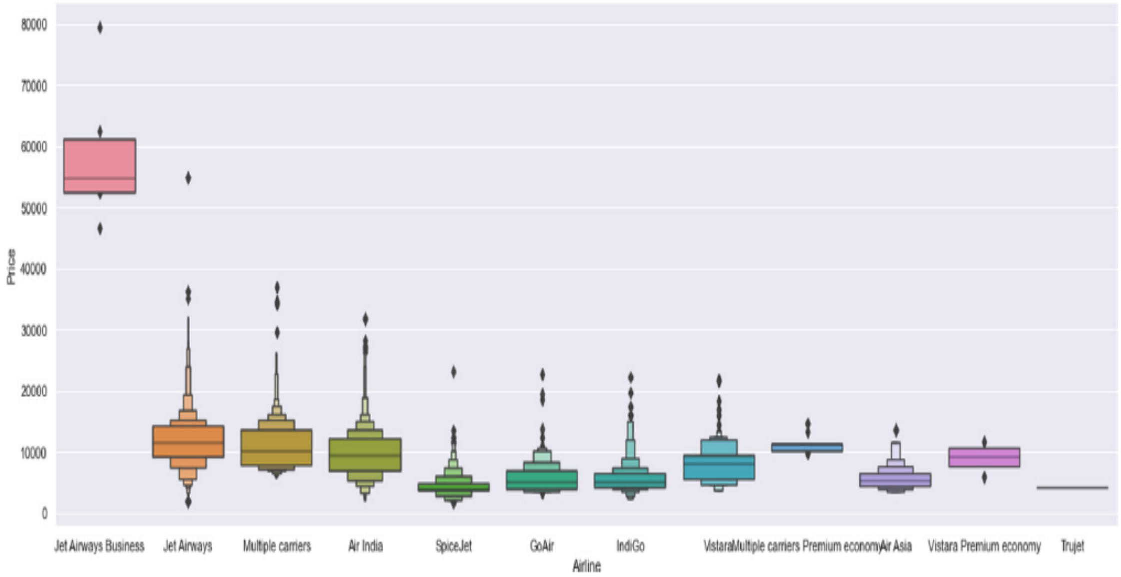

Fig. 1. Average price of all the airlines

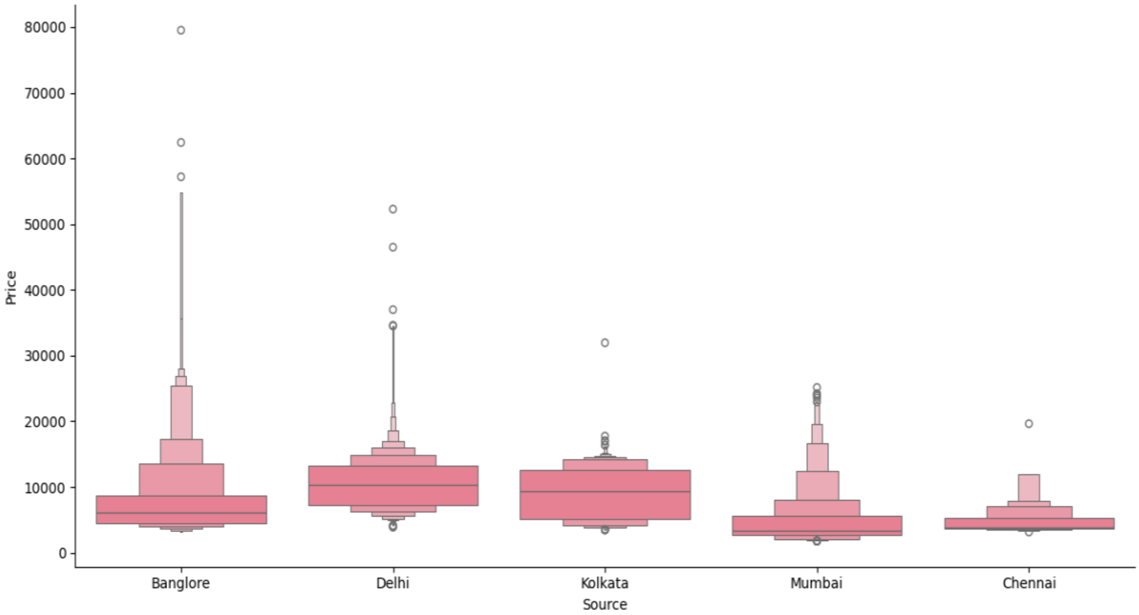

Fig. 2. Average price vs source

4. METHODOLOGY

4.1 Random Forest Regressor

A group of decision trees referred to as Random Forest is included in decision tree ensembles called Random Forests (RF). RF starts by trying to divide the original dataset into smaller groups. Subsets of the training dataset are produced when feature and row sampling with replacement is finished. In the following stage, it creates a distinct decision tree for each subgroup and generates an output. The decision tree method has challenges due to the training dataset's low bias and the testing dataset's substantial variation. But Random

Forest gets around this problem by using the bagging technique, which lowers the variance of the testing dataset in an effort to solve the bias-variance problem[5]

Random forests uses the bootstrap aggregation technique for the training dataset in order to train the tree learners. Fitting trees to random samples is done by bagging repeatedly (B times) given a training set X = x1,..., xn and its outputs as Y = y1,..., yn. It generates n training instances from X, Y, known as Xb and Yb, for b = 1,..., B, with replacement. After that, it uses Xb and Yb to train a regression or classification tree fb. Predictions for unobserved samples x 'can be made by averaging the predictions from each individual regression tree on x '.Given that this is a regression problem, the final result is determined by averaging all of the decision tree's outputs, which is represented as[5]

$$\widehat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(x')$$

4.2 Decision Tree

Classification and regression The two main types of decision trees are regression and classification. Regression is used for continuous data, and classification is used for categorical values. To make decisions, decision trees use independent variables from datasets as decision nodes. The dataset is divided into multiple sub-sections, and the section to which each info point is given determines the output of the model once test data is entered. The output of the decision tree also depends on which sub-section the information point is in, as it is the average of all the information points in that sub-section[4].

4.3 Bagging Regression Tree

Decision trees have a big variation with complicated trees and a large bias with basic trees. Bootstrap aggregating, a technique for choosing a random subset of data from a dataset with replacement, is where bagging originates. Its main purpose is to lessen the tree's variation. It is evident from the literature that random forest and gradient boosting techniques yield the highest accuracy[6].

4.4 XG Boost Regressor

The Extreme Gradient Boosting (XGBoost) technique is an efficient way to use gradient boosting to regression predictive modeling. This method is used when a forecast requires a large amount of precise data. Boosting is basically an ensemble of learning approaches that combines the predictions of multiple estimators to increase resilience over a single estimator. To produce a strong predictor, it combines several mediocre or weak predictors. This methodology's dual-part objective function is the main way in which it varies from earlier gradient boosting techniques. The first is the training loss, while the second is the regularization term[5].

4.5 Extra Trees Regressor

Extra Trees Regression is an ensemble learning method, meaning it combines multiple individual models to produce a more robust and accurate prediction. It utilizes the wisdom of crowds by aggregating the predictions of multiple models, which often leads to better generalization performance than individual models. It tends to be faster to train compared to traditional Random Forests because of its increased level of randomization. It often requires less tuning of hyperparameters compared to other ensemble methods, making it easier to use out of the box. Extra Trees can perform well even with noisy data and datasets with a large number of features.

5. Results

5.1 Evaluation Metrics

Metrics like as root mean square error (RMSE), mean absolute error (MAE), mean square error (MSE), and R2 are used to evaluate the models. Mean absolute error (MAE), or the difference between the goal and expected values, as shown in Equation. MAE penalizes errors less severely and is less susceptible to outliers. It gives equal weight to every single difference[7].

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \widehat{y_i}|$$

The average of the squares of the errors, or the average squared difference between the estimated and actual values, is the mean squared error (MSE) or mean squared deviation (MSD) of an estimator (of a process to estimate an unobserved variable). Even tiny errors are punished, leading to an overestimation of the model's performance. Since a lesser value denotes a smaller magnitude of error, a smaller MSE score suggests a better match. The MSE is provided in the equation[8], [9].

$$MSE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \widehat{y_i}|^2$$

Taking the square root of the difference between the target value and the anticipated value yields the root mean squared error (RMSE), as seen in the equation. Before averaging, an error is squared, and a penalty is imposed for significant errors. Hence, when significant errors are not expected, RMSE is useful[9].

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} |y_i - \widehat{y_i}|^2}$$

When forecasting the result of an event, the coefficients of determination (R2), as indicated in the equation, are a statistical metric that look at how changes in one variable can be explained by a change in another[7].

$$R^2 = 1 - \frac{RSS}{TSS}$$

where,

$\boldsymbol{y_i}$=actualvalues,

$\boldsymbol{\widehat{y}_\imath}$= predicted values,

n = total number of data points,

RSS = Residual sum of squares and

TSS = Total sum of squares

5.2 Accuracy Comparison of Models

The preprocessed data was used to build training and testing datasets. The remaining eighty percent of the dataset was utilized for training, and twenty percent was used for testing. All Four machine learning techniques were then applied to the training and testing datasets, and the models' accuracy was assessed using the evaluation metrics. The accuracy comparison of every model employed in this investigation is displayed in Table 2. For the random forest, decision Tree ,XGBoost regressor, and ExtraTrees regressor the corresponding R2 values were 81%, 75%, and 87%,88%. Following ExtraTrees regressor hyperparameter adjustment, we obtain a model with the lowest RMSE of roughly 1807.59 and the highest R2 of almost 89%. As a result, after hyperparameter adjustment, the XGBoost regressor performed better than the other machine learning algorithms in this study.

Table 2. Accuracy of all the models

Sl.S.No.	Model	MAE	MSE	RMSE	R2
1	Decision Tree	1961.25	8432989.82	2903.96	75%
2	Random forest regressor	1190.76	4011116.33	2002.77	81%
3	XGBoost regressor	1159.96	3307070.72	1818.53	87%
4	ExtraTrees regressor	1190.63	3267398.24	1807.59	88%

6. Conclusion

The prices of various airline tickets in India are examined in this essay. It was found that flights run by Air India, Jet Airways, and Multiple Carriers had the highest average cost, while flights run by IndiGo and Air Asia had the lowest average cost. The average cost was almost the same from all of the providers. There was also a little drop in flight costs in April. To predict airline fares, Four different machine learning (ML) techniques were utilized: random forest, XGBoostregressor,decision tree and ExtraTrees regressor.Using feature selection, redundant and superfluous variables were removed in order to determine which feature was most important. Therefore, the lowering of dimensionality was also discussed. Based on the assessment measures, the ExtraTreesregressor with hyperparamater tuning had the highest accuracy in estimating the price of an airline ticket. As a result, hyperparameter tweaking increased accuracy.

In the future, we can expand our framework to incorporate air ticket transaction data, which can offer more specifics about a certain itinerary, including the time and date of departure and arrival, the placement of the seat, covered supplemental products, etc.Such data can be used to create a more potent and thorough daily or even hourly airfare price prediction model by merging it with the current framework's macroeconomic characteristics and market segments.

References

[1] N. Shukla, A. Kolbeinsson, K. Otwell, L. Marla, and K. Yellepeddi, “Dynamic Pricing for Airline Ancillaries with Customer Context,” Feb. 2019, [Online]. Available: http://arxiv.org/abs/1902.02236

[2] . K., K. Wiweka, A. Parantika, N. Wahyuni, and P. P. Adnyana, “A Time Series Analysis of Airline Pricing Behavior Case Study Jakarta (CGK) - Denpasar (DPS) Market,” *Journal of Economics, Management and Trade*, pp. 1–10, Mar. 2019, doi: 10.9734/jemt/2019/v22i630105.

[3] J. A. Abdella, N. M. Zaki, K. Shuaib, and F. Khan, “Airline ticket price and demand prediction: A survey,” *Journal of King Saud University - Computer and Information Sciences*, vol. 33, no. 4. King Saud bin Abdulaziz University, pp. 375–391, May 01, 2021. doi: 10.1016/j.jksuci.2019.02.001.

[4] K. Rathi, A. Kumar, and M. Yadav, “AIRLINE FARE PRICE PREDICTION,” *International Research Journal of Engineering and Technology*, 2022, doi: 10.23919/EUSIPCO.

[5] A. Kumar, “Airline Price Prediction Using XGBoost Hyper-parameter Tuning,” in *Communications in Computer and Information Science*, Springer Science and Business Media Deutschland GmbH, 2023, pp. 239–248. doi: 10.1007/978-3-031-28183-9_17.

[6] S. Rajankar, N. Sakharkar, and O. Rajankar, “Predicting The Price Of A Flight Ticket With The Use Of Machine Learning Algorithms,” *INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH*, vol. 8, 2019, [Online]. Available: www.ijstr.org

[7] W. Groves and M. Gini, “A regression model for predicting optimal purchase timing for airline tickets A regression model for predicting optimal purchase timing for airline tickets A regression model for predicting optimal purchase timing for airline tickets,” 2011.

[8] J. A. Abdella, N. M. Zaki, K. Shuaib, and F. Khan, "Airline ticket price and demand prediction: A survey," *Journal of King Saud University - Computer and Information Sciences*, vol. 33, no. 4. King Saud bin Abdulaziz University, pp. 375–391, May 01, 2021. doi: 10.1016/j.jksuci.2019.02.001.

[9] P. Bambroo, "Analysis of dynamic pricing in airlines and predicting least fare," 2019. [Online]. Available: www.IJARIIT.com

BIOGRAPHIES OF AUTHORS

	Sarwath Unnisa- Sarwath Unnisa is an accomplished author and Assistant Professor in Dept. of Computer Science, Mount Carmel College Autonomous, Bangalore, Karnataka, India with a passion for interdisciplinary research, sarwath has made Significant contribution to various fields within computer Science. Her research endeavors have not only contributed to the academic Community But have also addressed real world challenges, fostering, innovation and technological advancement
	Rupam Kumari – Successfully completed BSc.B.ed at Tagore Govt College of education,Port Blair, Andaman Nicobar Islands. Currently pursuing MSc degree, Rupam continues to demonstrate a thirst for Knowledge and a commitment to excellence, is poised to make significant Contributions to both academic and society.