# Prediction of Spaceship Titanic's Passenger's Travelling Direction

Vennela Chadalavada, Rohitha Reddy Muppidi, Shirisha Avadoota
Kent State University

***Abstract*—Spaceship Titanic was a spaceship that traveled between stars and carried passengers from one planet to another. It can accommodate over 13,000 passengers. Emigrants from their home planet were being transported by the spacecraft on its maiden mission to three newly habitable exoplanets orbiting nearby stars. The interstellar passenger liner met with a spacetime anomaly concealed under a dust cloud while its route to its first destination, the arid 55 Cancri E, rounding Alpha Centauri. Over half of the passengers were sent to a different dimension as a result of the accident, which disturbed the orbit.Using data obtained from the interstellar spaceship's malfunctioning computer system, the goal is to find the best suitable machine learning algorithm which provides us the better accuracy.**

## I. Introduction

The Titanic disaster is one of the most horrific endings in human history, and the majority of us can still clearly remember it. It sank in less than three hours after four days of its inaugural journey as a result of hitting an iceberg. Only 1517 out of 2240 passengers were able to escape the wreckage, resulting in a large death toll. There were only 705 passengers who were saved. Unfortunately, we would have been powerless to prevent the Titanic ship disaster and save the people. Recent developments involve conducting large trials and developing cutting-edge algorithmic techniques for difficulties that could arise in the present and the near future. Spaceship Titanic is one such issue we face.

Based on mathematical ideas, machine learning algorithm techniques produce predictions. It focuses on looking at various angles of a problem from a specific circumstance. Nevertheless, the same solution may not always be effective and frequently leads to performance decrease.Predicting which passengers would be affected by the anomaly will enable rescue teams and help passengers be retrieved. To do this, they will need to use records that were salvaged from the interstellar spaceship's damaged computer system.

According to statistics, this work is a classification problem, thus we will use a variety of classification algorithms and evaluate the outcomes. Our projections must be accurately classified. True positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) are the four major statistical concepts that rely on it (FN). The Performance Evaluation will provide more information about this.We shall examine the underlying idea of this technique since we asserted that this task was a classification category for supervised learning.

## II. Literature Survey

Some machine learning and data mining tasks used many common classification methods as a classifier. For modelers conducting research, creating high-performance binary classification models is a difficulty. This paper reviews a number of current works on classification strategies and how well they perform when classifying binary data.

In referred paper[1], we referred An efficient XGBoost–DNN-based classification model for network intrusion detection system where the proposed XGBoost-DNN model uses the XGBoost strategy for feature selection before classifying network incursion using deep neural networks (DNN). Three steps comprise the XGBoost-DNN model: normalization, feature selection, and classification. During DNN training, the Adam optimizer is used to maximize learning rate, while the Softmax classifier is used to categorize network intrusions. Tensor flow and Python were used to implement the tests, which were correctly carried out on the benchmark NSL-KDD dataset. Cross-validation is used to validate the suggested model, and it is contrasted with other shallow machine learning techniques like logistic regression, SVM, and naive Bayes. The assessment metrics for classification, including accuracy, precision, recall, and F1-score, are computed and contrasted with the current shallow approaches. The suggested method fared better than the dataset's existing shallow methods.

Both the logistic regression and the decision tree models used in the research [2] use predictor features that are a mix of categorical and numerical information, and both models are prone to overfitting. In this project, a dataset with all the characteristics of a binary categorical scale will be used to develop a ridge logistic regression and decision tree model. The decision tree model's accuracy performance is 81%, whereas the ridge logistic regression model's accuracy performance is 84%.Only 2 features make up the majority of the dataset, accounting for almost 80% of the feature importance.

In this paper[3] we studied possible paths for additional research into the efficiency of deep learning networks for stock market analysis and forecasting. Deep learning methods come in a wide range of network structures, activation functions, and other model parameters, and it is well recognized that the way data is represented has a significant impact on how well these algorithms perform. Our research aims to offer a thorough and unbiased evaluation of the benefits and limitations of deep learning algorithms for stock market analysis and forecasting. We investigate the impact of three unsupervised feature extraction techniques—principal component analysis, autoencoder, and the restricted Boltzmann machine—on the network's overall capacity to forecast future market behavior using high-frequency intraday stock returns as input data. Yet, when the autoregressive model is applied to the network residuals, it is not possible to say that deep neural networks can extract additional information from the residuals of the autoregressive model and increase prediction performance. When the predictive network is used to do a covariance-based market structure study, covariance estimation is also much enhanced.

In the cited paper[4], decision tree classifiers were mentioned as one of the most well-known approaches to classifier representation in data. The issue of extending a decision tree from accessible data, such as machine learning, pattern recognition, and statistics, has been studied by several academics from a variety of professions and backgrounds. The use of decision tree classifiers has been suggested in a variety of domains, including medical disease analysis, text categorization, user smartphone classification, pictures, and many more. Gain in Information and Entropy Entropy is used to quantify the unpredictability or impurity of a dataset. Entropy's value is constantly between 0 and 1.Decision tree classifiers are renowned for providing a more comprehensive picture of performance results. All reputable data classifiers use upgraded tree pruning approaches (ID3, C4.5, CART, CHAID, and QUEST) and optimized splitting parameters because of their high precision. The precision of the test set is impacted by the employment of the distinct datasets for training samples from a sizable data collection. Decision trees may be vulnerable to issues with robustness, scalability adaption, and height optimization. However unlike other data classification techniques, decision trees produce an effective and understandable rule collection.

In reffered paper[5]we have seen the issue of imbalanced datasets before presenting evaluation metrics for this classification challenge, which are distinct from standard classification metrics. The issue of class imbalance has been addressed using a variety of strategies. These methods can be divided into two groups: internal methods, which develop new algorithms or alter existing ones to take the class-imbalance problem into account and external methods, which preprocess the data to lessen the impact of its class imbalance. However, cost-effective educational methods using both the data. The learning job is not hindered by skewed class distributions alone, as stated. but typically, several issues connected to this problem arise. where we display the SBAG's performance with the various datasets used in the preceding section, arranged according to the IR, to look for any noteworthy areas of excellent or bad behavior. We can see that there is no consistent pattern of behavior for.They have reviewed the topic of classification with imbalanced datasets and have concentrated on two main issues which are to present the main approaches for dealing with this problem, namely, cost-sensitive learning, ensemble techniques, and preprocessing of instances, and to develop a thorough discussion on the effect of data intrinsic characteristics in learning from imbalanced datasets.

## III. Background

In our project the main task technically refers to a classification problem, where we will put different classification methods to use and evaluate the outcomes.So, for time being we are considering the algorithms like decision tree,Random Forest,Gradient Boosting. One of the efficient techniques frequently employed in several disciplines, including machine learning, image processing, and pattern recognition, are decision trees.A prominent classification model in data mining is the decision tree. Each tree is made up of nodes and branches. Each node represents a feature in a classification category, and each subset specifies a value the node may accept. Decision trees have several implementation domains due to their straightforward analysis and accuracy on various data formats.

One more Machine learning algorithm is the Random Forest which is user-friendly,dynamic and typically yields excellent results even without hyper-parameter adjustment. As a result of its versatility and simplicity, it is also one of the most popular algorithms used in both classification and regression tasks.The ability to construct trees from subsamples of the training dataset greedily was a key discovery in bagging ensembles and random forests.With gradient boosting models, the correlation between the trees in the sequence can be decreased using the same advantage.

## IV. System Architecture

### A. Define Task

In the flow diagram the first step is defining the task which means we consider the problem statement, divide it into smaller tasks and assigning a time interval to complete the task.

### B. Collect Data

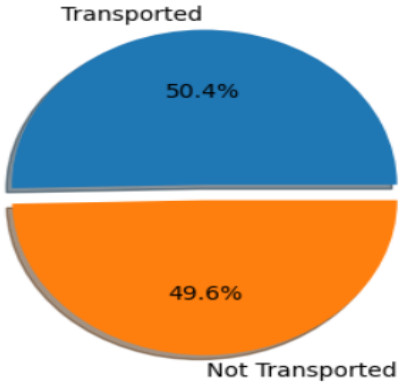
Collection of data is done from Kaggle where we have considered 2 datasets for this project such as test.csv and train.csv.We have 8693 rows and 14 columns in the training dataset and 4277 rows and 13 columns in the test dataset.

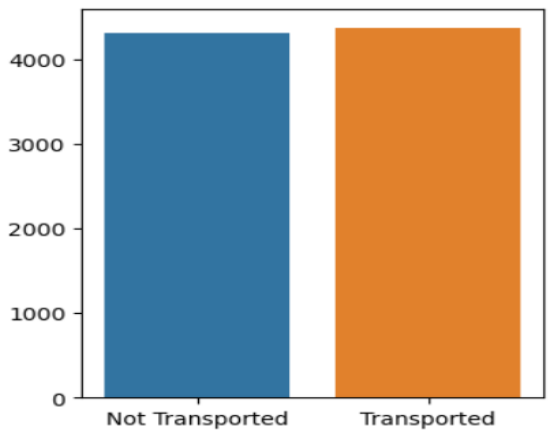
### C. Data Cleaning

Data Cleaning means the process of correcting or deleting duplicate, incomplete, or improperly formatted data from a dataset.So, here we will identify the null values, duplicate values from the given dataset and eliminate them.

### D. Exploratory Data Analysis

We are using some visualization techniques in the form of graphs like Bar plot, Boxplot, Histograms ...etc where with help of this it is very easy to identify challenges we are facing, and we can decide what are the algorithms we can consider.



### E. Model Refinement

An abstract data model is refined with data to create usable data structures. Operation refinement changes a system operation specification into an executable program (such as a procedure).

### F. Test and Evaluation

With the help of Test Evaluation (TE) ,we can assist the risk management in the stages of developing , producing and operating.We have taken 6-7 machine learning algorithms into consideration which will be discussed in the below section and implemented them to get best accuracy.

### G. Deployment

The 'Deployment' phase includes putting your model in use and in front of actual people.


Fig. 1: Architecture Diagram.

## V. Methodology

In our project we have used machine learning algorithms such as Random forest,Extra Trees, Gradient boosting, Adaptive boosting,KNN,decision tree and Logistic Regression to find the accuracy levels of each and predict the best algorithm.

### A. Random Forest

Random Forest is used for classification, regression, and other applications. It functions by building several decision trees and then integrating the output to get a more precise prediction.Due to its accuracy and robustness, it is a strong and adaptable algorithm that is frequently utilized in numerous fields.

### B. Logistic Regression

It is the algorithm applied to binary classification work, where the objective is to estimate the likelihood that an event will occur, such as whether or not a consumer might buy a product.The logistic function, which is a mathematical formula, is used in the method to convert input information to a probability value between 0 and 1.

### C. k-nearest neighbors

For classification and regression tasks, this algorithm is utilized. With the goal of trying to predict the new data point, it locates the K training data points that will be most comparable to the new data point and makes use of their labels (for classification) or values (for regression) to do so.The choice of the value of K, the distance measured metric, and the level of quality of the training data all determine the way it performs.

### D. Decision Tree

It operates by generating a model that appears like a tree and show a number of options according to what is given as input.A powerful and understood algorithm referred to as a decision tree has the ability of collecting complicated connections between features and decision boundaries.

### E. Adaptive Boosting

It performs by integrating weak learners (like decision trees) to generate an effective learner.AdaBoost is a reliable and powerful algorithm that can boost the performance of lagging learners as well as handle unbalanced datasets.

### F. Gradient Boosting

It can be utilized for developing better models for prediction by merging a number of less valuable models. It functions with the iterative process of including fresh models to the ensemble, each one aiming to address the flaws of the preceding.By implementing several weak models into a powerful one through an iterative process of correcting mistakes, it improves the accuracy of machine learning models.

### G. Extra Trees

It is used for difficulties that involve classification and regression. It functions by generating numerous decision trees, each of them using a random subset of its characteristics and training data.To make predictions, it constructs multiple decision trees with randomly splits. This the randomization minimizes over fitting and improves generalization.

## VI. Evaluation

As we have considered seven machine learning algorithms above, now we compared the accuracy of each algorithm as shown in below (fig2)and it is displayed as bar graph.From this figure we can say that the highest accuracy is acquired by Gradient Boosting which is followed by Random Forest,Extra Trees, Logistic Regression , KNN, Decision Tree, Ada Boost.

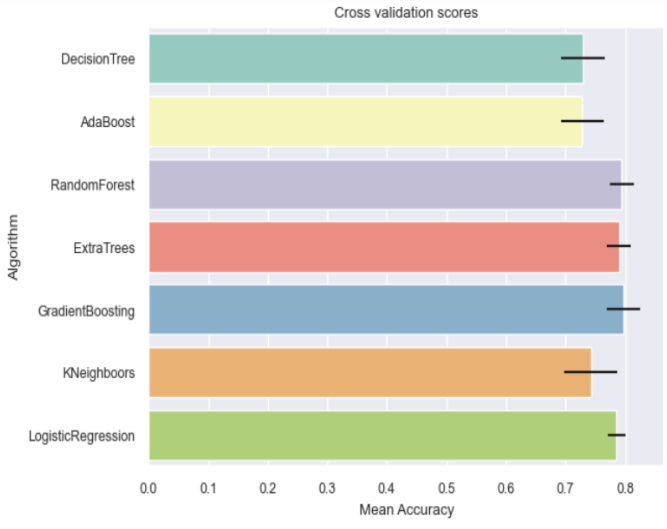
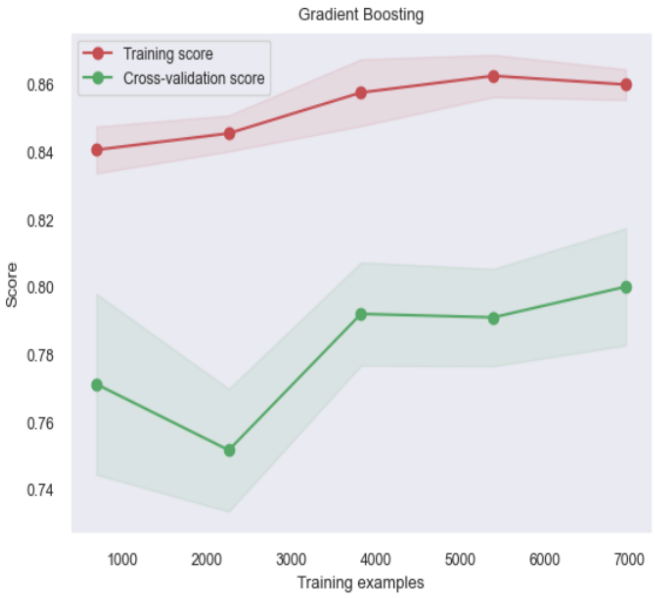
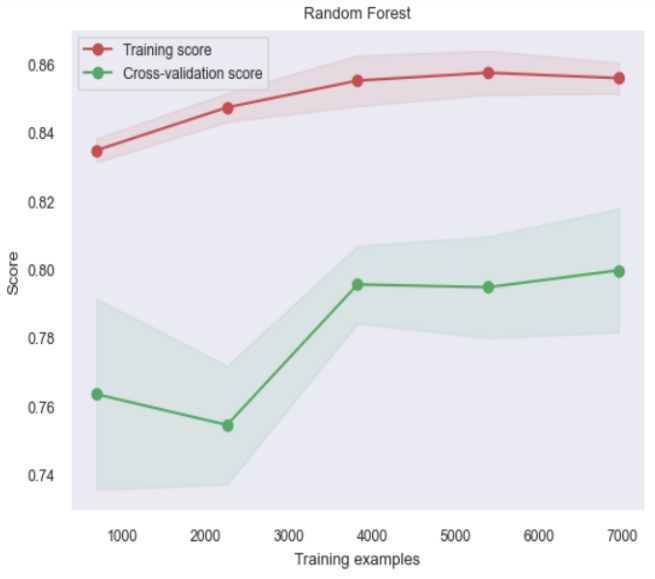


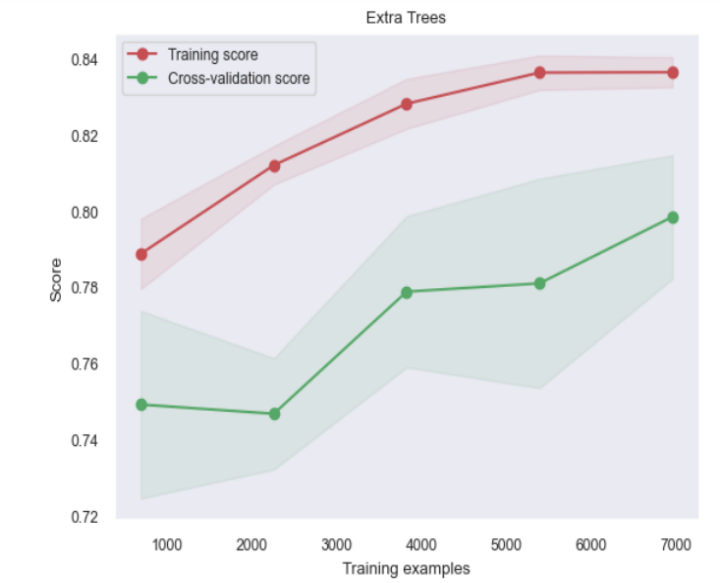

Fig. 2

| | Model | Score | Std |
|---|---|---|---|
| 4 | GradientBoosting | 0.796514 | 0.027979 |
| 2 | RandomForest | 0.793982 | 0.020305 |
| 3 | ExtraTrees | 0.789379 | 0.019352 |
| 6 | LogisticRegression | 0.785810 | 0.015385 |
| 5 | KNeighboors | 0.742340 | 0.045064 |
| 0 | DecisionTree | 0.728530 | 0.036931 |
| 1 | AdaBoost | 0.727726 | 0.036162 |

Now, we have considered the top three algorithms which are Gradient Boosting, Random Forest and Extra Trees when compared to others and performed tuning to get better accuracy.For Gradient Boosting we got 0.80 accuracy , 0.799 accuracy for Random forest and 0.797 for Extra Trees after tuning which can be seen in below figures.




## VII. Conclusion

Basically, the Spaceship Titanic was a spaceship that traveled between stars and carried passengers from one planet to another which encountered a spacetime anomaly that was concealed by a dust cloud. The collision disturbed the orbit, sending nearly half of the passengers into a different dimension.So, our main goal is to find the best suitable machine learning algorithm which provides us the better accuracy.Finally, after performing tuning we conclude that the Gradient boosting algorithm is the best suitable algorithm here as we got better accuracy(i.e. 0.80) when compared to other algorithms that we have taken into consideration.

## VIII. References

1) Devan, P., Khare, N. An efficient XGBoost–DNN-based classification model for network intrusion detection system. Neural Comput Applic 32, 12499–12514 (2020).
2) Marji, Marji Handoyo, Samingun. (2022). PERFORMANCE OF RIDGE LOGISTIC REGRESSION AND DECISION TREE IN THE BINARY CLASSIFICATION. Journal of Theoretical and Applied Information Technology. 100. 4698-4709.
3) Eunsuk Chong, Chulwoo Han, Frank C. Park,Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies,Expert Systems with Applications,Volume 83,2017,Pages 187-205,ISSN 0957-4174
4) Jijo, Bahzad Mohsin Abdulazeez, Adnan. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. Journal of Applied Science and Technology Trends. 2. 20-28.
5) Victoria López, Alberto Fernández, Salvador García, Vasile Palade, Francisco Herrera,An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics, Information Sciences,Volume 250,2013,Pages 113-141,ISSN 0020-0255