CSE445: Titanic Project Report

PREPARED FOR

Intisar Tahmid Naheen
Lecturer,
Department of Electrical and Computer Engineering

PREPARED BY

Samiur Rahman Prapon
ID: 1712666642

CSE445, Section: 4

Summer, 2021

Submission Date

22th September, 2021

Abstract

The sinking of the Titanic is one of the most iconic tragedies in history. The Titanic sank after hitting an iceberg while on its way to Southampton. This incident killed many passengers and crew members. The tragedy shocked the international community, leading to increased ship safety standards. One of the reasons for such a high death toll was a lack of lifeboats for both passengers and crew. Although luck played a part in surviving the sinking, certain persons had a higher chance of survival than others. Everyone else is prohibited, including women, children, and the upper class. We will do a predictive analysis of the kind of people who were most likely to survive the disaster in this report, as well as utilize machine learning approaches to identify which passengers survived.

Introduction

Technology's unavoidable advancement has facilitated our lives while also creating certain challenges. One of the advantages of technology is that it makes it simple to access a wide range of data when needed. However, getting the appropriate information isn't always possible. Raw data obtained just from online sources makes no sense and must be processed in order to serve an information source. In this case, feature engineering approaches and machine learning algorithms are important. The goal of this report is to use machine learning and feature engineering approaches to generate as accurate results as possible from raw and missing data. As a result, Titanic, one of the most prominent datasets in data science, is utilized. This dataset contains information about passengers aboard the Titanic, such as who survived and who did not. It was discovered that the performance of prediction was hampered by several missing and non - linear information. The impact of the characteristics has been examined for a comprehensive data analysis. As a result, some new features have been added to the dataset, while others have been deleted.

Literature Review

There are several studies in the literature that compared different classification algorithms on multiple datasets.

Meyer et al., compared SVM implementation to 16 classification algorithms and for titanic dataset they achieved 20.81% and 21.27% error rates with neural networks and SVM respectively as minimum errors. Rutsch et al. compared Adaboost classifiers to SVM and RBF classifiers. For titanic dataset, %22.4 error rate is obtained from SVM as the minimum error rate. Li et al. used SVM as a component classifier for Adaboost. They used titanic dataset as one of the experimental data and the minimum error rate they obtained is %21.8. Chatterjee applied multiple logistic regression and logistic regression to check whether a passenger is survived. He reported performance metrics across different cases comparison and concluded that, the maximum accuracy obtained from Multiple Linear Regression is 78.426%. The maximum accuracy obtained from Logistic Regression is 80.756%. Datla compared the results of Decision tree and Random Forests algorithms for Titanic dataset. Decision tree is resulted 0.84% correctly classified instances, while Random Forests resulted 0.81%. As the feature engineering steps, they created new variables such as “survived”, “child”, “new fare”, “title”, “Family size”, “Family Identity” which are not included in feature list of Titanic dataset and also replaced a missing value by the mean value of a given feature.

Methodology

A. Logistic Regression

One of the most often used algorithms for classifying binary data is Logistic Regression. LR is founded on the idea that independent variables can predict the value of a dependent variable. Observing X, the input or collection of independent variables (x_i, ..., x_k), we try to predict Y, the dependent variable. The value of Y that corresponds to the number of persons who survived (Y = 1) or did not survive (Y= -1) and is summarized by (X=x). The conditional probability follows a logistic distribution given by (P (Y = 1 |X =xi) from this definition. We need to predict Y using this function, which is known as a regression function.

B. Naïve Bayes

In machine learning applications, the Naive Bayes method, also known as an effective inductive learning algorithm, delivers efficient and rapid classification. The technique is based on the Bayes theorem, which assumes that all characteristics are independent of the class variable's value. This is the assumption of conditional independence, which holds true in real-world situations. Because of this assumption, NB performs effectively on datasets with high - dimensional and complexity.

C. Support Vector Machines

Vapnik invented SVM in 1995, which is based on the structural risk reduction concept and has high generalization ability. It is proposed to use SVM to identify an optimum separation support vectors across classes by concentrating on the support vectors. This classifier divides the training data by the greatest possible distance. By mapping data points into a high-dimensional space, SVM addresses nonlinear challenges.

D. Decision Tree

One of the most commonly used classifiers is decision trees, which have a very easy structure to construct. A decision tree is a model comprising decision and prediction nodes in a tree form. Branching is done with decision nodes, and class labels are specified via prediction nodes. C4.5 is a decision tree algorithm that uses information gain to create a decision tree from training data. C4.5 employs a divide-and-conquer strategy while creating decision trees.

E. Random Forest

Breiman and Cutler created Random Forest, a classification method that employs an array of tree predictors. For many datasets, it is one of the most accurate learning algorithms. It creates a classifier that is extremely accurate. Each tree in RF is built by bootstrapping the training data and using a randomly selected subset of attributes for each split. Splitting is done according to purity. Despite the fact that a large portion of the data is missing, this classification technique remains accuracy.

Experiments

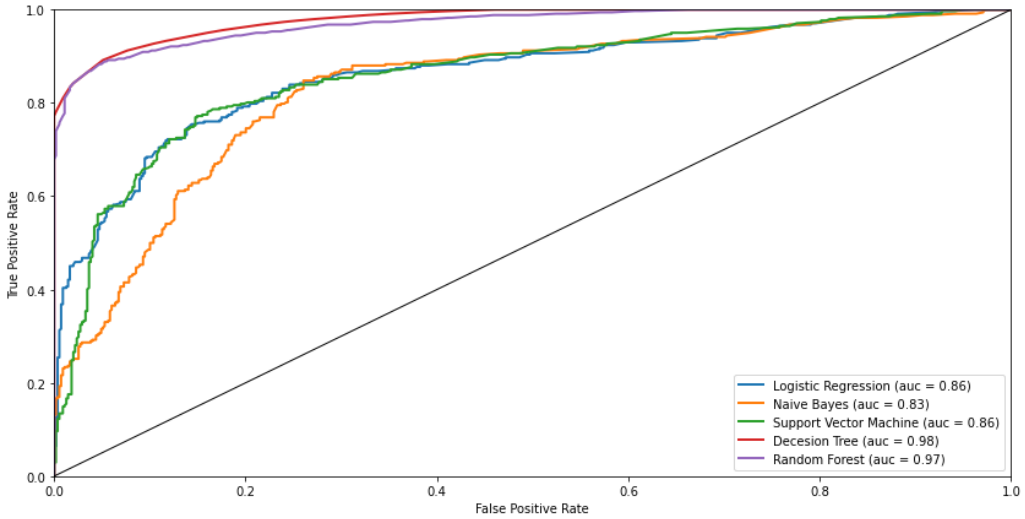
Dataset

The original data was divided into two groups: the training dataset (70 percent) and the test dataset (30 percent). Our machine learning models are built using the training set. Our objective variable, passenger survival status is included in the training set, along with other independent characteristics such as gender, class, fare, and Pclass. The test set should be used to assess how well our model works with data that has never been seen before. There isn't anything in the test set that you can use. Passengers' chances of survival We'll use our model to forecast whether or not a passenger will survive. The exam set should be used to determine how well you know your material. The model works with data that hasn't been seen before. We do not offer the ground truth for each passenger in the test set. It is our responsibility to predict.

	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	Number of siblings / spouses aboard the Titanic	
parch	Number of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

Evolution metrics

ROC Curves



Decision Tree has highest true positive rate. Higher is better. Hence, Decision tree gives more accurate result than other classifiers.

Results

All algorithms are performed in order to determine the chance of passenger and crew survival and to determine which features have a link with passenger and crew survival. When applying algorithms to the Titanic dataset, we discovered that some model parameter modifications are necessary to make the method correct.

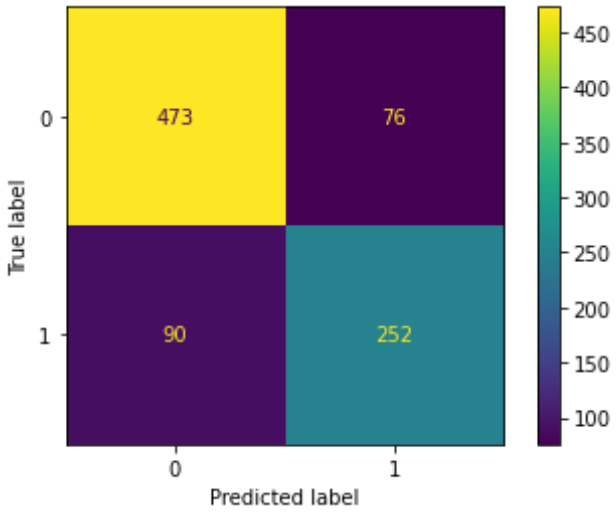
Algorithms are analyzed based on their accuracy and F1 score.

Accuracy & F1 Score

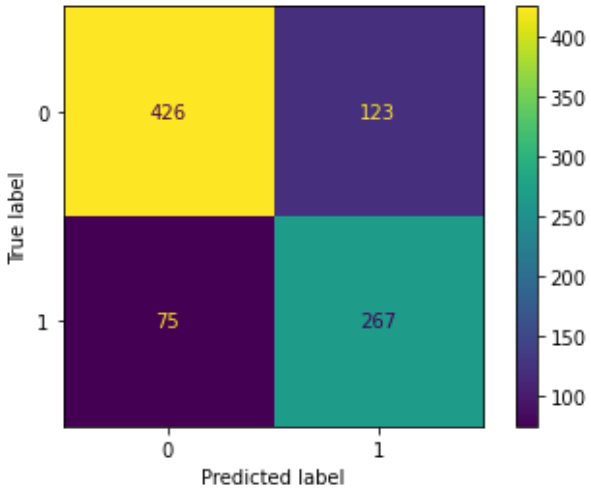
Model	F1 Score	Accuracy
Decision Tree	**0.709**	**92.37**
Random Forest	**0.7446**	**92.37**
Naive Bayes	**0.726**	**82.60**
Support Vector Machines	**0.748**	**82.60**
Logistic Regression	**0.736**	**81.37**

Confusion Matrices

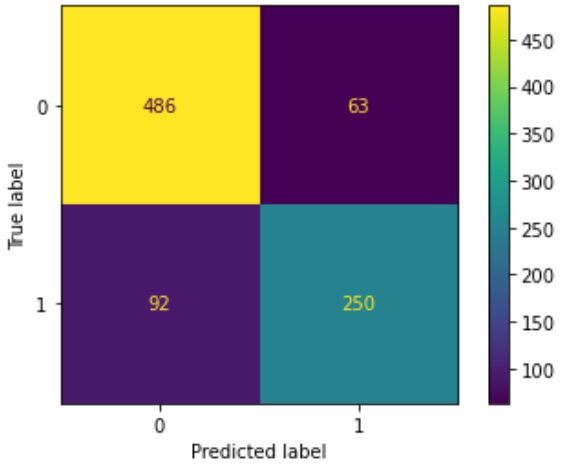
A. Logistic Regression



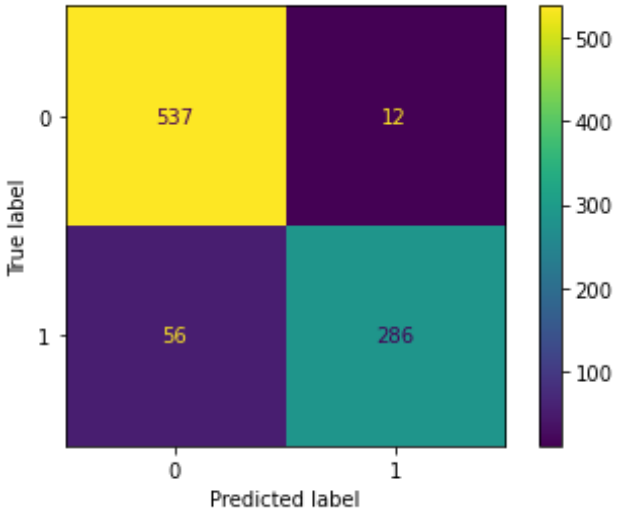
B. Naïve Bayes:



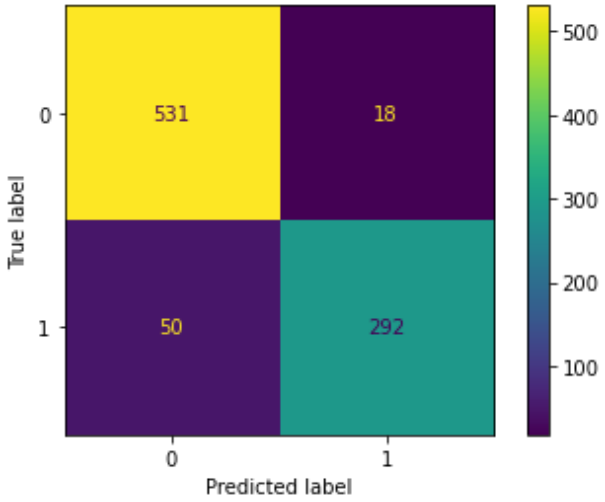
C. SVM:



D. Decision Tree:



E. Random Forest:



Conclusion

Between the five approaches we tested, there were no significant variations in accuracy. We were unable to obtain an accuracy rate that differed significantly from the classifier using simply *sex* as a feature, despite testing every combination of features. The other characteristics appeared to be only moderately predictive of survival, since sex seemed to exceed the others in terms of accuracy in predicting survival. We weren't able to make much progress even with more advanced algorithms. This demonstrates the significance of selecting relevant parameters and collecting high-quality data.