Compare support vector machines to a 3 layer neural networks with Titanic dataset

SUNNY PAMNANI
Machine learning Engineer Intern

at

AI technology & systems
Uttar Pradesh, India
Pamnani9@gmail.com

***Abstract*— Titanic disaster occurred 100 years ago on April 15, 1912, killing about 1500 passengers and crew members. The fateful incident still compel the researchers and analysts to understand what can have led to the survival of some passengers and demise of the others. With the use of machine learning methods and a dataset consisting of 891 rows in the train set and 418 rows in the test set, the research attempts to determine the correlation between factors such as age, sex, passenger class, fare etc. to the chance of survival of the passengers. These factors may or may not have impacted the survival rates of the passengers. In this research paper, comparing machine learning algorithm like support vector machines w.r.t 3-layer neural networks have been implemented to predict the survival of passengers. In particular, this research work compares the algorithm on the basis of the percentage of accuracy on a test dataset.**

Keywords—support vector machine, neural network, prediction, python, classification, Confusion matrix, Machine learning.

I. INTRODUCTION

The field of machine learning has allowed analysts to uncover insights from historical data and past events. Titanic disaster is one of the most famous shipwrecks in the world history. Titanic was a British cruise liner that sank in the North Atlantic Ocean a few hours after colliding with an iceberg. While there are facts available to support the cause of the shipwreck, there are various speculations regarding the survival rate of passengers in the Titanic disaster. Over the years, data of survived as well as deceased passengers has been collected. The dataset is publically available on a website called Kaggle.com [1]. This dataset has been studied and analyzed using machine learning algorithm like SVM and MLP Neural networks etc. Various languages and tools are used to implement these algorithms are Python. Our approach is centered on Python for executing algorithms- support vector Machines and MLP neural network. The prime objective of the research is to analyze Titanic disaster to determine a correlation between the survival of passengers and characteristics of the passengers using machine learning algorithm.

In particular, we will compare the algorithms on the basis of the percentage of accuracy on a test dataset.

II. DATASET

The dataset we use for our paper was provided by the Kaggle website. The data consists of 891 rows in the train set which is a passenger sample with their associated labels [1]. For each passenger, we were also provided with the name of the passenger, sex, age, his or her passenger class, number of siblings or spouse on board, number of parents or children aboard, cabin, ticket number, fare of the ticket and embarkation. The data is in the form of a CSV (Comma Separated Value) file. For the test data, we were given a sample of 418 passengers in the same CSV format. The structure of the dataset with a sample row has been listed in the three tables below:

Table I: Kaggle Dataset

Passenger Id	Survived	Pclass	Name
1	0	3	Braund, Mr. Owen Harris
2	1	1	Cumings, Mrs. John Bradley

TABLE II: KAGGLE DATASET (CONTD.)

sex	Age	Sibsp	Parch	Ticket	Fare	Cabin	Embarked
male	22.0	1	0	A/5 21171	7.2500	NaN	S

TABLE III: ATTRIBUTES IN TRAINING DATASET

Attributes	Description
passenger ID	Identification no. of the Passengers.

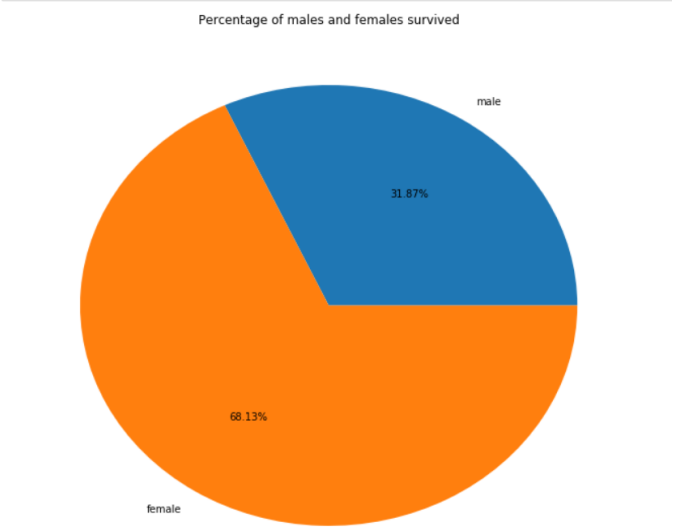
Name	Name of the passengers
Sex	Gender of the passengers (male or female)
Age	Age of the passenger
SibSp	Number of siblings or spouse on
the ship Parch	Number of or parents children on
the ship Ticket	Ticket number
Fare	Price of the ticket
Cabin	Cabin number of the passenger
Embarked	Port of embarkation (Cherbourg, Queenstown or Southampton)
Survived	Target variable (values 0 for perished and 1 for survived)
Pclass	Passenger class (1, 2 or 3)

III. INSIGHTS OF DATA

determine what all factors or attributes can prove beneficial while creating the classifier for prediction. We started with few X-Y generic plots to get an overall idea for each attribute. Few generic plots have been shown
below

Here 0 is for perished and 1 for survived

Before building a model, we explored the dataset to From the age plot in Figure below we came to a conclusion that maximum or majority of the passengers belonged to the age group of 20-40.



Similarly, we plotted a graph and performed some calculations for the sex attribute and found out that the survival rate of the female is 25.67% higher to that of the male. Similarly, we explored each of the attribute to extract those attributes or features which we will use later for prediction. We also explored the dataset to determine the number of people survived vs. number of people who could not survive. From the pie graph, it is clear that the number of people who survived is less

than the number of people who could not survive

IV. MISSING VALUES AND OUTLIERS TREATMENT

There are null values in the train and test dataset for Age, cabin, embarked, so it is replaced with mean or median or mode of particular column.

We performed some data cleaning in order to deal with the missing values. We saw that the dataset is not complete. There are various rows for which one or more fields are marked empty (especially age and

cabin). We think that age could an important attribute to predict the survival of passengers. So we used a technique to replace the NAs in the age column. The gender column has been changed to 0 and 1 (0 for male and 1 for female) to fit the prediction model in a better manner. We also introduced some new variables into the dataset to predict the survival more closely. And there are no outliers present in the dataset.

V. RELATED WORK

Many researchers have worked on the Titanic problem in order to compare various different machine learning techniques in terms of the efficiency of the algorithm to predict the survival of the passengers. Studies have tried to trade-off between different features of the available dataset to provide the best prediction results. Lam and Tang et al. used the Titanic problem to compare and contrast between three algorithms- Naïve Bayes, Decision tree analysis and SVM. They concluded that sex was the most dominant feature in accurately predicting the survival. They also suggested that choosing important features for obtaining better results is important. There were no significant differences in accuracy between the three methods they used [2]. Shawn Cicoria and John Sherlock et al. performed to suggest that sex is the most important feature as compared to other features in determining the likelihood of the survival of passengers [3]. Kunal Vyas and Lin et al. suggested that dimensionality reduction and playing more with the dataset could improve the accuracy of the algorithms. The most important conclusion provided by them was that more features utilized in the models do not necessarily make results better [4]. Although many researchers have worked hard to determine the actual cause of the survival of some passengers and demise of others, we attempt to get better results and accuracy by utilizing various different combination of features and different machine learning methods.

VI. METHODOLGY.

The first step in our methodology is to clean the training dataset. We start with exploring different attributes for NA values. We see that age column has 177 rows with NA values and cabin 687 rows with NA values. As most of the data in the cabin column is missing, we decided to drop this column from our analysis. We assume that age is a very important attribute. Hence, we decided to keep the age column for the analysis. We attempt to establish a relationship between the title of the passengers and their age. We believe that Ms. A is younger than Mrs. A and we also assume that the people having same titles are closer in age. Titles of the passengers have been extracted form the name of the passengers and we have replaced the name column with the extracted titles. The missing entries has been replaced by the average age of the particular titlegroup i.e. if there is a missing age value for a woman with title Mrs. then the missing value gets replaced with the average age of all the women with title Mrs.

In the past marine disasters, the policy of Women Children First (WCF) has been used by the crew members giving women and children survival advantage over men [5]. Based on this social norm we decided to introduce some new attributes to strengthen our dataset and improve our analysis These attributes are listed in the table below:

VII. SPLITTING THE TRAINING DATA

So by dividing the train data into training, validation data into 80:20 ratios. After that we predicts the model using SVM and Neural Network Model.

A. SUPPORT VECTOR MACHINES

Support Vector Machines' is like a sharp knife – it works on smaller datasets, but on them, it can be much stronger and powerful in building models Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in ndimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane
that differentiate the two classes very well

Classification accuracy was obtained by comparing the predictions to the class values of the test data [6]. We got a classification accuracy of 83.15%.

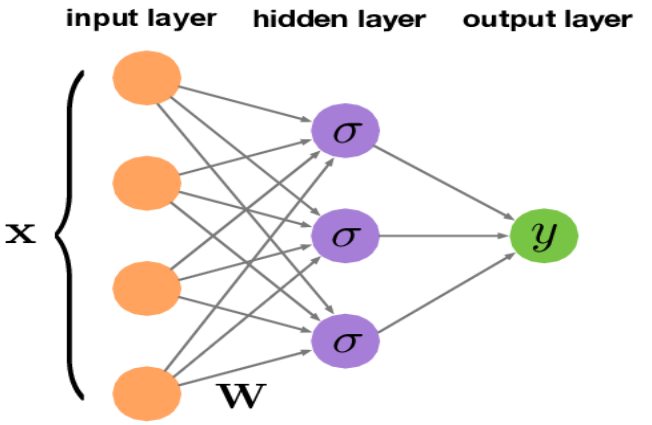
B. 3-layer neural network (using Keras Sequential Model with Relu and sigmoid Activation function on dense layers)

In this I have used Keras to make a 3 layer neural networks with all the dense layers.

I have used Relu and Sigmoid Activation function to make it good and I hav used Adam Optimizer for making loss very much optimized.

I have done the batch size of 10 and I have used 150 epochs for that .

Finally, the accuracy was 79.15% o the test data.



output.

Algorithm		Test accuracy
SVM		83.15%
Using Adam Optimizer and Relu		79.13%

VIII. Results.

[2] 2. Eric Lam, Chongxuan Tang. Titanic – Machine LearningFromDisaster.AvailableFTP: cs229.stanford.edu Directory: proj2012 File: LamTang-TitanicMachineLearningFromDisaster.pdf

[3] Analyzing Titanic disaster using machine learning algorithmsComputing, Communication and Automation (ICCCA), 2017 International Conference on 21 December 2017, IEEE.

[4] [2] Eric Lam, Chongxuan Tang, "Titanic Machine Learning From Disaster", LamTang-Titanic Machine Learning From Disaster, 2012.

[5] 3. Cicoria, S., Sherlock, J., Muniswamaiah, M. and Clarke, L, "Classification of Titanic Passenger Data and Chances of Surviving the Disaster," pp. 4-6, May 2014

[6] 4. Vyas, K., Zheng, Z. and Li, L, "Titanic-Machine Learning From Disaster," pp. 6, 2015.

[7] 5. Mikhael Elinder. (2012). Gender, social norms, and survival in maritime disasters [Online]. Available: http://www.pnas.org/content/109/33/13220.full.

[8] Corinna Cortes, Vlasdimir Vapnik, "Support-vector networks", Machine Learning, Volume 20, Issue 3,pp 273-297.

Aaron Whitley, 2015.

[11] [9] Kunal Vyas, Zeshi Zheng, Lin Li, Titanic- Machine Learning From Disaster- 2015.

[9] Cortes, Corinna; and Vapnik, Vladimir N.; "SupportVector Networks", Machine Learning, 20, 1995.

[10] MICHAEL AARON WHITLEY, Using statistical learning to predict survival of passengers on the RMS Titanic by Michael

So finally w conclude that SVM is predicting good when compare to the Sequential keras model using adam optimizer and relu and sigmoid activation function.

REFERENCES

.

[1] 1. Kaggle, Titanic: Machine Learning Form Disaster [Online]. Available: http://www.kaggle.com/

[12] [10] EECS 349 Titanic- Machine Learning From Disaster, Xiaodong Yang, Northwestern University.

[13] [11] Prediction of Survivors in Titanic Dataset: A Comparitive Study using Machine Learning Algorithms, Tryambak Chatterlee, IJERMT-2017.

[14] https://en.wikipedia.org/wiki/Multilayer_perceptron