Compare Support Vector Machine to a 3 Layer Neural Networks on the Titanic Dataset

Sumit Singh
Machine Learning Intern
AI Technology and Systems
sumitssheoran@gmail.com
www.ai-techsystems.com

Abstract- Titanic disaster occurred about 100 years ago on April 15, 1912, killing more than 1500 passengers and crew members during its maiden voyage. After so many years, this accident still compels researchers and analysts to understand what could have led to the survival of some passengers while demise of others. In this paper two Machine Learning Algorithms are used to compare the accuracies of the models-

(i.) Support Vector Machine

(ii.) 3 Layer Neural Networks

Keywords- support vector machine, neural networks,dataset,classification, accuracy,confusion matrix.

1. INTRODUCTION

The dataset used for the paper is provided by the Kaggle website. The data consists of two CSV files- train set and test set. Train set contains information of 891 passengers.For each passenger, 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, embarkation and survival are provided. For the test set, a sample of 418 passengers is provided and asked to predict their survival. The list of all the attributes is given in a tabular form below-

Attributes	Description
PassengerID	Identification no. of the passengers.
Pclass	Passenger class (1, 2 or 3)
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 parents or 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)

2. DATA PREPROCESSING

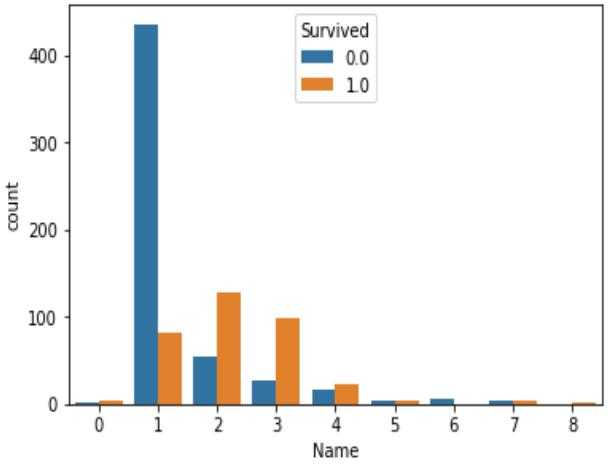
The dataset provided to us is in raw form and there are high chances that it may contain incorrect datatypes, NaN values and unexpected symbols. Before moving forward, the dataset is cleaned column-wise.For the sake of simplicity, the training set rows and test set rows are concatenated.

⇔Name

A person's chances of survival can't be predicted using his/her name but after observing closely, the designations were separated from their names and categorised as follows-

- Mr.=1
- Miss =2
- Mrs.=3
- Master =4
- Dr.=5
- Rev =6
- Major, Capt, Col, Mlle =7
- Lady, Ms =8
- The countess,Sir, Mme, Jonkheer, Don=0

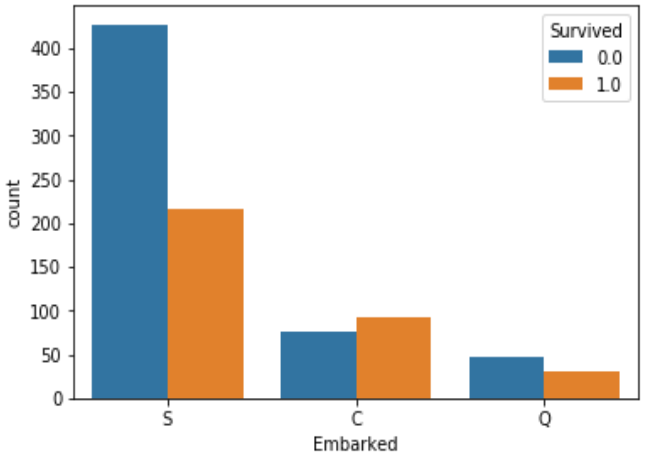
Thereafter a graph was plotted between the survivors and their names-



As it can be clearly seen that the count of children and women which survived are more as compared to the men.

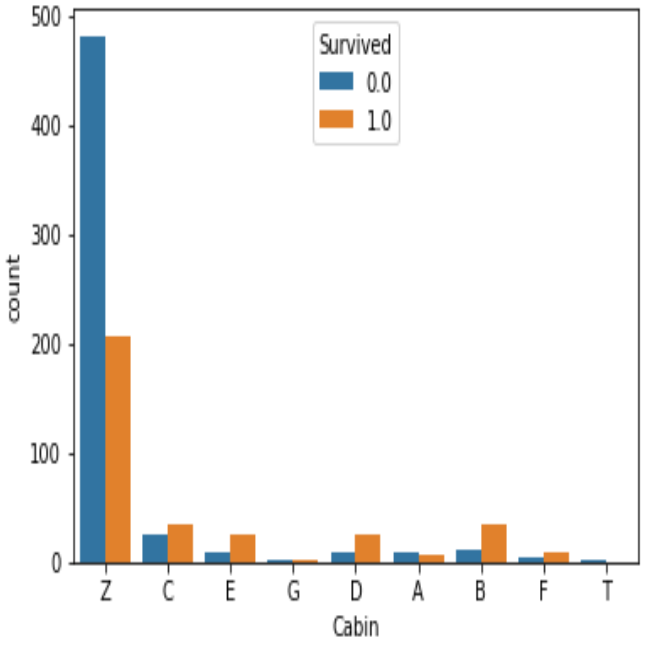
⇔ Embarked

The passengers embarked on the ship from three places namely-Cherbourg(C), Queenstown(Q), Southampton(S). The passengers from Cherbourg survived more than the other two and thus it is used as an important aspect in this paper.



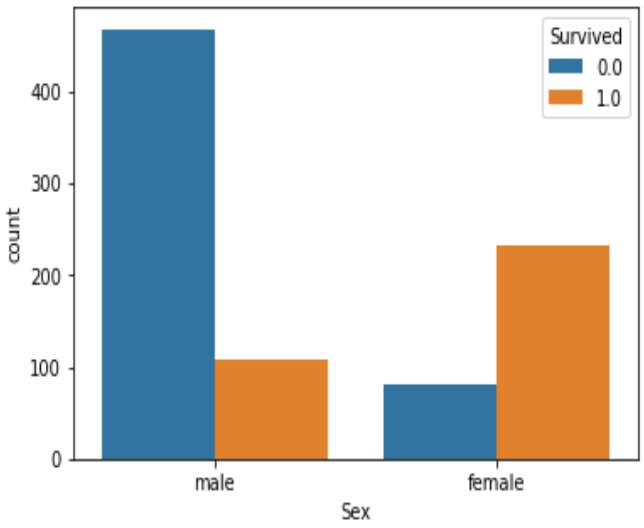
⇔ Cabin

The ship had a total of 8 cabins. This predictor contains the highest number of NA values. For simplicity, all the unknowns are filled with "Z". The passengers of cabin C,E,D and B survived more than any other cabin.



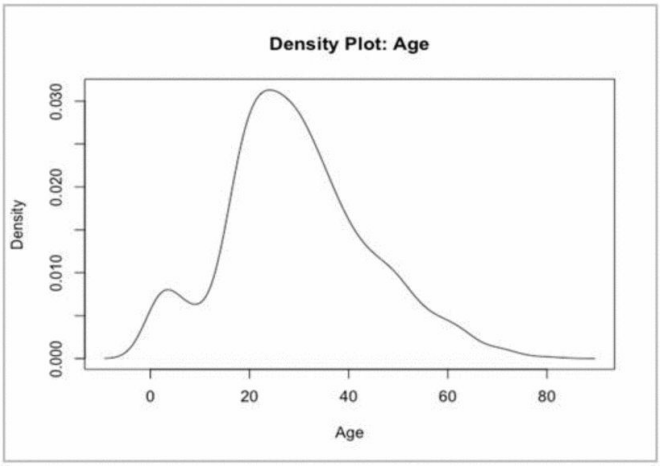
⇔Sex

There were more male passengers as compared to female passengers. But in case of survival, graph showed that more females survived.



⇔Age

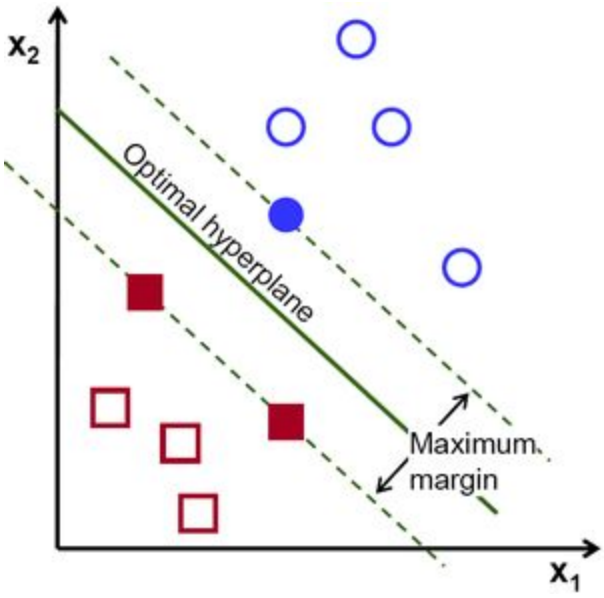
The ship had passengers of age between <1 year and 80 years.The maximum people were of age between 18 years to 40 years as shown in the density plot-



3. MODELS

⇔ SUPPORT VECTOR MACHINE

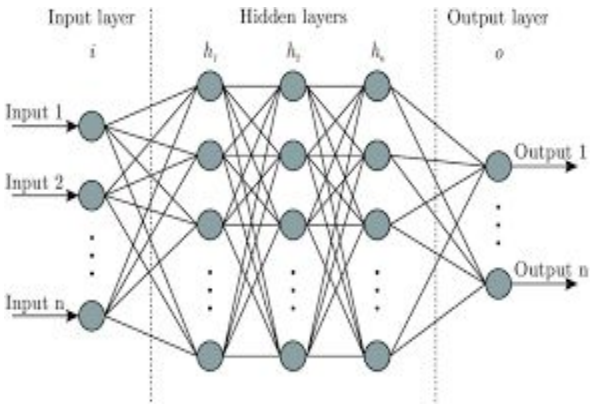
Support Vector Machine, abbreviated as SVM is a simple algorithm that can be used for both regression and classification tasks. The objective of svm algorithm is to find a hyperplane in an N-dimensional space (N-Number of features) that distinctly classifies the data points.To separate the two classes of data points, there are many possible hyperplanes that could be chosen. Our objective is to find a plane that has the maximum margin, i.e the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence.



⇔ 3-LAYER NEURAL NETWORKS

An artificial neuron network (ANN) is a computational model based on the structure and functions of biological neural networks.ANNs are considered nonlinear statistical data modeling tools where the complex relationships between inputs and outputs are modeled or patterns are found.ANNs have three layers that are

interconnected. The first layer consists of input neurons. Those neurons send data on to the second layer, which in turn sends the output neurons to the third layer.The whole process can be explained using the image below-



4. RESULTS

For the SVM classifier 'rbf' kernel was used and the accuracy obtained from the model using confusion matrix was 84.511%. For the detailed observation, confusion matrix is provided below-

Predicted↔ Actual ↕	Not Survived	Survived
Not Survived	495	54
Survived	84	258

For the 3-Layer Neural Networks setting 'uniform' as kernel_initializer and 'relu' as activation function, the model was compiled using 'adam' optimizer. The dataset was then fitted using a batch size of 10 and epochs was set to 1000. The accuracy obtained was slightly higher than the SVM as 84.96%. For detailed observation, confusion matrix is provided below-

Predicted↔ Actual ↕	Not Survived	Survived
Not Survived	507	42
Survived	92	250

5. CONCLUSIONS

After performing Exploratory Data Analysis on the Titanic Dataset and applying both the algorithms, it was observed that 3-Layer Neural Network outperformed SVM by a slight margin.

6.REFERENCE

[1.] https://www.kaggle.com/c/titanic
[2.]https://ieeexplore.ieee.org/abstract/document/8229835/figures#figures
[3.]https://www.techopedia.com/definition/5967/artificial-neural-network-ann
[4.]http://www.statsoft.com/textbook/support-vector-machines