Compare Support Vector Machines to a 3 layer Neural Networks on the titanic dataset

Pragnya Konakalla
AITS(ai-techsystems.com)
Mumbai,India
pragnyak09@gmail.com

***Abstract*—Titanic disaster occurred 100 years ago on April 15,1912,killing about 1500passengers and crew members is one of the deadliest maritime disasters in history. Although it has been many years, this fateful incident still compel the researchers and analysts to understand what could 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 study, I propose to compare Support Vector Machine(SVM) to a 3 layer Neural Networks on the Titanic dataset, which is publicly available, to analyze likelihood of survival and learn what features have a correlation towards survival of passengers and crew.**

Keywords—Titanic, Prediction, Classification, SVM, Artificial Neural Network

I. INTRODUCTION

Titanic disaster is one of the most infamous shipwrecks in the history. During her maiden voyage en route to New York City from England, she sank killing 1500 passengers and crew on board. Various information about the passengers was summed up to form a database, which is available as a dataset at Kaggle platform[1]. The historical data has been split into two groups, a 'training set' and a 'test set'. For the training set, we are provided with the outcome (whether or not a passenger survived). . The data analysis will then be done and then this set is used to build the model using SVM and ANN to generate predictions for the test set.The prediction outcomes will be checked for accuracy. The accuracy will then be compared in order to suggest the better performing algorithm.

II. DATASET

The dataset used in this paper is taken from the Kaggle website.The Titanic dataset consist of a training set that includes 891 passengers and a test set that includes 418 passengers which are different from the passengers in training set. A description of the features is given in Table I.

Feature	Description	Characteristic
PassengerID	Identification no. of the Passenger	Integer
Pclass	Passenger class (1,2 or 3)	Integer
Name	Name of Passenger	Object
Sex	Gender of Passenger(Male/Female)	Object
Age	Age of Passenger	Real
SibSp	Number of siblings or spouse on the ship	Integer
Parch	Numbr of parents or children on the ship	Integer
Ticket	Ticket Number	Object
Fare	Price of the Ticket	Real
Cabin	Cabin number of the passenger	Object
Embarked	Port of embarkation (Cherbourg, Queenstown or Southampton)	Object
Survived	Target variable (values 0 for perished and 1 for survived)	Integer

III. METHODOLOGY

A.Data Preprocessing

The data we collected is still raw-data which is very likely to contains mistakes ,missing values and corrupt values. Before drawing any conclusions from the data we need to do some data preprocessing which involves data wrangling and feature engineering . Data wrangling is the process of cleaning and unify the messy and complex data sets for easy access and analysis .Feature engineering[2] process attempts to create additional relevant features from existing raw features in the data and to increase the predictive power of learning algorithms.

While the features such as PassengerId, Survived, Pclass, Age, SibSp, Parch and Fare are numeric values, Name, Sex and Embarked can take nominal values; the features such as Ticket, Cabin can take numeric and nominal values.

The missing values of Age and Fare features are filled by median values of these features. The missing values of Embarked and Cabin feature are filled by “C” and “Z” value repectively.

For a detailed feature engineering we first analyzed the features.

1)Sex: When we consider the distribution of the “Sex” feature, there are 314 female and 577 male passengers. 233 of female passengers have been rescued and others have lost their lives. On the other hand, 109 of male passengers have been rescued and others have lost their

lives. If we analyze these distributions it is realized that the survival rate of women is higher than that of men. It has been concluded that the effect of this feature on predicting the class label is significant. The gender column has been changed to 0 and 1(0 for male and 1 for female).

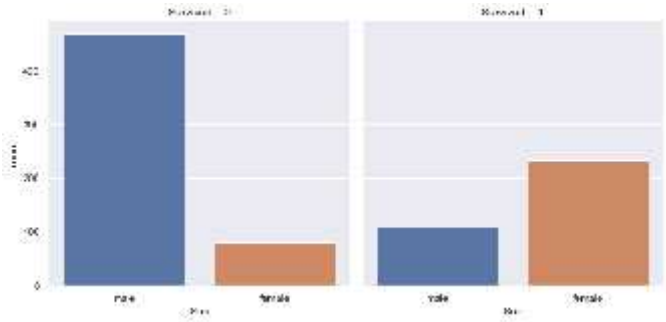


Fig.1 Distribution of gender feature

2)Pclass: Pclass feature describes three different classes of passengers. There are 216 passengers belong to the class 1, 184 passengers in class2 and finally 491 passengers in class 3.The survival rates of passengers due to Pclass feature are given in Fig. 2. The passengers with the highest survival rates are the first class passengers with 63%. This ratio also shows that wealthy people are alive.

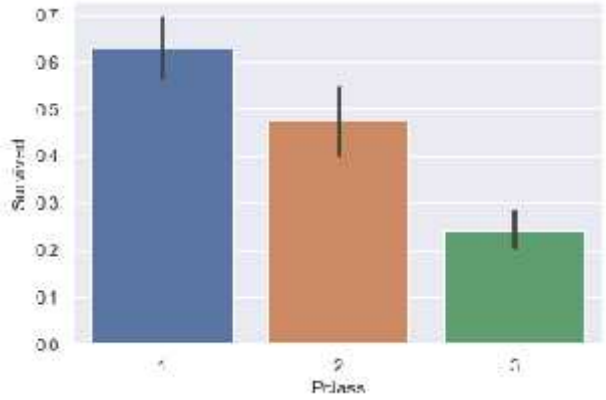


Fig.2 Pclass vs Survival

3) Embarked: When we consider the distribution of the Embarked feature, there are 644, 168, 77 passengers boarding from the port "S", "C" and "Q" on the ship respectively. The survival rates of passengers boarding from these ports are given in Fig. 3. When this figure is analyzed, C is the port with the highest survival rate of 55%. Thus, this can be interpreted like Embarked feature gives important clues about survival.

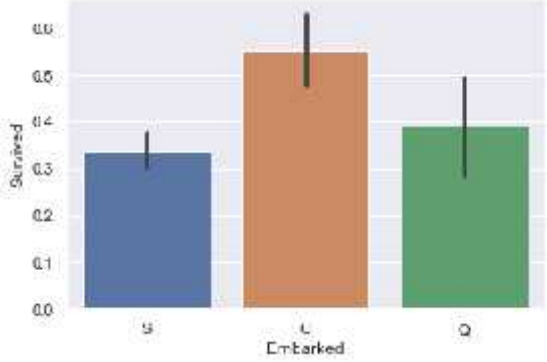


Fig.3 Embarked Vs Survival

In the dataset S,C,Q are mapped to 3,1,2 respectively that is in the increasing order of survival.

4)*Pclass,Embarked,Gender:*On further study of the relation between Embarked and Pclass it is found that port C has higher number of class 1 passengers therefore causing it to have more survival rate and on the contrary port S has most number of class 3 passengers therefore having least survival rate.Also number of male passengers are greater than female passengers in port S which may also be a factor of less survival rate.

5) Age: When the "age" feature" is considered it is seen that, the age of passengers are range from 0 to 80. If we group the passengers by specific age ranges such as 0-16, 17-40 , 41-60 and >60 then we realized that most of the passengers in the 0-16 age group are survived and a large majority of passengers in the age group 61-80 lost their lives. This statistical information proves that the first children were rescued when the ship started to sink.

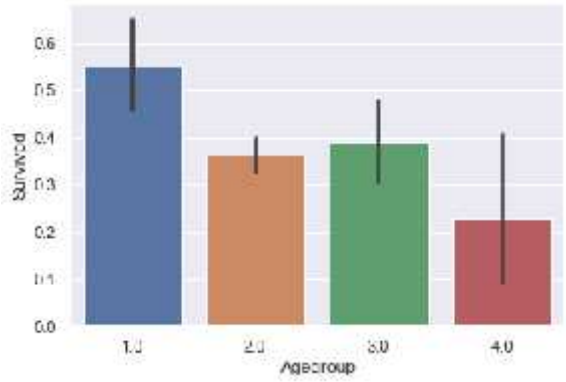


Fig.4 Agegroup vs Survival

Agegroup 1(0-16),2(17-40),3(41-60),4(>60) is mapped in increasing order of survival as 1,3,2,4 respectively.

5) Fare: The "fare" feature specifies the fare paid by the passenger and it changes between 0-512.Fig.5 shows the relation between fare and survival.

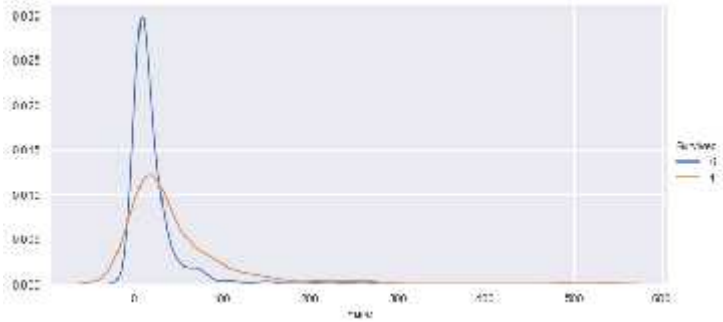


Fig 5 Fare vs Survival

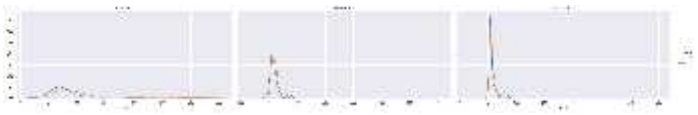


Fig 6. Pclass vs Fare vs Survival

From a glance, it appears to be that the lower fare has lower survival rate.But when broke down to each Pclass, the pattern is no longer obvious.Therefore Fare by itself is not

really carry in much information about survival,the information is mostly contained within Pclass.Therefore Fare is not included in the model.

*6)Family size:*In machine learning applications, features extension methods as well as feature reduction methods can also improve the classification performance. In this study, a feature named “Family_size” is created in addition to the existing features. This feature is calculated by adding the value of Sibsp feature to the value of Parch feature.After that, we have distinguished this feature with 3 groups.The first group consist of passengers whose family_size is 0 ,the second group consist of passengers whose family_size is 1-3,the third group consists of passengers whose family size is greater than 4. It is observed that most of the first group lost their lives and the majority of the second group is survived and there is a huge decrease in survival rate for group 3. These results show that the number of family members strengthens the possibility of survival.

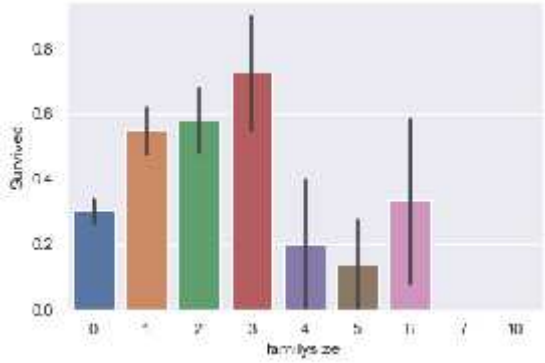
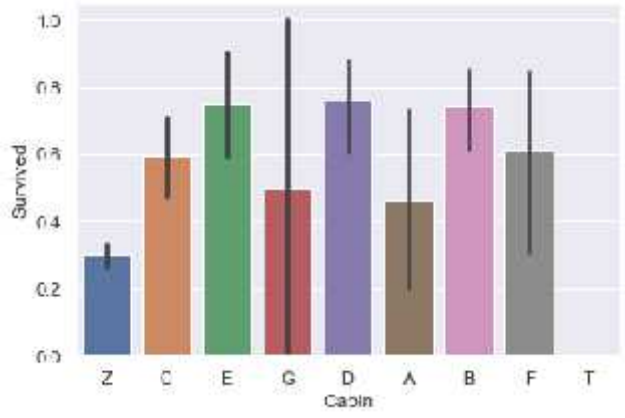


Fig 7 Family size vs Survival

7)Cabin: The Cabin names starts with a Letter. The Initial letter of each Cabin is taken and then a graph between cabin and survival is plotted.



Cabin is categorized in 4 groups :

1)E,B,D are mapped to 1

2)C,F are mapped to 2

3)A,G,are mapped to 3

4)Z ,T are mapped to 4

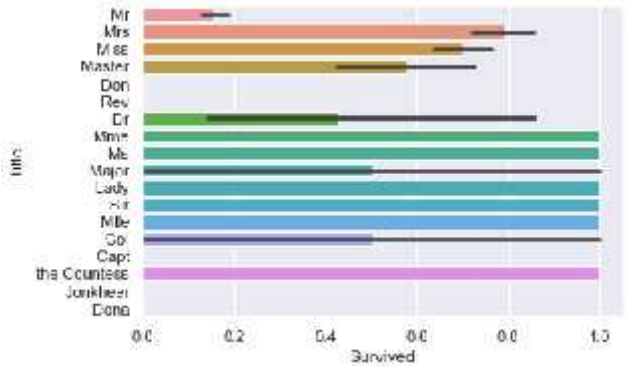
*8)Name:*The name of the passenger does not affect the survival but the title does.The title is extracted from the name and categorized in increasing order of survival as the follows:

1)'Mme','Ms','Lady','Sir','Mlle','the Countess' are mapped to 1

2) 'Mrs','Miss' are mapped to 2

3) 'Mr','Master','Dr','Major','Col' are mapped to 3

4) 'Don','Rev','Capt','Jonkheer','Dona' are mapped to 4



The PassengerId, Name , Ticket and all other unwanted features are removed from the feature set.

B.Support Vector Machine

SVMs (Support Vector Machines) are a useful technique for data classification. The foundations of Support Vector Machines have been developed by Vapnik (1995) and are gaining popularity due to many attractive features, and promising empirical performance. The SVM belongs to a class of machine learning algorithms that are based on linear classifiers and the “kernel trick”. A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. This hyperplane separates the training data by a maximal margin. SVM solves nonlinear problems by mapping the data points into a high-dimensional space.

SVMs offer advantages over other types of classifiers. SVMs are free of the optimization headaches of neural networks because they present a convex programming problem. It guarantees for finding a global solution. These classifiers are much faster to evaluate than density estimators, as they make use of only the relevant data points, instead of looping over each point regardless of its relevance to the decision boundary [3,4,5].

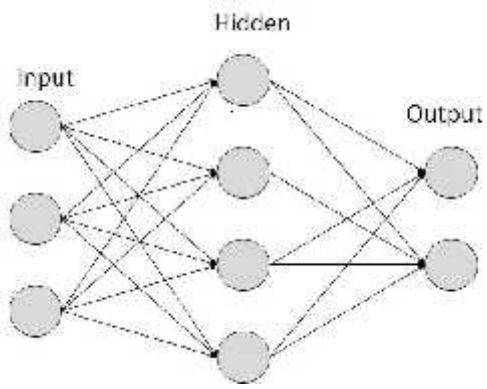
The model was built with all the variables of our cleaned dataset,that are Pclass,Age,Embarked,FamilySize,Cabin,Title and Gender.In this model I used the Radial basis function kernel(RBF) or Gaussian Kernel. Grid Search was used for hyperparameter tuning.The regularization parameter was set to 3 and degree was set to 1.

C.Artificial Neural Network

The idea of ANNs is based on the belief that working of human brain by making the right connections,can be imitated using silicon and wires as living neurons and dendrites. ANNs are composed of multiple nodes, which imitate biological neurons of human brain. The neurons are connected by links and they interact with each other. The

nodes can take input data and perform simple operations on the data. The result of these operations is passed to other neurons. The output at each node is called its activation or node value.

Each link is associated with weight. ANNs are capable of learning, which takes place by altering weight values. The following illustration shows a simple ANN −



For the titanic dataset an Artificial Neural Network with two hidden layers is used to compare the results and accuracy.The output is extracted in Binary format i.e 1s(survived) and 0s(deceased).The data is preprocessed. The final training dataset consists of 7 features that are Pclass, Age, Embarked, FamilySize, Cabin, Title and Gender. The ANN model consists of 2 hidden layers with 4 and 3 features each. Rectified Linear Unit(ReLU) is used as an activation function in both hidden layers.

Due to small amount of data there are chances that the neural network will converge at the local minima and not the global minima in the loss function. This can be overcome by optimizing the model, with increase in the momentum or increasing the learning rate slightly.

In this model, the loss function was stuck at the local minima even after playing with the learning rate. The optimizer that I used was an adaptive optimizer called AdamOptimizer. It works on the principle of tuning the hyperparameter(learning rate). It uses a large step size and the algorithm will converge to the minima without fine tuning. But the downside of the AdamOptimizer is that it requires large computation and to be performed on each parameter in the training steps.

IV. Result

For the purpose of this research paper, the training data from the Kaggle website will be divided into two parts,80% for training and creating the model and 20% for predicting.The testing data from the Kaggle website will not be used.

Confusion matrix for ANN model

Predicted	Survived:NO	Survived:YES
Survived:NO	92	18
Survived:YES	17	52

From the above confusion matrix we can see that out of total 179 predictions, this model made 144 correct predictions,giving an accuracy of 80.48%

Confusion matrix for SVM model

Predicted	Survived:NO	Survived:YES
Survived:NO	96	14
Survived:YES	18	51

From the above confusion matrix we can see that out of total 179 predictions,this model made 147 correct predictions,giving an accuracy of 82.1%.

V. Conclusion

It is observed that Support Vector Machine gives a slightly higher accuracy than Artificial Neural Network.The most significant features for the prediction of survival are Pclass,Gender,Age,Family Size,Title,Embarkment,Cabin.

It would be interesting to play more with dataset and introduce more features which might lead to good results.Various other machine learning techniques like Logistic Regression,Random Forest can be used to solve the problem.

References

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

[2] https://machinelearningmastery.com/discover-feature-engineering-how-to-engineer-features-and-how-to-get-good-at-it/

[3] S N Sivanandam, S Sumathi and S N Deepa, “Introduction to Neural Networks”,sixteenth edition, 2012, McGraw hill.

[4] Aprajita Sharma, Ram Nivas Giri, “Automatic Recognition of Parkinson’s diseases via Artificial neural Network and Support vector machine”, IJITEE, 2278-3075, Vol-4, Issue-3, Aug2014

[5] Damodhar Shahare, Ram Nivas Giri “Comparative Analysis of Artificial Neural Network and Support Vector Machine Classification for Breast Cancer Detection ,” International Research Journal of Engineering and Technology (IRJET)

[6] Ekin Ekinci, Sevinç İlhan Omurca, Neytullah Acun “Title A Comparative Study on Machine Learning Techniques using Titanic Dataset ,” 7th International Conference on Advanced Technologies (ICAT'18)

[7] Dr. Prabha Shreeraj Nair "Analyzing Titanic Disaster using Machine Learning Algorithms" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-2 | Issue-1, December 2017, pp.410-416, URL: https://www.ijtsrd.com/papers/ijtsrd7003.pdfM. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.