Compare Support vector machine to three layer Neural Network on Titanic dataset

Rituraj Singh

Machine learning Intern

AI-Tech Systems

Ai-techsystems.com

Rituraj.nitrkl@gmail.com

Abstract —Titanic disaster occurred 100 years ago on April 15, 1912, killing about 1500 passengers and crew members. With the use of machine learning methods and a dataset provided by Kaggle consisting of 891 rows in the train set and 418 rows in the test set, we attempt to determine the correlation between factors such as age, sex, passenger class, fare etc. to the chance of survival of the passengers. In particular, compare two different machine learning technique SVM and neural network

Keywords — Machine learning, Support Vector Machine, Neural Network.

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. There was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class. Using data provided on https://www.kaggle.com/c/titanic our goal is to apply machine-learning techniques to successfully predict which passengers survived the sinking of the Titanic. The method use in this project include SVM and three layer neural network. Various tools are used to implement these algorithms including Python, Pandas, Sklearn, TensorFlow etc.

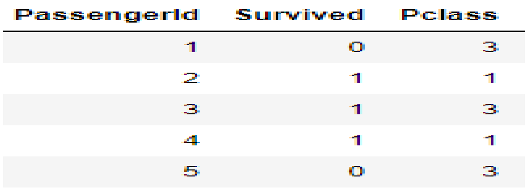
II. DATASET

The data I have used for this project is provided on the Kaggle website [1]. Data consists 891 passenger sample for training set and their associated labels of whether or not the passenger survived. For each passenger, his/her passenger class, name, sex, age, number of siblings/spouses aboard, number of Parents/children aboard, ticket number, fare, cabin and embarked are given. In Table I training dataset sample is given.

TABLE I: ATTRIBUTES IN TRAINING DATASET

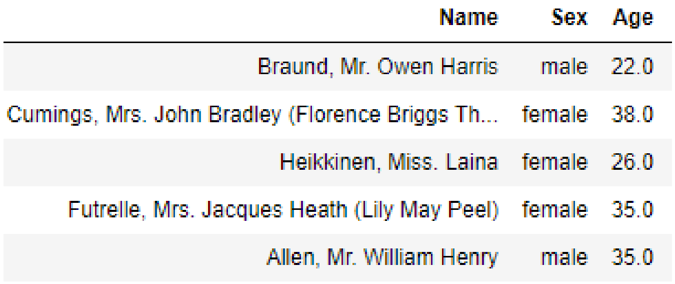
Attributes	Description
PassengerId	Id given to each traveler on the boat
Survived	Survival (0 = No, 1 = Yes)
Pclass	Ticket class. It has three possible values: 1,2,3 (first, second and third class)
Sex	Gender of the passengers (Male or Female)
Age	Age of the Passengers
Sibsp	Number of siblings and spouses traveling with the passenger
Parch	number of parents and children traveling with the passenger
Ticket	Ticket number
Fare	Passenger fare
Cabin	Cabin number
Embarked	Port of Embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

TABLE II: KAGGLE DATASET



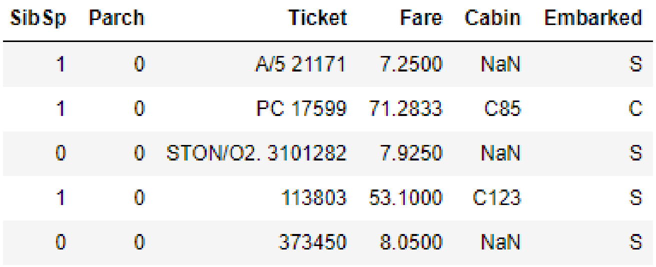
PassengerId	Survived	Pclass
1	0	3
2	1	1
3	1	3
4	1	1
5	0	3

TABLE III: KAGGLE DATASET (Contd...)



Name	Sex	Age
Braund, Mr. Owen Harris	male	22.0
Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0
Heikkinen, Miss. Laina	female	26.0
Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0
Allen, Mr. William Henry	male	35.0

TABLE IV: KAGGLE DATASET (Contd...)



SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	A/5 21171	7.2500	NaN	S
1	0	PC 17599	71.2833	C85	C
0	0	STON/O2. 3101282	7.9250	NaN	S
1	0	113803	53.1000	C123	S
0	0	373450	8.0500	NaN	S

III. DATA ANALYSIS

We need to explore dataset to consider potential data input for the solution. This step is very important because the quality and quantity of data determine how good our model can be. Ticket feature may not be a correlation with survival. It contains 210 duplicates values. We may drop ticket feature. Cabin feature may be dropped as it is highly incomplete or contains many null values in training. PassengerId is not correlated with survival so we may be drop it from training dataset. Out of all passengers in training dataset 38% survived. From Fig.2 (a) and Fig.2 (b) we can see that there is significant difference is survival between Female (74.20%) and male (18.89%).

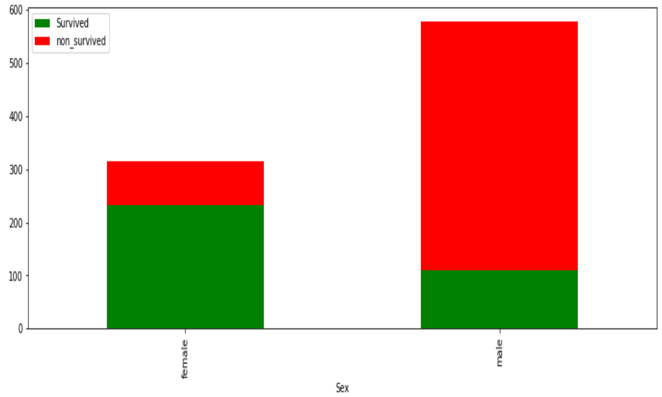
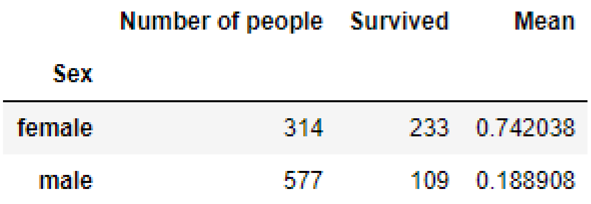


Fig.2 (a) Sex vs Survival



Sex	Number of people	Survived	Mean
female	314	233	0.742038
male	577	109	0.188908

Fig.2 (b) Sex vs. Survival

From Fig. 3 (a) and (b) we found out that female from first and second class have more than 90% survival chance and from third class only 50% survived while male has a much higher survival rate (36.88%) from first class then from second class(15.74%) and then from third class(13.54%).

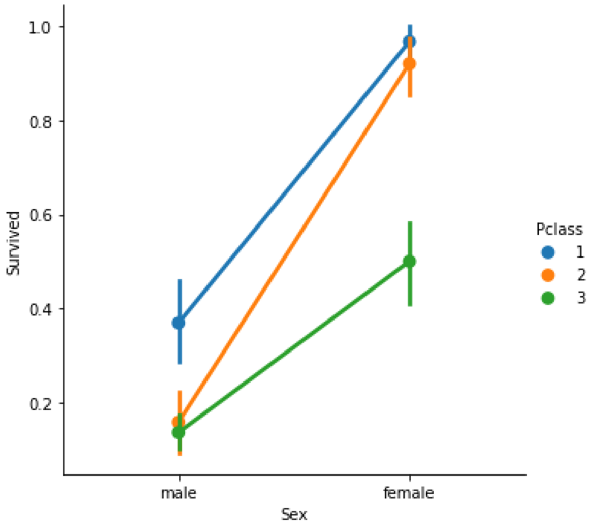
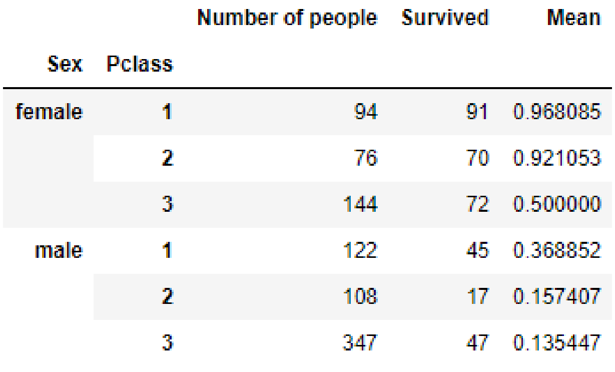


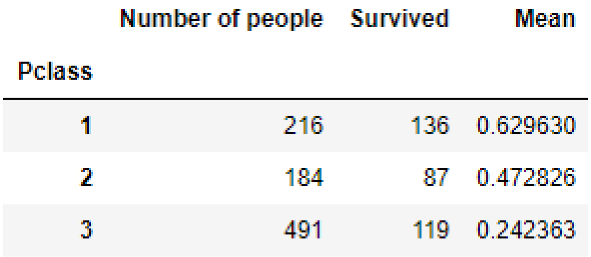
Fig.3 (a) Pclass and Sex vs. Survival



Sex	Pclass	Number of people	Survived	Mean
female	1	94	91	0.968085
	2	76	70	0.921053
	3	144	72	0.500000
male	1	122	45	0.368852
	2	108	17	0.157407
	3	347	47	0.135447

Fig.3 (b) Pclass and Sex vs. Survival

Average fare of Pclass 1 is high and also more people has survival chance from this class (136 people (63%)survived out of 216 people from Pclass 1) 87 people (47%) survived out of 184 people from Pclass 2 119 people (24%) survived out of 419 people from Pclass 3.



	Number of people	Survived	Mean
Pclass			
1	216	136	0.629630
2	184	87	0.472826
3	491	119	0.242363

Fig.4 (b) Pclass vs. Survival

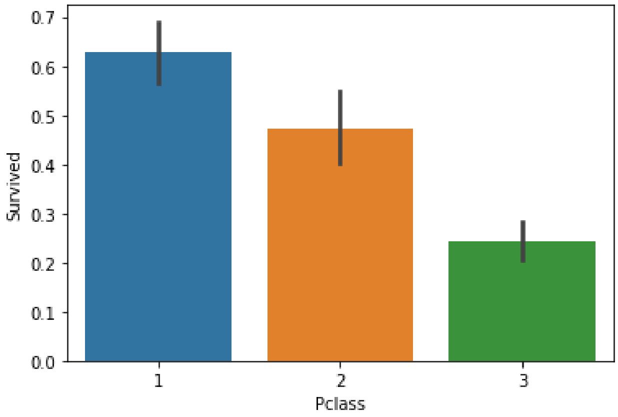


Fig.4 (b) Pclass vs. Survival

From Fig.5 (a, b, c) we can see that younger male of age range 5-10 year tend to survive as depicted by green histogram male of age range 20year-40year has more tend to die. Women survived more than men, as depicted by the larger female green histogram.

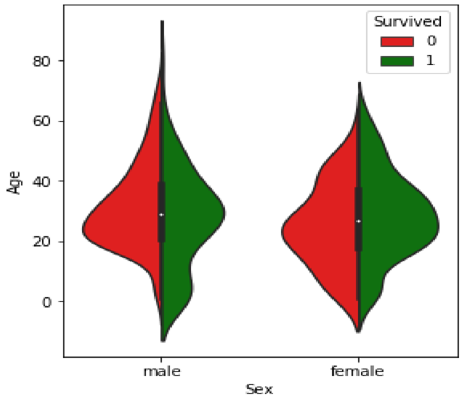


Fig.5 (a) Age and Sex vs. Survival

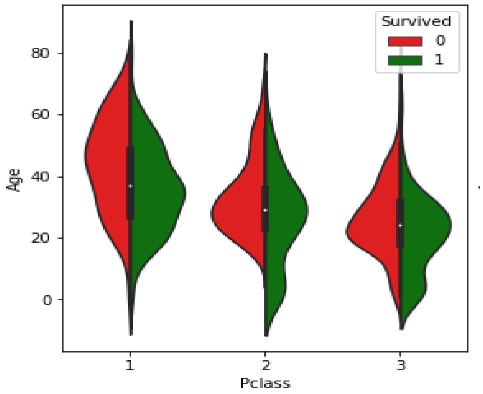


Fig.5 (b) Age and Pclass vs. Survival

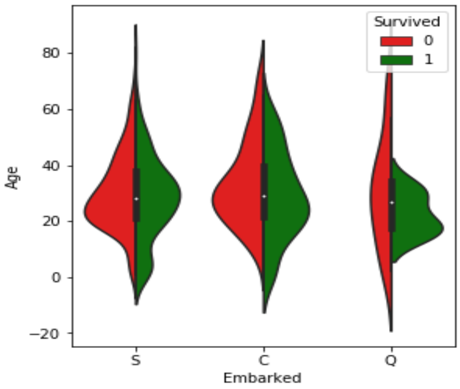


Fig.5 (b) Age and Embarked vs. Survival

IV. FEATURE ENGINEERING

First I have combined train and test data for feature engineering to prevent any information mismatch in train and test data set. From the name extract title that can give additional information about the social status. There are total 18 title in data, map these title with Mr., Miss., Mrs., Officer, Master and Royalty, we need to convert these feature in binary format. Sex features were mapped to 0 as male and 1as female. Next feature is age, null value of age has been filled with mean value of given age data. Now I combined the Sibsp and Parch to get family size as we can see from Fig.6 family size with 2, 3 and 4 members were high survival chance. Family has been converted into three group as singleton, small family and large family. Fare features ware converted into three group as low, medium and high fare range. In Embarked feature two null values were replaced by most frequent value S. After completion of feature engineering data were separated into two part first 891 data as train data and rest 418 data as test data. Then engineered feature were selected for modeling.

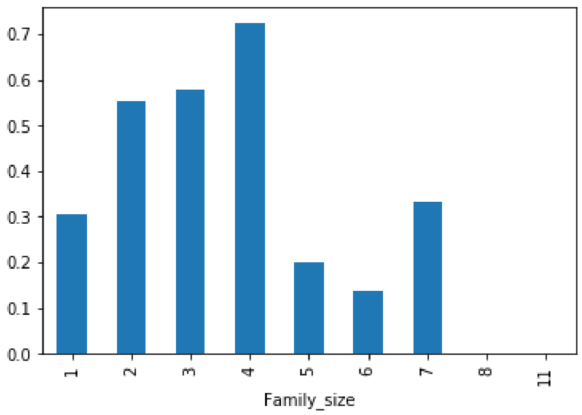


Fig.6 Family Size vs. Survival

V. MODELING

Models uses in this project were SVM and neural network model with training data. The following feature were selected for data modeling a.) Pclass, b.) Sex, c.) Age, d.) Embarked, e.) Title, f.) Family, g.) Fare_data. Train data was split into train and test set for modeling purpose.

The SVM model was implemented for classification on train dataset. rbf function was used as kernel in SVM model. We were able to achieve an accuracy rate of 88.88% on test data set.

The neural network build was a three-layer neural network. Two layer with relu function and third layer with sigmoid function. We were able to achieve an accuracy rate of 90.00% on test data set

TABLE IV: COMPARISON OF ALGORITHM

ALGORITHM	**ACCURACY**
SVM	88.88%
Neural Network	87.77%

VI. CONCLUSION

In this project machine learning algorithm SVM and neural network has been successfully Implemented. We also determined the feature that were most the most significant for the prediction. We were observed that shows SVM higher accuracy rate than neural network model.

VII. REFERENCES

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

[2] Cortes, Corinna; and Vapnik, Vladimir N.; "Support Vector Networks", Machine Learning, 20, 1995

[3] Eric Lam, Chongxuan Tang. Titanic Machine Learning From Disaster