Titanic Disaster

Abhishek Mogili
University of Missouri-St. Louis

8 May 2023

Contents

1 Introduction

The Titanic dataset is a widely used dataset in the field of data science and machine learning. It contains information about passengers who were aboard the RMS Titanic, which was one of the most famous and tragic maritime disasters in history. The Titanic was a British passenger liner that collided with an iceberg on its maiden voyage from Southampton, England to New York

City, USA on April 15, 1912. The dataset is often used as a starting point for learning and practicing data analysis and predictive modeling techniques.
The Titanic dataset consists of information about 1309 passengers, including their age, gender, ticket class, cabin, port of embarkation, and whether they survived or not.

2 Dataset

The dataset was obtained from kaggle Data Science website called the "titanic project". This dataset has been made publicly available. It contains the number of passengers with their names, their age, ticket number, fare of ticket, sex.

2.1 Dataset Description

There are 1309 rows and 12 columns. I will extract 12 columns and add 1 column for target variable. The label can be '0' or '1'. To draw that passangers can be "1" for survived or "0" for died. These columns (features we will use to predict) are as follows:

- passenger id
- survived
- pclass
- name
- sex
- age
- sibsp
- parch
- fare
- cabin
- embarked

2.2 Data Visualization

The bar graph plot shows that how many passengers survived and how many died. which can be seen in below figures.

Using Bivariate analysis is a statistical method used to investigate the relationship between two variables. The analysis aims to determine whether there is a significant association or correlation between the two variables and to quantify the strength and direction of this relationship.

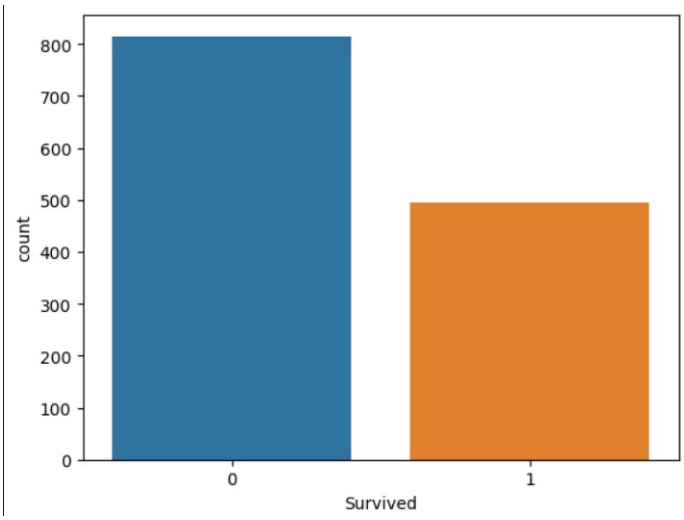

Figure 1: Input Data Distribution Histograms (Part One)

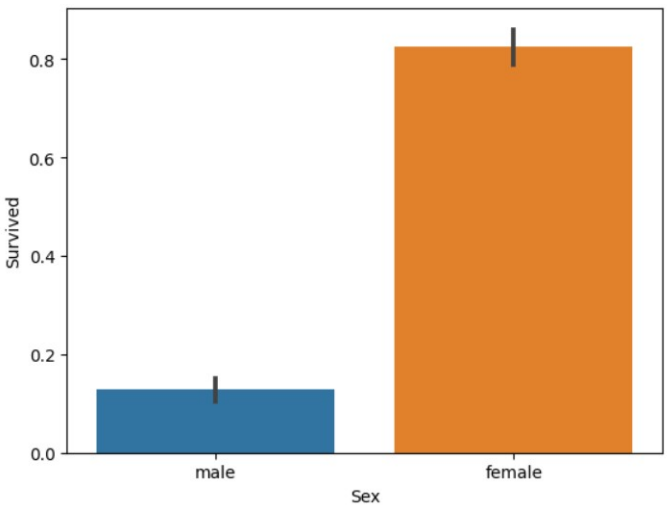

Figure 2: Plot shows which has better chance of survival

Figure 3: Input Feature Statistics

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
count	1309.000000	1309.000000	1309.000000	1309	1309	1046.000000	1309.000000	1309.000000	1309	1308.000000	295	1307
unique	NaN	NaN	NaN	1307	2	NaN	NaN	NaN	929	NaN	186	3
top	NaN	NaN	NaN	Kelly, Mr. James	male	NaN	NaN	NaN	CA. 2343	NaN	C23 C25 C27	S
freq	NaN	NaN	NaN	2	843	NaN	NaN	NaN	11	NaN	6	914
mean	655.000000	0.377387	2.294882	NaN	NaN	29.881138	0.498854	0.385027	NaN	33.295479	NaN	NaN
std	378.020061	0.484918	0.837836	NaN	NaN	14.413493	1.041658	0.865560	NaN	51.758668	NaN	NaN
min	1.000000	0.000000	1.000000	NaN	NaN	0.170000	0.000000	0.000000	NaN	0.000000	NaN	NaN
25%	328.000000	0.000000	2.000000	NaN	NaN	21.000000	0.000000	0.000000	NaN	7.895800	NaN	NaN
50%	655.000000	0.000000	3.000000	NaN	NaN	28.000000	0.000000	0.000000	NaN	14.454200	NaN	NaN
75%	982.000000	1.000000	3.000000	NaN	NaN	39.000000	1.000000	0.000000	NaN	31.275000	NaN	NaN
max	1309.000000	1.000000	3.000000	NaN	NaN	80.000000	8.000000	9.000000	NaN	512.329200	NaN	NaN

3 Data Processing

3.1 Data Splitting

The most common approach to data splitting is to randomly divide the dataset into a training set and a testing set. Here I have split it into 20 percent of data for testing and 80 percent of data for training.

3.2 Data Scaling

The below method first computes the mean and standard deviation of the data in the columns and then scales the data using the formula.

$$\frac{value - min}{sd}$$

4 Data analysis

In this after normalizing the input data,the output data need to preprocessed into right format. I have mapped 'sex' column as for male equals to '1' and female equals to '0'. I have used dummies method from pandas dataframe with the column name 'Embarked' column and resulting dataframe will have corresponding values as '0' or '1'.

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Gender	Embarked_C	Embarked_Q	Em
1148	1149	0	3	Niklasson, Mr. Samuel	male	28.000000	0	0	363611	8.0500	1	0	0	
1049	1050	0	1	Borebank, Mr. John James	male	42.000000	0	0	110489	26.5500	1	0	0	
982	983	0	3	Pedersen, Mr. Olaf	male	29.881138	0	0	345498	7.7750	1	0	0	
808	809	0	2	Meyer, Mr. August	male	39.000000	0	0	248723	13.0000	1	0	0	
1195	1196	1	3	McCarthy, Miss. Catherine Katie""	female	29.881138	0	0	383123	7.7500	0	0	1	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1095	1096	0	2	Andrew, Mr. Frank Thomas	male	25.000000	0	0	C.A. 34050	10.5000	1	0	0	
1130	1131	1	1	Douglas, Mrs. Walter Donald (Mahala Dutton)	female	48.000000	1	0	PC 17761	106.4250	0	1	0	
1294	1295	0	1	Carrau, Mr. Jose Pedro	male	17.000000	0	0	113059	47.1000	1	0	0	
860	861	0	3	Hansen, Mr. Claus Peter	male	41.000000	2	0	350026	14.1083	1	0	0	
1126	1127	0	3	Vendel, Mr. Olof Edvin	male	20.000000	0	0	350416	7.8542	1	0	0	

1309 rows × 14 columns

Figure 4: Table shows converted dummies embarked column

5 Modelling

A neural network model that can learn from data and make predictions based on that learning is created using an Artificial Neural Network (ANN).

5.1 Using different Neural Network Architectures

Here I am using various architectures starting with one input layer,one hidden layer and one output. But there should be significant improvement further upon adding more number of layers.

5.1.1 Performance comparison

Here we apply logistic regression with one layer and shows the accuracy of percentages for various layers increasing epochs for better performance.

Table 1: Performance comparison for varying hidden layers

Hidden Layers	Training Acc
64-32-16-8-1	85.97
32-16-8-1	85.44
16-8-1	86.21
8-1	87.36
4-1	86.97
2-1	86.21

As we can see from the above table that the basic architecture with just one hidden layer is performing better than the other architecture

5.2 Learning Curve of Neural Network

The given below images show the learning curve for the neural network model.

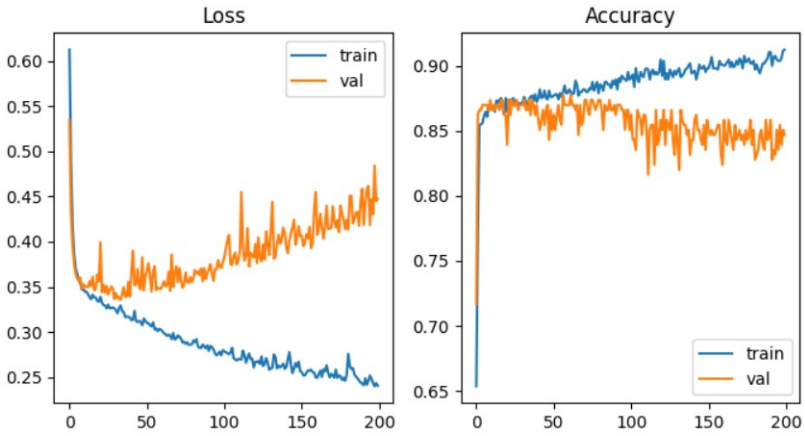

Figure 5: Two graphs shows in accuracy, loss vs epoch for model '64-32-16-8-1'

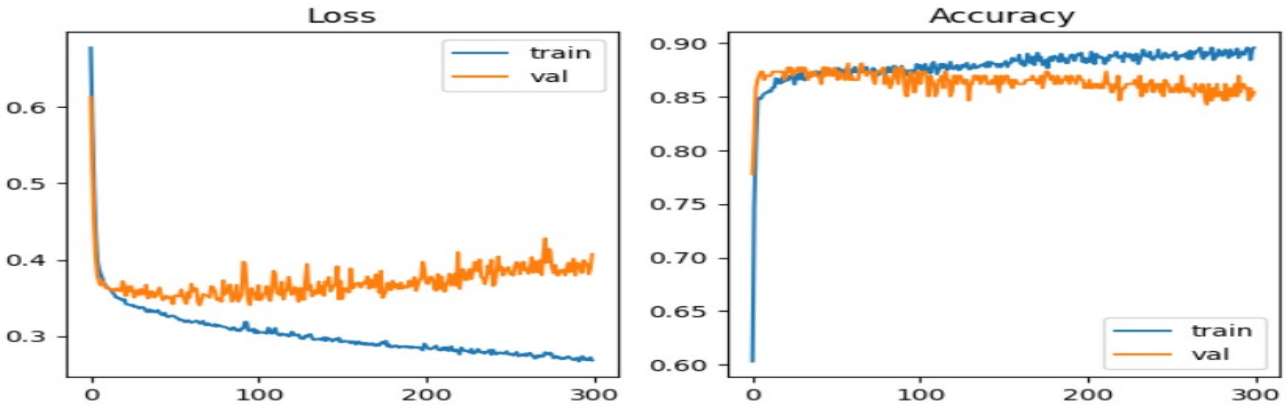

Figure 6: curve showing change in loss/accuracy vs epoch for model '32-16-8-1'

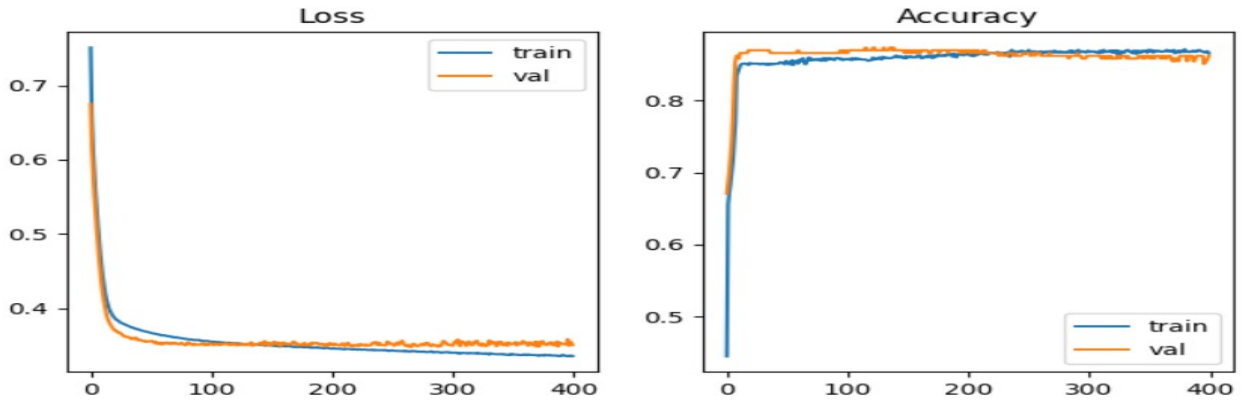

Figure 7: curve showing change in loss/accuracy vs epoch for model '16-8-1'

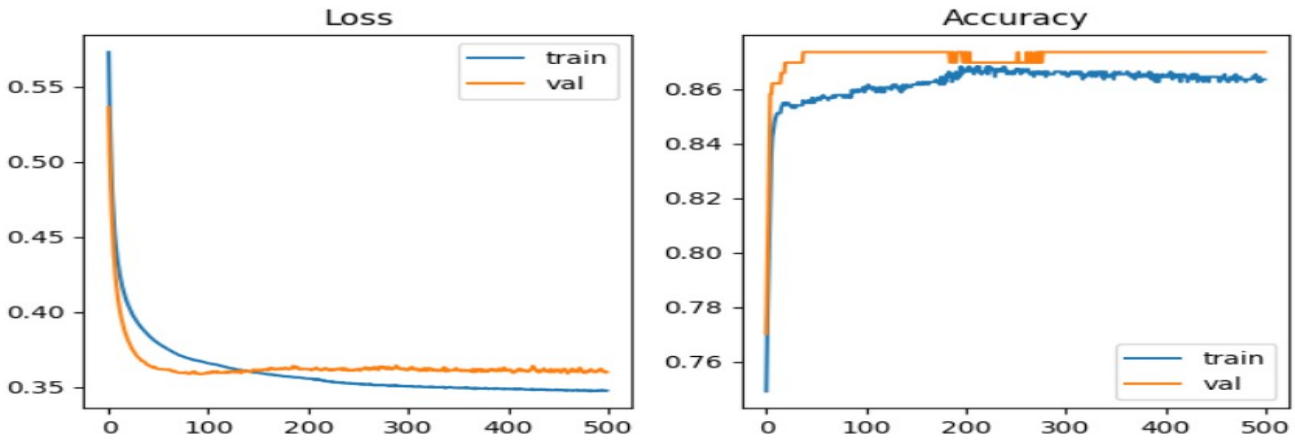

Figure 8: curve showing change in loss/accuracy vs epoch for model '8-1'

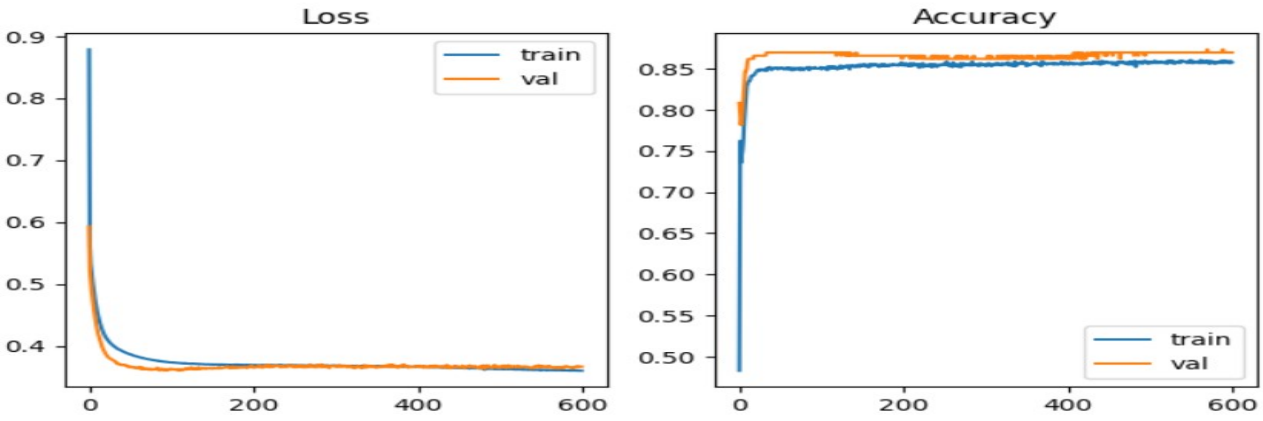

Figure 9: curve showing change in loss/accuracy vs epoch for model '4-1'

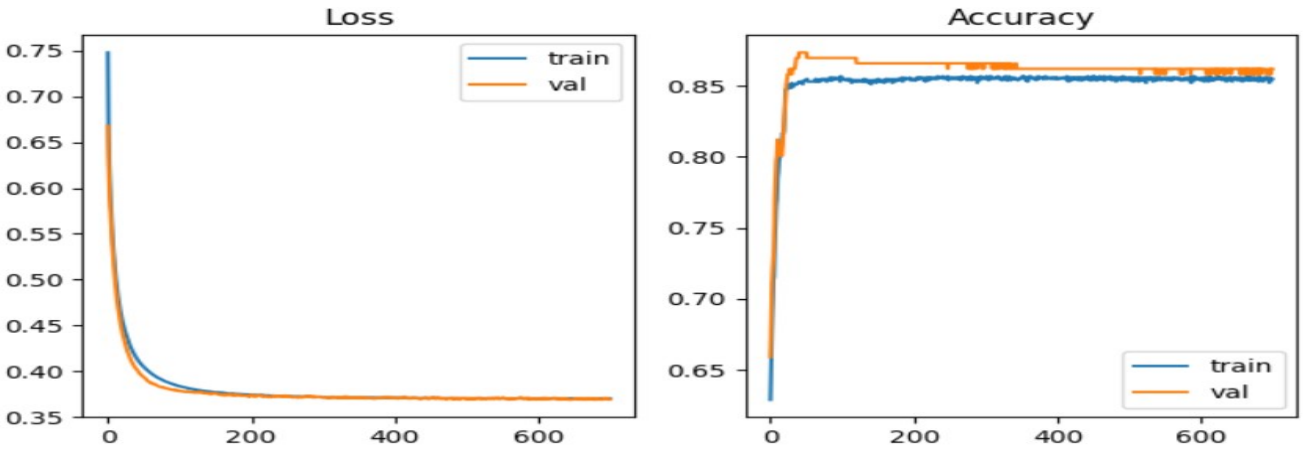

Figure 10: curve showing change in loss/accuracy vs epoch for model '2-1'

6 Logistic regression Model

Using Logistic regression model verify that if it gives best accuracy for testing data and compare with the other models In logistic regression, the relationship between the predictor variables and the binary outcome is modeled using a logistic function, which maps the input values to a value between 0 and 1. This value represents the probability of the binary outcome being true or false, given the predictor variables

Table 2: The table given below shows the performance of the logistic regression model.

Test accuracy
86.21

7 Random Forest

In Random Forest, each decision tree is trained on a subset of the training data and a subset of the input features. The output of the model is then determined by aggregating the predictions of the individual decision trees.The accuracy is increased slightly when compared to logistic regression model.

Table 3: The table given below shows the performance of the Random Forest model.

Test accuracy
88.12

8 Challenges Faced

The dataset contains missing data for several features, including Age, Cabin, and Embarked. This missing data can affect the performance of the machine learning model, as it may not have enough information to make accurate predictions.

The dataset is imbalanced, as there are more passengers in the Survived=0 category than the Survived=1 category. This can affect the performance of the machine learning model, as it may be biased towards predicting the majority class.

9 Future Improvement

- Regularization Method
- Ensemble Learning
- The dataset is relatively small, which increases the risk of overfitting. Overfitting occurs when the machine learning model becomes too complex and learns the noise in the training data

10 Conclusion

In this project, I developed a machine learning model to predict the survival of passengers aboard the Titanic using their characteristics such as age, gender, cabin class, and fare. After analyzing the dataset, it was found that gender, cabin class, and age were the most significant factors in predicting survival rate. By addressing the challenges associated with the dataset and implementing appropriate techniques, we can create accurate and robust machine learning models that can be used to predict survival rate based on passenger characteristics.