Predicting the Survival of Titanic Passengers

Introduction

The sinking of the Titanic is one of the most infamous shipwrecks in history.

On April 15, 1912, during her maiden voyage, the widely considered “unsinkable” RMS Titanic sank after colliding with an iceberg. Unfortunately, there weren’t enough lifeboats for everyone onboard, resulting in the death of 1502 out of 2224 passengers and crew.

While there was some element of luck involved in surviving, it seems some groups of people were more likely to survive than others. Let's find out the factors which help to survive.

The Titanic dataset contains information about the passengers on board the Titanic, including their age, gender, and ticket class, among other features. In this project, we used a Multi-Layer Perceptron (MLP) neural network to make predictions on this dataset. We explored different configurations of MLPs with varying numbers of layers and neurons per layer, as well as different optimizers and learning rates. Our goal was to find the best model configuration that maximizes the accuracy of our predictions.

Dataset Information:

The dataset contains 891 records in below format :

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Here is a brief summary of the features included in the Titanic dataset:

- **PassengerId:** unique identifier for each passenger
- **Survived:** whether or not a passenger survived (0 = no, 1 = yes)
- **Pclass:** ticket class (1 = first class, 2 = second class, 3 = third class)
- **Name:** name of passenger
- **Sex:** gender of passenger
- **Age:** age of passenger
- **SibSp:** number of siblings/spouses aboard the Titanic
- **Parch:** number of parents/children aboard the Titanic
- **Ticket:** ticket number
- **Fare:** price of ticket
- **Cabin:** cabin number
- **Embarked**: port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

Exploratory Data Analysis

We conducted exploratory data analysis to gain insights into the dataset. We plotted various charts to visualize the relationship between different features and survival status. We observed that women had a higher chance of survival. Additionally, passengers in higher ticket classes also had a higher chance of survival. Plots and their interpretation can be found in the code file.

Data Preparation

We started by loading the Titanic dataset into a pandas DataFrame and performing some basic data cleaning and preprocessing steps. This included filling in missing values, encoding categorical variables, and scaling numerical features.

We split the dataset into training and testing sets, with a 80/20 split, respectively.

Model Selection

We explored four different configurations of MLP models, each with a varying number of layers and neurons per layer. For each model configuration, we trained the model with six different optimizers and learning rates.

Model Evaluation

We evaluated the performance of each model by computing its accuracy score, f1 score, and confusion matrix on the testing set. We also computed the precision and recall scores for each model configuration. Here is the Result:

	model	optimizer	learning_rate	accuracy	f1_score	neurons
0	1-L (64) N	Adam	0.001	0.737430	0.680272	[64]
1	1-L (64) N	Adam	0.010	0.743017	0.671429	[64]
2	1-L (64) N	RMSprop	0.001	0.698324	0.630137	[64]
3	1-L (64) N	RMSprop	0.010	0.726257	0.662069	[64]
4	1-L (64) N	SGD	0.001	0.586592	0.000000	[64]
5	1-L (64) N	SGD	0.010	0.798883	0.746479	[64]
6	2-L-(64, 32)-N	Adam	0.001	0.743017	0.676056	[64, 32]
7	2-L-(64, 32)-N	Adam	0.010	0.748603	0.689655	[64, 32]
8	2-L-(64, 32)-N	RMSprop	0.001	0.715084	0.627737	[64, 32]
9	2-L-(64, 32)-N	RMSprop	0.010	0.776536	0.705882	[64, 32]
10	2-L-(64, 32)-N	SGD	0.001	0.586592	0.000000	[64, 32]
11	2-L-(64, 32)-N	SGD	0.010	0.798883	0.746479	[64, 32]
12	3L-(64, 32, 16)-N	Adam	0.001	0.748603	0.676259	[64, 32, 16]
13	3L-(64, 32, 16)-N	Adam	0.010	0.759777	0.666667	[64, 32, 16]
14	3L-(64, 32, 16)-N	RMSprop	0.001	0.698324	0.619718	[64, 32, 16]
15	3L-(64, 32, 16)-N	RMSprop	0.010	0.726257	0.662069	[64, 32, 16]
16	3L-(64, 32, 16)-N	SGD	0.001	0.586592	0.000000	[64, 32, 16]
17	3L-(64, 32, 16)-N	SGD	0.010	0.804469	0.768212	[64, 32, 16]
18	4L-(64, 32, 16, 8)-N	Adam	0.001	0.748603	0.661654	[64, 32, 16, 8]
19	4L-(64, 32, 16, 8)-N	Adam	0.010	0.586592	0.000000	[64, 32, 16, 8]
20	4L-(64, 32, 16, 8)-N	RMSprop	0.001	0.715084	0.627737	[64, 32, 16, 8]
21	4L-(64, 32, 16, 8)-N	RMSprop	0.010	0.709497	0.675000	[64, 32, 16, 8]
22	4L-(64, 32, 16, 8)-N	SGD	0.001	0.586592	0.000000	[64, 32, 16, 8]
23	4L-(64, 32, 16, 8)-N	SGD	0.010	0.787709	0.724638	[64, 32, 16, 8]

Conclusion

In conclusion, we successfully used an MLP neural network to predict survival on the Titanic dataset. We explored different model configurations and optimizers and learning rates, and found the best model configuration to be a 3-layer MLP with 64, 32, 16 neurons per layer wich optimized by SGD with learning rate 0.01. Our best model achieved an accuracy of 81% on the testing set.

Further improvements to this project could include exploring more advanced machine learning algorithms, such as ensemble methods or gradient boosting, and performing more extensive hyperparameter tuning to further improve the accuracy of our predictions.