Machine Learning - Titanic Survival Prediction Analysis

Meixuan Li *

Department of Information Science and Technology, Nanjing Forestry Univerity, Nanjing, China

* Corresponding Author Email: wimp74231@gmail.com

Abstract. The most well-known shipwreck in history, the Titanic, happened in April 1912. The dataset contains comprehensive details regarding everyone that stepped onto the ship, and more than 1,500 passengers and crew members perished in the catastrophe. This study investigates the effectiveness of logistic regression, multi-layer perceptron support vector machines, and XGBoost algorithms in forecasting passenger survivability using the Titanic passenger dataset. Numerous factors, including passengers' financial status, gender, age, class, and others, influence their chances of surviving, and these factors are included in the Titanic passenger data collection. This work creates the classification model using logistic regression and other methods, preprocesses and analyzes the characteristic data, and compares the model's performance. The variables that significantly affect the survival rate are found. Experiments are used to assess each algorithm's accuracy, precision, and recall rate in order to determine which prediction model is optimal. Every algorithm has a unique optimization space and application context. In real-world use, it ought to be chosen and tailored based on the features of particular issues.

Keywords: Titanic; machine learning; logistic regression; support vector machine.

1. Introduction

On April 15, 1912, the most well-known being the Titanic tragedies historically [1], sank as a result of an iceberg colliding with it as it was during its inaugural trip. There were 2,224 people on board, including crew, and 1502 of them died. The shipwreck is gaining international recognition. While some people's life depends only on luck [2], other people always have a stronger chance of surviving than others, including women, children, and members of upper-class individuals.

In order to find the groups that share a lot of information [3] and modify the data, this approach first mines extra valuable features using feature engineering and other techniques, followed by the application of the same group effect [4]. The survival rate of passengers is predicted by comparing the logistic regression [5], support vector machine (SVM) [6], and multi-layer perceptron (MLP) techniques in order to identify the optimal model.

2. Algorithm design

2.1. Data processing

2.1.1 Data analyzing

To comprehend the fundamental circumstances and distributional properties of the information, a thorough examination of the information must be done prior to the design and implementation of the algorithm. Numerous characteristics of the Titanic data collection, including passenger survival, cabin distribution, age distribution, fare distribution, and the correlation [7] between port of embarkation and survival, are statistically analyzed.

Mixed data types features include Alphanumeric and numeric data in a single feature. These could be the targets for the correction. Ticket contains both alphabetic and numeric data types. The cabin uses alphanumeric characters. Due to a large percentage of duplicates (22%) and the possibility that there is no relationship between tickets and survival, the ticket component was removed from the analysis. Since passenger id is not necessary for survival, it's taken out of the training dataset. Since the name characteristic is somewhat unusual and does not immediately aid in survival, it will be removed.

Less than half of the passengers may have survived the accident, according to statistical analysis of the survival condition. The percentage of people traveling in first class is comparatively low, but the percentage of passengers traveling in third class is the greatest.

Between the ages of 20 and 40, there is a comparatively high proportion of travelers, with an age distribution that spans from 0 to 80 years old. The fee for passengers is dispersed unevenly, ranging from $512 to $0. A box plot was used to display the fare distribution by class, and it was discovered that the median fare for first class was substantially higher than that of other classes. Three distinct ports are where passengers board: Southampton (S), Cherbourg (C), and Queenstown (Q). Out of all of these, Southampton was where the most passengers boarded. Additionally, it examined the correlation between various embarkation ports and survival, concluding that passengers sailing from Cherbourg had a greater survival rate.

2.1.2 Data processing

The data is preprocessed, including missing value processing, feature coding, and feature selection, in accordance with the findings of the data analysis [8].

The mean filling approach is employed to cover the absent values within the age field.

The mode filling method fills in the port of embarkation field's missing value.

It decided to remove the cabin number field as it had so many missing values.

Label coding is used to translate category variables—like gender and port of embarkation—into numerical features that can be employed in further model training [9].

The features that were most strongly connected with survival are those that included cabin class, sex, age, number of parents/children, number of siblings/spouses, fare, and port of embarkation. These findings are based on the data analysis results [10].

A cleaned data set is obtained, any missing values are processed, and the class features are converted to numerical type by the aforementioned data processing processes. This provides a strong basis for the ensuing model training and evaluation.

2.2. Method

2.2.1 Logistic

One type of linear model that is frequently employed in binary classification issues is logistic regression. According to the theory of logistic regression, the target variable, y, is the outcome of a linear combination of eigenvectors X that have been run through a sigmoid function.

Mathematical formula

The prediction function of logistic regression model is:

$$h_\theta(X) = \frac{1}{1+e^{-\theta^T X}} \tag{1}$$

Where X is the input feature vector.

The cost function is:

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} [y^{(i)}\log(h_\theta(X^{(i)})) + (1-y^{(i)})\log(1-h_\theta(X^{(i)}))] \tag{2}$$

The cost function is optimized by gradient descent method and the parameters are updated

$$\theta := \theta - \alpha\frac{\partial J(\theta)}{\partial\theta} \tag{3}$$

2.2.2 MLP

A feed-forward neural network called a multi-layer perceptron (MLP) uses many hidden layers and nonlinear activation functions to record complex feature interactions. A back propagation algorithm updates the weights in the MLP in order to minimize the loss function.

The output of each layer is:

$$a^{(l)} = f(W^{(l)}a^{(l-1)} + b^{(l)}) \tag{4}$$

Where $W^{(l)}$ and $b^{(l)}$ are the weight and bias of layer l, respectively, and f is the activation function (such as ReLU).

The loss function (for the cross entropy loss) is:

$$J(W,b) = -\frac{1}{m}\sum_{i=1}^{m} [y^{(i)}\log(a^{(L)}) + (1-y^{(i)})\log(1-a^{(L)})] \tag{5}$$

Back propagation to update the parameters:

$$\frac{\partial J}{\partial W^{(l)}} = \delta^{(l)} a^{(l-1)^T}$$
$$W^{(l)} := W^{(l)} - \alpha \frac{\partial J}{\partial W^{(l)}} \tag{6}$$

2.2.3 SVM

When it comes to high-dimensional data classification problems, the support vector machine (SVM) finds the best hyperplane through which to classify the data. Maximizing the distance between each side of the classification hyperplane is the SVM's objective.

To find the optimal hyperplane： $w^T x+b$ where w and b are the parameters that need to be solved, the objective function is：

$$\min_{\mathbf{w},b} \frac{1}{2} \| \mathbf{w} \|^2 \tag{7}$$

The constraint is：

$$y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) \geq 1 \tag{8}$$

Using the Lagrange multiplier method, it can be converted into：

$$\mathcal{L}(\mathbf{w}, b, \alpha) = \frac{1}{2} \| \mathbf{w} \|^2 - \sum_{i=1}^{m} \alpha_i [y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) - 1] \tag{9}$$

The final classification decision function is obtained by solving the Lagrange dual problem.

2.2.4 XG Boost

A tree model called XG Boost uses gradient lifting and incorporates several weak classifiers, or decision trees, to enhance the model's classification performance. To reduce the total loss, each tree was continually iterated through the residuals of the preceding tree.

The XG Boost's objective function is:

$$L(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{10}$$

$$\Omega(\mathrm{f}) = \gamma \mathrm{T} + \frac{1}{2}\lambda \| \mathrm{w} \|^2 \tag{11}$$

is a regularization of the complexity terms in the model.

The loss function's approximate second-order Taylor expansion is as follows:

$$L^{(t)} \approx \sum_{i=1}^{n} [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) \tag{12}$$

$\mathrm{g_i} = \partial_{\hat{\mathrm{y}}^{(\mathrm{t}-1)}} \mathrm{l}(\mathrm{y_i}, \hat{\mathrm{y}}^{(\mathrm{t}-1)})$ and $\mathrm{h_i} = \partial^2_{\hat{\mathrm{y}}^{(\mathrm{t}-1)}} \mathrm{l}(\mathrm{y_i}, \hat{\mathrm{y}}^{(\mathrm{t}-1)})$(13) Are the first and second derivatives of the loss function, respectively.

3. Results and analysis

3.1. Data analysis results

In the data analysis step, the Titanic data set is subjected to a thorough statistical analysis and visualization in order to comprehend the data distribution and the correlation between the features and survival conditions.

There were roughly 38.38 percent of surviving passengers in the Titanic data set. This indicates that roughly 40% of the travelers had made it out of the mishap.

There were 491 people in third class, 184 in second class, and 216 in first class in the data set. First-class travelers are the lowest, and third-class travelers are at the top.

First-class passengers exhibit the highest survival rate, whilst third-class passengers have the lowest survival rate.

The passengers' age range is somewhat broad, with the majority of them falling between 20 and 40. The passenger's range in age from a few months old for the smallest to over eighty years old for the oldest.

There was also a tendency in the association between age and survival, with younger and younger passengers having greater survival rates and older passengers having lower survival rates.

The range of passenger tickets was $0 to $512, with first class having a substantially higher median fare than second and third class.

The result that first-class passengers had better survival rates is consistent with the fact that passengers paying higher fees often have greater survival rates.

Three ports are where most passengers board: Southampton (S), Cherbourg (C), and Queenstown (Q). Southampton port has the greatest number of passengers.

Passengers boarding ships at various ports have varying survival rates; passengers aboard ship from Cherbourg had a better survival rate.

3.2. Experimental results

After data processing, the models were trained and evaluated using four machine learning algorithms. Table 1 is training set results. Table 2 is testing set result.

Table 1 Training set results

algorithm	Accuracy	Precision	Recall	F1-score
logistic regression	0.7972	0.7684(Class0) 0.8542 (Class1)	0.9125(Class0) 0.6508(Class1)	0.8343(Class0) 0.7387(Class1)
MLP	0.7692	0.7975(Class0) 0.7344 (Class1)	0.7875(Class0) 0.7460(Class1)	0.7925(Class0) 0.7402(Class1)
SVM	0.6364	0.6321(Class0) 0.6486 (Class1)	0.8375(Class0) 0.3810(Class1)	0.7204(Class0) 0.4800(Class1)
XG Boost	0.7413	0.7471(Class0) 0.7321 (Class1)	0.8125(Class0) 0.6508(Class1)	0.7784(Class0) 0.6891(Class1)

Table 2 Testing set result

	logistic regression	MLP	SVM	XG Boost
Predicted value	0.7972	0.7947	0.6385	0.7296

Table 2 shows that the logistic regression model has the highest prediction accuracy (0.797), followed by the MLP model (also above 0.79), the XGboost model (above 70%), and the SVM model (comparatively low, 0.639).

3.3. Final results

The following conclusions can be reached by looking at each model's evaluation.

Especially in the classification task of survival or not, logic regression performs well in terms of accuracy and many assessment indicators, and its overall performance is the best balanced. When processing unbalanced data, the logistic regression's higher F1 score suggests a stronger classification capacity.

Despite being marginally less accurate than logistic regression, MLP demonstrated excellent precision and recall. Although it requires a lengthy training period, its intricate network structure and nonlinear properties offer it some advantages in capturing complicated correlations in the data.

SVM performed poorly, particularly when it came to recall, notably for class 1. The reason for this could be the significant computational burden involved in managing vast amounts of data and the limited capacity of SVM to handle nonlinear relationships in high-dimensional space.

XG Boost's accuracy, which demonstrates good precision and recall, lies between logistic regression and MLP. Its integrated learning features enable it to handle complex relationships and vast amounts of data with relative ease; nevertheless, the model's complexity and parameter tweaking are more involved.

In this trial, logic regression fared the best, exhibiting a balanced evaluation index and good accuracy. Although XGBoost and multi-layer perceptrons are better at capturing intricate data correlations, they also take longer and more processing power to implement. In this experiment, SVM underperformed other algorithms; this could mean more parameter tweaking or better feature selection techniques.

4. Conclusion

In real-world applications, it is very crucial to choose and optimize the problem based on the features of unique difficulties. Additionally, optimizing algorithms and updating computing resources have become essential issues due to the increase in data amount and complexity.

In order to increase the accuracy of the model, this paper can also optimize the model through feature mining, screening, and missing value filling. Model optimization involves choosing a range of models, comparing them initially based on model scores, and then taking into account a number of performance metrics to choose an appropriate prediction model. while it comes to feature mining and screening, new features can be mined, and model prediction accuracy can be evaluated while choosing various features, all leading to the eventual training model's feature set. Missing value filling also has an impact on the model's final prediction outcomes.

References

[1] Aastha PH, Liu Y. 2020. DeepCompete: A deep learning approach to competing risks in continuous time domain. AMIA Annu. Symp. Proc. 2020:177–86.

[2] Haque A, Shivaprasad G, Guruprasad G. Passenger data analysis of Titanic using machine learning approach in the context of chances of surviving the disaster. IOP Conference Series: Materials Science and Engineering, 2021, 1065(1): 12042.

[3] Sherlock J, et al. Classification of Titanic passenger data and chances of surviving the disaster. Cornell University, 2018.

[4] Singh K, Nagpal R, Sehgal R. Exploratory data analysis and machine learning on Titanic Disaster Dataset. IEEEXPLORE, 2020.

[5] Tabbakh A, Rout J K, Rout M. Analysis and prediction of the survival of Titanic passengers using machine learning. In Lecture notes in networks and systems, 2020, 297-304.

[6] Fei Zhe, Li Yi. Estimation and inference for high dimensional generalized linear models: A splitting and smoothing approach. Journal of Machine Learning Research, 2021, 22(58): 1-32.

[7] Shekhar S, Arora D, Sharma P. Classifying Titanic Passenger Data and Prediction of Survival from Disaster. In Lecture notes in networks and systems, 2020, 181-187.

[8] Kakde Y, Agrawal S. Predicting survival on Titanic by applying exploratory data analytics and machine learning techniques. International Journal of Computer Applications, 2018, 179 (44): 32-38.

[9] Khalid S, Khalil T, Nasreen S. A survey of feature selection and feature extraction techniques in machine learning Proc. Int. Conf. on Science and Information (London), 2014, 372-378.

[10] Guyon I, Elisseeff A. An introduction to feature extraction. Feature extraction: foundations and applications. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006: 1-25.