Proceedings of the 3rd International Conference on Financial Technology and Business Analysis
DOI: 10.54254/2754-1169/128/2024.18267

Prediction of the Titanic Survival Probability Based on Boosting Strategies

Xinan Zhang[1,a,*]

[1]*School of Mathematics, Hefei University of Technology, Hefei, China*
a. 2021212365@mail.hfut.edu.cn
**corresponding author*

Abstract: According to the unsatisfactory rescue outcomes in natural disasters such as marine accidents and the challenges in survival prediction at nowadays, this study employs Gradient Boosting, XGBoost model, and Optimized XGBoost model (based on RandomizedSearchCV method) to forecast the survival of Titanic passengers. The research begins by processing data on various passenger characteristics from the Titanic. Through comparative analysis of experimental results from these three predictive models, it manifests that the Optimized XGBoost model demonstrates highest precision and accuracy, while being less prone to overfitting and underfitting.

Keywords: Marine accident, Gradient Boosting, XGBoost, RandomizedSearchCV.

1. Introduction

With the rapid development of industrial standards in modern society, an increasing number of machine learning models like decision tree model, XGBoost model and Random forest model are attracting widespread interest due to its various application in the field of predicting human survival in natural disasters. This approach allows for more nuanced and accurate predictions in similar situations across various fields. Hakkal and Ait employ XGBoost model to enhance learner performance prediction models such as Item Response Theory (IRT), Performance Factor Analysis (PFA) and DAS3H [1]. Punitha et al. have developed a detection model utilizing the XGBoost classifier [2]. Their model analyzes muscle fatigue progression by examining geometric features of surface electromyography (sEMG) signals and demonstrates impressive results, achieving a balanced accuracy of 96.83% and an F-score of 95.25%. Abolhosseini et al. utilize their CRDT algorithm outperforms ID3 and other decision tree variants like CART and Random Forest in terms of accuracy and efficiency across ten real datasets [3]. Jing et al. present a novel method that combines deep learning with XGBoost model to estimate battery SOH with remarkable accuracy, leading to an RMSE and MAE below 1% [4]. Shobhit et al. develop a metasurface-based sensor that leverages graphene and gold layers to achieve high sensitivity, the integration of the XGBoost regressor into the optimization of their sensor has significantly reduced computational demands while maintaining high prediction accuracy [5]. Zhang provides a comprehensive evaluation of explainable machine learning models including Random Forest model and CatBoost model for urban traffic flow prediction, demonstrating and contrasting their performance and explainability [6].

Previous research of this tragedy mainly focuses on its social and economic impacts. However, by utilizing the machine learning and data mining technology, researchers are able to analyze the Titanic

DOI: 10.54254/2754-1169/128/2024.18267

dataset from a new perspective, which may lead to more nuanced and precise predictions. Cicoria performs cluster analysis and decision tree models, suggesting the Sex serves as the most influential factor [7]. Singh employs Logistic Regression and Naïve Bayes on the dataset, indicating that the Logistic Regression has the best accuracy [8]. Ma compares the performance of K-nearest neighbors, Random Forest and Support Vector Machine algorithms on the Titanic dataset, his study emphasizes the importance of hyperparameter tuning and the potential for ensemble algorithms to yield more accurate predictions [9]. Park et al. explore the educational aspects of the Titanic tragedy, their research discusses the integration of science and history in teaching about disasters [10]. Ao and Yu discuss the application of decision tree models, AdaBoost models, and random forest prediction models to predict the survival of Titanic passengers, their study concludes that the random forest model offers high precision and accuracy with the lowest risk of overfitting and underfitting [11].

This research aims to establish a binary classification model to predict the survival probability of each passenger in the Titanic shipwreck disaster of 1912 by analyzing factors such as gender, age and passenger class. This dissertation describes a new approach of using the XGBoost model to predict the survival rate of Titanic passengers. This dissertation begins by data preprocessing and establishing the Gradient Boosting model, which manifests the low relevance to the original dataset and poor accuracy in prediction. Issues regarding the XGBoost model and its optimization are discussed in later sections.

2. Data

Data feature description of the original dataset is shown in Table 1. he number of missing values in each column can be revealed in both training and test sets by using isnull() function and sum() method, the results are shown in Table 2. From the missing value statistics, it can be concluded that the numerical features Age and Fare have missing values in both original datasets, so this research directly replaces the missing values with the mean value of the corresponding columns. By counting different values of the Embarked column, it can be concluded that passengers embarking from port ‘S’ account for over 70% of the total, therefore this research fills the missing values in training set’s port column with ‘S’. The Cabin column has nearly 80% missing values, hence this research doesn’t consider this feature and remove this column. Since the columns of Id, Name and Ticket are highly individual and have no apparent relation to survival rate, this research also doesn’t consider these features and eliminate them.

Table 1: Feature Description

Feature	Description
‘PassengerId’	PassengerId
‘Survived’	Passenger Survival Status
‘Pclass’	Passenger Class
‘Name’	Passenger Name
‘Sex’	Passenger Gender
‘Age’	Passenger Age
‘SibSp’	Number of Siblings/Spouses Aboard
‘Parch’	Number of Parents/Children Aboard
‘Fare’	Passenger Ticket Fare
‘Ticket’	Passenger Ticket Number
‘Cabin’	Passenger Cabin Number
‘Embarked’	Port of Embarkation

DOI: 10.54254/2754-1169/128/2024.18267

When dealing with non-numeric categorical data, a commonly used method is to encode it so that machine learning models can correctly understand and process this data. There are two main encoding methods can be used to handle this kind of data: One of them is known as Label Encoding method, which can convert categorical data into consecutive integers ranged from 0 to n-1(where n represents the overall number of unique categories). Another method is One-Hot Encoding method, which can convert categorical data into a vector. Each category corresponds to a dimension where the existing category dimension is set to 1 and other dimensions are set to 0. This research uses the get_dummies() function to perform One-Hot Encoding for both 'Sex' and 'Embarked' column. For the Sex column, it contains two unique genders: male and female, which are converted into two new columns: Sex_male and Sex_female with corresponding values of 0 or 1. For the Embarked column, it contains three unique features: S, C and Q, which are converted into three new columns: Embarked_S, Embarked_C and Embarked_Q to better utilize this categorical information in subsequent model training.

Table 2: Missing Data

Test data column	Missing value count	Predict data column	Missing value count
PassengerId	0	PassengerId	0
Survived	0	Survived	0
Pclass	0	Pclass	0
Name	0	Name	0
Sex	0	Sex	0
Age	177	Age	86
SibSp	0	SibSp	0
Parch	0	Parch	0
Fare	0	Fare	1
Ticket	0	Ticket	0
Cabin	687	Cabin	327
Embarked	2	Embarked	0

3. Results and Discussion

3.1. Gradient Boosting

Gradient Boosting is an ensemble learning method that builds a series of weak learners (typically decision trees) sequentially. The core idea of Gradient Boosting method is that each new tree is trained to correct the errors of the previous ensemble. The model can be represented as:

$$F(x) = \sum_{i=1}^{T} \gamma_i \, h_i(x) \tag{1}$$

where $F(x)$ is the final model, $h_i(x)$ are the weak learners aiming to predict residuals between the model's predicted value and the actual value, γ_i are the weights for each weak learner computed to control the contribution of each predictor. Key parameters include:

- n_estimators: Number of boosting stages
- learning_rate: Shrinkage applied to each tree
- max_depth: Maximum depth of individual trees

The study uses the function feature_importances_ attribute to reveal feature importance for Gradient Boosting model. These importance scores help to identify which features contribute most to the model's predictions and are shown in Table 3. The Feature Importance Table manifests a clear

DOI: 10.54254/2754-1169/128/2024.18267

understanding of the feature importance, with 'Sex_male' being the most influential feature accounting for 46.4%, indicating that gender especially male has a significant impact on survival rate. The research uses the GradientBoostingClassifier() function to build the model with a fixed random_state of 42. The gb_model.fit() function is then used to train the model based on training dataset, and it shows the accuracy of Gradient Boosting Classifier on the training set was 0.9073. This high accuracy manifests that the Gradient Boosting Model fits the training dataset well. To verify whether the model is overfitted, this research utilizes accuracy_score() function to get the accuracy of the model's predictions, and it manifests a model accuracy of 0.7988, and the average F1 score of the model is only 0.71. The sns.heatmap() function is used to visualize the resulting confusion matrix is shown in Fig. 1, and the plot_tree() function is utilized to show the exact decision tree of the Gradient Boosting Model in Fig. 2.

Table 3: Feature Importance of Gradient Boosting Model

Feature	Importance
Sex_male	0.4644
Fare	0.1799
Pclass	0.1527
Age	0.1492
Sibsp	0.0295
Embarked_S	0.0223
Parch	0.0020

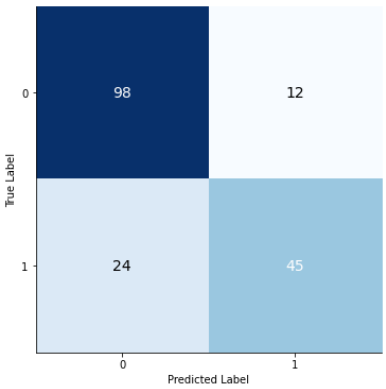

Figure 1: Confusion matrix of Gradient Boosting Model (Photo/Picture credit: Original).

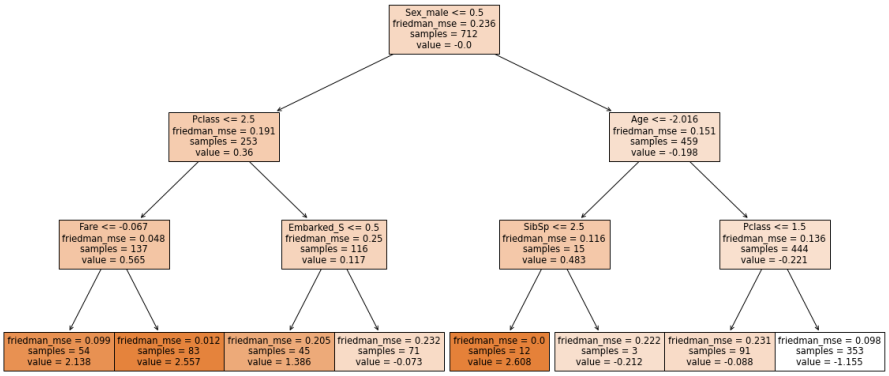

Figure 2: Decision tree of the Gradient Boosting Model (Photo/Picture credit: Original)

DOI: 10.54254/2754-1169/128/2024.18267

Figure 1 shows 12 passengers that were actually alive were predicted to be dead, and 24 passengers that were dead were predicted to be alive. It can be concluded that the precision of Gradient Boosting model is 0.89, and the Recall rate of this model is 0.69, which indicates that the Gradient Boosting model is overfitted.

3.2. XGBoost (Extreme Gradient Boosting)

XGBoost model is a more sophisticated version of Gradient Boosting. It uses a more regularized model formalization to control overfitting by adding new trees that correct the errors made by the existing trees. The objective function in XGBoost is:

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{i=1}^{t} \Omega(f_i) \quad (2)$$

where $l(y_i, \hat{y}_i)$ is the loss function, $\Omega(f_i)$ is the regularization term. Key parameters include:

- max_depth: Maximum tree depth
- learning_rate: Step size shrinkage
- colsample_bytree: Subsample ratio of columns for each tree
- subsample: Subsample ratio of the training instances

This study uses the model.get_score( )function to obtain feature importance from the model , the results are output in Table 4. The Feature Importance Table of XGBoost model manifests that “Fare” is the most important feature, accounting for 45.91% of the importance, followed by “Age” at 26.16%. This differs from the Gradient Boosting model, which considered 'Sex_male' as the most crucial feature. The XGBoost model highlights the significance of ticket price and age in survival prediction, suggesting that economic status and physical condition may be key factors influencing survival chances.

Table 4: Feature Importance of XGBoost Model

Feature	Importance
Fare	0.4591
Age	0.2616
Pclass	0.0667
Parch	0.0566
Sibsp	0.0526
Embarked_S	0.0491
Sex_male	0.039

This research uses the function xgb.train() to build the XGBoost model, setting a key parameter including max_depth of 6 to limit the maximum depth of each tree to 6 levels, helping to control model complexity and prevent overfitting; learning_rate of 0.1 to limit the contribution of each new subtree; colsample_bytree of 0.8 to limit each tree randomly samples 80% of the features when growing, which helps to increase the randomness of the model and prevent overfitting; subsample of 0.9 to ensure each tree uses only 90% of the training instances. Additionally, this research uses cross-validation (cv) to determine the optimal number of boosting rounds. It evaluates the model’s performance over 100 boosting rounds using 5-fold cross-validation (nfold=5) and selects the round of 37 with the minimum test log loss as the best round. This research then uses gb_model.fit() function based on the best selected round of 37 to train the model, and the output shows a model accuracy of 0.81 with the average F1 score of 0.73. This result indicates that the effect of XGBoost model is slightly better than Gradient Boosting model. Fig. 3 shows 12 passengers that were actually alive

DOI: 10.54254/2754-1169/128/2024.18267

were predicted to be dead, and 22 passengers that were dead were predicted to be alive. It can be concluded that the precision of Gradient Boosting model is 0.89, and the Recall rate of this model is 0.67. Learning Curve of XGBoost Model is shown in Fig. 4, which manifests that as boosting rounds increases, the log loss for both training and test sets decreases. These results shows that the XGBoost model keeps a stable learning process without significant overfitting or underfitting.

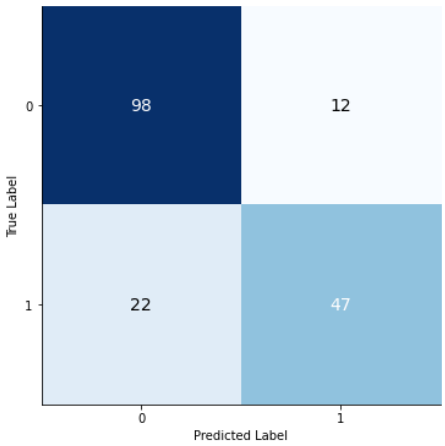

Figure 3: Confusion matrix of XGBoost Model (Photo/Picture credit: Original).

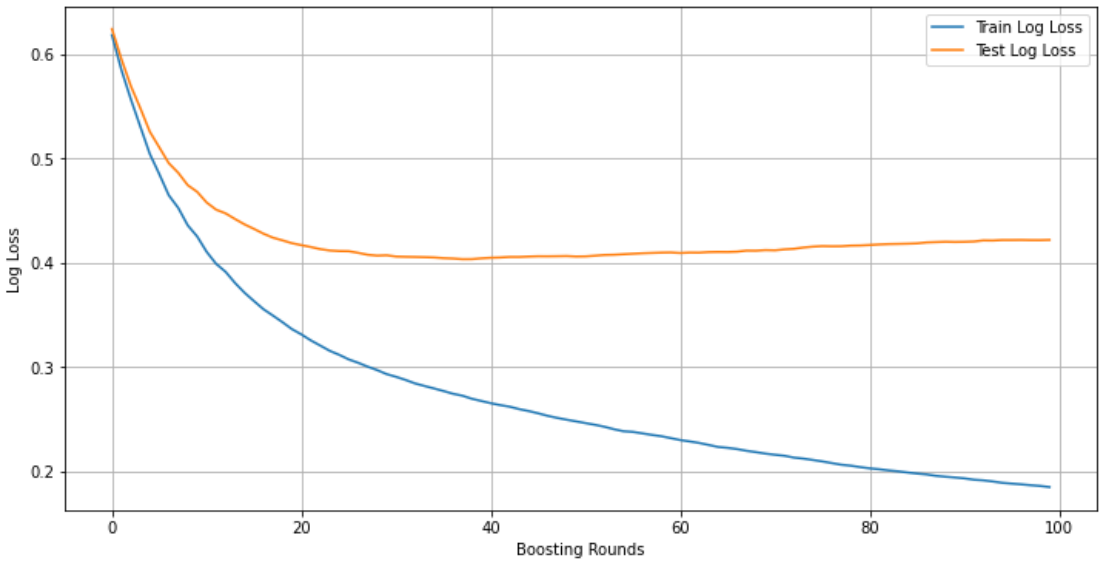

Figure 4: Learning Curve of XGBoost Model (Photo/Picture credit: Original).

3.3. RandomizedSearchCV-based Optimization

RandomizedSearchCV is a hyperparameter tuning method provided by scikit-learn that combines a randomized search with cross-validation. To further optimize model performance, the research employes this method for hyperparameter tuning by setting a paradigm of:

- max_depth: Randomly selected from the integer ranged from 3 to 10.
- learning_rate: Uniformly sampled between the range of 0.01 and 0.21.
- n_estimators: Randomly chosen from the integer ranged from 100 to 500.
- subsample: Uniformly sampled between the range of 0.5 and 1.0.
- colsample_bytree: Uniformly sampled between the range of 0.5 and 1.0.

DOI: 10.54254/2754-1169/128/2024.18267

The mean test score of each parameter is shown in Fig. 5 and Fig. 6.

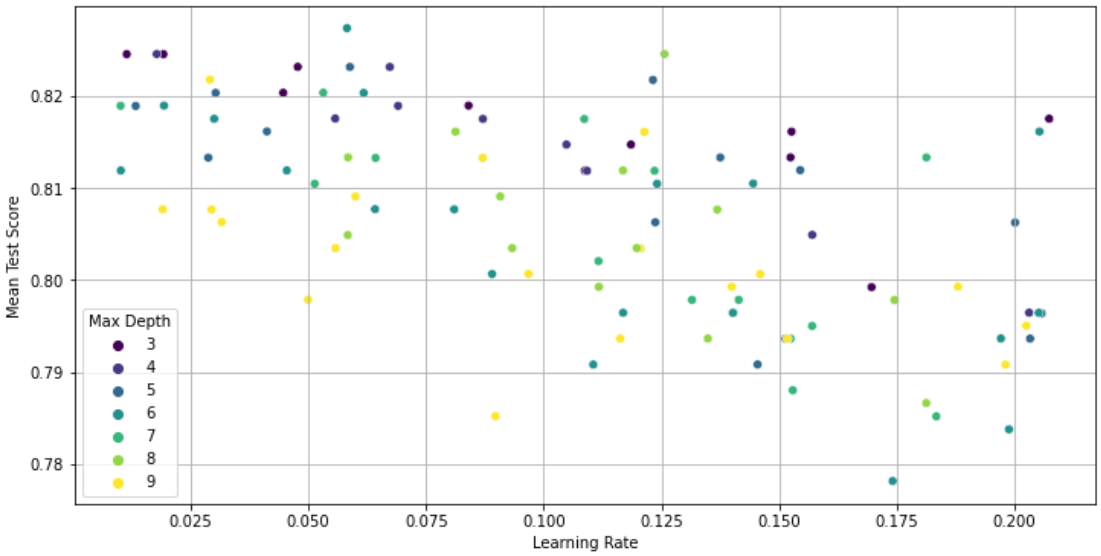

Figure 5: Learning Rate vs Mean Test Score (Photo/Picture credit: Original).

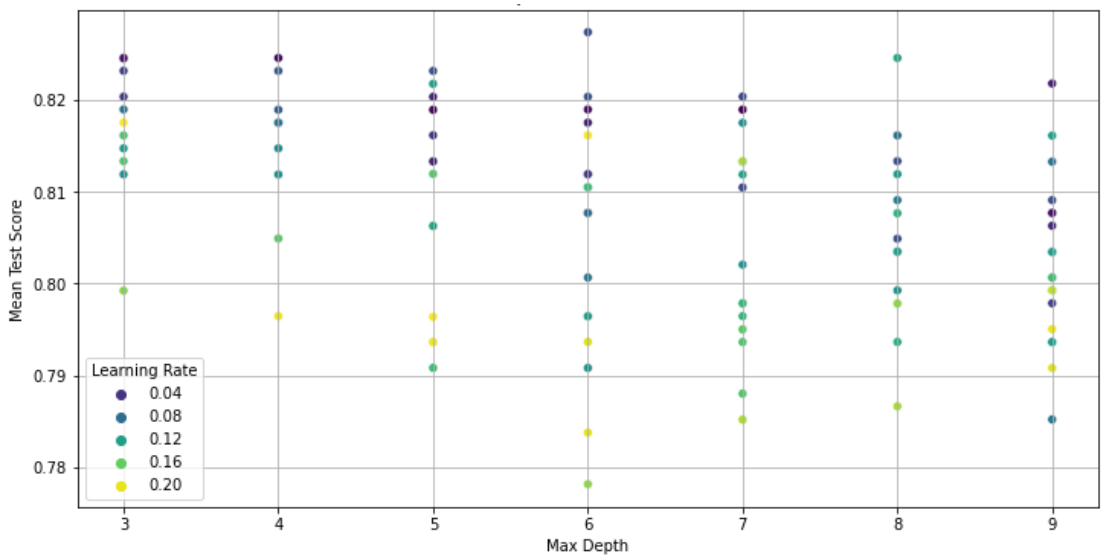

Figure 6: Max Depth vs Mean Test Score (Photo/Picture credit: Original).

4. Comparisons, Limitations and Prospects

This study employed several machine learning models including Gradient Boosting, XGBoost model and Optimized XGBoost model (based on RandomizedSearchCV) into the prediction of the survival rate of Titanic passengers. Through comparative analysis of the experimental results from these three predictive models, this research indicates that the Optimized XGBoost model demonstrates the highest precision and accuracy of 82.7% while being less prone to overfitting and underfitting issues. The Gradient Boosting model achieved an accuracy of 79.89%, and the XGBoost model outperformed it with an accuracy of 81.00%. This could be attributed to XGBoost's regularization parameters and its ability to handle a larger number of boosting rounds without overfitting. The Gradient Boosting model shows that 'Sex_male' is the most important feature, followed by 'Fare' and 'Pclass', indicating that gender especially male has a significant impact on survival rate. While the

DOI: 10.54254/2754-1169/128/2024.18267

feature importance table from XGBoost model manifests that 'Fare' and 'Age' were the most significant factors, which suggests that passengers who could afford higher fares and were at a certain age had a higher chance of survival.

Despite the promising results, this study has its limitations. The dataset used is historical and may not fully represent the complexity of modern datasets, which might affect their scalability to larger datasets. While this study provides valuable insights into predicting Titanic passenger survival probabilities, there is significant room for improvement and expansion. Future studies should integrate domain-specific knowledge with more in-depth hyperparameter tuning method to enhance model performance and accuracy.

5. Conclusion

In conclusion, this survival prediction model can be utilized to predict human survival in various natural disasters. According to the prediction results, corresponding rescue measures can be taken to improve the efficiency of rescue operations in natural disaster accidents. Although this research provides valuable prospects into the predictive modeling for the survival rate in the Titanic dataset, there is ample room for improvement and expansion. Future research could involve larger datasets, more complex models, and deeper integration with domain expertise to enhance the applicability of such models.

References

[1] *Hakkal, S. and Lahcen, A.A. (2024) XGBoost To Enhance Learner Performance Prediction. Computers and Education: Artificial Intelligence, 100254, 1-10.*

[2] *Punitha, N., Divya B.K., Manuskandan, S.R. and Karthick, P.A. (2024) Analysis of Muscle Fatigue Progression Using Geometric Features of Surface Electromyography Signals and Explainable XGBoost Classifier. Journal of Medical and Biological Engineering, 44(2), 191-197.*

[3] *Abolhosseini, S., Khorashadizadeh, M., Chahkandi, M. abd Golalizadeh, M. (2024) A modified ID3 decision tree algorithm based on cumulative residual entropy. Expert Systems with Applications, 255, 124821.*

[4] *Sun, J., Fan, C and Yan, H. (2024) SOH estimation of lithium-ion batteries based on multi-feature deep fusion and XGBoost. Energy, 132429.*

[5] *Patel, S.K., Wekalao, J., Mandela, N. and Al-Zahrani, F.A. (2024) Design of encoded graphene-gold metasurface-based circular ring and square sensors for brain tumor detection and optimization using XGBoost algorithm. Diamond and Related Materials, 111439.*

[6] *Zhang, X. (2023) Traffic Flow Prediction Based on Explainable Machine Learning. Highlights in Science, Engineering and Technology, 56, 56-64.*

[7] *Sherlock, J., Muniswamaiah, M., Clarke, L. and Cicoria, S. (2018) Classification of Titanic passenger data and chances of surviving the disaster. arXiv preprint arXiv:1810.09851.*

[8] *Singh, A., Saraswat, S. and Faujdar, N. (2017) Analyzing Titanic disaster using machine learning algorithms. In 2017 International Conference on Computing, Communication and Automation (ICCCA), 406-411.*

[9] *Ma, C. (2021) Comparison of Machine Learning Algorithms over Prediction of Titanic Database. Data Analysis, 18.*

[10] *Park, W., Shaby, N. and Newman, R. (2024) 'We Often Forget It Was a Disaster': Cross-Curricular Teacher Collaboration to Develop a Curriculum Unit on the Titanic Disaster. Science & Education, 1-28.*

[11] *Ao, H. and Yu, K. (2023) Design of Titanic Survival Prediction Model. Information Techniques, 11, 6-12.*