Predicting Passengers' Survivability in Shipwrecks: A Logistic Regression and SVM Approach for Search and Rescue Prioritization

Faldho Wirianandа Ho (A0244119J), Vincent Aurellio Budianto (A0244496U)

Abstract

Shipwrecks have been a common occurrence throughout history, often leading to the loss of many lives. One of the factors contributing to this is the issue of late rescue, where search and rescue teams may not prioritize their efforts to find survivors who can survive, resulting in reduced survival rates. To address this problem, we have implemented Support Vector Machine (SVM) and Logistic Regression, to analyze passenger information and predict whether a victim is likely to survive or not. This will allow search and rescue teams to prioritize their efforts to find survivors who require immediate medical attention, increasing their chances of survival.

Introduction

Shipwrecks have been tragic events throughout history, claiming the lives of numerous individuals. One significant factor that has contributed to these unfortunate outcomes is the issue of late rescue. In the aftermath of a shipwreck, it is important for search and rescue teams to promptly locate and aid survivors. However, when the rescue efforts are not efficiently targeted or prioritized, valuable time and resources may be wasted searching wider areas where many of the casualties were not expected to be alive. This can lead to a delay in finding survivors who are likely to survive and require urgent medical attention, resulting in late assistance and reducing their chance of survival. Therefore, it is important to prioritize searching for survivors who can survive in the event of a shipwreck, potentially reducing the unfortunate deaths associated with late rescue.

To address the problem of late rescue in shipwreck situations, we intend to use Machine Learning (ML) techniques to aid in the targeted and efficient search for survivors. Specifically, we will implement machine learning algorithms for classification, such as Support Vector Machine (SVM) and Logistic Regression. We used SVM and Logistic Regressions as both are reliable ML techniques for binary classification, which means it can distinguish between two categories or classes, such as survivor or non-survivor. These algorithms will analyze passengers' information to predict whether a victim is likely to survive or not. This will allow search and rescue teams to prioritize their efforts to find survivors and increase the survival rate in the event of a shipwreck.

This project is highly relevant to the module of IT1244 as it demonstrates the application of AI/ML techniques in solving a real-world problem related to the maritime industry. Furthermore, this project provides an opportunity for students to apply their knowledge of machine learning to develop a practical solution for a real-world problem and gain insights into the challenges and limitations of implementing AI/ML techniques in a specific domain.

Several studies have been conducted in recent years to predict the probability of survival for passengers in the event of a shipwreck. One significant result was made by a group of researchers led by Ekin Ekinci from Kocaeli University, which was published in 2018. The researchers implemented machine learning techniques on the Titanic dataset to predict the number of people who survived the incident. They found that the maximum accuracy obtained from Multiple Linear Regression was 78.426%, and the maximum accuracy obtained from Logistic Regression was 80.756%. However, their method had several drawbacks. One limitation was the lack of diversity in the dataset. The researchers used the Titanic dataset for the analysis, but it may not be representative of all shipwreck scenarios. This might have resulted in the model performing well on the training data, but not generalizing well to new data. Additionally, the dataset contained missing values that were replaced with median values, and the researchers did not perform feature scaling. Considering these factors would be crucial for improving our performance accuracy

To address the limitations of the previous study, we intend to utilize SVM and logistic regression models with feature scaling and different preprocessing techniques to handle the missing values. We expect that applying these techniques will enhance the comprehensiveness of our models, resulting in improved accuracy and greater robustness for predicting survival rates in shipwreck scenarios.

Disaster Dataset

A. Dataset

We have chosen to use the Disaster dataset for our analysis, which contains detailed information about the passengers who were on board when the incident occurred. The dataset includes various features such as passenger ID, ticket class, number of siblings or spouses on board, number of parents or children on board, ticket ID, ticket price, cabin number, departure location, gender, and age. Additionally, the dataset consists of 866 raw observations.

B. Preprocessing Steps

The dataset provided appears to have been cleaned previously, hence not much data cleaning is required. However, we did discover some missing values in the seat, departure_from, and age variables. Since we will not be using the Seat variable, we decide to remove it from the dataset. Additionally, there were only two missing entries in the departure_from variable, so we can remove those two entries without significantly impacting our model.

On the other hand, we found that the age variable proved to be a bit tricky to handle. We discovered that some age intervals have a higher survival chance than other as shown in the figure 2.1 . This means that the age variable has a significant impact on the prediction, therefore we cannot simply remove the variable from our models.

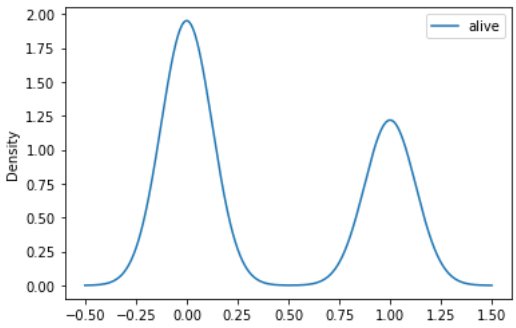

Figure 2.1

We considered three methods for filling in the missing values and concluded that using the means of the existing data is the best estimator for the age variable. This approach provides a more accurate estimate of the age variable compared to the previous works by Ekin, which used the median.

After obtaining the complete dataset, we created a correlation table to identify any potential multicollinearity between variables that could introduce noise into the model.

	class	num_sibling/spouse	num_parent/child	price	alive	age
class	1.000000	0.083923	0.013164	0.550570	0.331218	0.328359
num_sibling/spouse	0.083923	1.000000	0.424664	0.159181	0.039721	0.234771
num_parent/child	0.013164	0.424664	1.000000	0.219934	0.086437	0.192061
price	0.550570	0.159181	0.219934	1.000000	0.254530	0.084406
alive	0.331218	0.039721	0.086437	0.254530	1.000000	0.074360
age	0.328359	0.234771	0.192061	0.084406	0.074360	1.000000

Figure 2.2

Since none of the correlations were found to be significant (i.e., >0.75) enough to impact the model, we decided to keep all the variables in our analysis. We also noted that three of the variables (gender, class, and departure_from) are categorical variables. Thus, we created indicator variables (dummy variables) for each variable to transform them into numerical variables. Finally, We split the dataset into training and testing sets, with 80% allocated for training and 20% for testing, to facilitate further analysis and evaluation.

Methods

After experimenting with several models, we selected logistic regression and SVM as the most suitable methods for our model.

A. Logistic Regression

Our first model is based on logistic regression, which uses the sigmoid function to calculate the predictors, to produce binary outputs of "survived" or "not survived.". Our motivation behind logistic regression is the binary output it produces matches the distribution we initially observe when plotting some of the predictors. Furthermore, the sigmoid function allows for a probabilistic interpretation of the predictions, enabling us to understand the confidence level of our predictions.

Logistic regression works by minimizing the logistic loss function using the gradient descent algorithm. We implemented the model using the Scikit-learn library in Python, assigning a weight distribution of (0.6, 0.4) and a maximum iteration of 500. This weight distribution was chosen after a series of testing shown below to reduce false negative predictions while maintaining accuracy since we do not want the rescue team to overlook a potential survivor.

`

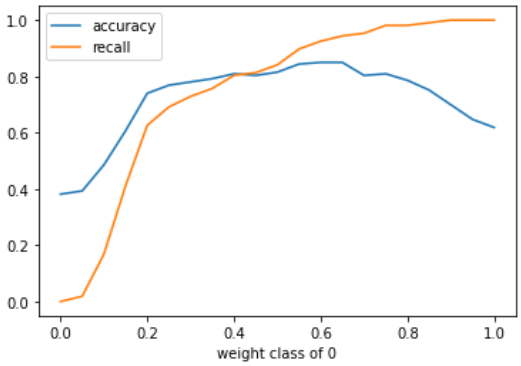

Figure 3.1.1

After considering the trade-off in figure 3.1.1 we decided to use maximum accuracy and maximize the recall afterward. It will be discussed later regarding the decision of why recall specifically. The model obtained provides a good estimate and prediction for most individuals, as demonstrated by the following confusion matrix.

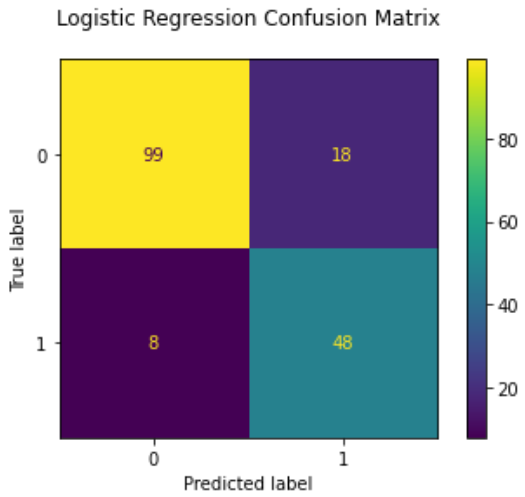

Figure 3.1.2

B. *SVM*

SVM is practically a similar model to logistic regression, but with a different method for outputting values. While logistic regression uses a sigmoid function, SVM uses a hyperplane to separate all points in the dataset. It makes SVMs effective in dealing with high-dimensional data, where there are many corresponding features. In the case of the Disaster dataset, there are several features that might be relevant to predicting survival, such as age, gender, class, and ticket price. SVMs can handle such high-dimensional data better than logistic regression.

Unlike logistic regression, SVM does not require the use of interaction terms or splines to optimize the model. It is capable to transform data that cannot be easily separated into a more complex space, where it becomes easier to find a decision boundary. This allows SVM to work well with complex datasets where simple decision boundaries cannot be easily identified.

To create an SVM model, we used the same Scikit-learn library used to build the logistic regression model. We specified the kernel as linear with a C value of 1. We tried other kernels such as sigmoid and radial basis, but the linear kernel worked best in this case. Surprisingly, the linear kernel works better than the sigmoid kernel, meaning that the decision boundary between the classes is better represented by a linear separator than a sigmoidal one. However, It is important to note that SVM and logistic regression have different ways of optimizing their models, which can result in different outcomes.

We also specify the class weight to consider in this model although it is different from class weight distribution in logistic regression. We considered the trade-off in the following graph regarding accuracy and recall and decided to use (0.55,0.45) as the class weight for SVM.

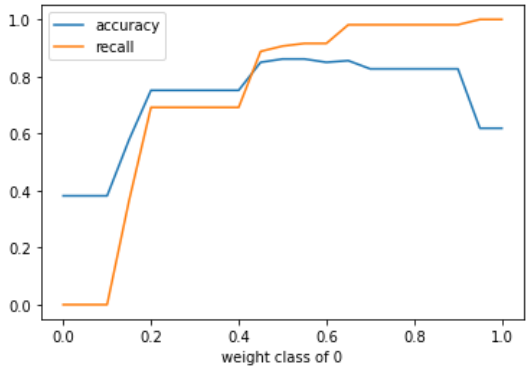

Figure 3.2.1

The confusion matrix for the SVM model with a linear kernel is shown below.

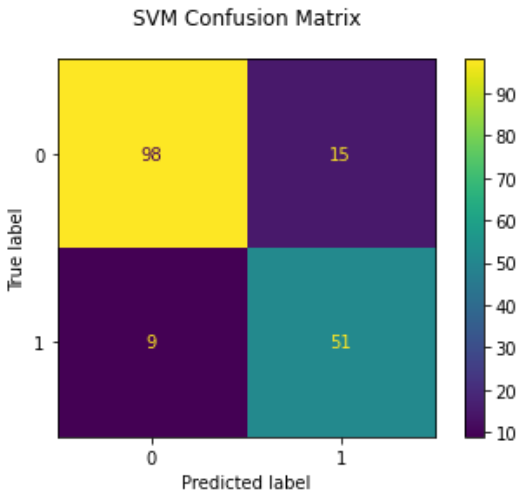

Figure 3.2.2

Results & Discussions

To predict survivability, we conducted multiple experiments by evaluating logistic regression and SVM models on the Disaster dataset. Specifically, we performed five experiments using different approaches for both logistic regression and SVM. We utilized several performance metrics, such as precision, recall, and F1-score, to assess the accuracy of the models. Subsequently, we carefully analyzed each experiment to determine the optimal preprocessing techniques and parameters that yielded the highest accuracy.

To fine-tune our performance, we experimented with different hyperparameters for each SVM and logistic regression model. We tested different kernel functions, C values, and class weights in our SVM model, while in logistic regression, we experimented with different weight classes and maximum iteration. Additionally, we also performed feature engineering on both models to select the most important features that have the highest impact on the model's performance. This would also enable us to overcome the challenges of the previous work.

The table below shows the final result of our experiment :

Method	**Accuracy**	**Precision**	**Recall**	**F1-score**
SVM	0.8613	0.8673	0.9159	0.8909
Logistic Regression	0.8497	0.8462	0.9252	0.8839

Based on the table of results above, our findings suggest that the SVM model outperforms the logistic regression models in terms of accuracy, precision, and F1-score. This can be attributed to the fact that SVM can handle non-linearly decision boundaries by mapping the features to a higher-dimensional space using the kernel function. This allows SVM to identify more complex patterns in our Disaster data that may be missed by logistic regression, which uses a linear decision boundary. Furthermore, our SVM model achieved a higher accuracy rate of 86.13% compared to the previous work by Ekin, which had a maximum accuracy of 80.756%.

However, when it comes to recall score, logistic regression has a higher score than SVM, with a margin of 0.0093. This is associated with a lower number of false negative predictions in logistic regression. Considering ethical considerations, it is crucial to minimize false negative predictions from the evaluation matrix, especially if the model is used to allocate resources or prioritize rescue efforts. We could tolerate non-survivors being predicted as survivors, but we cannot tolerate survivors being predicted as non-survivors. Therefore, in this particular case, logistic regression is more appropriate than SVM because it has a higher recall score.

In terms of accuracy, SVM is better for predicting survivability in a shipwreck event. However, in view of moral justification, it is important to have a 100% recall score (0 false negative) ensuring all alive passenger is prioritized, but it is impossible to have a 100% recall score without sacrificing a significant amount of accuracy. Although logistic regression produces less false negatives, it doesn't mean that it is morally justified, therefore there is a need to find a new model with 0 false negatives without sacrificing accuracy (as shown in Figures 3.1.1 and 3.2.1). In the end, neither logistic regression nor SVM is the ideal solution, but SVM is still preferred due to its higher accuracy score. Furthermore, both proposed model is still prone to overfitting as the dataset consisted of small observations from 866 passengers. Hence a larger dataset is needed to improve the models' performance in general shipwreck scenarios.

In conclusion, our experiments demonstrate that machine learning algorithms can effectively predict survival rates in shipwreck scenarios. We have selected the Disaster dataset, and after performing some preprocessing steps, we have applied logistic regression and SVM as our modeling techniques. This project demonstrates the practical application of AI/ML in solving real-world problems within the maritime industry, with the hope of improving survival rates in such scenarios. While our experiments have shown promise in predicting survival rates, ethical considerations must be taken into account, and further research is needed to find a new model with 0 false negatives without sacrificing accuracy.

References

Ekinci, E., Acun, N., & Omurca, S. İ. (2018). A Comparative Study on Machine Learning Techniques Using Titanic Dataset. 7th International Conference on Advanced Technologies- ICAT'18, April, 411–416. https://www.researchgate.net/publication/324909545