Predictive Model for Disaster Dataset

Ang Lin Xuan, Chew Yu Cai, Yune Thiri Khin

Abstract

This project explores different models used to predict the survivability of passengers of a ship that met with a disaster. This allows for the possible development of a scoring system for the survivability of passengers which could allow for the crew to save those at risk, as well as recommendation of actions by the passengers or authorities to maximise survivability. Based on our research on relevant works, this is similar to the prediction models healthcare systems use, such as the Trauma and Injury Severity Score (de Munter et al., 2018). The use of such predictive models allows for the better management of the disaster based on factors provided in the dataset (Acheme & Vincent, 2021).

Introduction

This project aims to create a predictive model for passengers in a ship that met with a disaster. We have used the k-nearest neighbours (kNN), logistic regression (LR), decision tree (DT), and neural network (NN) as our models to predict survivability, which have been covered through the course of the module. To assess which of the four is best at predicting the survivability of the passengers on board the ship, we will use the performance metric of accuracy, which is the ratio of correctly predicted outcomes to total outcomes.

We chose kNN classifier as it is simple yet reliable, with high accuracy in making predictions. It is especially suitable as our dataset is small, and has only 6 attributes to consider after we processed the data. This ensures that computations are fast, and the low dimension reduces noise and allows the model to perform more optimally (Tokuç, 2021). We chose NN because it is able to model and adapt to non-linear and complex relationships between different variables. It also allows for the output of results without the complete inputs after it has been trained, allowing for the model to be generalised to other similar datasets or disasters. We chose LR as it allows us to make predictions on the binary survival feature based on the other independent attributes, by estimating probabilities based on a logistic function. Furthermore, it is easy to implement, and efficient to train.

Additionally, a DT was implemented as it will allow us to predict the survival of the passenger based on multiple variables from our given dataset (Song & Lu, 2015). On top of being quick and easy to compute, one of the advantages of DTs is the ability to visualise the decision flow of the model. This visualisation allows for the quick understanding of which factors impact the survivability of the passengers the most.

Dataset Used

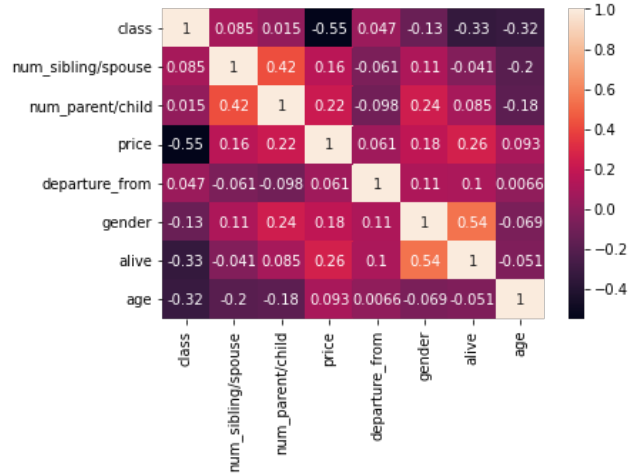
The dataset contains the ship's passenger details, which presented some issues upon exploration using pandas. The dataset is multivariate with a total of 866 instances and 12 attributes, containing missing values. Firstly, there were features that may not be essential to our data analysis. Secondly, there are missing values for some of the features. Lastly, some features are categorical which would require a change to numerical values to fit into the models.

Methodology

As our objective is to determine which of the four models would be the best in predicting the survivability of the passengers, we decided to directly compare their performances. However, before we could proceed with the models, we needed to process the data due to the challenges presented in the dataset.

To deal with irrelevant features, the "unnamed", "id" and "ticket" columns were dropped. In addition, the "price" column was removed as we found it was related to the "class" attribute, as observed from our correlation heatmap (Diagram 1). This correlation implies dependency between the two variables which may give rise to potential problems and challenges (Frost, 2020).

Diagram 1
Correlation heatmap of different features



To deal with the missing values for the features "departure_from" and "age", we cleaned the data and filled in the missing values by making predictions based on the other available data. As "departure_from" was missing two values, and is categorical, the data was completed by filling it with the most common occurrence. For "age", as it is numerical and continuous, random integers were generated between the mean and standard deviation of the ages instead.

Finally, we converted the categorical features, "gender" and "departure_from", into numerical values. We then split the data into training and testing sets, with 20% of the data as the test set. Following that, to ensure that the data does not have differing scales, we standardised both the data and test sets.

Once these were done, we then proceeded to train the four models based on the training data. To prevent overfitting and reduce prediction bias, we performed k-fold cross-validation. For our dataset, we chose k to be 5. This also allowed us to maximise the use of our relatively small dataset, in training and testing our models. Furthermore, in our kNN model, we ran an algorithm to ensure that the best k-value providing the highest accuracy is used in each iteration of the model, which ensured that the model is providing the best prediction for our testing dataset. We then assessed the accuracy of the four models based on its predictions on the testing set features. Then, we proceeded to iterate through this process 50 times so as to gain a more reliable median of the accuracies for all four models, which would be our determining factor for the best model to predict.

Results & Discussion

Table 1
Average accuracy of the models, from 50 iterations

Model	Average Accuracy
kNN	83.58%
NN	81.38%
LR	81.30%
DT	77.37%

After repeating the process 50 times, our results (Appendix 1) indicated that kNN algorithm had the highest accuracy averaged at 83.58% (Table 1), making it the optimal model to use for prediction of the ship passengers' survival. Meanwhile, the LR model had 81.30% average accuracy, the NN model had 81.38% average accuracy, and DT had 77.37% average accuracy.

As mentioned in our introduction, a reason kNN had the performance would likely be due to our small dataset and low dimensionality, making it exceptionally suitable for our project. However, from what we have learned through the module and related works, the more the data, the better NN should perform compared to other machine learning models. This is because there is more training data for the NN to learn from, enabling NN to automatically identify features to predict outcomes. If we were to run this same experiment on a significantly larger dataset, we may see that NN could be the most optimal instead. Hence, for future works, it may be beneficial to investigate the performance of the different models with respect to different sizes of data. We could also use other performance metrics such as precision, in assessing the performance of the models, as we only focused on accuracy.

Conclusion

We would choose kNN as the best model to predict the survivability of the passengers. However, this may only be so in smaller datasets with few attributes such as this dataset. In much larger datasets with more data points or higher dimensionality, the deployment of a NN model may be more effective and efficient instead, and this is something we can further analyse in future projects.

References

Acheme, I. D., & Vincent, O. R. (2021). Machine-learning models for predicting survivability in COVID-19 patients. *Data Science for COVID-19*, 317–336. https://doi.org/10.1016/b978-0-12-824536-1.00011-3

de Munter, L., Ter Bogt, N., Polinder, S., Sewalt, C. A., Steyerberg, E. W., & de Jongh, M. (2018). Improvement of the performance of survival prediction in the ageing blunt trauma population: A cohort study. *PloS one*, *13*(12), e0209099. https://doi.org/10.1371/journal.pone.0209099

Ranganathan, P., Pramesh, C. S., & Aggarwal, R. (2017). Common pitfalls in statistical analysis: Logistic regression. *Perspectives in clinical research*, *8*(3), 148–151. https://doi.org/10.4103/picr.PICR_87_17

Tokuç, A. A. (2021, October 13). *K-nearest neighbors and high dimensional data*. Baeldung on Computer Science. Retrieved October 31, 2022, from https://www.baeldung.com/cs/k-nearest-neighbors

Appendix 1

Results of accuracies of the different models from the 50 iterations of running the models

Iteration	kNN	LR	NN	DT
1	85.06%	81.61%	80.46%	78.16%
2	83.91%	82.18%	80.46%	75.86%
3	83.91%	81.03%	80.46%	75.29%
4	82.76%	81.61%	81.03%	75.29%
5	83.91%	81.03%	83.33%	77.59%
6	82.18%	81.61%	79.89%	78.74%
7	83.91%	80.46%	83.33%	78.74%
8	82.18%	79.89%	79.89%	76.44%
9	83.91%	80.46%	81.61%	75.86%
10	83.33%	81.61%	81.03%	75.29%
11	82.18%	82.18%	82.18%	74.71%
12	85.06%	83.91%	81.61%	75.86%
13	83.91%	81.03%	81.03%	73.56%
14	83.91%	80.46%	79.89%	77.59%
15	82.76%	79.89%	81.03%	78.16%
16	82.18%	81.61%	81.61%	78.16%
17	83.91%	81.03%	81.03%	77.01%
18	82.18%	81.03%	81.03%	78.74%
19	83.91%	79.89%	80.46%	77.59%
20	82.76%	82.18%	81.03%	77.01%
21	85.63%	80.46%	82.18%	77.59%
22	83.33%	82.18%	82.76%	78.16%
23	83.33%	81.03%	81.61%	77.01%
24	82.18%	81.03%	78.74%	76.44%
25	84.48%	81.61%	78.16%	75.86%
26	83.33%	80.46%	79.89%	78.16%
27	83.91%	80.46%	82.76%	79.31%
28	84.48%	81.03%	83.91%	78.74%
29	85.06%	82.18%	81.61%	75.29%
30	84.48%	81.61%	81.03%	79.89%
31	84.48%	81.61%	83.33%	74.71%
32	82.18%	81.03%	76.44%	77.01%
33	83.91%	81.03%	81.03%	73.56%
34	83.91%	81.03%	81.61%	79.31%
35	83.91%	81.61%	82.18%	76.44%
36	82.76%	81.61%	81.61%	78.74%
37	83.33%	81.61%	82.18%	79.31%
38	83.33%	82.18%	81.61%	81.61%
39	82.76%	81.61%	81.61%	78.16%
40	82.76%	82.18%	82.18%	76.44%
41	83.33%	80.46%	82.18%	73.56%
42	83.91%	81.61%	81.61%	81.61%
43	85.06%	81.03%	81.61%	79.31%
44	83.91%	81.03%	83.33%	78.16%
45	85.06%	81.03%	82.18%	75.86%
46	82.76%	81.61%	81.61%	79.31%
47	82.76%	81.03%	83.33%	78.74%
48	84.48%	81.03%	82.18%	79.31%
49	83.33%	82.18%	81.03%	75.86%
50	82.76%	81.61%	81.03%	79.31%

	Average
kNN	83.58%
LR	81.30%
NN	81.38%
DT	77.37%