Datathon-2 Report

Introduction

Heart failure is a common and severe cardiac condition with high morbidity and mortality.[1] Around half of the heart failure patients who have reduced ejection fraction (EF ≤ 40%) exhibit worse clinical manifestations and prognosis.[1,2] Previous research have showed that age, male, low systolic pressure, anemia, hyponatremia, kidney diseases, and other risk factors are negative predictors.[2] We aim to predict the mortality of heart failure population using logistic regression and K-nearest neighbour (KNN) models and evaluate their application in clinical practice.

Data Engineering Process

The data set used in this study is the morality data with death event as the outcome. We first imported and read the mortality dataset, then checked for missing values. We followed by exploratory data analysis, including data summary and pair plots, box plots and bar plots to examine the relationships between features.

We selected features based on a comprehensive consideration on clinical evidence and multicollinearity. Since follow-up time is strongly associated with death event, we removed time for both models. We then examined variance inflation factor (VIF) of all features in the data set, where age, serum sodium, and ejection fraction had high VIF values. After excluding the features that were not of interests in this study, namely smoking, diabetes, and platelets, we saw a decrease in the VIF of age, serum sodium, and ejection fraction. Furthermore, based on the pair plots, all continuous features approximately followed normal distribution, and thus, we used a standard scaler to normalize the data and to be used in our models. For fitting the model, we split the data into 80% training set and 20% testing set.

Analysis

Since our dataset is properly labelled with 0 or 1 for the patient's death status**,** it is appropriate to apply supervised learning models such as logistic regression and KNN. We first fitted a logistic regression model using the selected variables, age, sex, ejection fraction, anaemia, high blood pressure status, creatinine phosphokinase, serum creatinine and sodium, to predict the binary outcome death event. To prevent overfitting, we used the L2 penalty and regularization strength of 1. Since our outcome is unbalanced (203 non-death versus 96 death events), we further fitted a separate weighted logistic regression model, which assigned an inversely proportional weight to the minor class, to compensate for this class imbalance. In addition, we looked at the confusion matrixes with precision and recall scores to provide us with additional information on our model performance. Furthermore, we included summaries of the full model with all features and our fitted model for estimations of coefficients and p-values of each feature included.

Alternatively, we utilized KNN to predict the occurrence of death event in the dataset. To select the optimal K, we plotted error rate against the K values ranging from 1 to 50, and we found that the error rate was minimized to 0.35 at K = 5 before increasing to an average error rate around 0.38. Thus, the optimal K selected for the KNN model in this study was 5. We then evaluated the performance of our model using classification report and confusion matrix and compared the results between KNN and logistic regression models.

Findings

Our sampled patients were middle aged (mean: 60.83 ± SD 11.90). 194 (64.88%) were male and 105 (35.12%) were female. A substantial number of patients had reduced ejection fraction (mean: 38.8% ± SD 11.83%). The average follow-up time was 130 days (range: 4 to 285 days) indicating our dataset could be used to predict mortality within one year. We found that features with stronger associations with heart failure mortality in this dataset were age, ejection fraction, serum creatinine. Other clinically significant predictors (e.g., NT-ProBNP[2]) were not available in our dataset.

The overall accuracy of our logistic regression model is 0.65 on the testing dataset. It performed relatively satisfactory in terms of precision (no death event = 0.63, death event = 0.75). However, we noticed that the recall score for death event is lower (0.24) compared to that of no death event (0.94), which means our model performs less well in predicting true mortality and would potentially miss some positive cases. Nevertheless, the weighted logistic model gave higher overall accuracy (0.75) and average precision (0.75), as well as a 40% higher outcome class 1 recall (0.64). From both full and fitted model summaries, the p-values for features age, ejection fraction and creatinine phosphokinase are shown to be statistically significant at a level of 0.05.

Using a KNN with K = 5, we obtained a training accuracy of 0.78 and a test accuracy of 0.65. The precision for observing and not observing a death event are 0.63 and 0.83, respectively. However, the recall scores of two outcomes are less comparable, where the model would achieve a 0.97 and 0.20 recall score for having or not having a death event, respectively. Again, we obtained a low recall for outcome class 1, meaning that the model is conservative in predicting true for death event. To compare the models, both non-weighted logistic regression and KNN models have a moderate overall accuracy and a low recall for positive death event. However, the precision for death event occurring in KNN model is the highest among all models, meaning that KNN model makes relatively few false positive predictions. Noticeably, while the recall score of the logistic model with a balanced weight increases substantially compared to the two unadjusted models, the resulting recall for death event occurrence is still moderate.

Conclusion

The models in this study have the capacity to accurately detect true positives, making them potentially useful in the identification of patients who may require enhanced surveillance and medical care. They serve as supplementary tools for clinicians' decision-making, though they should not be the primary factor in clinical determinations. The lower recall score exhibited by both logistic and KNN models in recognising death events suggests a possibility of overlooking actual positive occurrences, making them limited in clinical practice. Clinicians should be aware of this limitation and not rely solely on the models to identify high-risk patients.

Contributions

Yutong Lu: Coding for KNN, writing analysis, findings and conclusion, and preparing slide
Feifan Xiang: Coding for logistic regression, writing data engineering, analysis, and preparing slide.
Xue Man: Coding for logistic regression, writing introduction, findings, and preparing slide.
Github: https://github.com/Yutong-Lu/Datathon-2
Presentation: https://tinyurl.com/27z9sdpr

References

1. Murphy SP, Ibrahim NE, Januzzi JL. Heart Failure With Reduced Ejection Fraction: A Review. JAMA : the journal of the American Medical Association. 2020;324(5):488–504.
2. Jones NR, Hobbs FR, Taylor CJ. Prognosis following a diagnosis of heart failure and the role of primary care: a review of the literature. BJGP open. 2017;1(3):bjgpopen17X101013–bjgpopen17X101013.