Team 14: Abdulaziz Sherif, Priyonto Saha, Rohini Datta
October 2nd, 2023

Datathon 2: Early Prediction of Cardiovascular Events using Logistic Regression

Introduction:
Cardiovascular disease refers to health conditions impacting the blood circulatory system.[1] Given its prevalence globally, this study aimed to explore the factors associated with the occurrence of a cardiovascular event (heart disease, stroke, and/or hypertension). This study also aimed to explore the health factors and lifestyle factors associated with mortality among patients with heart failure.

Data Engineering Process:
The *Cardiovascular Event* dataset (CED) and the *Mortality for Cardiovascular Disease Complications* dataset (MD) were used to analyze the first and second objective, respectively. We started with an initial exploratory data analysis which consisted of missing data analysis, summary statistics, pairwise contingency matrices for categorical predictors, and a correlation matrix for continuous predictors. Histograms were plotted to visualize the distribution of the predictors, along with histograms for the distribution of continuous predictions subpopulations stratified by the categorical predictors.

For MD, there was not much preprocessing required since all the variables were encoded as numeric, there was no missing data, and all variables were either binary or continuous which works well with logistic regression models. A box plot was created to visualize any outliers, then the data was cleaned by removing those outliers. For CED this was not the case as many variables were encoded as text and there were nominal variables that needed to be analyzed.

For CED, the unique identifier ‘id’ was stripped from the dataset immediately and a single entry with the gender “Other” was dropped as an outlier. There were 201 entries with missing ‘bmi’ values, and 1544 entries with unknown “smoking_status”. Due to the high amount of missingness for “smoking_status”, and due to being a mix of both ordinal and categorical, this predictor was removed. Mean imputation was used to deal with the missing data for ‘bmi’. We note there should be correlation between age and work type, as “children” implies a young age. However, we find that there are 169 individuals under the age of 18 who are not classified as “children”, and that all of the 22 “Never_worked” individuals are under the age of 23. Thus the “work_type” variable was removed. A feature called “event” was created to be the response variable, with value 1 if any cardiovascular event is experienced, and value 0 otherwise. This accounts for the comparatively low number of patients with hypertension, heart disease, or that have had a stroke.

Analysis:
Logistic regression (LR) was the machine learning technique selected for this analysis given the binary classification of both outcomes, cardiovascular event (0 or 1) and mortality status (0 or 1). Using LR permitted the odds ratio for the outcome of interest while controlling for multiple covariates to be obtained. The assumptions for logistic regression were verified and the most clinically relevant variables were selected to minimize the presence of multicollinearity. Finally, outliers for continuous variables were visualized using a box-plot and subsequently removed from the final model. To prepare the data, each dataset was split into a training set (80%) and test set (20%) and the data was standardized. A logistic model was fitted for each outcome of interest including its respective covariates using the training data. Using the test data for each model, we predicted the outcome and then evaluated the model using the confusion matrix and classification report. The results of the models were visualized and summarized for interpretation.

Team 14: Abdulaziz Sherif, Priyonto Saha, Rohini Datta
October 2nd, 2023

Findings:

Logistic Regression for Cardiovascular Event Data

The processed variables were analyzed for multicollinearity, and age, BMI, and the reencoded work status variable had a VIF > 10. Given that BMI and work status tend to be correlated with age, conceptually, BMI and work status were not included in the final model. The remaining continuous variables of age and average glucose level were verified for linearity with log odds. Outliers were identified for average glucose level, and those observations were removed.

The final LR model for the event of heart disease, stroke, and/or hypertension included the covariates of gender, age, marital status, residence type, and average glucose level (Table 1). Gender (OR=1.504, $p<0.001$), age (OR=1.068, $p<0.001$), and average glucose level (OR=1.004, $p=0.002$) are significantly associated with the outcome of interest at the 95% level. The accuracy was 0.88, indicating that 88% of predictions made by the LR model were correct. The pseudo R-squared value was 0.2163 which indicates that the model explains 21.63% of the variation in the outcome of interest.

Logistic Regression for Mortality Data

Overall, the mortality logistic regression model suggested that age, ejection fraction, serum creatinine, and time were significant predictors of mortality status while the other features did not have a significant impact in explaining the variation in mortality status based on their p values being greater than 0.05 (Table 2). The pseudo R-squared value was 0.4361 which suggests that the model explains approximately 43.61% of the variation in the mortality status. The LLR p-value was 8.007e-19. The low p-value suggests that the model is significantly better than a null model with no features.

The prediction model had an accuracy of 0.82 which indicates 82% of the predictions are correct. The model was very good at predicting when death would not occur but it performed poorly for predicting a death event with a recall score of 0.45.

Conclusion:

The cardiovascular event model and the mortality logistic regression model were able to interpret which features are significant predictors of a cardiovascular event and death respectively. Gender, age and average glucose levels significantly increase the log odds of a cardiovascular event. As age and serum creatinine increase, the log odds of death increase significantly and as ejection fraction and duration of follow-up period increase the log odds of death decrease significantly. Health practitioners can use these factors to assess the risk of death or to determine risk factors of cardiovascular health. Despite the accuracy being very high for both prediction models, caution should be taken when using it for prediction. The recall for the mortality model was low for the event of death (0.45) most likely due to the fact that there was a low test sample size of death. Similarly, the cardiovascular event prediction model had a high accuracy of 88% but a low recall of the outcome of interest, meaning the model missed a lot of people with a cardiovascular event.

The datasets have contrasting limitations. CED has many observations, but they may be considered irrelevant due to missing or unknown data and inconsistencies in variables. Comparatively, MD has very few observations, with less than a third experiencing the outcome of interest. Since the models are limited by the data, our modeling process may produce better results given more complete data.

Team 14: Abdulaziz Sherif, Priyonto Saha, Rohini Datta
October 2nd, 2023

Individual Contributions:

Abdulaziz Sherif: Logistic Regression for MD, Analysis, Findings, Conclusion
Priyonto Saha: Exploratory Data Analysis, Preprocessing, Feature Engineering, Conclusion (All Data)
Rohini Datta: Logistic Regression for CED, Presentation, Introduction, Analysis, Findings

Code and Presentation:
All materials pertaining to Datathon 2 for Team 16 is hosted on Github: Datathon 2 Github

The presentation slides are provided at the following link: 14-CHL5230-F23

References:

1. Thiriet M. (2019). Cardiovascular Disease: An Introduction. Vasculopathies: Behavioral, Chemical, Environmental, and Genetic Factors, 8, 1–90. https://doi.org/10.1007/978-3-319-89315-0_1

Appendix:

Table 1. Summary of logistic regression for cardiovascular event dataset

	Coefficient	Standard Error	p-value	95% CI
Constant	-5.7219	0.232	0.000	(-6.177, -5.266)
Gender	0.4082	0.096	0.000	(0.220, 0.596)
Age	0.0657	0.003	0.000	(0.060, 0.072)
Marital status	-0.0054	0.147	0.971	(-0.293, 0.283)
Residence type	-0.0679	0.095	0.476	(-0.255, 0.119)
Average glucose level	0.0038	0.001	0.002	(0.001, 0.006)

Table 2. Summary of logistic regression for mortality dataset

	Coefficient	Standard Error	p-value	95% CI
Constant	3.4922	8.673	0.687	(-13.507, 20.492)
Age	0.0499	0.020	0.012	(0.011, 0.089)
Anaemia	0.3694	0.447	0.409	(-0.507, 1.246)
Creatinine Phosphokinase	0.0007	0.001	0.336	(-0.001, 0.002)
Diabetes	0.0981	0.436	0.822	(-0.756, 0.952)
Ejection fraction	-0.0713	0.020	0.000	(-0.110, - 0.033)
High blood pressure	-0.2434	0.455	0.592	(-1.135, 0.648)
Platelets	-1.115e-06	3.28e-06	0.734	(-7.55e-06, 5.32e-06)
Serum Creatinine	2.2455	0.739	0.002	(0.796, 3.695)
Serum Sodium	-0.0378	0.060	0.0530	(-0.158, 0.080)
Sex	-0.5191	0.541	0.337	(-1.580, 0.541)
Smoking	0.6158	0.532	0.247	(-0.427, 1.658)
Time	-0.0220	0.004	0.000	(-0.029, -0.015)