Statistical Model for Predicting CVD in Patients

Deepak Dhankani (ded4001)

Weill Cornell Medicine

ABSTRACT

Cardiovascular diseases account for 1 in 4 deaths every year in the US as well as Europe and there are significant medical costs associated with treatment of cardiovascular conditions. Early detection and prevention by classifying patients at high risk can lead to prolonged life expectancy of patients and ease off significant medical procedure costs. Patient data related to cardiovascular tests was collected for a one-month period from four tertiary care cardiac hospitals (two in the US, two in Europe). This data was used to build an ordinal logistic regression model to evaluate the role of different test results associated to the stages of vessel blockage among patients. A binominal logistic regression model was built to predict the patient disease outcome. Lastly, the models were adjusted based on the costs associated to each test.

OBJECTIVE

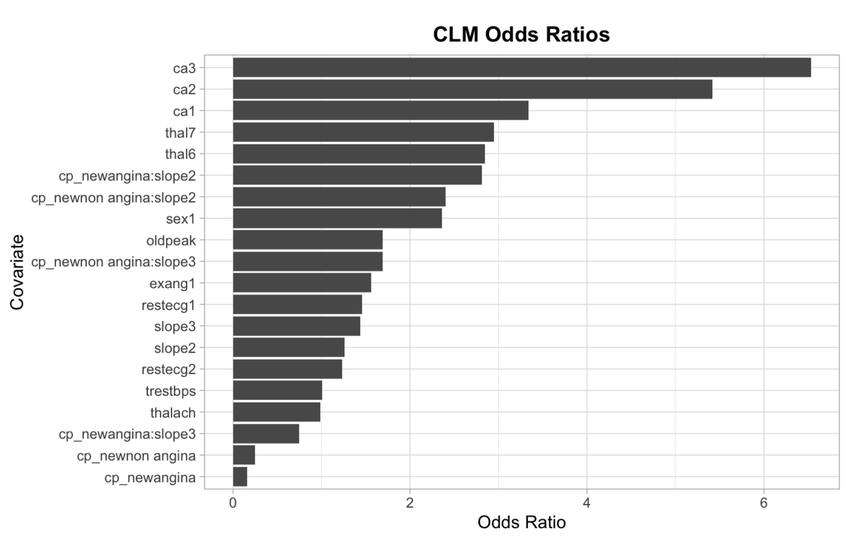
Analyze patient data from four tertiary care cardiac hospitals to identify the effectiveness of test results in predicting the cardiovascular vessel blockage stage among patients. The prediction model also needs to encompass the associated costs and lag times for better resource allocation.

METHOD

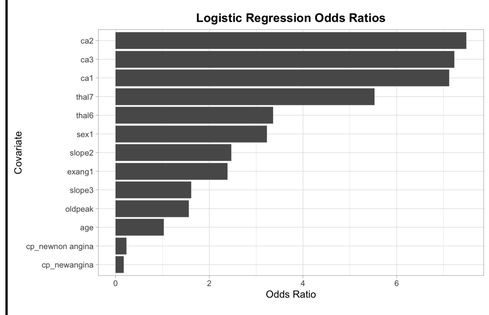
- **Merge Data**: Understand the general traits patients from each hospital to assess if the data can be merged. (see if the health of patients differs between samples).
- **Exploratory Data Analysis**: Explore the univariate relationships between each of the test results (predictor) and the diagnosis of the patients (outcome) by visualization and simple logistic regression for significance.
- **Interactions**: Visually inspect the trends between predictors that are likely to have an interactions based on existing research and evaluate correlations between the variables.
- **Model Selection**: A continuation-ratio logit model (CLM) since it suitable for prediction the odds associated to hierarchical ordinal variable (CV vessel blockage has five stages).
- **Model Tuning**: Backward selection is done to trim down the full model to features with significance and reduced randomness. Proportional Odds assumption is tested for the features and violations are inspected.
- **Binomial Model**: Categories for the outcome variable are collapsed to binary (yes/no) and a simple logistic regression model with logit link is built with the same process to predict the the risk of any stage of CV vessel blockage among patients.
- **Result Exploration**: Final models are compared to the initial model and the associated costs are considered for final adjustments.

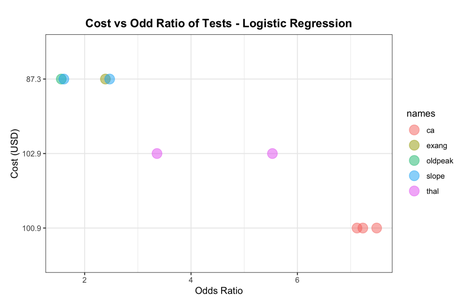
RESULTS

CLM Model for CV Vessel Blockage



Logistic Regression for CV Vessel Blockage





CONCLUSION

- **Data Merge**: All the hospitals had different age distributions, sample sizes as well as diagnosis proportions. Hence, the hospital code was initially included as a feature in the model, however only one of the hospital codes was significantly different from the reference category. Additionally, t-SNE dimensional reduction showed that the test results of patients from different hospitals did not form any specific clusters, so the hospital code feature was removed from the revised models.
- **CLM Feature Selection**: The categories typical and atypical Angina merged into one category due to lack of significance in the model. Resting Blood Pressure and resting ECG were retained in the model despite their lack of significance as research has shown that blood pressure and ECG are crucial indicates of cardiovascular abnormalities.
- **CLM Interactions:** Interaction effect between slope of peak exercise and chest pain was included into the model after reviewing relevant research in the domain. The interaction between downward slope and Angina related chest pain was significant in the model. [1]
- **Proportional Odds Assumption**: The assumption was not met for chest pain, exercise induced Angina, Thallium stress test, and number of vessels in fluoroscopy. However, the effect plot showed that the trends for all four variables were consistent in positive CV vessel blockage diagnosis (code: 1-4) and varied in negative CV vessel blockage. Additionally, the univariate plots for these variables showed trends between stages which gave further validation on their value addition in the model and hence all the variables included.
- **Logistic Regression Model:** The model has a high classification power as the area under the curve is 0.93 which is significantly better than a random chance model (AUC: 0.50). Most of the predictors between the two models were similar except the interaction in CLM and age in logistic regression.
- **Cost Effective Analysis**: The scatter plot with the odds ratios and costs of different tests shows that slope of peak exercise, depression by exercise/rest, and exercise induced Angina have higher costs and relatively lower odd ratios, however removing them significantly reduce the model performance and it is best to keep the tests despite their cost.
- **Limitations and Future Work**: There is scope for further feature engineering by understanding the clinical aspects of the tests to make the model better. Use of machine learning classification models like decision tree, random forest, K- Nearest Neighbors would be helpful in boosting classification performance.

BIBLIOGRAPHY

[1] Hill J, Timmis A. Exercise tolerance testing. BMJ. 2002;324(7345):1084–1087. doi:10.1136/bmj.324.7345.1084