Key Indicators in predicting a Heart Attack

Arpit Dang
Data Science Institute at Brown University
December 11th, 2024
https://github.com/adang66/Data1030-Project

Recap

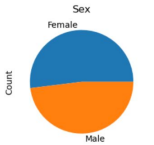
What: Telephone surveys among 400k+ American adults that collected variables that may contribute to heart attacks

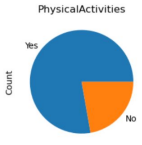
Who: Collected by the Centers for Disease Control and Prevention (CDC)

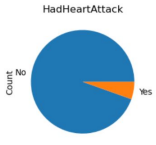
When: It was collected in the year 2022 (relatively recent)

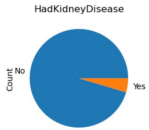
- The authors removed the data points with missing values leaving **246,022 data points** with **39 variables**
- **Target Variable:** If the participant had a heart attack; (Yes OR No); making this a classification problem

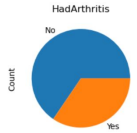
Binary Features

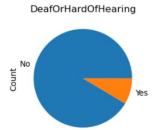


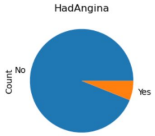


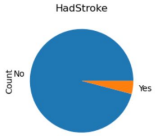


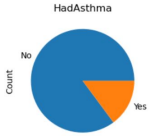


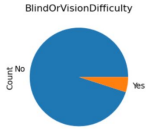


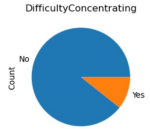


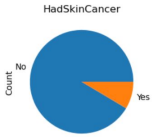


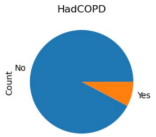


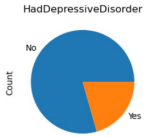


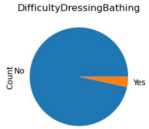


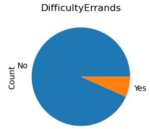


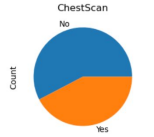






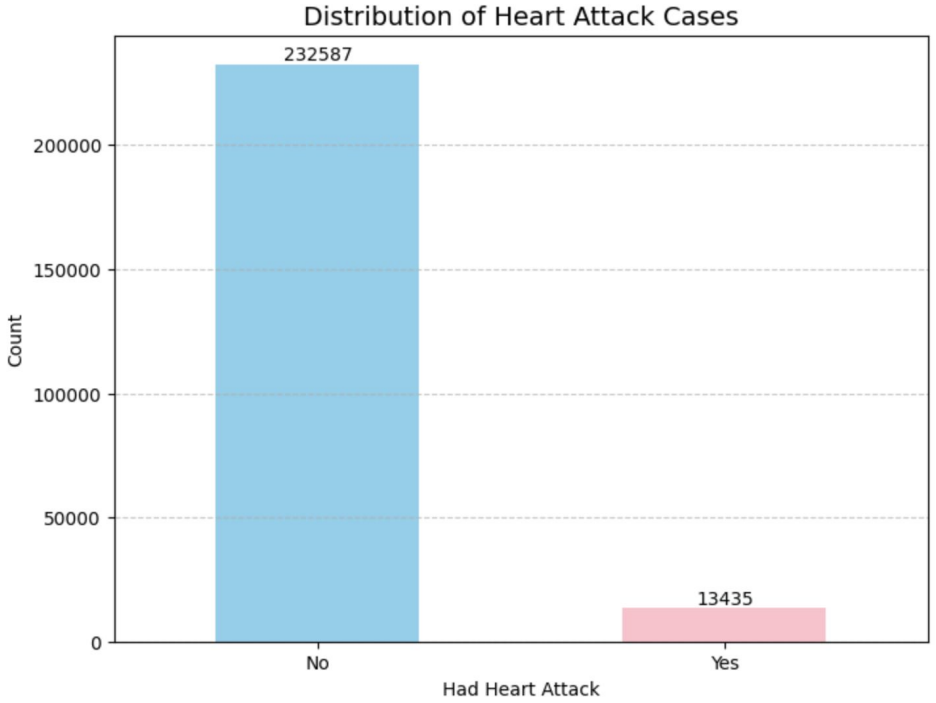






23/39 features are binary features.
Limitation: lack of qualitative data and variability in data, it can lead to reduced model accuracy or interpretability

Distribution of Heart Attack Cases (target variable)



- 5.5% of the participants has an incident of a heart attack, making it an imbalanced dataset
- Solution: stratified split
- 60% training set, 20% validation set, 20% test set

Machine Learning Algorithms

Machine Learning Algorithm	**Parameters**	**‘class_weight’**
Logistic Regression	penalty : [l1, l2, elasticnet (l1_ratio = 0.5)] C : [0.001, 0.01, 0.1, 1, 10, 100]	class_weight = ‘balanced’
Random Forest Classifier	max_depth : [1, 3, 10, 30, 100, 300, None] max_features : [0.25, 0.5, 0.75, 1.0]	class_weight = ‘balanced’
Support Vector Classifier	C : [0.01, 0.1, 1, 10, 100] gamma : [scale, auto]	class_weight = ‘balanced’
XGBoost	learning_rate : [0.001, 0.01, 0.1, 1, 10, 100] gamma : [0, 1, 5, 10, 50, 100]	scale_pos_weight = 17.31

Evaluation Metric for ML algorithms

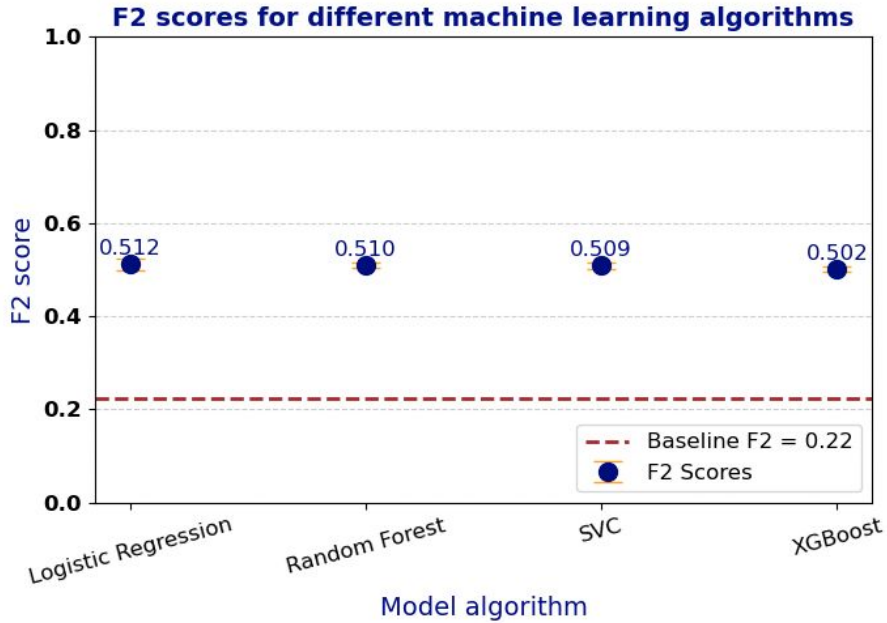
- Because it is an imbalanced dataset, F2 was chosen because it does not take True Negative values into account.

- It is also a medical diagnostic problem (where missing a heart attack prediction (FN) is far worse than predicting a heart attack incorrectly (FP)), F2 score was chosen to put more emphasis on recall.

- Baseline F2 Score: 0.2241 (Assume that all predicted points belong to ‘Yes’ for ‘HadHeartAttack’)

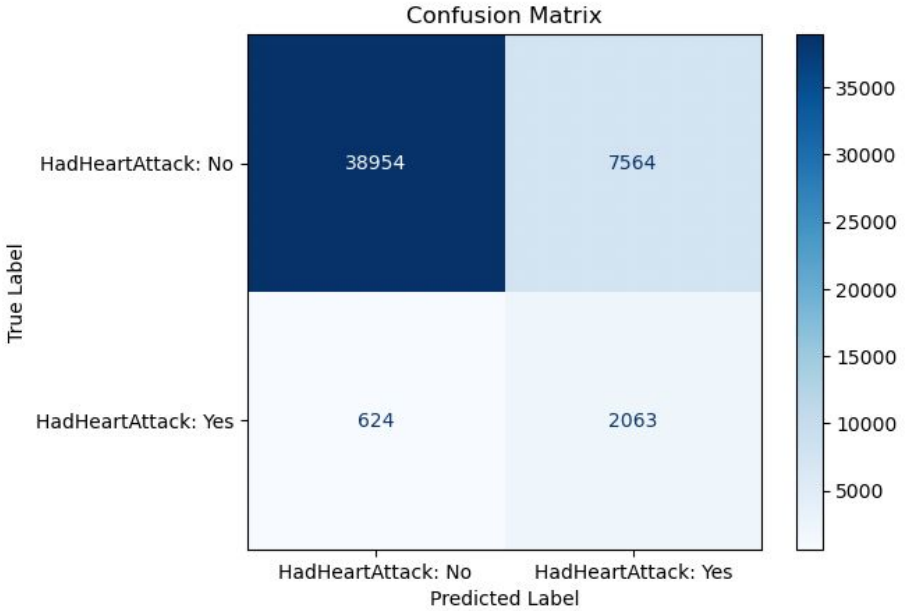
F2 Score for different ML Algorithms

- Uncertainties from splitting were addressed by performing the model training and testing over 5 random states.



*20% of the data points were used to train, validate, and test the SVC model (due to the large dataset, and limited computational resources)

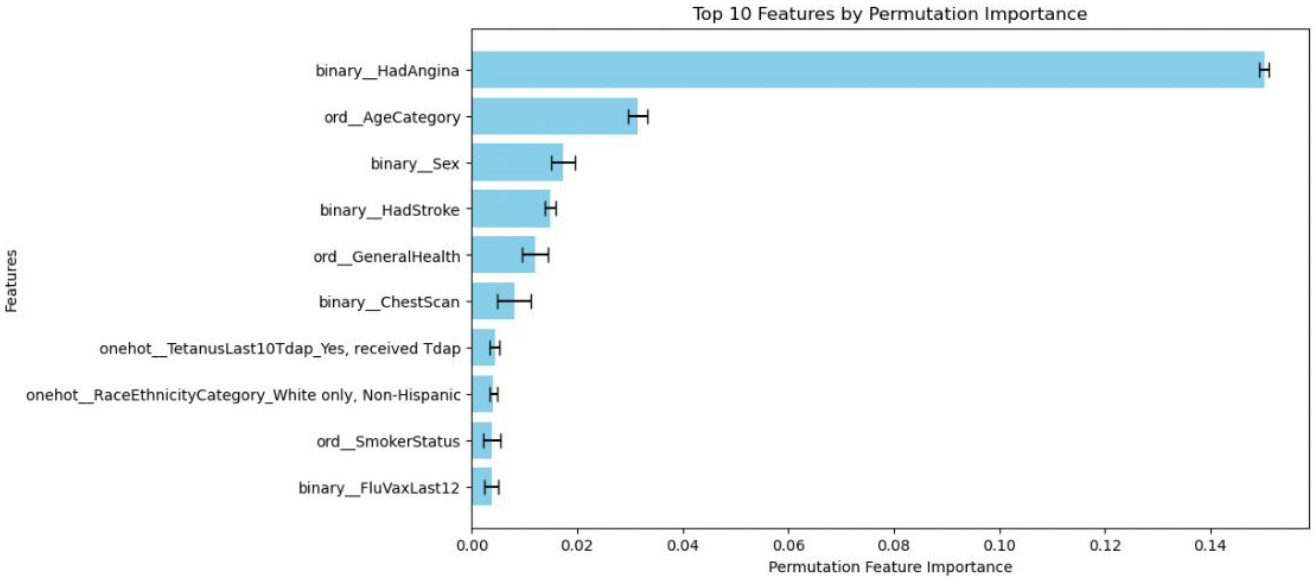
Test Set confusion matrix using the best performing Logistic Regression model



- Parameters:
  - penalty : elasticnet
  - l1_ratio : 0.5
  - C : 100 (other C values were 0.1 or 1)
  - Random_state = 168

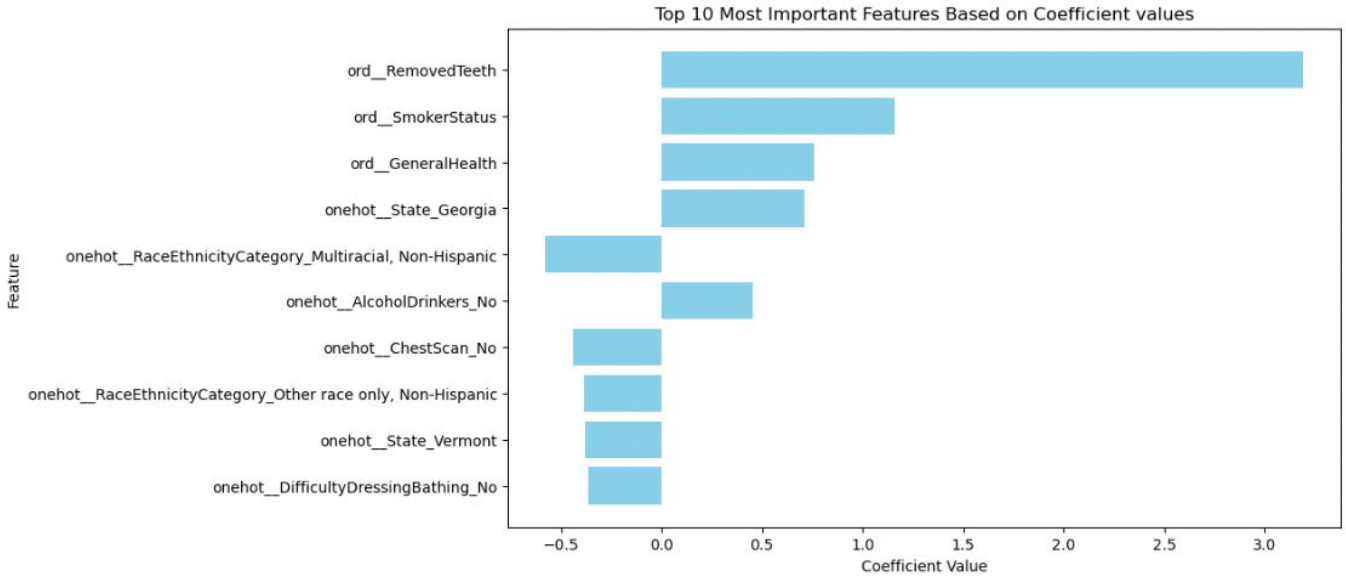
- Accuracy: 0.8336
- Precision: 0.2143
- Recall: 0.7678
- F1 Score: 0.3351
- F2 Score: 0.5312

Top 10 most important features according to permutation importance (test set)



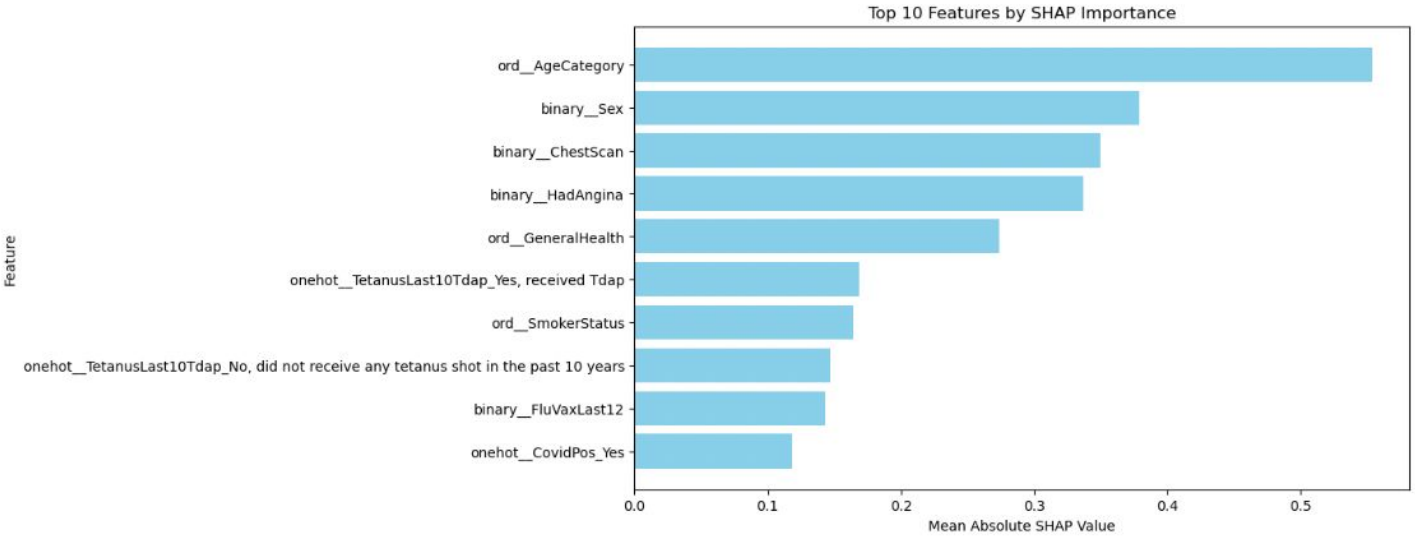
- Angina (chest pain)[1]
- Age (Ranges of age; ordinal not continuous)[2]
- Sex [3]
- Had a stroke [1]
- General Health (Poor, Fair, Good, Very good, Excellent)

Top 10 most important features according to coefficient values (test set)



- Removed Teeth (None of them, 1 to 5, 6 or more but not all, all) [4]
- Smoker Status (Never smoked, Former smoker, Current smoker - some days, Current smoker - every day) [2]
- General Health
- State of Georgia

Top 10 most important features according to mean Shap value (test set)



- Age
- Sex
- Chest Scan
- Agina (chest pain)
- General Health

Some overlap of features that were most important

- Age
- Angina (chest pain)
- Sex
- General Health
- Smoker Status
- Chest Scans

Outlook - how can we improve?

- Use other types of machine learning algorithms (K-Nearest Neighbors, Naive Bayes, Bagging Classification, and Boosting Classification)
- Maybe put even more emphasis on recall (F3 or F5 Score) because there were still 624 cases in the test set that are False Negatives (1.27%)
- Given more computational power, perform the SVC ML algorithm on 100% of the data
- Estimate feature importance (permutation/global shap) after dropping correlated variables OR perform a permutations on two features at a time

Thank you for listening!

Questions?

References

Aminoshariae, A., Nosrat, A., Jakovljevic, A., Jaćimović, J., Narasimhan, S., & Nagendrababu, V. (2024b). Tooth loss is a risk factor for cardiovascular disease mortality: A systematic review with Meta-analyses. *Journal of Endodontics*, *50*(10), 1370–1380. https://doi.org/10.1016/j.joen.2024.06.012

Mall, S. (2024). Heart attack prediction using machine learning techniques. *2024 4th International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, 1778–1783. https://doi.org/10.1109/icacite60783.2024.10617300

Man, J. J., Beckman, J. A., & Jaffe, I. Z. (2020). Sex as a biological variable in atherosclerosis. *Circulation Research*, *126*(9), 1297–1319. https://doi.org/10.1161/circresaha.120.315930

Pallangyo, P., Mkojera, Z.S., Komba, M. *et al.* Public knowledge of risk factors and warning signs of heart attack and stroke. *Egypt J Neurol Psychiatry Neurosurg* **60**, 12 (2024). https://doi.org/10.1186/s41983-023-00780-x