Springboard Capstone 2
Abdul Hannan
06/22

PERSONAL KEY INDICATORS OF HEART DISEASE

Introduction

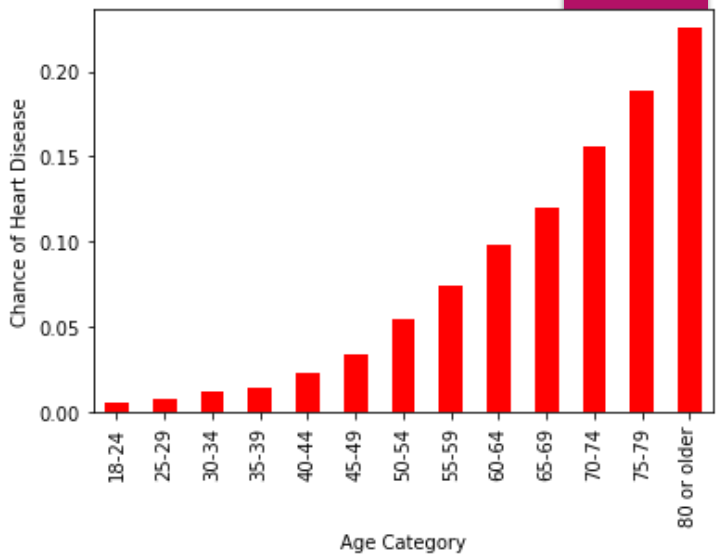
- Behavorial Risk Factor Surveillance System.
- Risk factors of heart disease - blood pressure, cholesterol, smoking.
- Original dataset contained nearly 300 variables and the version acquired from Kaggle had 20 of the 300 variables.
- Unbalanced classes.

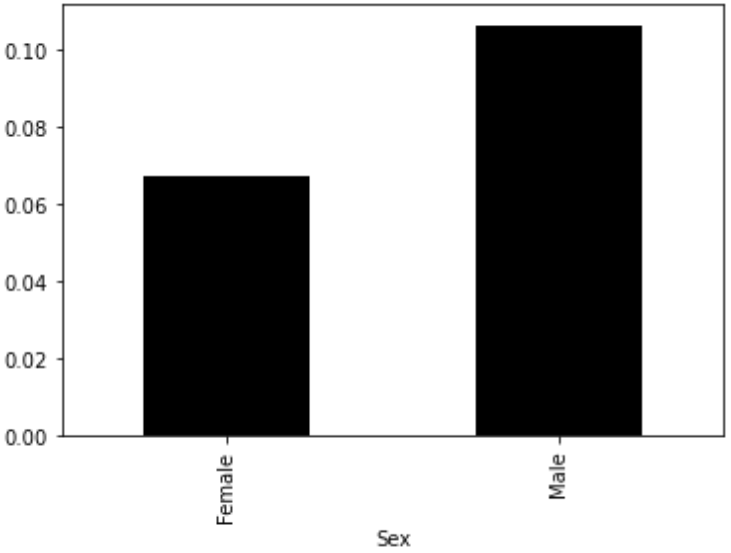
Data Wrangling

- 14 Categorical Variables & 4 Numerical variables.
- 319,795 entries with no missing data.
- Having classes with heterogeneous sizes meant resampling to gain better score for our target feature.

Exploratory Data Analysis

- Older Patients are at more risk of a heart disease.
- Male Patients are at more risk of a heart disease.
- Smoking doubles the chances of you contracting a heart disease.
- Diabetic, Strokes, skin cancer, kidney disease, poor lifestyle significantly increase the risk of a heart disease.
- No exercising – Difficulty walking – can increase your chances by 4X.





Pre Processing and Training

- After getting dummies at the end of EDA, we saved the dummy data and applied the train test split with 75% data for training set.
- The imbalance of our target variable corroborates with the imbalance in our train and test sets.
- The problem: the opportunity cost on passing up on a client is high, and the cost of acting is low, thus we will use recall to compare model scores.
- The logistic regression model from the LogisticRegression class corroborated that resampling is needed. One class did significantly better than the other.

Modeling

- To target imbalance – Resampling
  - SMOTE – Over
  - ADASYN – Over
  - Random Under Sampler – Under
- Random Forest – XGBoost – ADABoost
- Under sampling technique performed better than over sampling.
- In total, we had 9 models – XGBoost with Random Under Sampler performed the best with the recall score of 0.82.
- Lowering the threshold made our model better, Precision: 0.72, Recall: 0.88

Conclusion & Future Work

- Successfully created a model on the imbalanced dataset to predict the likelihood of a heart disease.
- Apply SHAP values that represent a forward approach to interpret predictions, BMI, age, etc.
- Distinguish the characteristics that lead to a wrong outcome.
- Further collection of patients medical history can be beneficial to the current model.

Recommendations

- Suggestions:
  - Strong Positive: X >= 0.5
  - Strong Negative: X <= 0.35
  - Active Surveillance / Borderline: 0.35 <= X <= 0.5
- SHAP + Bayesian Optimization that show how a strong probability (of disease) can turn into negative, and vice versa.
- Understand what changes to adapt and avoid while doing the above.

Thank You!