Springboard Data Science Capstone

Capstone Project #2

Personal Key Indicators of Heart Disease

Abdul Hannan

05/2022

1 – Introduction

The Behavioral Risk Factor Surveillance System (BRFSS) conduct large health survey system that span over all 50 states of the United States . According to CDC, about 47% Americans have one of the following three risk factors for heart disease: high blood pressure, high cholesterol, and smoking. Other key factors include high obesity, diabetic status, etcetera. The goal of this project is to build models that estimate the probability that a patient might suffer of heart disease, or not, as a function of several health metrics.

The data we are working on is heavily skewed (91% no, 9% yes). To build model on skewed data would not be ideal as one class will perform way better than the other class, so to have better scores for our target variable, we needed to apply resampling techniques, and use the latter to generate models. After applying the resampling techniques, my tuned XGBoost Classification model was able to achieve a recall of 0.82, and by lowering the threshold of when a positive class is chosen over a negative from 0.5 to 0.4, the recall for the model increased to 0.90.

2.1. – Data Acquisition and Wrangling

The dataset we worked with had 20 features, and along with our target variable ("heart disease"), most of the variables we found were also skewed, such as, stroke, kidney disease, diabetic, skin cancer. Having classes with heterogeneous sizes meant resampling to gain better score for our target feature. Also, we moved to the next stage of Exploratory Data Analysis as we found that we have no missing values and thus we were ready to look for insights.

2.2 – Storytelling and Inferential Statistics

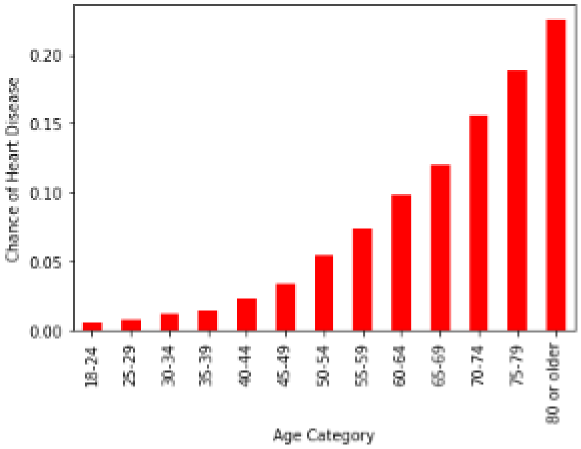


Figure A

Since we were provided with bins of feature age, we gained insight that old patients are at more risk of a heart disease (Figure A), along with kidney disease and skin cancer. Also, that if you are a smoker, your likelihood of a heart disease doubles from 6% to 12%. Your chances of heart disease go up as the previous health issues keep increasing, such as having kidney disease, skin cancer, stroke, if you have all 4, your chances of a heart disease can be as much as 60%.

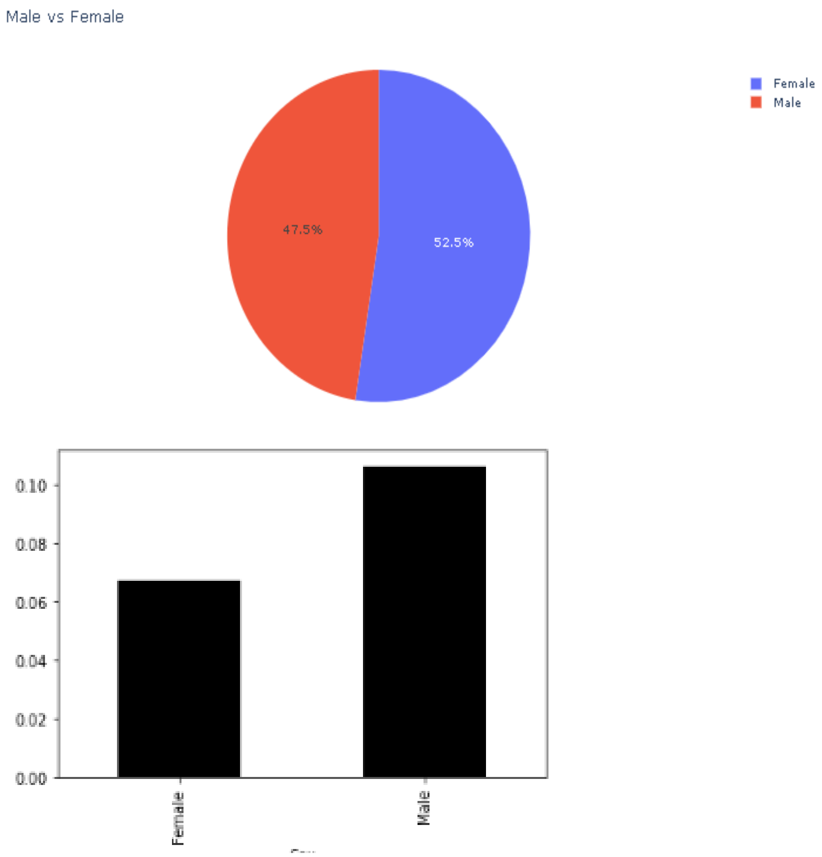


Figure B

We also find that while more females are reported than males, Most heart patients are Male and that Male patients are 1.6 times more likely to have a heart problem than females (Figure B). Another key insight that is important to share is that most patients in the dataset are White (about 77%) and they have about 9% chance of a heart disease, whereas American Indian/Alaskan Native account for only 1.5% and they have a 10% chance of a heart disease. At the end, we added dummies variables to our data and plotted a heatmap to gain insight into the pairwise correlation among features and the target (Figure C).

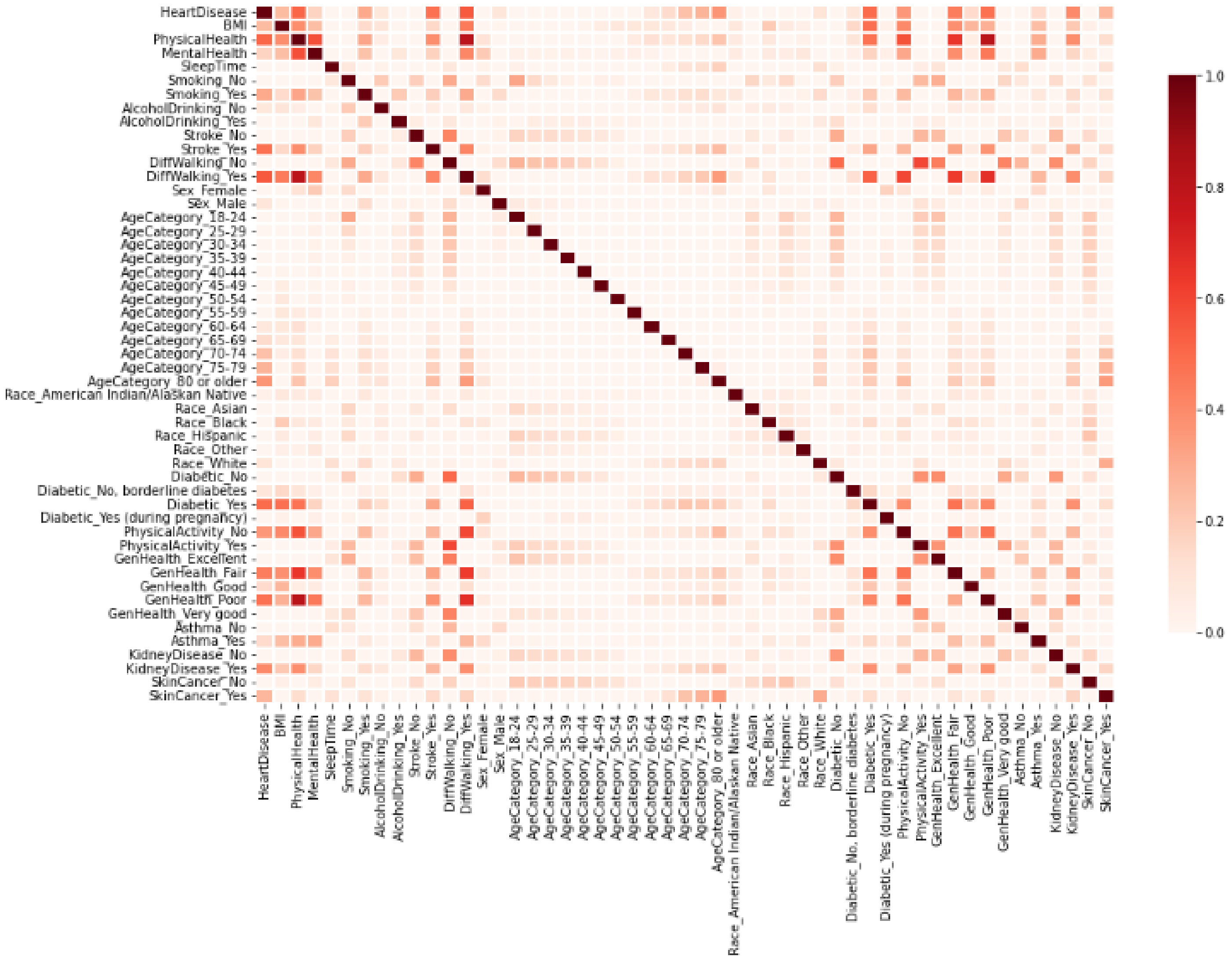
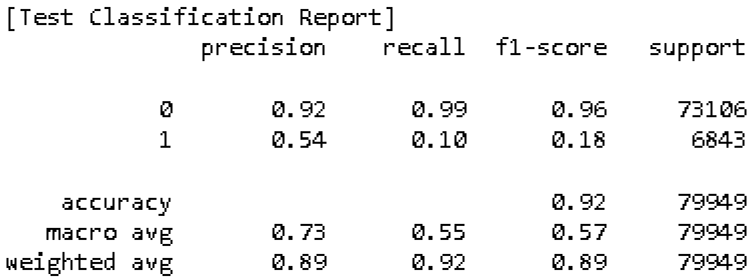


Figure C

2.3 – Baseline Modeling

For our next step, we preserved the dummy variables and applied the train test split with 75% data for train set and 25% for test set. The data in the training set corroborates the fact that the data is skewed with respect to the number of elements in each class with 91.4% belonging to the class of patients with no heart disease, and 8.6% belonging to the class of patients with heart disease. Our first model we generated a logistic regression model using the LogisticRegression class from scikit-learn using default parameters corroborated with the fact that resampling needed to be done if we do not want one class to do better than the other.



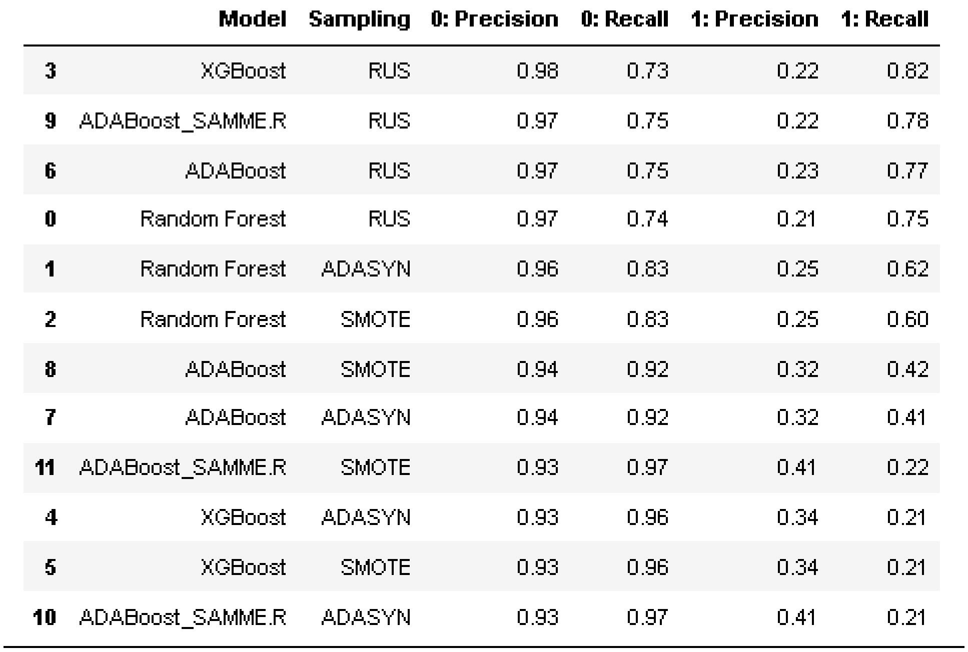
```
[Test Classification Report]
              precision    recall  f1-score   support

           0       0.92      0.99      0.96     73106
           1       0.54      0.10      0.18      6843

    accuracy                           0.92     79949
   macro avg       0.73      0.55      0.57     79949
weighted avg       0.89      0.92      0.89     79949
```

2.4 – Extended Modeling

Blindly building models with our skewed dataset would have kept producing the same results for the minority class (the class with less elements), thus we decided to apply the resampling techniques in combination with classification algorithms as an approach to address the skewness of the dataset. More specifically, we applied three resampling techniques implemented in the package […..] , Random under sample (Under sampling technique), SMOTE (Over-sampling technique), ADASYN (Over-sampling). We then combined these resampling approaches with thefollowing classification algorithms, Random Forest Classifier, XGBoost Classification, ADABoost classification. For our problem, the opportunity cost on passing up on a candidate is high and cost of acting is low, that is correctly identifying a true positive case and keeping the cost at minimum, thus we will use the metric recall to measure our models.

3 – Findings

Model Comparison



	Model	Sampling	0: Precision	0: Recall	1: Precision	1: Recall
3	XGBoost	RUS	0.98	0.73	0.22	0.82
9	ADABoost_SAMME.R	RUS	0.97	0.75	0.22	0.78
6	ADABoost	RUS	0.97	0.75	0.23	0.77
0	Random Forest	RUS	0.97	0.74	0.21	0.75
1	Random Forest	ADASYN	0.96	0.83	0.25	0.62
2	Random Forest	SMOTE	0.96	0.83	0.25	0.60
8	ADABoost	SMOTE	0.94	0.92	0.32	0.42
7	ADABoost	ADASYN	0.94	0.92	0.32	0.41
11	ADABoost_SAMME.R	SMOTE	0.93	0.97	0.41	0.22
4	XGBoost	ADASYN	0.93	0.96	0.34	0.21
5	XGBoost	SMOTE	0.93	0.96	0.34	0.21
10	ADABoost_SAMME.R	ADASYN	0.93	0.97	0.41	0.21

Figure D

Different resampling techniques result in different metric numbers for the same model. In our case, RandomUnderSampler (RUS) performs better than our over-sampling techniques, one reason could be that RUS technique contained about 40k observations whereas oversampling techniques contained over 300k observations, so less observations are yielding a better recall metric score, that is over all the positive cases in the data, how many the classifier predicted correctly. Note that these results are for the test set only.

Our client has asked to completely avoid a situation where a patient has heart disease but we do not provide a treatment because our model predicted so, that is the opportunity cost our client wants to avoid. The tuned XGBoost model with the random under sampler performs the best with recall score of 0.82, way better than our generic model of logistic regression did with the recall score of 0.10. That is, our model was able to correctly identify a heart disease patient 0.82 times with the test set, and low precision means that we have a lot of false positive cases but as discussed acting on false positive will cost less than not acting on a true positive.

This is a case where we have high recall and low precision and threshold is 0.5, that is input is classified as belonging to class A if PROB(A) >= PROB(B), however if we lower the threshold from 0.5 to 0.4 (the score where a positive class is chosen over a negative), our recall goes up to 0.88, and our precision jumps from 0.22 to 0.72, making our model even better.

4 – Conclusion & Future Work

Given the skewed dataset of heart disease, with features taken from patient's medical history such as kidney disease, skin cancer, general health, and others, we applied resampling techniques to our skewed data and created model that help us avoid the situation a true positive situation (where a patient has heart disease, but our model predicted otherwise). With the help of under sampling method and tuned XGBoost, the model gave us a recall score of 0.82, and after lowering the threshold 0.4, the precision for the model went up by 0.50 (from 0.22 to 0.72) and recall of 0.88.

Few recommendations for next steps of work can be to apply the SHAP values that represent a forward approach to interpret predictions made by our best model, XGBoost in our case. With the help of SHAP, we can have different values for variables that affect the outcomes, an example can be when we found that older patients are at more risk, the SHAP approach can give us a value relative to the age, the difference between 80-year-old compared to a 60-year-old. Along with age, we can have the BMI score give us a SHAP value relative to the patients BMI, which will in turn help us supplement clinical intuition for risk stratification. Having a way to tell us whether a patient is a high, very high, or low risk can help us accommodate the persons treatment better and plan on whether he should undergo treatment or just active surveillance. The SHAP framework will not only benefit in interpreting predictions, rather also in visualizing relationships between linear and nonlinear features and can prove very beneficial in this journey. Further collection of patients medical history can be beneficial, such as cholesterol, measure of blood pressure, sugar level, etc.

5 – Recommendations for the clients

- Some suggestions that doctors can consider:
  - Strong Positive: X >= 0.5
  - Strong Negative: X <= 0.35
  - Active Surveillance / Borderline: 0.35 <= X <= 0.5
- Use the SHAP package and Bayesian Optimization to show how to compute counter-factual cases that show how a patient with a strong probability can make changes to obtain a strong negative probability, and the opposite to understand what changes to avoid (for the worst).