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 1 of the following 3 risk factors for heart disease, high blood pressure, high cholesterol, smoking. Other key factors include high obesity, diabetic status, etcetera. To find pattern with the computational power at hand, within the given data to predict a patient's condition.

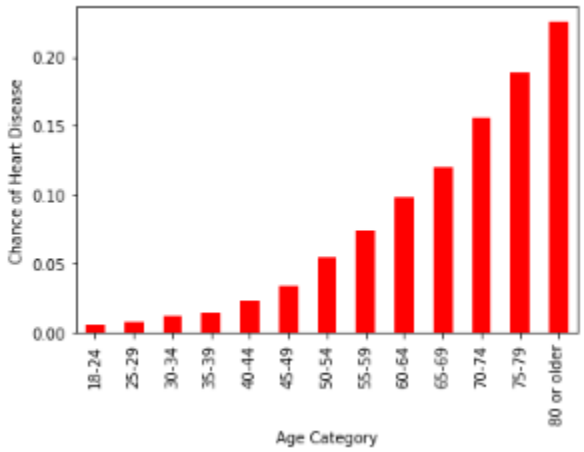
Along the process we learn some insights that will go over soon, and that the dataset is heavily skewed. In order to have better scores for our target variable, we needed to apply resampling techniques, and use the latter to generate models. After applying the resampling techniques, my tuned XGBoost Classification model was able to achieve a recall of 0.79.

2.1. – Data Acquisition and Wrangling

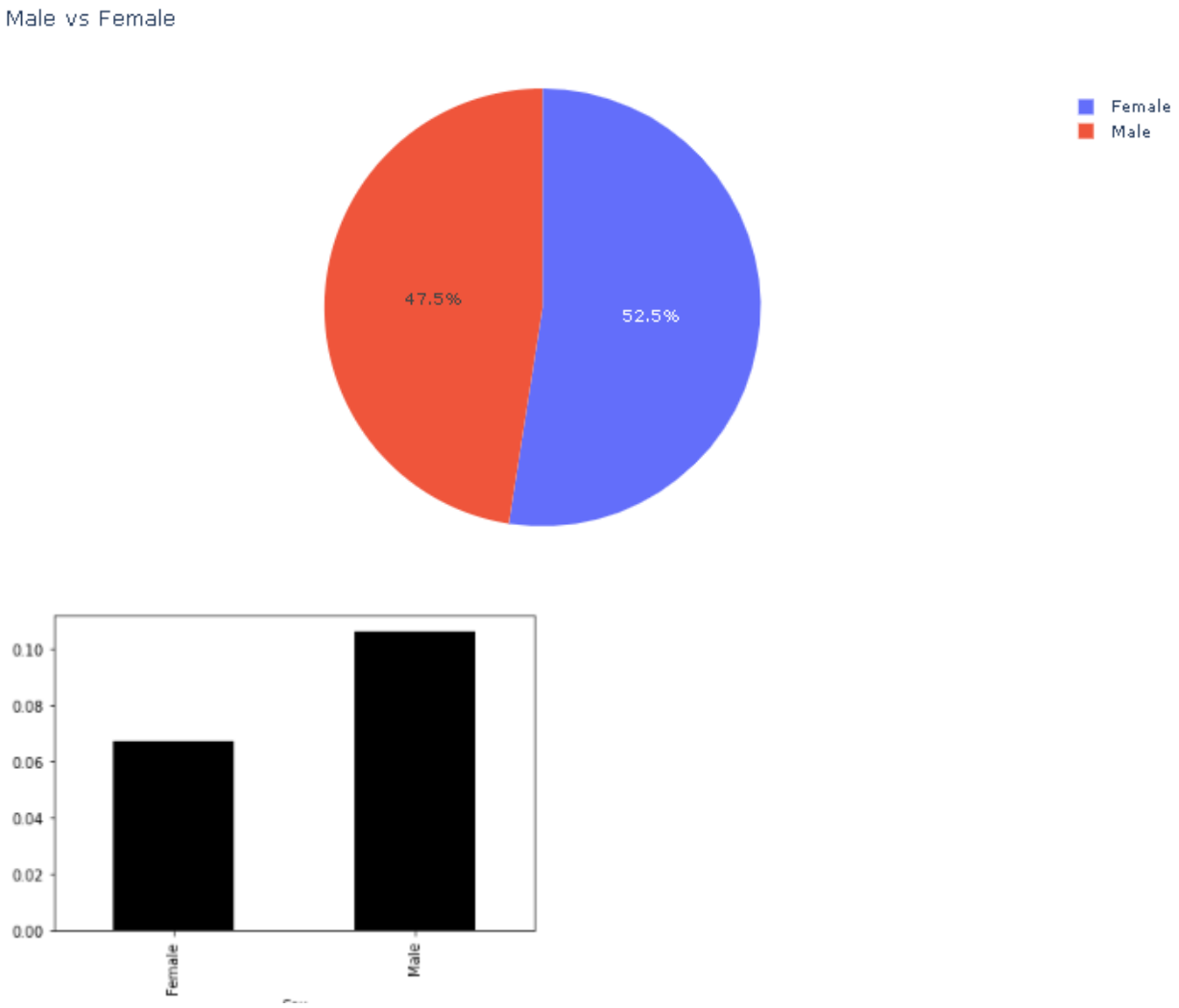
The original dataset contained nearly 300 variables but the version of that was acquired from Kaggle had only 20 features. Along with target variable (heart disease) most of the variables we found were also skewed, such as, stroke, kidney disease, diabetic, skin cancer. Having skewed

data meant that we needed to apply resampling in order 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 are ready to look for insights.

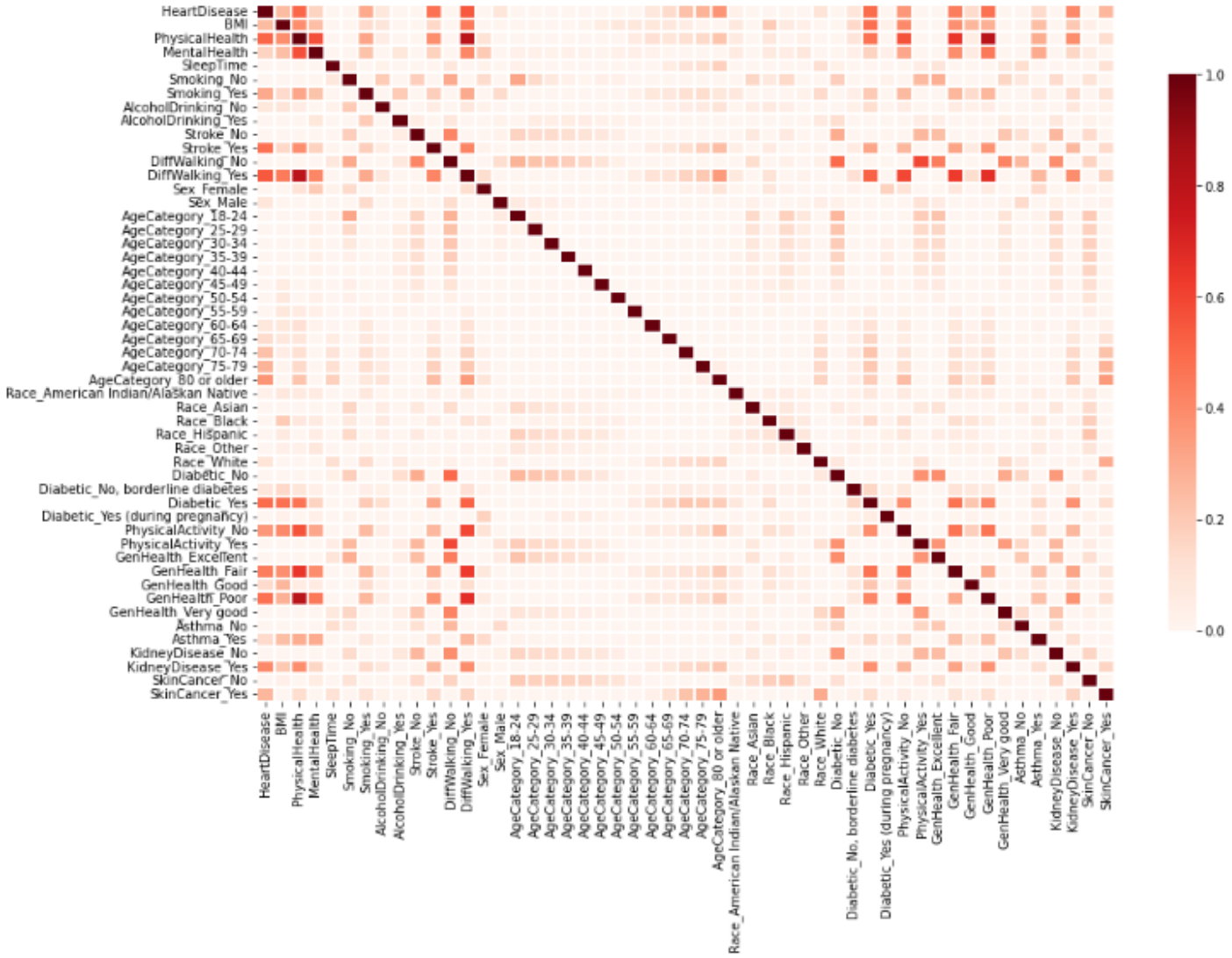
2.2 – Storytelling and Inferential Statistics



Since we were provided with bins of feature age, we gained insight that old patients are at more risk of a heart disease, 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 goes 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%.



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. Another key insight that is important to share is that most patients in the dataset are White (about 77%) & they have about 9% chance of a heart disease, whereas American Indian/Alaskan Native account for only 1.5% abd they have a 10% chance of a heart disease. At the end, we added dummies to our data and plotted a heatmap to gain insight into the correlation between two features.



2.3 – Baseline Modeling

For our next step, we preserved the dummies data and applied the train test split with 75% data for train set and 25% for test set. The data in train set corroborates with the skewed data in our target variables with 91.4% no values and 8.6% yes values for heart disease. After generating a generic logistic regression model the recall score of our target feature was 0.10, a sign that we are on the right track but have a long way to go.

2.4 – Extended Modeling

To generate models with our skewed data, our models would have kept under performing as one class performs significantly better than the other, thus applying the resampling techniques and having a proportion of 0.5 for both classes would help our models perform better. We applied three resampling techniques, Random under sample (Under sampling technique), SMOTE (Over-sampling technique), ADASYN (Over-sampling). We then created each model three times, one for each resampling technique, and three different models in total, Random Forest Classifier, XGBoost Classification, ADASYN classification. We prefer using the recall method because cost of acting on a result is low, but the opportunity cost of passing up on a candidate is high. That is, we want to quantify the number of correct predictions made out of all predictions, opportunity cost of passing up on a correct prediction, that is not responding when the model predicts someone has a heart disease.

3 – Findings

	Model	Sampling	Precision	Recall
0	Random Forest	RUS	0.22	0.74
1	Random Forest	ADASYN	0.25	0.60
2	Random Forest	SMOTE	0.26	0.59
3	XGBoost	RUS	0.22	0.79
4	XGBoost	ADASYN	0.35	0.19
5	XGBoost	SMOTE	0.36	0.19
6	ADABoost	RUS	0.23	0.77
7	ADABoost	ADASYN	0.32	0.42
8	ADABoost	SMOTE	0.32	0.42
9	ADABoost_SAMME.R	RUS	0.23	0.77
10	ADABoost_SAMME.R	ADASYN	0.41	0.21
11	ADABoost_SAMME.R	SMOTE	0.41	0.22

In the case of over sampling methods, we achieve a not so high precision and low recall whereas with under sampling we achieve low precision and a decent recall. Our clients want the model to correctly identify a true positive case and keep the opportunity cost at the minimum, the preferred metric we will use will be recall. For example, our client does not want 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 0.79, 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 0.79 times, there is room to make our model better, but we have come far from the baseline model.

4 – Conclusion

Given the skewed data of heart disease, with features such as patients medical history such as kidney disease, skin cancer, general health, and others, we were able to turn the skewed data into normal using resampling techniques and create model that help us avoid the situation where a patient was not giving proper treatment when the model predicted wrong. With the help of under sampling method and tuned XGBoost, the model gave us a recall score of 0.79.

Future work

5 – Recommendations for the clients

- 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 make a decision 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.

- Some of the variables can be better, instead of having one variable for general health, we can collect more data on the type of pain a patient is feeling (Chest, pain, numbness, upper back, etc), their blood pressure, the amount of sugar intake for a day, perhaps a model with better features will surely be beneficial. We can also have the other two main factors found in 47% Americans, a measure of their blood pressure, and cholesterol can prove vital features in predicting the outcome.