CAPSTONE PROJECT: FINAL REPORT

Oliver Leach | April 11, 2023

Problem Statement and Background:

Cardiovascular disease (CVD) is the leading cause of death worldwide. In the USA alone, CVD is estimated to cost $555 billion per year, but early intervention is proven to reduce the financial burden. For every $1 spent on preventive measures an estimated $7 can be saved in the long run.
Given these statistics, it seems sensible to create an app that acts as a preventative measure and empowers the general public to take control of their heart health.
CardioCare is a machine learning aided app that aims to predict an individual's risk of developing CVD based on their personal data and provides users with a detailed report outlining their primary risk factors and personalized recommendations to improve their lifestyle.

Data Collection:

To build this app I needed to answer a few key questions:

1. What are the Key Indicators of heart disease?
2. How do these Key Indicators affect our risk of cardiovascular disease?
3. How can we use this to provide personalised, actionable advice?

To accurately answer these questions, I need a dataset containing a mix of cardiovascular specific metrics along-side lifestyle features. From the outset, it was clear that finding this data set would be difficult, personal health data is often inaccessible due to data protection regulations.
The only dataset employed in this project was a 'lifestyle' based one from the yearly CDC BRFSS survey. The original survey contained 279 columns containing data related to respondents' health-related risk behaviours, chronic health conditions, and use of preventive services. As the CDC is recognized as a highly credible authority, the data provided by them can be deemed reliable and trustworthy. This can also be confirmed by cross referencing EDA results with medical literature.
The dataset provided to me had been 'pre-distilled' down to just 18 features from the original 279. While it did save me time, one disadvantage was that the task of feature engineering had been taken care of for me.

Data Cleaning and EDA:

The dataset was generally clean, and few data pre-processing steps were necessary. Specifically, ~18,000 duplicated rows were eliminated. Furthermore, 0.7% of the entries in

the SleepTime column were considered invalid due to being either less than 3 or greater than 14 hours of sleep per night. These were consequently removed. Additionally, the values in the age column were represented as string objects that displayed age ranges. To convert these values to integers, the average age for each unique range was calculated and converted.

An important observation derived from early EDA was the presence of a significant class imbalance in the target variable, HeartDisease, with the negative class making up 90% of the feature.

Integrity of the dataset required the EDA into the target variable, to be in line with the extensive volume of medical literature surrounding the topic. Fortunately, during the EDA process, all features that could be cross-referenced with medical literature, such as age, BMI, mental health, and sleep time, returned congruent results, confirming the trustworthiness of the data source.

Analysis of the categorical data vs the target variable found that an increased proportion of heart disease was present in the following:
- People that smoke.
- People that have had a stroke.
- People that have difficulty walking.
- Males.
- White and native American/Alaskan.
- People that have not exercised in the last 30 days.
- People of poor health.
- People with kidney disease.
- Very slight CVD rate increase in people with Asthma.
- People with skin cancer.

The difference in the distributions within these populations was confirmed to be statistically significant by a Chi Squared test with a Holm Bonferroni Correction.

Following cleaning and EDA, the data was processed. This involved dummy/one hot encoding the categorical variables into binary features. The final result of data cleaning was one dataset to be used for modelling.

Modelling:

Class imbalance was the main source for concern when planning out the modelling process. Prior to addressing this, I created a logistic regression baseline model using the statsmodel package. The baseline model test accuracy was ~92% whilst recall and precision were ~10% and 50% respectively. The baseline model further highlighted the issue with class imbalance in my target variable. To solve this, I employed a range of

sampling methods that I believed were suitable, ADASYN, SMOTE, RandomOverSampling and RandomUnderSampling.
As the final output of the models required an ordered hierarchy of feature importance with regards to CVD risk, a high importance was placed on the interpretability of the models created. As such, SVM and KNN models were excluded, PCA was also excluded as a method for dimensionality reduction for the same reason. Modelling focused on creating optimised logistic regression, Decision Tree, and Random Forest models, all of which have a large degree of interpretability. The models relied on L1/L2/elasticnet regularisation to control collinear/statistically insignificant features. As a final note, the baseline model highlighted poor performance in model recall and precision, as such models were refitted using the f1 scoring metric.

All models were assessed using 5-fold cross validation. F1, accuracy, recall and precision scores were calculated based on a test subset of the original data. Hyperparameters were tuned using GridSearchCV. Confusion matrices, ROC, and Precision Recall curves were created and analysed for all models. The model with the best performance on the test data was a Random Forest model using the RandomUnderSampler() sampling method.

The lifestyle features of importance to the model, in order of most to least importance (least just means less in this case), were BMI, Smoking, Mental Health, Sleep time and Physical activity.
It's worth noting that improving mental health is potentially more useful than trying to improve on one of these metrics. People are far more likely to succeed in quitting smoking if their mental health is good.
As a final thought exercise, let's say we have a user of our app that presents as a smoker with a healthy BMI, high volume of physical activity, poor mental health, and sleeps 5 hours a day.
The app would calculate heart disease likelihood and return a personalised recommendation to focus on improving mental health, focus on quitting smoking and get more sleep.

Further Application and next steps:

Metrics like blood pressure, resting heart rate and cholesterol level are much better predictors of heart disease. To improve the current app concept, I'd like to include another model based on a dataset containing these granular stats. I placed a lot of weight on model interpretability due to the need to find the key actionable indicators of CVD. In this new model, interpretability would not be required. As such, the full range of machine learning methods could be applied, and an optimal model found.
This new model would become the soul predictor of heart disease in users and the CDC data trained model would be used purely for aiding in lifestyle change recommendations. In addition to this, I'd like to retrain new models based on fresh lifestyle data from the CDC. I'd also like to use data with the original columns still available, having a dataset with pre-engineered features limited the spectrum of potential lifestyle habits I could explore. For example, there were no features based on a user's diet.