Springboard: Data Science Career Track Program

Capstone Project 2 Proposal

Muhammad A. Hannan

03/2022

Heart Disease Indicators Proposal

1) **What is the business problem?**

   - Detecting and preventing the factors that have the greatest impact on heart disease is very important in healthcare. About half of all Americans (47%) have at least 1 of 3 key risk factors for heart disease: high blood pressure, high cholesterol, and smoking. Other key indicators include diabetic status, obesity (high BMI), not getting enough physical activity, or drinking too much alcohol. Thus, the purpose of this project is to build a model that can predict if a respondent or instance is at risk of a heart disease.

2) **Who are the intended stakeholders, and why is this problem relevant to them?**

   - The Behavioral Risk Factor Surveillance System (BRFSS). They conduct the world's largest health survey system, consisting of all 50 states and U.S territories. There are factors that directly or indirectly influence heart disease so having access to models that can explain these influences can be very beneficial.

3) **Where are the datasets available from?**

   - Kaggle

4) **What data science approaches do you anticipate you will use to model the business problem as a data science problem?**

   - The business problem will be modeled as a data science classification problem. Using classification algorithms, models will be built and their performance will be compared, according to relevant metrics−with an emphasis on the metrics that better model the business goals. These models will then be used to conduct interpretability analyses, to study how the variation of features affects the increase/decrease of the likelihood of heart disease. It is anticipated that the dataset might be imbalanced--which might require the use of resampling techniques, for instance. Some of the classification algorithms we can anticipate will be used a[1]re Logistic Regression, Random Forests, Support Vector Machines (SVM), XGBOOST, LGBM, and others that might be considered.

[1] As described, the available dataset can be found at https://www.kaggle.com/kamilpytlak/personal-key-indicators-of-heart-disease

5) How do you anticipate that the intended clients will use the results of your CP2 to address the original business problem?

- After establishing the alignment between the business goals and the appropriate performance metrics, models will be ranked with respect to these metrics, which will lead to explanations to the clients on how to manage false positives and false negatives
- Interpretability analyses will hopefully lead to identifying features that impact the likelihood of heart disease positively and negatively--including pairwise interactions;

  Examples of the use of counter-factual explanations will be discussed as a proxy for suggesting feature variations that lead to diminishing the likelihood of heart disease.