Springboard--Data Science Career Track Program

Capstone Project 2 Proposal

ECG Anomaly Detection

By: Armaghan Azarbarzin

October, 2022

(1) What is the business problem?

A study was conducted to evaluate the safety and efficacy and side effects of long-term oral milrinone therapy over 100 patients with a variety of conditions with major public health implications, including life-threatening arrhythmias, and congestive heart failure. A long-term ECG recording from these 100 subjects were recorded during long-term follow-up which each of them are labeled as normal or abnormal. There is a desire to classify new ECGs as normal or abnormal, automatically.

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

Patients: The results have affected patients' health indirectly by revealing the safety and efficacy or any side effects of this medicine.

Hospitals: The result might change the guidelines taking this medicine decreasing/increasing rate of death/ clinical improvement. Better use of medicines can improve health outcomes and reduce the use of costly medical care.

(3) Where are the datasets available from?

The original dataset for "ECG5000" is a 20-hour long ECG downloaded from Physionet. 5,000 heartbeats were randomly selected including biomedical signals from healthy subjects and from patients with severe congestive heart failure.

https://www.kaggle.com/code/mineshjethva/ecg-anomaly-detection/data

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

As data are labeled and the output of the model should be classified as two discrete values, the business problem can be modeled as supervised learning, which leads to the use of classification algorithms.

I anticipate that the following algorithms will be used:

Logistic Regression, Naïve Bayes, Stochastic Gradient Descent, K-Nearest Neighbors, Random Forest Classifiers, Support Vector Machine Classifiers, XGBOOST Classifiers, LGBM Classifiers, and classical Neural Networks.

It should be mentioned that each sample of the dataset consists of a sequence taken at successive equally spaced points in time. Then, using time series-based models might be helpful to solve this problem as well. In particular I have started to read about the following time series-based algorithms that might be useful:

- Distance-based (KNN with dynamic time warping).
- Interval-based (TimeSeriesForest) adapts the random forest classifier to series data.
- Dictionary-based (BOSS, cBOSS).
- Frequency-based (RISE — like TimeSeriesForest but with other features).
- Shapelet-based (Shapelet Transform Classifier).

In addition, interpretability approaches will be used to identify and quantify the impact of each of the features on the target, for each of the models built.

(5) How do you anticipate that you will evaluate the performance of each of the data science approaches that you envision?

The metrics to evaluate the performance of the classification models that will be built, include (but might be limited to) the following: ROC curves, AUC values, Confusion Matrices, and Classification Reports showing precision, recall, F-1, and support.

As mentioned above, with respect to interpretability, we will also explore the impact and dependencies between all features and the target for a selected set of models.

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

As an ECG is recorded, all other required information belonging to the patient such as medicine dosages are recorded. So, after classifying ECGs by a proper model into normal and abnormal groups, it might be possible to hypothesize that a specific amount of this medicine affects patient health, and potentially more specifically with respect to the abstract features.