

Risk Predictor - Cardiovascular Diseases

Rohit Bansal(rbansal3)

Alpesh Darji (adarji2)

Kathan Al Jewary (ksa2)

Paul Nel (paulnel2)

Abstract—This electronic document is a proposal for a project to develop a risk predictor for the incidence of heart disease using on some basic health metrics based on publicly available heart disease data set.

Keywords—machine learning, map reduce, cardiovascular,

I. BACKGROUND & MOTIVATION

According to WHO, Cardiovascular diseases (CVD) take the lives of 17.9 million people every year, 31% of global deaths. This is the leading cause of human deaths in the world. Often the old cliché of “prevention better than cure” is very much applicable when it comes to CVD. There are multiple factors which trigger these diseases and often there are multiple indicators before the one gets into grip of these diseases. There are no tools publicly available which could track these indicators and create a risk profile for an individual before the individual is in grip of one of these diseases.

Our idea is to indicate the risk level for a user based on the user data. This risk level will indicate the urgency of taking precautionary measures. As part of this project, we will analyze the CVD dataset to find the leading cause of these diseases using ML algorithms. Once our ML model is trained, we will predict the risk level of an individual by analyzing the individuals corresponding dataset. This tool will be available on cloud so that users can use it to find the risk level corresponding to CVD.

II. DATA SET

We will use the Heart Disease Data Set available from the UCI Machine Learning Repository. This data set contains four distinct databases from Budapest, Switzerland and Cleveland respectively. A total of 75 attributes are described in these data sets including metrics such as age, gender, smoking history amongst others. We will also try considering any other publicly available datasets and include them if time permits.

III. PROPOSED OUTCOME

We propose to build a simple predictor that will provide a user their risk of heart disease based on basic health metrics such as age, gender etc. This predictor will make use of machine learning techniques to provide a user with the risk score within a defined certainty.

IV. METHODOLOGY

We will apply machine learning algorithm(s) on the publicly available dataset to train our model. This way our model will be able to establish good understanding of the underlying data. To predict any individual's risk level for a heart disease, we will gather an individual's vitals and perform a predictive analysis using the model we developed. The work is divided in four major areas research, development, validation and reporting. Since all the team members are capable of performing work in

all the four specified areas, we will attempt to allocate work in such a way that everyone gets a chance to work on all four domains.

V. TIMELINES

Milestone 1: Due Mar 9

Data Collection and Manipulation

By this milestone we should be done with gathering and processing data for selecting and training a machine learning model. We will ensure that we are able to host the data in a cloud environment and that we are able to work with it effectively and efficiently. Concepts learned regarding cloud environment, hosted instances and virtualization, will be extremely useful in achieving this milestone.

Milestone 2: ~~Due Mar 30~~ Due Apr 10

ML Model Development

By this milestone we expect that our ML algorithm will be effective and mature. The data fed to the ML model should provide enough correlation for the model to predict an outcome. We hope to refine our model during this phase and attempt to implement concepts like server hosting, big data are presented during the CCA course.

Milestone 3: Due Apr 27

Real Time Prediction Using ML Model

By this milestone we should be able to analyze test data and predict result using our ML model in a reasonable time. To achieve this, the concepts concerning parallel processing, Hadoop Map-Reduce will be utilized to the fullest. By this time we will have project almost ready. Final piece (Milestone 4) would be to add more documentation, finalize project report, setup instructions and build presentation.

Milestone 4: Due Apr 30

Project Submission along with Final Report & Presentation

Our team will work together to distribute work regarding preparation of project report and presentation. In this phase code is documented and setup instructions are clear enough for anyone to follow.

VI. RESOURCES

We propose to develop the tools using Python and for machine learning we propose to use publicly available libraries such as sklearn or similar.

- Language: pySpark
- Cloud Technologies Used: IaaS, PaaS, Map-reduce, Hadoop, Hortonworks
- Algorithms & Concept material: Well-Known Machine Learning (ML) Concepts
- Code Management System: Github
- Libraries: Scikit-Learn, scipy, pandas, numpy

VII. PROJECT PROGRESS REPORT

We made significant progress in our project goals. The following section includes brief updates regarding progress, challenges, and changes made. It also outlines the prototype of our final project outcome.

VIII. CHANGES

There was a change in second milestone date for the project. We found that building and training the model using PySpark MLlib takes more time and requires knowledge presented week 12 of course. This delay will not propagate to the third milestone as real-time prediction is almost instantaneous once the trained model is available. As a result, the delay in second milestone does not affect third milestone and overall the project is on track.

IX. PROGRESS

Milestone 1: Due Mar 9 (completed)

Data Collection and Manipulation

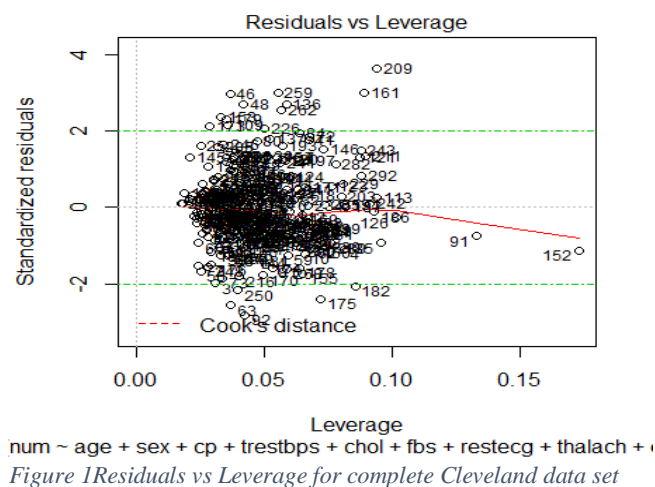
We used the Heart Disease Data Set from the UCI Machine Learning Repository for this project. This data set consists of four distinct data sets (i) Cleveland, (ii) Hungary, (iii) Switzerland and (iv) VA data sets. After some research we found that the Cleveland data set is the most complete. It has been used for some other machine learning research as well.

We have therefore opted to create two different datasets, one for Cleveland only and the second one a combination of all four datasets but using only the most common attributes across the four sets. Both data sets were initially cleaned to remove all samples that contained erroneous data. The sets were then inspected for outliers based on comparison between standardized residuals and leverage, as well as Cook's distance.

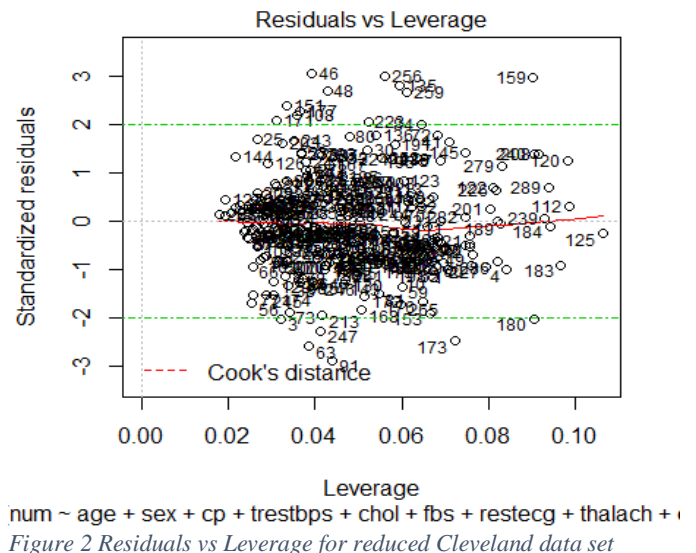
A. Cleveland Data Set

For the Cleveland data set, the following attributes were used (refer to UCI reference for a detailed explanation):

1. #3 (age)
2. #4 (sex)
3. #9 (cp)
4. #10 (trestbps)
5. #12 (chol)
6. #16 (fbs)
7. #19 (restecg)
8. #32 (thalach)
9. #38 (exang)
10. #40 (oldpeak)
11. #41 (slope)
12. #44 (ca)
13. #51 (thal)
14. #58 (num) (the predicted attribute)



Preparation of the data consisted of two parts. Firstly, all rows containing erroneous values were removed. This reduced the data sample count from 303 to 296 which was then used to construct a linear regression model against all attributes using R. Figure 1 above shows the resulting plot with labelled data points. Data point 209 was identified as a possible outlier due to higher than expected deviation from the mean and points 91 and 152 due to relatively high leverage and Cook's distance. Figure 2 shows the resulting plot after these three data points had been removed.



B. Combined Data Set

Combining data sets proved to be somewhat of a challenge due to the high number of missing attributes in data sets (ii) – (iv) when compared with (i). Evaluating all attributes across all four data sets, the final subset decided upon included:

1. #3 (age)
2. #4 (sex)
3. #9 (cp)
7. #19 (restecg)
8. #32 (thalach)
9. #38 (exang)
14. #58 (num) (the predicted attribute)

Since these attributes will be the ones measured and recorded immediately on admission of a patient to the hospital it therefore suggests that this is a logical grouping of attributes.

In addition to consolidating the attributes across the four different data sets, we also decided to add an additional attribute, *set_id*, that identified the dataset per sample. Since each data set represents a different geographical area, we will test to see if this has a significant outcome in predictions. Combining these data sets resulted in a total sample count of 920. In total, 57 samples were removed that contained at least one erroneous attribute value resulting in an interim data set of 863 samples. We constructed a linear regression model in R to determine any obvious outliers resulting in the plot shown as figure 3. Figure 3 suggests point 212 as a possible outlier and was removed. The result is presented in Figure 4.

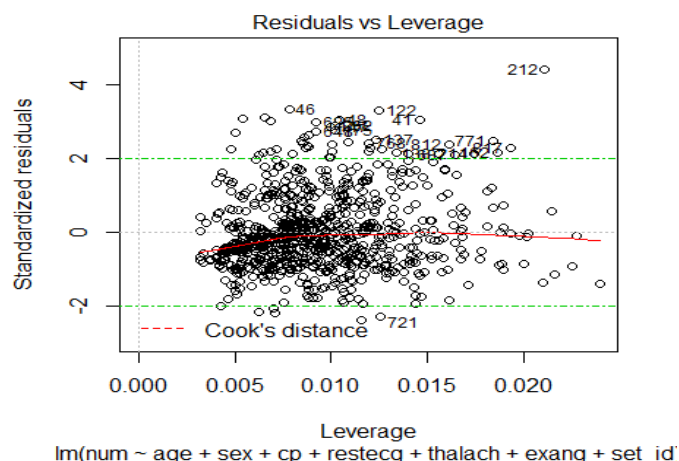


Figure 3 Residuals vs Leverage for complete Combined data set

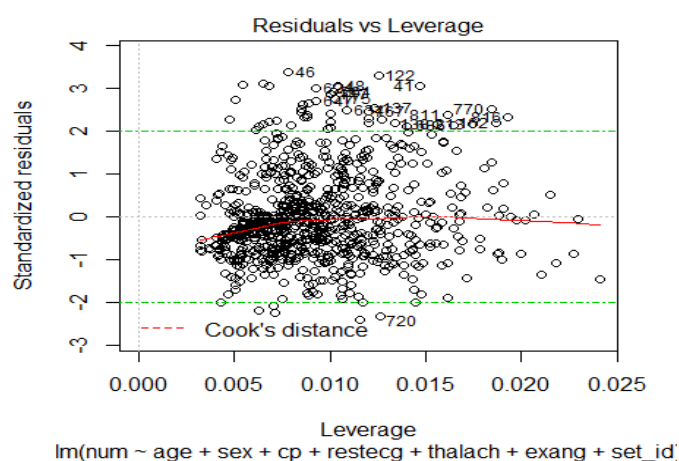


Figure 4 Residuals vs Leverage for reduced Combined data set

Milestone 2: ~~Due Mar 30~~ Due Apr 10 (in Progress)

ML Model Development

After thorough research we found that linear regression will be the best model for our use case. We have implemented the ML model using python-scikit learn to test the accuracy of the predictions and results look promising. We are working further to optimize the prediction accuracy.

X. CHALLENGES & SOLUTION

In early phase of project bootstrap team faced following challenges:

Data Availability: Finding enough data on which the ML model could be trained for predicative analysis of risk of developing a Heart Disease turned out to be a challenge. We searched on various open source platforms like Google Big Data & Kaggle to explore the data in various categories which could fit our criteria.

Data Quality: Finding reliable data which can be trusted and is backed by authorities or recognized organizations was also a challenge. After some research, we found data from few reliable

data sources. We detected data outliers using R and performed some data cleanup to make data usable.

ML Algorithm Understanding: we had to develop deep understanding of various Machine Learning Algorithm suitable for the project. To solve this challenge we did a lot of research. Also, we went through ML course content offered by UIUC and found that linear regression technique could be first step to get started.

Remote Team: Our team is scattered all over the globe. We have one person working from Europe while on another has travel job. To meet this challenge, we emphasized on team collaboration by:

- Doing weekly sprints (every Saturday 10AM)
- Utilizing online collaboration tools (like git, zoom, git bot on slack) to the fullest
- Pushing project related communication updates in real time by enabling git bot on our slack channel.

XI. PROTOTYPE

After having thorough discussions team decided on the following prototype:

Step 1: Preprocess and data cleanup

```
>> python preProcess.py
```

In this batch processing step application would cleanup data and formatted data will be used to train the prediction model.

Step 2: Reading user input and providing prediction result

Console:

```
>> python riskAnalysis.py input.json
```

Step3: predict the risk level based on user input using trained model

Console:

```
Computing and matching your data ...
```

```
Output of analysis:
```

```
riskResult: {
  outcome: "MEDIUM"
}
```

The input.json file will have user input in following format

```
userData: {
  age: 30,
  sex: "M",
  cholesterol: 135.65,
  sugar: 153,
  ecg: 23.45
  ...
}
```

XII. RELATED WORK

We found following article in which the author talks about disease prediction tool. However, there is no implementation provided. Also, our research concluded that there is no implementation that scales. Our implementation will utilize Spark ML pipeline capability which can run on any arbitrary Spark cluster.

<https://medium.com/@jonbaldie/predicting-heart-disease-diagnoses-with-machine-learning-2e1a8f5213f8>