Heart Failure & Heart Disease Predictor

Douglas H Waxler

Western Governors University

**Table of Contents**

1. Project Overview3
2. Project Plan3
3. Methodology3
4. **Data Extraction & Preparation4**
5. **Methods and Tools for Analysis4**
6. **Results5**
7. **Key Takeaways5**
8. **Appendices6**

**A. Project Overview**

This project aimed to find out to what degree of accuracy we could predict a diagnosis of heart disease based on combinations of known contributing factors. For Eastern Medical Center’s need, we leveraged patient data to create a predictor to help prioritize patients for testing and care.

In order to successfully accomplish this, we leveraged a dataset of admitted patients, inclusive of 918 male and female patients, and with twelve total features for each. The dataset was relatively clean from the start, with null values only for tests for which not all patients would be submitted.

The project scope included the gathering of patient data, review and analysis including data visualization and reports in R, and the testing of models developed against this data.

Our solution included analysis of the data, transformation of the data types as needed, and validating a model with appropriate levels of accuracy as prescribed in the project proposal. The finished product includes a report generated in R Studio with data visualizations, and test evidence from the machine learning model running on the data to prove its acceptable level of accuracy.

**B. Project Plan**

Data was obtained for the fictitious patient set for Eastern Medical Center, the customer for this proposal. The plan included data analysis and data engineering to be completed. The dataset came with few anomalies or errors, and the analysis in R confirmed this. A few items stood out in the analysis which will be discussed in the next section.

Required cleanup of the data was not as originally anticipated. The features within the dataset had few outliers or apparent outliers, and so the analysis and cleaning step included primarily converting the character or text values into numeric values in a copied data file which can be used for easier correlation and machine learning functions.

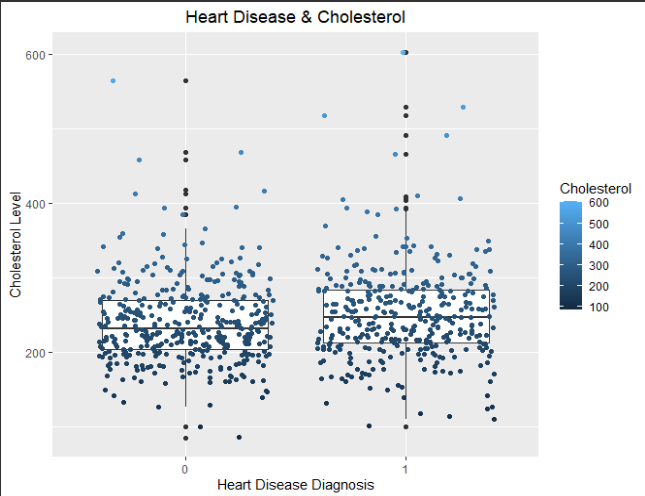
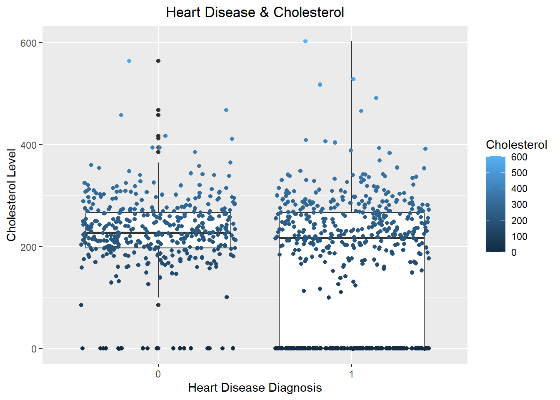
The methodology weighted more heavily toward analysis and correlation of variables, rather than detailed preparation for a machine learning model build.

The anticipated timeline was shortened overall, and therefore the billing may need to be adjusted from the original quoted amounts. The data analysis took longer than scheduled, while the modeling was sourced more quickly than originally planned.

**C. Methodology**

**Data gathering and selection went precisely as planned in this instance. The data set was briefly reviewed prior to project planning to ensure this went smoothly, and we did not have any issues with the records in the set that caused any deviation from the plan.**

**The only obstacles included in the data collection and review were related to metadata clarifications. The visualizations for distributions of cholesterol values against heart disease diagnosis in the bivariate analysis revealed that there were many 0 values. This turned out to be for patients whose cholesterol hadn’t been measured upon admission. You can see how this was handled via limiting the Y axis from the original to the final charting here:**

****

An additional oddity in the data was found with the blood sugar values. Upon reviewing a few records, values of 0 and 1 were found. I had to read into the data dictionary information to find that this was a Boolean value representing patients who had a fasting blood sugar registered at or above 120 mg/dL.

These findings did not necessitate any major changes in our plan and did not represent any major data governance issues. In charting some of these, it was helpful to change the 0 values to NULL for the character fields, so that they would not be included in the plots.

The biggest limitation to the dataset was its size, despite being packaged up nicely and relatively clean from the start. Several major factors referenced in scholarly articles on the topic of heart disease reference many of the features in this dataset as significant contributing factors to a person’s diagnosis of heart disease.

**D.  Data Extraction and Preparation**

For the patient dataset in this project, data was collected from Kaggle to represent the patient population. The data was reviewed in R Studio to analyze the distribution of data within each feature, as well as the correlation between features and particularly the correlation between features and a positive heart disease diagnosis.

Since the data had been reviewed by peers in the industry and cited or reviewed as good data, I had reason to believe it was clean per their feedback. For this reason, I thought it appropriate to examine the data in R based on its powerful capabilities and ease of building visualizations, including plotting correlations between variables.

I started by conducting brief analysis of the individual features’ distribution, and remarked findings along the way. I conducted additional bivariate analysis by starting with a plot of the correlations between variables, and comparing the strong positively or negatively correlated features.

Once this analysis was done, I was able to compare multiple features in a few multivariate plots, and see what conditions mixed more strongly to suggest heart disease was likely, or were more commonly correlated (rather than causing) to the diagnosis.

Two models were reviewed which aimed to predict heart disease diagnoses. In those, the data is transformed using LabelEncoder from SciKitLearn in one of them, and turned into Boolean values in separate features or fields in the other. It seems that LabelEncoder was simpler, but did not provide for the same level of accuracy in the end. Additionally, this model used ADABoost which I previously hypothesized would not be as accurate as linear regression or random forest for this dataset.

**E.  Methods and Tools for Analysis**

To analyze the data, I progressed through a three-step analysis in R Studio. I started by conducting univariate

analysis to get a view of the individual features. I reviewed the distribution of the data, and this

started to uncover a need to further investigate the features in the dataset. As previously mentioned in this report, some features had values which represented things which weren’t immediately obvious, such as fasting blood sugar over a reading of 120 which was represented in Boolean fashion.

I next conducted bivariate analysis. I used the knowledge gained from reference materials I found online, along with my nearly 20 years’ experience in emergency medicine and response to review various features’ correlation with heart disease. I created a correlation plot using numerical representations for character or string data. I reviewed in detail the strong positive and negative correlations from this chart and those are included in the report with some analysis.

Lastly, I conducted some multivariate analysis to see how different features interacted with each other, centering still around the diagnosis of heart disease. These results are again available in the included report.

By analyzing the data with R, I was able to quickly view statistics and summaries of the data. I was able to see the upper and lower limits, mean, and median. I also was able to view a few rows of data to see what the records looked like. R Studio doesn’t handle character data well when it comes to correlating, but it does give the ability to easily display and label the data in charts.

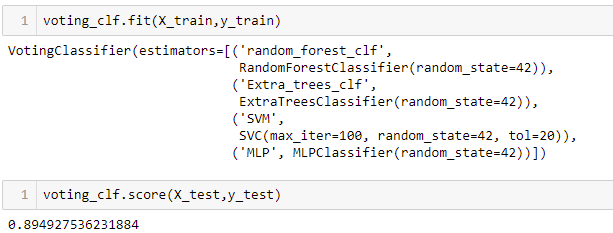
By correlating this data, I was able to easily verify that my hypothesis was true, and that certain factors are strong predictors of heart disease not only in real life application, but also in a machine learning model. This was proven by stronger correlations between variables and our target. In my analysis in R, I did also suggest that additional features and a larger dataset would confirm or deny that these trends remain true across a larger population, and would prove or disprove the trend found in the male or female sexes where males more commonly were diagnosed, but the female population was much smaller.

**F. Results**

With regard to the purpose of this project from the onset, I believe it is proven that if you had a

number of patient vitals and statistics, you could feasibly predict the likelihood that the patient in

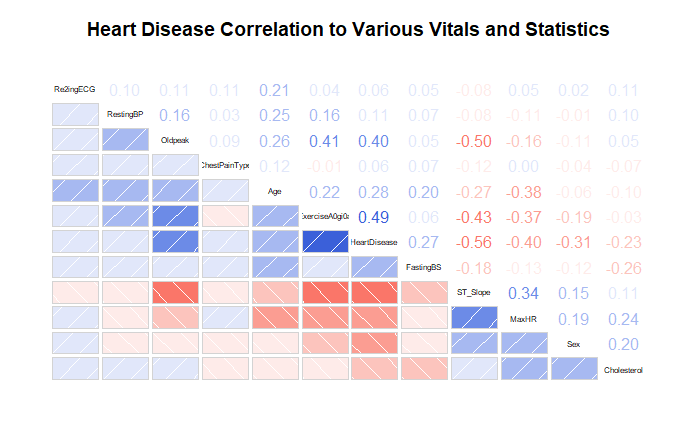
question would or would not be diagnosed with heart disease, and with an acceptable level of accuracy. This information could be used to prioritize testing and treatment of higher risk patients in an emergency medicine scenario.



It was found that you can in fact predict heart disease with recall of greater than 85%, which was our initial requirement. This level of accuracy would increase medical personnel’s ability to prioritize treatments, and likely would have life-saving capability in a scenario where the volume of patients is high. Patients could be assessed, information entered into the digital health record as is done today, and a feature could be added to the application to automatically take those entries, run them through this model, and provide a score which the personnel could use to prioritize care.

**G.  Key Takeaways:**

As hypothesized, there are variables which more strongly correlate with heart disease diagnosis. Since some of the variables correlated in this plot were character values interpreted as numerical values, both strong negative and positive correlations were reviewed.



You can see in this plot and in the report generated in R that the factors that most strongly correlate in some way with heart disease are ST Slope, particularly flat as found in the multivariate portion of my analysis, as well as exercise angina and old peak.

R Studio allows you to easily view statistical data and features’ limits and averages. It is easy to build a story from the ground up, starting by variable, then combining two or three of them.

My first recommendation for Eastern Medical Center is to integrate this prediction with their electronic health record system. This would allow for simple integration into their daily operations. The model could be run and a score returned at a click of a button once the patients’ initial assessments are entered. The time to run such a model can also be minimal, allowing for this to add very little time to the process.

Additionally, I would recommend that Eastern Medical Center hire or leverage existing data analysts and engineers to accomplish two tasks. First, they should report on and monitor the results of this data, and monitor the ongoing accuracy of the model. Second, they should reassess the model and retrain it periodically, especially once a larger sample set is available. I would even be willing to come back and re-run my analysis at no cost if the data becomes available.

****H.** **Appendices****

* Panopto Recording:

<https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=60b471bb-c618-4adb-9c05-adf4011f5198>

* Code, Data Files, Documentation, Additional Materials

<https://github.com/d2waxler/WGU_DMDA_Capstone>

**References –**The following resources were used throughout the capstone project, whether referenced directly or indirectly.

Chicco, Davide. (03 Feb 2020). Machine learning can predict survival of patients with heart failure from

serum creatinine and ejection fraction alone. BMC Medical Informatics and Decision Making.

<https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-1023-5>

DKDKDKDKW. (2021). Practice with some explanation. Kaggle.

<https://www.kaggle.com/dkdkdkdkw/practice-with-some-explanation>

Fedesoriano. (2021). Heart Failure Prediction Dataset. Kaggle.

<https://www.kaggle.com/fedesoriano/heart-failure-prediction>

Mayo Clinic Staff. (n.d.). Heart Disease. May Clinic.

<https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118>

Tarukon5. (2021). Simple Analysis: Recall 89.5%. Kaggle.

<https://www.kaggle.com/tarukon5/simple-analysis-recall-89-5/notebook>

Wilson, Peter W. F.. (12 May 1998). Prediction of Coronary Heart Disease Using Risk Factor

Categories. American Heart Association Journals.

<https://www.ahajournals.org/doi/10.1161/01.CIR.97.18.1837>