Report of INFO411 Assignment 2

Gregory Jackson, Guangjie Guo, Mohan Zhang

This is a report for INFO411 Assignment 2. What we did is data analysis and modelling. The datasets we used is the Heart Disease Data Set ("DS2") [1], which contains heart disease datasets from different locations, Cleveland, Switzerland, Va and so on. The Cleveland dataset is relatively cleaned with very few missing values, so we build our models based on the Cleveland dataset. In addition, we imputed Switzerland and Va datasets to fill many missing values by using KNN and Random Forest strategies, and then tested our models on these imputed datasets. At last, we created several dashboards to demonstrate some interesting issues. We use four sections to elaborate our work: visualization, imputation, modelling, and dashboards. Please look at the details in the main body.

Main body

1 Visualization

In this part, we explore the statistical properties of the data. We build nice pictures to illustrate the properties of the data, such distribution of each feature and target, correlations between features and target, PCA, t-SNE, etc.

1.1 Analysis of Data

As previously indicated, the Heart Disease Data Set ("DS2") has three (3) datasets where each of the datasets have thirteen exploratory features and one response feature. The response feature indicates the degree of heart deterioration using a scale between 0-4. However, the t-sne analysis indicated that partitioning the datasets between five (5) categories did not improve data classification. The decision was made, therefore, to bifurcate the response feature into Healthy and Malignant categories such that categories values > 0 where reclassified as malignant.

The size of each dataset varied: Cleveland (203x14), Switzerland(123x14) and VA (200x14) as well as the number of missing values per dataset: Cleveland (six total), Switzerland(e.g. ca-

118, thal-52, slope-17, oldpeak-6) and VA (e.g. ca-198, thal-166, slope-102). From this analysis, Cleveland is the largest, and most complete; there it was used as the baseline for imputing missing values. Additionally, imputing missing values for Switzerland and VA required a supplicated imputation model.

Finally, heatmap of missing values is available in the Visualization Notebook.

1.2 Statistics Summary

The role of the statistics summary is to give a high-level idea to identify whether the data has any outliers, data entry error, distribution of data such as the data is normally distributed or left/right skewed. The numerical summary used was generated using describe(<dataset>) and was augmented with a wide range of visual graphic.

1.3 EDA Univariate Analysis

At the start of EDA, the feature set was partitioned between Continuous Features ("age", "trestbps", "chol", "thalach", "oldpeak") and Categorical Features (“sex","cp”,,"fbs","restecg", "exang", ”slope", "ca", "thal") where all continuous features have type Float64/32 and categorical features have type Int64. However, most data analysis converted all features to Float64. For continuous features, available visualization are: boxplot, violin, histogram, and KDE where the data points are categorized by Health and Malignant.

Continuous Feature Observations:

- "age" – viewing the *Histogram*, the data suggest that heart malignant rates increase as the dataset population approaches 60 years of age; after 60, an increasing percent of the population is tagged as Maligant. This observation is reinforced with the KDE visualization.
- "trestbps", - give data on resting blood pressure between 90-200. In Cleveland, the clients that are tagged as Healthy or Maligant have value in the full range. This is confirmed by KDE visualization. Therefore, this is not a strong candidate for *predictive strength*. Additionally, the *BoxPlots* that there maybe outliers in both response categories.

- "chol" – indicates serum cholestoral. BoxPlot indicates that there are several data points that may be outliers. But like “trestbps”, the histogram and KDE shows that this feature does not partition the dataset for heart disease.
- “oldpeak" gives the level of ST depression. Both histogram and KDE support the hypothses that this feature maybe a good predictive feature. Also, the BoxPlot indicates that there are possible outliers in the Begnin category.
- “thalach” – give a person maximum heart achieved. Both Histogram and KDE indicate this feature might be a good predictive candidate. The data shows that Healthy clients are able, on average, to achieve a higher maximum heart rate.

Categorical Feature Observations:

The visualization for this class of feature creates a categorical Histogram separating data point between Healthy and Malignant.

- "cp” – give chest pain from 0-3; it is clear that client with cp=4 have a high probability of a Malignant tag.
- "fbs" – fasting blood sugar is not a strong indicator of heart health.
- "restecg" – indicates that an ecg value of 2 indicating ST-T abnormality is a reasonable indicator of heart malignancy
- "exang" – indicate if exercise induced angina was induced; yes is a reasonable indicator of heart malignancy.
- ”slope"- indicate slope of the peak exercise ST segment: a value of 2 (downward slope) is a strong indicator of of heart malignancy.
- "ca"," – number of major vessels occulted; any value above 0 indicates an increasing risk of heart malignancy.
- thal"- thalassemis; a value of 7 (reversible defect) is a reasonable indicator of heart malignancy

1.4 Bi-variant Analysis for Continuous Features

All bi-variant Scatter and Contour plots are generated holding Age as a constant. Age was considered to be the primary factor contributing to heart malignancy. The Contour plot provide a good visual on the distribution of a feature associated with age.

1.5 Multi-variant Analysis

The principal method used for multi-variant analysis was the correlation matrix and related correlation HeatMap. There are two major observation: “thalach “is moderately to strongly correlated with several other feature: “age”, “cp”, “exang”, “oldspeak”, “slope”, “ca”, and “thal”: and “slop” is strongly correlated with “oldspeak”.
This provide possible opportunity for feature elimination.

1.6 Correlations

The most five correlated attributes to the target are 'ca', 'thal', 'oldpeak', 'thalach', and 'cp'. According to their histograms for all kinds of disease types, we have some interesting findings.

1.7 t-SNE:

t-SNE performs well on Cleveland dataset, and it can essentially distinguish between begnin and malignant without considering the malignant category. However, t-SNE cannot distinguish malignant categories. The reason could be too few and imbalanced data points for different malignant categories.
Gregory contributes 1.1-1.5, and Guangjie contributes 1.6, 1.7.

2 imputations

Firstly, we checked the types of missing data, which can be roughly divided into the following two categories. Among the 14 attributes, most of them are MAR (missing at random), and a few are MCAR (missing completely at random). MAR data represents that the loss of data is random and has nothing to do with the data itself, but is related to part of the observed data, such as blood pressure and resting electrocardiographic. MCAR means that the loss of data does not depend on any other variables, such as age, gender, blood sugar, etc. Fortunately, our missing values are basically MAR. Based on the above conclusions, we have considered several options. We finally chose two methods: KNN(K-Nearest-Neighbors) and Random Forest.
Why don't we use mean or SRS (simple random sampling)? The main reason is that most of our missing data is concentrated in columns 12 and 13, which are ca (vessels fluoroscopy) and thal (thallium scintigraphy) respectively, both of which have a certain correlation with

other variables (serum cholesterol). The former one will cause a lot of duplication and the latter one loses the meaning of the data.

Why we use KNN and Random Forest? The main advantage of KNN is that for missing values with strong correlation, other attributes can be used to more reasonably infer the missing values. For example, elderly patients with poor electrocardiogram results have a high probability of developing some complications. However, KNN also has shortcomings, it may lose some creativity for missing values that are not strongly related to other attributes.

The advantages of Random Forests are as follows. Random Forest can handle thousands of input variables without variable deletion. It also able to evaluate the importance of each feature in the classification problem (feature importance). During the generation process, an unbiased estimate of the internal generation error can be obtained. Therefore, it has an effective method for estimating missing data and maintains accuracy when a large proportion of the data are missing.

We tested both methods and found that the data passed through random forest imputation is more accurate than KNN in decision tree model testing.

<u>This part is mainly contributed by Mohan.</u>

3 Modelling

3.1 Classification

In this part, we build our classification models based on the cleaned Cleveland dataset. We fine-tuned and compared several kinds of models, such as Decision tree, SVM, KNN, Random Forest, etc. The results from cross-validation show the best two models belong to KNN and Random Forest. Furthermore, these two models are trained on the whole Cleveland dataset and tested on other locations' imputed datasets. Overall, the Random Forest model wins by a margin of approximately 1.5%. Therefore, RF becomes our final classification model for the further use.

3.2 Regression

In this part, we build our regression models based on the cleaned Cleveland dataset. Three models are fine-tuned and evaluated: Ridge regressor, Ensemble Ridge regressor, Random Forest regressor. The results show: (1) Ensemble Ridge regressor performs more stable than single Ridge regressor; (2) Ensemble Ridge regressor performs a little better than Random Forest regressor.

Guangjie contributes 3.1, and Gregory contributes 3.2.

4 Dashboards

In this part, we build several dashboards to show some interesting stuff.

4.1 Interface of Heart Disease Diagnosis

You can fill the measurement results into the corresponding tables, and a prediction (benign or malignant) will be received.

4.2 KNN modelling

This is to show the performance considering different numbers of K. It can be found that when using K equals 1, the accuracy is not good. By considering 2 or more neighbors, the accuracy can be improved to a large extent. Furthermore, the performance does not differ much as K more than 10 are considered.

4.3 t-SNE

This is to show t-SNE under different number of neighbors considered. It can be found that when using 1 neighbor, it is not enough to distinguish the areas of benign and malignant cases. With 2 or more neighbors considered, it is possible to largely separate benign and malignant cases in different areas. In addition, with more neighbors considered, the data points t-SNE-mapped tend to become more dispersed.

4.4 Dashboards in EDA

Several dashboards are created in EDA part to show distributions of some features. Please see the Pluto notebook EDA_Heart.jl for details.

Guangjie contributes 4.1, 4.3. Mohan contributes 4.2. Gregory contributes 4.4.

Conclusion

We work as a team, and this experience is precious. Through this assignment, we have learned and reviewed many knowledge and skills, such as the skill of analyzing data, the skill of modelling, the skill of imputation, and the skill of creating dashboards. These will be very beneficial for our future work or research.