Data Mining and Decision Systems 600092 Assigned Coursework Report

Student ID: 538629

Date: 07 October 2019

Due Date: 6 January 2020

Report must be within 8 page maximum. Strict page limits will be enforced. Any extra pages will be ignored and no marks awarded for any work on these. Exclusions to this limit are the front page, the references section, and any appendices. Please keep to the given section headings and format; subsections are permitted.

Methodology

The methodology used to execute the data mining of the legacy data was the CRISP-DM methodology. The CRISP-DM model stands for the Cross Industry Standard Process for Datamining and is considered one of the best methodologies out there currently due to the model's ability to be flexible based on the organizations business goals. The CRISP-DM model breaks down the data mining project into 6 sections. Each section has a specific purpose. These 6 sections within the CRISP-DM methodology is not fixed and does not have to be followed in a linear fashion. CRISP-DM model allows you to move back and forth between various stages without having any issues. The 6 sections within the CRISP-DM model are:

- Business Understanding – What you would like to accomplish from a business perspective.
- Data Understanding – Includes collecting data, and then interpreting the collected data.
- Data Preparation – Preparing the collected data for modelling, this includes any form of cleaning or reconstructing data for data to be valid for the next stage of the CRISP-DM cycle.
- Modelling – The objective of the modelling stage is to select a model that can then be trained with the data in order to a valid outcome to be achieved and assessed.
- Evaluation – The outcomes generated for each model is then assessed against the business intentions. This can then lead to new objective as new information and patterns have been discovered.
- Deployment – The information and data discovered during the previous steps are then presented to stakeholders for a strategy to be determined for deployment.

Business Understanding

Set Objectives and success criteria: The main objective of this data mining project is to create a model which would accurately predict weather a given patient is considered a risk or not based on the previous attributes fed into the model during the CRISP-DM stage. This project is considered a success if it performs better than previous models/approaches done before it.

Project Plan: The plan that will be executed for this project to be a success is to clean the data in such a manner to remove any imputes and inaccuracies from the data. Various forms of cleaning will be executed on the data for there to be more data available to be compared and assessed increasing the chances of the project to be a success. This data will be fed into a model, meaning there will be different models all with different accuracies and the most appropriate data will be selected in the deployment stage.

Potential Problems: The potential problems that could occur during the project include cleaning of data in order to find and filter out any imputes within the data frame. If any of the imputes within the data frame remain during the modelling stage, then this could potentially inaccurately give an accuracy score which isn't representative of the true data frame all due to that single impute.

Data Understanding

Description of The Data:

The data frame contains the following attributes for all the patients within the given data frame:

A. Random Column – The random column contains unique integer values for every single patient, meaning every value in this column is unique and can be used to uniquely identify a row within the dataset.
B. ID Column – The ID column contains a unique integer values for every same patient. Meaning every patient has a unique integer value for themselves but that ID value doesn't change for them. This allows the ability to unique identify the patient as they always have the same ID number associated with themselves.
C. Indication Column – The indication column contains 4 categorical data, with the given patient being one of ASx, CVA, TIA, A-F.
D. Diabetes Column – The diabetes column in a Boolean column for the given patient and states weather or not the given patient has diabetes.
E. Hypertension Column – The hypertension column in a Boolean column for the given patient and states weather or not the given patient has hypertension.
F. IHD Column – The IHD column in a Boolean column for the given patient and states weather or not the given patient has IHD.
G. History Column – The history column in a Boolean column for the given patient and states weather or not the given patient has history of a given illness.
H. Arrhythmia Column – The arrhythmia column in a Boolean column for the given patient and states weather or not the given patient has arrhythmia.
I. IPSI Column – The IPSI column contains continuous data which is between 0-100. The IPSI value is linked to brain ischemia which represents the amount of blood flow going to the brain. Insufficient blood flow received to the brain could be linked to various illnesses including a stoke; due to lack of oxygen meaning no ATP can be produced meaning various biochemical reactions cannot occur.
J. Contra Column – The contra column contains continuous data ranging from 0 – 100 and represents the chances of the given patient having contralateral cerebral.
K. Label Column – The label column contains 2 categorical data values which is either risk or no risk. This is based off all the previous column data combined to decide weather the patient is at risk of a given illness based off their medical conditions in the previous columns.

Exploration of Data:

During the data understanding stage, it became apparent that some of the columns within the dataset were not required since they could potentially ruin the results when it came to the modelling phase. These columns were the Random column and ID column. As they increase the search space when clustering in modelling and that each column has no link to the target column they would skew the data causing there to be a dramatic drop in data accuracy as the model is trying to find a potential link between the target column and the Random and ID column. For this reason, I believe that it's best to remove the ID and Random column **before** the modelling stage.

Verification of Data Quality:

There was various potential issue when going to the original dataset. These problems will be explained as I go through each column to explain:

Random Column: The random column did not match the data description for the column, as random column was supposed to be unique and therefore no repeating values. But instead the random column contained many repeating values. This could be a potential issue as it prevents users from uniquely identifying a row within the dataset.

ID Column: The ID column contained no repeated data within the column. Which is fine as it could be that the dataset contained information of patients on a given single day, therefore no repeated patients.

Indication Column: The indication column had two main issues which were one of the column names ASx was mis-spelled to Asx, causing an extra option to be available. Also, there were null values.

Diabetes Column: The diabetes column contained 2 null rows in the dataset.

IHD Column: The IHD column was found to have no imputes within the column.

Hypertension Column: The hypertension column contained null values in 3 of the rows in the dataset.

Arrhythmia Column: The arrhythmia column was found to contain no imputes in the dataset.

History Column: The history column contained two imputes which were null in the dataset.

Contra Column: The contra column contained 2 issues, the first one was it contained a missing value for a single row in the dataset. The second problem was that Contra was found to have the d-type: object, which is incorrect as it should be numeric, meaning it has to be changed to int32.

IPSI Column: The IPSI column was found to contain 4 missing values in the dataset.

Data Preparation:

During the data preparation it was decided to use not one choice of cleaning but 3. The reason for this is it allows greater comparison between a varied approach to how to clean data, allowing for the best method of cleaning data to be decided during modelling.
Cleaning Approaches:

Cleaning Method 1.0 – This employs a system by which every impute found in the dataset will be removed from the dataset. This is the simplest form of cleaning. The main downside to such an approach is that there is reduced number of rows available to be trained on. Meaning less accuracy for the modelling phase.

Cleaning Method 2.0 – This employs a more advanced system by which the mode and in some cases the mean are found of similar rows for a more educated answer to the imputes can be found. This means no rows in the dataset are dropped. Allowing there to be more data available than cleaning method 1.0. The main problem with this approach is that in order to find the mean or mode of a similar row there are quiet a few steps in order to achieve this then find the mode/mean of those given rows. This can be a long and tedious process, as this is done for every impute in the dataset.

Cleaning Method 3.0 – This is the most advanced cleaning approach and uses a decision tree model in order to find the missing values of the imputes in the dataset. This model is trained on 70% of the original dataset.

Selecting Data:
During the selection of data, it was decided that the arrhythmia column, IHD column and ID column all matched the specification of what the column should be, and therefore didn't require to be cleaned.
The random and ID column were dropped in all the cleaning stages from 1.0 – 3.0 mainly due to the reason that they are not required at all in the modelling stage and so as a result must be removed. The reason for why they are not required is stated in exploration of data about under data understanding.

Cleaning the Data

Cleaning 1.0:

During cleaning 1.0 it was found that Indication had 3 null values, Hypertension had also 3 null values, History contained 2 null values, IPSI contained 4 null values, Contra contained 1 null value and label all contained missing data within the dataset. Every single null row in the dataset during cleaning 1.0 was dropped and removed from the dataset, as cleaning 1.0 was the simplest approach to cleaning.

Indication also contained mis-spelt category as instead of ASx there were found to be some rows which were spelt as Asx. These rows with the mis-spelt Asx were all found and removed from the dataset during cleaning 1.0 as well. This is because the mis-spelt ASx was creating another category in indication causing rows which should have been of ASx to not be when it came to modelling.

Contra was also found to be of the incorrect d-type. As in the data specification contra should be of type in as it contains numbers, but all the values in contra were of type object. Instead of dropping the contra column it was decided to instead to just change the d-type of contra as dropping the whole column could potentially ruin the dataset when it came to modelling for 1.0.

Label contained not only missing data but some of the rows in label were to contain unknowns which were slightly different to Nan's and were removed from the dataset in cleaning 1.0.

Cleaning 2.0:

Cleaning 2.0 found all the same imputes and issues with the dataset as in cleaning 1.0, with the only difference was that in cleaning 2.0 all the imputes in the dataset were corrected using either the mode or mean of similar rows. Initially the random column was fixed. As in the data description the random column contains unique values for each row which wasn't the case. Instead every duplicate value in the random column was found then re-assigned a value which was checked to ensure it was unique between 0-1 then added back to the dataset. This allowed random to meet the data specification.

In Indication all the rows that were found to be of Asx were renamed correctly to ASx then added back to the dataset. Ensuring that the mis-spelt error was gone from the dataset. Indication, Diabetes, Hypertension, History, Contra, IPSI and Label were all found to contain missing data. In cleaning 2.0 none of the rows were dropped but instead similar rows which contained the same data for all the other columns were found and either the mean or mode were taken of the data based on weather that given column contained categorical or continuous data. For IPSI and Contra the mean was calculated for the missing rows in the data from all the similar rows.

Contra is found to be of the incorrect d-type as all the items in contra column are of type object. In order to fix this in cleaning 2.0 the d-type of the column is changed to a numeric d-type in order to match the data specification.

Cleaning Model 3.0 – Cleaning 3.0 used a decision tree classifier in order to predict all the missing values in the dataset. When setting up the decision tree classifier there were two available criteria, Gini or entropy. It was decided that the Gini criterion would be used as its less computationally intense compared to entropy, as entropy utilizes logarithms making it a bit slower. The decision tree was trained on 70% of the dataset, then tested on the 30% to try and prevent any overfitting. Overall the decision tree worked with a 99% accuracy.

Why Use Decision Tree – When it came to chose a model during the cleaning stage for cleaning 3.0 there were two routes that could have been taken. This was either not doing a model in cleaning at all or using an Apyori model or use a Decision Tree. Firstly, it was decided to use a black box model for cleaning over/alongside a white box cleaning approach as the author believed that having a method in which it was not fully known how the model makes its specific predictions was a good approach to one method of cleaning. As links between columns in the dataset that was either too complicated or that the author didn't see could potentially be a better method of cleaning. Instead of a simple repetitive cleaning a black box cleaning approach allows for more varied and interesting results too.
The reason for choosing Decision Tree over Apyori was because, firstly, Decision Trees tend to work better on labelled data whereas Apyori tends to work better on un-labelled datasets. As the dataset that is being used on this project is labelled a decision tree would be more preferred over apyori algorithm. Another reason for implementing the Decision Tree over Apyori algorithm is a Decision Tree is pure classification techniques, whereas association analysis using Apyori algorithm creates associations between items with no focus on the target item meaning you would have to go through all the rule sets it creates in order to find the particular rule set for the target item. This seems a more long-winded more complicated approach compared to the decision tree.

Domain Knowledge

Hypertension – Hypertension is highly linked to diabetes as firstly from the dataset it found that if a person has yes to diabetes there is a very high likelihood that they also have

Hypertension. From medical research carried out this finding is also backed up by research conducted. ‘High blood pressure (hypertension) can lead to many complications of diabetes’, (WebMD 2019, Diabetes and High Blood Pressure).

Diabetes – Diabetes is linked to many of the columns in the dataset as conditions such as hypertension and heart attack and stroke. This is because diabetes lowers your bodies natural immunity due to the reduced number of white blood cells increasing the chance of getting many other diseases, (Rowan Hillson, Diabetes and the blood – white cells and platelets)

Indication – TIA is highly linked to many other diseases meaning if a patient has TIA there is a good chance, they have hypertension, diabetes and history as true. This is because A transient ischemic attack (TIA) is a brief interruption of blood flow to part of the brain that causes temporary stroke-like symptoms, (MedicineNet, 2019). A-F is linked to many heart related deaths, (British Heart Foundation, Atrial fibrillation (AF) - Causes, Diagnosis, Symptoms & Treatments). Meaning if you have A-F there is a higher chance you’re type risk in label column. This is because A-F causes an uncommon heartbeat as electrical impulses are fired different places in the atria of the heart.

History – History of a disease is a strong indication that a patient might also have the same disease, based only on history. It was found that most patients that have history also have CVA as type indication. This is because these two diseases are heavily linked as if you have a stroke, you’re also likely in the future to have heart issues/heart disease as they have common risk factors. (American Stroke Association, How Cardiovascular Stroke Risks Relate).

IHD – IHD (Ischemic heart disease) is a condition in which the heart is not receiving enough blood and oxygen and therefore doesn’t have enough energy to contract during a heartbeat. Throughout the dataset it has become apparent that there is a strong link between having IHD and having indication as type A-F. This is also backed up by medical findings that prove there is a link between having IHD and having A-F. AF is associated with increased risk of death only in patients with ischemic heart disease,(pubmed.gov, Atrial fibrillation, ischemic heart disease, and the risk of death in patients with heart failure, 2006). This finding means that there would be an increased link between mortality and IHD and A-F, as suffers of these conditions are more likely to be labelled as a risk.

Modelling

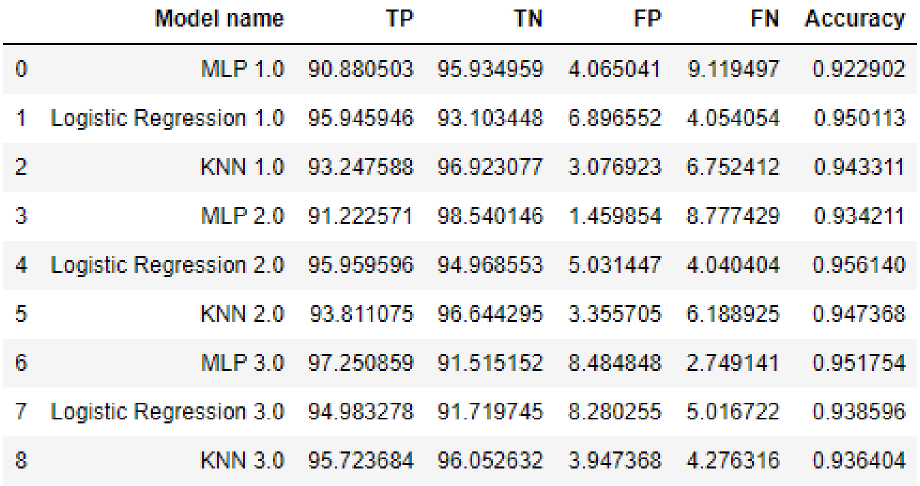
MLP – The multilayer perceptron is a feedforward neural network which contains many hidden layers to allow various complex links/associations between various neurons. The MLP uses weights created between each node in order to find the error using this method repetitively its able to learn within a short period of time. There were many hyperparameters that needed to be set up correctly in order to maximize the accuracy and efficiency of the model. One of the hyperparameters was the number of layers; there were 10 hidden layers in the MLP with a total of 550 neurons. When creating an MLP model using more neurons and more layers allows more complex links to be created, but the downside is that the MLP become more computationally intense meaning a balance must be stuck between complexity of the model and computational intensity. It was also decided to limit the maximum number of iterations to 50000, this is to allow a reasonable links to be made without the model being too computationally intense on the computer. The author also decided to play about with the epsilon value as the MLP solver that was being used is Adam and whilst playing with the epsilon value the accuracy and confusion matrix was always worse when manipulating it, so it was decided to leave the epsilon value. The reason it was decided to use Adam instead of sgd was since adam was more optimized compared to sgd.

Logistic Regression – Logistic regression is a machine learning algorithm that classifies items based on probability gained from previous testing on a percentage of the dataset. The only hyperparameter that was altered with this model was the number of iterations as it was originally 100 iterations, it was changed to 10000 in order to limit the number of iterations and also is to allow the logistic regression model to calculate accurate and reliable probability for the target variable in order for the model to perform at a decent level. If the model didn't not perform on enough iterations this could bottleneck the performance of the model as predictions that should not be made will be based on the lack of iterations the model goes through.

KNN – The only hyperparameter that was altered with KNN model was the n_neighbors which was altered to 5. n_neighbors represents how close each neighbor node is to one-another in the KNN as the model clusters the items together in a given space and if the items in that given space are considered too close it can make it difficult for the model to distinguish one type of item from another. That is why having a relatively distinguishable distance from each node allows them all to be classified with a higher level of accuracy. As when the model was run initially without altering the n_neighbors, the accuracy was a lot less then when it was run with the altered n_neighbors of 5.

Results

In the python file there is a section at the end of the python code called overall results from all models on line 271 and line 272 a dataset has been created which contains all the information from all the models created which compares all aspects of the sensitivity and accuracy of each model. This dataset was created in order to easily compare and decide which model was overall the best model to choose from.



	Model name	TP	TN	FP	FN	Accuracy
0	MLP 1.0	90.880503	95.934959	4.065041	9.119497	0.922902
1	Logistic Regression 1.0	95.945946	93.103448	6.896552	4.054054	0.950113
2	KNN 1.0	93.247588	96.923077	3.076923	6.752412	0.943311
3	MLP 2.0	91.222571	98.540146	1.459854	8.777429	0.934211
4	Logistic Regression 2.0	95.959596	94.968553	5.031447	4.040404	0.956140
5	KNN 2.0	93.811075	96.644295	3.355705	6.188925	0.947368
6	MLP 3.0	97.250859	91.515152	8.484848	2.749141	0.951754
7	Logistic Regression 3.0	94.983278	91.719745	8.280255	5.016722	0.938596
8	KNN 3.0	95.723684	96.052632	3.947368	4.276316	0.936404

Evaluation & Discussion

When going through the dataset in order to find which model is the most optimized for this problem, I believe that the best model is MLP from Modelling 3.0. It comes as no surprise that modelling 3.0 models perform better compared to modelling 2.0 and modelling 1.0 as the dataset which the model was trained upon uses a lot more complex approaches from the decision tree in cleaning 3.0 in order to find complex links between the various columns. This means that when it comes to modelling, modelling 3.0 is already at an advantage.

The reason MLP 3.0 is considered the best model is because from the image MLP has the greatest true positive percentage whilst also having the lowest false negative percentage. This is particularly important with this problem set as potential lives could be on the line you ideally would like a model which is able to given the greatest number of true results, but you would never want to give a patient a false negative result as they would go on the rest of their lives without know they have an illness which means they would also not receive potentially life saving treatment if checked in the early stages. The fact that MLP Modelling 3.0 has the greatest true positive percentage and also lowest false negative percentage whilst also having a decent accuracy score of around 95% is enough justification to chose it as the most ideal model and one that should be picked.

Improvements

There are various ways of doing things looking back that would have been done differently second time round. With some of these being:

- Researching and altering more hyperparameters in order to further understand what they do and weather they could further improve the model from where it's at already.
- Research the field a bit more in order to gain more domain knowledge as I would understand just from the domain knowledge weather an outcome from a model or cleaning would make sense as I would understand how each column is linked to every other column much better that I originally did.
- Potentially not doing 3 cleaning methods as I believe it was a bit too much to not only code up but to also talk about as there were potential slip ups that could have easily been made with the strict 8-page limit (excluding main page and reference page).

References

Any references used throughout the report should be included here in Hull Harvard Style. If no references used, remove this section.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5514868/
https://www.sv-europe.com/crisp-dm-methodology/#dataunderstanding
https://www.webmd.com/diabetes/high-blood-pressure
https://www.practicaldiabetes.com/wp-content/uploads/sites/29/2016/04/Diabetes-and-the-blood-%E2%80%93-white-cells-and-platelets.pdf
https://www.medicinenet.com/transient_ischemic_attack_tia_mini-stroke/article.htm#what_is_a_transient_ischemic_attack_tia
https://www.bhf.org.uk/informationsupport/conditions/atrial-fibrillation
https://www.stroke.org/en/about-stroke/stroke-risk-factors/how-cardiovascular-stroke-risks-relate
https://www.ncbi.nlm.nih.gov/pubmed/17101637