Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

STU33002: Statistical Analysis III Project Report

Author: Nishan Chatterjee

Overview

Cardiovascular diseases (CVDs) are the number 1 cause of death globally, taking an estimated 17.9 million lives each year [1]. The two major causes of CV's are High Blood Pressure (HBP) and Coronary Heart Disease (CHD). There are two types of CHD's: Myocardial Infarctions (MI, or heart attack), and Angina Pectoris (AP, or chest pain). MI's form about 54.5% of the cases among CHD's in the United States of America [2]. Myocardial infarctions occur when the blood flowing to a part of the heart decreases or stops [3]. This causes damage to the heart muscle.

The dataset used in this classification study was originally published by the University of Leicester with the aim to solve two research questions: Predicting the problems of Myocardial Infarction based on information of the patient (i) at the time of admission and (ii) on the third day of the hospital period. Cluster analysis of the disease and disease mapping are other avenues of research which can be done with this dataset.

The publishers of the dataset also note that MI is one of the most challenging problems of modern medicine and that Acute myocardial infarction is associated with high mortality in the first year after it. The incidence rate as noted earlier is high in all countries, especially in the developing ones where its population is exposed to chronic stress factors and irregular and unbalanced nutrition. Out-of-Hospital cardiac arrest is another major problem [2] where 200-300 thousand people die from acute MI before arriving at the hospital every year [4] in the United States. There's also a note of attention disparity with regards to race [5] and gender [6] in the United States which necessitates further analysis of the dataset under analysis. Therefore, predicting MI in order to timely carry out preventative measures is a very important task.

Dataset

This Myocardial Infarction Complications Database was collected in the Krasnoyarsk Interdistrict Clinical Hospital No20 named after I. S. Berzon (Russia) during the time period 1992-1995 [4]. The dataset has 112 input columns consisting of 111 patient details which correspond to a unique ID. There are in total 1700 patient records with 7.6% of missing data. This is not explained but from studies of myocardial infarctions, it can be concluded that this is because certain observations and measurements are relevant and were recorded with only select patients [7]. If we look at the data, we also notice that there's a gender disparity with the female records proportionate to only 37.4% of the total records, while the male record proportion was 62.6%. There was also 0.5% of data missing with no indication of why they were missing. If we look at the Age of the people, then the Mean age was 61.9 with a variance of 11.3. The patient ages ranged from 26 to 92 with the interquartile range being 16. This also indicates that we should be careful while creating models and make sure that there isn't a gender bias towards the method used. The dataset also mentions a record of the number of MI incidents a patient has had over their life-time based on hospital records or the patient's personal recollection if no record was present (INF_ANAM) where 62.5% or the patients had no previous incidents, while 24.2%, 8.7%, and 4.7% of the patients had 1, 2, or 3+ incidents respectively. Here the missing data corresponds to 0.2%. A look at the Lethal Outcome for these values revealed no pattern (My hypothesis was that the results were not known because the patients were no longer alive and had nobody who knew their personal details, but it seems to be simply a case of missing data). Another important consideration to undertake in this situation would be that sometimes noise is added to datasets to not reveal sensitive information about patients [8], but it's not evident if the authors did this in this specific attribute case or to the whole dataset.

There are 12 complications noted in this dataset, 11 of them being binary predictor variables and 1 being a multinomial variable which can be also interpreted to a binary outcome. They are (the proportions mentioned correspond to no complication - complication in order): Atrial Fibrillation (90-10), Supraventricular Tachycardia (98.82-1.18), Ventricular Tachycardia (97.53-2.47), Third-degree AV block (96.65-3.35), Pulmonary edema (90.65-9.35), Myocardial rupture (96.82-3.18), Dressler Syndrome (95.59-4.41), Chronic Heart Failure (76.82-23.18), Relapse of Myocardial Infarction (90.65-9.35), Post-Infarction Angina (91.29-8.71), and Lethal Outcome (84.06-15.94). The Lethal Outcome variable also indicates the type of complication that the living patients had with 6.47% of the patients with Cardiogenic Shock, 1.06% having Pulmonary edema, 3.18% having Myocardial Rupture, 1.35% having progression of congestive heart failure, 0.71% having Thromboembolism, 1.59% having Asystole, and 1.59% having Ventricular Fibrillation. There were no missing values in the complications list.

Research question

The original research questions were the same as the ones listed on the UCI Machine Learning Repository where this dataset was also posted: There are four times for analysing moments of complications which are the time of admission to the hospital, the end of the first day, the end of the second day, and the end of the third day.

This was narrowed down to analysis of the moments of complications during the time of admission to the hospital during the exploratory analysis phase.

Finally, based on time constraints of the study period, this paper discusses only two of the moments of complications which are Chronic Heart Failure (ZSN) and Lethal Outcome (LET_IS) at the time of admission to the hospital. The selection reasons of these two complication targets are discussed in the Methodology section.

Methods

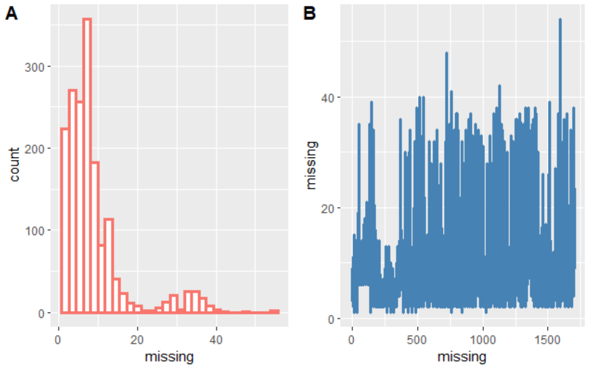
Based on the distribution of the data of no-complication with respect to the complication in the complications observations, we see that there's a polarization due to the less number of complication cases present. So it was the easiest choice to go with the complications which had the most case frequencies. The top five complications predictor variables in descending order are Chronic Heart Failure (394 cases), Lethal Outcome (271), Atrial Fibrillation (170), Pulmonary Edema (159), and Relapse of Myocardial Infarction (159). Here I started with the top two complications: Chronic Heart Failure (ZSN) and Lethal Outcome (LET_IS). This also allowed me to observe both binary and multinomial cases.

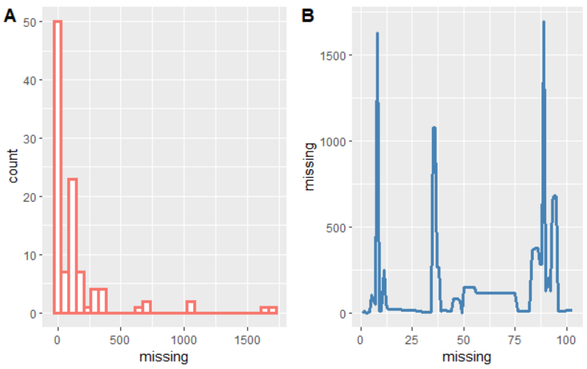
For our input variables, since we're looking at the first moment of complication in this case, we look at all the 111 dependent variables barring 9 columns which aren't relevant for the time of admission as was listed in the data description. After subsetting the data to the relevant variables, we take a look at the missing data. There are two ways to deal with missing data:

1. If they are continuous values, fill them with the mean/median depending on what the data is about. If they are categorical or binary, then replace them keeping the ratio of the categories intact [9].
2. Work with a subset of the dataset by removing and working with only select observations, i.e. start with a dense dataset and then include the sparse variables [9].

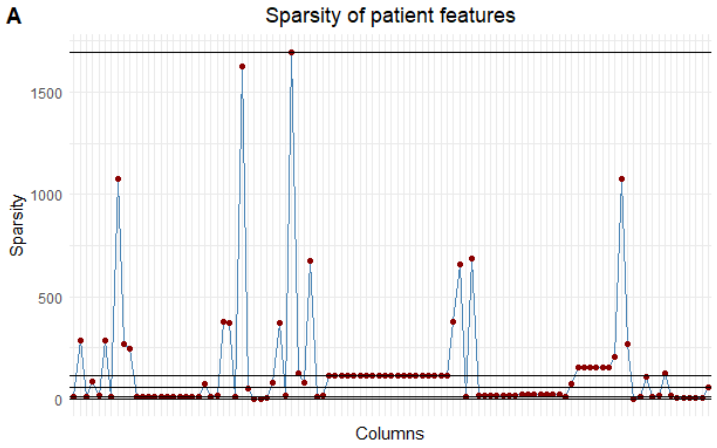
The analysis so far explores working with case 2.

We first look at the dataset by assigning the number of missing values in each row (left) as compared to the columns with the most missing values (right).





We notice that there’s a much bigger sparsity in column values, so we look deeper into this area:

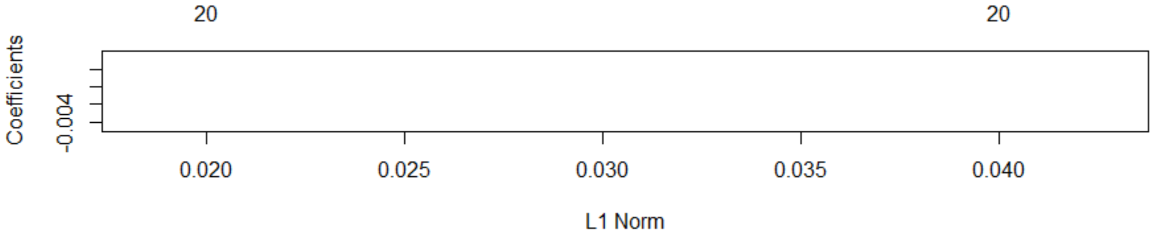


The number of columns with missing values less than or equal to 10 was 22. This looked to be a decent number of features for analysis to carry out to preserve our dense feature space. A similar row analysis was again carried out which revealed that dropping only 50 rows would ensure that our feature space had no missing values. The relevant categories observed are: Age, Sex, Quantity of myocardial infarctions in the anamnesis (INF_ANAM), Presence of an essential hypertension (GB), Symptomatic hypertension (SIM_GIPERT), Obesity in the anamnesis (endocr_02), Thyrotoxicosis in the anamnesis (endocr_03), Chronic bronchitis in the anamnesis (zab_leg_01), Obstructive chronic bronchitis in the anamnesis (zab_leg_02), Bronchial asthma in the anamnesis (zab_leg_03), Chronic pneumonia in the anamnesis (zab_leg_04), Pulmonary tuberculosis in the anamnesis (zab_leg_06), Presence of a right ventricular myocardial infarction (IM_PG_P), Fibrinolytic therapy by Đ¡Đµliasum 750k IU (fibr_ter_01), Fibrinolytic therapy by Đ¡Đµliasum 1m IU (fibr_ter_02), Fibrinolytic therapy by Đ¡Đµliasum 3m IU (fibr_ter_03), Fibrinolytic therapy by Streptase (fibr_ter_05), Fibrinolytic therapy by Đ¡Đµliasum 500k IU (fibr_ter_06), Fibrinolytic therapy by Đ¡Đµliasum 250k IU (fibr_ter_07), Fibrinolytic therapy by Streptodecase 1.5m IU (fibr_ter_08), Use of liquid nitrates in the ICU (NITR_S), and Use of lidocaine in the ICU (LID_S_n).

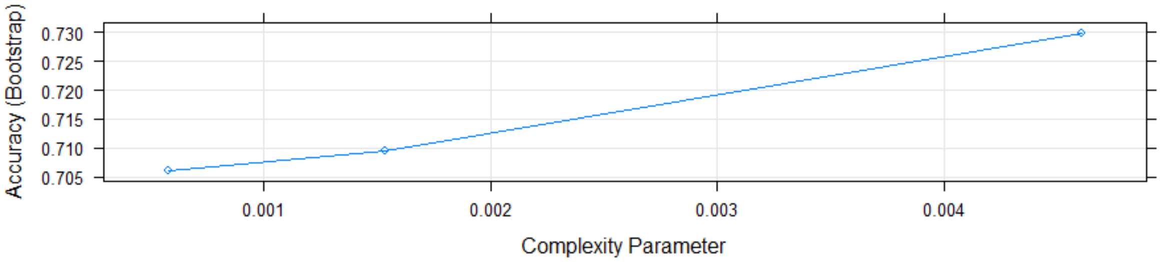
From these variables, we can notice that there's a number of variables that should correspond to one of the complications listed (although not under the immediate analysis), Atrial Fibrillation.

The use of dense features in a health-care database also indicates that the features are general features which are common in treatment for all cases. It could be the case that the variables don't showcase good predictions but since this is case specific, it seemed to be a good avenue to follow up on.

The lasso-mod produced no coefficients for a GLM after cross-validation and fitting with a glmnet. The ridge-mod also produced no interpretable results after using the same method as shown below. We also notice that the fibr_ter_08 was deemed unnecessary in this modeling for ZNS.

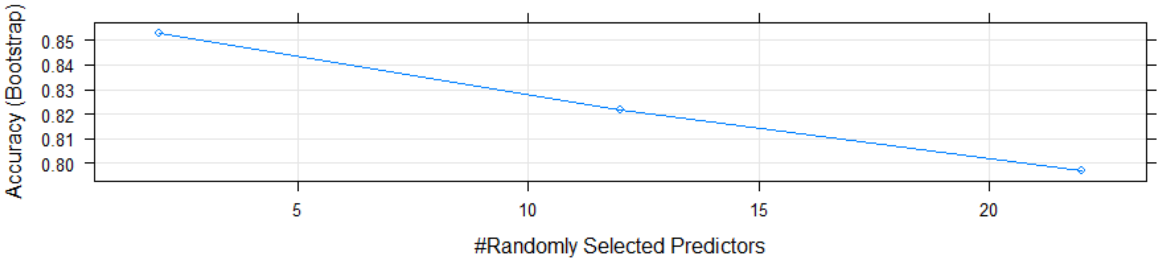


Using the Classification and Regression model with binary family modeling shows the following performance. But it should be noted that the model wasn't run for more than 5 recursions and that since the feature distribution was around 76.82%, this may be a bad combination of features to model with. The CART model is used to rule out defective or unnecessary features.

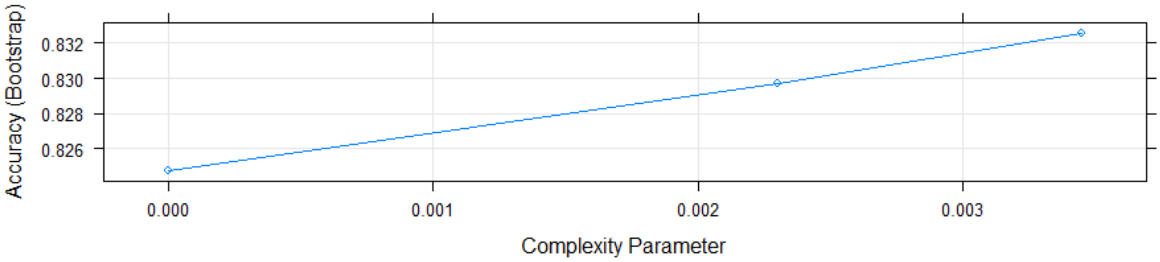


A similar analysis was run on the LET_IS complication with family as multinomial. The three models that were used here are:

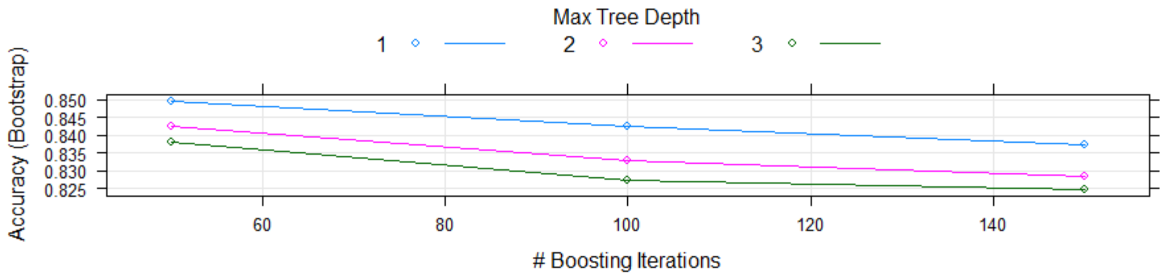
1. A vanilla implementation which shows poor performance.



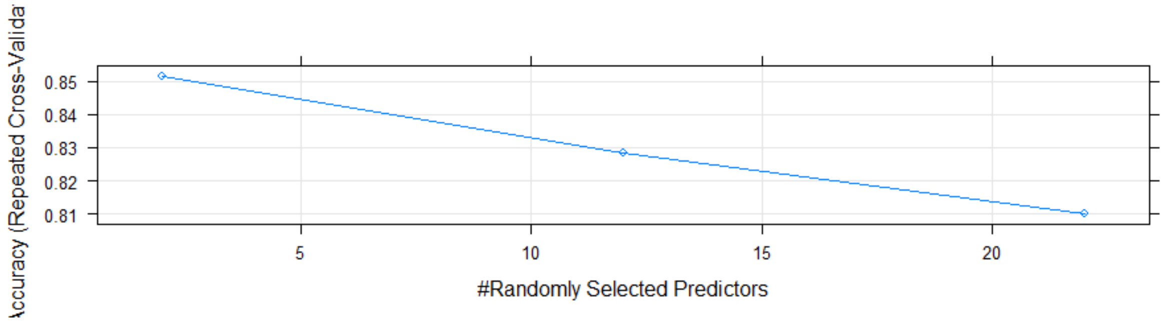
2. A Classification and Regression Model which shows improving performance.



3. A Stochastic Gradient Boosting method shows decreasing accuracy. This indicates that the CART model might not improve with this data-subset, but that it needs additional specialized features to model it. There are 3 tree depths which are done in this comparison which are 50, 100, and 150.

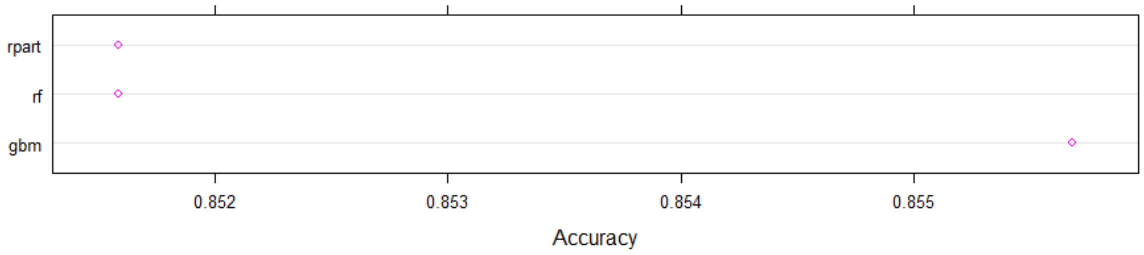


4. Finally, the Random Forest model.

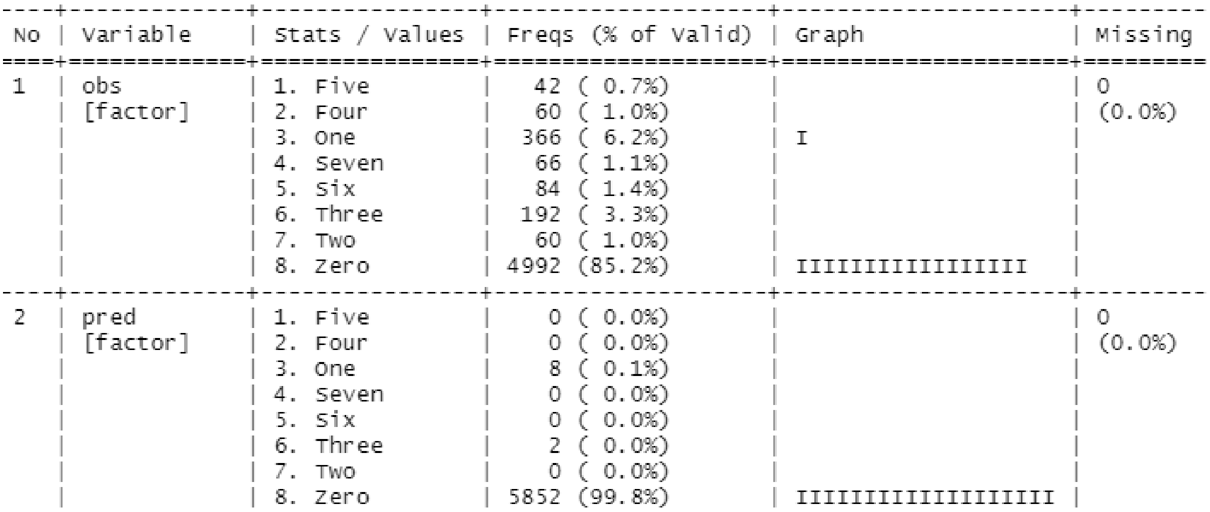


Any checks of if these models were overfitting haven’t been carried out yet.

As a final part of the analysis so far, we compare the models and to check how they perform.



We can see that the GBM model performs the best so far.



```
----+-------------+----------------+--------------------+---------------------+---------
 No | Variable    | Stats / Values | Freqs (% of Valid) | Graph               | Missing
====+=============+================+====================+=====================+=========
 1  | obs         | 1. Five        |   42 ( 0.7%)       |                     | 0
    | [factor]    | 2. Four        |   60 ( 1.0%)       |                     | (0.0%)
    |             | 3. One         |  366 ( 6.2%)       | I                   |
    |             | 4. Seven       |   66 ( 1.1%)       |                     |
    |             | 5. Six         |   84 ( 1.4%)       |                     |
    |             | 6. Three       |  192 ( 3.3%)       |                     |
    |             | 7. Two         |   60 ( 1.0%)       |                     |
    |             | 8. Zero        | 4992 (85.2%)       | IIIIIIIIIIIIIIIII   |
----+-------------+----------------+--------------------+---------------------+---------
 2  | pred        | 1. Five        |    0 ( 0.0%)       |                     | 0
    | [factor]    | 2. Four        |    0 ( 0.0%)       |                     | (0.0%)
    |             | 3. One         |    8 ( 0.1%)       |                     |
    |             | 4. Seven       |    0 ( 0.0%)       |                     |
    |             | 5. Six         |    0 ( 0.0%)       |                     |
    |             | 6. Three       |    2 ( 0.0%)       |                     |
    |             | 7. Two         |    0 ( 0.0%)       |                     |
    |             | 8. Zero        | 5852 (99.8%)       | IIIIIIIIIIIIIIIIIII |
```

We can also notice that the predictions were heavily biased for the non-lethal outcome case as shown here.

Results

The feature space chosen and modeled in this work so far reveals that the direction that was adopted of how and which input features to choose from, that is,

“Work with a subset of the dataset by removing and working with only select observations, i.e. start with a dense dataset and then include the sparse variables”

don’t seem to be the ideal choice. But they are necessary features which should help other dependent variables.

Conclusions and Future Work

In conclusion, there needs to be a continuation of work in the avenues of:

1. Other complications
2. Choosing more features while modeling
3. Different approaches to handling missing data
4. Exploring medically informed views of which predictor variables have been found to correlate with which complication and then simulating those results to check for consistency.