Prediction of cavernous malformations based on clinical data of a patient

Pavlína Ružičková

January 30, 2024

1 Introduction

In my project, I work with data from the following study: Modifiable Cardiovascular Risk Factors in Patients With Sporadic Cerebral Cavernous Malformations. Data comes from 1219 patients and the main goal of the research is to explore whether obesity influences the risk of brain bleeding. Data also includes risk factors such as arterial hypertension, diabetes, hyperlipidemia, nicotine abuse, and obesity... This research is a classic example of survival analysis on clinical data where the time until the event - bleeding, is analyzed. **My goal is to predict whether a patient will suffer hemorrhage according to their clinical data**. This could be very useful when monitoring patients and prescribing treatment.

2 Data

We have data from 1219 patients, followed in 5 years, monitoring 20 parameters about them. First thing, we removed patients with NAs in any column except in the targeted event column. We are now left with only 339 rows however this was the safest way to do this because all the columns I dropped the NAs at were factor (boolean) data. This is a big challenge in our data because all except age (patients were diagnosed from age 18 to 79) are factors (1 indicates the person has diagnoses or is a smoker or is obese... and 0 means the person doesn't meet the criteria in this column). I visualized Age at diagnoses, being the only numerical feature, with the information of rebleeding. If there was a correlation, events would only show on the upper part of the graph 1. Also the correlation was only -0.007. In the correlation matrix in 2 we see there aren't strong correlations with Event.

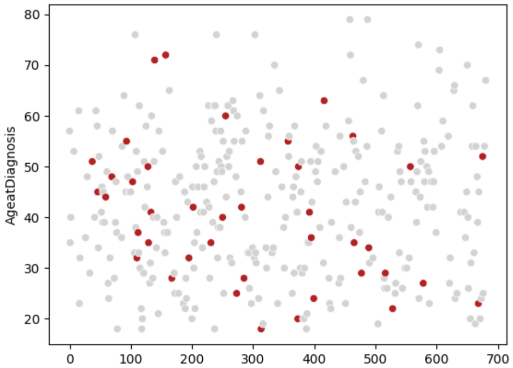


Figure 1: Age at diagnosis and visualized events

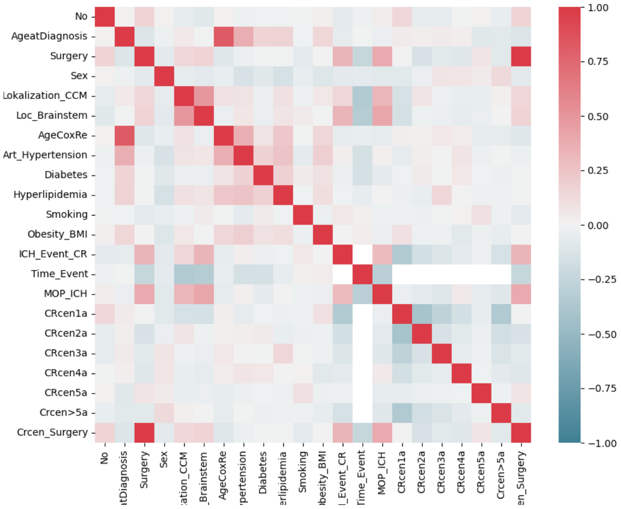


Figure 2: Enter Correlation matrix

3 Logistic regression

I applied a logistic regression model to predict the event of bleeding. So, it is a simple binary classifier, where I put all medical patient information and predict 1 or 0.

I did a stratified split of data so that we have 30 percent of people with positive event in the testing set. With parameters: $solver =' liblinear', C = 3, class_weight = ''balanced''$, I was able to achieve relatively good classification. In the runs, I usually got none or only one false negative from 13 positive patients in testing set, and about 12 percent of false positives, as shown on the confusion matrix 3. **When I performed this 1000 times, I achieved an average accuracy of 0.92, average recall 0.95, average F1 score 0.76, and average precision 0.63.** Precision is quite low, but as it is true positives/all positives I think it is not as relevant as other metric scores. With different parameters I was able to achieve higher accuracy and precision, however, this improved the classification of 0 events but worsened the prediction of bleeding. In this case (I assume), a false positive is much better than a false negative. I suspect a false positive could tell us we need to take a look at a patient who in the end won't suffer bleeding. But false negatives could only mean we are more cautious about a "healthy" patient. So I opted for the above-mentioned model.

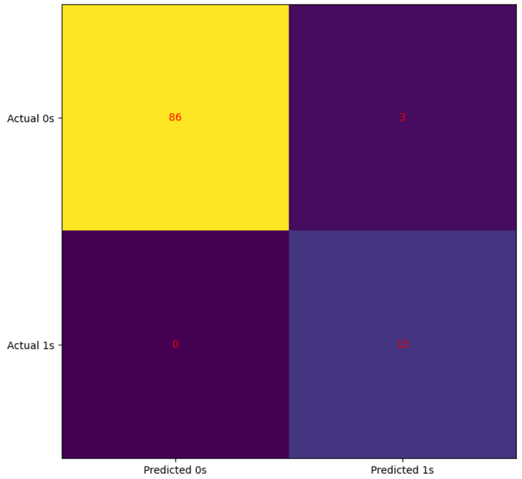


Figure 3: Confusion Matrix of logistic regression model

4 Decision tree

Similarly, I used a decision tree to predict the outcome of O/1 event. After validating I have following model parameters: $class_{w}eight = ”balanced”, min_{s}amples_{s}plit = 15, max_{d}epth = 10$. Once again, maximizing the depth or minimizing samples in a split could improve my overall accuracy, but I was aiming to improve the number of false negatives. **When I performed this 1000 times, I achieved an average accuracy of 0.90, average recall 0.88, average F1 score 0.70, and average precision 0.58**. Which are overall worse scores than in Logistic regression and I was bot able to train a better decision tree model. When we look at histograms of the percentage of false positives(5), it is mostly around 10 percent. False negatives (4) are mostly under 10 percent however there were some cases when the prediction of bleeding was very bad. I find this decision tree to be inadequate, however I can look at the importance of features. 6 Obesity, different CRcen measures, Surgery...were importnant

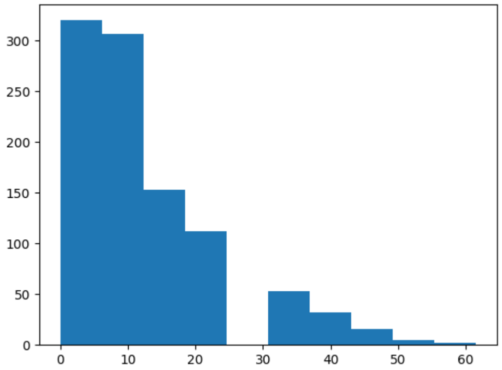


Figure 4: Percentage of bad classification of positive patients

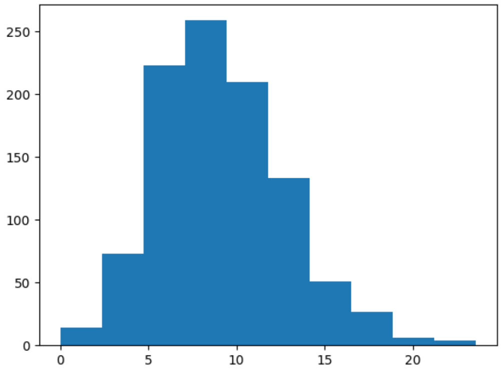
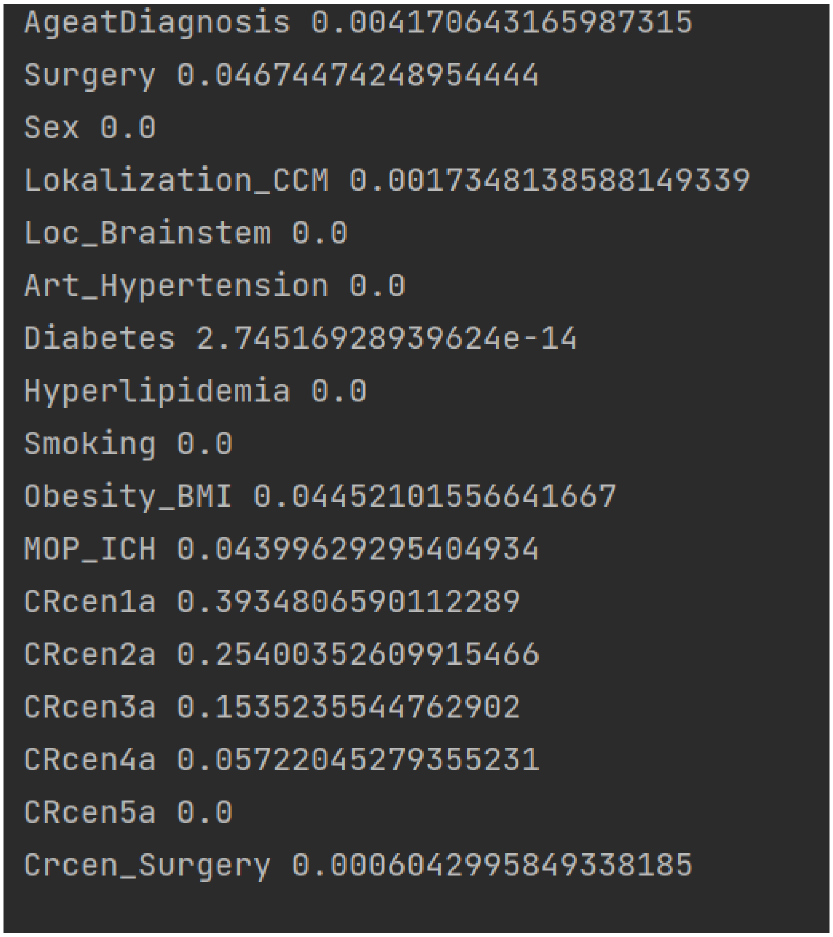


Figure 5: Percentage of bad classification of negative patients

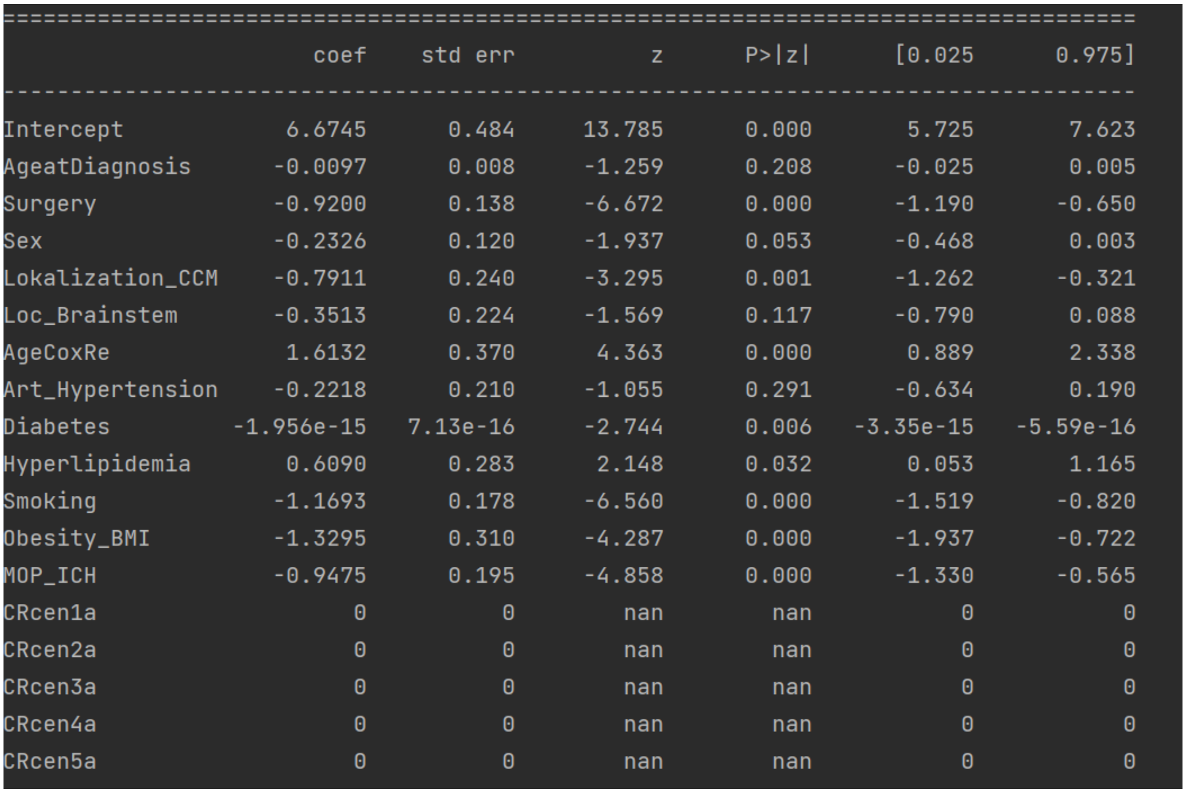


```
AgeatDiagnosis 0.004170643165987315
Surgery 0.04674474248954444
Sex 0.0
Lokalization_CCM 0.0017348138588149339
Loc_Brainstem 0.0
Art_Hypertension 0.0
Diabetes 2.74516928939624e-14
Hyperlipidemia 0.0
Smoking 0.0
Obesity_BMI 0.04452101556641667
MOP_ICH 0.0439962929540493 4
CRcen1a 0.3934806590112289
CRcen2a 0.25400352609915466
CRcen3a 0.1535235544762902
CRcen4a 0.05722045279355231
CRcen5a 0.0
Crcen_Surgery 0.000604299584933818 5
```

Figure 6: Importance of features in decision tree model

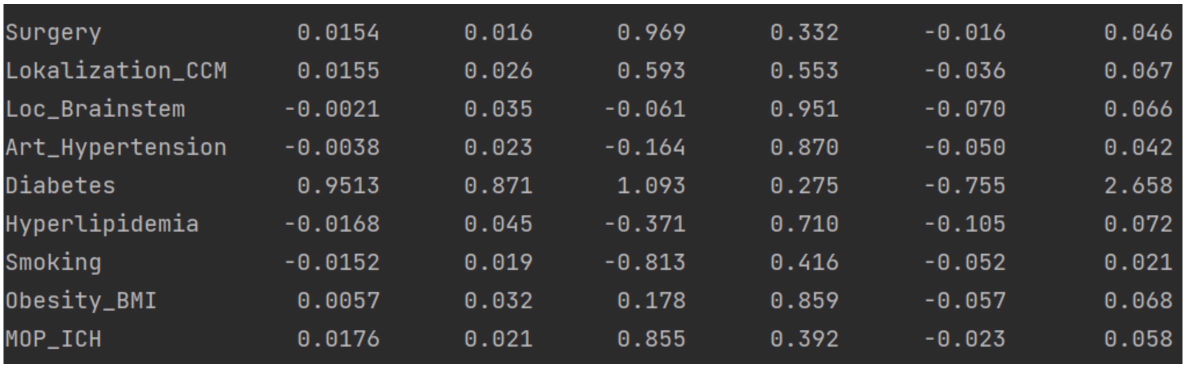
5 Generalized regression model

I used glm model to predict months until the bleeding will occur. As it would be difficult to include people who did not suffer bleeding at all (0 months means the person suffered bleeding immediately after data collection, so we would have to put > 5 years to those with e negative vent as we do not know what happened to them after the research was done and that would be very imprecise), I only predicted these months for people we know would suffer the bleeding. In reality, we could take the outcome of logistic regression, and if the prediction was positive, we would use this model to predict months. I started with a full model 7, and gradually removed the features that had the lowest p-value. I ended up with a model with features shown on 8. We can see variables such as factors of brainstem neurons, hypertension or obesity were influential. This model, in 100 runs had a mean of absolute differences between prediction and actual time event equal to **21 months**. We can also see histogram 9, the difference is usually around two years but there were runs where the difference was quite big.



```
=====================================================================================
                        coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------
Intercept             6.6745      0.484     13.785      0.000       5.725       7.623
AgeatDiagnosis       -0.0097      0.008     -1.259      0.208      -0.025       0.005
Surgery              -0.9200      0.138     -6.672      0.000      -1.190      -0.650
Sex                  -0.2326      0.120     -1.937      0.053      -0.468       0.003
Lokalization_CCM     -0.7911      0.240     -3.295      0.001      -1.262      -0.321
Loc_Brainstem        -0.3513      0.224     -1.569      0.117      -0.790       0.088
AgeCoxRe              1.6132      0.370      4.363      0.000       0.889       2.338
Art_Hypertension     -0.2218      0.210     -1.055      0.291      -0.634       0.190
Diabetes          -1.956e-15   7.13e-16     -2.744      0.006   -3.35e-15   -5.59e-16
Hyperlipidemia        0.6090      0.283      2.148      0.032       0.053       1.165
Smoking              -1.1693      0.178     -6.560      0.000      -1.519      -0.820
Obesity_BMI          -1.3295      0.310     -4.287      0.000      -1.937      -0.722
MOP_ICH              -0.9475      0.195     -4.858      0.000      -1.330      -0.565
CRcen1a                    0          0        nan        nan           0           0
CRcen2a                    0          0        nan        nan           0           0
CRcen3a                    0          0        nan        nan           0           0
CRcen4a                    0          0        nan        nan           0           0
CRcen5a                    0          0        nan        nan           0           0
```

Figure 7: Full GLM model



```
Surgery               0.0154      0.016      0.969      0.332      -0.016       0.046
Lokalization_CCM      0.0155      0.026      0.593      0.553      -0.036       0.067
Loc_Brainstem        -0.0021      0.035     -0.061      0.951      -0.070       0.066
Art_Hypertension     -0.0038      0.023     -0.164      0.870      -0.050       0.042
Diabetes              0.9513      0.871      1.093      0.275      -0.755       2.658
Hyperlipidemia       -0.0168      0.045     -0.371      0.710      -0.105       0.072
Smoking              -0.0152      0.019     -0.813      0.416      -0.052       0.021
Obesity_BMI           0.0057      0.032      0.178      0.859      -0.057       0.068
MOP_ICH               0.0176      0.021      0.855      0.392      -0.023       0.058
```

Figure 8: Final GLM model features

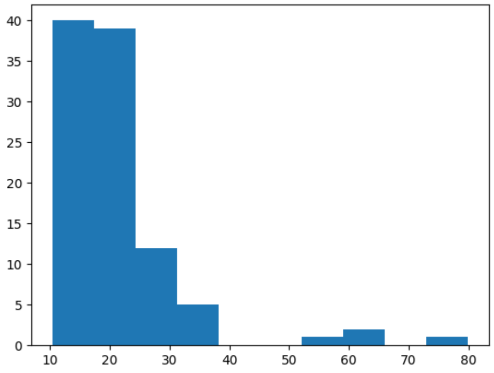


Figure 9: Histogram of absolute differences between prediction and reality in months

6 PCA

I tried to visualize our patients using PCA. In many runs, data always fell in a similar pattern as shown on 10. Data was separated alongside the x-axis into two groups, where the upper group always included more positive events. This could be used to categorize higher and lower-risk patients. Also, it could explain why some positive events fail to be classified by models as they have features more similar to healthier people. I tried the kmeans algorithm on this PCA data and analyzed the correctness of prediction into these two groups. However, I was not successful in finding a good model, and as the separation is not perfect I did not think this model would be efficient. I believe the problem with PCA in our data is, that we only have one numerical variable and other boolean or factor variables do not have big variances, so this method is not very helpful.

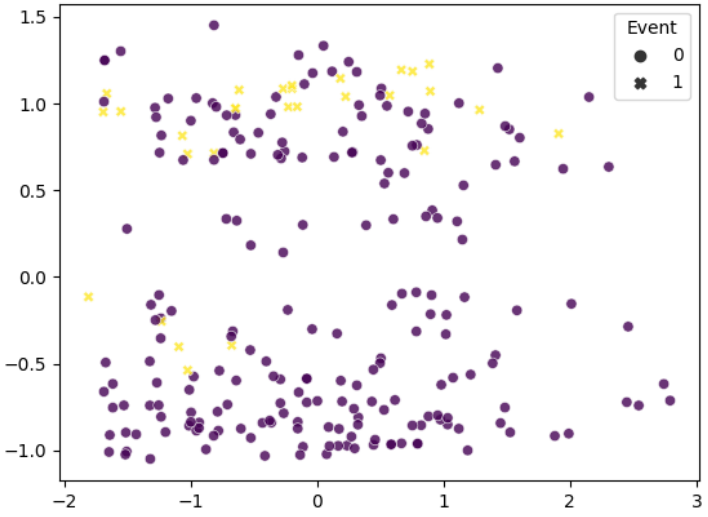


Figure 10: PCA, events visualized with yellow color

7 Conclusion

We were able to create a predictor of the rebleeding of patients according to their clinical data. Although the results were not perfectly precise, logistic regression performed the best. In reality, it is very difficult to predict the medical behavior of individuals, as I read in the official article risk of rebleeding is not currently assigned to specific patients, however, they concluded that factors such as obesity contribute to the chances. We saw this in our model as well, as obesity was an influential feature in all models. So we have ended up with the same conclusion as the survival analysis research with our ML approach. In practice, doctors could not use this model to confidently separate patients, however, I think with the accuracies we achieved, medical professionals could use this model to highlight higher-risk patients. A generalized regression model could shine more light on the prediction of time until the rebleeding. We could predict a two-year interval of high chance of rebleeding. It is also worth pointing that all of our data is from patients, from people who once suffered hemorrhage, so their state might be more volatile than of healthy people. And even if the rebleeding didn't occur during the study, it might have happened after the research, which could account for some of the differences in our predictions. And sometimes, obviously, unfortunate medical tragedies can occur in otherwise healthy-marking people.