LEPL1109 - Statistics and Data Sciences
HACKATHON 2 - Diabetes health indicators

Group n°46 November 29, 2024

Lastname	Firstname	Noma
Bellens	Romain	30642100
Bien	Jonathan	43742200
Jacques	Louis	90802200
Bognar	Bence	32392200
Peduzzi	Lionel	69692200
Zaoudi	Ismaël	60552200

Please, read carefully the following guidelines:

- Answer in English, with complete sentences and correct grammar. Feel free to use grammar checker tools such as LanguageTools free and open-source plugin;
- Do not modify questions, and input all answers inside `\begin{answer}...\end{answer}` environments;
- Each question should be followed by an answer;
- At the end of each question, there is the length of the expected answer. This is for your information but it is not too important if you do not respect these recommendations.
- Clearly cite every source of information (even for pictures!);
- Whenever possible, use the `.pdf` format when you export your images: this usually makes your report look prettier[1];
- Do not forget to also submit your code on Moodle.
- **Reminder:** You need to belong to a group to submit your project on Moodle.

Contents

[1] This is because `.pdf` is a vector format, meaning that it keeps a perfect description of your image, while `.png` and other standard formats use compression. In other words, this means you can zoom as much as you want on your figure without decreasing image resolution. For simple plots, vector formats can also save a lot of memory space. On the other hand, we recommend using `.png` when you are plotting many data points: large scatter plots, heatmap, etc.

1 Description of the project

1.1 Your objective

You work in the diabetology department at **Saint Luc University Hospital**. The head of the department has asked you to find a solution for classifying and predicting **whether patients are at high risk of developing diabetes**. This will enable them to schedule an appointment with these patients to set up prevention tools. To do this, you have a database of patients who have passed through the department in recent years. In addition, the head of the department feels that the poll is too long, and would like to **reduce the number of questions while maintaining the reliability and quality of the results**. The attached `.ipynb` file will guide you in this process.

Your aim is to determine which characteristics are relevant and enable reliable patient classification. **Be careful**, don't let a potential diabetic patient slip through the cracks.

1.2 The dataset

The dataset is a real dataset based on a questionnaire carried out in the USA some ten years ago. It contains around 70 000 entries and is a collection of 22 features individually defined in table 1.

Features name	Description	Range
`Diabetes`	Diabetes (0:no diabetes; 1:diabetes)	{0, 1}
`HighBP`	High blood pressure (0:no; 1:yes)	{0, 1}
`HighChol`	High cholesterol (0:no; 1:yes)	{0, 1}
`CholCheck`	Cholesterol check in 5 years (0:no; 1:yes)	{0, 1}
`BMI`	Body mass index in [kg/m^2]	/
`Smoker`	Smoked at least 100 cigarettes in your life (0:no; 1:yes)	{0, 1}
`Stroke`	Stroke (0:no; 1:yes)	{0, 1}
`HeartDisease`	Heart disease (0:no; 1:yes)	{0, 1}
`PhysActivity`	Physical activity in past 30 days (0:no; 1:yes)	{0, 1}
`Fruits`	Consume fruit 1 or more times per day (0:no; 1:yes)	{0, 1}
`Veggies`	Consume vegetables 1 or more times per day (0:no; 1:yes)	{0, 1}
`Alcohol`	Heavy alcohol drinkers (0:no; 1:yes)	{0, 1}
`AnyHelathcare`	Health insurance (0:no; 1:yes)	{0, 1}
`NoDocbcCost`	No doctor because of cost (0:no; 1:yes)	{0, 1}
`GenHlth`	General health (1:excellent; 5:poor)	{1, . . . , 5}
`MenHlth`	Number of days out of the last 30 when mental health was poor	{0, . . . , 30}
`PhysHlth`	Number of days out of the last 30 when physical health was poor	{0, . . . , 30}
`DiffWalk`	Serious difficulty for walking (0:no; 1:yes)	{0, 1}
`Sex`	0:female; 1:male	{0, 1}
`Age`	Age category (1:18-24; . . . ; 13:80 or older)	{1, . . . , 13}
`Education`	Education level (1:never; 6:university)	{1, . . . , 6}
`Income`	Income scale (1:less than $10,000; . . . ; 8:$75,000 or more)	{1, . . . , 8}

Table 1: Data set features

2 Questions and answers (4/10)

Question 2.1:

(1/10) What happens to the precision and recall (of any method) when the threshold tends to 0? And when it tends to 1? How can you explain it?
Expected answer length : 8 lines.

Answer to 2.1:

When the threshold tends to 0 recall increases and precision decreases, while when the threshold tends to 1 the method's precision increases but recall decreases.
This can be explained by looking over the definitions of precision and recall and seeing how the threshold will influence the classifications we obtain.
Recall is defined as the number of true positive (TP) over real positive (RP) $Recall = \frac{TP}{RP}$ and Precision as the number of true positive (TP) over perceived positive (PP) $Precision = \frac{TP}{PP}$ cases for a given threshold.
When the threshold tends to 0 the number of PP cases increases as more and more negative cases get classified as positive by our method decreasing Precision but in the meanwhile lets us correctly identify all positive cases correctly increasing TP cases and thus Recall as well.
Tending the threshold to 1 has the opposite effect decreasing TP cases and increasing false negative (FN) cases while reducing PP cases causing low Recall and high Precision.

Question 2.2:

(1/10) Explain which precision/recall trade-off you prefer to have for the specific task asked in this hackathon: don't let a potential diabetic slip through the cracks. How should you adjust the threshold of your model to bring it closer to the desired trade-off? Should it be above or below the default threshold value of 0.5?
Expected answer length : 5 lines.

Answer to 2.2:

Since we are working in the medical field, our main objective is to make prevention. Therefore, we prefer to detect as many patients at high risk of developing diabetes as possible, which means we want **high recall**. As we have seen in the course, the recall curve decreases monotonically as the threshold increases. To maximize recall, we need to **set the threshold lower than 0.5**.

Question 2.3:

(1/10) Based on your code, select a final model that you will keep as classifier. **Justify.**
Expected answer length : 5 lines.

Answer to 2.3:

We take as model **linear classifier** with $\tau = 0.28$ because this model meet specifications (i.e. recall at least 95% and F1 score at least 75%). Specifically, when using this model with the full questionnaire, which includes 21 features, we observe a recall of 0.957 and an F1 score of 0.757. It is better than logistic and knn classifiers for the same threshold and questionnaire length.

Question 2.4:

(1/10) Could you reduce the length of the questionnaire? If so, how many questions? Which questions? **Justify.**
Expected answer length : 6 lines.

Answer to 2.4:

We can see that the F1-score and the recall are above the limits fixed with 5 or more features. In order to have a model not to complex and to reduce the time of compilation of our code, we chose to use only 5 most correlated features. More than 5 features does not improve the F1-score and the recall. The limits were 95% for recall and 75% for F1-score. We need a high recall and F1-score to minimize the number of people who has a high risk of developping diabete not detected.

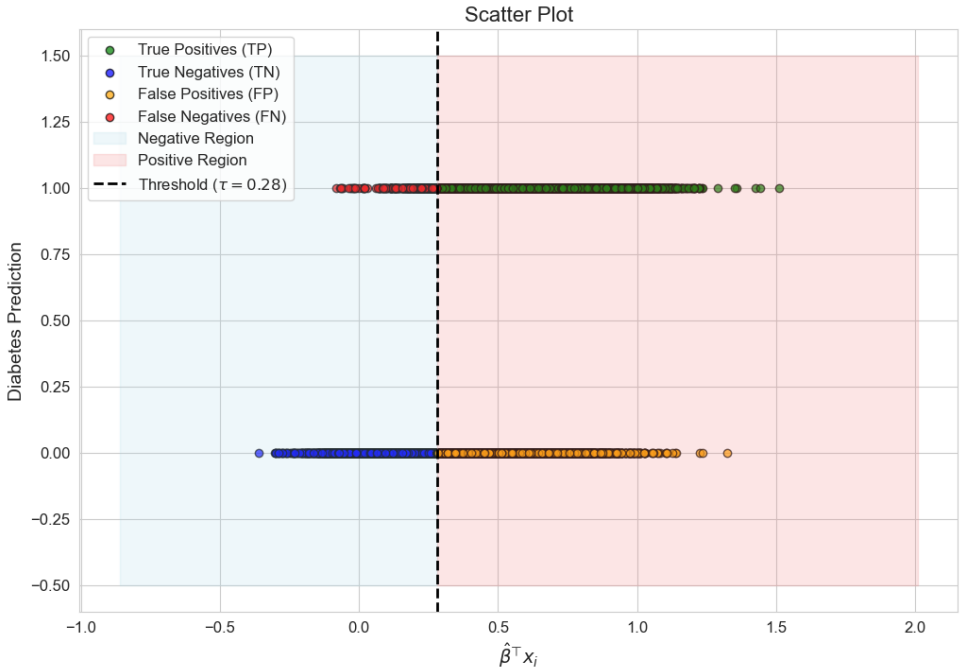
3 Visualization (2/10)

Question 3.1:

(2/10) To answer this question, we ask you to produce a clear, clean figure expressing a result or giving an overall vision of your work for this hackaton. Please feel free to do as you wish. Be original! The clarity, content and description of your figure will be evaluated.
Expected answer length : 4 lines + 1 figure.

Answer to 3.1:

We choose to illustrate our final model (i.e. the linear classifier with $\tau = 0.28$) as a scatter plot. Thanks to this visualization, we are able to see the behaviour of the classifier in function of the given threshold (e.g. $\tau = 0$). It allows us to have an overall feeling about what recall and precision are. For the aesthetic aspect, we relied on ChatGPT's assistance.


Scatter Plot
True Positives (TP)
True Negatives (TN)
False Positives (FP)
False Negatives (FN)
Negative Region
Positive Region
Threshold ($\tau = 0.28$)
Diabetes Prediction
1.50
1.25
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
$\hat{\beta}^\top x_i$