Lab 1. Diabetes Prediction lab

OIDD 2550, Professor Tambe

Learning Objectives for this lab

- Using a familiar tool - Excel — to predict a target variable.
- To start to develop a common understanding about machine learning models and the tradeoffs that arise when using them.
- Understand the concept of error when using prediction models.

Data Context

The setting for the exercise is healthcare. Machine learning models are becoming widespread in healthcare diagnostics, and using machine learning for diabetes prediction has become quite common. [1, 2]

This lab uses a popular and common machine learning model — “logistic regression” — to predict outcomes from patient data. In this context, given access to other indicators of patient health (that may be readily available or easy to collect), the goal would be to predict whether a patient has diabetes (or will have it soon).

We have provided an Excel spreadsheet with a logistic regression model already built into the spreadsheet for the PIMA Diabetes data. For a large sample of patients, these data contain information on a variety of health indicators for patients as well as whether they have diabetes, denoted by a 1 (if they have diabetes) or a 0. If you would like to

read more about the diabetes and health issues associated with the PIMA people (a community of Native Americans from Arizona and Northwestern Mexico), read here (optional).

Modeling

It is not essential that that you understand ML concepts for this exercise - that comes later. This lab is meant to get you to go hands-on with some key concepts before we formally cover them in class.

Regression and prediction

It is important to know, however, that different types of models can be used to make predictions. You may have come across "regressions" in the context of fitting a line (or another shape) to a set of known data points. Once these models are fitted, they allow modelers to take new data and predict outcomes for unknown data points. In other words, we can use a set of known data points (*x*'s) with outcomes (*y*'s) to fit or "train" a regression model, and then use the fitted model to predict outcomes for data points (*x*'s) where we do not know the outcomes.

In the diabetes context, we can use data on patient indications and known diabetes outcomes to find how the patient indications can be combined to predict whether a patient is likely to have diabetes. We can then use this fitted model to predict whether new patients, for whom we do not know a diabetes diagnosis, might have diabetes.

Logistic regression

Logistic regression is commonly used when the outcome variable to be predicted is "binary" (i.e. it takes either the value 0 or 1). For instance, you might use logistic regression to predict whether someone is likely to be approved for a home loan (which can only take the value 0 or 1), but not for predicting a house price (which can take any positive value).

Logistic regression-based prediction proceeds in several steps.

- First the logistic regression model is fitted using data on the available features (i.e., your "x" variables) and known outcomes for the patients. This step computes the best weights (i.e. coefficients) to apply to the features to predict the outcome values as well as possible.

- The output of the fitted logistic regression model is the *predicted likelihood* (a probability value that lies between 0 and 1) that the outcome variable takes the value 1.

- The final step in using the prediction model to predict new outcomes is to choose a *threshold value* that the predicted likelihood value must exceed for the observations to take a predicted value of 1.

In this lab, we use a logistic-regression based prediction model to understand the tradeoffs that arise when building predictive models. If you would like to learn more about logistic regression, there are many resources available online. For example, you can read here about logistic regression (optional).

Spreadsheet

For this lab, you are provided with a spreadsheet (click here to access the spreadsheet). In this spreadsheet, you are provided with data on the PIMA people, including the health indicators for each patient (**labeled as x1 - x8 in columns A-H in the spreadsheet**) and whether they have diabetes (**column I**).

You are also given a column for predictions from a fitted model (**column M**) and another column for the predicted value for whether the patient has diabetes (**Column Q**). Note! The model has already been fit for you, using the data available on this sheet. You do not need to do anything further with the model coefficients.

Notice that predictions from the fitted model are all ***probabilities*** that an individual has the diabetes condition. As described above, a threshold value is required to convert these probabilities into a ***binary indicator (***1 or 0) of whether an individual is predicted to have diabetes. You can see in this spreadsheet, an initial threshold value of 0.75 is used (**cell Q2**), so using this model, all patients with a predicted probability greater than 0.75 would be predicted to have diabetes.

The data scientists/ML engineer's challenge is to find models that fit the data well, and to adjust parameters like the threshold value to make accurate predictions. Here, we have given you the model (i.e., we have computed the best logistic regression coefficients) and will be asking you to adjust the threshold value for the best prediction.

Deliverables

Please submit a response document (e.g. a Word or Google Doc that is converted to a .pdf) with answers to each of the questions below. You do not need to submit the Excel document in which you do the work, and you do not need to paste the original question into your response document, and your answers need not be overly lengthy. For instance, when a question asks for a number, reporting the number is fine. When it asks for an explanation, one to three sentences are usually sufficient.

Some Excel functions are recommended in the text for those that are new to Excel, but you are welcome to use other methods. Assignments are to be submitted through Canvas. Please see Canvas for the due date.

Please answer the following questions.

Question 1.

What fraction of patients in the data provided have a known actual diagnosis of the diabetes condition? The Excel function *AVERAGE* may be useful here, which generates the average value of a column of numbers.

Question 2.

The spreadsheet allows you to adjust the threshold used to convert the logistic regression probabilities from the fitted model to predictions of whether a patient has diabetes. It is preset to 0.75. When using this **0.75** threshold, what fraction of 1/0 predictions is correct (i.e. matches the patient's actual condition)?

This number is known as "accuracy". To answer this question, you will first need to generate a new column which indicates whether the prediction matches the patient's condition. The "*IF"* function in Excel may be useful here.

Question 3.

Of course, this model does not always predict the correct answer. Assuming we chose the best modeling technique and fit the data appropriately, why might the model still produce incorrect predictions for some individuals?

Question 4.

Read here about false positives and false negatives. (These terms have now become part of the public vocabulary thanks to the pandemic and COVID-19 tests!)

- In language that might be relevant to a medical practitioner, what do false positives and negatives indicate for this diabetes diagnostic context?
- What fraction of predictions are "false negatives" when using the 0.75 threshold?
- What fraction of predictions are "false positives" when using the 0.75 threshold?
- In your view, should a medical practitioner prefer a prediction tool that minimizes false positives or false negatives? State the assumptions needed to justify your decision.

Question 5.

Does a threshold value of 0.75 maximize the accuracy of the prediction model? If not, specify a value (between 0 and 1) to three digits that maximizes this accuracy.

(Excel experts might use “Solver” to see if you can find a good value. If you choose this approach, use the “Evolutionary” Solving Method when setting up Solver. Alternatively, you can use trial-and-error to find this value by changing it by hand until you find which values produce the highest accuracy.)

Question 6.

A trained medical doctor could have provided you with a detailed explanation of how they use patient health indicators to diagnose diabetes. You have configured this prediction tool to predict diabetes outcomes, and you — presumably — have almost no medical knowledge. In fact, you have not even been told what data are stored in columns X1 - X8.

How is this possible? Medical experts learn by going to school. In what sense are these machine models “learning, and what are they learning from?

Question 7.

A second sheet in the spreadsheet called *TEST* contains data for additional patients. Unlike the data on the main sheet, these data were not used to build the model we are using. Apply our model to the data on the TEST sheet by pasting the coefficients in the same place in the new sheet and pasting or entering the same threshold value.

What is the **prediction accuracy** when applying your model (i.e. the same coefficients and threshold value) to data from this group?

Intuitively, should we expect the accuracy to be higher or lower than it was for the data on the main sheet? Justify your answer.

Question 8.

Now, place yourselves in the shoes of a healthcare provider (e.g. a nurse) who is considering using such a tool on patients.

a. From a patient outcomes perspective, provide an argument against using **accuracy** as the metric that the algorithm should be optimized on.

b. What is an alternative (single) metric that you might want to use here instead and why? Justify your decision. There is no single correct answer here, but your justification should match the metric you recommend. You could also consider creating a metric that combines several simpler metrics into a single measure.

Question 9.

It turns out that many major hospital systems — e.g., Penn Medicine, Cedar Sinai, and the Mayo Clinic — have invested heavily in building their own data science capabilities from the ground up, rather than relying on outside technology vendors, even though

technology vendors have highly sophisticated prediction tools that they are happy to provide for a price. In the context of your answers above, as well as HIPAA rules and any other information you might want to include, discuss why many hospitals may choose to “build” instead of “buy” prediction tools.

Question 10.

Now suppose you — again as a healthcare practitioner — built such a prediction tool and found it to be quite accurate. Your data science team suggests that you can run this tool on the data from all the patients that have been through the hospital over the last five years and proactively notify them of the likelihood they will contract diabetes within the next decade.

What are some ethical arguments for and against proactively using this model in this way? This is an open-ended question. Thoughtful and well-considered answers receive full credit.