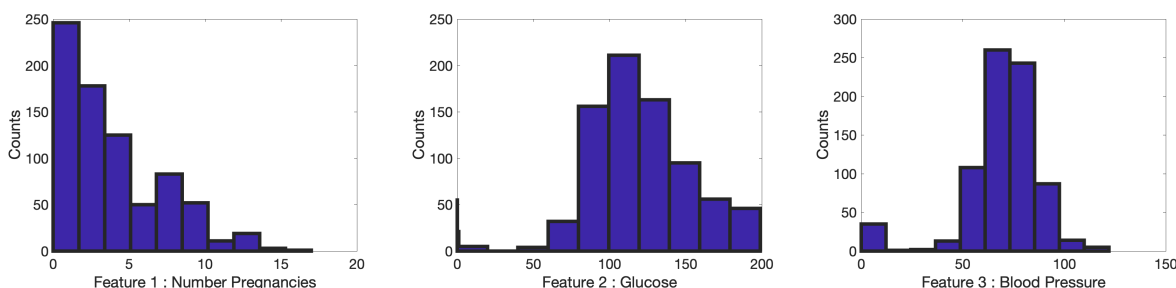


# Notes on Exploratory Data Analysis, Mostly ROC Curves

## Duke Course Notes

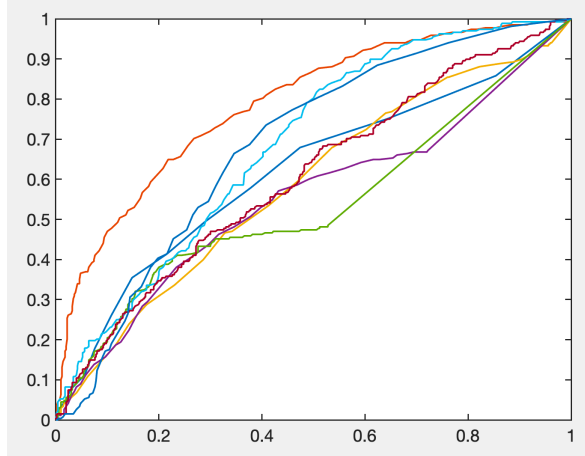
Cynthia Rudin

When I have a new dataset for a binary classification problem, two of the first things I do (after checking that there are no missing values and so on) is to plot histograms and then ROC curves for the individual features. Here are histograms for the first three features of the pima-indians-diabetes dataset. Histograms are useful for sanity checking, to ensure (for instance) that the points are not concentrated at one feature value, and that there are not outlier values that are nonsensical.



Luckily, these histograms look good.

Now I'll plot the ROC curve for each individual feature. In other words, I create a predictive model that is just one feature, and I look at its ROC curve with respect to the label; I do this for all features, and put them on the same plot. That is, if I issue a command to plot the ROC curve of  $f(\mathbf{x})$ , it is usually like this: `PlotROCCurve(y, f(x))`. If I want to plot the ROC curve for feature  $j$ , I would issue the command `PlotROCCurve(y, x.j)`, where  $\mathbf{x}_{.j}$  is the  $j$ th feature vector. Below is a plot of the ROC curves for each of the 8 features of the pima-indians-diabetes dataset from the UCI Machine Learning Repository.

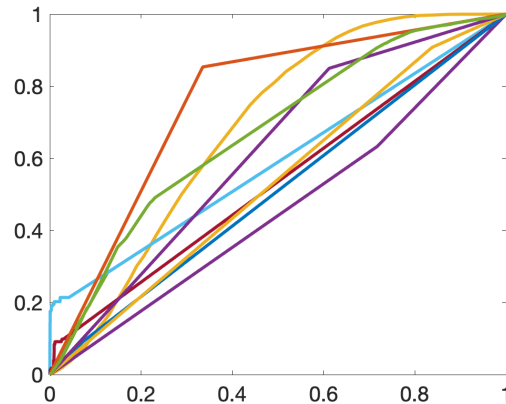


I now know a lot about this dataset based on this plot. I also can guess a lot about the model we will build from this dataset, even before we do any machine learning. Namely I can see that:

- One of the features dominates the others. That feature will be very useful in constructing the classification model. The model's ROC curve will be at least as good as that of the best feature (which happens to be the glucose feature for this dataset).
- Since the features have diverse ROC curves, the model will probably be much better than the best feature. This is because the features can each contribute to the model in a different way to improve it.
- You could also see that all of the features are better than random guessing at various points along the ROC curve; most datasets have some features that are not very good, so this dataset is unusual.
- The glucose feature is likely to be important to determine the high-scoring predictions. I can see that by looking at the left part of the ROC curve, where the glucose feature's curve has a steep upwards slope, indicating that individuals with the highest glucose levels tend to have label +1.
- I suspect that a couple of the features might be correlated since they have similar ROC curves (you could check that directly if you wanted to).
- In general, if you see diagonal lines on an ROC plot, it means there are probably a lot of tied scores. Here, some of the features start behaving as bad as random guessing towards the middle of the ROC curve. This is generally because the feature value is constant (probably 0) for many of the

data points. At that point, the feature is as bad as random guessing because its value is constant.

For comparison with the pima-indians-diabetes dataset, I plotted ROC curves for a few of the features for the adult dataset, which is derived from the 1994 census. The label is whether a person earned over \$50K in 1994 based on demographic and investment data (such as whether they had capital gains or losses).



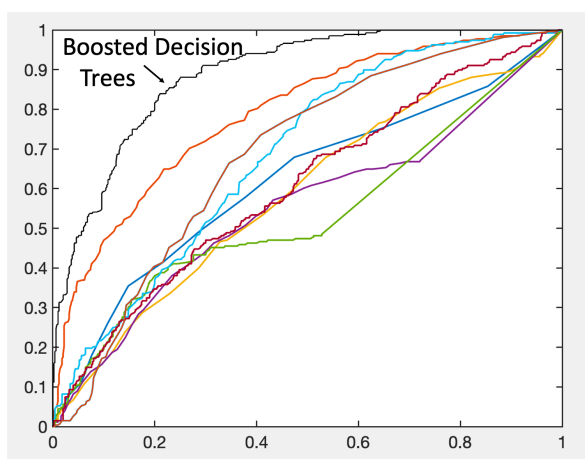
Here you will see that:

- One of the features (“workclass==Private”) has an ROC curve that is below the diagonal, which means that its negation will have an ROC curve above the diagonal and it will then perform better than random guessing.
- If a feature is binary, the ROC curve will look like two connected diagonal lines. For binary features, there are only three possible thresholds for classifiers on that feature value (predict all negative, predict 0 as negative and 1 as positive, and predict all as positive), so there are only 3 points on the ROC curve, and we connect them by diagonals.

Some of the interesting curves in this picture are age (the yellow curvy ROC curve), capital gains (the light blue curve that performs extremely well for large values, indicating that if someone has capital gains, they probably make a lot of money), and the feature “marital\_status==Married-civ-spouse” which has the very large AUC in red. I am not actually sure why that feature is so predictive, but now at least my exploratory data analysis leads to something I can investigate.

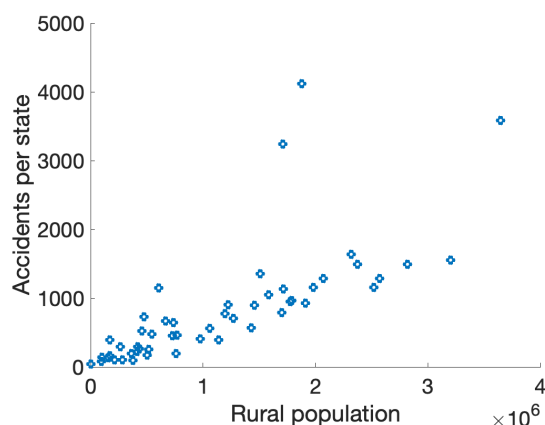
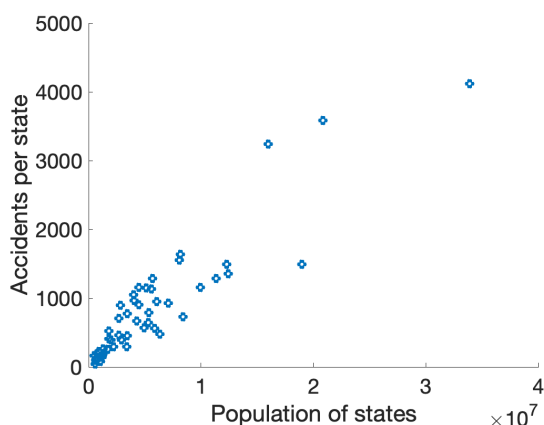
Hopefully by understanding the observations above, you will understand how powerful ROC curves are for exploratory data analysis!

Let's run boosted decision trees (one of the ML algorithms covered in the course) on the pima-indians-diabetes dataset to predict whether a patient has diabetes. I will then add its ROC curve to the plot.



The boosted models' ROC curve dominates that of the best feature. This model is a strong predictor of the outcome. (I might be concerned that I am overfitting since the predictions' ROC curve looks so good, so I might also use cross-validation to make sure that doesn't happen.)

If I have regression data, rather than classification data, it can be useful to scatter plot each feature against the label. For instance, in the highways dataset from Matlab, to predict the number of accidents per state, two of the features are the population of the state and the rural population of the state.



Both features are nicely correlated with the outcome, indicating that a regression model should perform well. However, there are three unusual points that have more accidents than their rural population would normally have. We could investigate these points to see if there's a reason.

There is a lecture later on dimension reduction for data visualization. That is also a useful tool for exploratory data analysis.

Thus, what I hope you get out of this lecture is that *if you know how to do exploratory data analysis, you can have some idea of what result you will get from machine learning before you do it*. You can also see and troubleshoot unusual datapoints or patterns before they are mixed up in a complicated machine learning model which would be more difficult to troubleshoot.