ROC & Logistic Regression Breakout

Galvanize DSI

Morning Objectives

- Understand the information contained in a confusion matrix
- Picking the right performance metric for the job
- Build a ROC curve by hand
- Pick a best model from a ROC plot

You have built a fraud prediction model

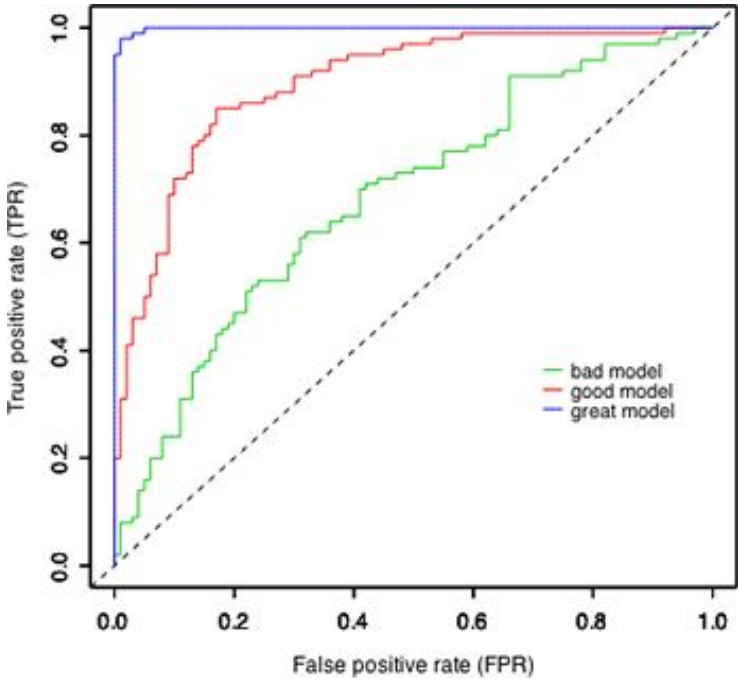
- Label each square with one of TP, FP, FN, TN.
- How many total data points do you have? How many are fraudulent? How many aren't fraudulent?
- Calculate accuracy, precision and recall.

	Predicted: Yes	**Predicted: No**
Actual: Yes	4	10
Actual: No	2	204

- Is the confusion matrix shown here representative of a good model?
- Which of the metrics you calculated above are most useful in determining how good the model is?
- What are cases where accuracy is useful? When do you need to be wary of using accuracy?

Assume we're dealing with predicting fraud...

1. In this scenario, do you think you'd care more about optimizing TPR or FPR?
2. What is a scenario where you'd care more about the other (TPR or FPR)?

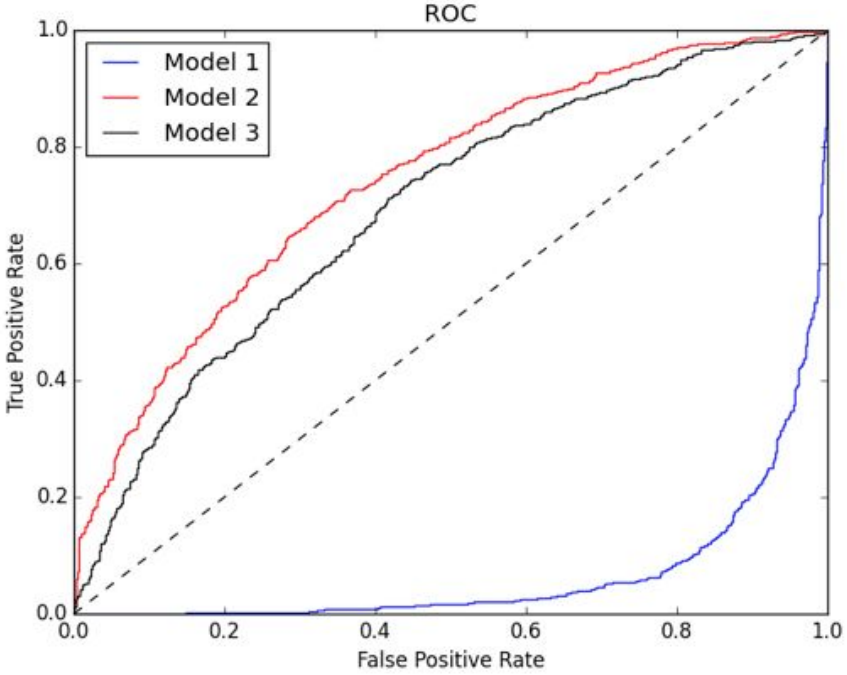


Construct a ROC curve only given the following predicted probabilities from a logistic regression and true labels

Predicted Probability	Actual customer churn?
0.99	Fraud
0.84	Fraud
0.70	Fraud
0.70	Not Fraud
0.51	Fraud
0.22	Fraud
0.14	Not Fraud
0.05	Not Fraud

Prompt: You have built 3 models to predict whether or not someone will default on a loan. You have 3000 data points and these features: age, gender, city, FICO score, highest education completed
Question: Which of the 3 ROC curves represents the model you should use?

Question: How would you pick between 50 models? 100 models? 1000 models?



Afternoon Breakout Objectives

- Understand what your chances are
- Discuss the benefits of the logit model
- Interpret logistic regression output

Understanding your chances

1. State what each of the following terms are:

   Probability, Odds, Log-Odds, Odds Ratio

2. Give an example to demonstrate what each of the 4 terms are

Why logistic and not just plain old linear?

1. What shape does the logistic function take?
2. Why is the logistic function a good, logical fit for binary classification?
3. Discuss the problems with using standard linear regression for modeling binary response.

Interpret the results from this logistic regression model

1. **What are my current chances of getting into grad school?**
2. **How would my chances change if I increased my GPA by 1.0 pts?**
3. **What score would I need on the GRE's to increase my chances to 95%?**

Logit Regression Results

Dep. Variable:	admit	No. Observations:	400
Model:	Logit	Df Residuals:	397
Method:	MLE	Df Model:	2
Date:	Fri, 02 Dec 2016	Pseudo R-squ.:	0.03927
Time:	16:43:29	Log-Likelihood:	-240.17
converged:	True	LL-Null:	-249.99
		LLR p-value:	5.456e-05

	coef	std err	z	P>|z|	[95.0% Conf.	Int.]
const	-4.9494	1.075	-4.604	0.000	-7.057	-2.842
gre	0.0027	0.001	2.544	0.011	0.001	0.005
gpa	0.7547	0.320	2.361	0.018	0.128	1.381

Model 1 and 2 are from the same dataset. Explain what you see.

Dep. Variable:	Survived	**No. Observations:**	712
Model:	Logit	**Df Residuals:**	709
Method:	MLE	**Df Model:**	2
Date:	Tue, 22 Nov 2016	**Pseudo R-squ.:**	0.2528
Time:	15:27:35	**Log-Likelihood:**	-359.02
converged:	True	**LL-Null:**	-480.45
		LLR p-value:	1.825e-53

	coef	std err	z	P>|z|	[95.0% Conf. Int.]
Intercept	0.6590	0.167	3.935	0.000	0.331 0.987
Sex[T.male]	-2.3711	0.189	-12.524	0.000	-2.742 -2.000
Fare	0.0121	0.003	4.595	0.000	0.007 0.017

Model 1

Dep. Variable:	Survived	**No. Observations:**	712
Model:	Logit	**Df Residuals:**	708
Method:	MLE	**Df Model:**	3
Date:	Tue, 06 Dec 2016	**Pseudo R-squ.:**	0.3013
Time:	08:33:07	**Log-Likelihood:**	-335.70
converged:	True	**LL-Null:**	-480.45
		LLR p-value:	1.852e-62

	coef	std err	z	P>|z|	[95.0% Conf. Int.]
Intercept	3.1335	0.399	7.863	0.000	2.352 3.915
Sex[T.male]	-2.5536	0.204	-12.528	0.000	-2.953 -2.154
Fare	0.0019	0.002	0.850	0.395	-0.002 0.006
Pclass	-0.9283	0.137	-6.788	0.000	-1.196 -0.660

Model 2