Additional Workshops (2 x 2hrs)

Final Project infos

REVIEW TOPICS

- Distances and how to apply them in the KNN algorithm
- Difference between the linear and logistic regression:
  - The output
  - Loss function (also called Cost function) v. Linear Reg.
  - Evaluation of your classification model v. Linear Reg.

Today's Plan

- Confusion matrix
- Threshold
- What's an ROC curve
- What do precision and recall mean? How are they similar and different to True Positive Rate and False Positive Rate?
- Communication of your results

ADVANCED CLASSIFICATION METRICS

ADVANCED CLASSIFICATION METRICS

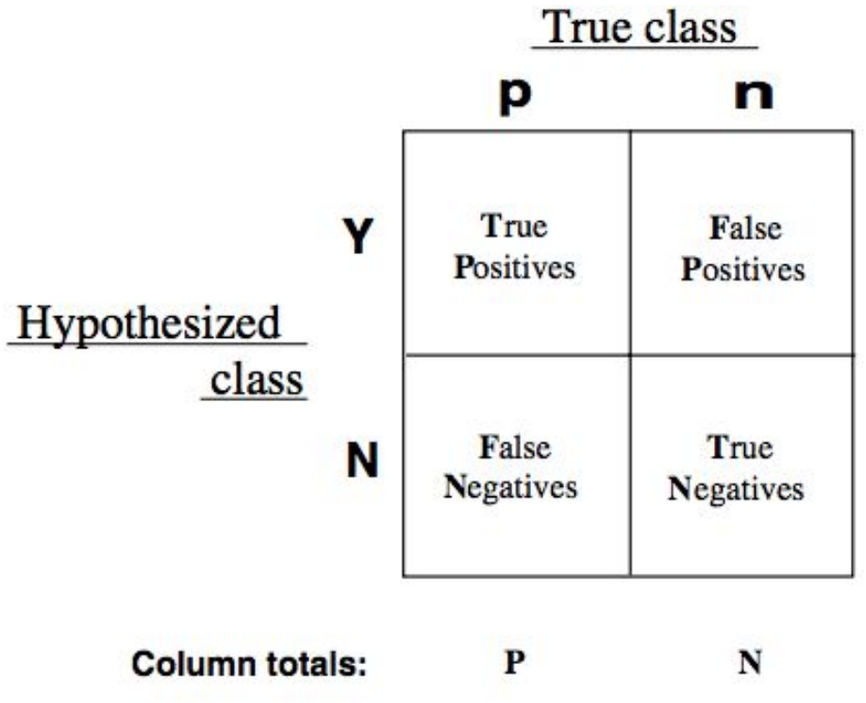
- Accuracy is only one of several metrics used when solving a classification problem.
- Accuracy = total predicted correct / total observations in dataset
- Accuracy alone doesn't always give us a full picture.
- If we know a model is 75% accurate, it doesn't provide *any* insight into why the 25% was wrong.

ADVANCED CLASSIFICATION METRICS

- Was it wrong across all labels?
- Did it just guess one class label for all predictions?
- It’s important to look at other metrics to fully understand the problem.

ADVANCED CLASSIFICATION METRICS

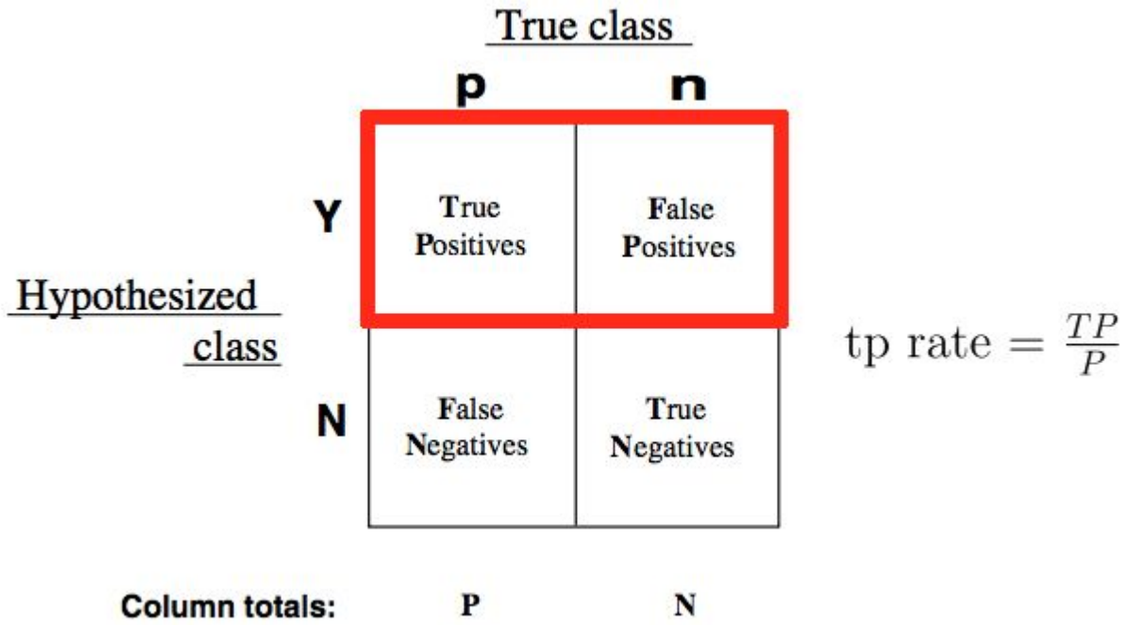
- Accuracy of each label formula with the *true positive rate* and the *false positive rate*.
- For each label, we can put it into the category of a true positive, false positive, true negative, or false negative.



ADVANCED CLASSIFICATION METRICS

- True Positive Rate (TPR) asks, “Out of all of the target class labels, how many were accurately predicted to belong to that class?”

- For example, given a medical exam that tests for cancer, how often does it correctly identify patients with cancer?

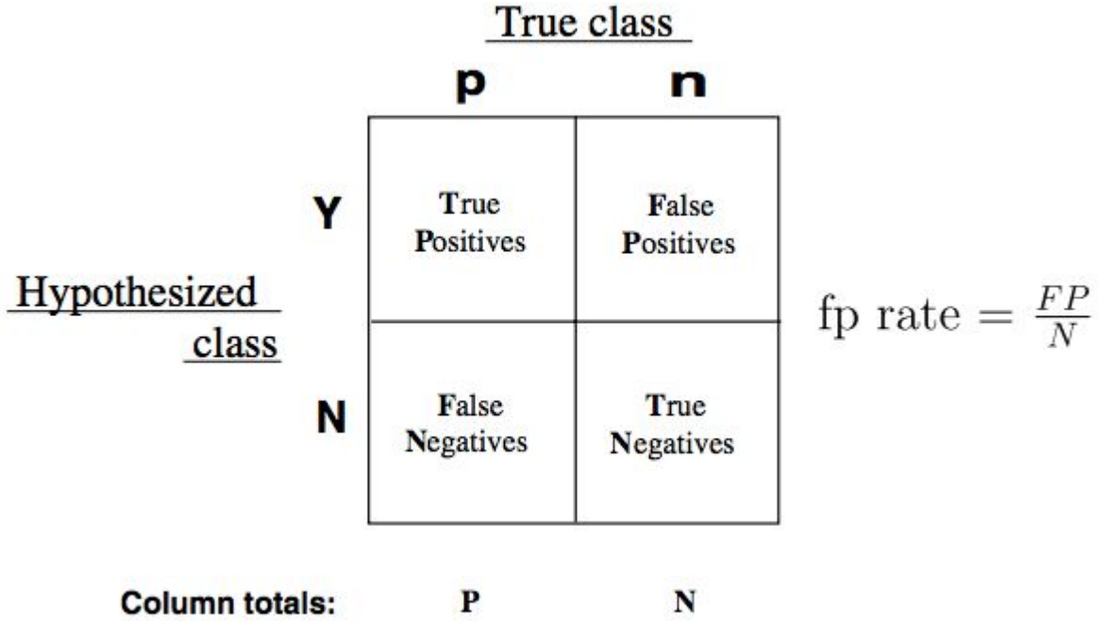


$$\text{tp rate} = \frac{TP}{P}$$

ADVANCED CLASSIFICATION METRICS

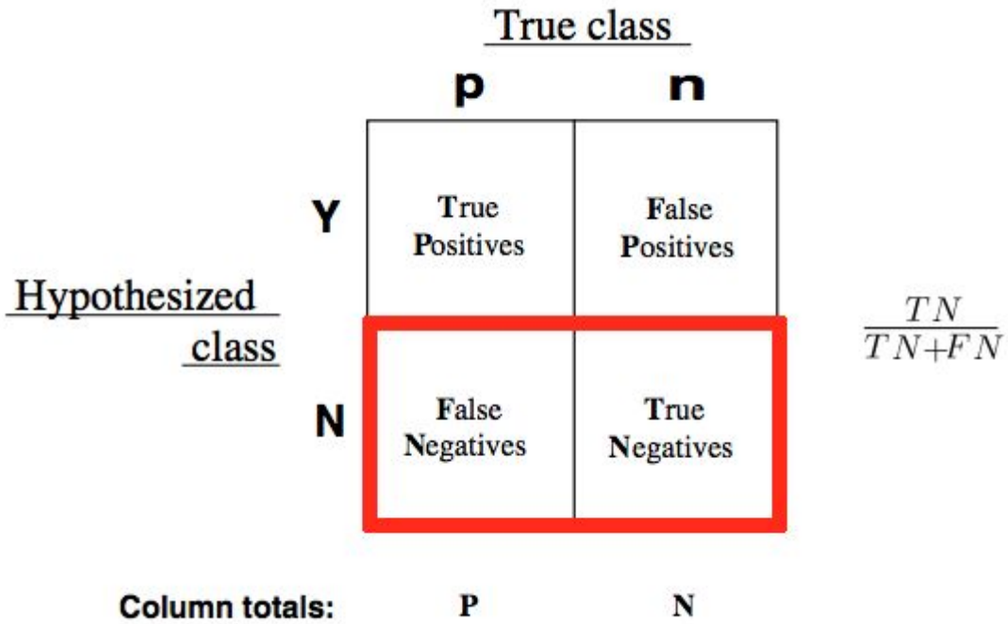
- False Positive Rate (FPR) asks, “Out of all items not belonging to a class label, how many were predicted as belonging to that target class label?”

- For example, given a medical exam that tests for cancer, how often does it trigger a “false alarm” by incorrectly saying a patient has cancer?



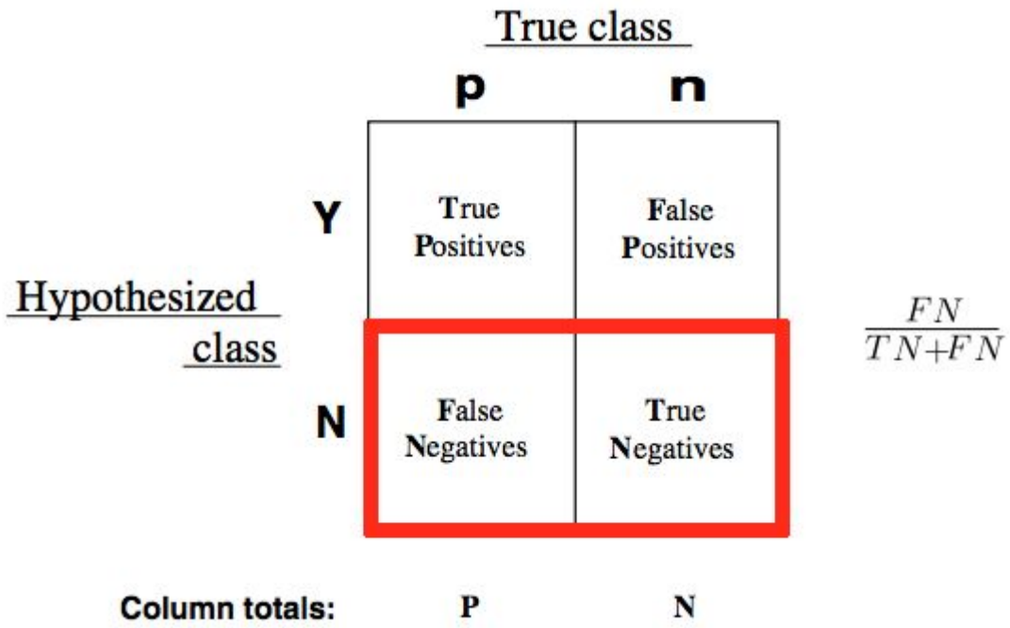
ADVANCED CLASSIFICATION METRICS

- These can also be inverted.
- How often does a test *correctly* identify patients without cancer?



ADVANCED CLASSIFICATION METRICS

- How often does a test *incorrectly* identify patient as cancer-free?



ADVANCED CLASSIFICATION METRICS

- The true positive and false positive rates gives us a much clearer pictures of where predictions begin to fall apart.

- This allows us to adjust our models accordingly.

ADVANCED CLASSIFICATION METRICS

- A good classifier would have a **true positive rate approaching** 1 and a **false positive rate** approaching 0.

- In our smoking problem, this model would accurately predict ***all* of the smokers as smokers** and **not accidentally predict any of the nonsmokers as smokers**.

INDEPENDENT PRACTICE

CODE EXERCISE: Humor-classification Part-1

ADVANCED CLASSIFICATION METRICS

- Go through and reply to questions up to 6 (included):
- http://localhost:8888/notebooks/GA/DS_PartTime/DS-course-materials/lessons/lesson-10/code/week04/day2_logistic_regression_and_classification_metrics/lab-2-evaluation-classifiers_confusion_matrix_roc/humor-classification.ipynb

INDEPENDENT PRACTICE

THRESHOLDS in CLASSIFIERS

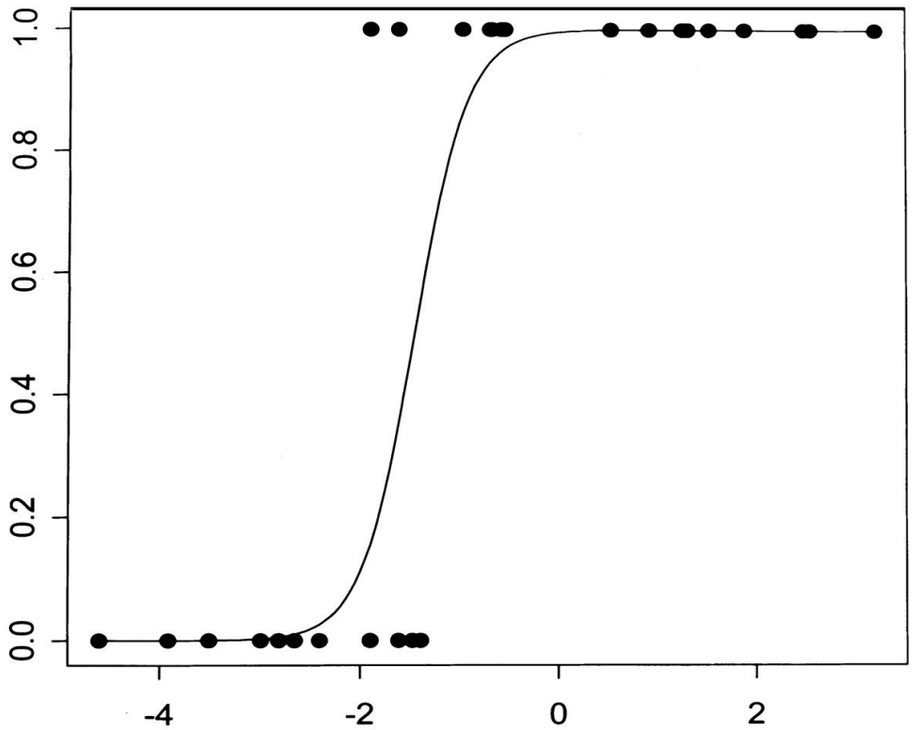
ADVANCED CLASSIFICATION METRICS

- Two systems have the following performance:
  - A: True Positive = 50%, False Positive = 20%
  - B: True Positive = 100%, False Positive = 60%

> Which one is better?

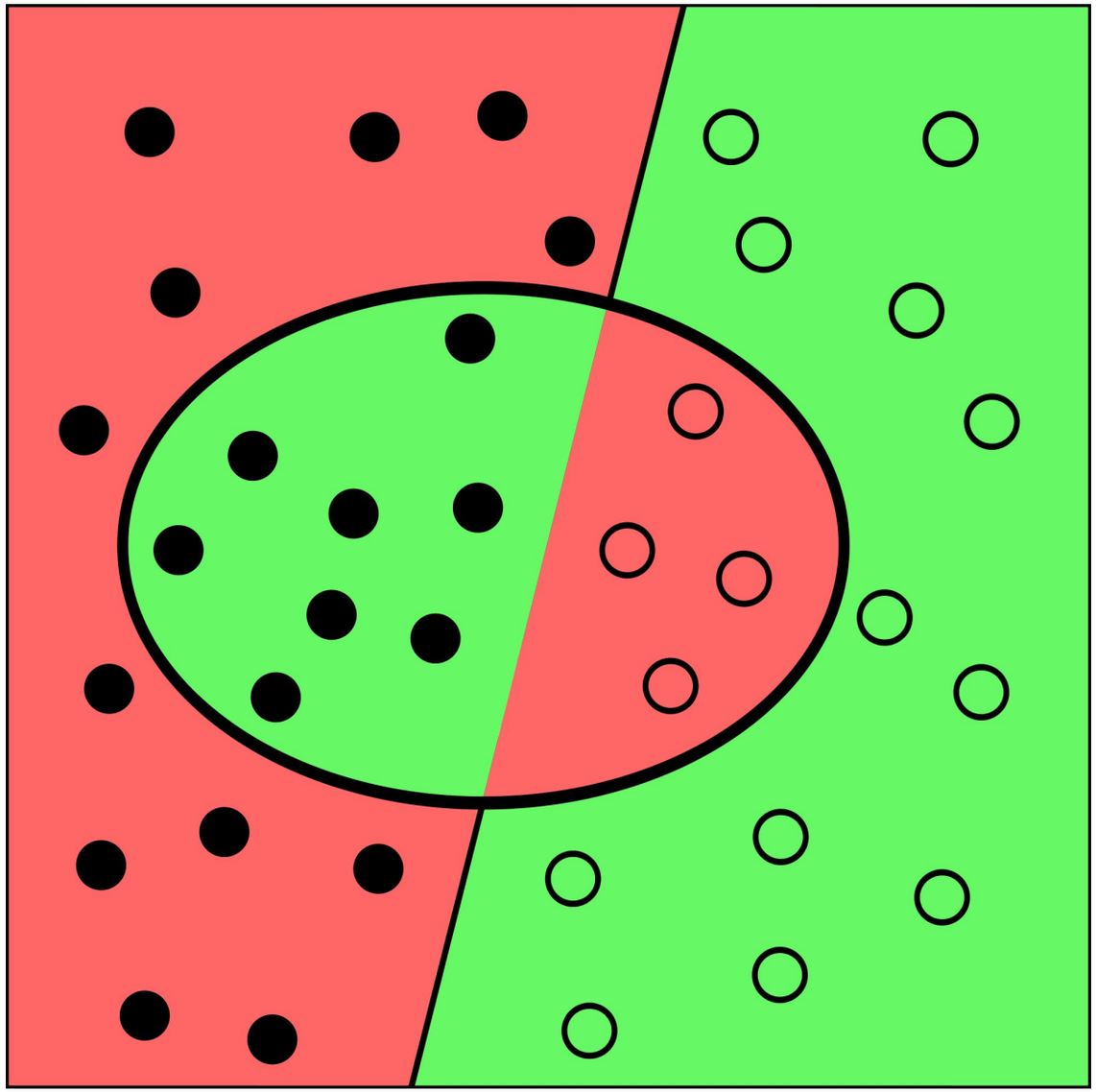
ADVANCED CLASSIFICATION METRICS

- Comes in the **Thresholds**:
  - Classifier compute the probability of new observation
  - Then compare it to a threshold



ANOTHER EXAMPLE

- The background shows the color predicted.
- A shaded circle on a green background represents a green marble that was predicted as green.
- An unshaded circle on a red background represents a red marble that was predicted as green.

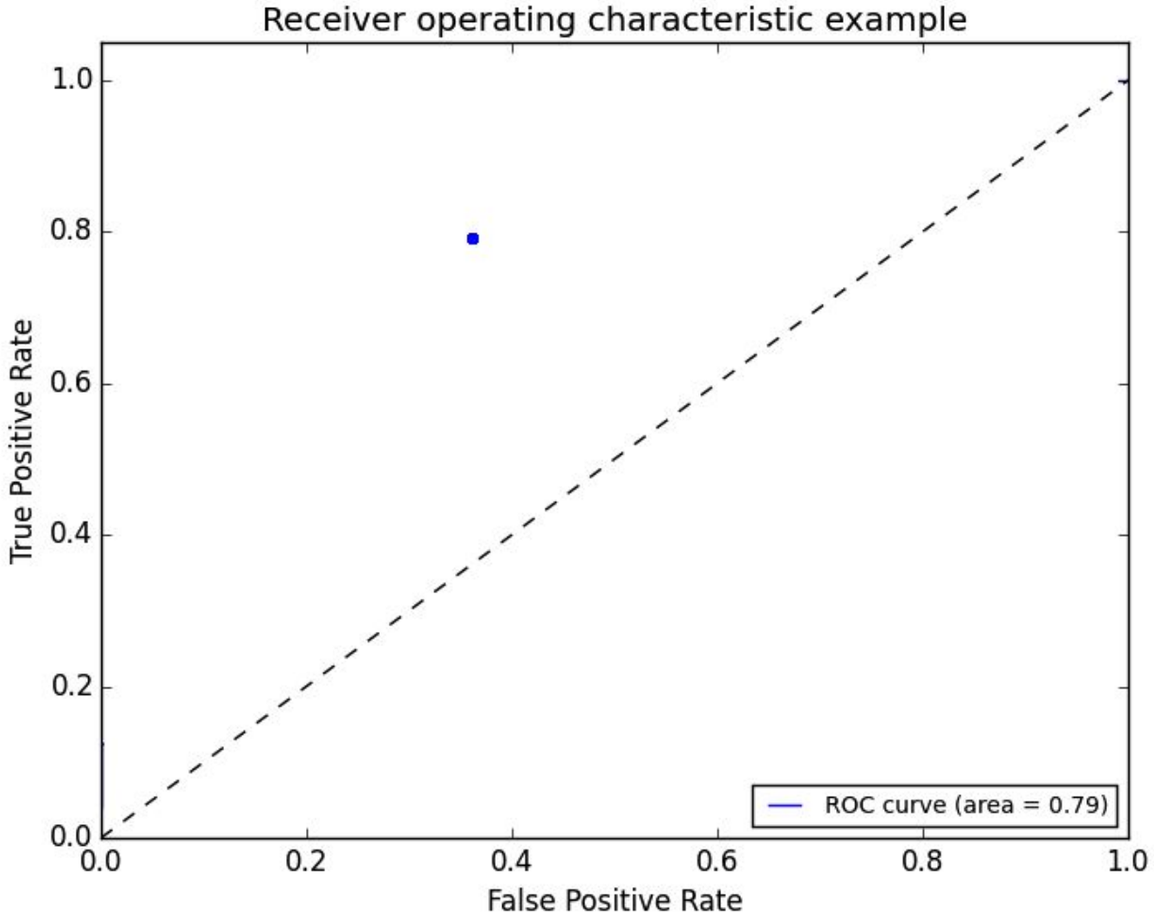


ADVANCED CLASSIFICATION METRICS

- Receiver Operation Characteristic (ROC) curve help optimizing 2 parameters
- The curve is created by plotting the true positive rate against the false positive rate at various model threshold settings.
- Area Under the Curve (AUC) summarizes the impact of TPR and FPR in one single value.

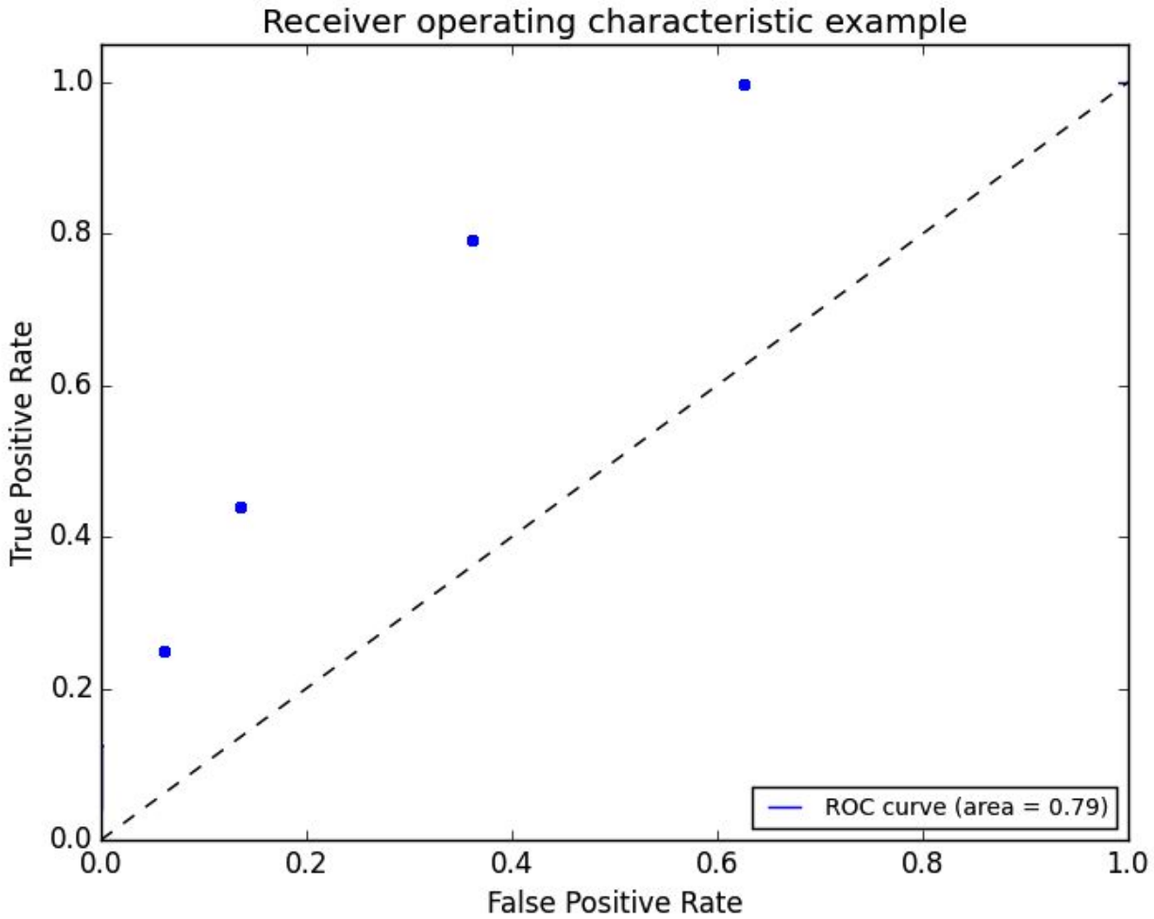
ADVANCED CLASSIFICATION METRICS

- We can begin by plotting an individual TPR/FPR pair for one threshold.



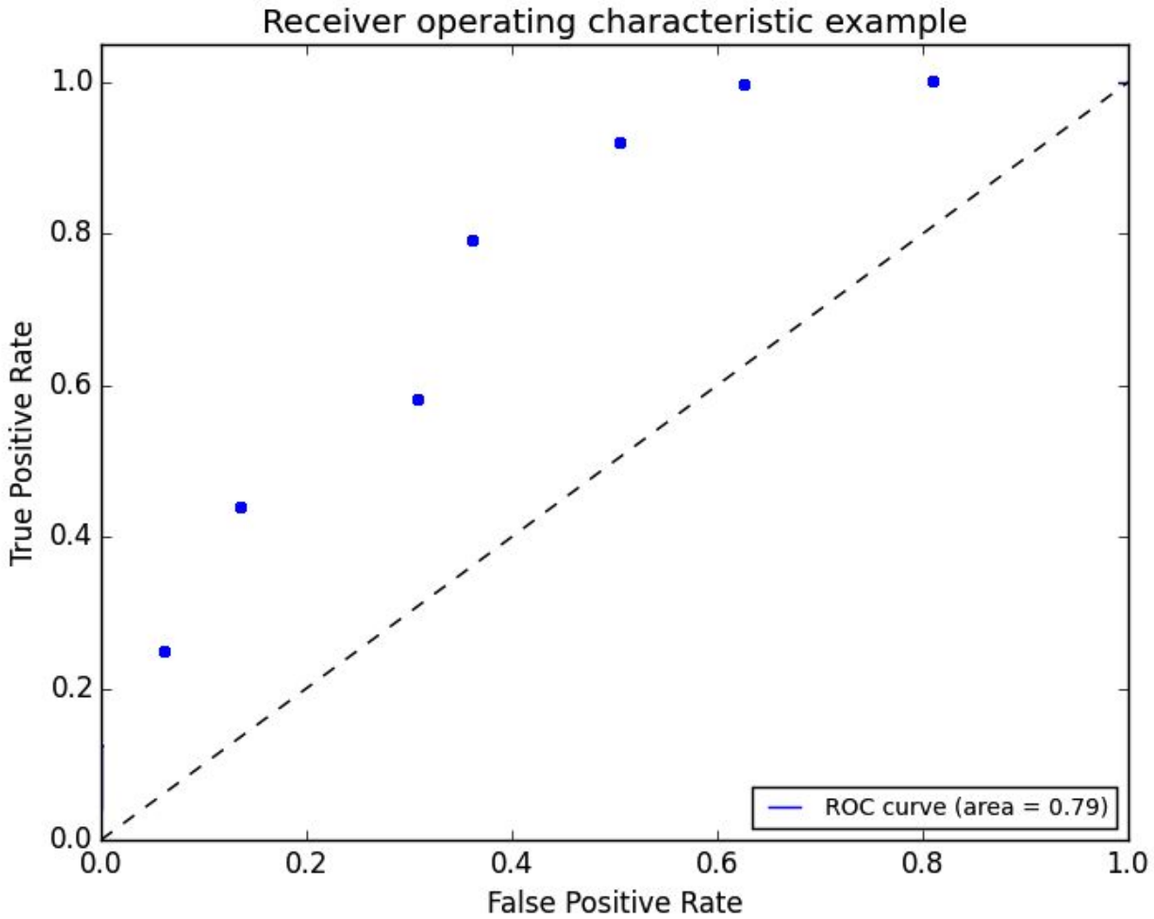
ADVANCED CLASSIFICATION METRICS

- We can continue adding pairs for different thresholds



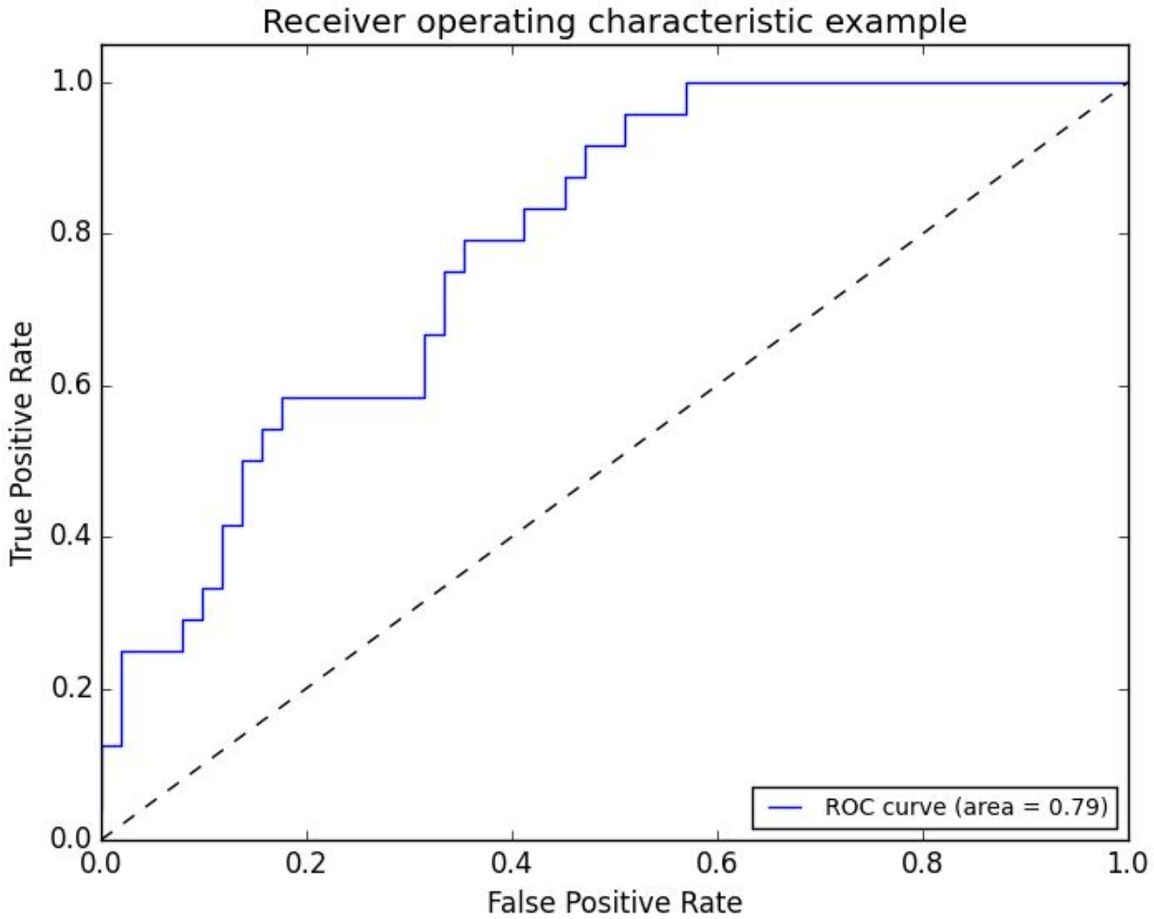
ADVANCED CLASSIFICATION METRICS

- We can continue adding pairs for different thresholds



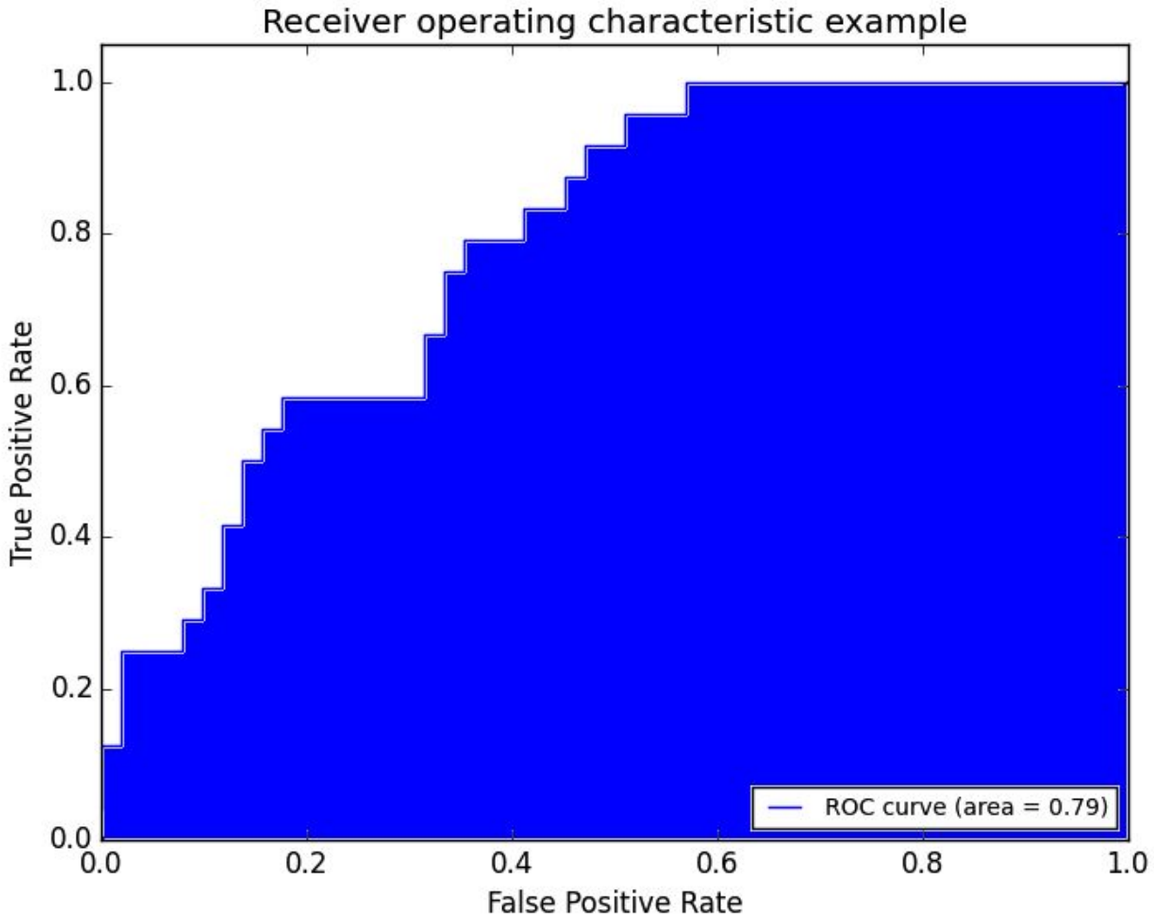
ADVANCED CLASSIFICATION METRICS

- Finally, we create a full curve that is described by TPR and FPR.



ADVANCED CLASSIFICATION METRICS

- With this curve, we can find the Area Under the Curve (AUC).



ADVANCED CLASSIFICATION METRICS

- Breakout: Play with this interactive visualization
- Reply to those questions:
  - What does the red curve represent
  - Why is the ROC moving
  - What is the curve of a perfect performance?
  - What is the curve of a classifier using a coin for prediction?
  - Where should we set the threshold for fraud detection?

ADVANCED CLASSIFICATION METRICS

- There are several other common metrics that are similar to TPR and FPR.

		True class: p	True class: n
Hypothesized class	Y	True Positives	False Positives
	N	False Negatives	True Negatives
Column totals:		P	N

$$\text{fp rate} = \frac{FP}{N} \qquad \text{tp rate} = \frac{TP}{P}$$

$$\text{precision} = \frac{TP}{TP+FP} \qquad \text{recall} = \frac{TP}{P}$$

$$\text{accuracy} = \frac{TP+TN}{P+N}$$

$$\text{F-measure} = \frac{2}{1/\text{precision}+1/\text{recall}}$$

- Sklearn has all of the metrics located on one convenient page.

ADVANCED CLASSIFICATION METRICS: Code Along

> Start above Titanic (ROC):
http://localhost:8888/notebooks/GA/DS_PartTime/DS-course-materials/lessons/lesson-09/solution-code/solution-code-9.ipynb

INDEPENDENT PRACTICE

CODE EXERCISE: Humor-classification Part-2

ADVANCED CLASSIFICATION METRICS

- Go through and reply to the other questions:
- http://localhost:8888/notebooks/GA/DS_PartTime/DS-course-materials/lessons/lesson-10/code/week04/day2_logistic_regression_and_classification_metrics/lab-2-evaluation-classifiers_confusion_matrix_roc/humor-classification.ipynb

OPENING

COMMUNICATING RESULT

WE BUILT A MODEL! NOW WHAT?

- We’ve built our model, but there is still a gap between your **Notebook with plots/figures and a slideshow** needed to present your results.

- Classes so far have focused on two core concepts:
  - developing consistent practices
  - interpreting metrics to evaluate and improve model performance

- But what does that mean to your audience?

WE BUILT A MODEL! NOW WHAT?

- Imagine how a non-technical audience might respond to the following statements:
  - The predictive model I built has an accuracy of 80%.
  - Logistic regression was optimized with L2 regularization.
  - Gender was more important than age in the predictive model because it has a larger coefficient.
  - Here's the AUC chart that shows how well the model did.

WE BUILT A MODEL! NOW WHAT?

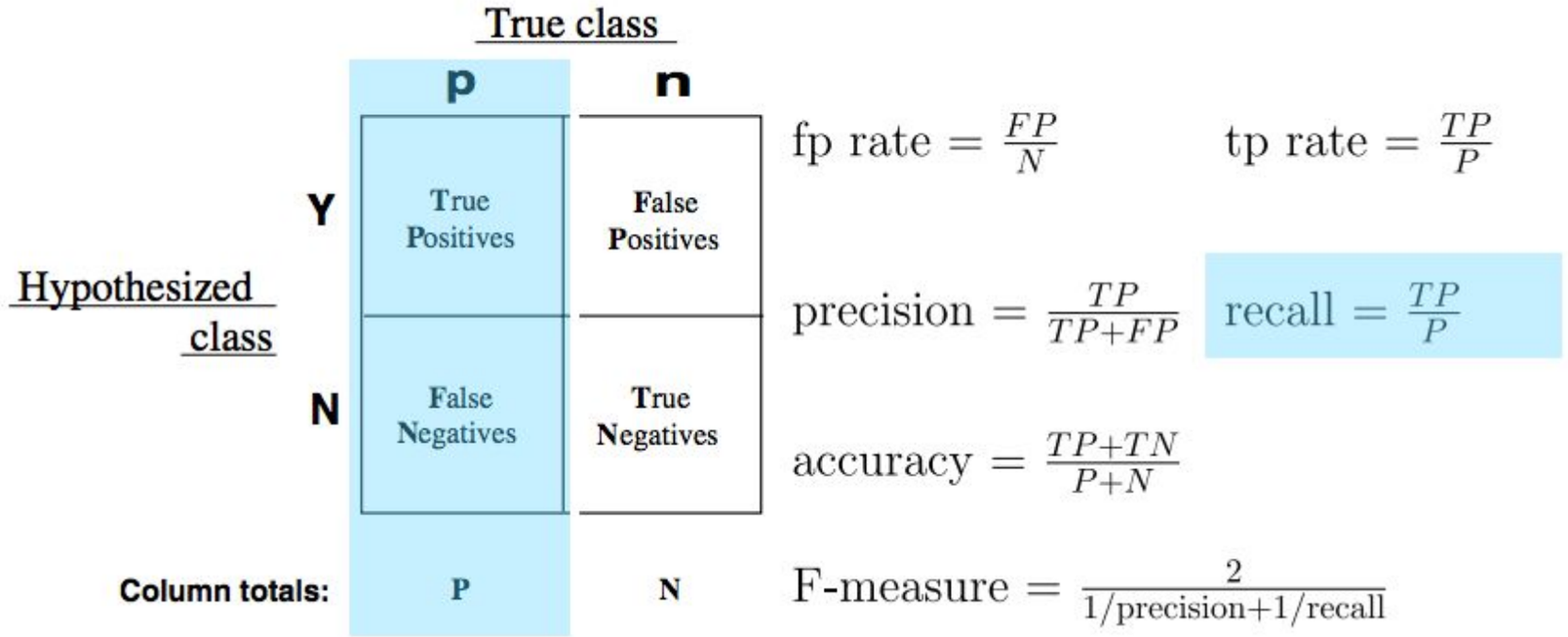
- Who is your audience? Are they technical? What are their concerns?
- You may be *the only person* who can interpret what you've built.
- You will likely have to do a lot of "hand holding".
- You need to be able to efficiently explain your results in a way that makes sense to **all** stakeholders (technical or not).

PRECISION AND RECALL

THE MATH FOR RECALL

- Recall is the count of predicted *true positives* over the total count of that class label.

- This is the same as True Positive Rate or *sensitivity*.

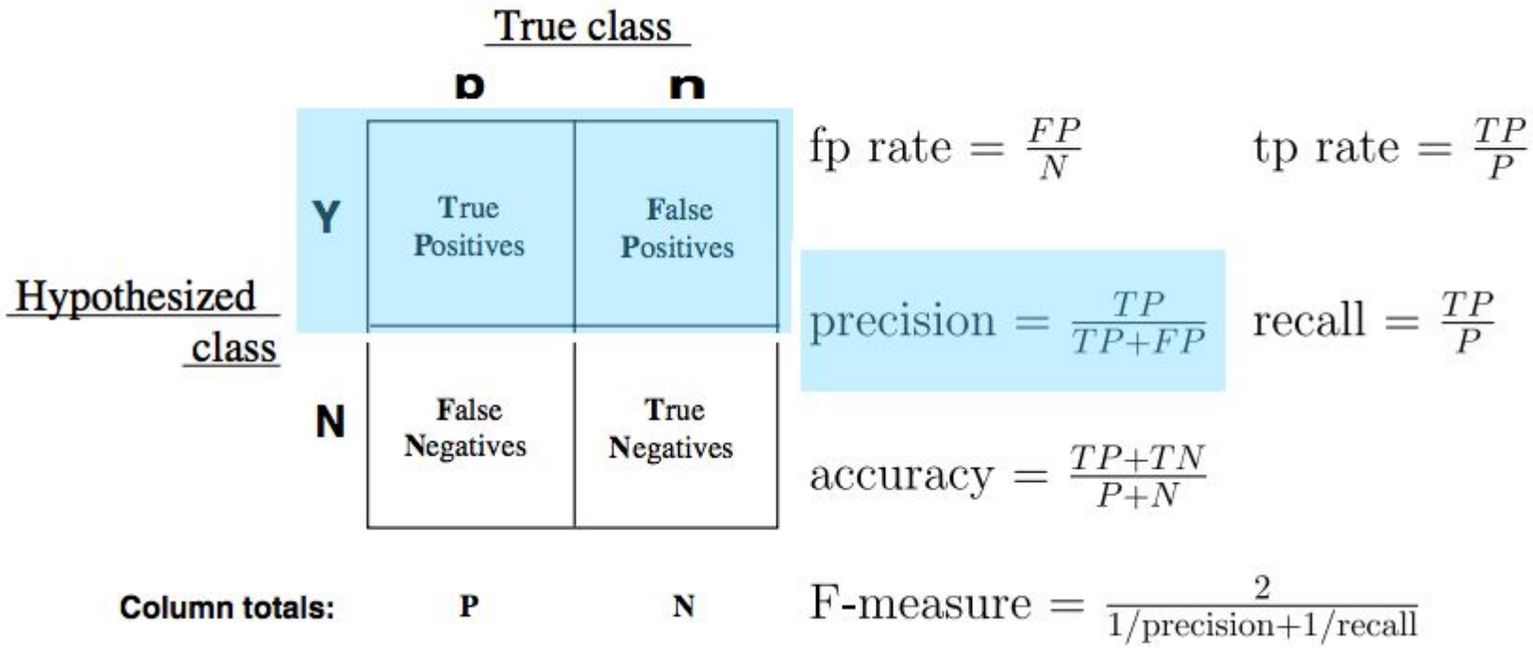


THE MATH FOR RECALL

- Imagine predicting the color of a marble as either red or green. There are 10 of each.
- If the model identifies 8 identifies 8 of the green marbles as green, the recall is 8 / 10 = 0.80.
- However, this says nothing of the number of *red* marbles that are also identified as green.

THE MATH FOR PRECISION

- Precision, or positive predicted value, is calculated as the count of predicted true positives over the count of all values predicted to be positive.



THE MATH FOR PRECISION

- Let’s use our marble example again.
- If a model predicts 8 of the green marbles as green, then precision would be 1.00, because all marbles predicted as green were in fact green.
- Let’s assume all red marbles were predicted correctly, and 2 green were predicted as red.
- The precision of red marbles would be 10 / (10 + 2) = 0.833.

ANOTHER EXAMPLE

- For this example, we would have the following confusion matrix.

		True Class	
		Green	Red
Predicted Class	Green	8	4
	Red	12	12

- We could calculate precision for green marbles as 8 / (8 + 4) = 0.6666.

- We could calculate recall for green marbles as 8 / (8 + 12) = 0.4000.

GUIDED PRACTICE

COST BENEFIT ANALYSIS

ACTIVITY: COST BENEFIT ANALYSIS

DIRECTIONS (15 minutes)

One tool that complements the confusion matrix is cost-benefit analysis, where you attach a *value* to correctly and incorrectly predicted data.

Like the Precision-Recall trade off, there is a balancing point to the *probabilities* of a given position in the confusion matrix, and the *cost* or *benefit* to that position. This approach allows you to not only add a weighting system to your confusion matrix, but also to speak the language of your business stakeholders (i.e. communicate your values in dollars!).

ACTIVITY: COST BENEFIT ANALYSIS

EXERCISE

DIRECTIONS

Consider the following marketing problem:

As a data scientist working on marketing spend, you've build a model that reduces user churn--the number of users who decide to stop paying for a product--through a marketing campaign.

Your model generates a confusion matrix with the following probabilities (these probabilities are calculated as the value in that position over the sum of the sample):

```
| TP: 0.2 | FP: 0.2 |
---------------------
| FN: 0.1 | TN: 0.5 |
```

ACTIVITY: COST BENEFIT ANALYSIS

DIRECTIONS (15 minutes)

In this case:

- The *benefit* of a true positive is the retention of a user ($10 for the month)
- The *cost* of a false positive is the spend of the campaign per user ($0.05)
- The *cost* of a false negative (someone who could have retained if sent the campaign) is, effectively, 0 (we didn't send it... but we certainly didn't benefit!)
- The *benefit* of a true negative is 0: No spend on users who would have never retained.

To calculate Cost-Benefit, we'll use this following function:

```
(P(TP) * B(TP)) + (P(TN) * B(TN)) + (P(FP) * C(FP)) + (C(FN) * C(FN))
```

which for our marketing problem, comes out to this:

```
(.2 * 10) + (.5 * 0) - (.2 * .05) - (.1 * 0)
```

or $1.99 per user targeted.

ACTIVITY: COST BENEFIT ANALYSIS

FOLLOW UP QUESTIONS

Think about precision, recall, and cost benefit analysis to answer the following questions:

1. How would you rephrase the business problem if your model was optimizing toward *precision*? i.e., How might the model behave differently, and what effect would it have?

2. How would you rephrase the business problem if your model was optimizing toward *recall*?

3. What would the most ideal model look like in this case?

DELIVERABLE

Answers to the above questions

DEMO

VISUALIZING MODELS OVER VARIABLES

VISUALIZING MODELS OVER VARIABLES

- One effective way to **explain your model over particular variables** is to plot the predicted values against the **most explanatory variables**.

- For example, in logistic regression, plotting **the probability of a class against a variable** can help explain the range of effect of the model.

VISUALIZING MODELS OVER VARIABLES

- We'll use the flight delay data for all following examples. Let's build our first model and plot.

- Open the starter code from the class repo and follow along.

http://localhost:8888/notebooks/GA/DS_PartTime/DS-course-materials/lessons/lesson-10/code/starter-code/starter-code-10.ipynb

VISUALIZING MODELS OVER VARIABLES

```
# read in the file and generate a quick model (assume we've done the data
exploration already)
import pandas as pd
import sklearn.linear_model as lm
import matplotlib.pyplot as plt

df = pd.read_csv('../../assets/dataset/flight_delays.csv')

df = df.join(pd.get_dummies(df['DAY_OF_WEEK'], prefix='dow'))
df = df[df.DEP_DEL15.notnull()].copy()
```

VISUALIZING MODELS OVER VARIABLES

```
# Build a model
model = lm.LogisticRegression()
features = ['dow_1', 'dow_2', 'dow_3', 'dow_4', 'dow_5', 'dow_6']
model.fit(df[features + ['CRS_DEP_TIME']], df['DEP_DEL15'])

df['probability'] = model.predict_proba(df[features + ['CRS_DEP_TIME']]).T[1]
```

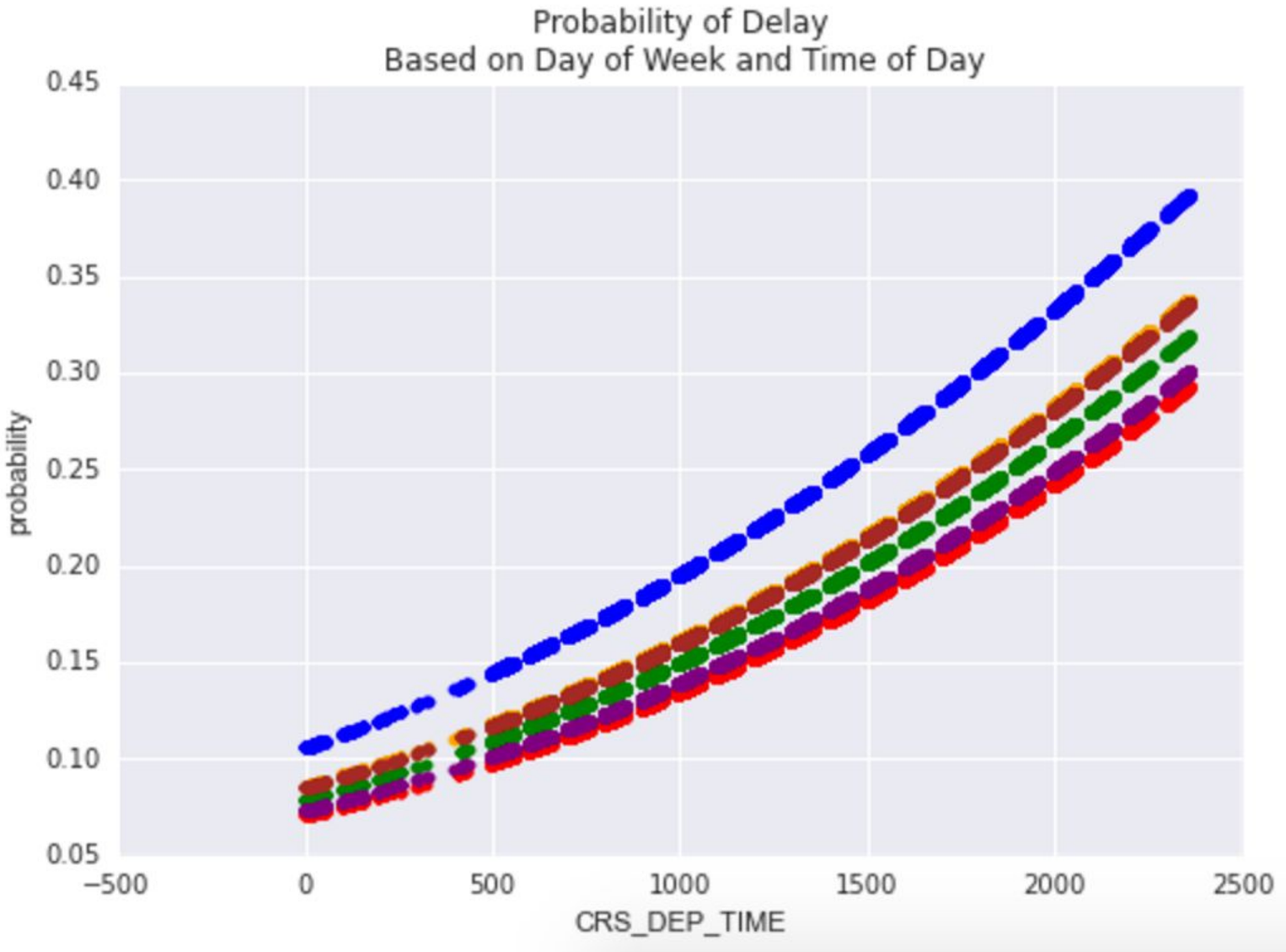
VISUALIZING MODELS OVER VARIABLES

```
# Create a plot
ax = plt.subplot(111)
colors = ['blue', 'green', 'red', 'purple', 'orange', 'brown']
for e, c in enumerate(colors):
    df[df[features[e]] == 1].plot(x='CRS_DEP_TIME', y='probability',
kind='scatter', color = c, ax=ax)

ax.set(title='Probability of Delay\n Based on Day of Week and Time of Day')
```

VISUALIZING MODELS OVER VARIABLES

- This visual can help showcase the range of effect on delays from both day of the week and time of day.
- Given this model, some days are more likely to have delays than others.
- The likelihood of delay increases as the day goes on.



ACTIVITY: TRY IT OUT

DIRECTIONS

1. Adjust the model to make delay predictions using airlines instead of day of week, and time, then plot the effect on CRS_DEP_TIME=1.

2. Try plotting the inverse: pick either model and plot the effect on CRS_DEP_TIME=0.

DELIVERABLE

The new plots

TOPIC REVIEW

REVIEW AND NEXT STEPS

- Confusion matrix
- Threshold
- What's an ROC curve
- What do precision and recall mean? How are they similar and different to True Positive Rate and False Positive Rate?
- Communication of your results

BEFORE NEXT CLASS

BEFORE NEXT CLASS

UPCOMING

- Project: Unit Project 3

LESSON

Q & A