# EXIT TICKETS

- 1. how to apply knn.predict to recommendation engines like Amazon's products based on what I have purchased, Linkedin jobs applied, in terms of making modifications to that code?
  2. at work, how does the accuracy variables are used when there are quite a few tests?
- In logistic regression, is it worth running something so complicated only to eventually arrive at a binary outcome?
- Is there a unified sheet or site where we can view the code related to the tool and explanation of the tool? E.g. LogisticRegression for SKLearn compared to others and explanations?
- It would be good to demonstrate the difference in the outcome that we get when we use (non-linear) linear regression versus using the logistic regression function on jupyter notebook.
- The different methods and syntax in implementing the classification and logistic regression models in Python

GENERAL ASSEMBLY

# COMMUNICATING RESULTS

*Tan Kwan Chong*

*Chief Data Scientist, Booz Allen Hamilton*

# LEARNING OBJECTIVES

- Explain the trade-offs between the precision and recall of a model while articulating the cost of false positives vs. false negatives
- Describe the difference between visualization for presentations vs. exploratory data analysis
- Identify the components of a concise, convincing report and how they relate to specific audiences/stakeholders

OPENING

# COMMUNICATING RESULTS

# 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?
- Remember: in a business setting, you may be *the only person* who can interpret what you've built.
- Some people may be familiar with basic visualization, but 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).

# WE BUILT A MODEL!  NOW WHAT?

- Today, we’ll focus on communicating results for “simpler” problems, but this applies to any type of model you may work with.

- First, let’s review classification metrics, review our knowledge, and talk about how we might communicate what we know.

REVIEW

# BACK TO THE CONFUSION MATRIX

# BACK TO THE CONFUSION MATRIX

- Confusion matrices allow for the interpretation of correct and incorrect predictions for *each class label*.

- It is the first step for the majority of classification metrics and goes deeper than just accuracy.

# BACK TO THE CONFUSION MATRIX

- Let's review our confusion matrix.

| | Prediction Positive | Prediction Negative |
|---|---|---|
| Condition Positive (P) | True Positive (TP) | False Negative (FN) |
| Condition Negative (N) | False Positive (FP) | True Negative (TN) |

**True Positive Rate aka Sensitivity, Recall**

$$TPR = \frac{TP}{P} = \frac{TP}{(TP + FN)}$$

**False Positive Rate aka Specificity**

$$FPR = \frac{FP}{N} = \frac{FP}{(FP + TN)}$$

**Positive Prediction Value aka Precision**

$$PPV = \frac{TP}{(TP + FP)}$$

**Accuracy**

$$A = \frac{TP + TN}{P + N}$$

# ACTIVITY: KNOWLEDGE CHECK

EXERCISE

## ANSWER THE FOLLOWING QUESTIONS

1. Without looking at the previous slide, how do we calculate the following?
    a. Accuracy
    b. True positive rate
    c. False positive rate

## DELIVERABLE

Answers to the above questions

# PRECISION AND RECALL

# PRECISION AND RECALL

- Our initial metrics (accuracy & misclassification rates) were primarily designed for less biased data problems:  we could be interested in both outcomes, so it was important to generalize our approach.

- For example, we may be interested if a person will vote for a Republican or Democrat.  This is a binary problem, but we're interested in both outcomes.

# PRECISION AND RECALL

- Precision and recall, metrics built from the confusion matrix, focus on *information retrieval*, particularly when one class is more interesting than the other.
- For example, we may want to predict if a person will be a customer. We care much more about people who will be a customer of ours than people who won’t.

# PRECISION AND RECALL

- *Precision* aims to produce a high amount of relevancy instead of irrelevancy.
- Precision asks, "Out of all of our positive predictions (both true positive and false positive), how many were correct?"
- *Recall* aims to see how well a model returns specific data (literally, checking whether the model can *recall* what a class label looked like).
- Recall asks, "Out of all of our positive class labels, how many were identified?"

# THE MATH FOR RECALL

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

- This is also referred to as True Positive Rate or *sensitivity*.

| | Prediction Positive | Prediction Negative |
|---|---|---|
| Condition Positive | True Positive (TP) | False Negative (FN) |
| Condition Negative | False Positive (FP) | True Negative (TN) |

**True Positive Rate aka Sensitivity, Recall**

$$TPR = \frac{TP}{P} = \frac{TP}{(TP + FN)}$$

# 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 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.

| | Prediction Positive | Prediction Negative |
| --- | --- | --- |
| Condition Positive | True Positive (TP) | False Negative (FN) |
| Condition Negative | False Positive (FP) | True Negative (TN) |

**Positive Prediction Value aka Precision**

$$PPV = \frac{TP}{(TP + FP)}$$

# 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.

# ACTIVITY: KNOWLEDGE CHECK

EXERCISE

## ANSWER THE FOLLOWING QUESTIONS

1. What would the precision and recall be for the following confusion matrix (with "green" being "true")?

| | **predicted_green** | **predicted_not_green** |
|---|---|---|
| is_green | 13 | 7 |
| is_not_green | 8 | 12 |

## DELIVERABLE

Answers to the above question

# THE DIFFERENCE BETWEEN PRECISION AND RECALL

- The key difference between the two is the attribution and value of error.
- Should our model be more picky in avoiding false positives (precision)?
- Or should it be more picky in avoiding false negatives (recall)?
- The answer should be determined by the problem you’re trying to solve.

DEMO

# UNDERSTANDING TRADEOFF

# UNDERSTANDING TRADEOFF

- Let’s consider the following data problem:  we are given a human resource dataset in order to identify factors that are indicative of employee turnover

- Optimizing toward recall, we could assume that every employee will leave the company

- The trade-off, a lower precision, is that you spend unnecessary resources on employees who were not intending to leave

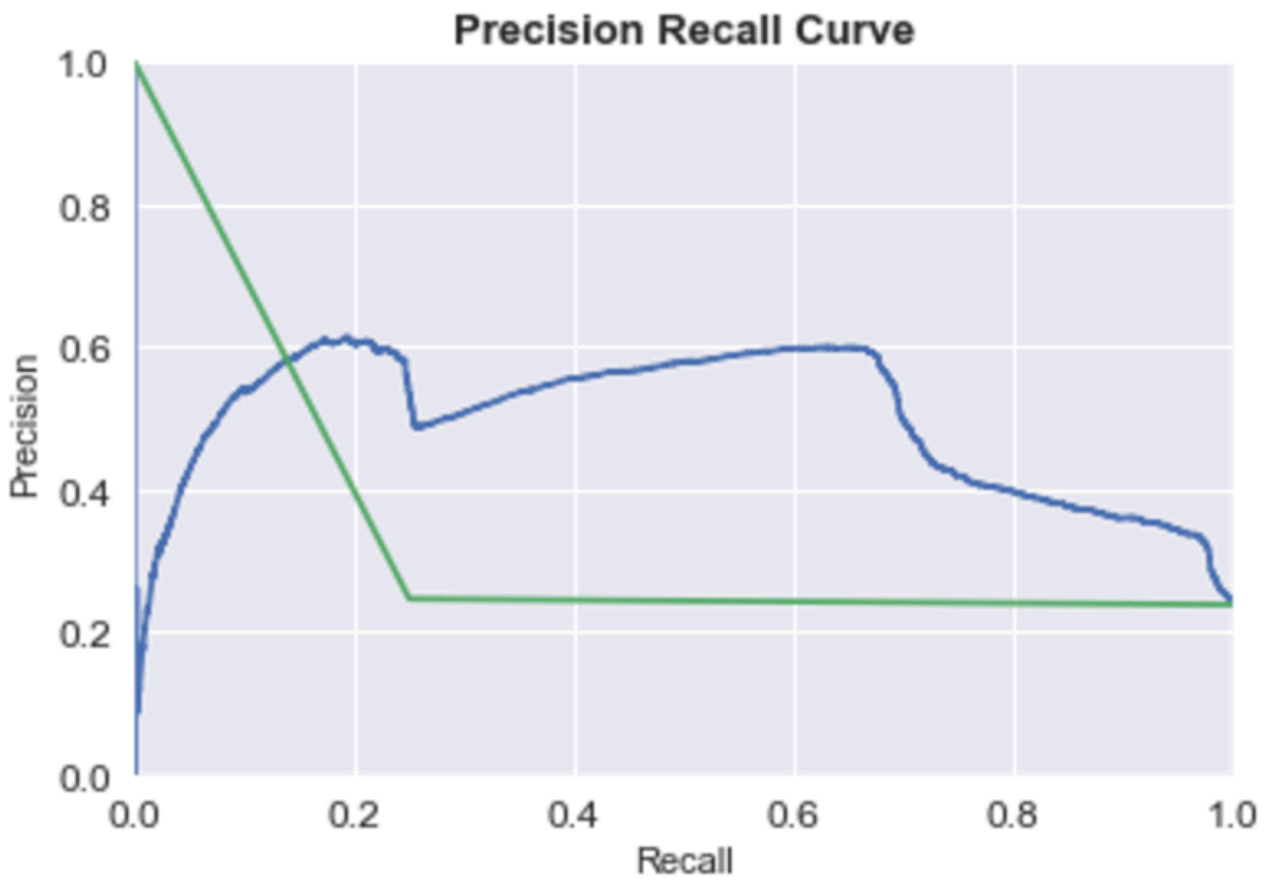
# UNDERSTANDING TRADEOFF

- Optimizing toward precision, we would focus on high risks individuals with higher probabilities of turnover

- The trade-off here would be lower recall.  We might miss certain employees who were at risk of leaving that we could have intervened

# UNDERSTANDING TRADEOFF

- Below is a sample plot that shows how precision and recall are related for a model used to predict employee attrition.



# UNDERSTANDING TRADEOFF

- This plot is based on choosing decision line thresholds, much like the AUC figure from the previous class.

- In terms of modeling delays, this would be like moving the decision line for leaving from a probability of 0.01 up to 0.99, and then calculating the precision and recall at each decision.

# UNDERSTANDING TRADEOFF

- Interpreting our plot, there's a few interesting nuggets compared to the benchmark (green line):
    - At a lower recall (below 0.15), there is a noticeable lower precision in the model.
    - Beyond 0.15 recall, the model outperforms the benchmark.
- Whether we're optimizing for recall or precision, this plot helps us decide based on the thresholds.

## GUIDED PRACTICE

# COST BENEFIT ANALYSIS

# ACTIVITY: COST BENEFIT ANALYSIS

EXERCISE

## 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

EXERCISE

## 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 $1.90 per user.

# ACTIVITY: COST BENEFIT ANALYSIS

EXERCISE

## 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 if 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

# SHOWING WORK

# SHOWING WORK

- We've spent a lot of time exploring our data and building a reasonable model that performs well.
- However, if we look at our visuals, they are most likely:
    - Statistically heavy:  Most people don't understand histograms.
    - Overly complicated:  Scatter matrices produce too much information.
    - Poorly labeled:  Code doesn't require adding labels, so you may not have added them.

# SHOWING WORK

- In order to convey important information to our audience, make sure our charts are:
    - Simplified
    - Easily interpretable
    - Clearly labeled

# SIMPLIFIED

- At most, you'll want to include figures that either explain a variable on its own or explain that variable's relationship with a target.
- If your model used a data transformation (like natural log), just visualize the original data.
- Try to remove any unnecessary complexity.

# EASILY INTERPRETABLE

- Any stakeholder looking at a figure should be seeing the exact same thing you're seeing.
- A good test for this is to share the visual with others less familiar with the data and see if they come to the same conclusion.
- How long did it take them?

# CLEARLY LABELED

- Take the time to clearly label your axis, title your plot, and double check your scales - especially if the figures should be comparable.

- If you're showing two graphs side by side, they should follow the same Y axis.

# QUESTION TO ASK

- When building visuals for another audience, ask yourself these questions:
  - **Who**:  Who is my target audience for the visual?
  - **What**:  What do they already know about this project?  What do they need to know?
  - **How**:  How does my project affect this audience?  How might they interpret (or misinterpret) the data?

## 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

- This visual can help showcase the range of effect on leaving from both satisfaction level and work accidents.

- Given this model, employees with work accidents are less likely to leave

- The likelihood of leaving decreases as the satisfaction goes up.

DEMO

# VISUALIZING PERFORMANCE AGAINST BASELINE

# VISUALIZING PERFORMANCE AGAINST BASELINE

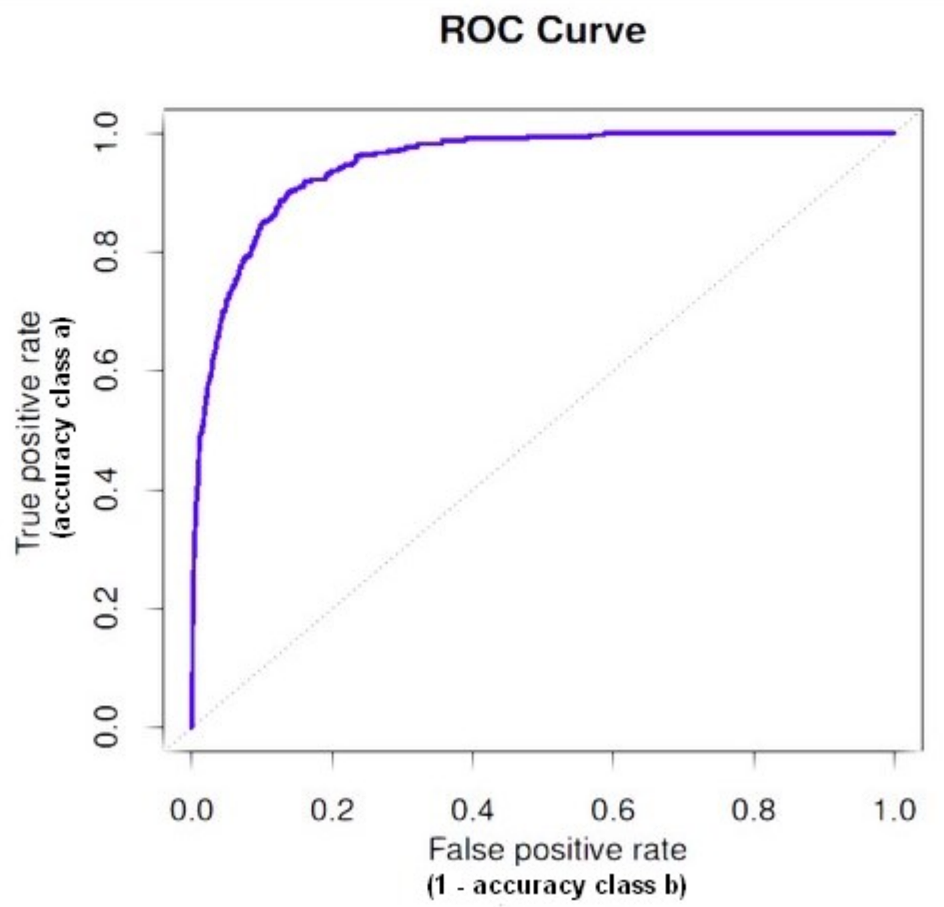
- Another approach of visualization is the effect of your model against a baseline, or - even better - against previous models.
- Plots like this will also be useful when talking to your peers - other data scientists or analysts who are familiar with your project and interested in the progress you've made.

# VISUALIZING PERFORMANCE AGAINST BASELINE

- For classification, we've practiced plotting AUC and precision-recall plots.  Consider the premise of each:

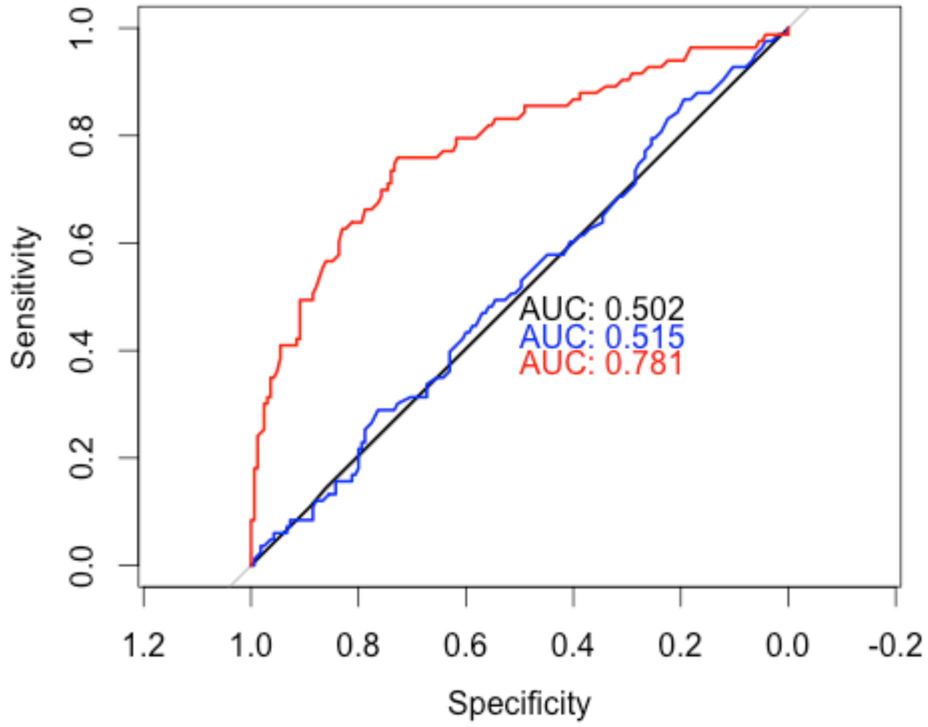
  - AUC plots explain and represent "accuracy" as having the largest area under the curve. Good models will be high and to the left.

  - For precision-recall plots, it will depend on the *cost* requirements.  Either a model will have good recall at the cost of precision or vice versa.



# VISUALIZING PERFORMANCE AGAINST BASELINE

- When comparing multiple models:
  - For AUC plots, you'll be interested in which model has the *largest* area under the curve.
  - For precision-recall plots, based on the cost requirement, you are looking at which model has the best precision given the same recall, or the best recall given the same precision.

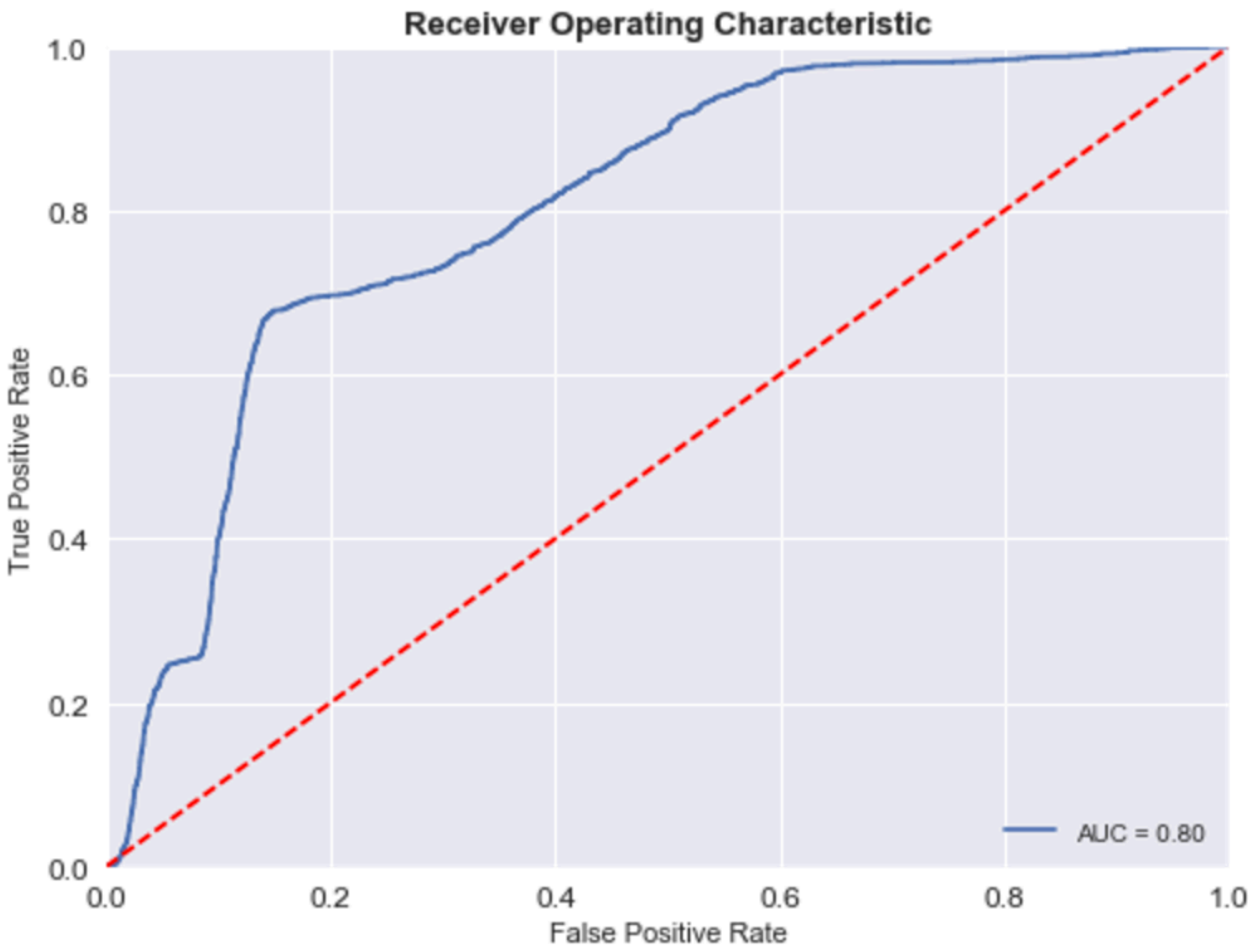


# VISUALIZING PERFORMANCE AGAINST BASELINE

- This plot showcases:

1. The model using data outperforms a baseline dummy model.

2. By adding other features, there's some give and take with probability as the model gets more complicated.



# TOPIC REVIEW

## REVIEW AND NEXT STEPS

- What do precision and recall mean? How are they similar and different to True Positive Rate and False Positive Rate?

- How does cost benefit analysis play a role in building models?

- What are at least two very important details to consider when creating visuals for a project's stakeholders?

- Why would an AUC plot work well for a data science audience but not for a business audience? What would be a more effective visualization for that group?

LESSON

# Q & A