GENERAL ASSEMBLY

COMMUNICATING RESULTS

Angelo Klin
Katra Analytics

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 visualisation for presentations vs. exploratory data analysis
- Identify the components of a concise, convincing report and how they relate to specific audiences/stakeholders

PRE-WORK

PRE-WORK REVIEW

- Understand results from a confusion matrix and measure true positive rate and false positive rate
- Create and interpret results from a binary classification problem
- Know what a decision line is in logistic regression

OPENING

COMMUNICATING RESULTS

WE BUILT A MODEL! NOW WHAT?

- We have 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 optimised with L2 regularisation

  - Gender was more important than age in the predictive model because it has a larger coefficient

  - Here is 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 have built
- Some people may be familiar with basic visualisation, 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 will 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

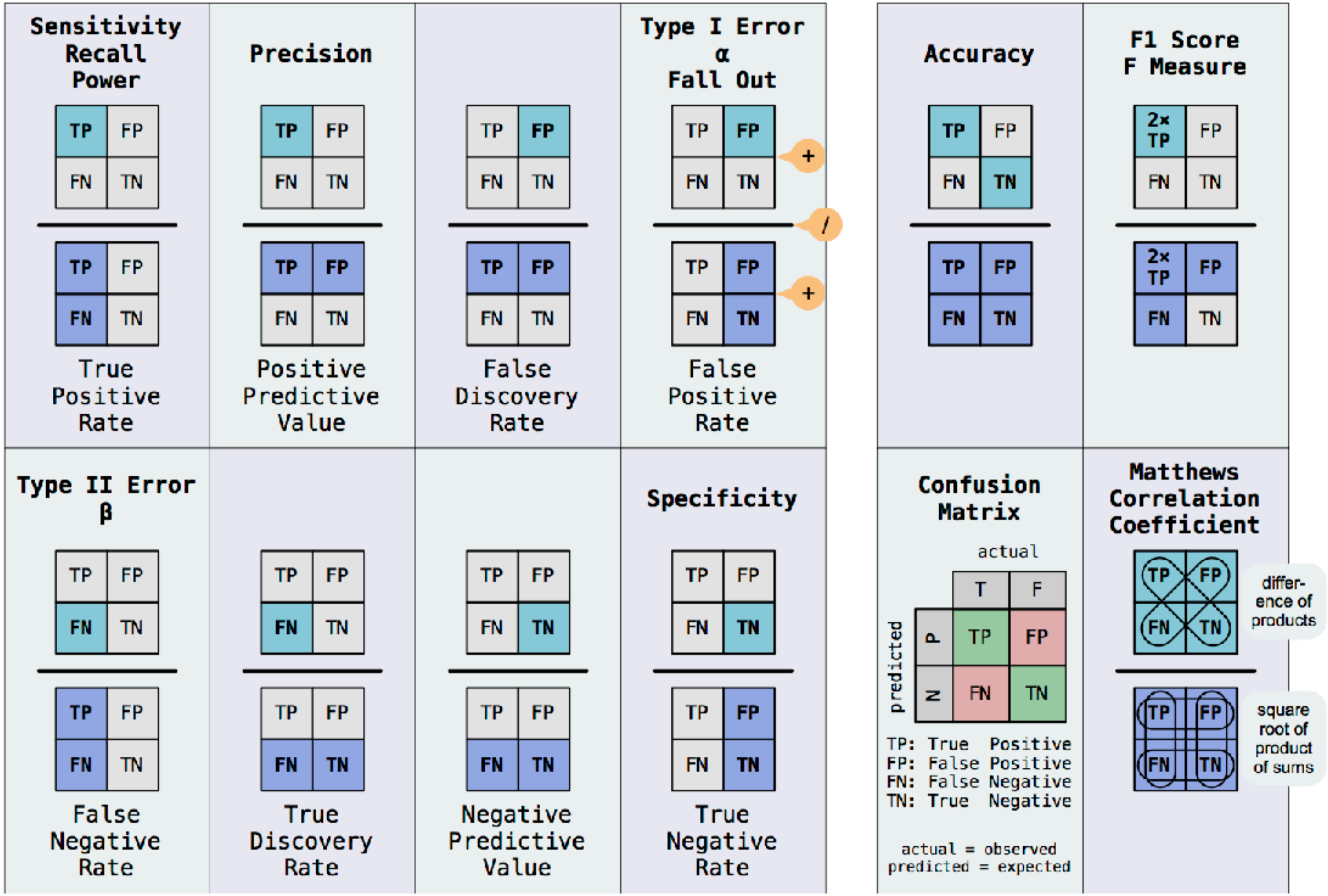
REVIEW

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

STATISTICAL CLASSIFICATION METRICS



ACTIVITY: KNOWLEDGE CHECK

EXERCISE

DIRECTIONS: 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

PRECISION AND RECALL

PRECISION AND RECALL

- Our previous metrics were primarily designed for less biased data problems
  - We could be interested in both outcomes, so it was important to generalise our approach

- For example, we may be interested if a person will vote for a Liberal or Labour
  - This is a binary problem, but we are interested in both outcomes

PRECISION AND RECALL

- Precision and Recall are 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 will not

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 correct?”

ACTIVITY: KNOWLEDGE CHECK

EXERCISE

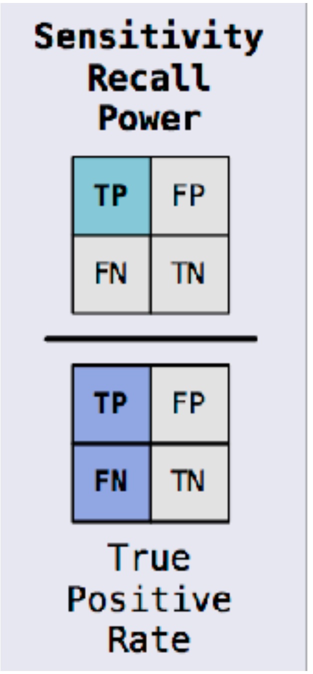
DIRECTIONS: ANSWER THE FOLLOWING QUESTIONS

1. If the goal of the “recall” metric is to identify specific values of a class correctly, what other metric performs a similar calculation?

THE FORMULA 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



$$recall = \frac{TP}{P} = \frac{TP}{TP + FN}$$

THE FORMULA FOR RECALL

- Imagine predicting the colour of a marble as either **red** or **green**
  - There are 10 of each

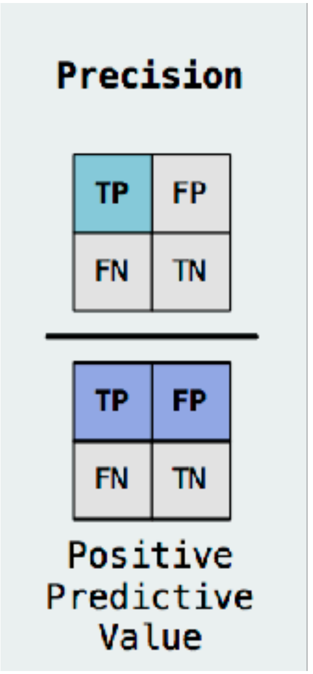
- If the model identifies 8 of the **green** marbles as **green**, the recall is

$$recall = \frac{TP}{P} = \frac{8}{10} = 0.8$$

- However, this says nothing of the number of **red** marbles that are also identified as **green**

THE FORMULA 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



$$precision = \frac{TP}{TP + FP}$$

THE FORMULA FOR PRECISION

- 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

$$precision = \frac{TP}{TP + FP} = \frac{10}{10 + 2} = 0.833$$

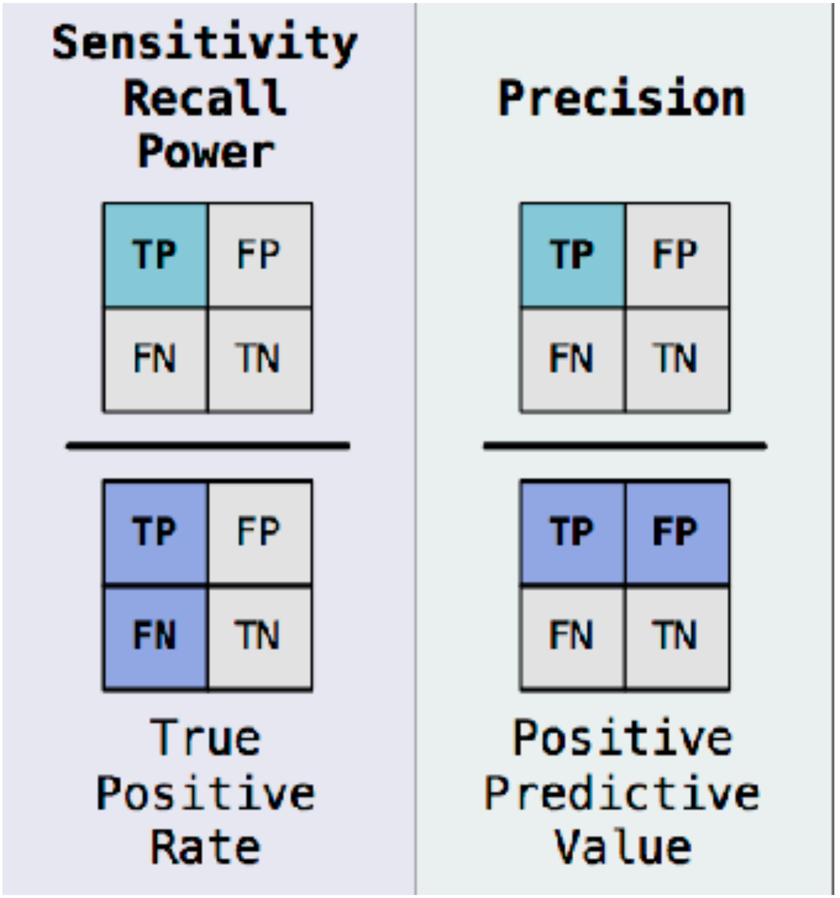
THE FORMULA FOR PRECISION

- Precision
  - % of selected items that are correct

$$precision = \frac{TP}{TP + FP}$$

- Recall
  - % of correct items that are selected

$$recall = \frac{TP}{TP + FN}$$



THE DIFFERENCE BETWEEN PRECISION AND RECALL

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

ANOTHER EXAMPLE

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

		Actual Class	
		Green	Red
Predicted Class	Green	8	4
	Red	12	12
	Total	20	16

- We could calculate precision for green marbles as

$$precision = \frac{8}{8+4} = 0.6666$$

- We could calculate recall for green marbles as

$$recall = \frac{8}{8+12} = 0.4000$$

ACTIVITY: ANOTHER EXAMPLE

EXERCISE

DIRECTIONS: ANSWER THE FOLLOWING QUESTIONS

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

		Actual Class	
		Is Green	Is Not Green
Predicted	Green	13	8
	Not Green	7	12
	Total	20	16

DEMONSTRATION

UNDERSTANDING TRADEOFF

UNDERSTANDING TRADEOFF

- Let’s consider the following data problem:
  - We are given a data set in order to predict or identify traits for typically late flights

- Optimising toward recall, we could assume that every flight will be delayed

- The trade-off, a lower precision, is that this could create even further delays, missed flights, etc

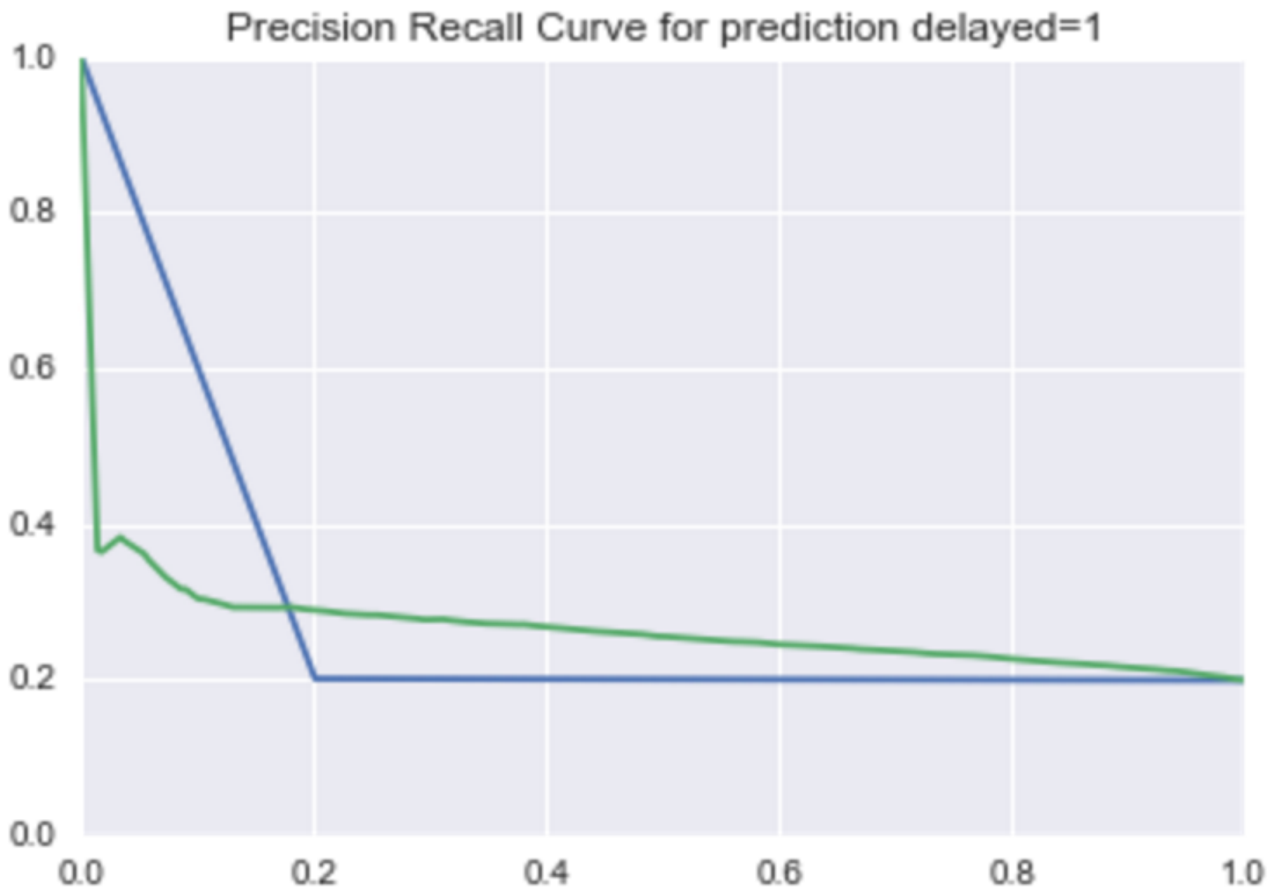
UNDERSTANDING TRADEOFF

- Optimising toward precision, we would specifically look to identify flights that will be late

- The trade-off here would be lower recall
  - We might miss flights that would be delayed, causing a strain on the system

UNDERSTANDING TRADEOFF

- Below is a sample plot that shows how precision and recall are related for a model used to predict late flights



UNDERSTANDING TRADEOFF

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

- In terms of modelling delays, this would be like moving the decision line for lateness 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 is a few interesting nuggets compared to the benchmark (blue line):
  - At a lower recall (below 0.2), there is a noticeable lower precision in the model
  - Beyond 0.2 recall, the model outperforms the benchmark
- Whether we are optimising for recall or precision, this plot helps us decide based on the 0.3 threshold

GUIDED PRACTICE

COST BENEFIT ANALYSIS

ACTIVITY: COST BENEFIT ANALYSIS

EXERCISE

DIRECTIONS: (15 MINUTES)

1. One tool that complements the confusion matrix is cost-benefit analysis, where you attach a value to correctly and incorrectly predicted data
2. 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
3. Consider the following marketing problem: As a data scientist working on marketing spend, you have 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)

1. The **benefit** of a true positive is the retention of a user ($10 for the month)
2. The **cost** of a false positive is the spend of the campaign per user ($0.05)
3. 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 did not benefit!)
4. The **benefit** of a true negative is 0: No spend on users who would have never retained
5. To calculate Cost-Benefit, we will use this following function:

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

6. Which for our marketing problem, comes out to this:

```
(0.2 * 10) + (0.5 * 0) - (0.2 * 0.05) - (0.1 * 0)
```

7. or $1.99 per user targeted

ACTIVITY: COST BENEFIT ANALYSIS

EXERCISE

DIRECTIONS: (15 MINUTES)

1. Think about precision, recall and cost benefit analysis to answer the following questions:
   a. How would you rephrase the business problem if your model was optimising toward precision? i.e., How might the model behave differently and what effect would if have?
   b. How would you rephrase the business problem if your model was optimising toward recall?
   c. What would the most ideal model look like in this case?

SHOWING WORK

SHOWING WORK

- We have 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 do not understand histograms

  - Overly complicated: Scatter matrices produce too much information

  - Poorly labeled: Code does not 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 will 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 visualise 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 are 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 are showing two graphs side by side, they should follow the same Y axis

QUESTIONS TO ASK

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

DEMONSTRATION

VISUALISING MODELS OVER VARIABLES

VISUALISING 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

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

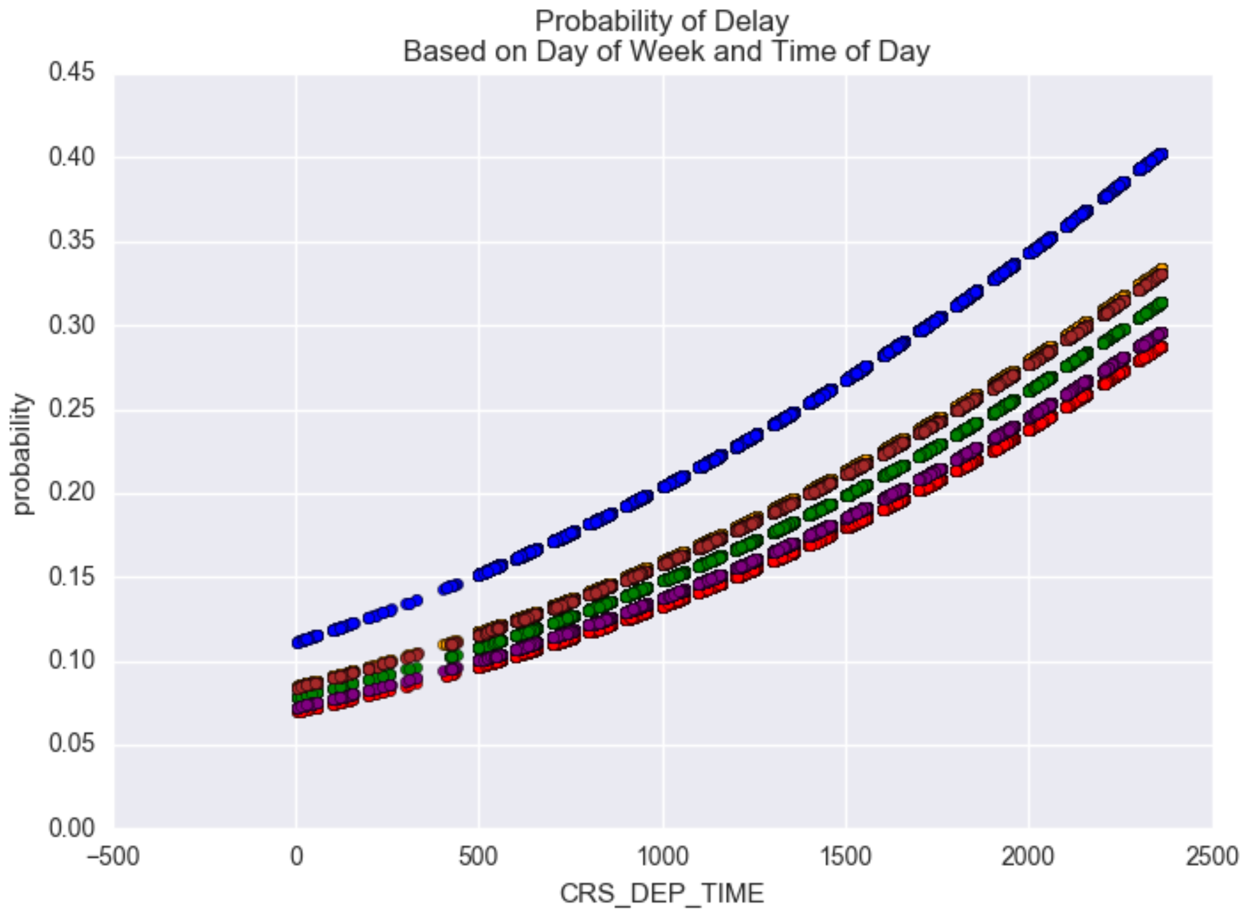
- Open the starter code from the class repository and follow along
  - ~/lessons/lesson-10/code/starter/starter-10.ipynb

VISUALISING 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

EXERCISE

DIRECTIONS: (5 MINUTES)

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

DEMONSTRATION

VISUALISING PERFORMANCE AGAINST BASELINE

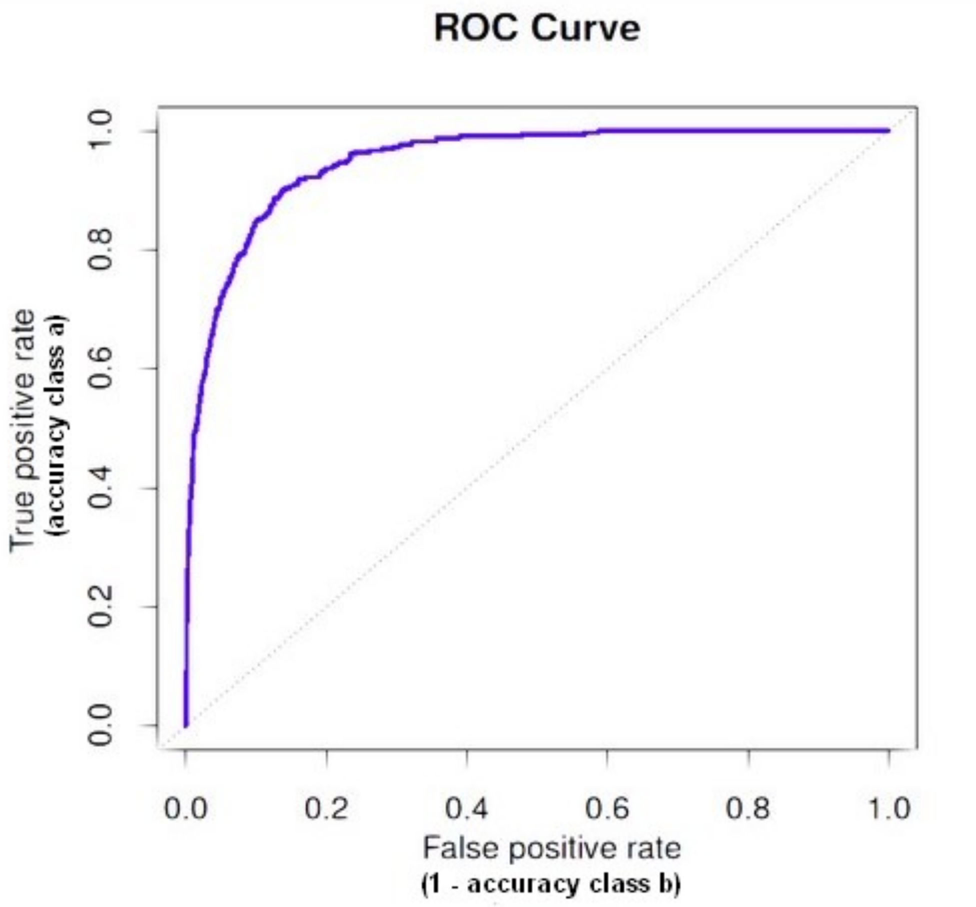
VISUALISING PERFORMANCE AGAINST BASELINE

- Another approach of visualisation is the effect of your model against a baseline
  - 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 have made

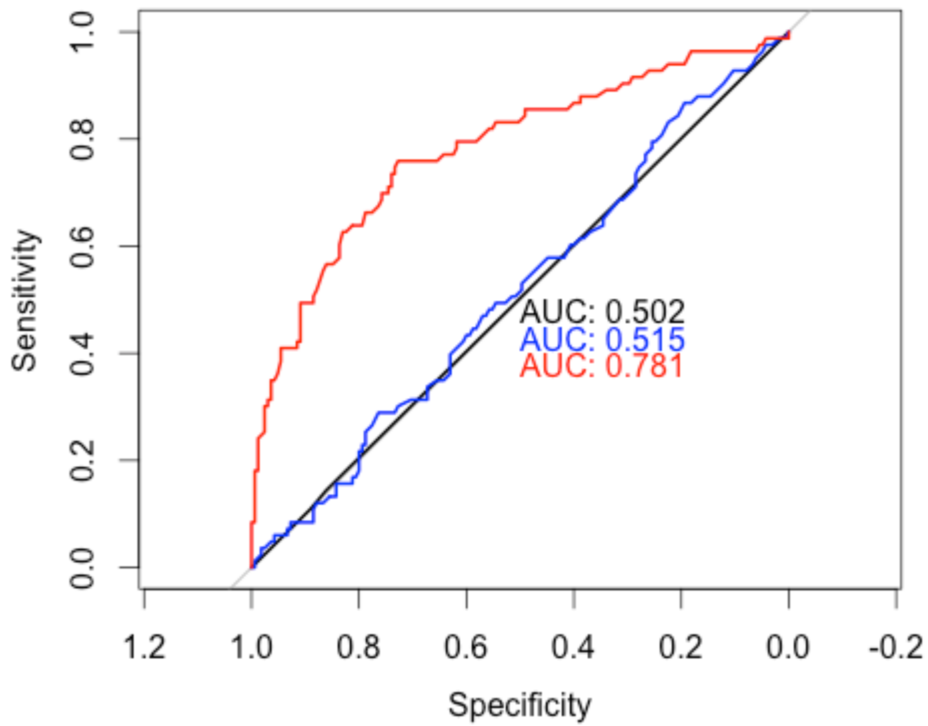
VISUALISING PERFORMANCE AGAINST BASELINE

- For classification, we have 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



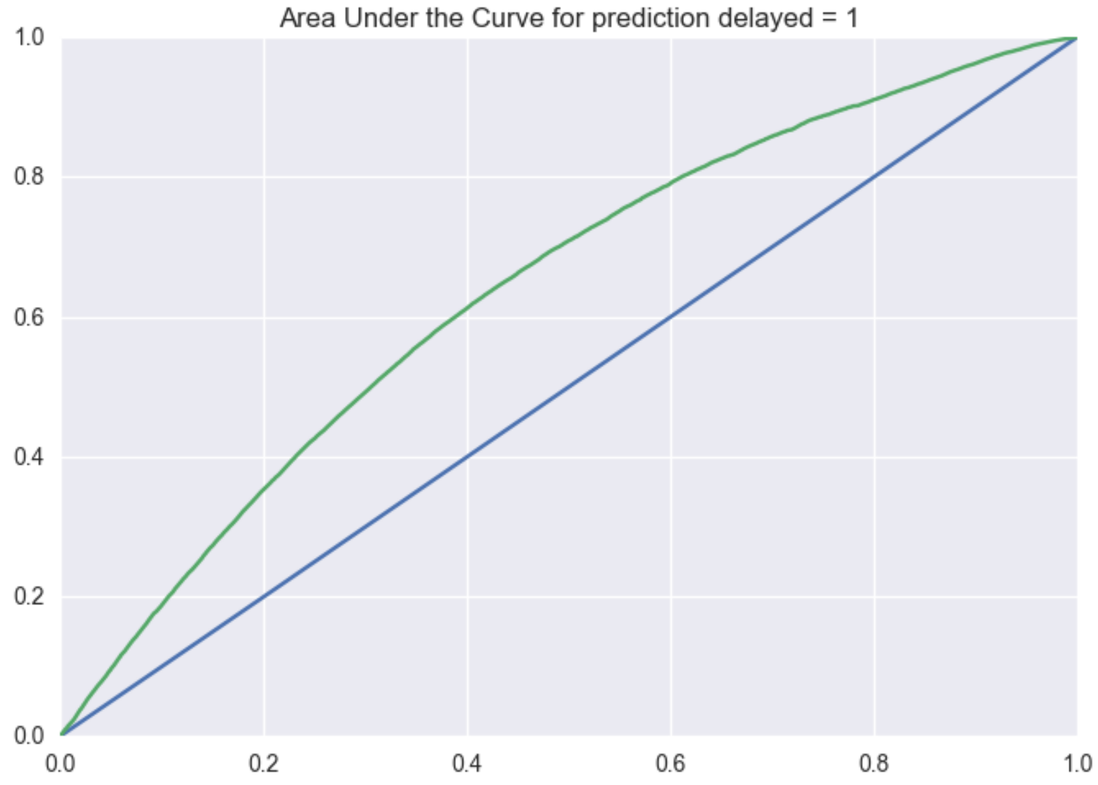
VISUALISING PERFORMANCE AGAINST BASELINE

- When comparing multiple models:
  - For AUC plots, you will 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



VISUALISING PERFORMANCE AGAINST BASELINE

- This plot showcases:
  - The model using data outperforms a baseline dummy model
  - By adding other features, there is some give and take with probability as the model gets more complicated



ACTIVITY: TRY IT OUT

EXERCISE

DIRECTIONS: (5 MINUTES)

1. In a similar approach, use the scikit-learn precision_recall_curve function to enable you to plot the precision-recall curve of the four models from above. Keep in mind precision in the first array is returned from the function, but the plot shows it as the y-axis.
2. Explain what is occurring when the recall is below 0.2
3. Based on this performance, is there a clear winner at different thresholds?
4. **Bonus**: Redo both the AUC and precision-recall curves using models that have been cross validated using kfold. How do these new figures change your expectations for performance?

PROJECT PRACTICE

ACTIVITY: PROJECT PRACTICE

EXERCISE

DIRECTIONS: (45 MINUTES)

1. Using models built from the flight data problem earlier in class, work through the same problems. Your data and models should already be accessible. Your goals:
   a. There are many ways to manipulate this data set. Consider what is a proper "categorical" variable and keep only what is significant. You will easily have 20+ variables. Aim to have at least three visuals that clearly explain the relationship of variables you've used against the predictive survival value.
   b. Generate the AUC or precision-recall curve (based on which you think makes more sense) and have a statement that defines, compared to a baseline, how your model performs and any caveats.
   c. For example: "My model on average performs at x rate, but the features under-perform and explain less of the data at these thresholds." Consider this as practice for your own project, since the steps you will take to present your work will be relatively similar.

CONCLUSION

TOPIC REVIEW

TOPIC REVIEW

- 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 visualisation for that group?

BEFORE NEXT CLASS

DUE DATE

- Project:
  - Unit Project 4

Q & A

EXIT TICKETS

DON'T FORGET TO FILL OUT YOUR EXIT TICKET

Exit Ticket Link

What's the lesson number?	10
What was the topic of the lesson?	Communicating Results

CREDITS AND REFERENCES

CREDITS AND REFERENCES

- Pandas plotting
- Text with Matplotlib
- Anatomy of Matplotlib