GENERAL ASSEMBLY

COMMUNICATING RESULTS

Haley Boyan

Data Science Instructional Associate, GA DC

COMMUNICATING RESULTS

LEARNING OBJECTIVES

- Describe the difference between exploratory data analysis and visualization for sales/presentations
- Identify the components of a concise, convincing report and how they relate to specific audiences/stakeholders
- Explain accuracy, precision, recall, and type errors of a model
- Articulate the cost of false positives vs. false negatives.
- Explain a confusion matrix

TITLE

PRE-WORK

PRE-WORK REVIEW

- Build linear and logistic regression models
- Use cross-validation and regularization to improve performance
- Explain model coefficients and key metrics

COMMUNICATING RESULTS

OPENING

WE BUILT A MODEL!

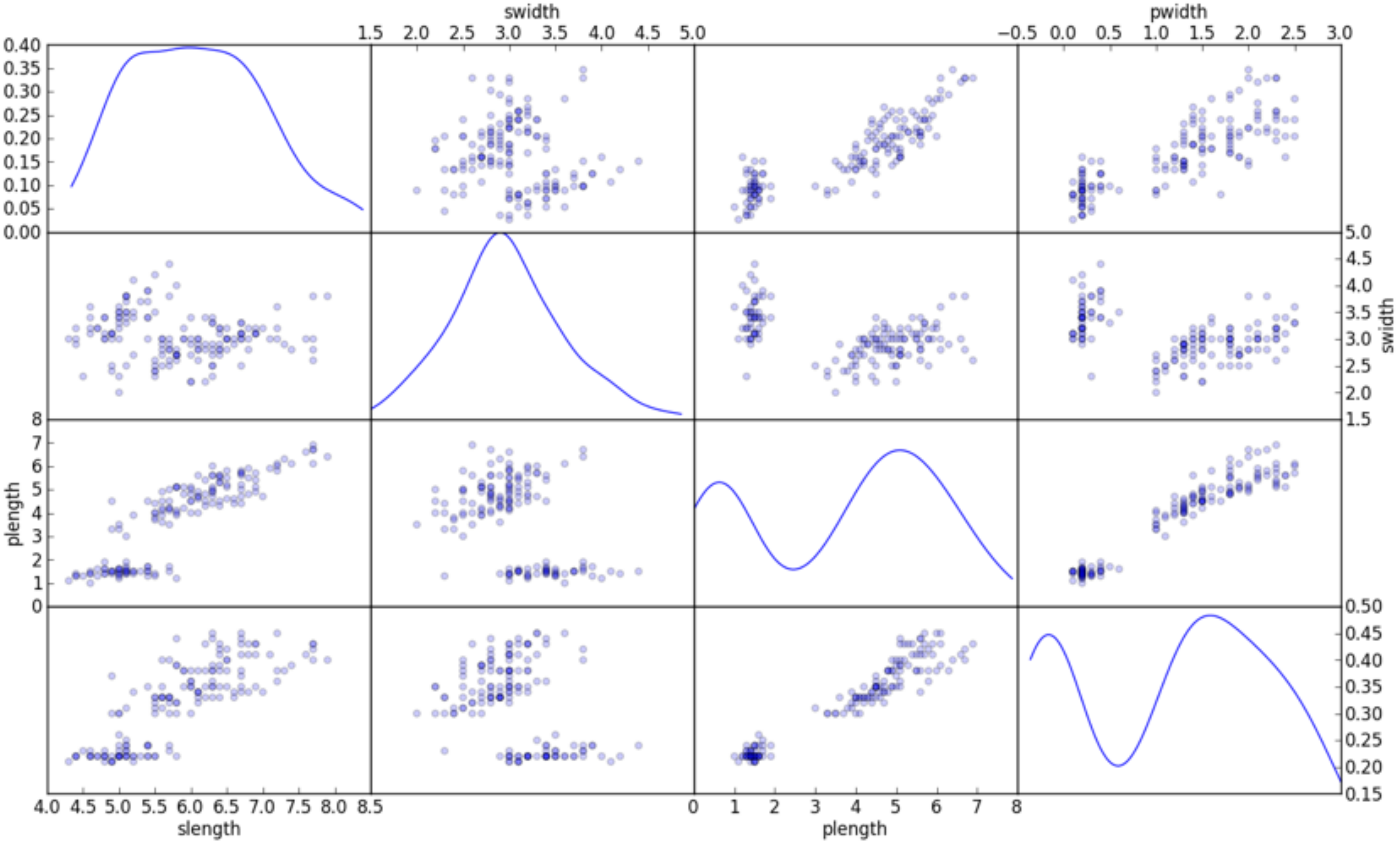
- We’ve been building models, but what we’ve made is not what we could show in a presentation

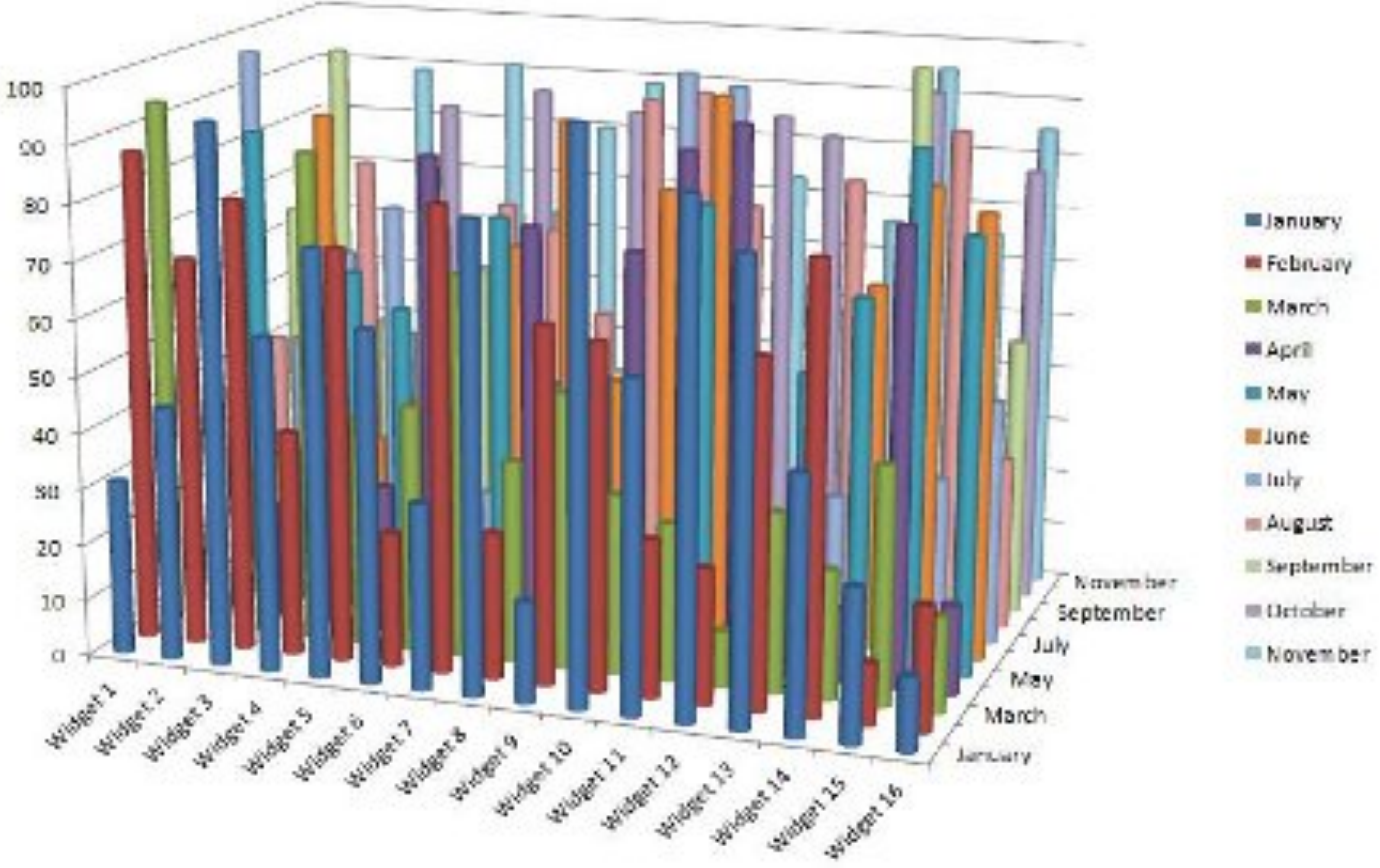
- Imagine user responses to some of the following statements:
    - The predictive model I built has an accuracy of 80%.
    - The logistic regression was optimized with L2 regularization, so you know it's good.
    - Gender was more important than age in the predictive model because it had a larger coefficient.
    - Here's the AUC chart that shows how well the model did.

COMMUNICATING RESULTS

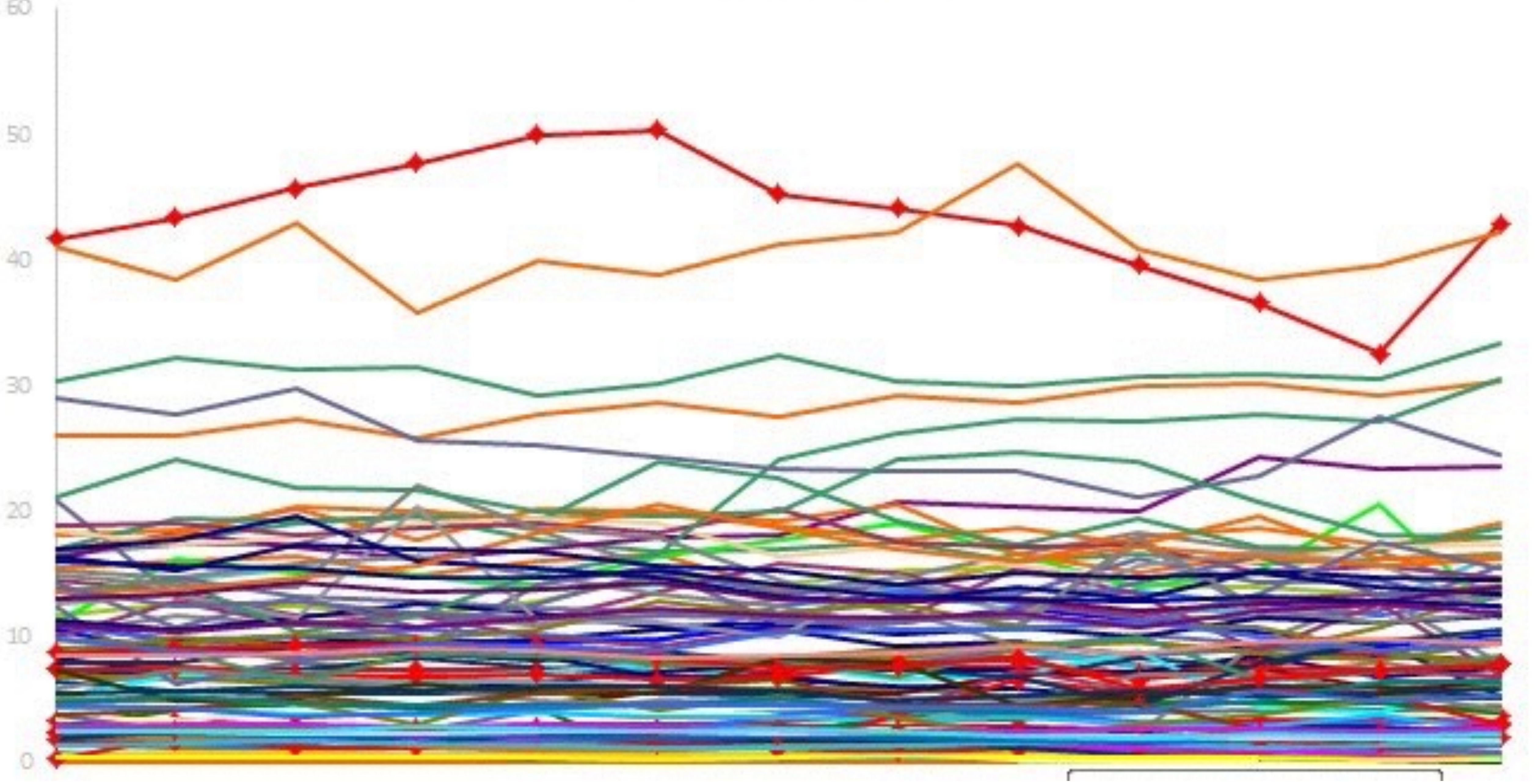
- We've spent time exploring data and building models that performs well.

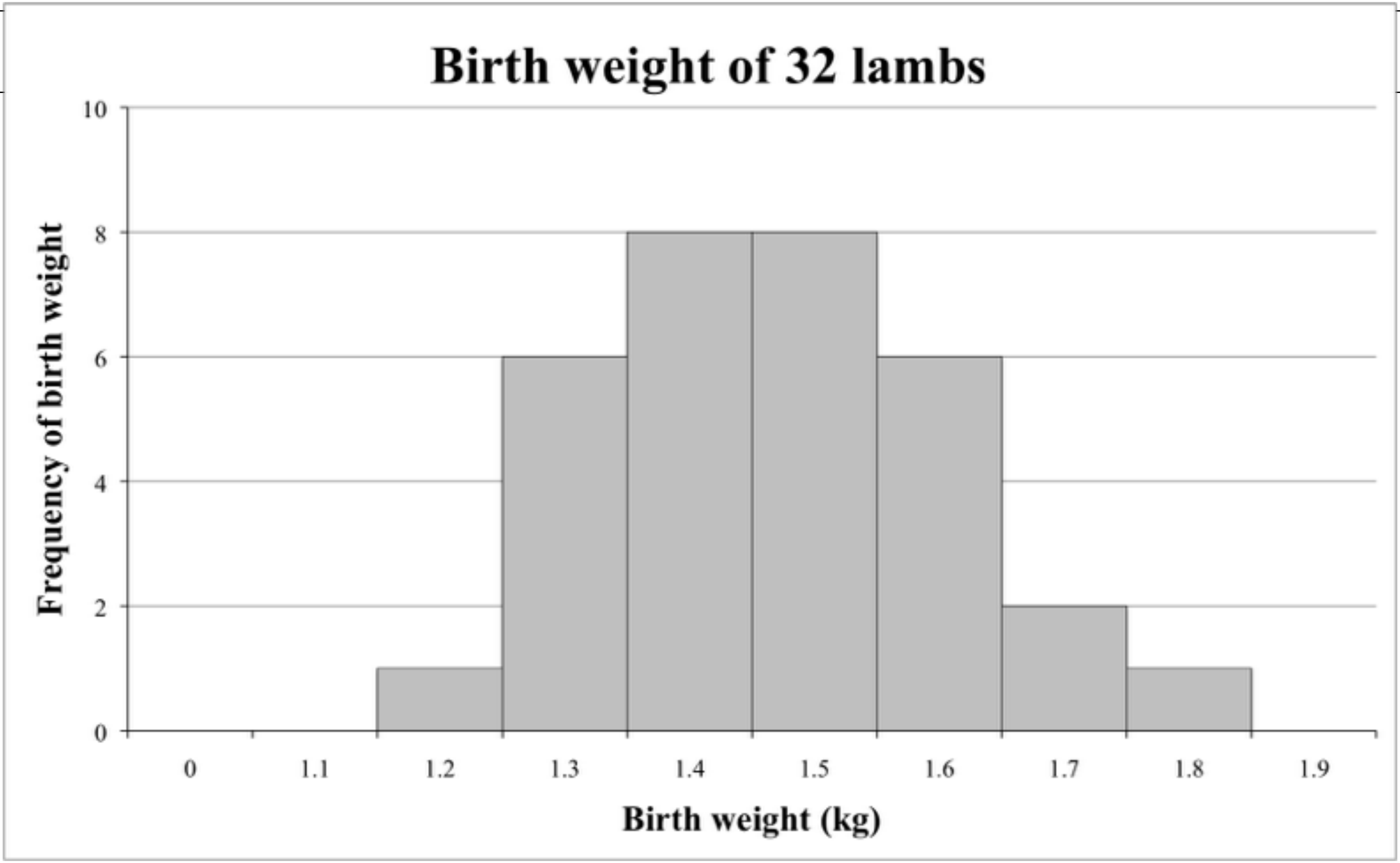
- However, our plots and visuals may be:
  - **Statistically heavy**
  - **Overly complicated**
  - **Poorly labeled**


swidth
pwidth
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
swidth
8
7
6
5
4
3
2
1
0
plength
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
8.5
slength
0
1
2
3
4
5
6
7
8
plength


100
90
80
70
60
50
40
30
20
10
0
Widget 1
Widget 2
Widget 3
Widget 4
Widget 5
Widget 6
Widget 7
Widget 8
Widget 9
Widget 10
Widget 11
Widget 12
Widget 13
Widget 14
Widget 15
Widget 16
January
March
May
July
September
November
January
February
March
April
May
June
July
August
September
October
November

Monthly Share (%)




Birth weight of 32 lambs
Frequency of birth weight
10
8
6
4
2
0
0
1.1
1.2
1.3
1.4
1.5
1.6
1.7
1.8
1.9
Birth weight (kg)

COMMUNICATING RESULTS

- What questions should we ask when building visuals for external, non-technical audiences?
- These are a great start:
    - Who: Who is my target audience
    - What: What do they already know about this project, and what do they need to know?
    - How: How does my project affect my audience? How might they interpret/misinterpret the data?

AGENDA

- How to begin the communication process
- Methods of communicating results:
    - Score/Accuracy Score/Cross Validation Score
    - Confusion Matrix
    - Classification Report
    - ROC / AUC Plots
    - Plotting variables, probabilities, and comparisons
    - Coefficients
    - Model Tuning Results

COMMUNICATING RESULTS

SCORE

WHEN TO USE IT

- Estimator score: Estimators have a score method providing a default evaluation criterion for the problem they are designed to solve. This is discussed in each estimator's documentation.

- Typically, this involves putting in a set of data and a target, and returns how frequently the model predicts the target using the data input.

- This gives a good **baseline** for any model on a given set of data. However, you need to be conscious of what data you're giving it (accuracy or cross validation scores are often more useful).

COMMUNICATING RESULTS

ACCURACY SCORE

WHEN TO USE IT

- When you are comparing the results of your prediction to true values directly and need a simple, overall metric

- Computes the accuracy, either the fraction (default) or the count (normalize=False) of correct predictions.

- The fraction of correct predictions over n samples is defined as:

$$\texttt{accuracy}(y, \hat{y}) = \frac{1}{n_\text{samples}} \sum_{i=0}^{n_\text{samples}-1} 1(\hat{y}_i = y_i)$$

    - y-hat_i: predicted value of the i-th sample
    - y_i: corresponding true value

- In multilabel classification, this function computes subset accuracy: the set of labels predicted for a sample must exactly match the corresponding set of labels in y_true.

INTERPRET

- If you returned a fraction (default), this is the PROPORTION of true positives - positives that your model predicted correctly

- If you returned a count (normalized=False), this is the COUNT of true positives that your model got right

VISUALIZE

- Pretty straightforward, returns either a ratio or an integer

EXPLAIN

- ‘Our model was able to correctly predict this result PROPORTION/ NUMBER of times.’

CONSIDERATIONS

- Great for multi class problems (potentially best), because trade offs in other metrics cause a lack of clarity

- Should not be used as a stand-alone evaluator otherwise

- Need to include confidence intervals and qualify with what certainty you are presenting your accuracy score

COMMUNICATING RESULTS

CROSS VALIDATION SCORE

WHEN TO USE IT

- In classification models to legitimize your accuracy score, and easily spot an overfit model

- Cross-validation is NOT an estimate of test-error conditional on the training set. Rather, it is an estimate of the unconditional test error.

- In other words, this is the expected test error if you are also randomizing over the world of possible training sets, rather than the precise training set you've been given.

INTERPRET

- This will return an array of scores, one for each cross fold
- Look at:
    - mean - get single score
    - std - how consistent is the model
- Can use along with score and/or accuracy score to support model quality

VISUALIZE

- Print array, or mean if you want a single number
- Would plotting be beneficial?

EXPLAIN

- This is a validation of how well your model performs, and has been used to make sure that it can perform well on multiple arrangements of data. It was a cautionary tactic to avoid unusual effects of your particular data set.

COMMUNICATING RESULTS

REGRESSION METRICS

ERROR IN REGRESSION

- While you likely don’t want to get into the math behind your error, it is a way of explaining how closely your model is able to match the data, and therefore predict values.

- This is a good way to showcase the accuracy of a regression model. If your audience has some statistical knowledge, it’s good to include this along with the accuracy score.

- Make sure to be clear about what error you are discussing and their differences

ERROR IN REGRESSION

RMSE

- The mean_squared_error function computes mean square error, a risk metric corresponding to the expected value of the squared (quadratic) error loss or loss.
- Present RMSE instead of simple MSE because the units/scale are the same as the variable.

R2 ERROR

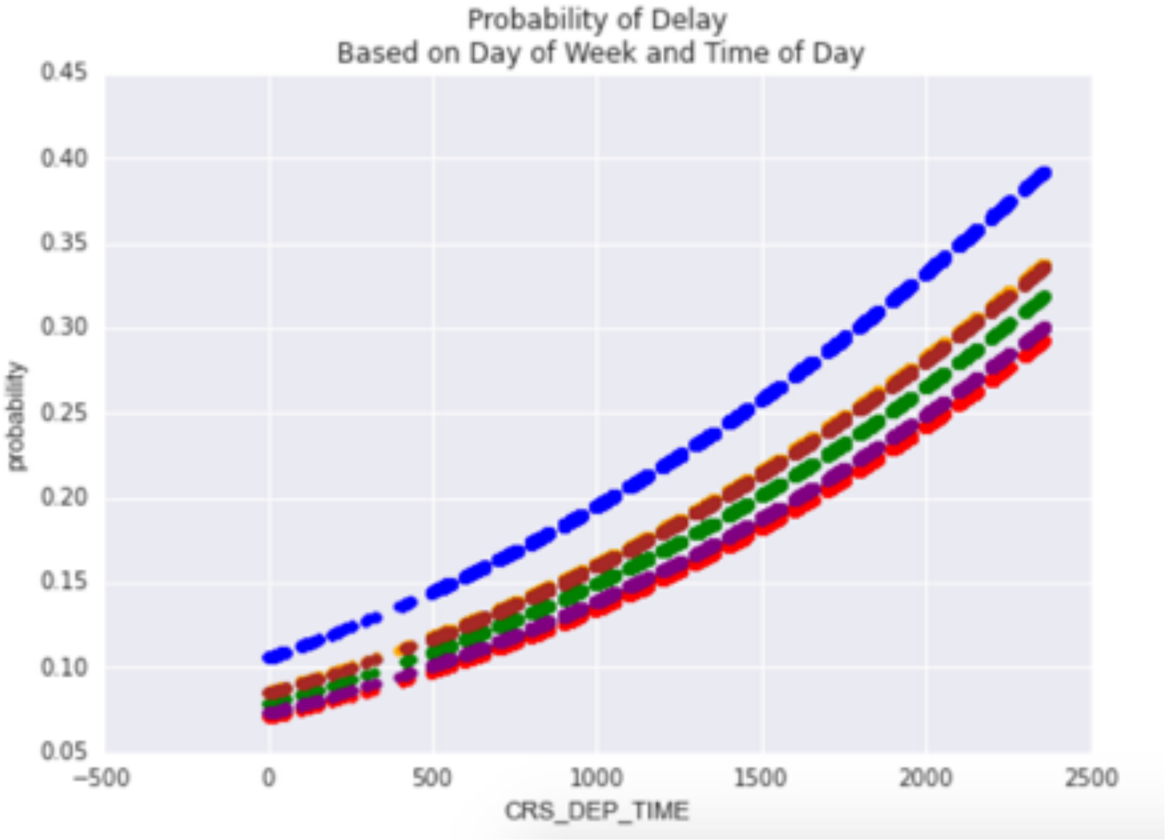
- The r2_score function computes R², the coefficient of determination. It provides a measure of how well future samples are likely to be predicted by the model. Best possible score is 1.0 and it can be negative (because the model can be arbitrarily worse). A constant model that always predicts the expected value of y, disregarding the input features, would get a R^2 score of 0.0.

WHEN TO USE IT

- If you care about the specificity of an individual feature (like coefficients)

- One effective way to explain your model over particular variables is to plot the predicted variables against the most explanatory variables. For example, with logistic regression, plotting the probability of a class against a variable can help explain the range of effect on the model.

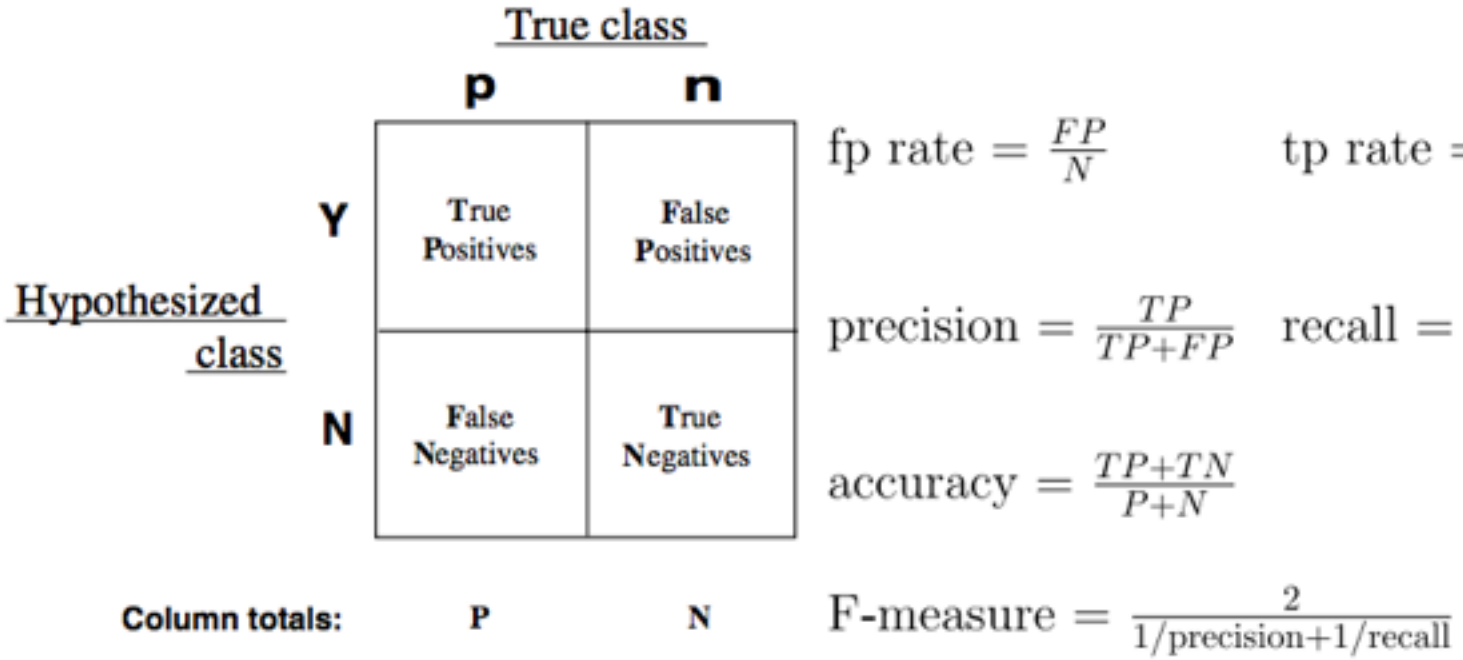
- A visual like this can help showcase the range of effect on delays from both the day of week and the time of day: given this model, some days are more likely to have delays than others, and the likelihood of a delay increases as the day goes on.



CONFUSION MATRIX

WHEN TO USE IT

- Evaluates the accuracy of **classification** problems
- Allow for the interpretation of correct and incorrect predictions for **each class label**.
- The confusion matrix is the beginning step for the majority of classification metrics, and gives our predictions deeper meaning beyond an accuracy score.



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

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

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

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

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

CONFUSION MATRIX REVIEW

- How do we calculate:
  - 1. Accuracy
  - 2. True Positive Rate
  - 3. False Positive Rate

- Check: What would the precision and recall be for the following confusion matrix (with "blue" being "true")?

	Predicted Blue	Predicted Not Blue
Is Blue	13	7
Is Not Blue	8	12

INTERPRET

- Obviously you want most values in your True Positive and True Negative cells

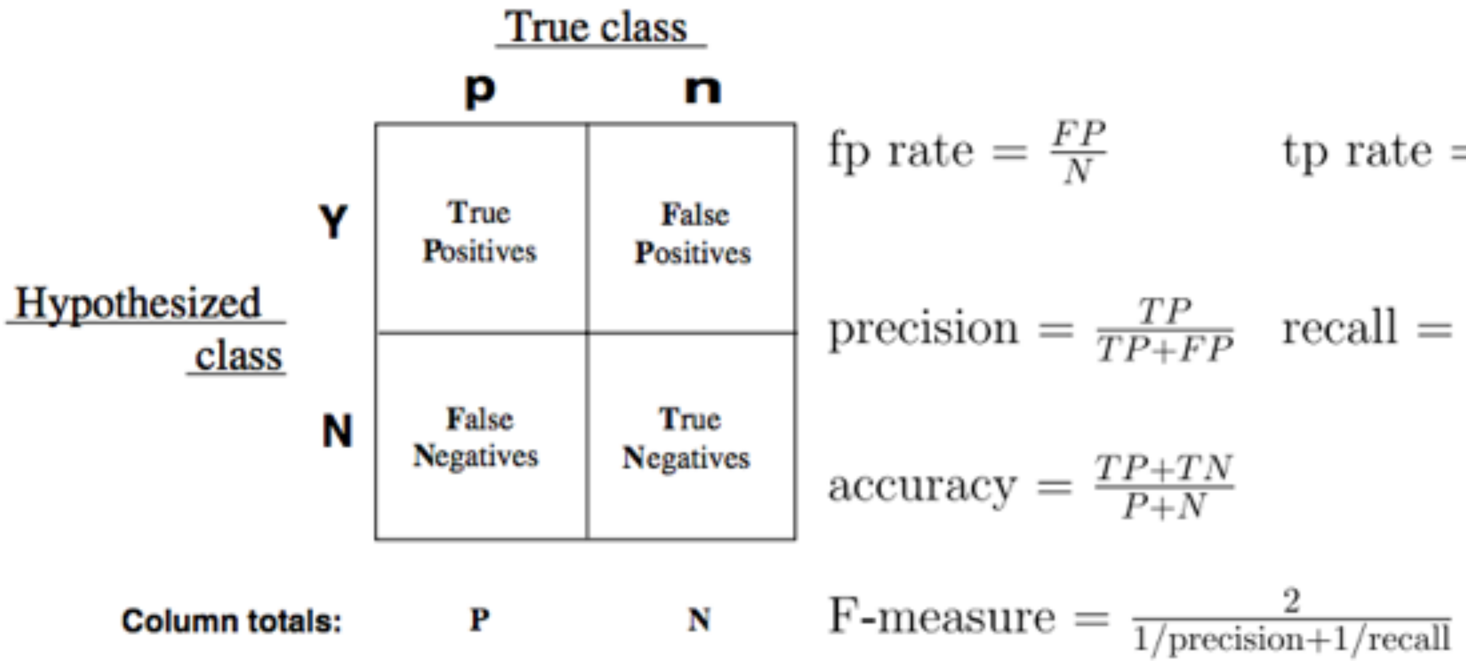
- Examine precision and recall, adjust model depending on preferences for those

- See if your model is avoiding any classes, and think about why

- See if your model is consistently putting one category in another

EXPLAIN

- Share true positive and false positive rates
- Indicate how frequently model was correct (true positives, true negatives)
- Discuss precision and recall (coming up)



True class

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

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

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

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

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

CONSIDERATIONS

- Precision and recall, bias and variance - always

COMMUNICATING RESULTS

CLASSIFICATION REPORT

WHEN TO USE IT

- Builds a text report showing the main **classification** metrics.

```
             precision    recall  f1-score   support

    class 0       0.67      1.00      0.80         2
    class 1       0.00      0.00      0.00         1
    class 2       1.00      1.00      1.00         2

avg / total       0.67      0.80      0.72         5
```

INTERPRET

- Recall is the ratio of a number of events you can **correctly predict** to a number of **all correct events**.

- Precision is the ratio of a number of events you can **correctly predict** to **all events you predict** (mix of correct and wrong recalls). In other words, it is how precise is your recall.

- If a machine learning algorithm is good at recall, it doesn’t mean that algorithm is good at precision. That’s why we also need F1 score which is the (harmonic) mean of recall and precision to evaluate an algorithm.

VISUALIZE

- Usually you would just pull out the overall precision and recall, and explain them, rather than presenting the actual classification report
- If you're going to print the classification report, make sure to highlight the values you will be discussing

EXPLAIN

- Explain your balance between precision and recall, why you prioritized which, and how well your model did at leaning in the right direction.

COMMUNICATING RESULTS

PLOTTING MODELS

WHEN TO USE IT

- 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.

- For classification, we can work with ROC-AUC and precision-recall plots.

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

    - For precision-recall plots, it'll depend on the cost requirements; either a model will have good recall at the cost of precision, or visa versa.

AUC

- AUC just means “area under the curve”

- This is a metric that is relevant in BOTH ROC and precision-recall curves

- Just because you hear it most often with ROC (as in ROC-AUC), doesn’t mean you don’t care about the area under the precision recall curve

- Typically:

  - For ROC-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 (essentially, the point where the curve is farthest from the baseline)

COMMUNICATING RESULTS

ROC-AUC

WHEN TO USE IT

- Total AUC is good measure of how well your model delineates between classes

- Justifying the trade off you chose (between whatever your curve’s axes are)
  - ROC curves: True Pos / False Pos
  - Precision - Recall: Precision / Recall

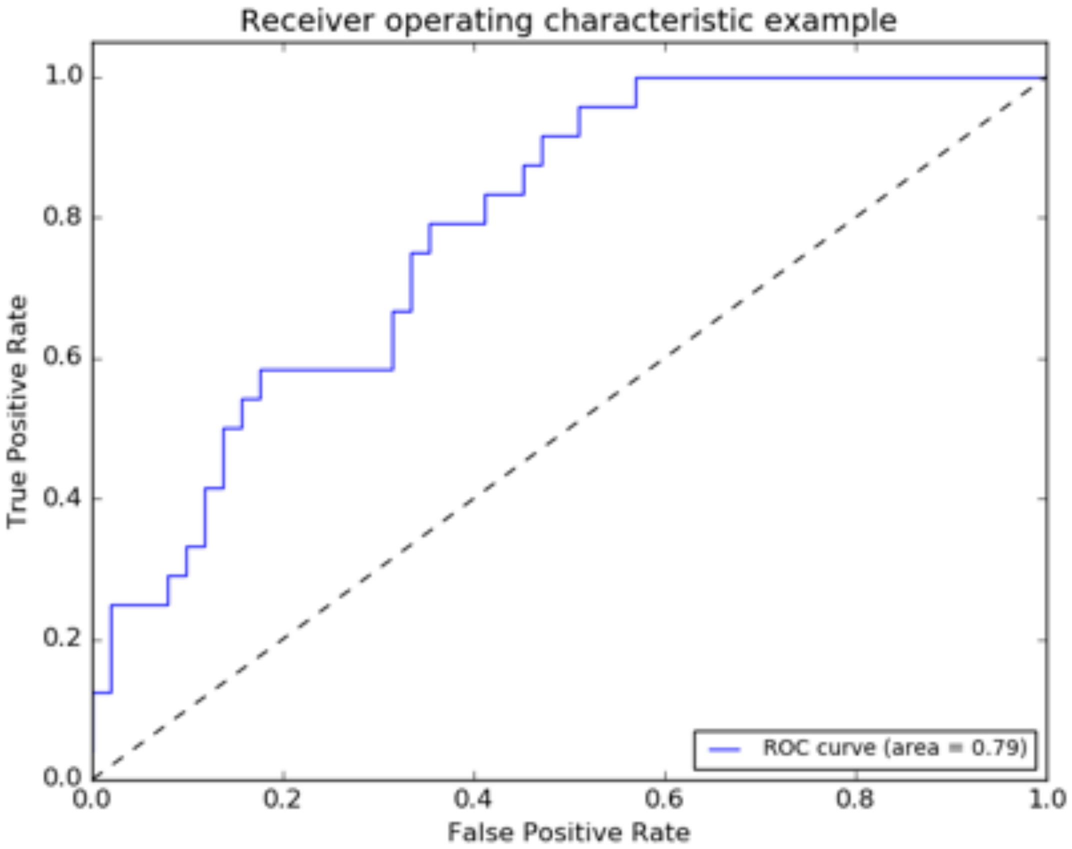
ROC CURVES

- The function roc_curve computes the receiver operating characteristic curve, or ROC curve.

- These curves show the change in performance of your model as you adjust the ‘discrimination threshold’ (decision boundary) when balancing recall and precision

- It is created by plotting the fraction of true positives out of the positives (TPR = true positive rate = sensitivity) vs. the fraction of false positives out of the negatives (FPR = false positive rate = 1-specificity (TNR)), at various threshold settings.

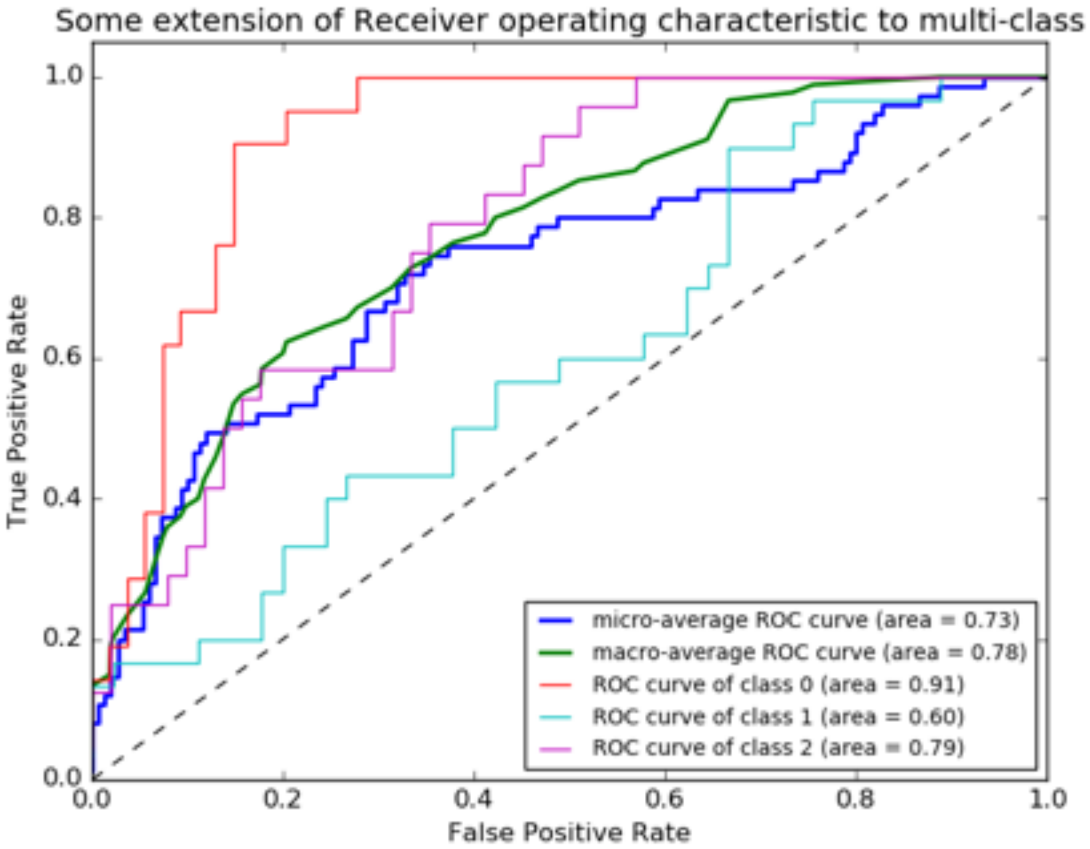
- This function requires the true binary value and the target scores, which can either be probability estimates of the positive class, confidence values, or binary decisions.



ROC CURVES

- In multi-label classification:
  - calculate one vs. rest roc_score for each target class
  - plot this
  - average across roc_scores
  - plot this (macro average)

- The roc_auc_score function can also be used in multi-class classification, if the predicted outputs have been binarized.



INTERPRET

- For ROC-AUC curve, look at TOTAL area under curve
  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. Try adding additional features (such as time of day) and compare models.

- For precision-recall curve, look at POINT where curve is farthest from baseline

CONSIDERATIONS

- Do we always want the exact middle? The exact high/low point?
- Why might it be useful to show different models, instead of just our final one?

COMMUNICATING RESULTS

PRECISION / RECALL CURVE

WHEN TO USE IT

- When presenting a classification model to other data scientists

- Precision is the ability of the classifier not to label as positive a sample that is negative, and recall is the ability of the classifier to find all the positive samples.

- Several functions allow you to analyze the precision, recall and F-measures score:
    - average precision score, f1 score, f beta score, precision score, recall score…
    - We will look at precision recall curve here, but all can be useful
    - We are using this curve because it produces one of the best visuals
    - The precision_recall_curve computes a precision-recall curve from the ground truth label and a score given by the classifier by varying a decision threshold.
    - Note that the precision_recall_curve function is restricted to the binary case

PRECISION/RECALL EXAMPLE

- Imagine we receive an airline flight dataset and we want to predict what causes delayed flights

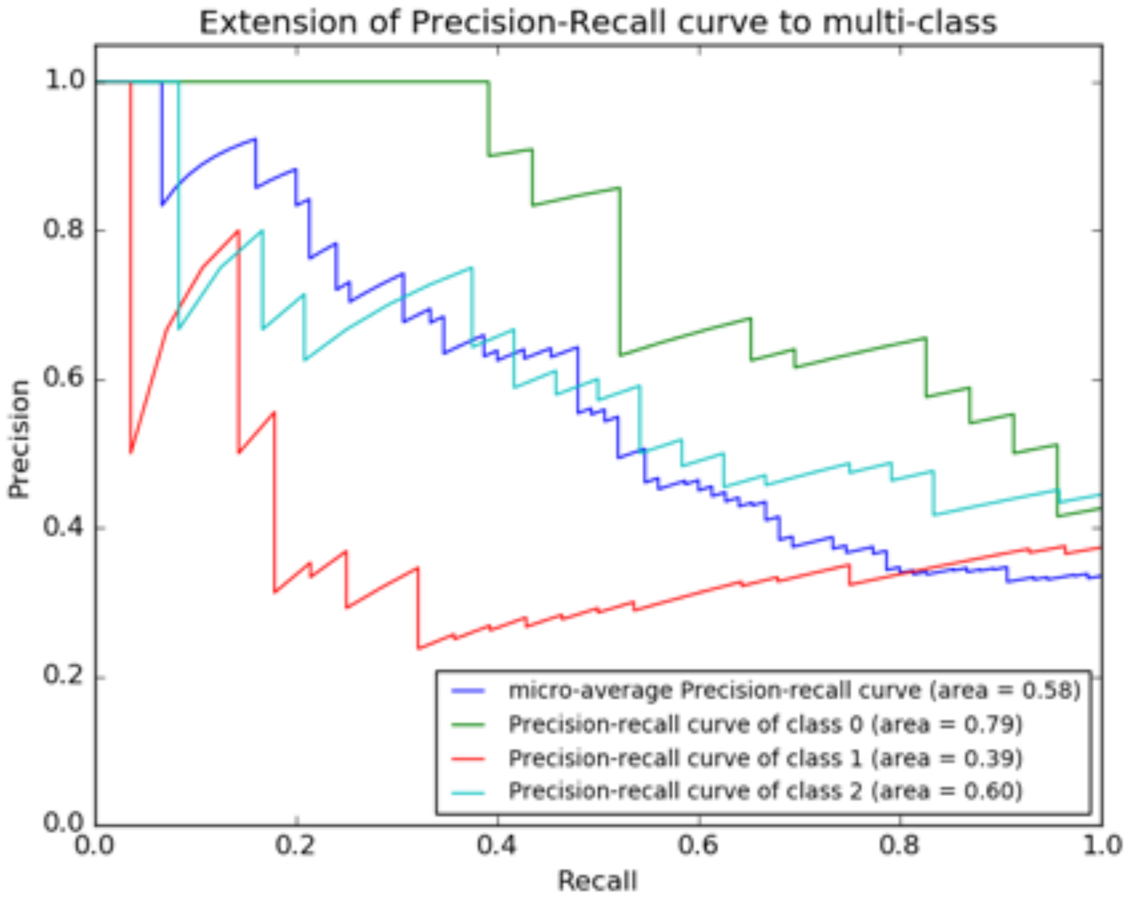
	Optimize toward PRECISION	Optimize toward RECALL
Bias toward	Flights being on time - only a few “problem” observations are labeled as delayed	Identifying all delayed flights
Trade Off	Lower recall, doesn’t give us visibility into problem flight patterns - poor model	Lower precision; we might wrongly predict on-time flights as delayed, artificially lowering the predicted performance of our network

VISUALIZE

- This plot is based on choosing decision line thresholds on the ROC curve. In terms of the model, this would be like moving the decision line for lateness from 0.01 up to 0.99, and then calculating precision and recall at each decision.

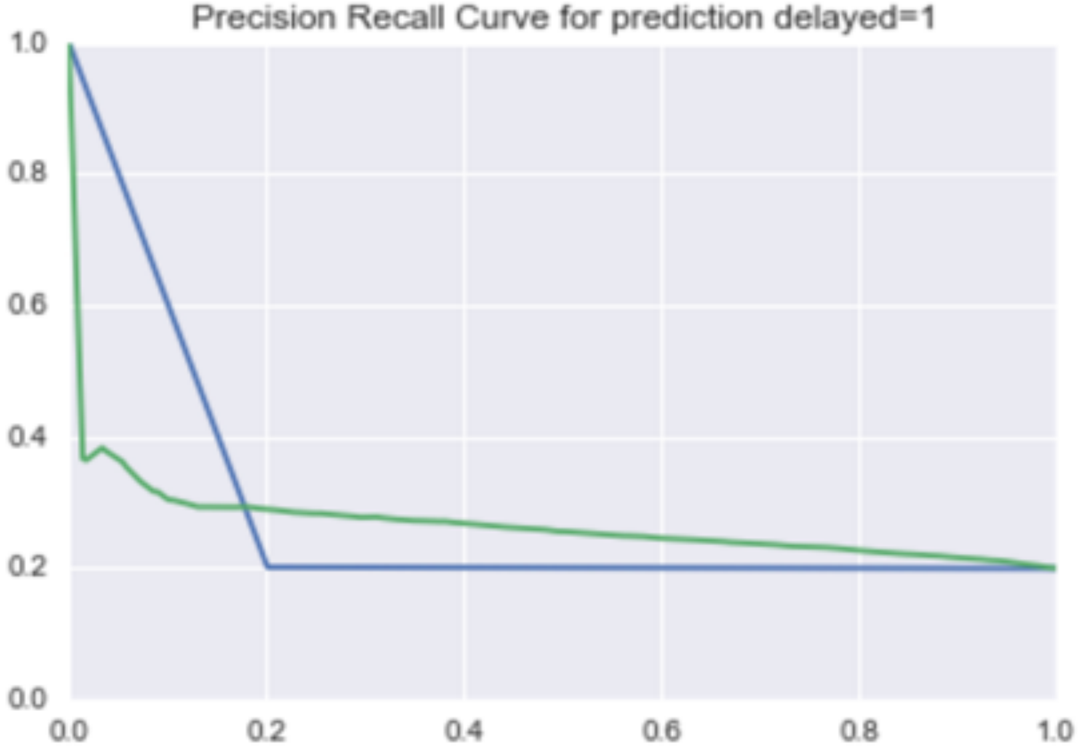
- Keep in mind precision in the first array is returned from the function, but the plot shows it as the y-axis.

- Local maxima are optimal choices, but our tradeoff involves deciding the balance between precision and recall and choosing the appropriate local maximum point



INTERPRET / EXPLAIN

- There are a few interesting takeaways compared to the AUC benchmark (blue):
  - 1. At a lower recall (below .2), there is a noticeable lower precision in the model.
  - 2. Beyond .2 recall, the model outperforms the benchmark.
- Whether we're optimizing for recall or precision, this plot will help us decide based on the .2 threshold.



CONSIDERATIONS

- When would it be better to use an ROC-AUC chart vs a precision-recall chart?