CPSC 330 Lecture 10: Regression Metrics

Varada Kolhatkar

Announcements

- Important information about midterm 1

  - https://piazza.com/class/m01ukubppof625/post/249

  - **Good news for you: You'll have access to our course notes in the midterm!**

- Note the change in the lecture schedule.

- HW5 will be released today. It's a project-type assignment and you get till Oct 28th to work on it.

UBC
Computer
Science

Recap: Confusion matrix

	predicted not Fraud	predicted Fraud
true not Fraud	59700	8
true Fraud	39	63

	predicted not Fraud	predicted Fraud
true not Fraud	TN	FP
true Fraud	FN	TP

- TN → True negatives
- FP → False positives
- FN → False negatives
- TP → True positives

Recap: Precision, Recall, F1-Score
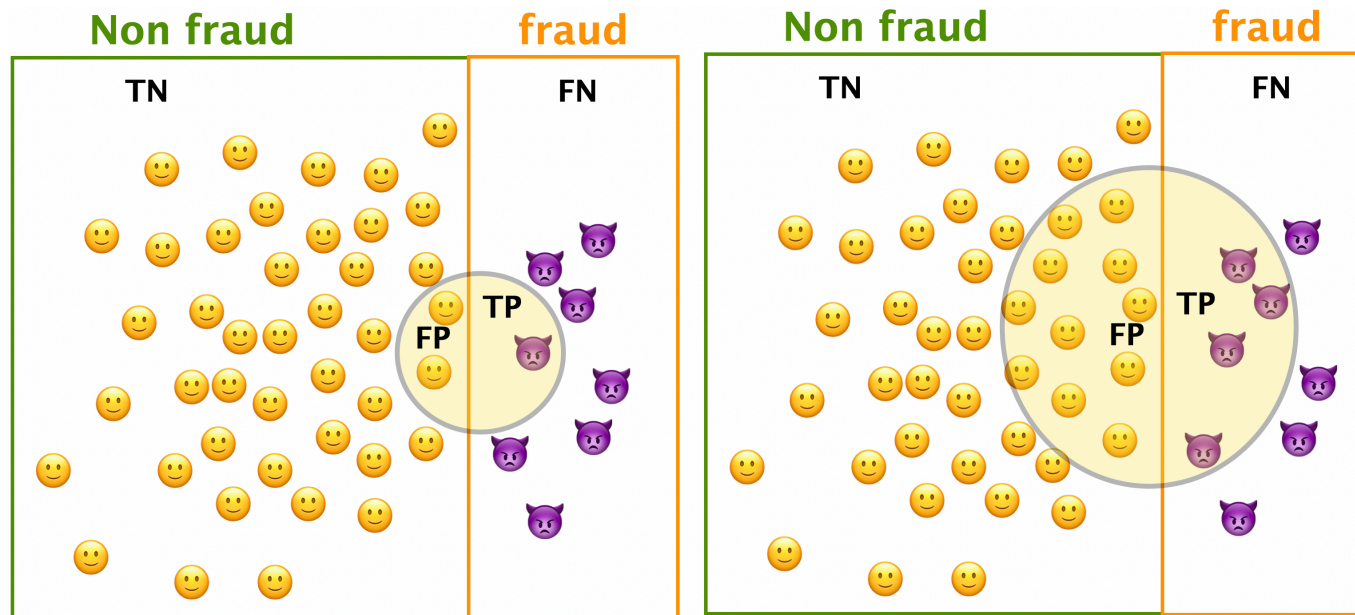
$$f1 = 2 \times \frac{precision \times recall}{precision + recall}$$

$$Precision = \frac{1}{3}$$

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

$$\downarrow Precision = \frac{4}{15}$$
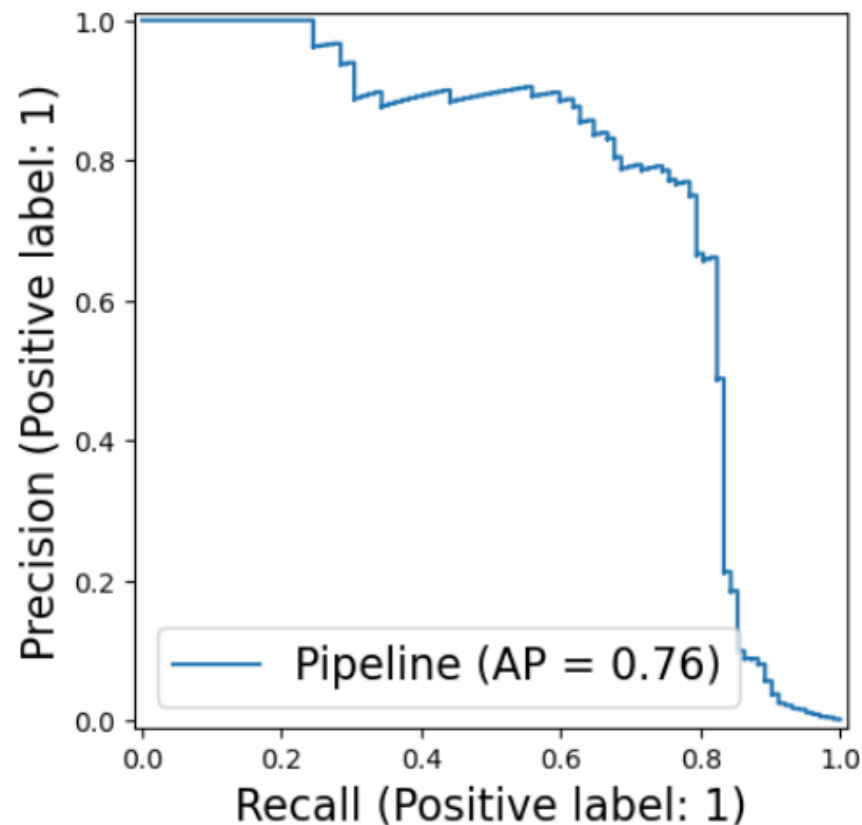
$$Recall = \frac{1}{8}$$

$$Recall = \frac{TP}{TP + FN}$$
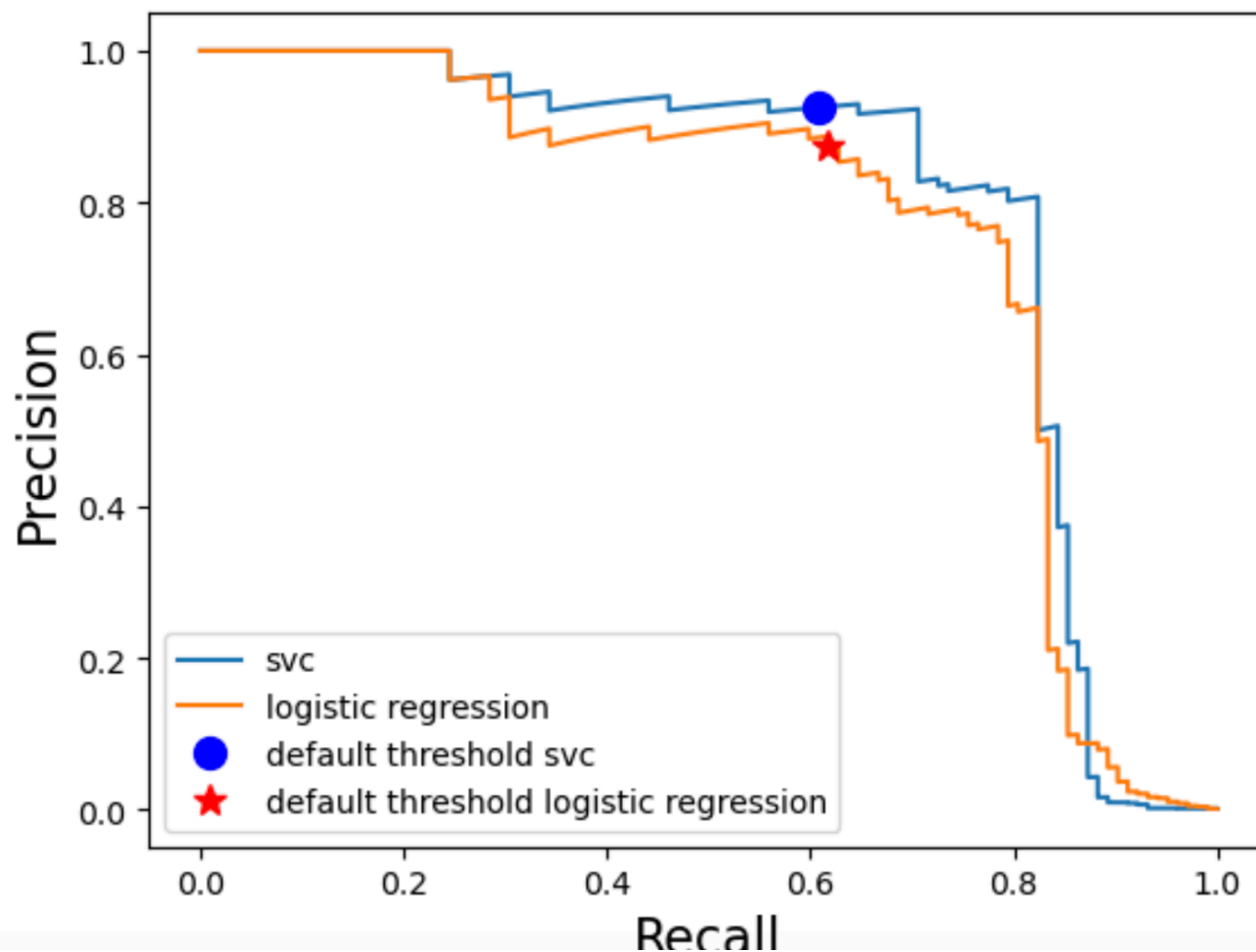
$$\uparrow Recall = \frac{4}{8}$$

Recap: PR curve

- Calculate precision and recall (TPR) at every possible threshold and graph them.

- Better choice for highly imbalanced datasets because it focuses on the performance of the positive class.

Questions for you

- What's the difference between the average precision (AP) score and F1-score?

- Which model would you pick?

Recap: ROC curve

- Calculate the true positive rate (TPR) and false positive rate (FPR) ($\frac{FP}{FP+TN}$) at every possible thresholding and graph TPR over FPR.

- Good choice when the datasets are roughly balanced.
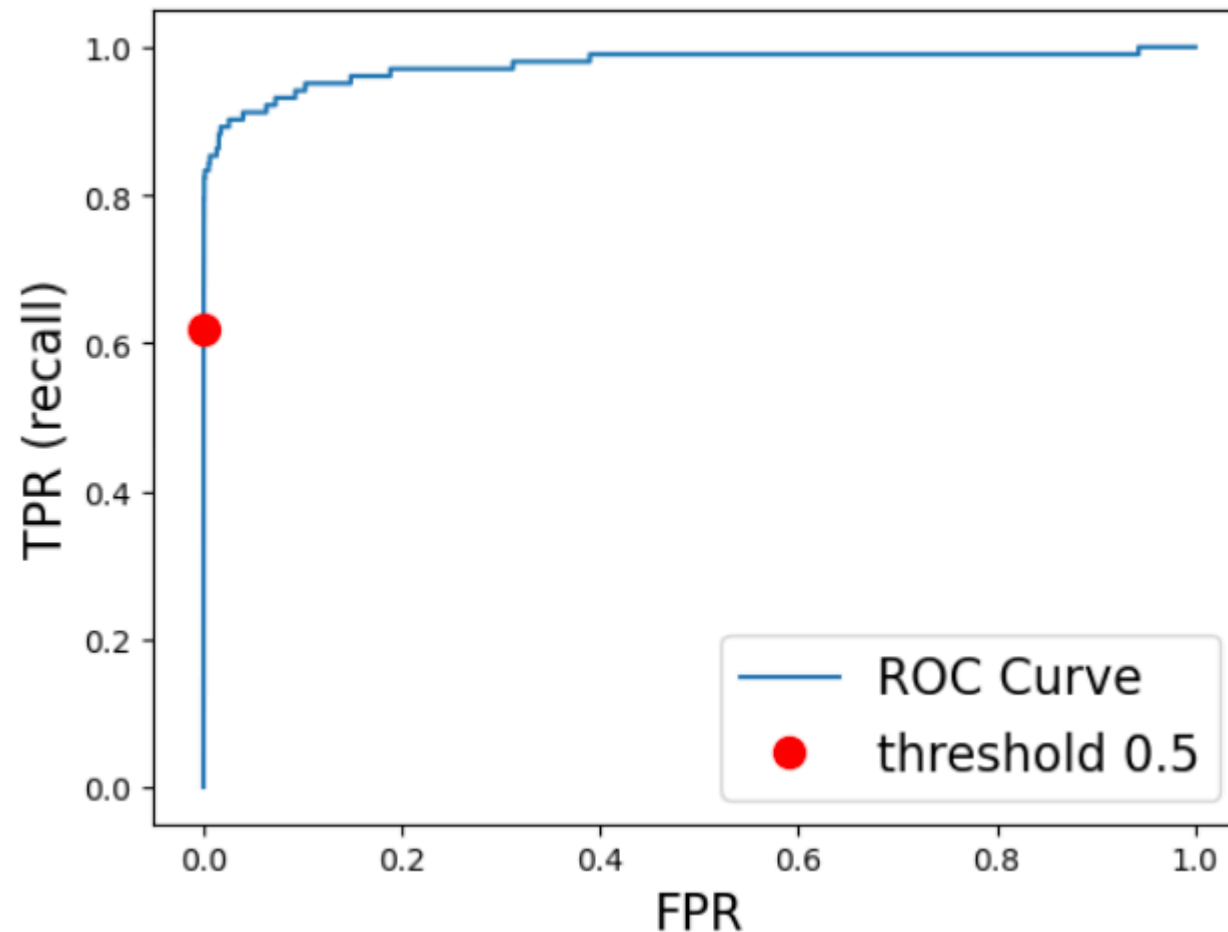
Recap: ROC Curve

- Not a great choice when there is an extreme imbalance because FPR can remain relatively low even if the number of false positives is high, simply because the number of negatives is very large.

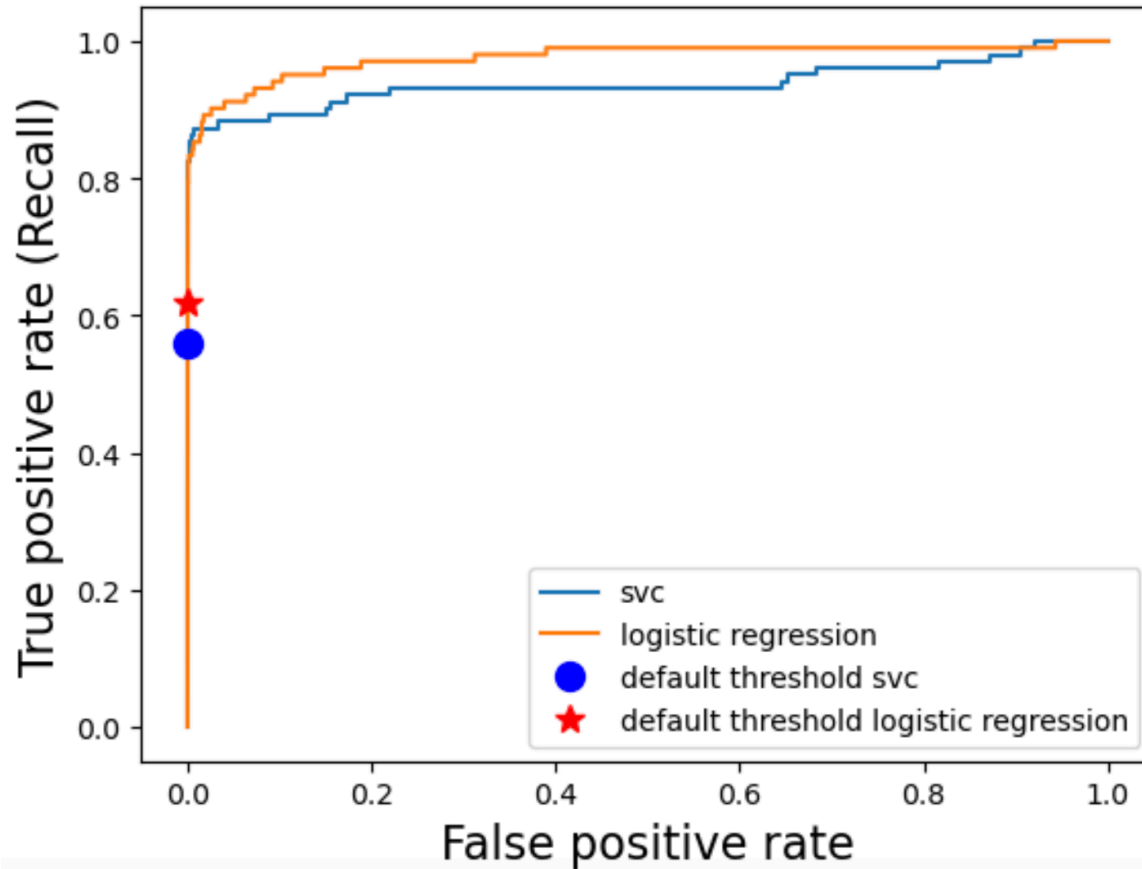$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

- The area under the ROC curve (AUC) represents the probability that the model, if given a randomly chosen positive and negative example, will rank the positive higher than the negative.

UBC
Computer
Science

Questions for you

AUC = 0.5
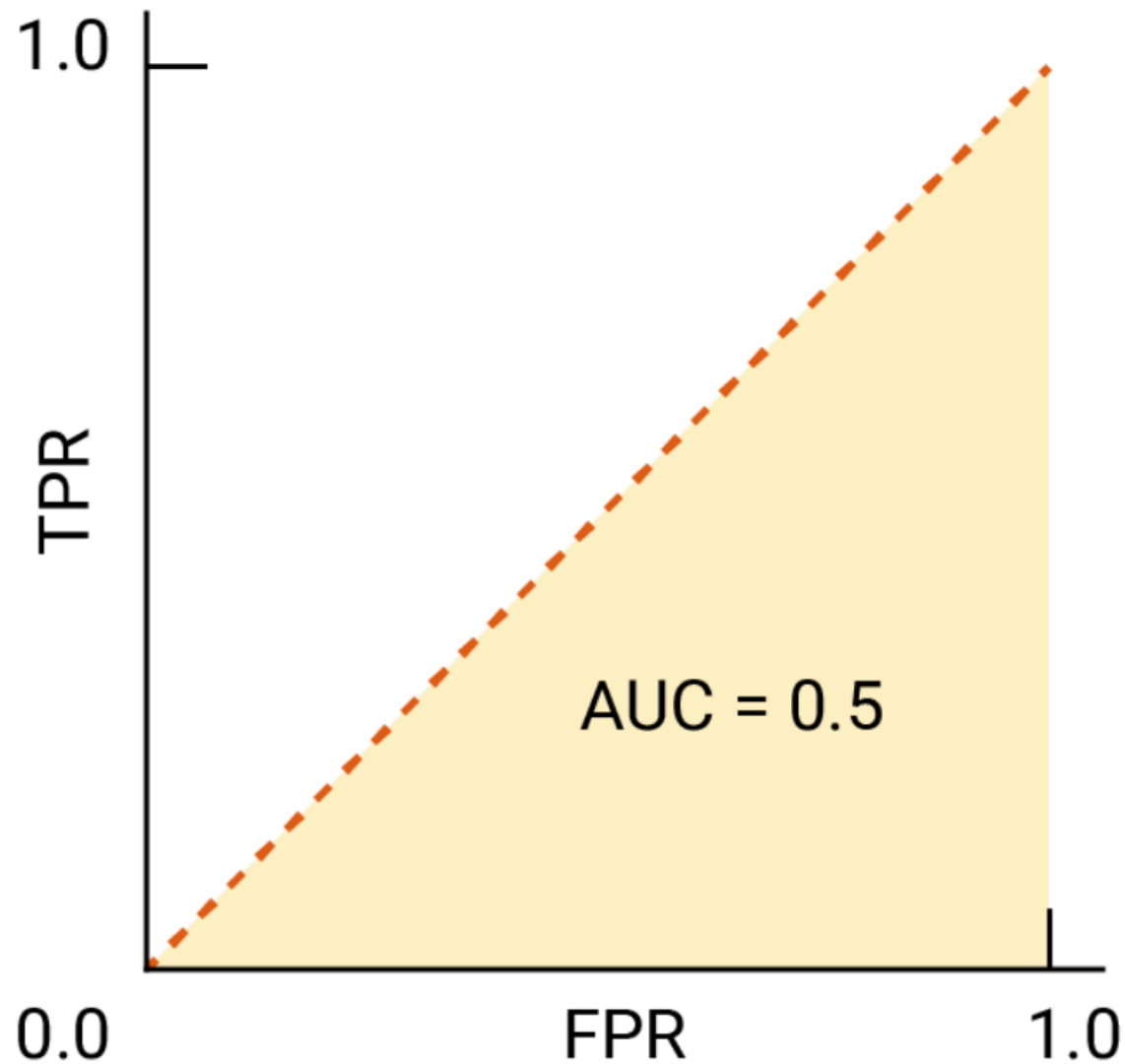
- What's the AUC of a baseline model?

Source

Questions for you

- Which model would you pick?

Dealing with class imbalance

- Under sampling

- Oversampling

- `class weight="balanced"` (preferred method for this course)

- SMOTE

UBC
Computer
Science

Classification metrics class demo

Regression metrics class demo

Ridge and RidgeCV

- **Ridge Regression**: `alpha` hyperparameter controls model complexity.

- **RidgeCV**: Ridge regression with built-in cross-validation to find the optimal `alpha`.

`alpha` hyperparameter

- **Role of `alpha`:**

  - Controls model complexity

  - Higher `alpha`: Simpler model, smaller coefficients.

  - Lower `alpha`: Complex model, larger coefficients.

Regression metrics: MSE, RMSE, MAPE

- **Mean Squared Error (MSE)**: Average of the squares of the errors.

- **Root Mean Squared Error (RMSE)**: Square root of MSE, same units as the target variable.

- **Mean Absolute Percentage Error (MAPE)**: Average of the absolute percentage errors.

Applying log transformation to the targets

- Suitable when the target has a wide range and spans several orders of magnitude
  - Example: counts data such as social media likes or price data
- Helps manage skewed data, making patterns more apparent and regression models more effective.
- `TransformedTargetRegressor`
  - Wraps a regression model and applies a transformation to the target values.

UBC
Computer
Science

iClicker Exercise 10.1

iClicker cloud join link: https://join.iclicker.com/VYFJ

Select all of the following statements which are TRUE.

- a. Price per square foot would be a good feature to add in our `X`.

- b. The `alpha` hyperparameter of `Ridge` has similar interpretation of `C` hyperparameter of `LogisticRegression`; higher `alpha` means more complex model.

- c. In `Ridge`, smaller alpha means bigger coefficients whereas bigger alpha means smaller coefficients.

iClicker Exercise 10.2

iClicker cloud join link: https://join.iclicker.com/VYFJ

Select all of the following statements which are TRUE.

- a. We can use still use precision and recall for regression problems but now we have other metrics we can use as well.

- b. In `sklearn` for regression problems, using `r2_score()` and `.score()` (with default values) will produce the same results.

- c. RMSE is always going to be non-negative.

- d. MSE does not directly provide the information about whether the model is underpredicting or overpredicting.

- e. We can pass multiple scoring metrics to `GridSearchCV` or `RandomizedSearchCV` for regression as well as classification problems.

UBC
Computer
Science