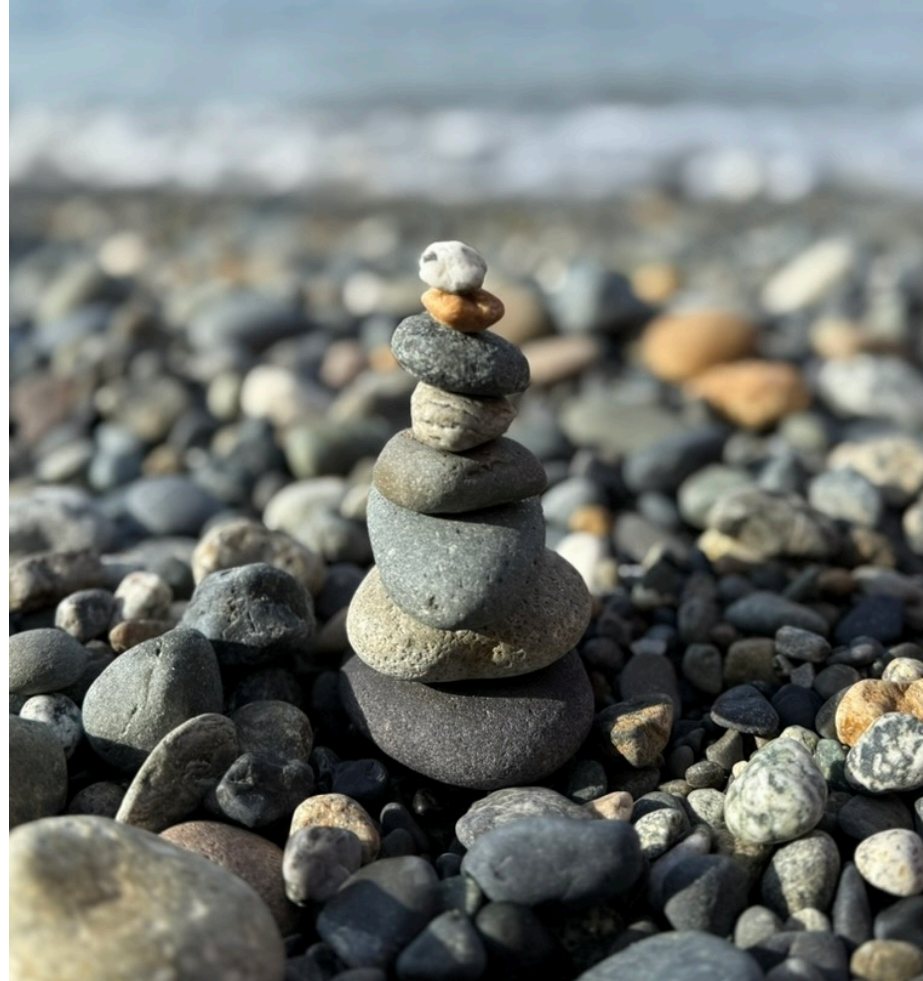


CPSC 330 Lecture 10: Regression Metrics

Varada Kolhatkar

Focus on the breath!



Announcements

- Midterm next week. Important information about midterm 1
 - <https://edstem.org/us/courses/104703/discussion/8335467>
 - **Good news for you: You'll have access to our [course notes](#) and sklearn, pandas, and numpy documentation in the midterm!**
- HW4 was due Monday, Oct 12 at 11:59 pm.
- HW5 is now available and due Oct 26 at 11:59 pm. This project brings together what you've learned. Start early!

Which metric fits best?

Scenario	Data Imbalance	Main Concern	Best Metric(s) / Curve
Email Spam Detection	10% spam	Avoid false positives	
Disease Screening	1 in 10,000	Avoid false negatives	
Credit Card Fraud	0.1% fraud	Focus on rare positive class	
Sentiment Analysis	50/50 balanced	Overall correctness	

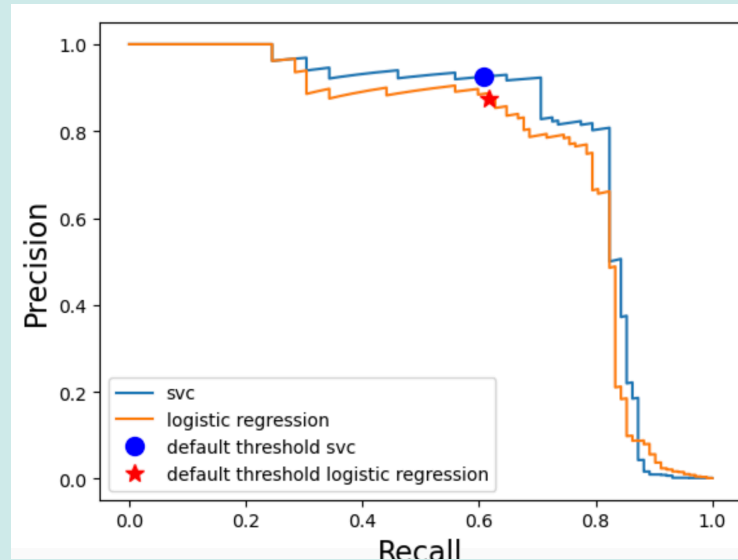
PR curve and AP: Key takeaways

- **PR curve:** x = recall; y = precision. Each threshold gives a precision–recall pair.
- **Precision:** Of the examples flagged positive, what fraction are positive?
- **Recall:** Of all actual positives, what fraction did we find?
- **AP:** A single summary of precision across recall levels; higher is better.

The curve helps choose an operating point; AP summarizes the ranking.

PR curves: Which model would you choose?

- What's the difference between the average precision (AP) score and F1-score?
- Which model would you pick if **high recall** is essential?

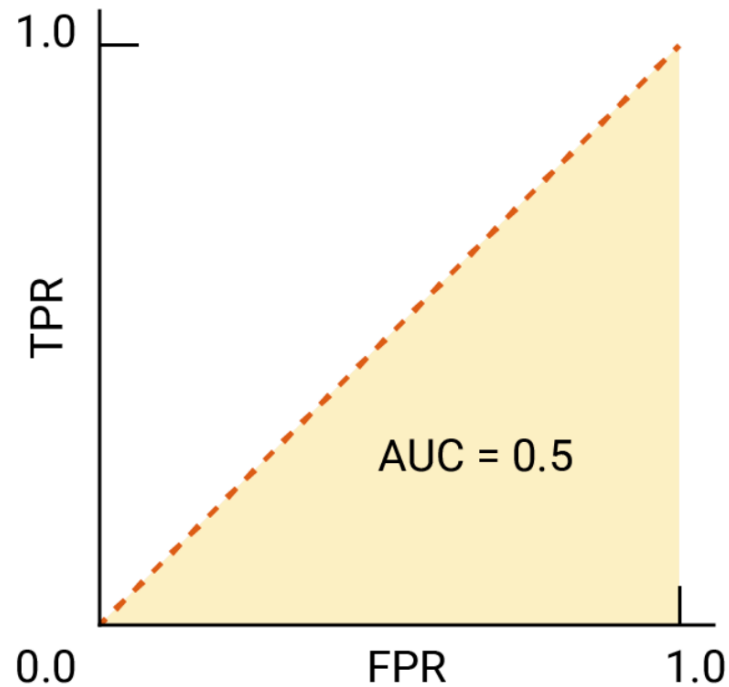


ROC curve and ROC AUC: Key takeaways

- **ROC curve:** x = false-positive rate; y = true-positive rate (recall).
- **TPR:** $TP / (TP + FN)$, the fraction of positives caught.
- **FPR:** $FP / (FP + TN)$, the fraction of negatives incorrectly flagged.
- **ROC AUC:** Area under the ROC curve; higher is better.

FPR uses all actual negatives; precision uses all predicted positives.

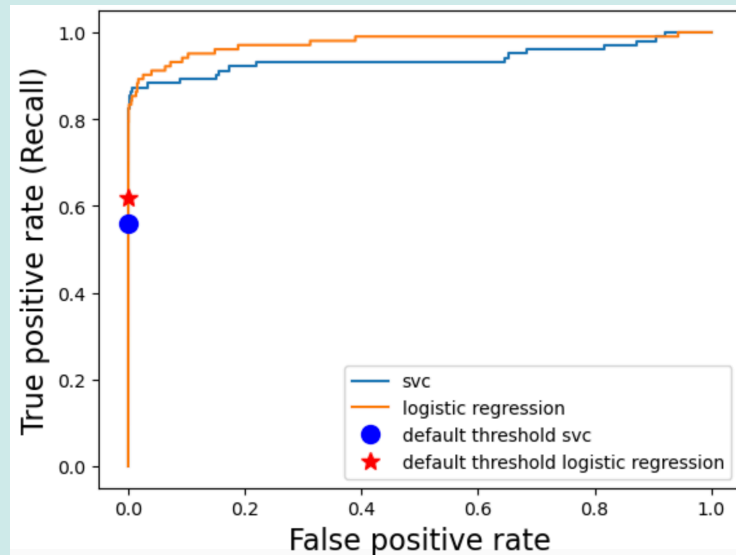
ROC AUC: Baseline and interpretation



- **AUC = 1:** Perfect separation in the evaluated sample.
- **AUC = 0.5:** Baseline ranking reference.
- **AUC < 0.5:** Ranking worse than that reference.

AUC is the probability that a randomly chosen positive scores above a randomly chosen negative.

ROC curves: Which model would you choose?



- Which model would you pick if **very few false alarms** are acceptable?
- Is the model with the largest overall AUC necessarily best there?

Why check precision when positives are rare?

Suppose there are **100 positives** and **9,900 negatives**. At one threshold, **TPR = 80%** and **FPR = 1%**.

- How many true positives and false positives do we get?
- What fraction of alerts are correct?

Summary: Choosing the right metric

Metric / plot	Question it answers
Precision, recall, F1	How good are the decisions at a chosen threshold?
PR curve	What precision can we get at the recall we need?
AP	How well is precision maintained across recall levels?
ROC curve	What fraction of negatives do we flag at the recall we need?
ROC AUC	How well does the model rank positives above negatives overall?

Choose a threshold from the task's costs and constraints. When positives are rare, inspect precision even if ROC AUC is high.

Dealing with class imbalance

- Under sampling
- Oversampling
- `class_weight="balanced"` (preferred method for this course)
- SMOTE

Revision: Changing class weights

`class_weight="balanced"` gives rarer classes more weight during fitting.

For fraud detection, what changes might you expect?

- False negatives and recall?
- False positives and precision?
- Does the validation set become balanced?

Today's big questions

1. How do **MAE**, **RMSE**, R^2 , and **MAPE** describe different aspects of prediction error?
2. How do we choose a metric and use it in **cross-validation** and **model selection**?
3. What changes when we **transform the target**?

Notebook demo: Housing and preprocessing

Predict **SalePrice** from information available for a new house.

- Inspect feature meanings and missing values.
- Distinguish an absent amenity from unknown information.
- Build and inspect a complete preprocessing pipeline.

[Chapter 10 notebook](#)

iClicker Exercise 10.1

Select all statements that are TRUE.

- a. $\text{SalePrice} / \text{GrLivArea}$ is a useful input when predicting a new house's `SalePrice`.
- b. Larger `alpha` in ridge means stronger regularization, opposite to larger `C` in logistic regression.
- c. Preprocessing must be fitted separately within each cross-validation training fold.

Ridge regression and α

- Ridge fits a linear model with a penalty on squared coefficients.
- **Larger α :** stronger regularization; the coefficient vector shrinks overall.
- **Smaller α :** weaker regularization; more freedom to fit the training data.
- Too much regularization can underfit; too little can overfit.

Notebook demo: Baseline and ridge tuning

- Compare `DummyRegressor` and ridge on the same CV folds.
- Search `ridge__alpha` around the **complete pipeline**.
- Inspect training/validation curves and selected coefficients.

Predict: What happens when `alpha` becomes very large?

Mean absolute error (MAE)

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

- Average absolute error, in the **target's units**.
- Each additional dollar of absolute error contributes equally.
- Smaller is better; perfect predictions have $\text{MAE} = 0$.

MAE = \$20,000 means an average absolute error of \$20,000 on the evaluated houses.

MSE and RMSE

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2, \quad \text{RMSE} = \sqrt{\text{MSE}}$$

- Squaring gives **large errors greater influence**.
- MSE has squared target units; RMSE has target units.
- Both have a best value of zero; smaller is better.

A \$20,000 error contributes **four times** as much to MSE as a \$10,000 error.

Your turn: MAE versus RMSE

Absolute errors on three houses:

Model	House 1	House 2	House 3
A	\$0	\$0	\$30,000
B	\$10,000	\$10,000	\$10,000

1. Which model has lower MAE?
2. Which model has lower RMSE?

Notebook demo: Error magnitudes and direction

- Calculate MAE, MSE, and RMSE for the same predictions.
- Inspect observed versus predicted prices.
- Count overpredictions and underpredictions.

Can a low average error hide a pattern of costly mistakes?

Coefficient of determination: R^2

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

The reference predicts the **mean of the evaluation targets**.

- **1**: perfect predictions.
- **0**: same squared error as that reference.
- **Negative**: worse than that reference.

For ridge, `.score(X, y)` returns R^2 by default.

Mean absolute percentage error (MAPE)

$$\text{MAPE} = \frac{1}{n} \sum_i \frac{|y_i - \hat{y}_i|}{|y_i|}$$

A \$30,000 error is 5% of \$600,000, but 50% of \$60,000.

- Use when proportional errors matter and targets are safely away from zero.
- Zero or near-zero targets can produce enormous values.
- sklearn returns a **proportion**: 0.093 means 9.3%.

iClicker Exercise 10.2

Select all statements that are TRUE.

- a. Precision and recall apply directly to continuous regression predictions.
- b. Ridge's default `.score()` agrees with `r2_score(y_true, y_pred)`.
- c. RMSE is always nonnegative.
- d. MSE alone does not show whether a model overpredicts or underpredicts.
- e. Regression searches can record multiple scoring metrics.

Which metric fits the scenario?

Scenario	What matters most?	Primary metric?
House prices, \$60K–\$800K	Error relative to each house's price	
Exam scores, 0–100	Each additional point of absolute error counts equally	
Energy consumption	Large errors deserve disproportionate emphasis	
Insurance claims	Squared error relative to predicting the evaluation mean	

What limitation would you check for each choice?

Regression metrics summary

Metric	Interpretation	Units	Best
MAE	Average absolute error	Target units	0
MSE	Average squared error	Target units squared	0
RMSE	Squared-error emphasis in familiar units	Target units	0
R^2	Squared error relative to evaluation mean	Unitless	1
MAPE	Average absolute relative error	Proportion	0

Choose a **primary metric before comparing models**; inspect errors as well as scores.

Error functions versus sklearn scorers

Metric functions return nonnegative errors: **smaller is better**.

sklearn scorers use **higher is better**, so error scores are negated.

```
1 scoring = {  
2     "MAE": "neg_mean_absolute_error",  
3     "RMSE": "neg_root_mean_squared_error",  
4     "R2": "r2",  
5     "MAPE": "neg_mean_absolute_percentage_error",  
6 }
```

An MAE score of **-20,000** is better than **-30,000**.

Notebook demo: Scoring and selection

- Record multiple metrics with `cross_validate`.
- Search with `GridSearchCV(..., scoring=scoring, refit="MAPE")`.
- Convert negative error scores to positive errors for reporting.

Does selecting with MAPE make ridge fit by minimizing MAPE?

Transforming the target

A log transformation compresses large target values more than small ones.

```
1 ttr = TransformedTargetRegressor(  
2     Ridge(alpha=best_alpha),  
3     func=np.log1p,  
4     inverse_func=np.expm1,  
5 )  
6 ttr_pipe = make_pipeline(preprocessor, ttr)
```

- `.fit()` transforms the training targets.
- `.predict()` returns predictions on the **original scale**.

Notebook demo: Does transforming the target help?

- Inspect price distributions before and after a log transformation.
- Fit the transformed-target pipeline.
- Compare approaches with **training-set CV**, scored in dollars or relative errors.

Reserve the test set for evaluating the final selected workflow.

Exit questions

1. When would you prefer MAE to RMSE?
2. What does a negative R^2 tell you?
3. How does changing the scoring metric differ from log-transforming the target?