Feature Engineering and Selection… Chapter 3: A Review of the Predictive Modeling Process pt. 1

Agenda

3.2 Evaluation metrics

- Regression metrics
- Robust metrics
- Classification (hard -- class prediction)
- Classification (soft – class probabilities)

3.3 Data splitting

- Training/testing
- Stratified sampling

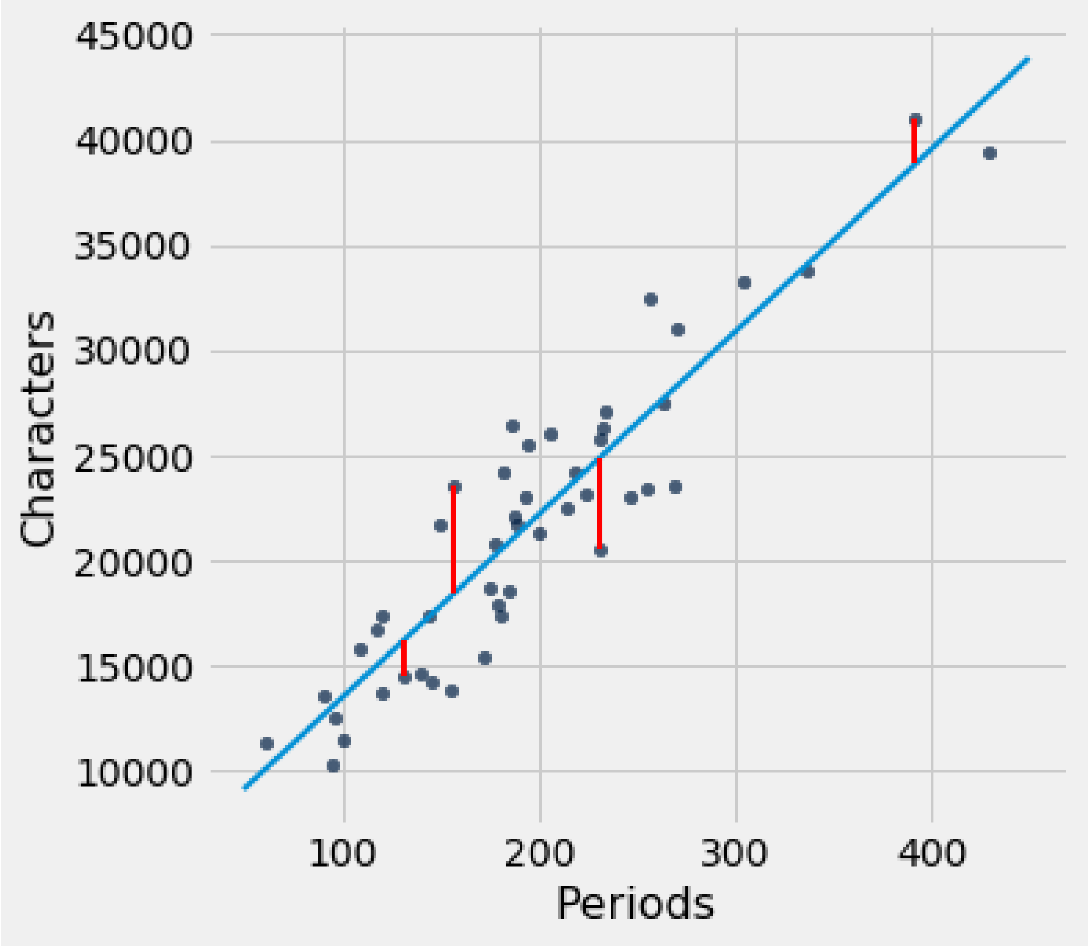
3.4 Resampling

- V-fold cross-validation; Monte Carlo; Bootstrap
- Independence in sampling
- Bias – variance of evaluation metrics
- Information leakage

3.2 Evaluation metrics

“Regression” (continuous target)

- *RMSE* (Root Mean Square Error)
  - *“Average distance of a sample from its observed value to its predicted value.”* (essentially)

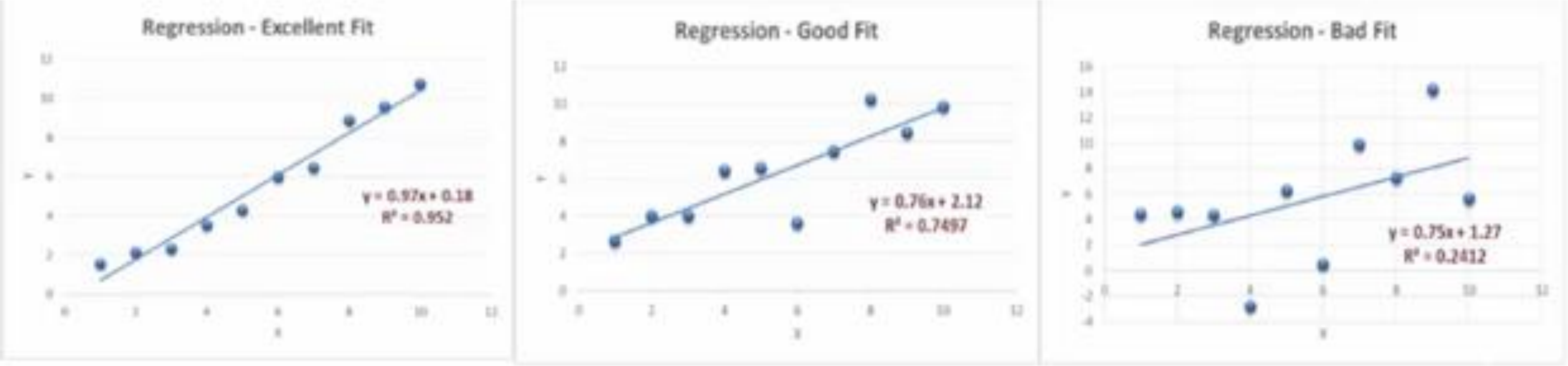


https://www.inferentialthinking.com/chapters/15/3/Method_of_Least_Squares.html

3.2 Evaluation metrics

"Regression" (continuous target)

- R^2 (R squared / coefficient of determination)
  - *"standard correlation between the observed and predicted values (a.k.a. R) and squares it."*

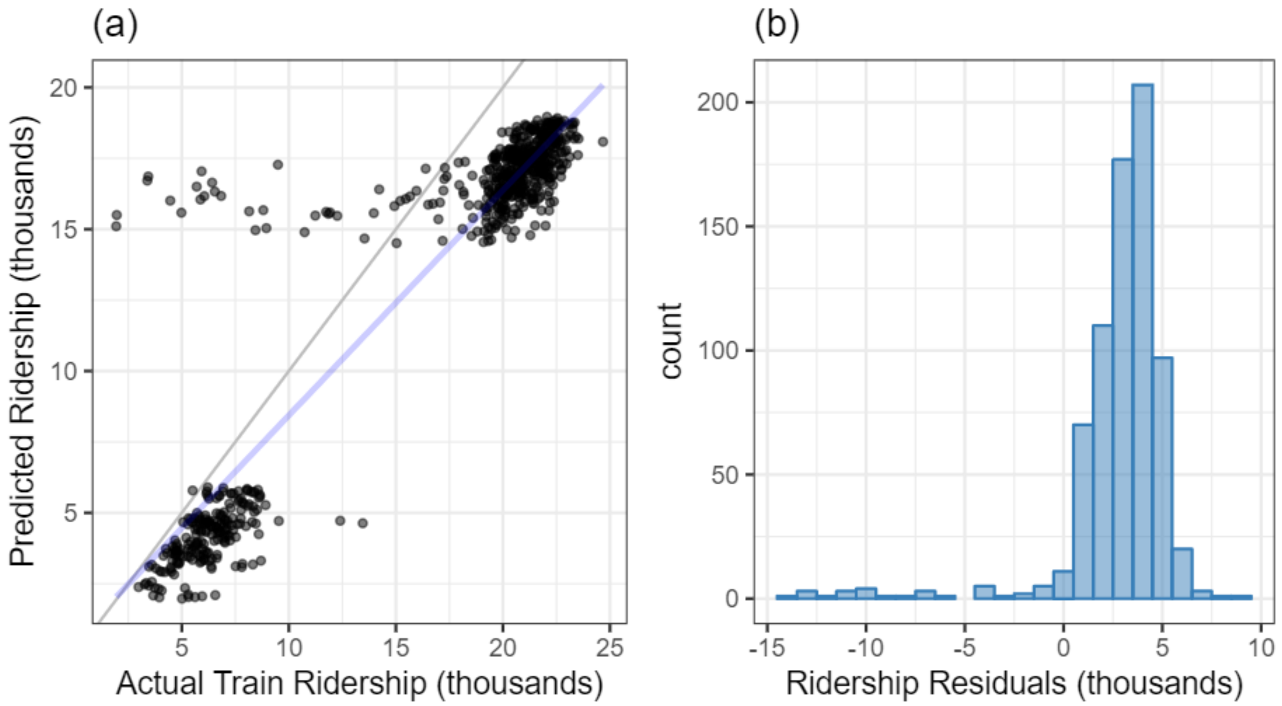


https://www.myaccountingcourse.com/financial-ratios/r-squared

3.2 Evaluation metrics

“Regression” (continuous target)

- R^2 (R squared / coefficient of determination)
  - *“standard correlation between the observed and predicted values (a.k.a. R) and squares it.”*



WARNING:

“Unfortunately, R^2 can be a deceiving metric. The main problem is that it is a measure of correlation and not accuracy.”

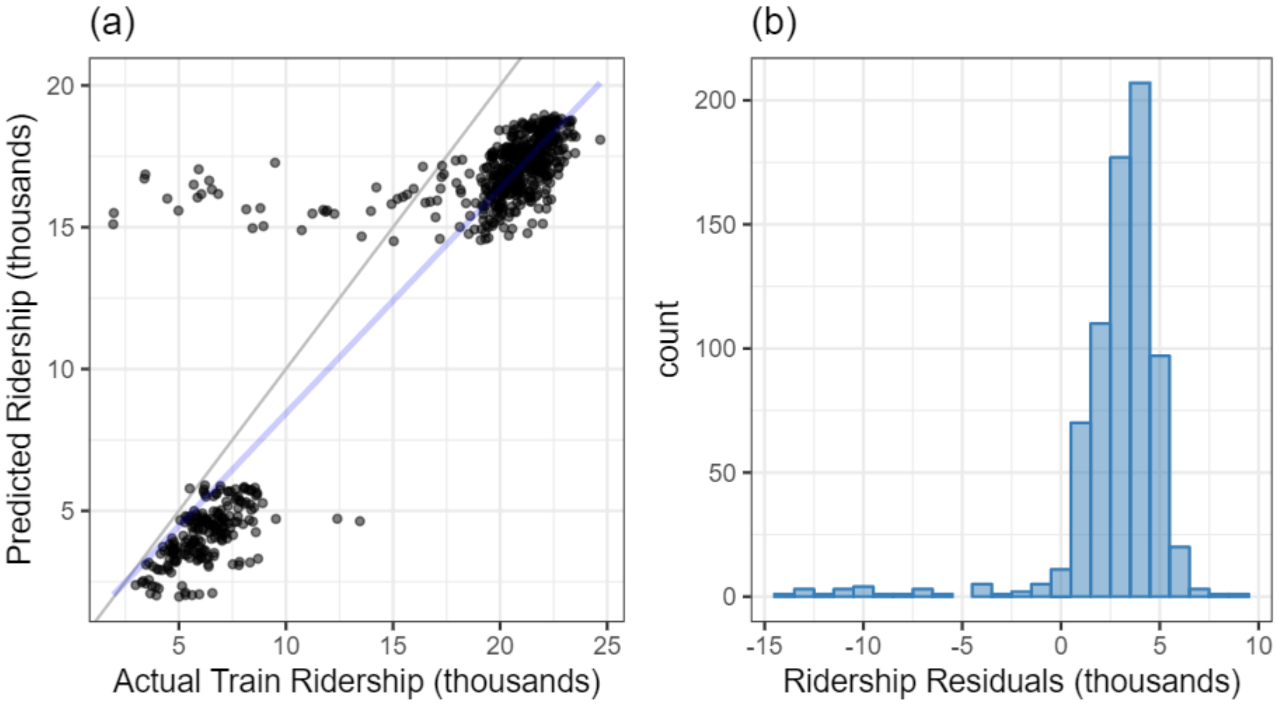
WARNING:

"Unfortunately, R^2 can be a deceiving metric. The main problem is that it is a measure of correlation and not accuracy."

3.2 Evaluation metrics

"Regression" (continuous target)

- R^2 (R squared / coefficient of determination)
  - *"standard correlation between the observed and predicted values (a.k.a. R) and squares it."*



CCC (Concordance Correlation Coefficient)

- "Product of the usual correlation coefficient and a measure of bias from the line of agreement… can be thought of as penalized version of the correlation coefficient… if the relationship between the observed and predicted values is far from the line of agreement"

3.2 Evaluation metrics

“Regression” (continuous target)

Which of these has the highest correlation?

- R^2 (R squared / coefficient of determination)
  - *“standard correlation between the observed and predicted values (a.k.a. R) and squares it.”*
  - *“Proportion of the total variability in the outcome that can be explained by the model.”*

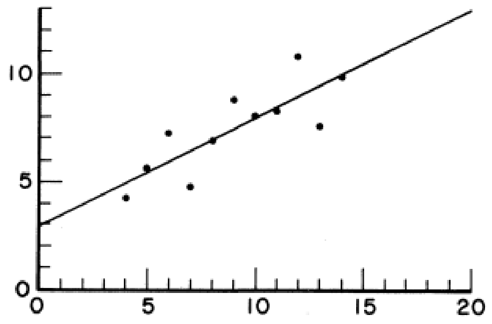


Figure 1

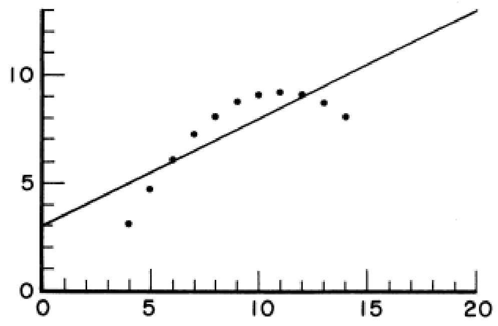


Figure 2

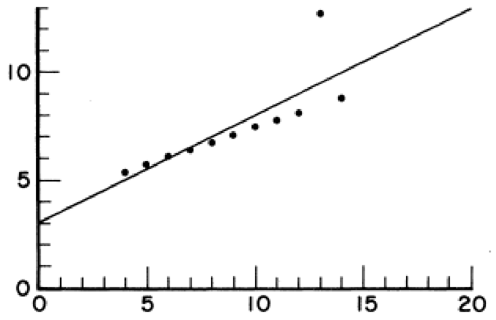


Figure 3

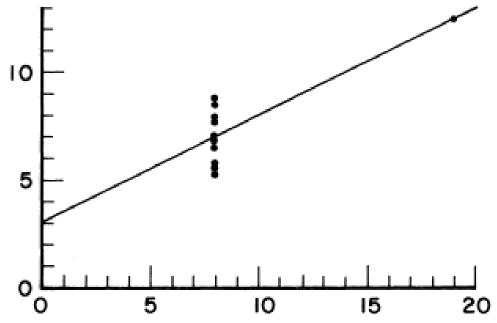


Figure 4

https://socratic.org/questions/is-a-model-with-a-high-r-squared-value-always-better-than-one-with-a-low-r-squar

3.2 Evaluation metrics

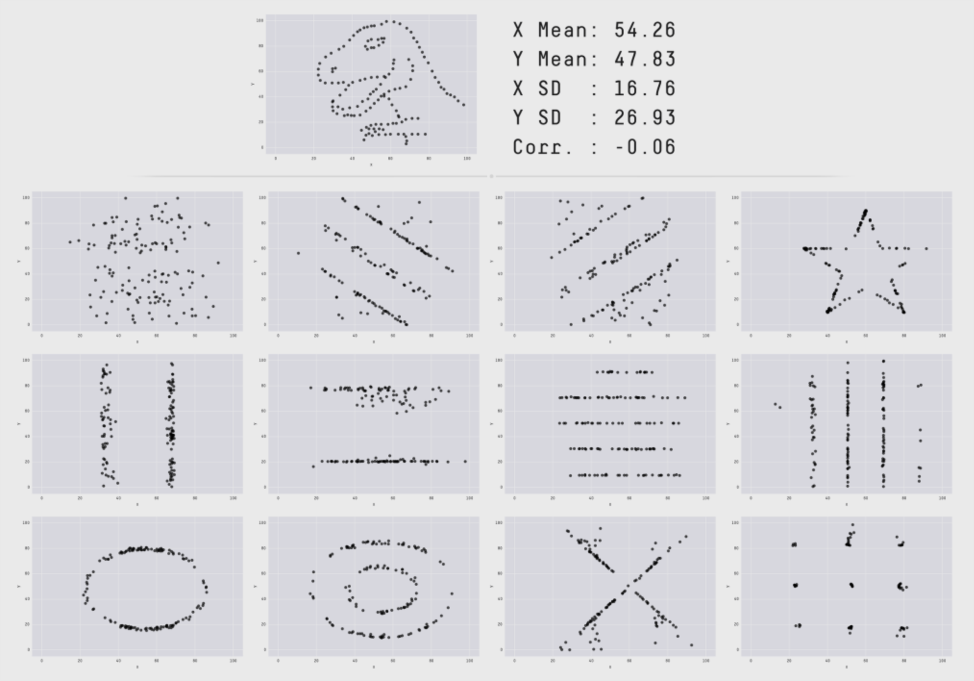
“Regression” (continuous target)

- R^2 (R squared / coefficient of determination)
  - *“standard correlation between the observed and predicted values (a.k.a. R) and squares it.”*
  - *“Proportion of the total variability in the outcome that can be explained by the model.”*

Common tips:

- Visualize your data, as well as your residuals
- Use multiple metrics
- Consider robust metrics…

Which of these has the highest correlation?



https://rweekly.org/2017-19.html

3.2 Evaluation metrics

"Regression" (continuous target)

- *"Both RMSE and R-squared are very sensitive to extreme values because each are based on the squared value of the individual samples' residuals. Therefore a sample with a large residual will have an inordinately large effect on the resulting summary measure."*
- **Robust Techniques**
  - *Try to be insensitive to outliers and extreme values. "Robust techniques seek to find numerical summaries for the majority of the data."*
  - Approaches may…
    - Down-weight extreme values
    - Focus on rank (correlation)
  - Some other metrics…
    - MAD (Median Absolute Deviation)
    - MAE (Mean Absolute Error) Similar interpretation to RMSE (actually more straightforward) – though less popular (math → tougher ; historical reasons)

(Warning: some software flips rows & columns)

3.2 Evaluation metrics

"Classification" (categorical target); class predictions

- Confusion Matrix (many metrics are just ways of combining these outcomes)

TRUTH

PREDICTED

	stem	other
stem	True Positive 5134	False Positive 6385
other	False Negative 2033	True Negative 25257

3.2 Evaluation metrics

"Classification" (categorical target); class predictions

- Confusion Matrix (many metrics are just ways of combining these outcomes)

PREDICTED

TRUTH

	stem	other
stem	True Positive; 5134 5134	False Positive; 6385 6385
other	False Negative; 2033 2033	True Negative; 25257 25257

ACCURACY: Proportion of events and non-events predicted correctly.

Cohen's Kappa: (Range: -1 to 1) Accuracy metric normalized to chance rate (to account for potential class imbalances). -1: worse than chance, 0: chance, 1: perfect

3.2 Evaluation metrics

"Classification" (categorical target); class predictions

- Confusion Matrix (many metrics are just ways of combining these outcomes)

PREDICTED

TRUTH

	stem	other
stem	True Positive; 5134 5134	False Positive; 6385 6385
other	False Negative; 2033 2033	True Negative; 25257 25257

Sensitivity/Recall:
Proportion of events ~~and non-events~~ predicted correctly.

3.2 Evaluation metrics

"Classification" (categorical target); class predictions

- Confusion Matrix (many metrics are just ways of combining these outcomes)

PREDICTED

TRUTH

	stem	other
stem	True Positive; 5134 5134	False Positive; 6385 6385
other	False Negative; 2033 2033	True Negative; 25257 25257

Specificity: Proportion of ~~events and~~ non-events predicted correctly.

3.2 Evaluation metrics

“Classification” (categorical target); class predictions

- Confusion Matrix (many metrics are just ways of combining these outcomes)

PREDICTED

TRUTH

	stem	other
stem	True Positive; 5134 5134	False Positive; 6385 6385
other	False Negative; 2033 2033	True Negative; 25257 25257

Specificity: Proportion of ~~events and~~ non-events predicted correctly.

3.2 Evaluation metrics

"Classification" (categorical target); class predictions

- Confusion Matrix (many metrics are just ways of combining these outcomes)

PREDICTED

TRUTH

	stem	other
stem	True Positive; 5134 5134	False Positive; 6385 6385
other	False Negative; 2033 2033	True Negative; 25257 25257

Precision: proportion of events that are predicted correctly out of the total number of predicted events.

3.2 Evaluation metrics

"Classification" (categorical target); class predictions

- Confusion Matrix (many metrics are just ways of combining these outcomes)

PREDICTED

TRUTH

	stem	other
stem	True Positive; 5134 5134	False Positive; 6385 6385
other	False Negative; 2033 2033	True Negative; 25257 25257

Precision: proportion of events that are predicted correctly out of the total number of predicted events.

Positive Predictive Value (PPV): Equal to precision if you determine "prevalence" based on your sample data… (though typically measure separately)

Assuming this, Negative Predictive Value (NPV) is TN / (TN + FN)

3.2 Evaluation metrics

“Classification”; [class probabilities]

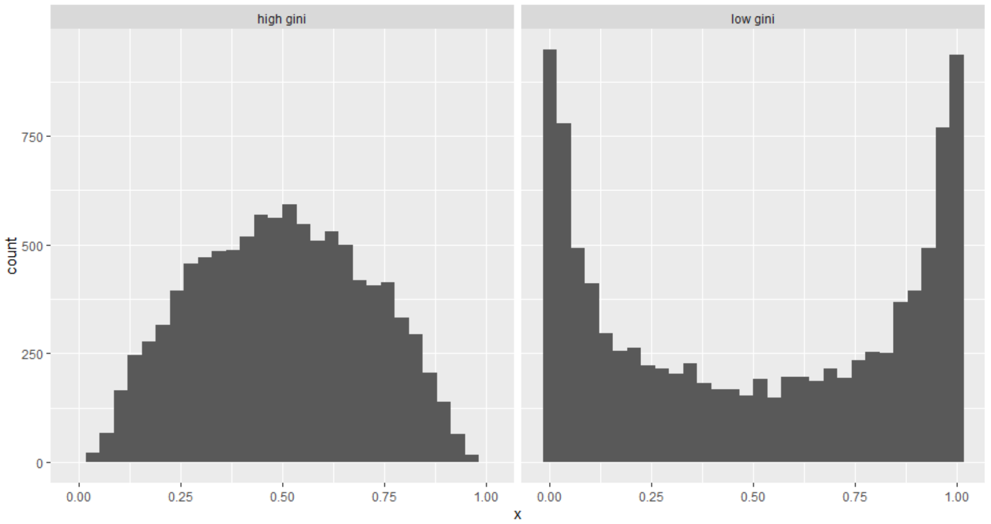
- *“The metrics discussed so far depend on having a hard prediction (e.g., STEM or other). Most classification models can produce class probabilities as soft predictions that can be converted to a definitive class by choosing the class with the largest probability. There are a number of metrics that can be created using the probabilities.”*

3.2 Evaluation metrics

“Classification”; class probabilities

Purity metrics

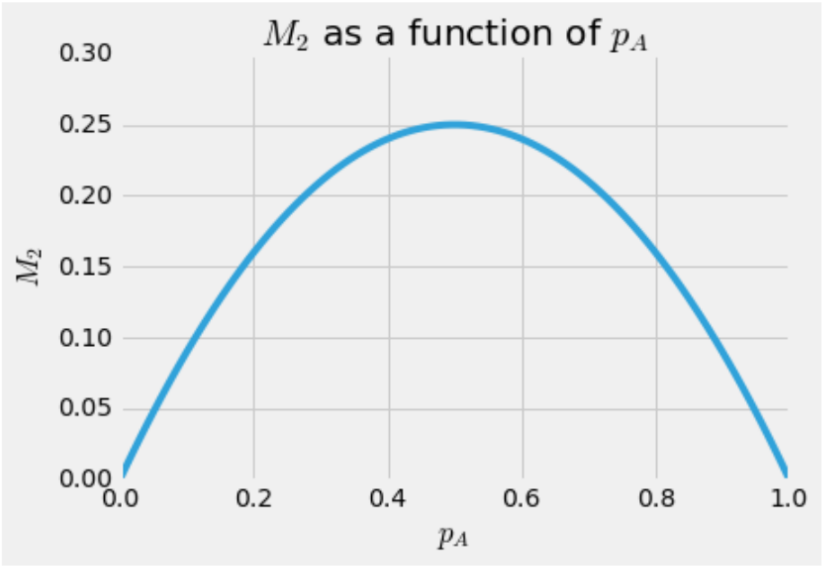
- Unsupervised method – is strictly a measure of the separation in your probabilities (does not vary depending on their accuracy) (e.g. Gini; Entropy)



- Plugging in more extreme probabilities (greater separation / purity) → smaller values
  - Want to minimize these metrics

$$G = \sum_{i=1}^{n} \sum_{j \neq j'} p_{ij} p_{ij'}$$

$$H = -\sum_{i=1}^{n} \sum_{j=1}^{C} p_{ij} \log_2 p_{ij}$$



https://www.quora.com/What-is-the-interpretation-and-intuitive-explanation-of-Gini-impurity-in-decision-trees

3.2 Evaluation metrics

"Classification"; class probabilities

Log-likelihood

- Supervised method – considers actual target classes in scoring
- Goal → maximize
- Magnitude of predicted probabilities matter not just class predicted
- "maximized if all samples are predicted with high probability to be in the correct class"
- More extreme misses are penalized more (on a log scale)

$$\log \ell = \sum_{i=1}^{n} \sum_{j=1}^{C} y_{ij} \log(p_{ij}),$$

https://www.quora.com/What-is-the-interpretation-and-intuitive-explanation-of-Gini-impurity-in-decision-trees

Example:

2 observations and predicted probabilities:

x_1 = 1

P_11 = 0.8

P_10 = 0.2

x_2 = 0

p_21 = 0.6

p_20 = 0.4

(PLUG IN and SUM)

Supervised vs unsupervised metrics performance

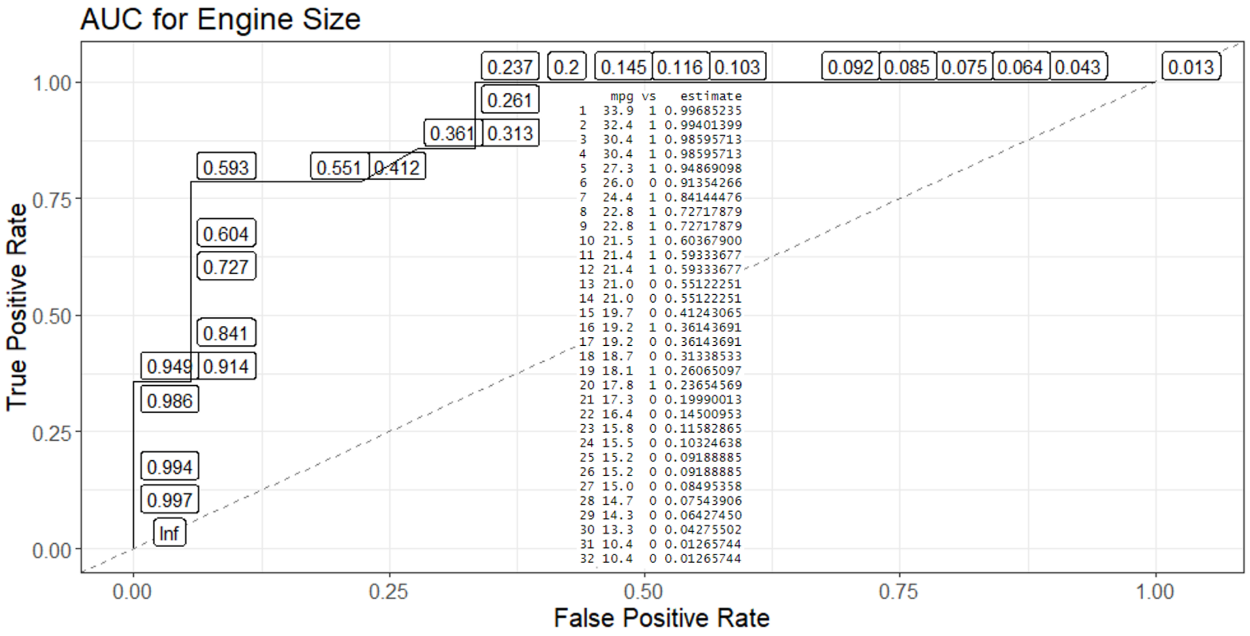
Purity metrics only penalize "equivocal" model

Table 3.2: A comparison of typical probability-based measures used for classification models. The calculations presented here assume that Class 1 is the true class.

	Probabilities		Statistics		
	Class 1	Class 2	Log-Likelihood	Gini	Entropy
Equivocal Model	0.5	0.5	-0.693	0.25	1.000
Good Model	0.8	0.2	-0.223	0.16	0.722
Bad Model	0.2	0.8	-1.609	0.16	0.722

3.2 Evaluation metrics

“Classification”; class probabilities



	mpg	vs	estimate
1	33.9	1	0.99685235
2	32.4	1	0.99401399
3	30.4	1	0.98595713
4	30.4	1	0.98595713
5	27.3	1	0.94869098
6	26.0	0	0.91354266
7	24.4	1	0.84144476
8	22.8	1	0.72717879
9	22.8	1	0.72717879
10	21.5	1	0.60367900
11	21.4	1	0.59333677
12	21.4	1	0.59333677
13	21.0	0	0.55122251
14	21.0	0	0.55122251
15	19.7	0	0.41243065
16	19.2	1	0.36143691
17	19.2	0	0.36143691
18	18.7	0	0.31338533
19	18.1	1	0.26065097
20	17.8	1	0.23654569
21	17.3	0	0.19990013
22	16.4	0	0.14500953
23	15.8	0	0.11582865
24	15.5	0	0.10324638
25	15.2	0	0.09188885
26	15.2	0	0.09188885
27	15.0	0	0.08495358
28	14.7	0	0.07543906
29	14.3	0	0.06427450
30	13.3	0	0.04275502
31	10.4	0	0.01265744
32	10.4	0	0.01265744

AUC (Area Under ROC curve):

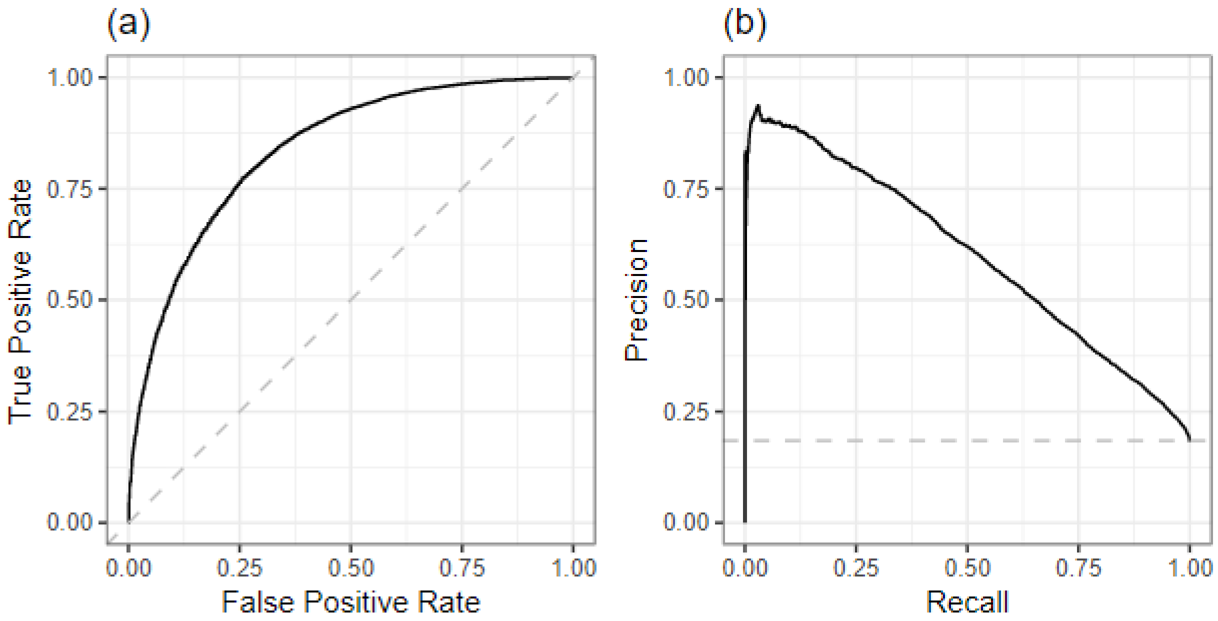
- Supervised method – considers actual target classes in scoring
- ORDER of predicted probabilities matters not just class predicted (magnitude of differences does not matter)
- More extreme misses are penalized more (on a log scale)

https://www.quora.com/What-is-the-interpretation-and-intuitive-explanation-of-Gini-impurity-in-decision-trees

Other notes on classification

1. Start w/ high-level “soft” metrics based on predicted probabilities (e.g. log-likelihood; AUC)

2. Use visualizations of model performance to review subtleties of model (e.g. ROC curve; precision-recall curve



Context specific metrics

- Frequently will need to design your own metrics
- Or modify an existing metric by applying custom weights