# Machine Learning and Forecasting

Seminar in Criminology, Research and Analysis– Crim 7301

Week 11, 11/3/16

Andrew Wheeler

# Class Overview

- Different goals – prediction vs inference
- Overview of different algorithms
  - Regression, e.g. logistic or linear
  - Random forest
  - Boosting and ensemble methods
- Evaluating predictions
  - Hold out sample and cross validation
  - Different cost functions
  - False negatives vs False positives
  - ROC Curve
  - Calibration
- Clinical vs actuarial decision making
- Potential Disparity due to prediction

# Prediction vs Inference

$$\hat{y} = f(x_1, x_2, x_3)$$

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

Machine learning *only* cares about how good a prediction $\hat{y}$ is.

Inference only cares about whether $\beta$ are unbiased.

# Different Algorithms

- Regression is still pretty standard for many problems:

- Logistic to predict classes

$$\hat{p} = \text{Logistic}(X_k \beta_k)$$

- Linear to predict continuous outputs

$$\hat{y} = X_k \beta_k$$

- Can use generalized linear to predict whatever type of outcome, may not give better predictions than linear though....

# Different Algorithms

- Random Forest
  - Generates a bootstrapped sample
  - Creates a decision tree
  - Repeats many times
  - The end prediction is the modal category

- Generalized Boosted Models
  a) Estimate base model
  b) Calculate residuals
  c) Train new model on residuals
  d) Updated model based on base model (a) and new model (c)
  e) Repeat many times

# Different Algorithms

- Other methods
  - SVM [support vector machine] – very similar to logistic regression in practice ($n < p$) [includes non-linear basis functions]
  - K-nearest neighbors
  - Neural networks – very similar to many different logistic regressions, and predict intermediate latent classes [Deep Learning], can predict many different classes

- How to choose each technique:
  - Additive and linear – regression will probably be best (for noisy data this is often true)
  - Highly non-linear and/or many interactions, random forests can work well
  - If you want a non-linear model (e.g. survival), boosted regressions can work well

# Evaluating Predictions

- Hold Out sample, Cross-validation
  - Estimate model on one sample, test the predictions on a new sample

- Different cost functions
  - For continuous inputs, typically try to minimize $(y - \hat{y})^2$, also see Brier Score for probabilities
  - For categorical inputs, try to minimize false positives and false negatives
  - Can give unequal weight though to over-predictions, or try to minimize false negatives

# Evaluating Predictions

| | Predicted Negative | Predicted Positive |
|---|---|---|
| True Negative | Correct Negative | False Positive |
| True Positive | False Negative | Correct Positive |

- False negative rate =

  False Negatives / [False negative + Correct Positive]

- False positive rate =

  False Positive / [Correct negative + False Positive]

# Evaluating Predictions

Cut-Off at predicted value of 0.5

| | Predicted Negative | Predicted Positive |
|---|---|---|
| True Negative | 1,560 | 556 |
| True Positive | 653 | 759 |

FNR = 653 / (653 + 759) = 0.46

FPR = 556 / (556 + 1560) = 0.26

# Evaluating Predictions

Cut-Off at predicted value of 0.2

| | Predicted Negative | Predicted Positive |
|---|---|---|
| True Negative | 357 | 1,759 |
| True Positive | 63 | 1,349 |

$$FNR = 63 / (63 + 1349) = 0.04$$

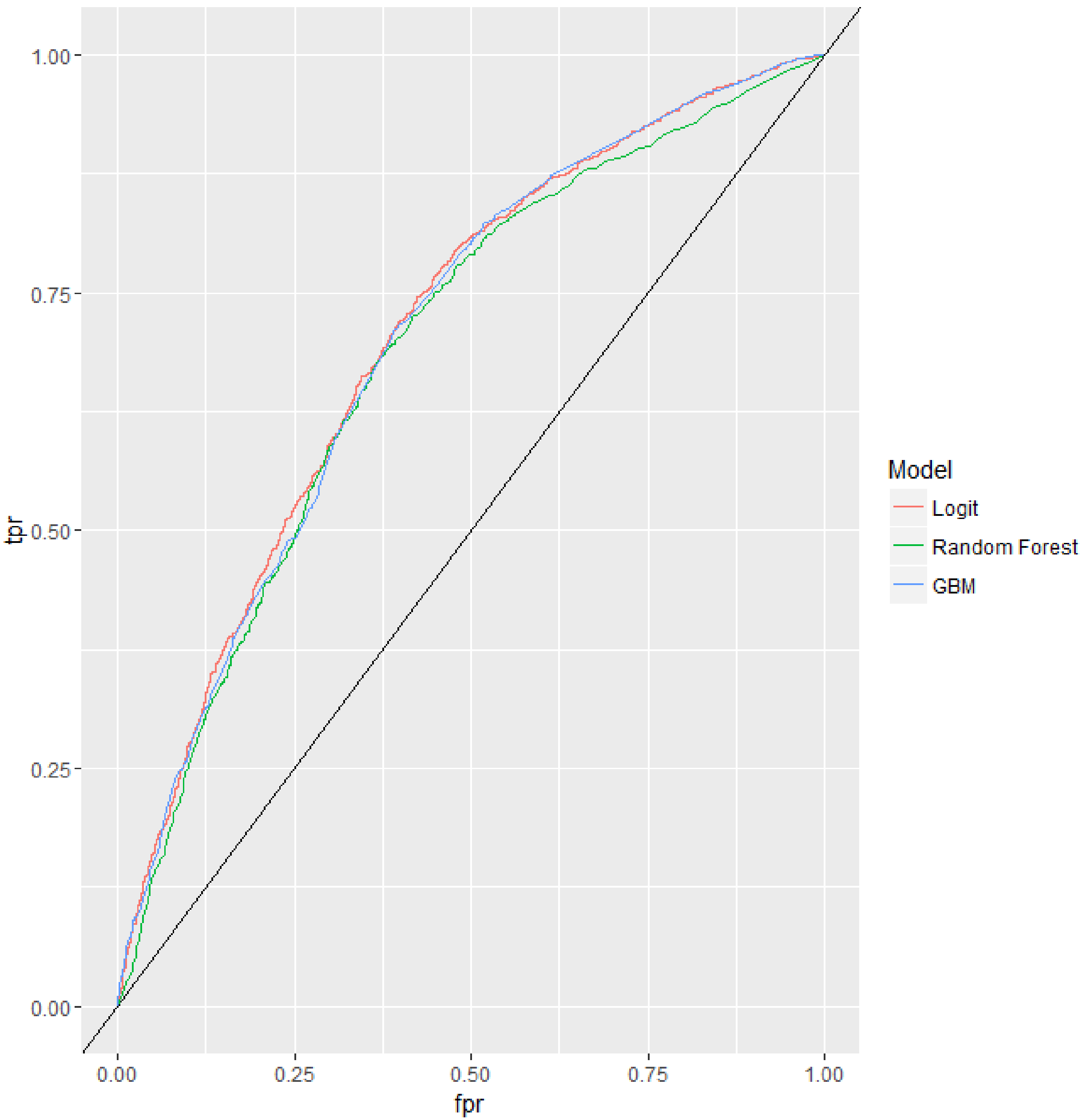
$$FPR = 1759 / (1759 + 357) = 0.83$$

# Evaluating Predictions

- ROC Curve, receiver operating characteristic
  - X axis is the false positive rate (sometimes labelled as 1 – specificity)
  - Y axis is the true positive rate (sometimes labelled as sensitivity)

- Area Under the Curve (AUC)
  - Can be used for model selection
  - Random classifier (taking into account baseline prevalence) has AUC of 0.5

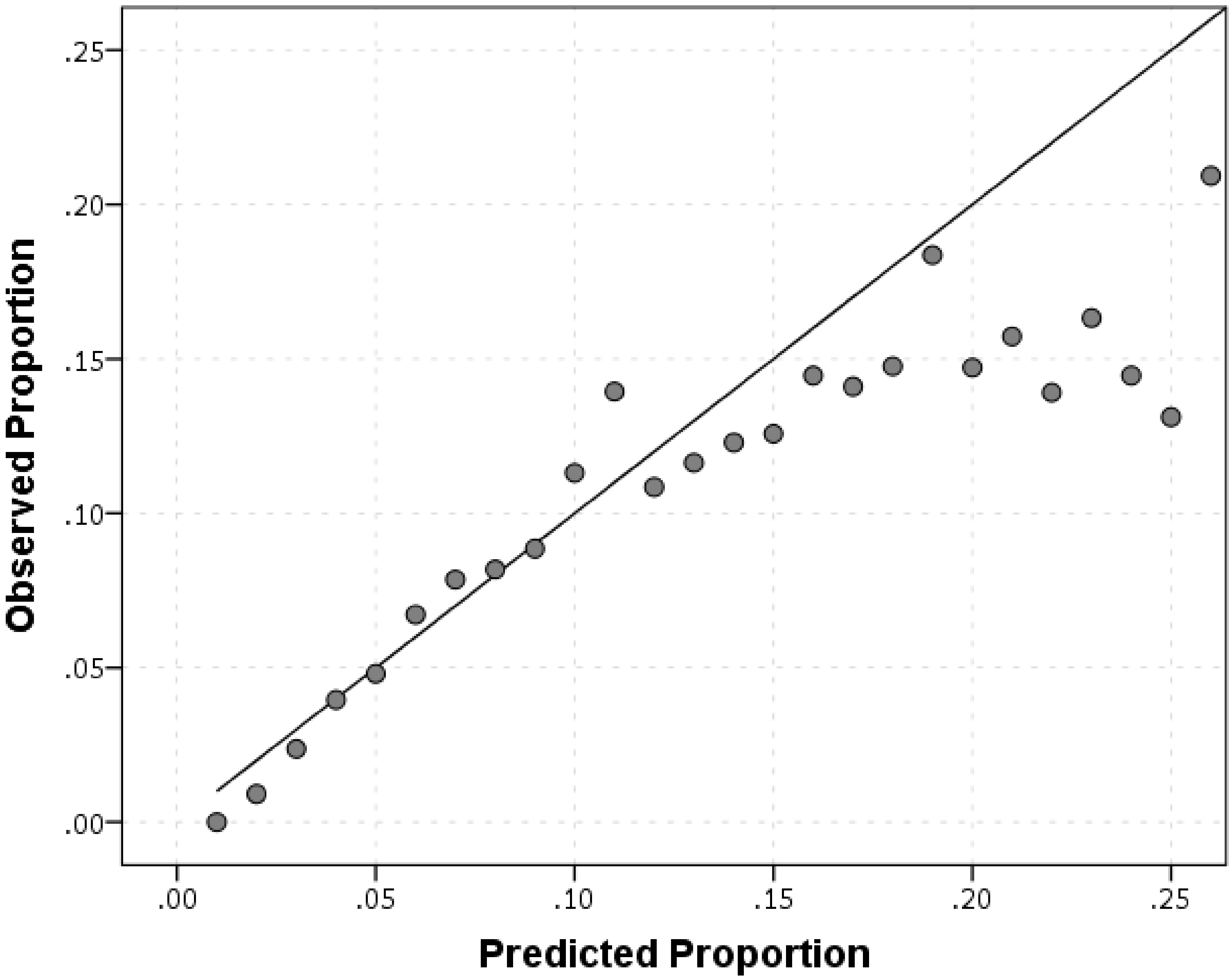
# Evaluating Predictions



# Evaluating Predictions

- Calibration
  - If the model predicts an outcome 5% of the time, does it happen 5% of the time, etc.

# Evaluating Predictions

- 

# Clinical vs Actuarial Decision Making

- Simple actuarial models are **always** better than human opinions
- The robust beauty of linear regression models, can swap out different variables and still get very similar predictions
- The past is the best predictor of the future
- Static vs dynamic indicators
- Can combine the two, see Tetlock's *Superforecasting* or structured clinical decision making

# Potential Disparity

- Many of the machine learning models are black boxes
  - Plausible deniability or ignorance?
- Should you be allowed to include different factors? Gender, Age, Race?
- Even if you don’t include race factors, biases can be carried via other mechanisms
  - Race can be proxied by other factors, such as where you live
  - Biases in one part of the criminal justice system are carried forward to others
  - The training data can only predict past instances

# Homework & Next Weeks Class

**Lab Assignment**

Predict recidivism using logistic regression, random forests, or generalized boosted models. Data taken from ProPublica series on racial bias in machine learning models. Code in R or SPSS (SPSS just calls R programs!). Stata has no machine learning capabilities.

Evaluate predictions using a hold out sample and ROC curves.

**For Next Week**

Come prepared to work on projects, get feedback & help if you need it.