Machine Learning & Business Value

By Kush Patel,
Data Scientist Resident at Galvanize

Outline

- Machine Learning
- Supervised vs Unsupervised
- Linear regression
- Decision Tree Classifier
- Random Forest Classifier
- Cost Benefit matrix
- ROC Curve
- Profit Curves

Machine Learning

“

Machine learning focuses on the development of computer programs that can teach themselves to grow and change when exposed to new data.

Machine learning is a subfield of computer science that evolved from the study of pattern recognition and computational learning theory in artificial intelligence

Machine Learning Technique

Supervised Machine Learning:

- Artificial neural network
- Random Forests
- Boosting
- Naive bayes classifier
- Support vector machines (SVM)
- Nearest Neighbor Algorithm

Unsupervised Machine Learning:

- Clustering (K-mean, hierarchical clustering)
- Blind Signal Separation Technique (PCA, SVD, NMF)

Simple Linear Regression

Definition

Population: The entire pool from which a **statistical** sample is drawn.

Sample: A group drawn from a larger population and used to estimate the characteristics of the whole population.

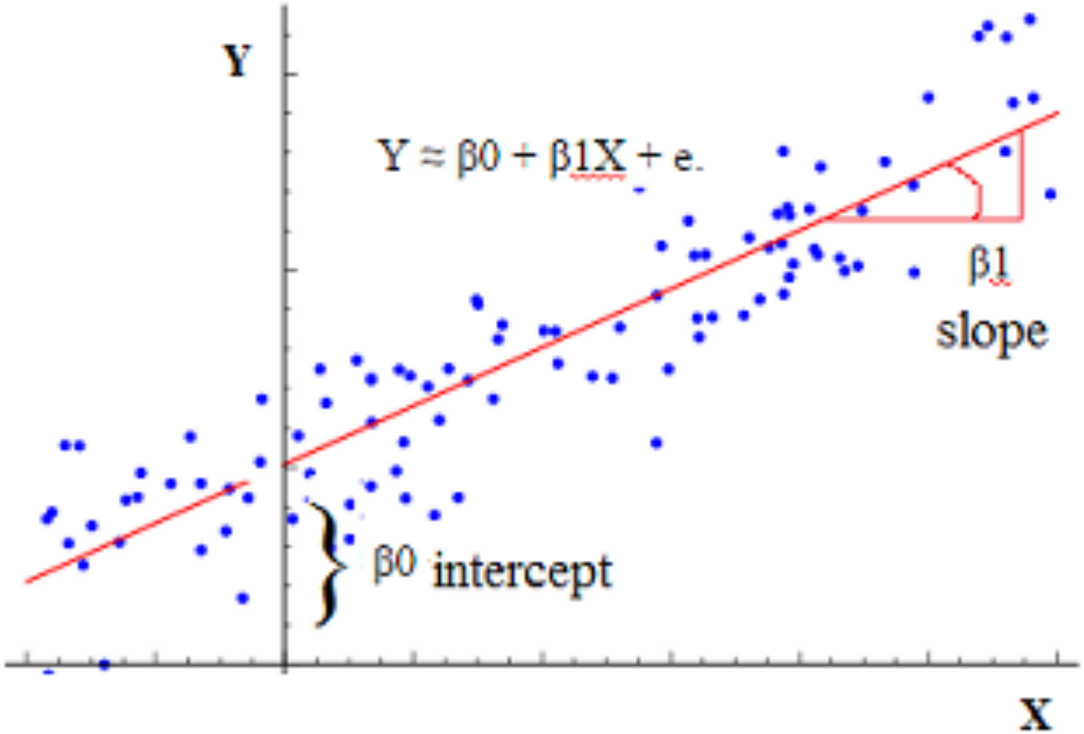
Training Set: The sample which used to train model.

Testing Set: The sample which used to evaluate model

Assumptions

1. Linearity
2. Constant Variance
3. Independence of errors
4. Normality of Errors
5. Lack of multicollinearity

Simple Linear Regression



β_0 is intercept -- constant
β_1 is intercept -- constant
e is error term

Simple Linear Regression

For population:

$Y = \beta_0 + \beta_1 X + e$

For sample:

$\hat{y}$ = estimated(β_0) + estimated(β_1)* x

where:

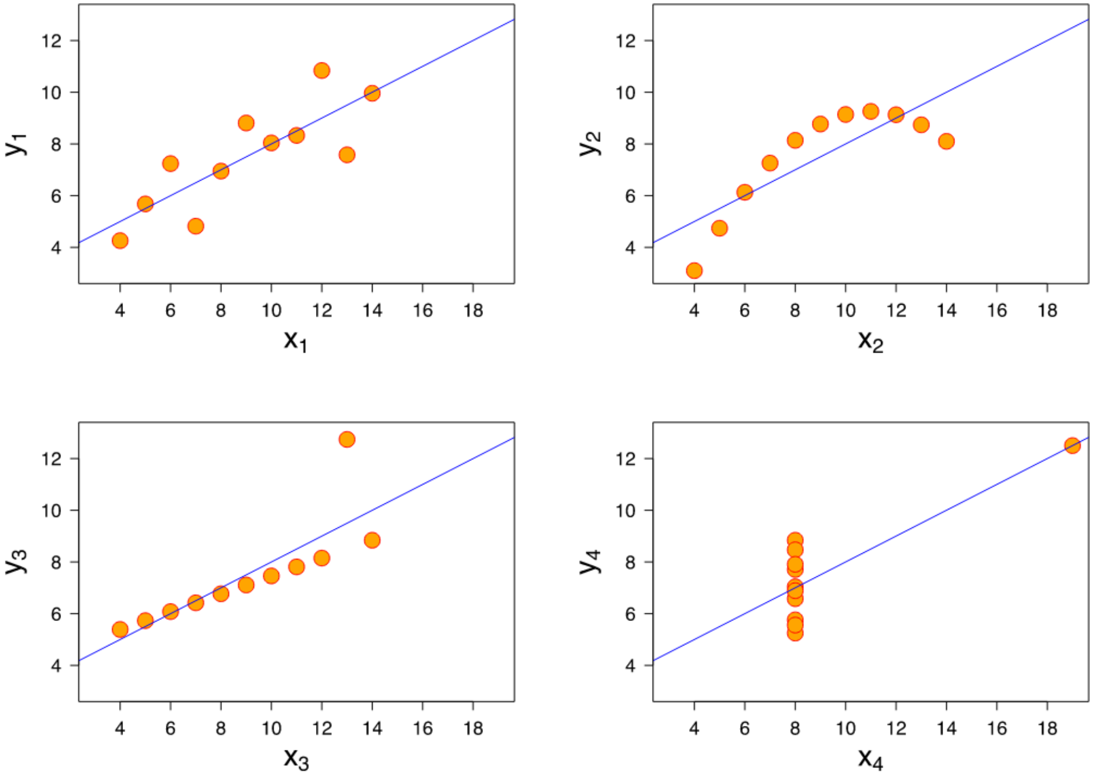
$\hat{y}$ is indicate prediction of Y when X = x

$\hat{y}$ is estimation of Y

Evaluation

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.933
Model:                            OLS   Adj. R-squared:                  0.928
Method:                 Least Squares   F-statistic:                     211.8
Date:                Mon, 03 Nov 2014   Prob (F-statistic):           6.30e-27
Time:                        14:45:06   Log-Likelihood:                -34.438
No. Observations:                  50   AIC:                             76.88
Df Residuals:                      46   BIC:                             84.52
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
x1             0.4687      0.026     17.751      0.000         0.416     0.522
x2             0.4836      0.104      4.659      0.000         0.275     0.693
x3            -0.0174      0.002     -7.507      0.000        -0.022    -0.013
const          5.2058      0.171     30.405      0.000         4.861     5.550
==============================================================================
Omnibus:                        0.655   Durbin-Watson:                   2.896
Prob(Omnibus):                  0.721   Jarque-Bera (JB):                0.360
Skew:                           0.207   Prob(JB):                        0.835
Kurtosis:                       3.026   Cond. No.                         221.
==============================================================================
```

R^2 -- useful ?



Alternatives:

- Use train/test to evaluate model

Linear Regression

Benefit

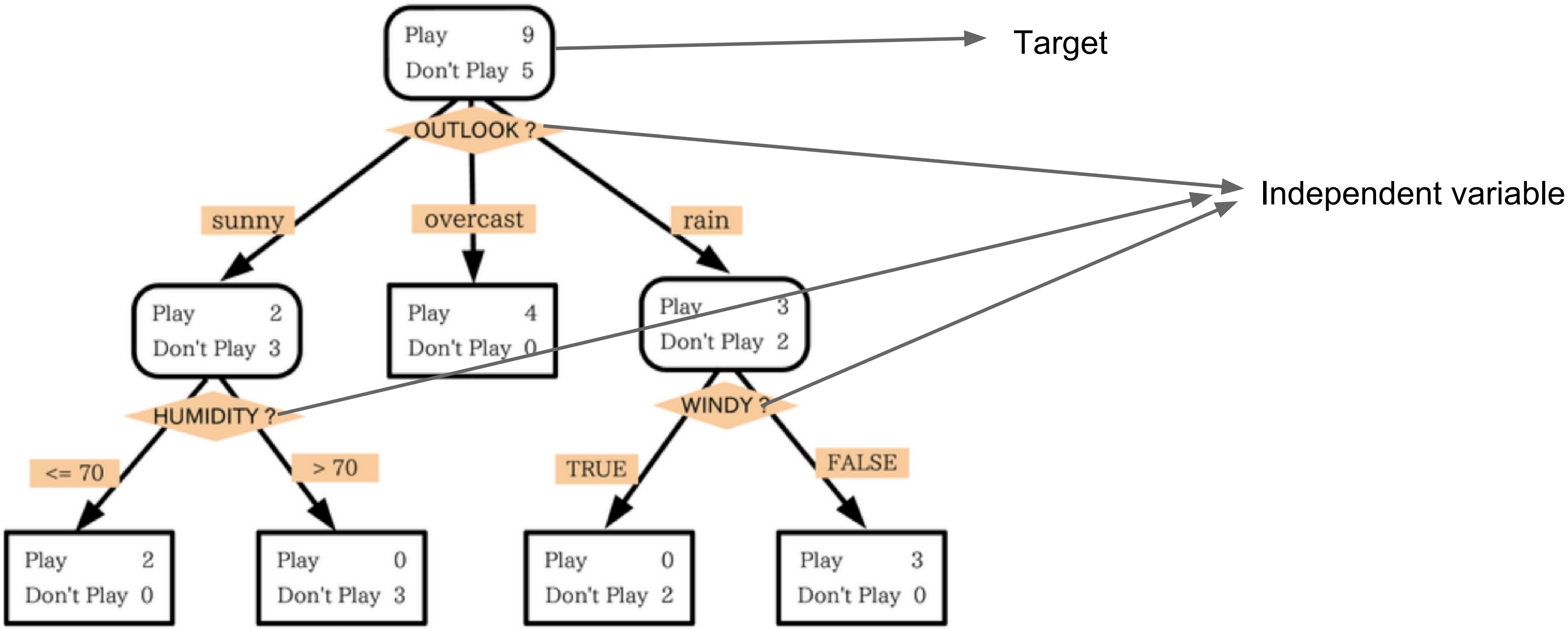
- Easy to interpret
- Computationally cheap to predict
- Computationally cheap to train
- Linear regression implements a statistical model that, when relationships between the independent variables and the dependent variable are almost linear, shows optimal results.

Disadvantage:

- Linear regression is often inappropriately used to model non-linear relationships.
- Linear regression is limited to predicting numeric output.
  -- logistic regression

Decision Tree

Decision Tree



Gini impurity
Information Gain

Tradeoffs of Decision Tree

Pros:

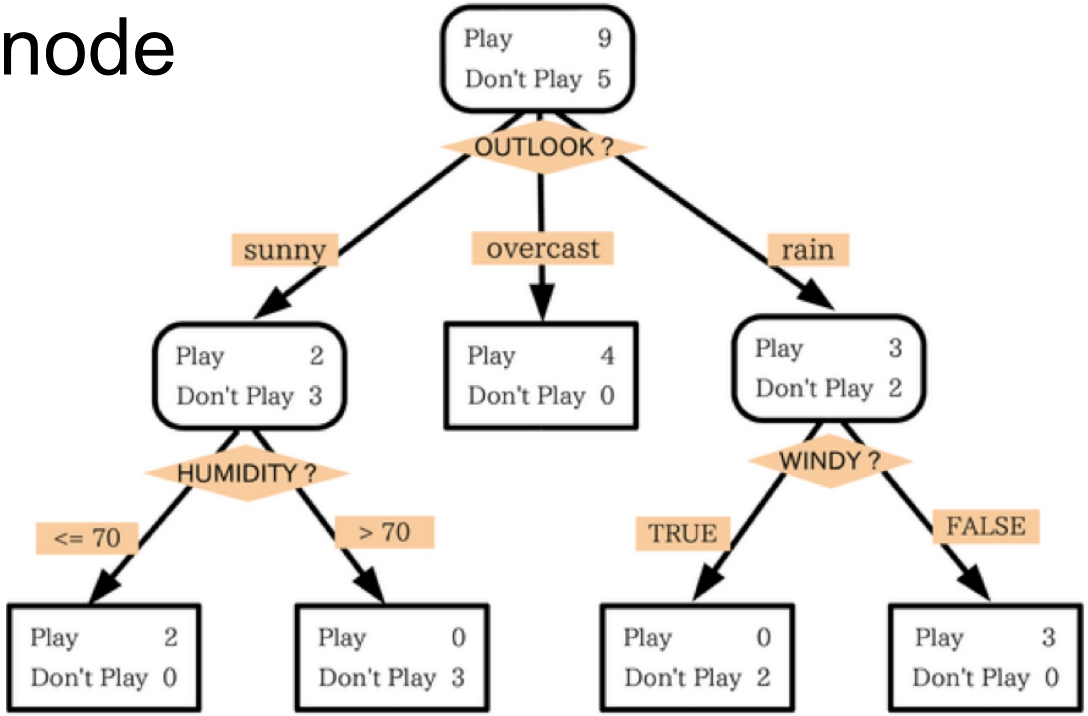
- Easily Interpretable
- Handles missing value and outliers
- Find more complex interaction
- Computationally cheap to predict
- Can handle irrelevant features
- Mix data

cons:

- Computationally expensive to train
- Greedy algorithm
- Very easy to overfit

Regularization

- Maximum Depth of tree
- Minimum sample split
- Minimum sample at leaf
- Maximum leaf node



Random Forest

Definitions

Bootstrap: can refer to any test or metric that relies on random sampling with replacement. (each random sample contains ⅔ of population)

Ensemble method: A technique for combining many weak learners in an attempt to produce a strong learner

Example:

5 completely independent classifier with accuracy of 70% for each.

Majority vote accuracy is 83.7%

How to build Random Forest

CreateRandomForest(***data***, ***num_trees***, ***num_features***):

Repeat **num_trees** times:

Create a random sample of the test data with replacement
Build a decision tree with that sample (**only consider num_features features at each node**)

Return the list of the decision trees created

Tradeoffs of Random Forest

Pros:

- Handles missing value and outliers
- Find more complex interaction
- Computationally cheap to predict
- Can handle irrelevant features
- Mix data
- Better accuracy
- One of best out of box algorithms
- Easy to Parallelize
- It runs efficiently on large databases

Cons:

- Can overfit
- Feature importance toward Continuous / categorical variable

Business Value

Confusion Matrix

		Condition (as determined by "Gold standard")	
	Total population	Condition positive	Condition negative
Test outcome	Test outcome positive	**True positive** (TP)	**False positive** (Type I error) (FP)
	Test outcome negative	**False negative** (Type II error) (FN)	**True negative** (TN)

Sensitivity & Specificity

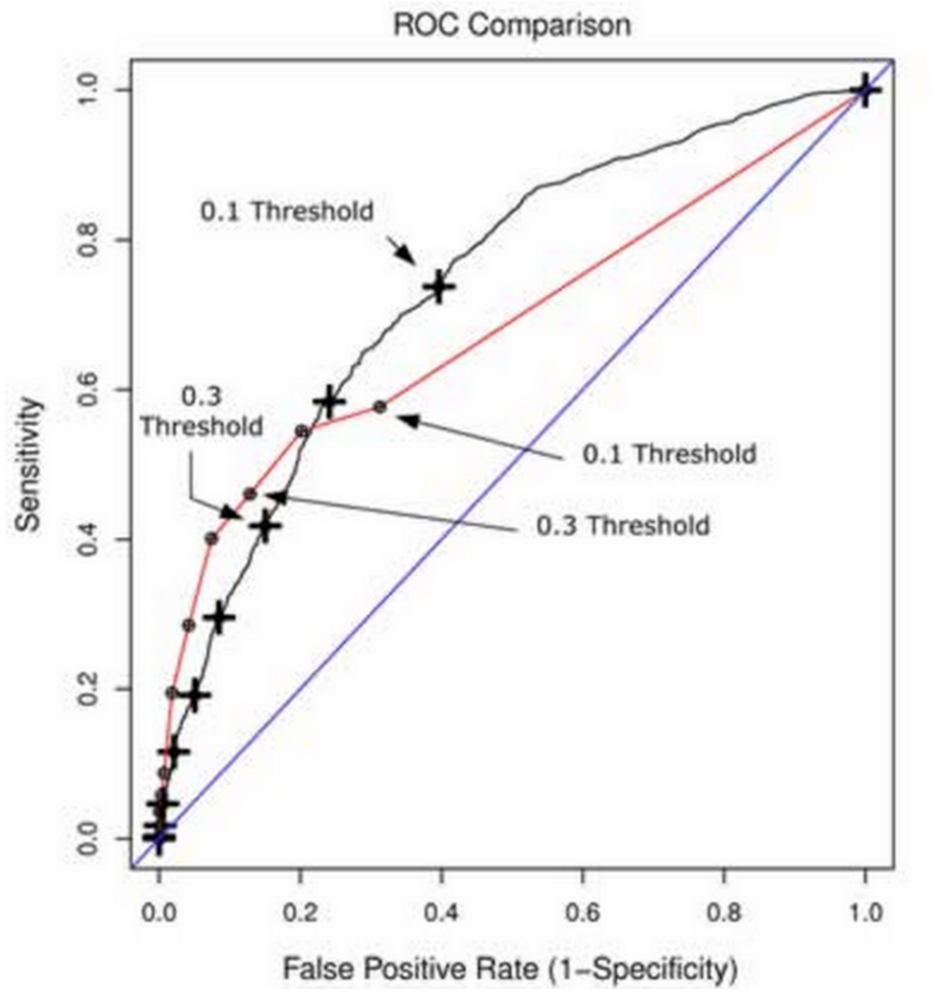
Sensitivity (also called the **true positive rate**, or the **recall** in some fields) measures the proportion of positives that are correctly identified as such (e.g., the percentage of sick people who are correctly identified as having the condition).

Sensitivity = TP/P = TP/(TP + FN)

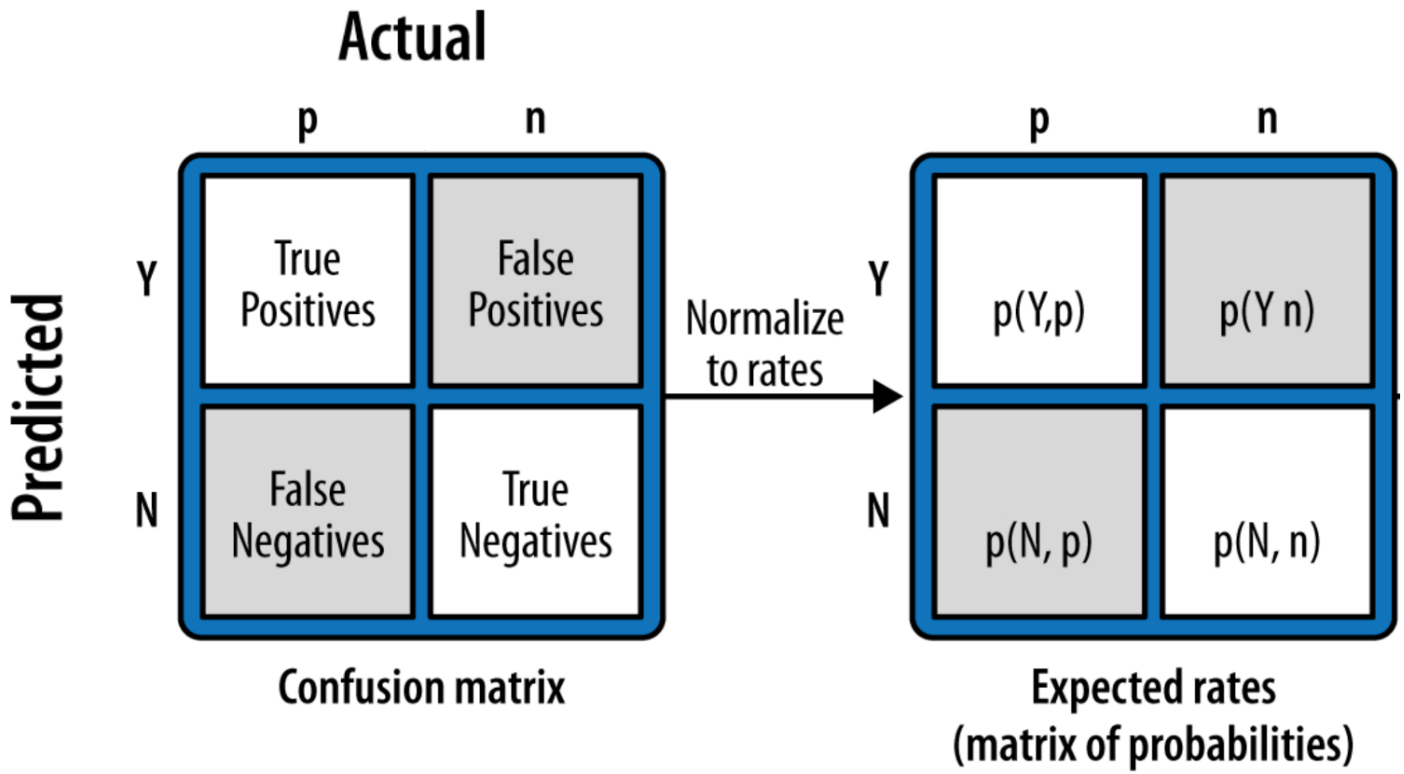
Specificity (also called the **true negative rate**) measures the proportion of negatives that are correctly identified as such (e.g., the percentage of healthy people who are correctly identified as not having the condition).

Specificity = TN/N = TN/(TN + FP)

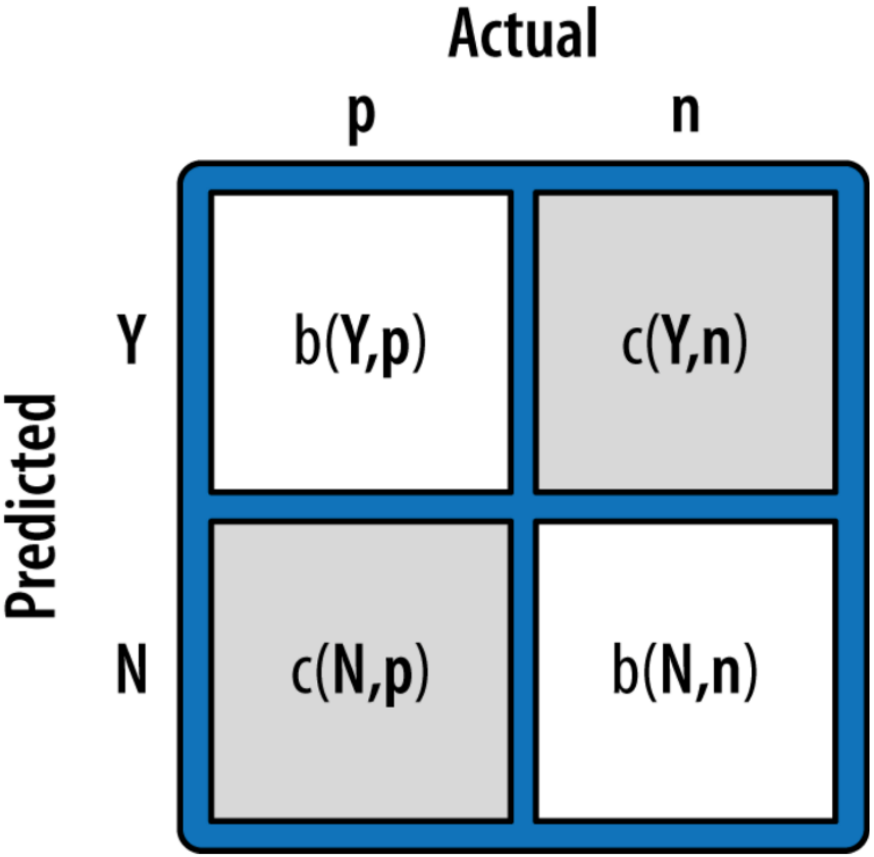
Receiver Operating Characteristic



Matrix Of Probability

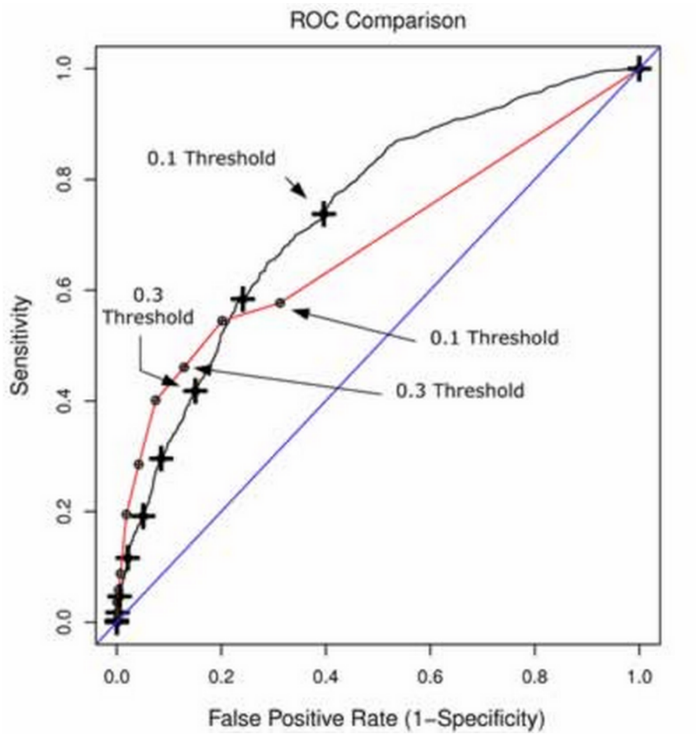


Cost-Benefit Matrix

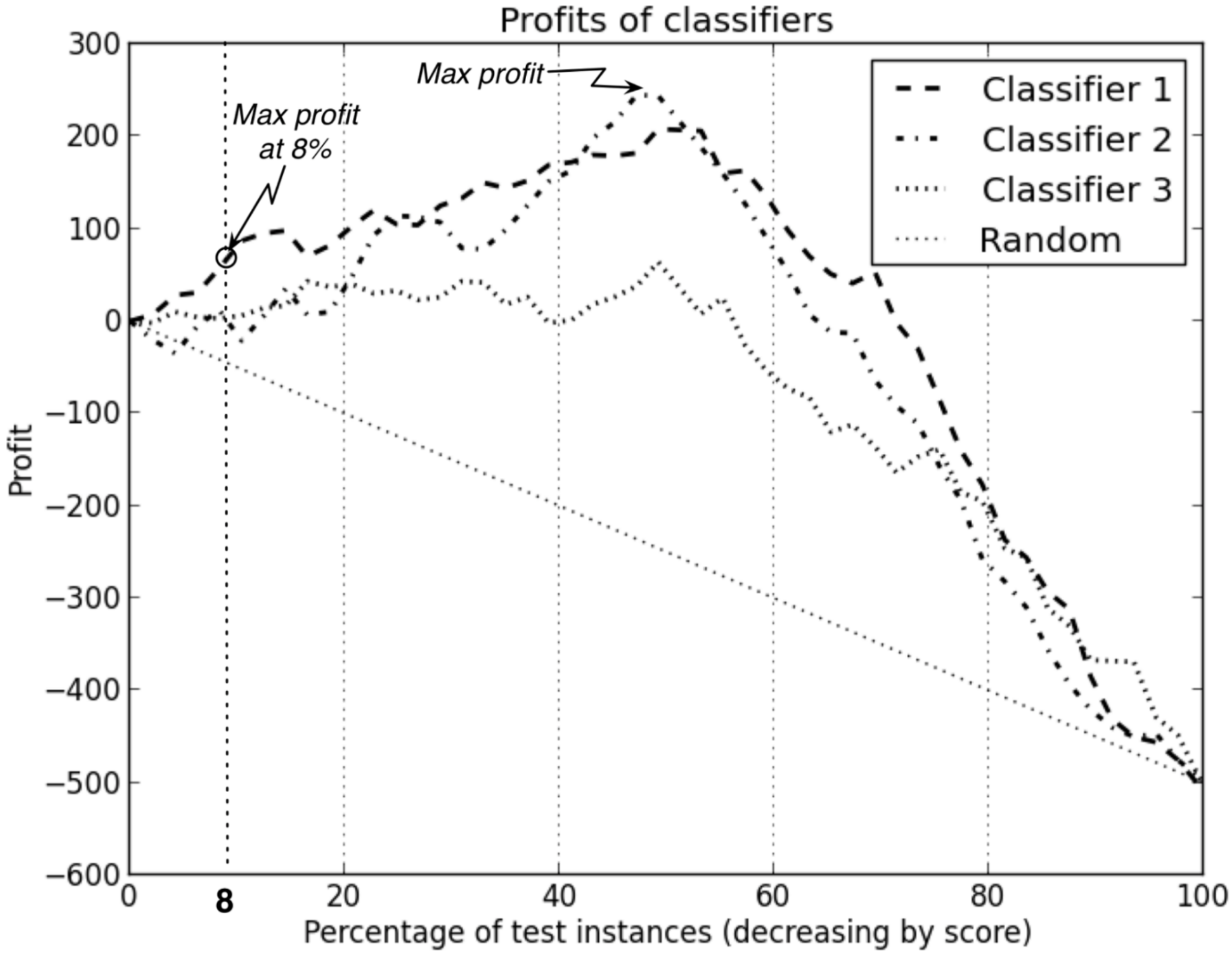


Expected Profit

$$E[Profit] = P(Y,p)\cdot b(Y,p) + P(Y,n)\cdot c(Y,n) + P(N,p)\cdot c(N,p) + P(N,n)\cdot b(N,n)$$



Profit Curve



Questions ???