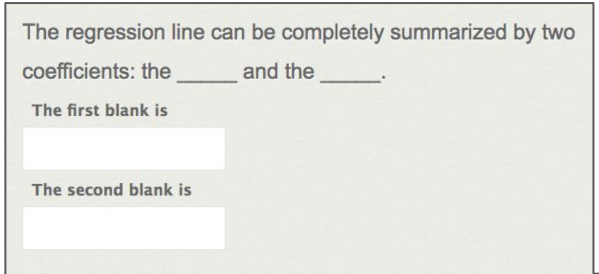
Week 5

Agenda

1. Async review and logistic regression discussion
2. Evaluation discussion
3. Regression notebook -> see notebooks folder in Github
4. Maybe: Case study "books2movies"

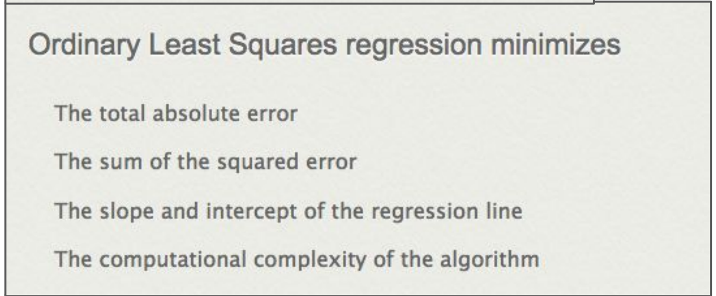
Quizzes as Warm-up



The regression line can be completely summarized by two coefficients: the _____ and the _____.

The first blank is

The second blank is



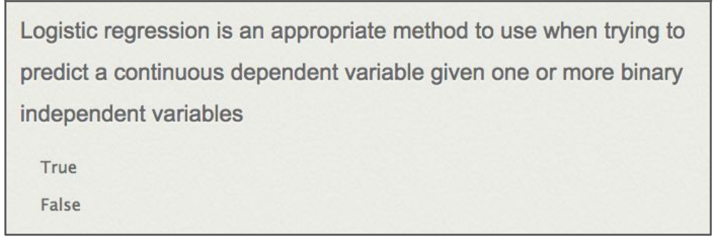
Ordinary Least Squares regression minimizes

The total absolute error

The sum of the squared error

The slope and intercept of the regression line

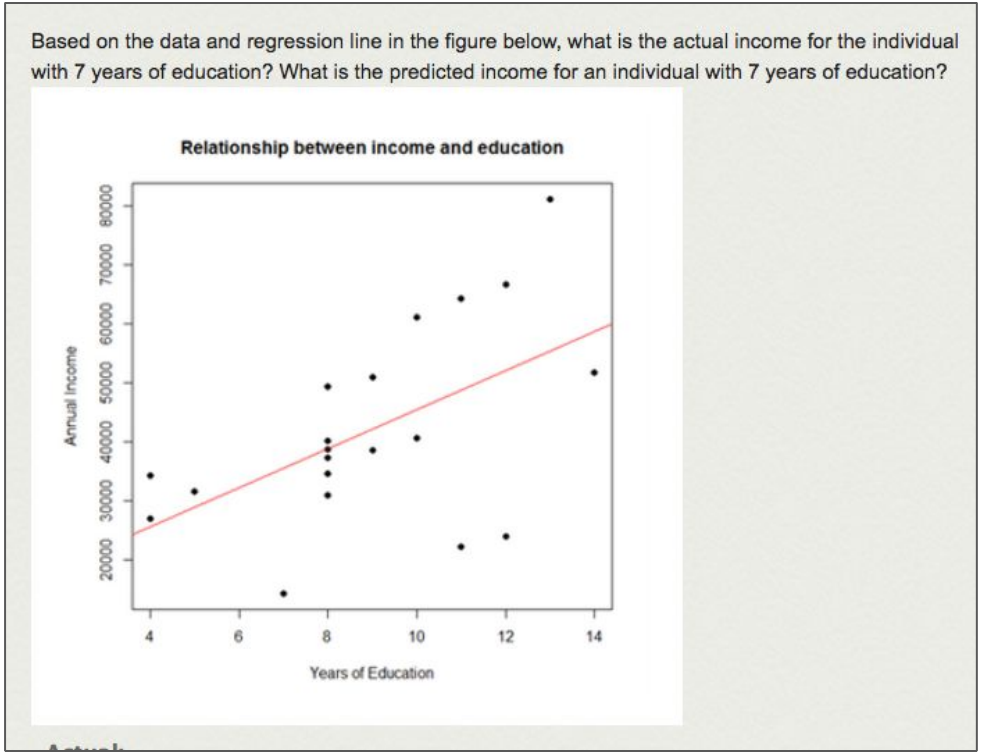
The computational complexity of the algorithm



Logistic regression is an appropriate method to use when trying to predict a continuous dependent variable given one or more binary independent variables

True

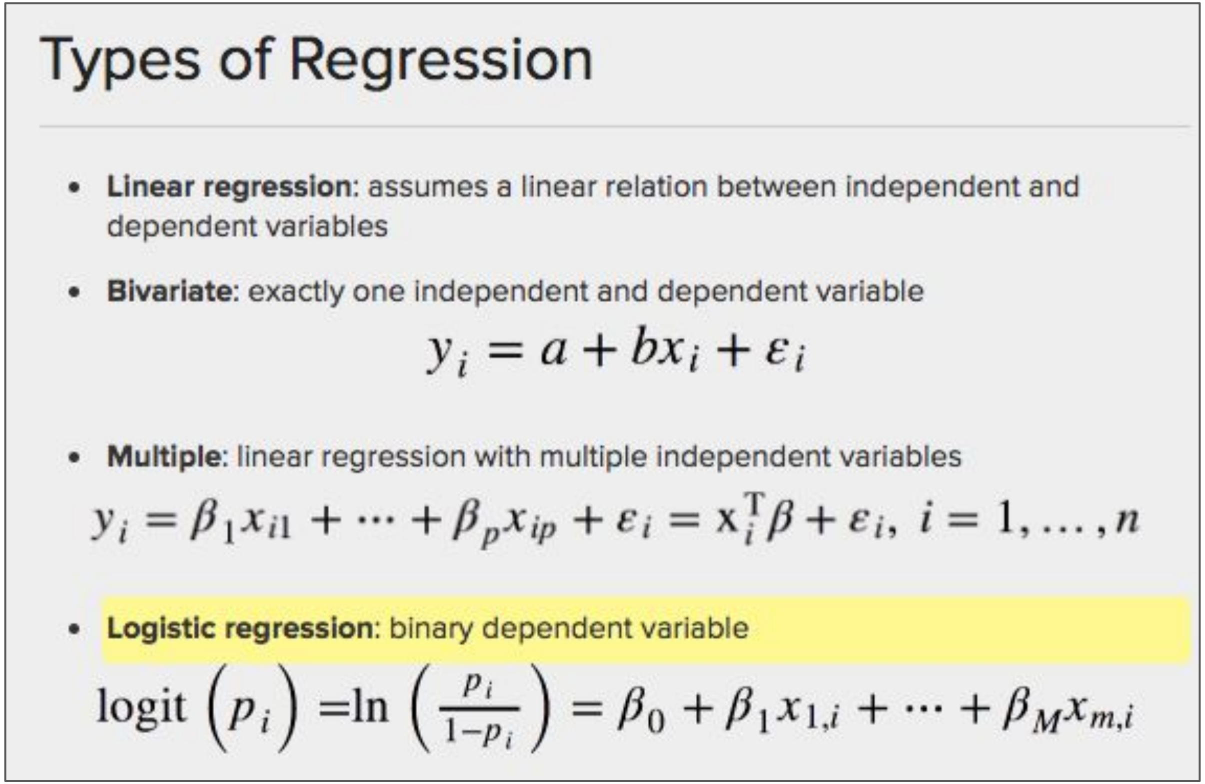
False



Based on the data and regression line in the figure below, what is the actual income for the individual with 7 years of education? What is the predicted income for an individual with 7 years of education?

Actual:

Linear Regression



Types of Regression

- **Linear regression**: assumes a linear relation between independent and dependent variables
- **Bivariate**: exactly one independent and dependent variable

$$y_i = a + bx_i + \varepsilon_i$$

- **Multiple**: linear regression with multiple independent variables

$$y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathrm{x}_i^{\mathrm{T}}\beta + \varepsilon_i, \quad i = 1, \ldots, n$$

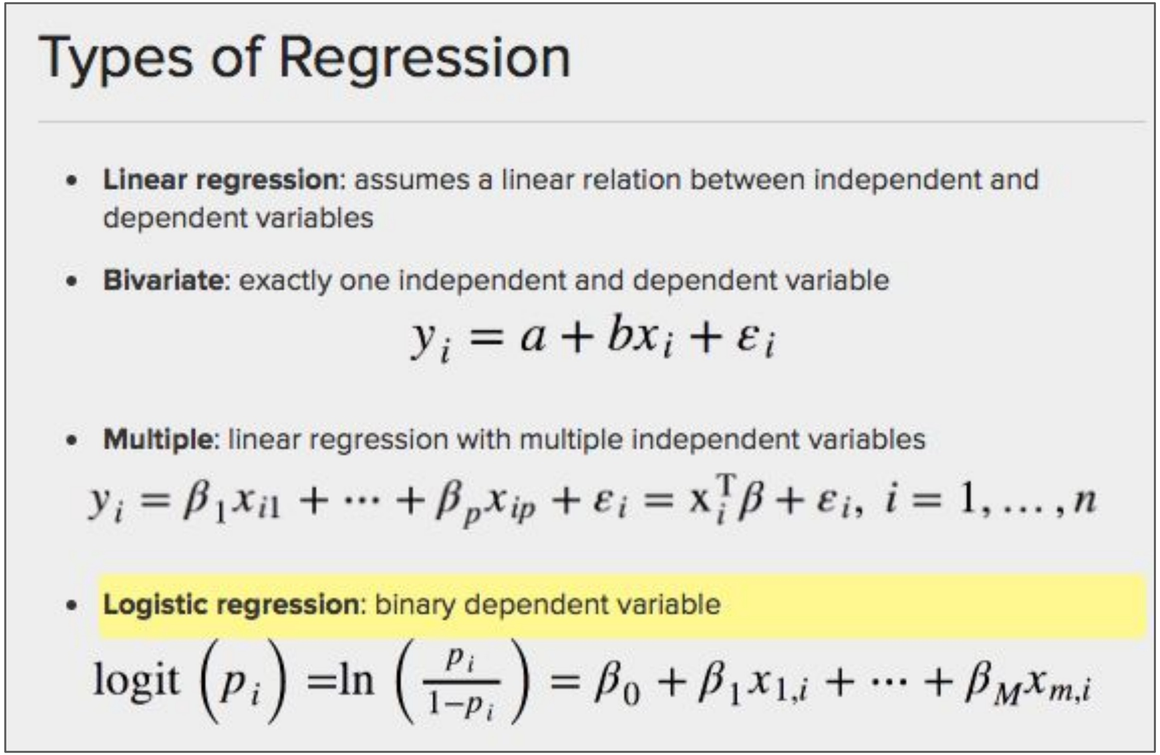
- **Logistic regression**: binary dependent variable

$$\operatorname{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_M x_{m,i}$$

What are the parameters?

Hyperparameters?

Multiple Regression



Types of Regression

- **Linear regression**: assumes a linear relation between independent and dependent variables
- **Bivariate**: exactly one independent and dependent variable

$$y_i = a + bx_i + \varepsilon_i$$

- **Multiple**: linear regression with multiple independent variables

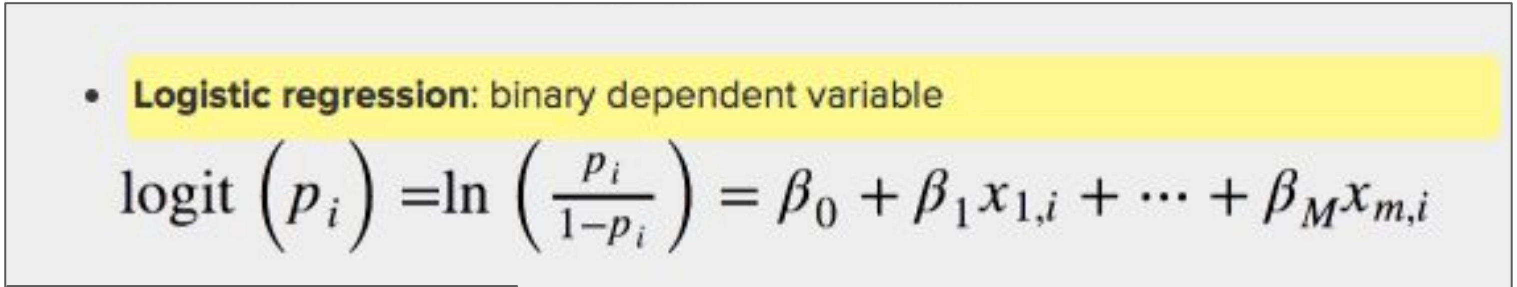
$$y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathrm{x}_i^{\mathrm{T}} \beta + \varepsilon_i, \; i = 1, \ldots, n$$

- **Logistic regression**: binary dependent variable

$$\mathrm{logit}\left(p_i\right) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_M x_{m,i}$$

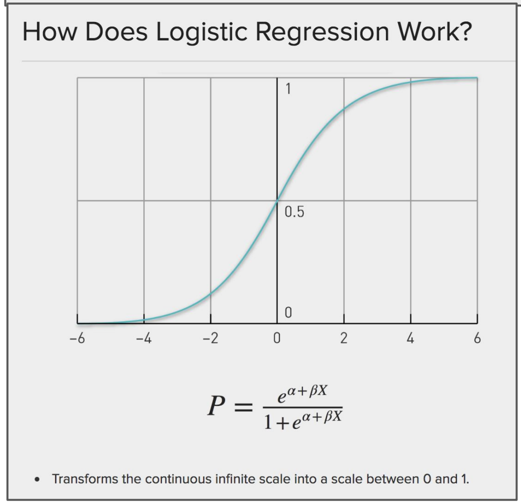
What is Multiple Regression?

Logistic Regression



- **Logistic regression**: binary dependent variable

$$\operatorname{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_M x_{m,i}$$



How Does Logistic Regression Work?

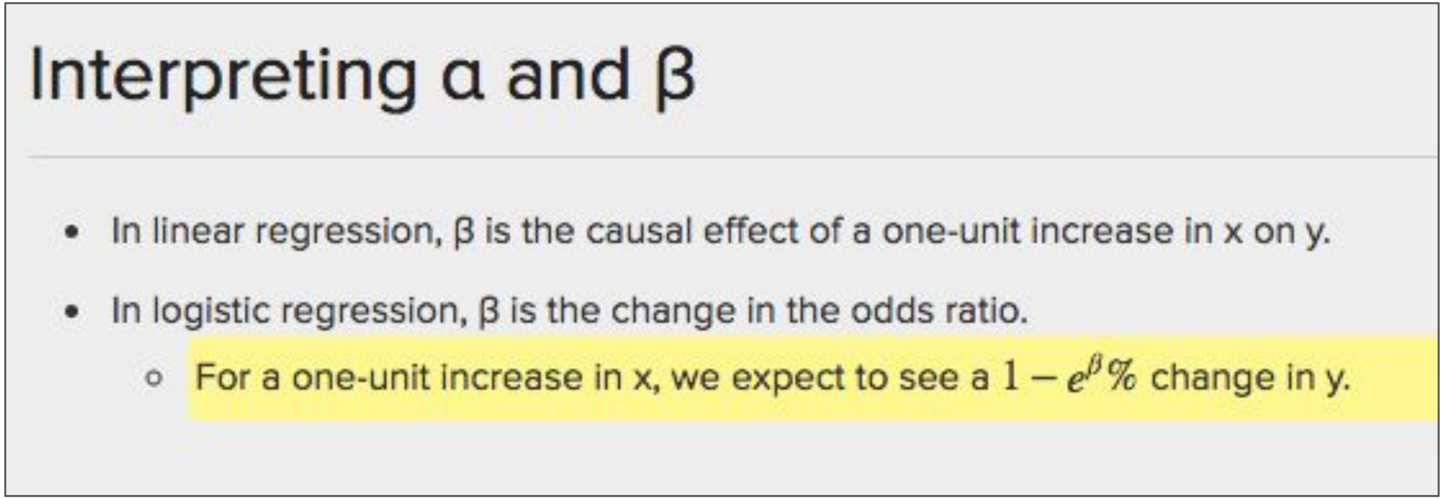
$$P = \frac{e^{\alpha+\beta X}}{1+e^{\alpha+\beta X}}$$

- Transforms the continuous infinite scale into a scale between 0 and 1.

Why is logistic regression necessary?

What are some of the characteristics of the logistic function?

Regression: Alpha and Beta

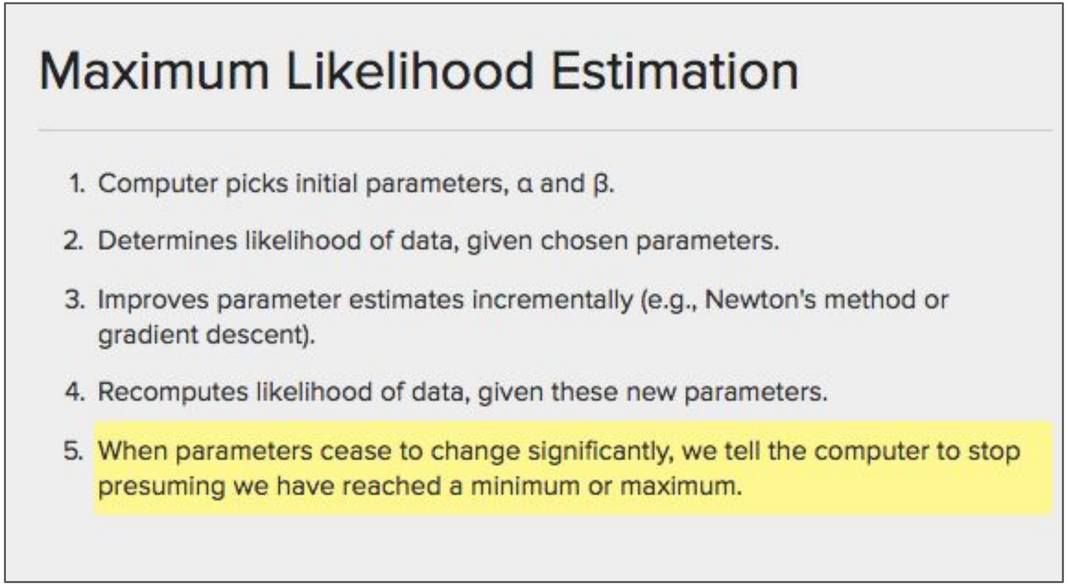


Interpreting α and β

- In linear regression, β is the causal effect of a one-unit increase in x on y.
- In logistic regression, β is the change in the odds ratio.
  - For a one-unit increase in x, we expect to see a $1 - e^{\beta}\%$ change in y.

How do you interpret the parameters?

Logistic Regression: MLE

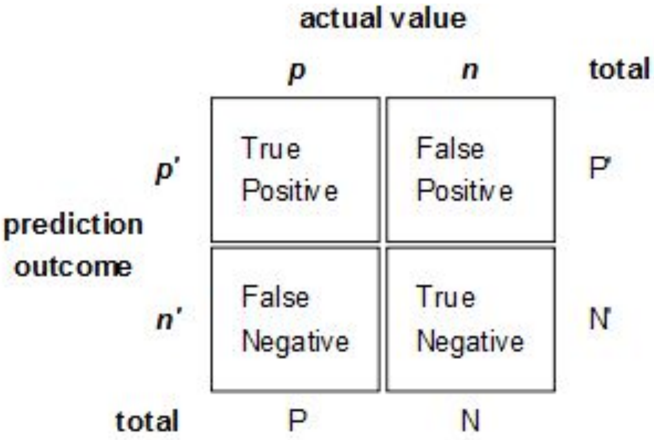


Maximum Likelihood Estimation

1. Computer picks initial parameters, α and β.
2. Determines likelihood of data, given chosen parameters.
3. Improves parameter estimates incrementally (e.g., Newton's method or gradient descent).
4. Recomputes likelihood of data, given these new parameters.
5. When parameters cease to change significantly, we tell the computer to stop presuming we have reached a minimum or maximum.

1. What makes one set of parameters better than another?
2. Describe the MLE approach to finding a good set of parameters.

Evaluation

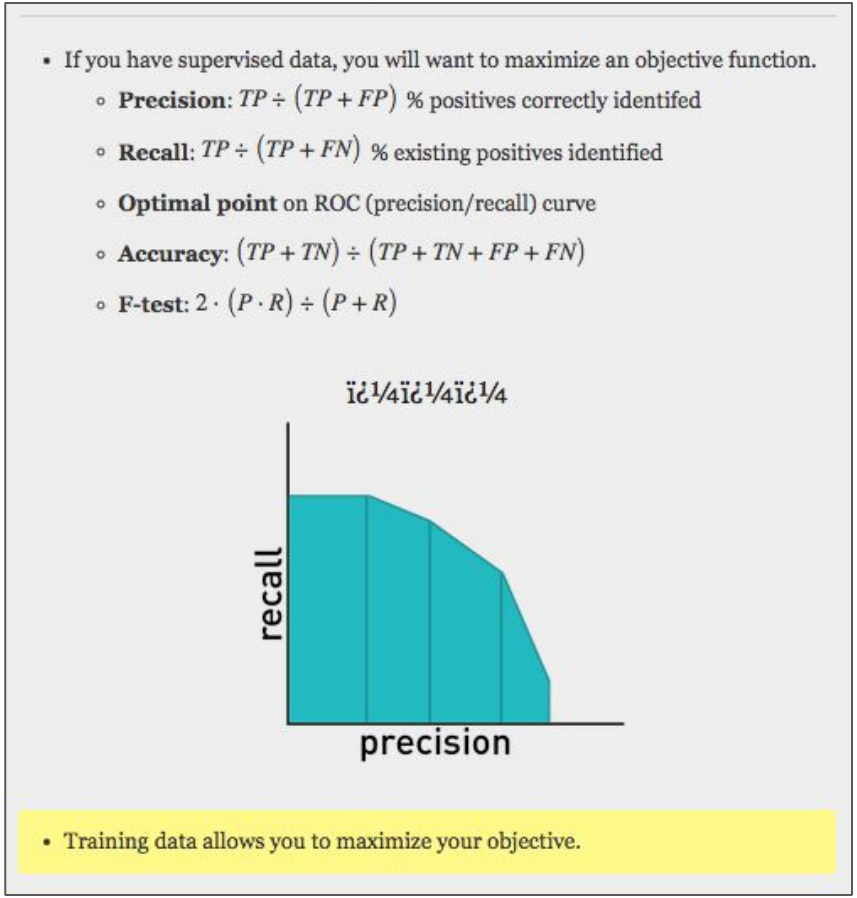


		actual value		
		p	*n*	total
prediction outcome	*p'*	True Positive	False Positive	P'
	n'	False Negative	True Negative	N'
	total	P	N	

Using and Evaluating Logistic Regression Models

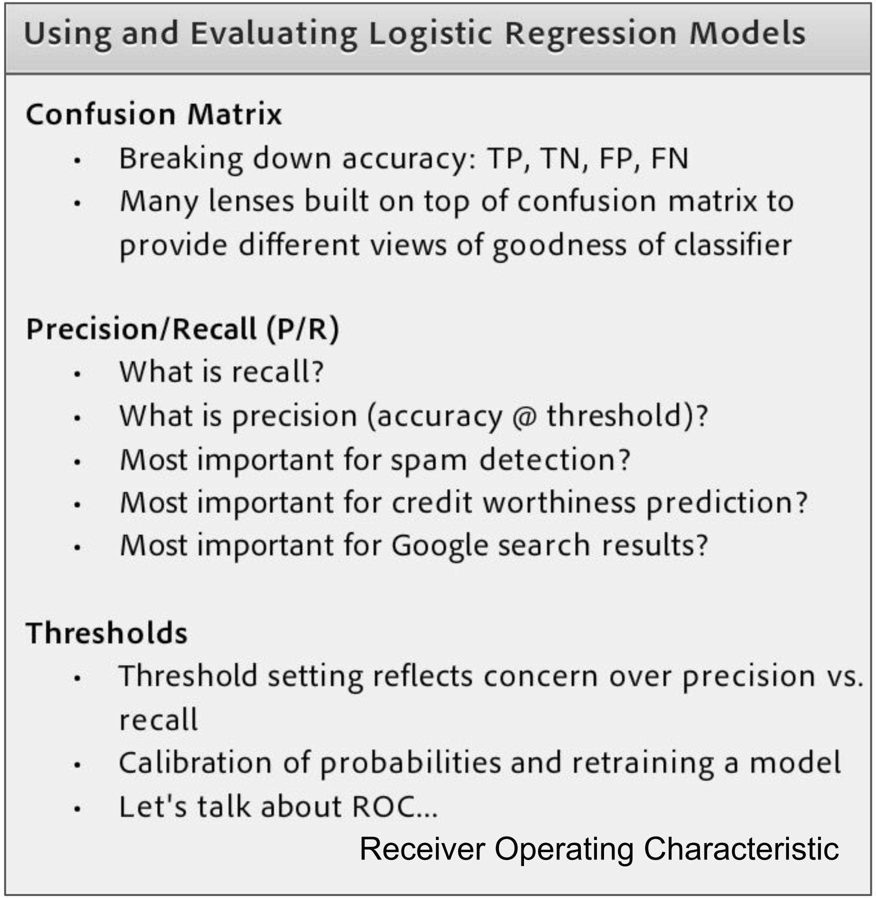
Confusion Matrix

- Breaking down accuracy: TP, TN, FP, FN
- Many lenses built on top of confusion matrix to provide different views of goodness of classifier



- If you have supervised data, you will want to maximize an objective function.
  - **Precision**: $TP \div (TP + FP)$ % positives correctly identifed
  - **Recall**: $TP \div (TP + FN)$ % existing positives identified
  - **Optimal point** on ROC (precision/recall) curve
  - **Accuracy**: $(TP + TN) \div (TP + TN + FP + FN)$
  - **F-test**: $2 \cdot (P \cdot R) \div (P + R)$

- Training data allows you to maximize your objective.



Using and Evaluating Logistic Regression Models

Confusion Matrix

- Breaking down accuracy: TP, TN, FP, FN
- Many lenses built on top of confusion matrix to provide different views of goodness of classifier

Precision/Recall (P/R)

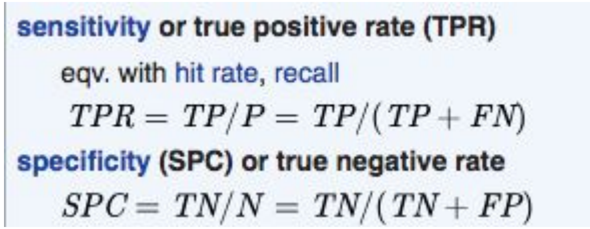
- What is recall?
- What is precision (accuracy @ threshold)?
- Most important for spam detection?
- Most important for credit worthiness prediction?
- Most important for Google search results?

Thresholds

- Threshold setting reflects concern over precision vs. recall
- Calibration of probabilities and retraining a model
- Let's talk about ROC...

Receiver Operating Characteristic

The ROC(Receiver Operator Characteristic)/AUC (Area under the curve)



sensitivity or true positive rate (TPR)

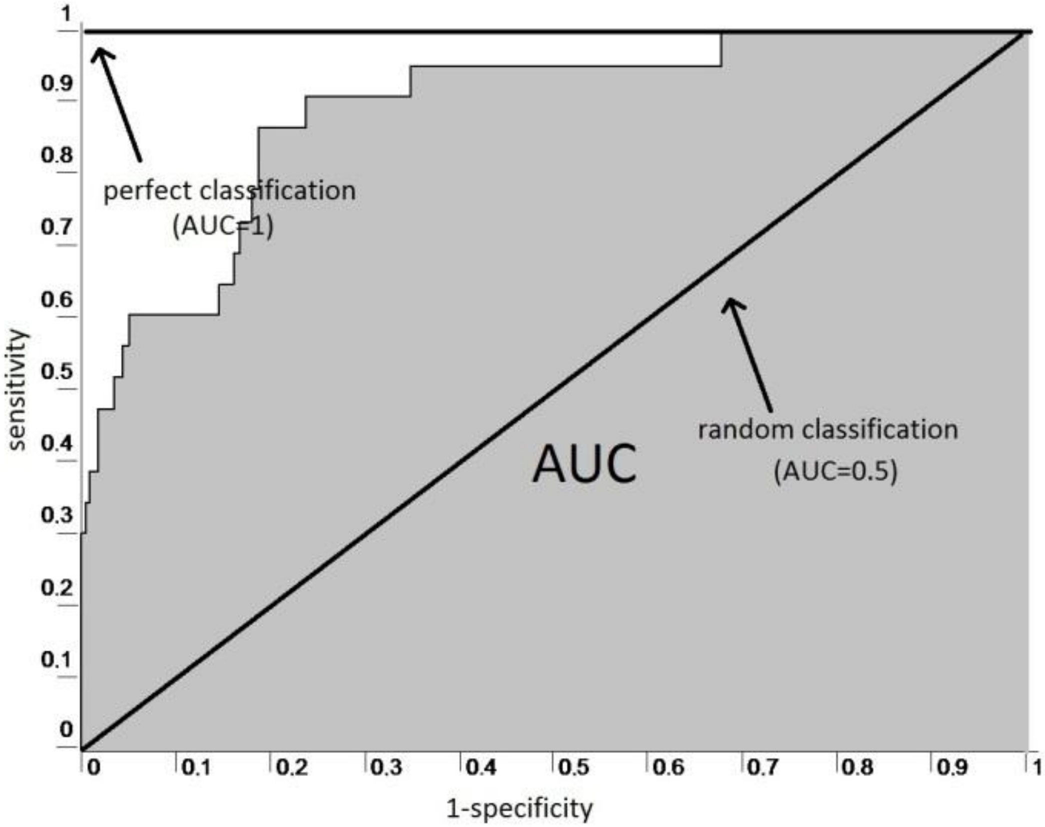
eqv. with hit rate, recall

$$TPR = TP/P = TP/(TP + FN)$$

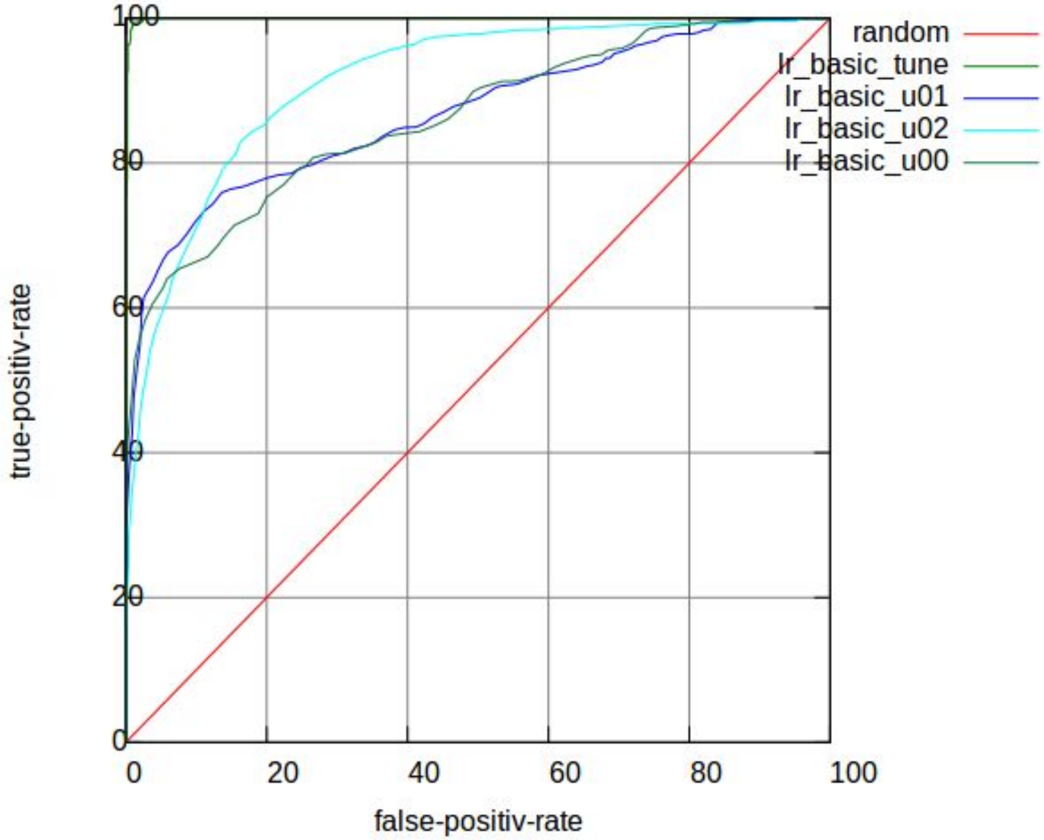
specificity (SPC) or true negative rate

$$SPC = TN/N = TN/(TN + FP)$$

1. What means perfect classification?
2. Describe the role of the threshold when comparing classifiers.
3. Why is the curve of TPR vs FPR generated by changing the threshold monotonically increase?
4. When is the AUC useful and when might it not be?

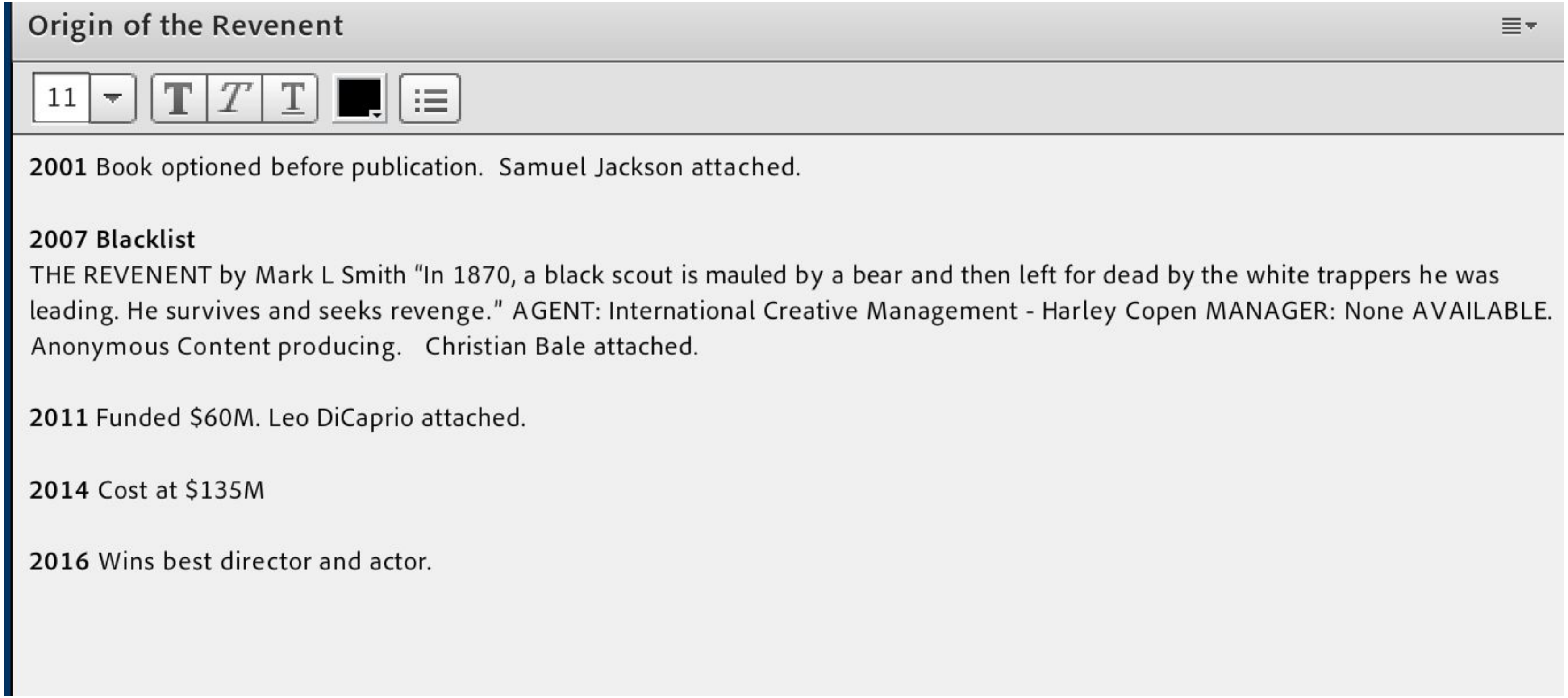


The ROC(Receiver Operator Characteristic)/AUC



Books2Movies Case study

Case study: Selecting Books to Option for Movies



Origin of the Revenent

2001 Book optioned before publication. Samuel Jackson attached.

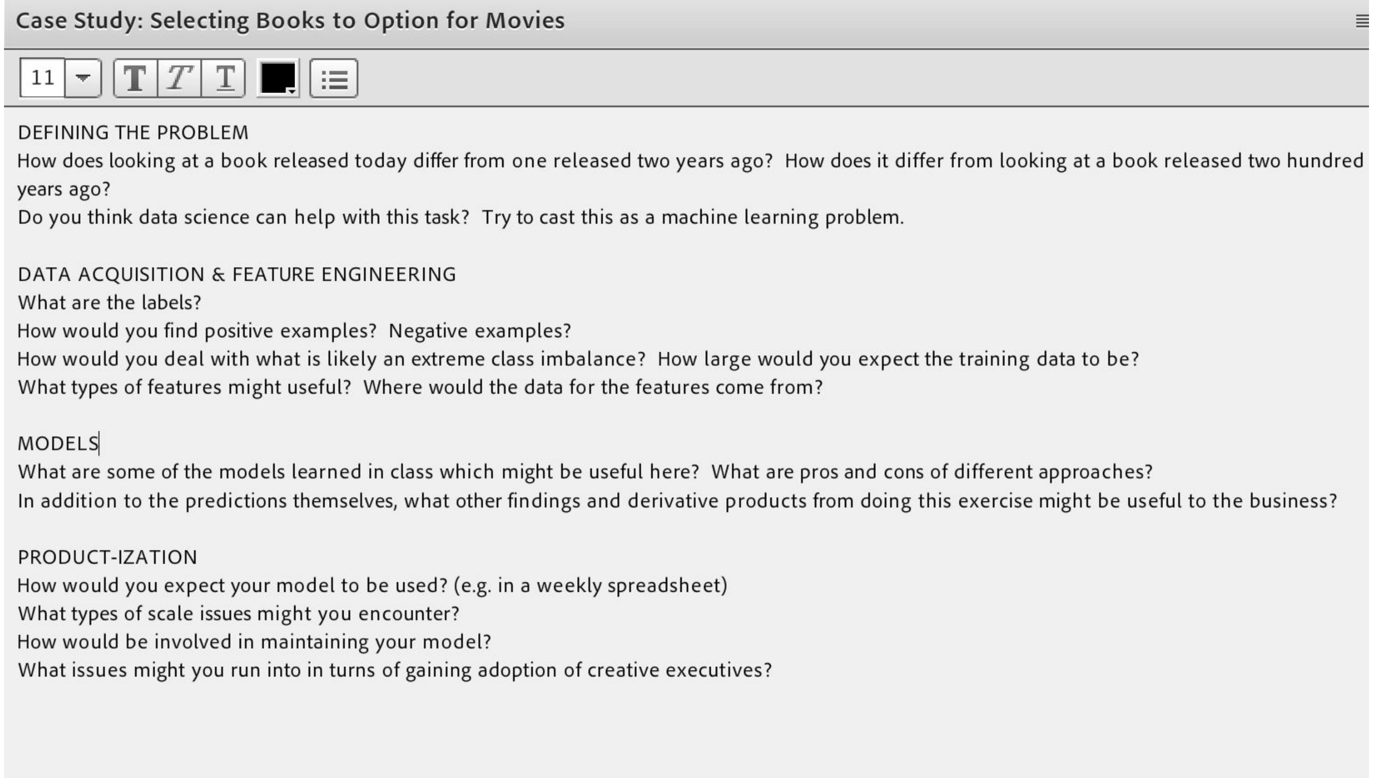
2007 Blacklist
THE REVENENT by Mark L Smith "In 1870, a black scout is mauled by a bear and then left for dead by the white trappers he was leading. He survives and seeks revenge." AGENT: International Creative Management - Harley Copen MANAGER: None AVAILABLE. Anonymous Content producing.   Christian Bale attached.

2011 Funded $60M. Leo DiCaprio attached.

2014 Cost at $135M

2016 Wins best director and actor.

Case study: Selecting Books to Option for Movies



Case Study: Selecting Books to Option for Movies

DEFINING THE PROBLEM
How does looking at a book released today differ from one released two years ago? How does it differ from looking at a book released two hundred years ago?
Do you think data science can help with this task? Try to cast this as a machine learning problem.

DATA ACQUISITION & FEATURE ENGINEERING
What are the labels?
How would you find positive examples? Negative examples?
How would you deal with what is likely an extreme class imbalance? How large would you expect the training data to be?
What types of features might useful? Where would the data for the features come from?

MODELS
What are some of the models learned in class which might be useful here? What are pros and cons of different approaches?
In addition to the predictions themselves, what other findings and derivative products from doing this exercise might be useful to the business?

PRODUCT-IZATION
How would you expect your model to be used? (e.g. in a weekly spreadsheet)
What types of scale issues might you encounter?
How would be involved in maintaining your model?
What issues might you run into in turns of gaining adoption of creative executives?

Case study: Selecting Books to Option for Movies

1. Preface: What are journeys from a book to a movie?
2. Problem definition:
    a. What changes for a book between release date and 2 years later? 100 years later?
    b. Can ML help with this? Try to formulate this as a machine learning problem.
3. Data Acquisition and Feature engineering:
    a. What set of books should you look at? How large would your training data be?
    b. What are the labels?
    c. How to deal with extreme class imbalance?
    d. What types of features would be useful? Where would the data for features come from?
4. Models:
    a. Which of the models we learned so far might be useful? Pros/cons of the approaches?
    b. What are findings beside the predictions that might be useful?
5. Product:
    a. How would you expect your model to be used?
    b. What types of scaling issues might come up?
    c. What issues might come up in terms of adoption of your model with creative execs?