Classification

- Classification is perhaps the most important form of prediction:
  - the goal is to predict whether a record is a 0 or a 1 *(phishing/not-phishing, click/don't click, churn/don't churn)*
  - or in some cases, one of several categories *(for example, Gmail's filtering of your inbox into "primary," "social," "promotional," or "forums")*
- *Often, we need more than a simple binary classification: we want to know the predicted probability that a case belongs to a class.*
- Rather than having a model simply assign a binary classification, most algorithms can return a probability score (propensity) of belonging to the class of interest.

*This presentation is based on the book 'Practical Statistics for Data Scientist' by Peter Bruce & Andrew Bruce

Naive Bayes

- To understand Bayesian classification, we can start out by imagining “non-naive” Bayesian classification.
- For each record to be classified:
  - Find all the other records with the same predictor profile (i.e., where the predictor values are the same).
  - Determine what classes those records belong to and which class is most prevalent (i.e., probable).
  - Assign that class to the new record.
- Why Exact Bayesian Classification Is Impractical:
  - When the number of predictor variables exceeds a handful, many of the records to be classified will be without exact matches.
- WARNING
  - Despite its name, naive Bayes is not considered a method of Bayesian statistics. Naive Bayes is a data–driven, empirical method requiring relatively little statistical expertise. The name comes from the Bayes rule–like calculation in forming the predictions

- In the naive Bayes solution, we no longer restrict the probability calculation to those records that match the record to be classified. Instead, we use the entire data set. The naive Bayes modification is as follows:
  - For a binary response Y = i (i = 0 or 1), estimate the individual conditional probabilities for each predictor ; these are the probabilities that the predictor value is in the record when we observe Y = i. This probability is estimated by the proportion of Xj values among the Y = i records in the training set.
  - Multiply these probabilities by each other, and then by the proportion of records belonging to Y = i.
  - Repeat steps 1 and 2 for all the classes.
  - Estimate a probability for outcome i by taking the value calculated in step 2 for class i and dividing it by the sum of such values for all classes.
  - Assign the record to the class with the highest probability for this set of predictor values.
- Why is this formula called “naive”?
  - We have made a simplifying assumption that the exact conditional probability of a vector of predictor values, given observing an outcome, is sufficiently well estimated by the product of the individual conditional probabilities
  - In other words, in estimating probability of one class seperetaly instead of all, we are assumi ng that is class is independent of all the other predictors.

Numeric Predictor Variables

- From the definition, we see that the Bayesian classifier works only with categorical predictors
- To apply naive Bayes to numerical predictors, one of two approaches must be taken:
  - Bin and convert the numerical predictors to categorical predictors
  - Use a probability model — for example, the normal distribution (see “Normal Distribution”) — to estimate the conditional probability .

Discriminant Analysis

- Discriminant analysis is the earliest statistical classifier
- While discriminant analysis encompasses several techniques, the most commonly used is linear discriminant analysis, or LDA.
- In addition, discriminant analysis can provide a measure of predictor importance, and it is used as a computationally efficient method of feature selection.

- To understand discriminant analysis, it is first necessary to introduce the concept of **covariance** between two or more variables.
- The covariance measures the relationship between two variables

$$s_{x,z} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(z_i - \overline{z})}{n-1}$$

- As with the correlation coefficient, positive values indicate a positive relationship and negative values indicate a negative relationship.
- Correlation, however, is constrained to be between –1 and 1, whereas covariance is on the same scale as the variables and .

Fisher's Linear Discriminant

- For simplicity, we focus on a classification problem in which we want to predict a binary outcome y using just two continuous numeric variables (x,z)
- Technically, discriminant analysis assumes the predictor variables are normally distributed continuous variables, but, in practice, the method works well even for non-extreme departures from normality, and for binary predictors.
- Fisher's linear discriminant distinguishes variation between groups, on the one hand, from variation within groups on the other.
- Specifically, seeking to divide the records into two groups, LDA focuses on maximizing the "between" sum of squares (measuring the variation between the two groups) relative to the "within" sum of squares (measuring the within-group variation).
- In this case, the two groups correspond to the records for which y = 0 and the records for which y = 1.
- Intuitively, by maximizing the between sum of squares and minimizing the within sum of squares, this method yields the greatest separation between the two groups.

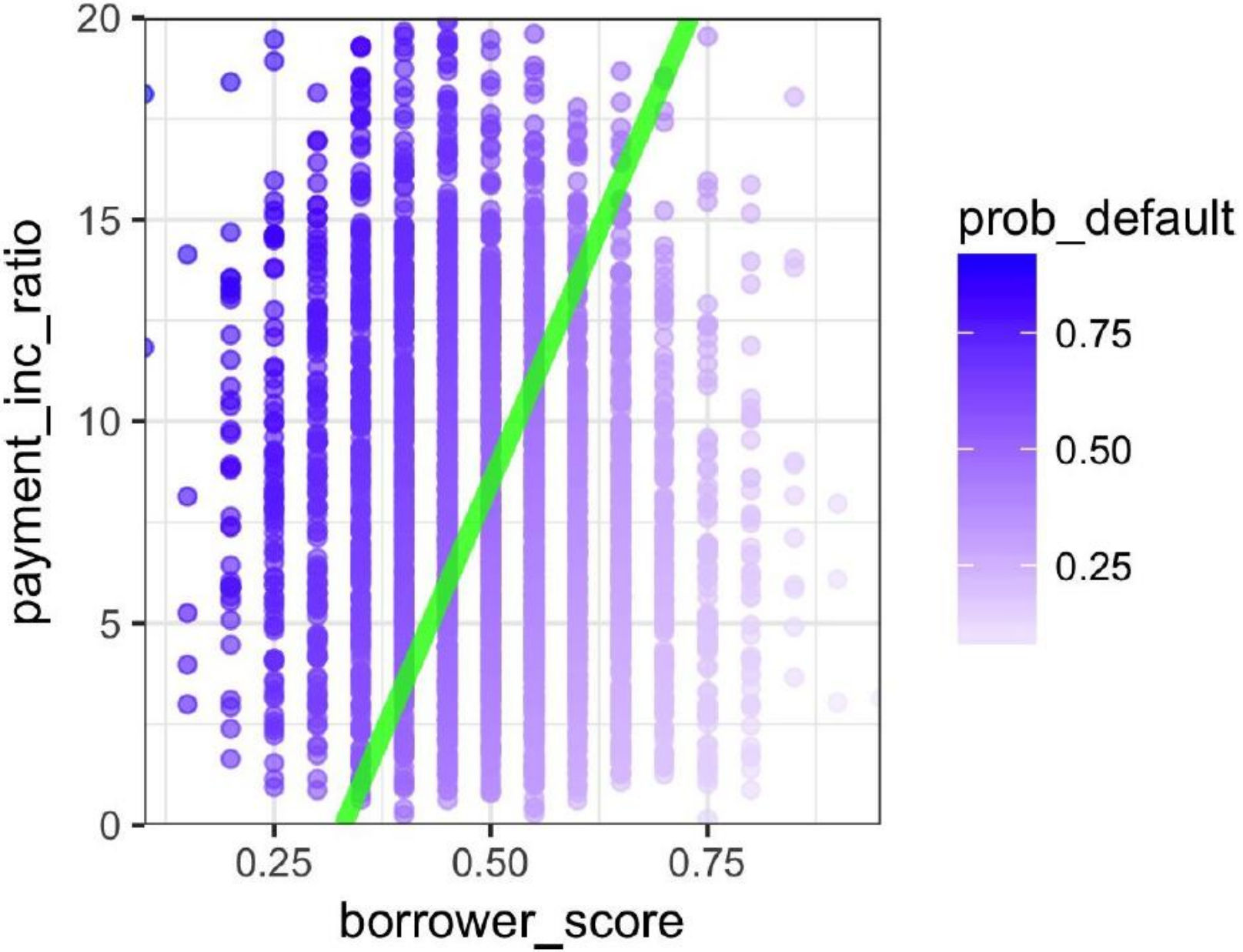


Figure 5-1. LDA prediction of loan default using two variables: a score of the borrower's creditworthiness and the payment to income ratio.

Logistic Regression

- Logistic regression is analogous to multiple linear regression, except the outcome is binary.
- Like discriminant analysis, and unlike K-Nearest Neighbor and naive Bayes, logistic regression is a structured model approach, rather than a data-centric approach.
- The key ingredients are the logistic response function and the logit, in which we map a probability (which is on a 0–1 scale) to a more expansive scale suitable for linear modeling.
- Naively, we might be tempted to model *p* as a linear function of the predictor variables:

$$p = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q,$$

- However, fitting this model does not ensure that p will end up between 0 and 1, as a probability must. Instead, we model p by applying a logistic response or inverse logit function to the predictors:

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q)}}.$$

- This transform ensures that the *p* stays between 0 and 1.
- In terms of probabilities, odds are the probability of an event divided by the probability that the event will not occur.

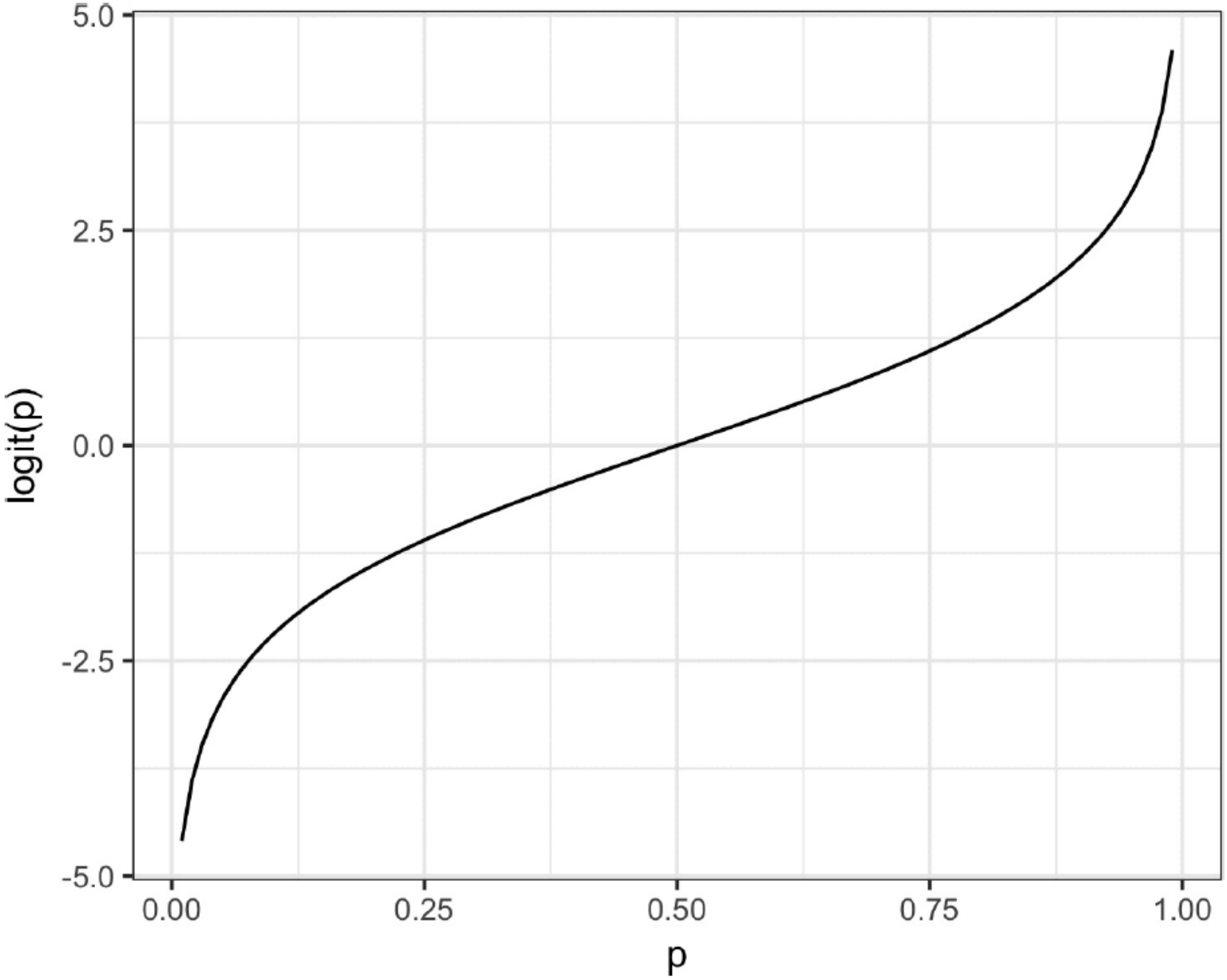
$$\text{Odds}(Y = 1) = \frac{p}{1 - p}.$$

$$\text{Odds}(Y = 1) = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q}.$$

- Finally, taking the logarithm of both sides, we get an expression that involves a linear function of the predictors:

$$\log\left(\text{Odds}(Y = 1)\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q.$$

- The *log-odds* function, also known as the *logit* function, maps the probability *p* from (0,1) to any value (-∞,+∞)


5.0
2.5
0.0
-2.5
-5.0
logit(p)
0.00
0.25
0.50
0.75
1.00
p

Logistic Regression and the GLM

The response in the logistic regression formula is the log odds of a binary outcome of 1. We only observe the binary outcome, not the log odds, so special statistical methods are needed to fit the equation. Logistic regression is a special instance of a *generalized linear model* (GLM) developed to extend linear regression to other settings.

In R, to fit a logistic regression, the `glm` function is used with the `family` parameter set to `binomial`. The following code fits a logistic regression to the personal loan data introduced in "K-Nearest Neighbors".

```
logistic_model

Call:  glm(formula = outcome ~ payment_inc_ratio + purpose_ + home_ +
    emp_len_ + borrower_score, family = "binomial", data = loan_data)

Coefficients:
               (Intercept)           payment_inc_ratio
                   1.26982                     0.08244
purpose_debt_consolidation    purpose_home_improvement
                   0.25216                     0.34367
    purpose_major_purchase             purpose_medical
                   0.24373                     0.67536
             purpose_other      purpose_small_business
                   0.59268                     1.21226
                  home_OWN                   home_RENT
                   0.03132                     0.16867
        emp_len_ < 1 Year              borrower_score
                   0.44489                    -4.63890

Degrees of Freedom: 46271 Total (i.e. Null);  46260 Residual
Null Deviance:      64150
Residual Deviance: 58530        AIC: 58550
```

Generalized Linear Models

- Generalized linear models (GLMs) are the second most important class of models besides regression.
- GLMs are characterized by two main components:
  - A probability distribution or family (binomial in the case of logistic regression)
  - A link function mapping the response to the predictors (logit in the case of logistic regression)
- Logistic regression is by far the most common form of GLM.
- The poisson distribution is commonly used to model count data (e.g., the number of times a user visits a web page in a certain amount of time).
- Other families include negative binomial and gamma, often used to model elapsed time (e.g., time to failure).
- In contrast to logistic regression, application of GLMs with these models is more nuanced and involves greater care. These are best avoided unless you are familiar with and understand the utility and pitfalls of these methods.

Predicted Values from Logistic Regression

$$\hat{p} = \frac{1}{1 + e^{-\hat{Y}}}$$

For example, look at the predictions from the model `logistic_model`:

```
pred <- predict(logistic_model)
summary(pred)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-2.728000 -0.525100 -0.005235  0.002599  0.513700  3.658000
```

Converting these values to probabilities is a simple transform:

```
prob <- 1/(1 + exp(-pred))
> summary(prob)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.06132 0.37170 0.49870 0.50000 0.62570 0.97490
```

- These are on a scale from 0 to 1 and don't yet declare whether the predicted value is default or paid off.
- We could declare any value greater than 0.5 as default, analogous to the K-Nearest Neighbors classifier.
- In practice, a lower cutoff is often appropriate if the goal is to identify members of a rare class

Interpreting the Coefficients and Odds Ratios

- One advantage of logistic regression is that it produces a model that can be scored to new data rapidly, without re-computation.
- Another is the relative ease of interpretation of the model, as compared with other classification methods. The key conceptual idea is understanding an odds ratio.

$$\text{odds ratio} = \frac{\text{Odds}(Y = 1 \mid X = 1)}{\text{Odds}(Y = 1 \mid X = 0)}$$

- This is interpreted as the odds that Y = 1 when X = 1 versus the odds that Y = 1 when X = 0. If the odds ratio is 2, then the odds that Y = 1 are two times higher when X = 1 versus X = 0.
- An example will make this more explicit. For the model fit in “Logistic Regression and the GLM”, the regression coefficient for purpose_small_business is 1.21226. This means that a loan to a small business compared to a loan to pay off credit card debt reduces the odds of defaulting versus being paid off by exp(1.21526) = 3.4 . Clearly, loans for the purpose of creating or expanding a small business are considerably riskier than other types of loans.

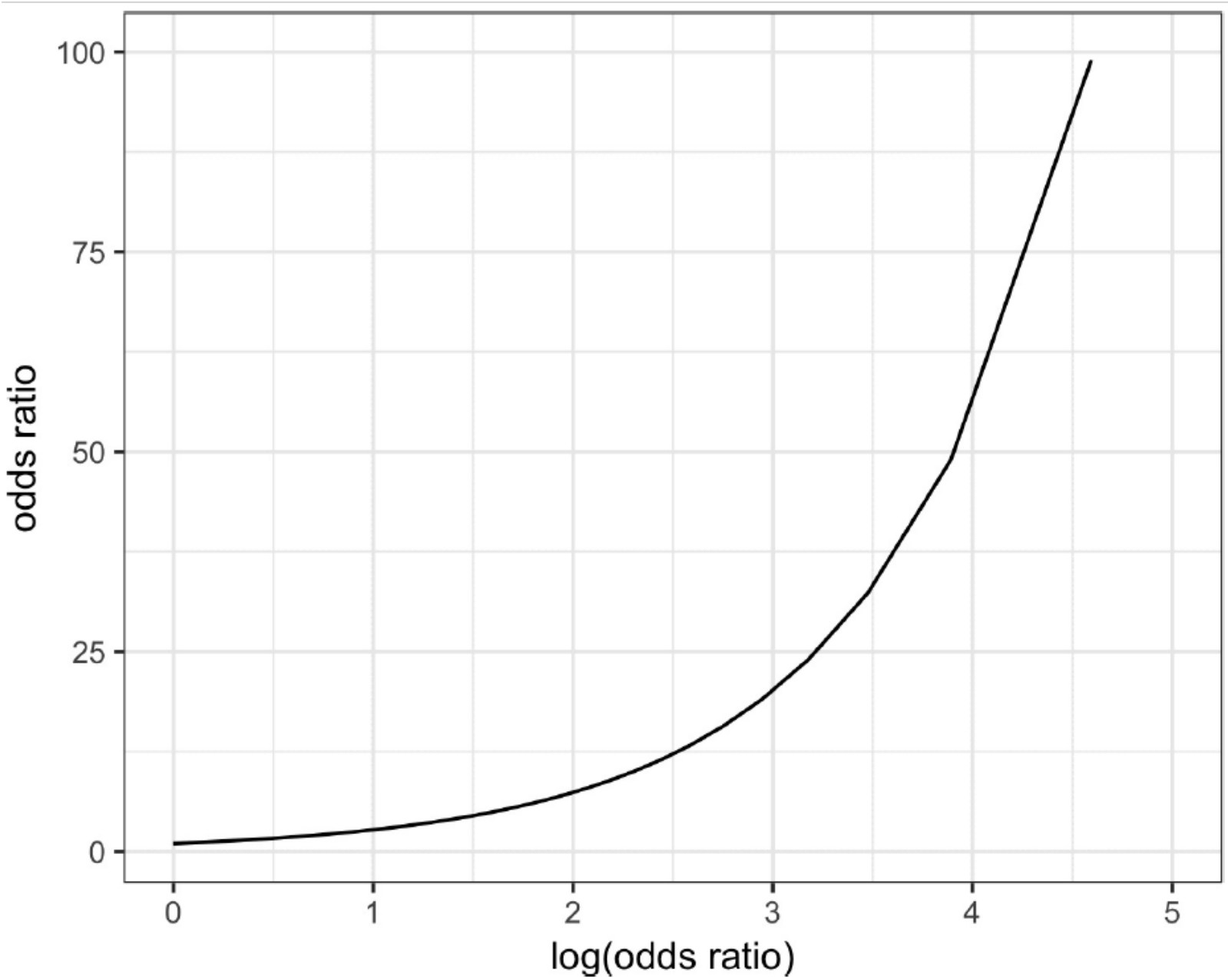


Figure 5-3. The relationship between the odds ratio and the log-odds ratio

Linear and Logistic Regression: Similarities and Differences

- Multiple linear regression and logistic regression share many commonalities. Both assume a parametric linear form relating the predictors with the response.
- Logistic regression differs in two fundamental ways:
  - The way the model is fit (least squares is not applicable)
  - The nature and analysis of the residuals from the model
- Fitting the model:
  - In logistic regression (unlike in linear regression), there is no closed-form solution and the model must be fit using maximum likelihood estimation (MLE).
  - Maximum likelihood estimation is a process that tries to find the model that is most likely to have produced the data we see.
  - In the logistic regression equation, the response is not 0 or 1 but rather an estimate of the log odds that the response is 1.
  - The MLE finds the solution such that the estimated log odds best describes the observed outcome.
  - In the fitting process, the model is evaluated using a metric called deviance:

$$\text{deviance} = -2\log\left(\mathcal{P}_{\hat{\theta}}(\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_n)\right)$$

  - Lower deviance corresponds to a better fit.

Assessing the Model

- Like other classification methods, logistic regression is assessed by how accurately the model classifies new data.
- Along with the estimated coefficients, R reports the standard error of the coefficients (SE), a z-value, and a p-value:

```
summary(logistic_model)

Call:
glm(formula = outcome ~ payment_inc_ratio + purpose_ + home_ +
    emp_len_ + borrower_score, family = "binomial", data = loan_data)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.71430  -1.06806  -0.04482   1.07446   2.11672

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                 1.269822   0.051929  24.453  < 2e-16 ***
payment_inc_ratio           0.082443   0.002485  33.177  < 2e-16 ***
purpose_debt_consolidation  0.252164   0.027409   9.200  < 2e-16 ***
purpose_home_improvement    0.343674   0.045951   7.479 7.48e-14 ***
purpose_major_purchase      0.243728   0.053314   4.572 4.84e-06 ***
purpose_medical             0.675362   0.089803   7.520 5.46e-14 ***
purpose_other               0.592678   0.039109  15.154  < 2e-16 ***
purpose_small_business      1.212264   0.062457  19.410  < 2e-16 ***
home_OWN                    0.031320   0.037479   0.836    0.403
home_RENT                   0.168670   0.021041   8.016 1.09e-15 ***
emp_len_ < 1 Year           0.444892   0.053342   8.340  < 2e-16 ***
borrower_score             -4.638902   0.082433 -56.275  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 64147  on 46271  degrees of freedom
Residual deviance: 58531  on 46260  degrees of freedom
AIC: 58555

Number of Fisher Scoring iterations: 4
```

- Interpretation of the p-value comes with the same caveat as in regression, and should be viewed more as a relative indicator of variable importance than as a formal measure of statistical significance.
- A logistic regression model, which has a binary response, does not have an associated RMSE or R-squared.
- Instead, a logistic regression model is typically evaluated using more general metrics for classification; see “Evaluating Classification Models”.

- Many other concepts for linear regression carry over to the logistic regression setting (and other GLMs).
  - For example, you can use stepwise regression, fit interaction terms, or include spline terms.
  - The same concerns regarding confounding and correlated variables apply to logistic regression (see “Interpreting the Regression Equation”).
  - You can fit generalized additive models (see “Generalized Additive Models”) using the mgcv package

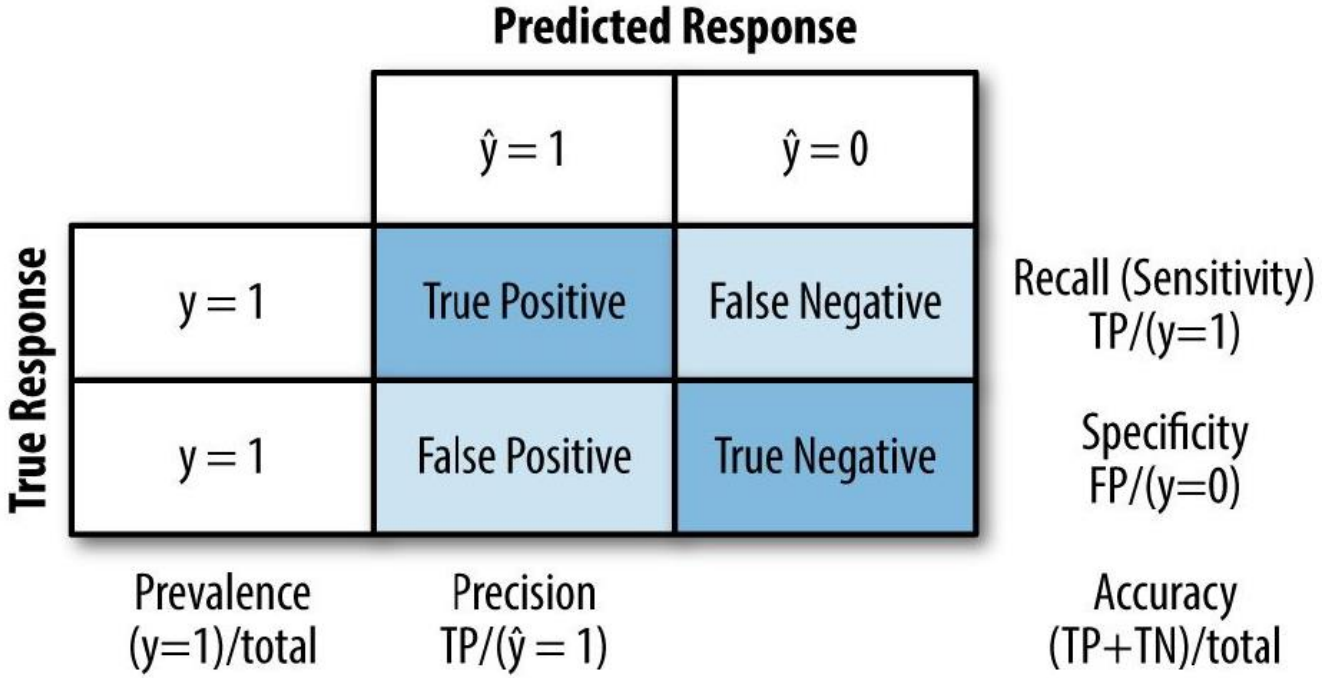
Evaluating Classification Models

- It is common in predictive modeling to try out a number of different models, apply each to a holdout sample (also called a test or validation sample), and assess their performance. Fundamentally, this amounts to seeing which produces the most accurate predictions.
- A simple way to measure classification performance is to count the proportion of predictions that are correct.

$$\text{accuracy} = \frac{\sum \text{TruePositive} + \sum \text{TrueNegative}}{\text{SampleSize}}$$

Confusion Matrix

- The confusion matrix is a table showing the number of correct and incorrect predictions categorized by type of response.
- When 1s are rare, the ratio of false positives to all predicted positives can be high, leading to the unintuitive situation where a predicted 1 is most likely a 0.
- This problem plagues medical screening tests (e.g., mammograms) that are widely applied: due to the relative rarity of the condition, positive test results most likely do not mean breast cancer. This leads to much confusion in the public.



The Rare Class Problem

- In many cases, there is an imbalance in the classes to be predicted, with one class much more prevalent than the other — for example, legitimate insurance claims versus fraudulent ones, or browsers versus purchasers at a website.
- The rare class (e.g., the fraudulent claims) is usually the class of more interest, and is typically designated 1, in contrast to the more prevalent 0s.
- In the typical scenario, the 1s are the more important case, in the sense that misclassifying them as 0s is costlier than misclassifying 0s as 1s.

- In such cases, unless the classes are easily separable, the most accurate classification model may be one that simply classifies everything as a 0.
- For example, if only 0.1% of the browsers at a web store end up purchasing, a model that predicts that each browser will leave without purchasing will be 99.9% accurate. However, it will be useless.
- Instead, we would be happy with a model that is less accurate overall, but is good at picking out the purchasers, even if it misclassifies some non-purchasers along the way.

Precision, Recall, and Specificity

- The precision measures the accuracy of a predicted positive outcome

$$\text{precision} = \frac{\sum \text{TruePositive}}{\sum \text{TruePositive} + \sum \text{FalsePositive}}$$

- The recall, also known as sensitivity, measures the strength of the model to predict a positive outcome - the proportion of the 1s that it correctly identifies

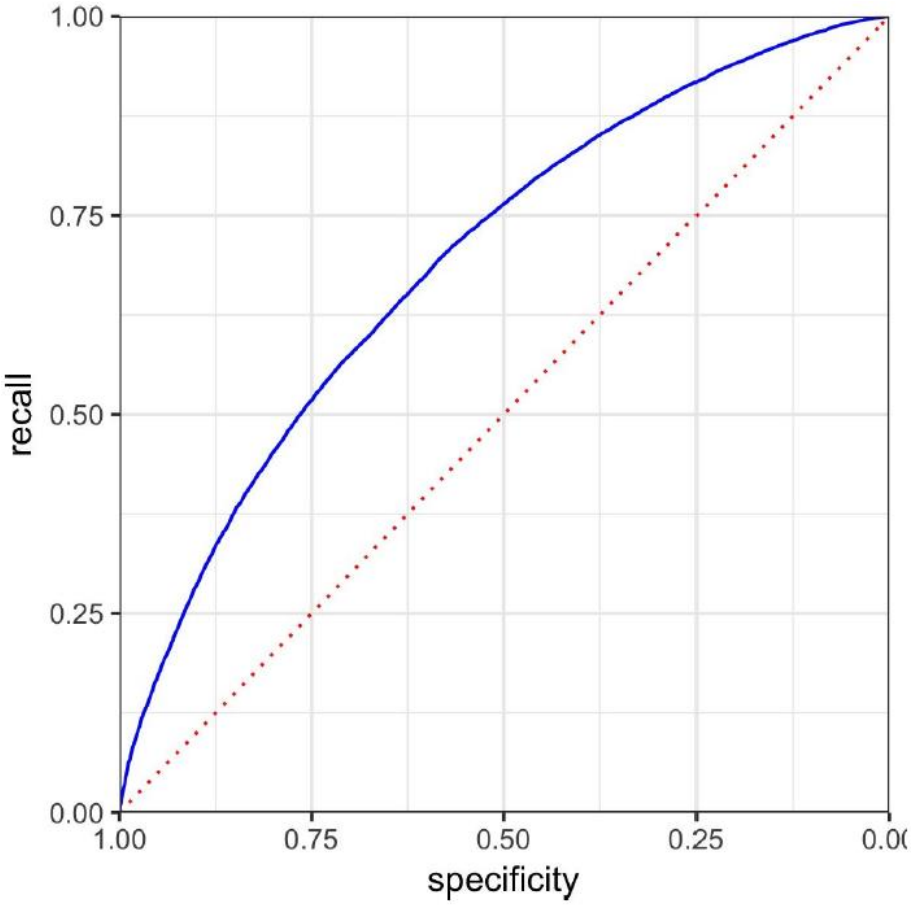
$$\text{recall} = \frac{\sum \text{TruePositive}}{\sum \text{TruePositive} + \sum \text{FalseNegative}}$$

- Another metric used is specificity, which measures a model's ability to predict a negative outcome:

$$\text{specificity} = \frac{\sum \text{TrueNegative}}{\sum \text{TrueNegative} + \sum \text{FalseNegative}}$$

ROC Curve

- You can see that there is a tradeoff between recall and specificity. Capturing more 1s generally means misclassifying more 0s as 1s.
- The metric that captures this tradeoff is the “Receiver Operating Characteristics” curve, usually referred to as the ROC curve.
- The dotted diagonal line corresponds to a classifier no better than random chance.



AUC

- The ROC curve is a valuable graphical tool but, by itself, doesn't constitute a single measure for the performance of a classifier.
- The ROC curve can be used, however, to produce the area underneath the curve (AUC) metric.
- AUC is simply the total area under the ROC curve.
- The larger the value of AUC, the more effective the classifier.
- An AUC of 1 indicates a perfect classifier: it gets all the 1s correctly classified, and doesn't misclassify any 0s as 1s.

Lift

- Using the AUC as a metric is an improvement over simple accuracy, as it can assess how well a classifier handles the tradeoff between overall accuracy and the need to identify the more important 1s.
- But it does not completely address the rare-case problem, where you need to lower the model's probability cutoff below 0.5 to avoid having all records classified as 0.
- In such cases, for a record to be classified as a 1, it might be sufficient to have a probability of 0.4, 0.3, or lower.
- In effect, we end up over-identifying 1s, reflecting their greater importance.
- Changing this cutoff will improve your chances of catching the 1s (at the cost of misclassifying more 0s as 1s). But what is the optimum cutoff?

Strategies for Imbalanced Data

- We look at additional strategies that can improve predictive modeling performance with imbalanced data.
- If you have enough data, as is the case with the loan data, one solution is to under sample (or down sample) the prevalent class, so the data to be modeled is more balanced between 0s and 1s.
- One criticism of the under sampling method is that it throws away data and is not using all the information at hand.
- If you have a relatively small data set, and the rarer class contains a few hundred or a few thousand records, then under sampling the dominant class has the risk of throwing out useful information.
- In this case, instead of down sampling the dominant case, you should oversample (up sample) the rarer class by drawing additional rows with replacement (bootstrapping).
- You can achieve a similar effect by weighting the data. Many classification algorithms take a weight argument that will allow you to up/down weight the data.
- Note that weighting provides an alternative to both up sampling the rarer class and down sampling the dominant class.

Cost-Based Classification

- In practice, accuracy and AUC are a poor man's way to choose a classification rule. Often, an estimated cost can be assigned to false positives versus false negatives, and it is more appropriate to incorporate these costs to determine the best cutoff when classifying 1s and 0s.
- For example, suppose the expected cost of a default of a new loan is **C** and the expected return from a paid-off loan is **R**.
- Then the expected return for that loan is:

$$\text{expected return} = P(Y = 0) \times R + P(Y = 1) \times C$$

- Instead of simply labeling a loan as default or paid off, or determining the probability of default, it makes more sense to determine if the loan has a positive expected return.
- For example, a smaller value loan might be passed over in favor of a larger one with a slightly higher predicted default probability.