1495 UNIVERSITY OF ABERDEEN

CELEBRATING
525 YEARS
1495 – 2020

ABERDEEN 2040

# Classification

Data Mining & Visualisation

Lecture 10

2025

# Today

- Classification
- Logistic Regression
- Evaluating a Classification Model

# Regression Vs Classification Problems

In the previous lecture, we outlined two broad types of problems that tend to come up in supervised learning:

- **Regression problems**, where our DV is quantitative.
- **Classification problems**, where our DV is categorical.

# Regression Vs Classification Problems

We have already discussed one technique for dealing with regression problems: Linear Regression.

In this lecture we’re going to focus on one technique for dealing with classification problems: *Logistic Regression.*

# A Quick Note on the Terminology

But before we start, just a quick point to clarify.

You might have noticed that we differentiate between classification problems and regression problems…

...before then going on to say that we’ll solve a **classification** problem with logistic **regression**

# A Quick Note on the Terminology

Importantly, in the context of statistical learning, the terms ‘classification problem’ and ‘regression problem’ are simply used to describe whether our DV is categorical or quantitative.

Separately, we have the family of *Regression techniques*, which are one approach to solving **regression problems**.

However, this family of techniques can also be modified and adapted to work within the context of classification.

# Logistic Regression

# Logistic Regression

Logistic regression is fairly similar to linear regression:

Our aim is to build a model such that our DV can be estimated from our IV(s).

However, in the case of **logistic regression**, we model the probability of a binary response (such that $0 \leq \Pr(DV) \leq 1$).

# Logistic Regression

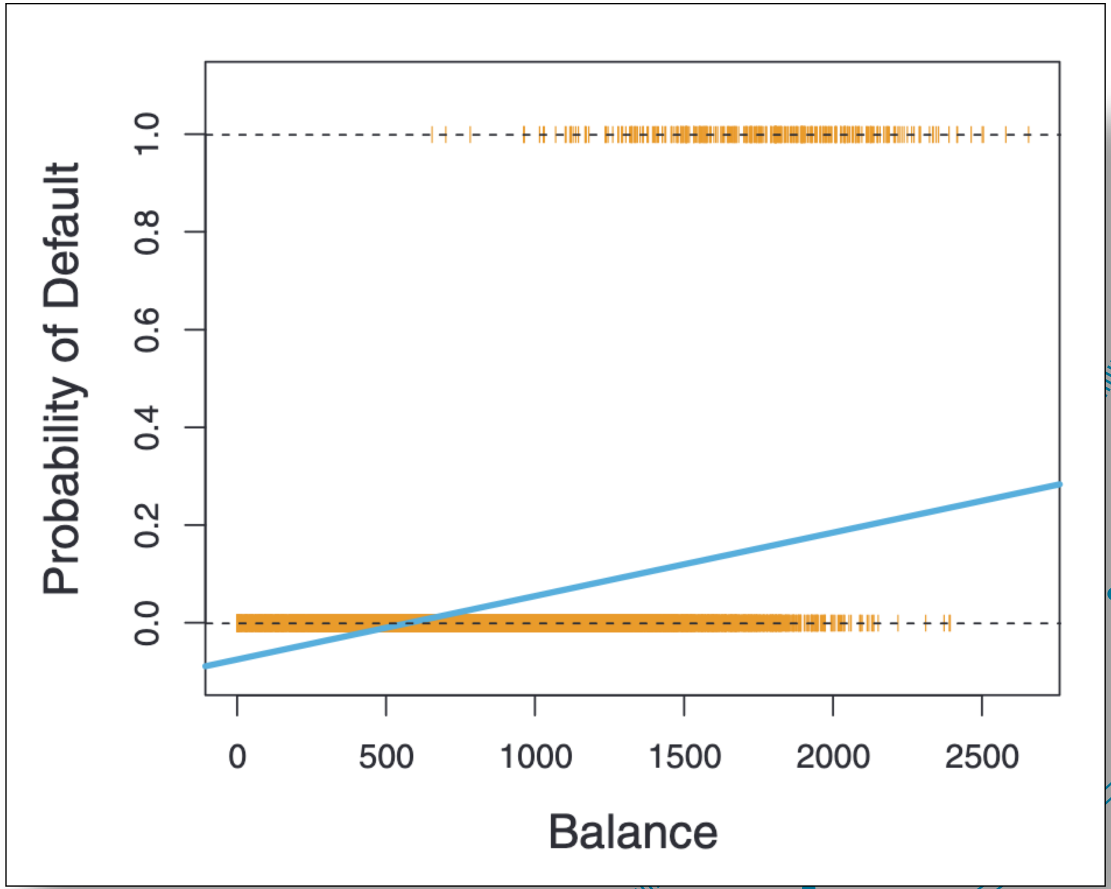
In other words, logistic regression let's us use our IV(s) to estimate the probability that our DV will be 1 (or 0).

We can use this approach in <u>some</u> classification tasks – as long as our IV is binary (it only has two possible outcomes):

- E.g., {True, False}; {0, 1}; {Spam, Not Spam} etc…

# Logistic Regression

- Why can’t we just use ordinary linear regression for this?
- If we do so, our linear model can predict < 0 and > 1
- For estimating probabilities, we don’t want this!



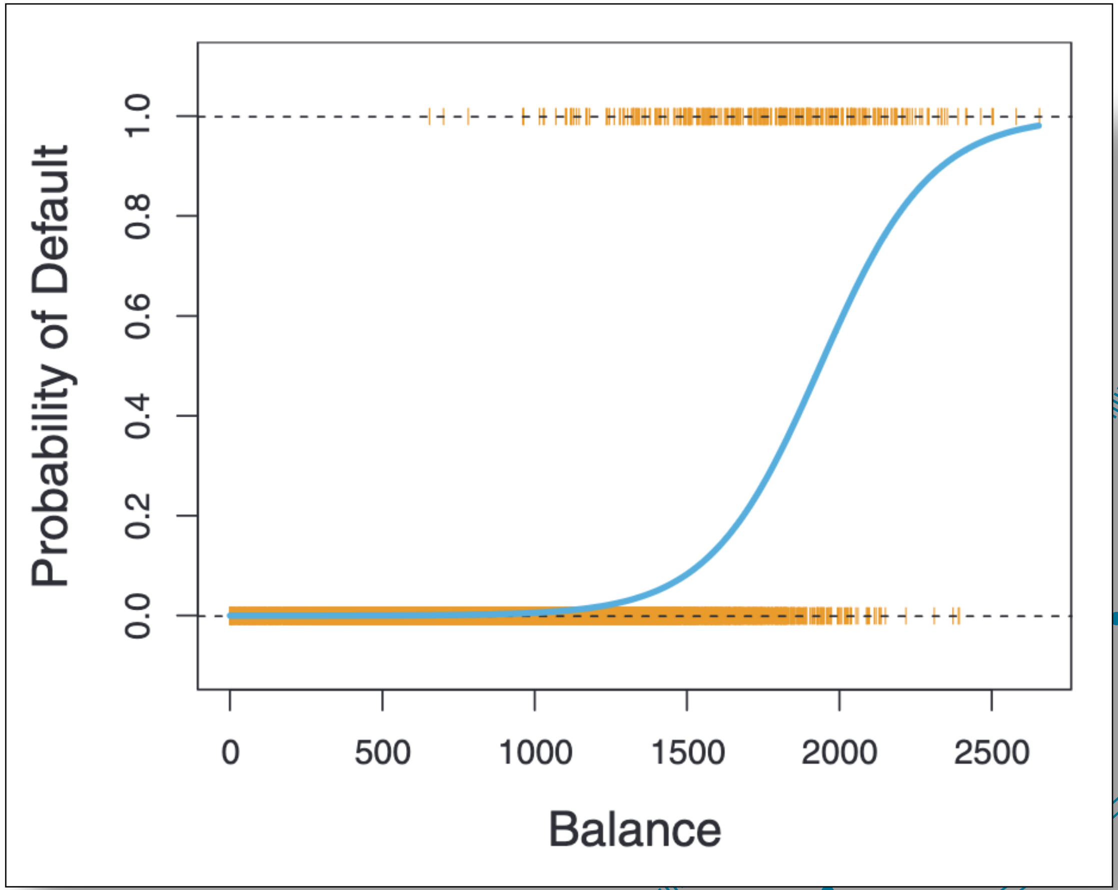
Source: Introduction to Statistical Learning. James et al. (2013). Springer.

# Logistic Regression

So if 'ordinary' linear regression involves a 'line of best fit'…

…what does the logistic regression process for a binary response variable look like?

This shape is what's known as the **Sigmoid Curve**.



Source: Introduction to Statistical Learning. James et al. (2013). Springer.

# Logistic Regression: Under the Hood

# Logistic Regression

More specifically, we're looking to estimate:

$$Pr(y = 1|x_1, ...x_n)$$

In other words, the probability that our DV is estimated as a particular class (in this case, y = 1), given our IV(s).

We typically estimate y = 1 whenever this probability > 0.5 (though we could technically choose a lower threshold).

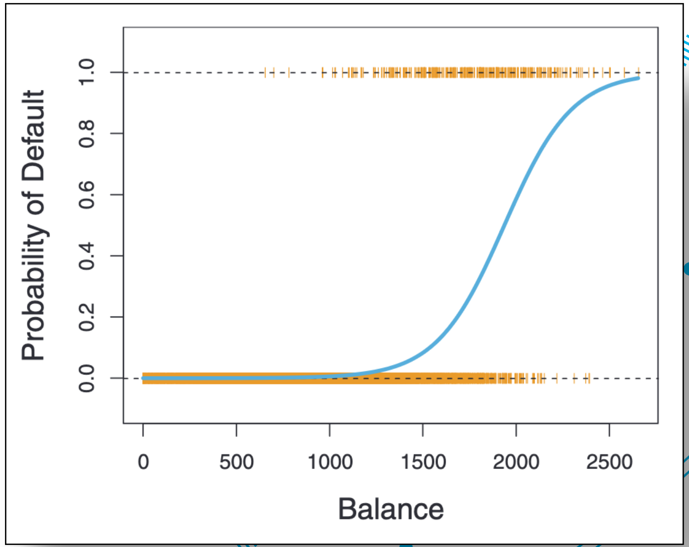
# Logistic Function

To transform our data into a probability (where 0 ≤ Pr ≤ 1) that the DV belongs to a particular class, we use a ‘logistic function’:

$$f(x) = \frac{1}{1+e^{-x}}$$

**(Note, you don’t need to learn this formula)**

This transformation is what gives us a sigmoid (or S-shaped) curve – and *logistic* regression its name.



Source: Introduction to Statistical Learning. James et al. (2013). Springer.

# Logistic Regression

So instead of using the linear regression formula:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

We now use:

$$P(y=1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n)}}$$

# What Can We Do With Logistic Regression?

# What Can We Do With Logistic Regression

Similar to linear regression, with our logistic regression model:

- We can use the model to predict the probability of Y being class 1 (prediction).

- We can also use the model to understand how the IVs influence the Y value (interpretation) – though this is a little bit more complex due to the non-linear nature of the model.

# Interpreting Logistic Regression

Since Logistic Regression uses the Logistic function, we can’t simply interpret the model as we do with linear regression.

Instead, our model gives us ‘odds ratios’, which are a bit more challenging to interpret and work with.

# Interpreting Logistic Regression

$$P(y=1) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_n x_n)}}$$

Simply put, if our $\beta_1$ coefficient weight was 1.5, that would tell us that an increase in $x_1$ by a value of 1 would increase the odds of y = 1 by 4.48 times:

$$Odds\ of\ (y=1) = e^{1.5} \approx 4.48\ times$$

**<u>But note that you won't be expected to calculate this in the exam.</u>**

# Evaluating a Classification Model

# Evaluating a Classification Model for Fit

In logistic regression, variations of (pseudo) $R^2$ calculations exist, serving a similar function to that of $R^2$.

These can be useful for quantifying the variance explained by your model,

However, quite often what you'll see in classification contexts, particularly when *predicting*, is a ***confusion matrix***.

# Confusion Matrix

Consider that when predicting with a binary classification model, you can have four types of results:

- **True positive** (correctly predicted a 1 as a 1)
- **False positive** (incorrectly predicted a 0 as a 1)
- **True negative** (correctly predicted a 0 as a 0)
- **False negative** (incorrectly predicted a 1 as a 0)

# Confusion Matrix

A confusion matrix is a table which reports these instances.

It can also be used to calculate a number of different metrics.

These include…

| | | Predicted | |
|---|---|---|---|
| | | 1 | 0 |
| Actual | 1 | TP | FN |
| | 0 | FP | TN |

# Confusion Matrix – Accuracy

**Accuracy**

= (TP + TN) / (TP + FP + FN + TN)

The proportion of all outcomes that were predicted correctly.

|  |  | Predicted |  |
|---|---|---|---|
|  |  | 1 | 0 |
| Actual | 1 | TP | FN |
|  | 0 | FP | TN |

# Confusion Matrix – Precision

**Precision**

= TP / (TP + FP)

The proportion of outcomes that were predicted as {1} that were predicted correctly.

Also known as **Positive Predictive Value**

| | | Predicted | |
|---|---|---|---|
| | | 1 | 0 |
| Actual | 1 | TP | FN |
| | 0 | FP | TN |

# Confusion Matrix – Recall / Sensitivity

**Recall / Sensitivity**

= TP / (TP + FN)

The proportion of outcomes that were actually {1} that were predicted correctly.

Also known as **True Positive Rate**

|  |  | Predicted |  |
|---|---|---|---|
|  |  | 1 | 0 |
| Actual | 1 | TP | FN |
|  | 0 | FP | TN |

# Confusion Matrix – Specificity

**Specificity**

= TN / (TN + FP)

The proportion of outcomes that were actually {0} that were predicted correctly.

Also known as **True Negative Rate**

| | | Predicted | |
|---|---|---|---|
| | | 1 | 0 |
| Actual | 1 | TP | FN |
| | 0 | FP | TN |

# Confusion Matrix – False Positive Rate

**False Positive Rate**

= FP / (FP + TN)

The proportion of outcomes that were predicted as {1} that were predicted incorrectly.

Also known as **Type I Error Rate**

| | | **Predicted** | |
|---|---|---|---|
| | | **1** | **0** |
| **Actual** | **1** | **TP** | **FN** |
| | **0** | **FP** | **TN** |

# Confusion Matrix – False Negative Rate

**False Negative Rate**

= FN / (FN + TP)

The proportion of outcomes that were predicted as {0} that were predicted incorrectly.

Also known as **Type II Error Rate**

| | | **Predicted** | |
|---|---|---|---|
| | | **1** | **0** |
| **Actual** | **1** | **TP** | **FN** |
| | **0** | **FP** | **TN** |

# Confusion Matrix – $F_1$ Score

The **$F_1$ Score** is a good ‘overall’ metric, calculated as the harmonic mean of precision and recall:

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

We can also simplify this equation to:

$$F_1 = \frac{2 \cdot \text{TP}}{2 \cdot \text{TP} + \text{FP} + \text{FN}}$$

| | | Predicted | |
|---|---|---|---|
| | | **1** | **0** |
| **Actual** | **1** | **TP** | **FN** |
| | **0** | **FP** | **TN** |

# Confusion Matrix – Warning!

| | | Predicted | |
|---|---|---|---|
| | | 1 | 0 |
| Actual | 1 | TP | FN |
| | 0 | FP | TN |

| | | Actual | |
|---|---|---|---|
| | | 1 | 0 |
| Predicted | 1 | TP | FP |
| | 0 | FN | TN |

| | | Predicted | |
|---|---|---|---|
| | | 0 | 1 |
| Actual | 0 | TN | FP |
| | 1 | FN | TP |

| | | Actual | |
|---|---|---|---|
| | | 0 | 1 |
| Predicted | 0 | TN | FN |
| | 1 | FP | TP |

Note that you will sometimes see confusion matrices transposed, or with different placements of 0 and 1.

All versions are valid, and all versions are commonly used.

**Always check carefully which one you are dealing with!**

# ROC Curve

When evaluating binary classifiers, we also have ROC curves.

A ROC (Receiver Operating Characteristic) curve is a visual way to show the predictive performance of binary classifiers.

It plots the True Positive Rate against the False Positive Rate at different classification thresholds.

Ideally, you want a ROC curve close to the top left corner!

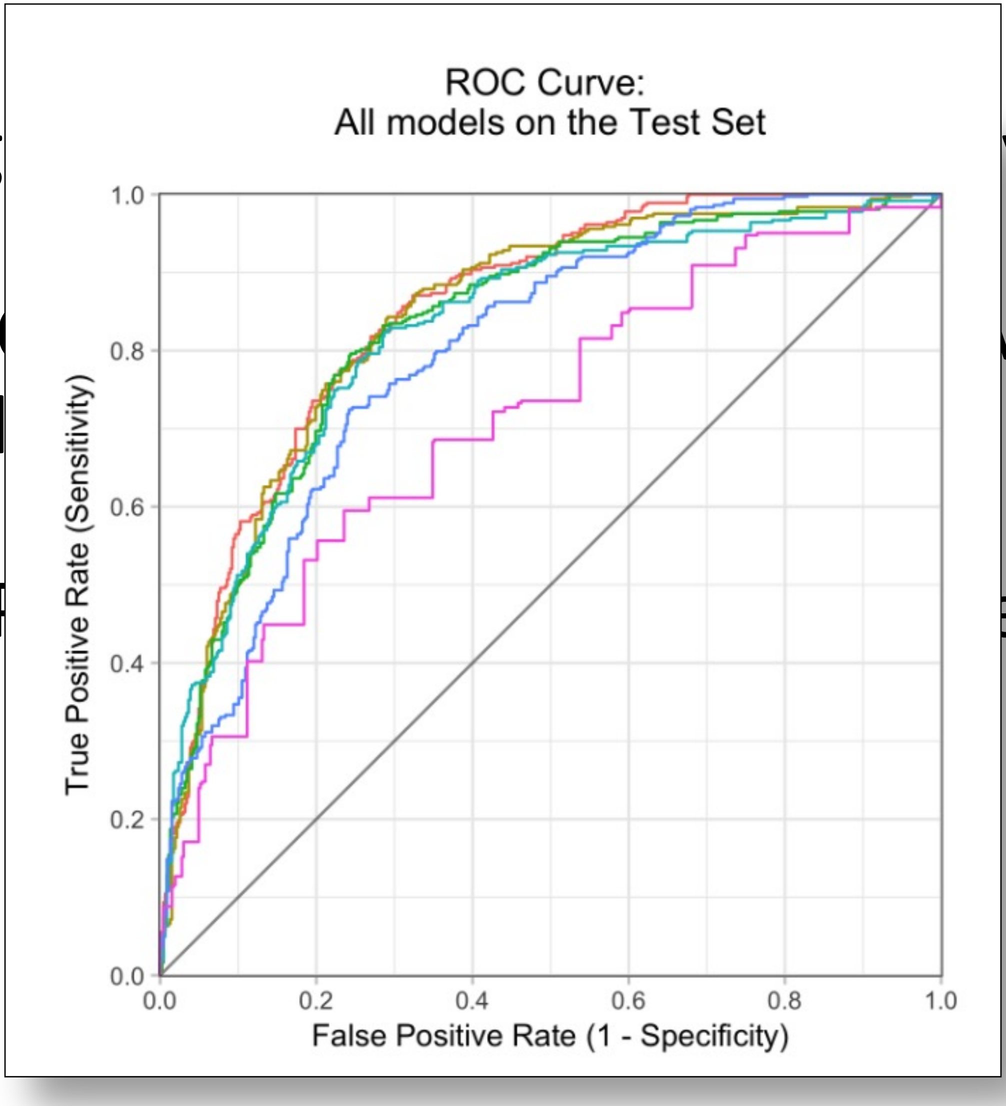
# ROC Curve

When evaluating ve ROC curves.

A ROC (Receiver ( ve is a visual way
to show the pred classifiers.

It plots the True F e Positive Rate at
different classific

Ideally, you want left corner!



# Different Types of Logistic Regression

# Different Types of Logistic Regression

Similar to linear regression, there are different variations of logistic regression (which we won’t cover in detail), including:

- Multiple logistic regression can be used with multiple IVs
- Multinomial Logistic Regression (can model 3+ classes)
- Ordinal Logistic Regression (can model ordinal variables)