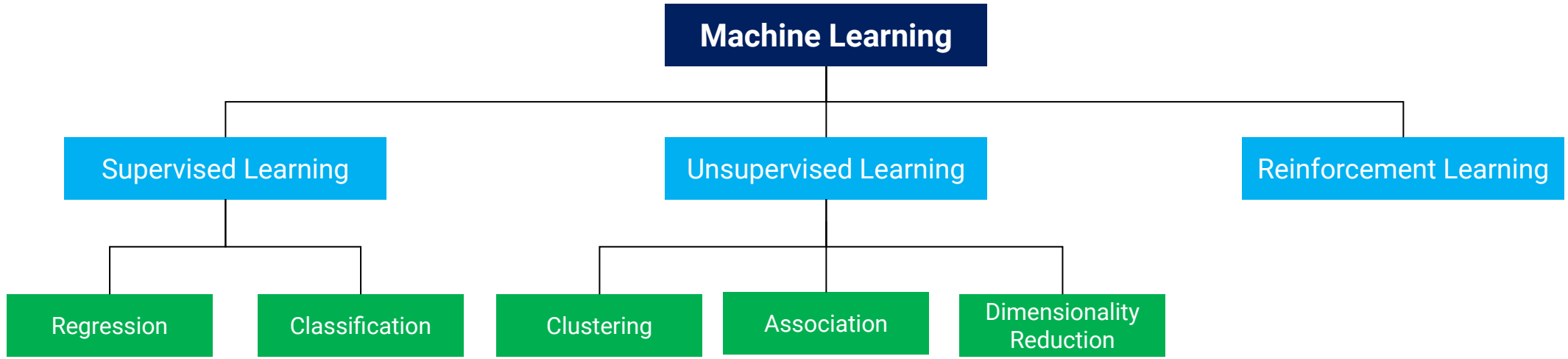
AI200: APPLIED MACHINE LEARNING

LOGISTIC REGRESSION

OVERVIEW & LITERATURE OF MACHINE LEARNING



1. Spectrum of supervision

2. Types of problems

3. Types of algorithm / models

Regression	Classification	Clustering	Association	Dimensionality Reduction
Linear regression	Logistic Regression	Neural Networks	Apriori Algorithm	Neural Networks
Polynomial regression	KNN	K-Means	FP-Growth Algorithm	SVD
Neural Networks	Naïve Bayes	Hierarchical Clustering	Eclat Algorithm	PCA
Decision Tree	SVM	Hidden Markov Models		
Random Forest	Neural Networks	Gaussian Mixture		
XGBoost	Decision Tree			
	Random Forest			
	XGBoost			

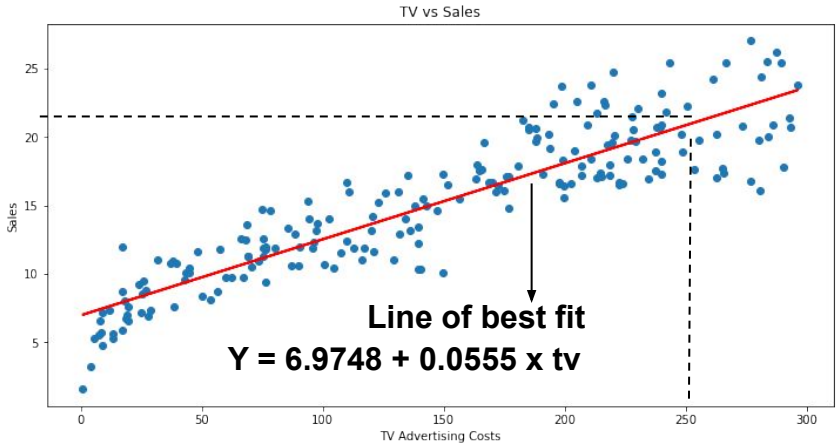
4. Types of applications

Forecasting	Fraud Detection	Recommendations	Market Basket Analysis	Feature Elicitation	Real-time Decisions
Predictions	Image Classification	Targeted Marketing		Structure Discovery	Game AI
Process optimization	Customer Retention	Customer Segmentation		Meaningful Compression	Learning Tasks
New Insights	Diagnostics			Big Data Visualisation	Skill Acquisition
					Robot Navigation

WHY DO WE NEED LOGISTIC REGRESSION?

- Recall that in simple linear regression, we use OLS to fit a line on the data, and thereafter use that line to predict outcomes?

Linear Regression (Regression Problem)

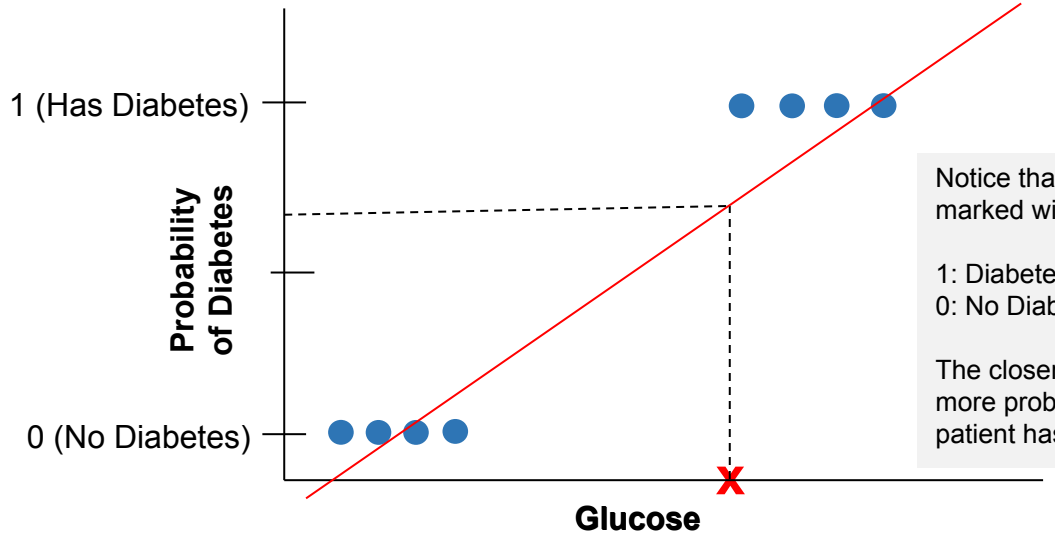


For a new data point with **TV ad cost of $250**, using the line of best fit the model predicts that the Sales is likely **21**

WHY DO WE NEED LOGISTIC REGRESSION?

- What if we tried to **apply linear regression to solve classification problems** as well?

Linear Regression (Classification Problem)



Notice that our **y-axis** is marked with

1: Diabetes
0: No Diabetes

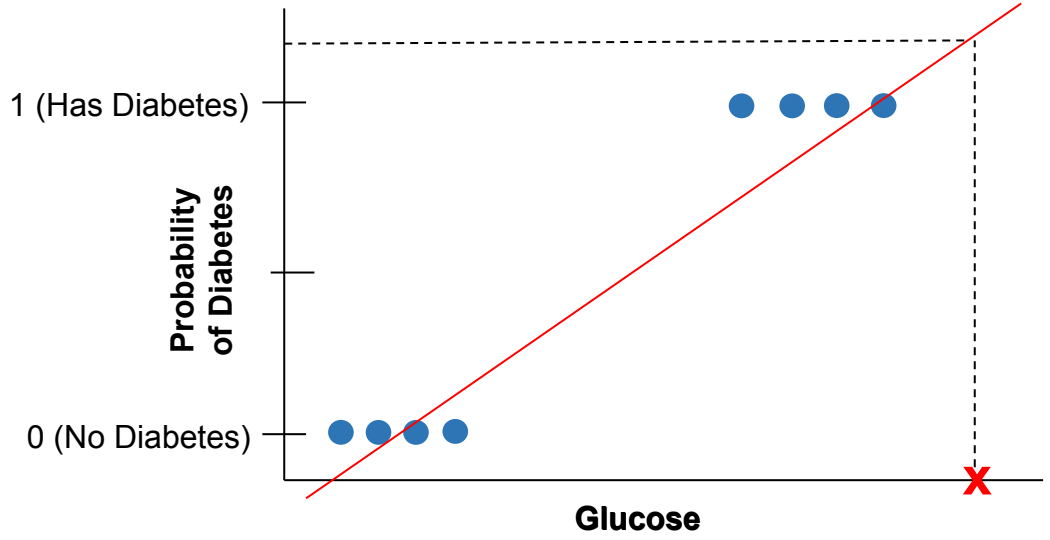
The closer the value is to 1, the more probable it is that the patient has diabetes.

The model seems to work well enough for now. Let's test it a bit more

WHY DO WE NEED LOGISTIC REGRESSION?

- Observes what happens we set our glucose level to the extreme. Our linear regression model predicts that the probability of the person having diabetes is more than 1, which does not make sense:
  - Linear regression **produces outputs of beyond the range of 0 and 1**, which is not suitable for classification problems.
  - But by extending the idea of linear regression, we can make it work!

Linear Regression (Classification Problem)

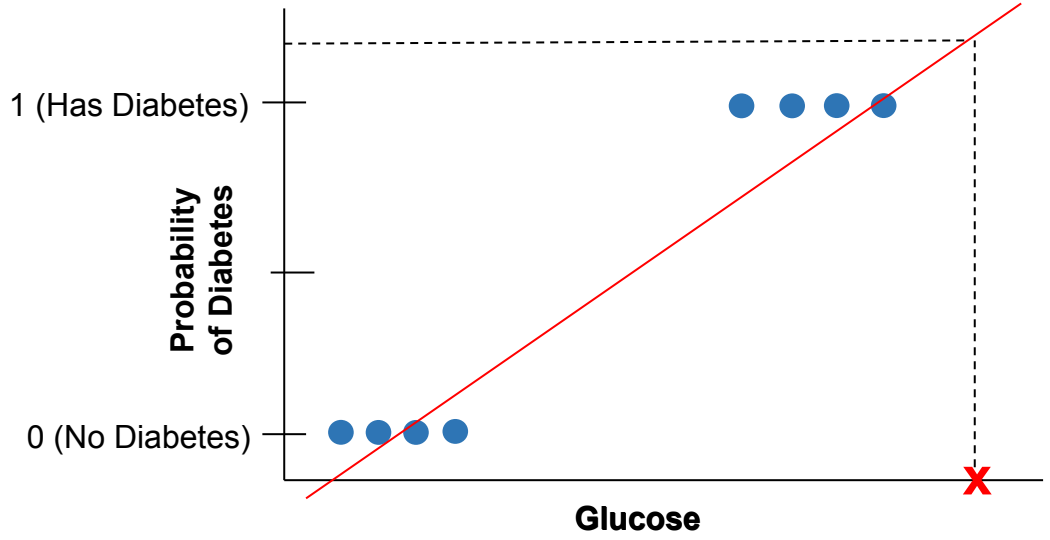


WHAT IS LOGISTIC REGRESSION: LAYMAN INTUITION

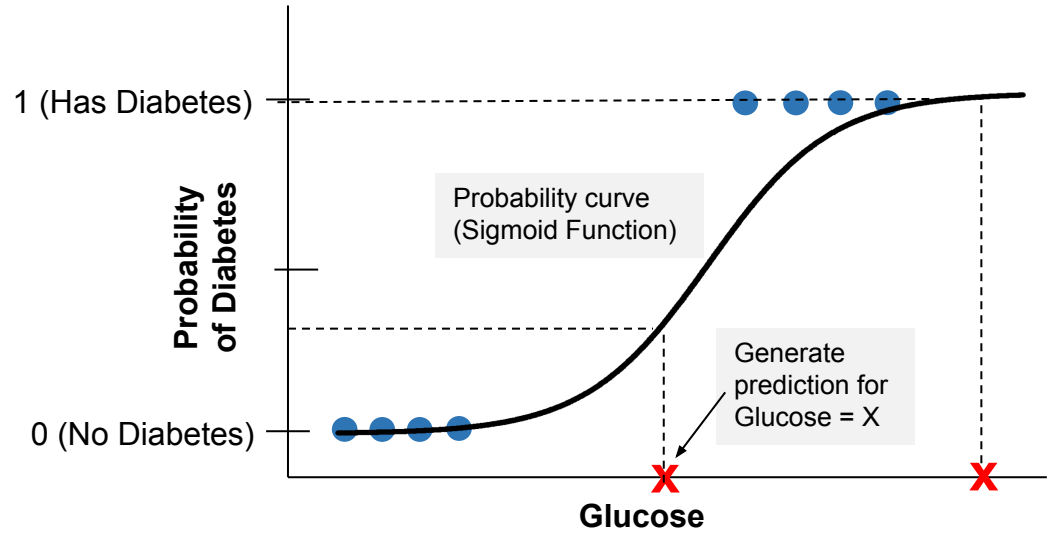
- *Training Step:* Instead of fitting the data on a straight line, we **fit the data on a s-curve (known as sigmoid function)** that is **constrained between 0 to 1** on the Y-axis.

- *Prediction Step:* When a patient with a glucose level of X walks in to get checked for diabetes, we provide this input to the **Logistic Regression model** to predict whether he has diabetes or not.

Linear Regression (Classification Problem)



Logistic Regression (Classification Problem)



The logistic regression returns a value between 0 to 1 (also known as a probability score). Notice that it is now able to handle extremely high glucose levels, unlike linear regression.

To achieve a clear outcome of 0 or 1 for classification output, let's first **set an arbitrary threshold of 0.5**. When the probability score of diabetes is less than the threshold of 0.5, we predict the patient will not have diabetes.

WHAT IS LOGISTIC REGRESSION: LAYMAN INTUITION

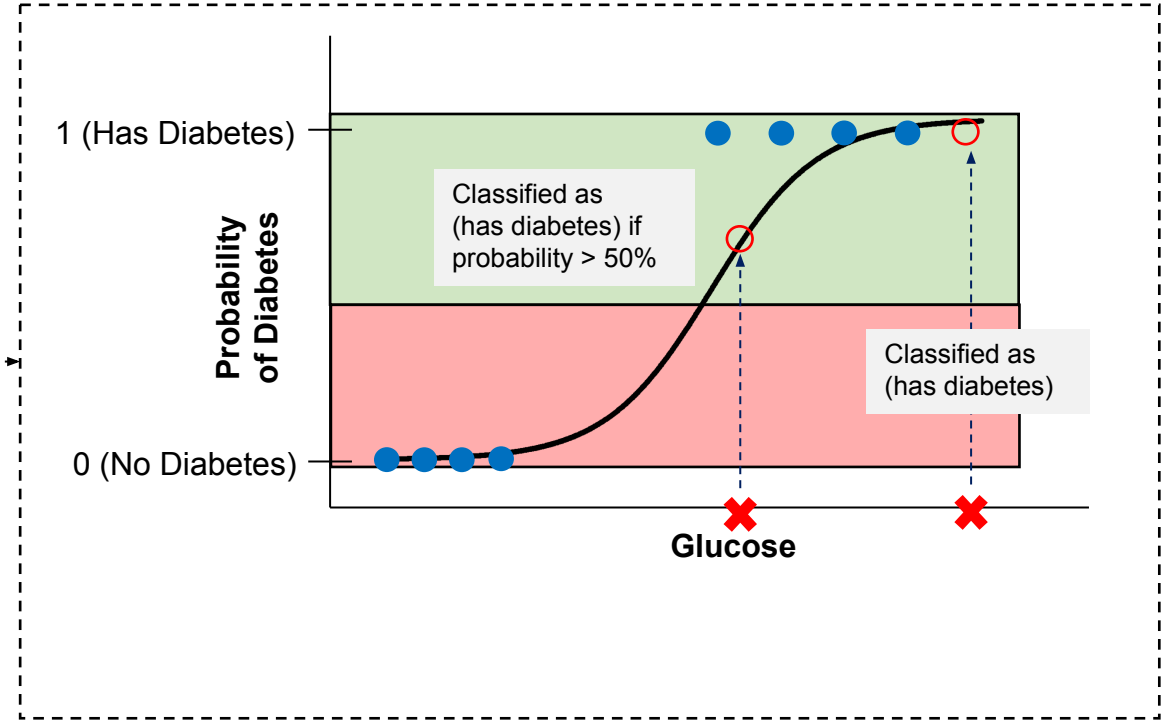
- In summary, Logistic Regression allows us to use a **logistic function** to learn from the dataset
  - The key property of a logistic function (or sigmoid function) is that it produces a s-curve **constrained between the values of 0 to 1**.

Features **Outcome**

Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
148	72	35	0	33.6	0.627	50	1
85	66	29	0	26.6	0.351	31	0
183	64	0	0	23.3	0.672	32	1
89	66	23	94	28.1	0.167	21	0
137	40	35	168	43.1	2.288	33	1

LOGISTIC FUNCTION

Here we **use a logistic function** to fit the data to produce the s-curve constrained between the values of 0 to 1.



WHAT IS LOGISTIC REGRESSION: LAYMAN INTUITION

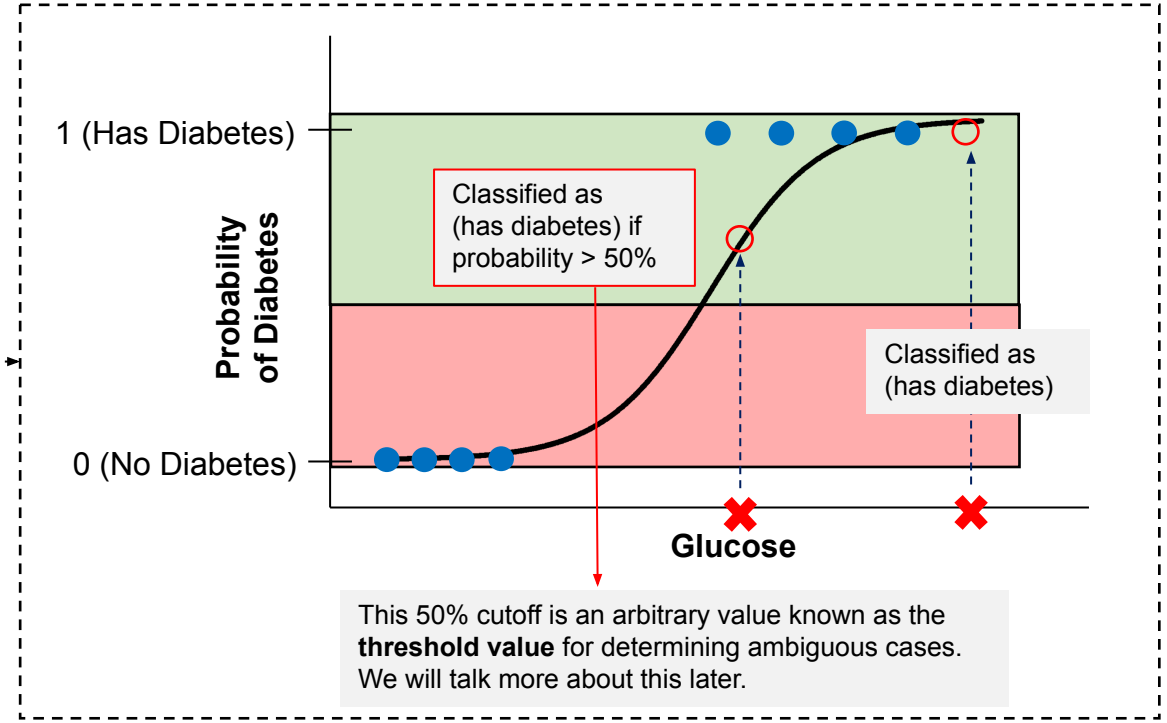
- In summary, Logistic Regression allows us to use a **logistic function** to learn from the dataset
  - The key property of a logistic function (or sigmoid function) is that it produces a s-curve **constrained between the values of 0 to 1**.
  - For now, we use an **arbitrary threshold value of 0.5**: if probability score exceeds the threshold, we classify that the patient has diabetes.

Features **Outcome**

Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
148	72	35	0	33.6	0.627	50	1
85	66	29	0	26.6	0.351	31	0
183	64	0	0	23.3	0.672	32	1
89	66	23	94	28.1	0.167	21	0
137	40	35	168	43.1	2.288	33	1

LOGISTIC FUNCTION

Here we **use a logistic function** to fit the data to produce the s-curve constrained between the values of 0 to 1.



LOGISTIC REGRESSION

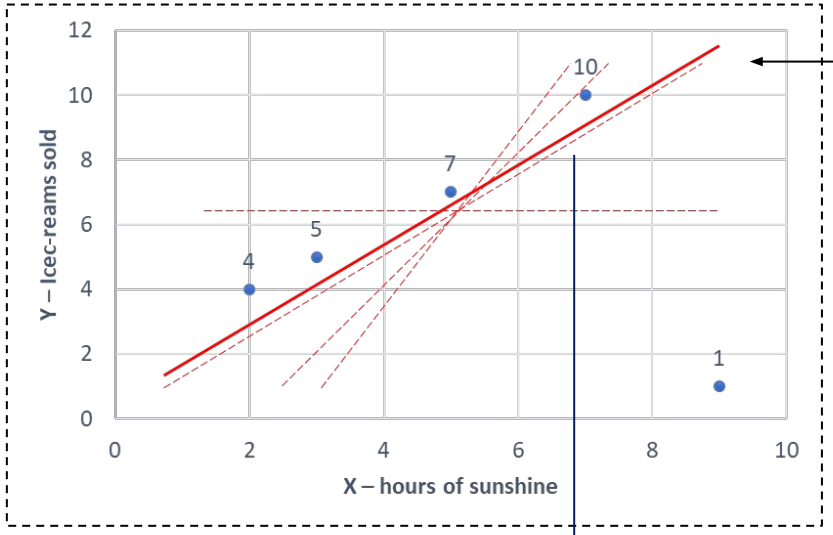
MECHANISM BEHIND MODEL

MECHANISM BEHIND LINEAR REGRESSION: ORDINARY LEAST SQUARES

- Recall that in linear regression, we performed the **Ordinary Least Squares iteratively** to find the most optimal for **α** and **β.**
  - For every set of **α** and **β** values, we will generate a regression line
  - We generate many different regression lines and measure the error of each regression line
  - We stop when we derive at a set of **α** and **β** values that gives a regression line with acceptable rate of error

Linear Regression (Best-fit)

$$y_i = \alpha + \beta x_i + \varepsilon_i$$



All the dotted lines are examples of the different regression lines generated as we seek to find the "best" regression line

We will repeatedly calculate the RSS for each set of α and β values

MECHANISM BEHIND LOGISTIC REGRESSION: MAXIMUM LIKELIHOOD ESTIMATION

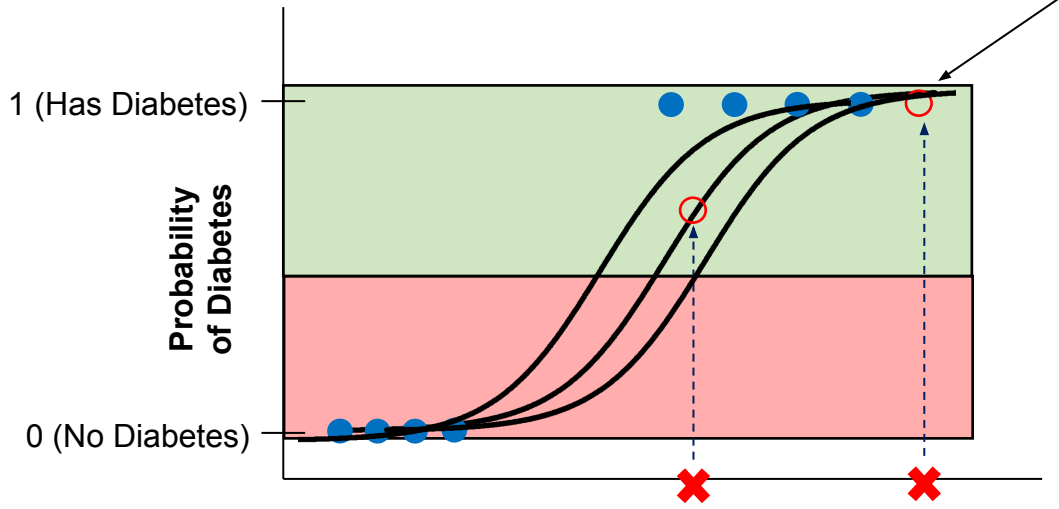
- The mechanism for logistic regression is similar. In this case we use a method called **Maximum Likelihood Estimation**.

Logistic Regression (Sigmoid Function)

$$\text{logOdds}(Y = 1) = \beta_0 + \beta_1 X_1$$

or

$$y = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$



Above is the formula for the sigmoid function of Logistic Regression.

We prefer this equation as it is easier to interpret than the left one.

MECHANISM BEHIND LOGISTIC REGRESSION: MAXIMUM LIKELIHOOD ESTIMATION

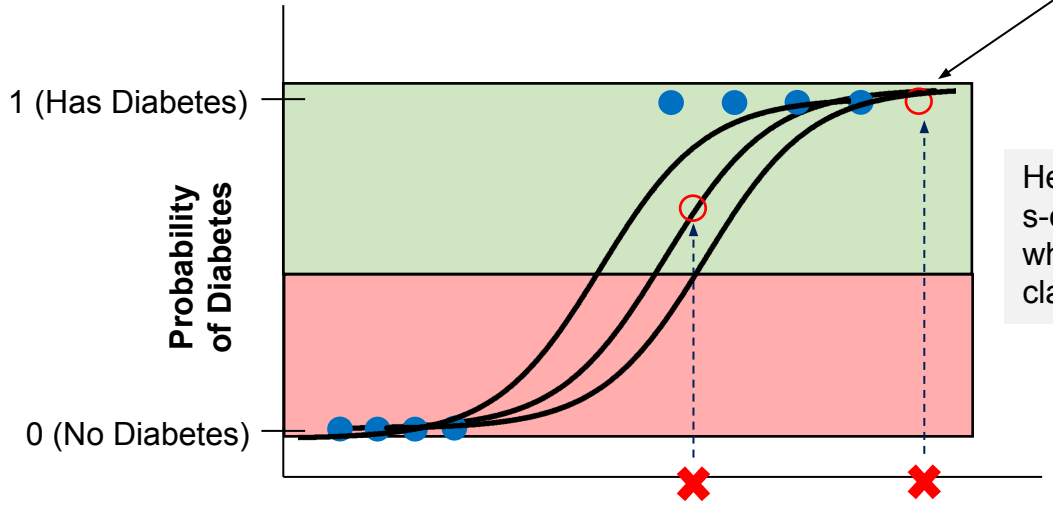
- **Maximum Likelihood Estimation** works through solving a math equation, but to help you visualise how it works:
  - **Step 1**: Sub in random values into β_0 and β_1 of the general formula to generate random sigmoid functions (s-curve)
  - **Step 2**: Identify how many % of outcomes was correctly classified by the sigmoid function
  - **Step 3**: We iteratively repeat step 1-2 to find the optimal value for β_0 and β_1 which has the maximum likelihood, i.e., so that the observed data is most probable.
  - **Step 4**: We stop this process once we have the optimal β_0 and β_1 with the least classification error rate.

Logistic Regression (Sigmoid Function)

$$\text{logOdds}(Y = 1) = \beta_0 + \beta_1 X_1$$

or

$$y = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$



Here we generate multiple s-curves, and try to find out which leads to the best classification results.

LINEAR REGRESSION

MULTIPLE LOGISTIC REGRESSION

MULTIPLE LOGISTIC REGRESSION

- Similar to linear regression, we can use more than one feature to predict an outcome in logistic regression.
  This is known as a multiple logistic regression model.

- Essentially you are only adding additional features to the equation (we re-arranged the formula to a more general form):

$$\hat{p} = \frac{\exp(b_0 + b_1X_1 + b_2X_2 + ... + b_pX_p)}{1 + \exp(b_0 + b_1X_1 + b_2X_2 + ... + b_pX_p)}$$ ***,where p = probability of outcome***

- **You need not worry much about the formula;** our focus is on the intuition behind the algorithm and not the math of the model.
  You will see during our hands-on session later that building a logistic regression model is as straightforward as calling library functions :)

LOGISTIC REGRESSION

STRATEGY FOR SELECTING A THRESHOLD VALUE

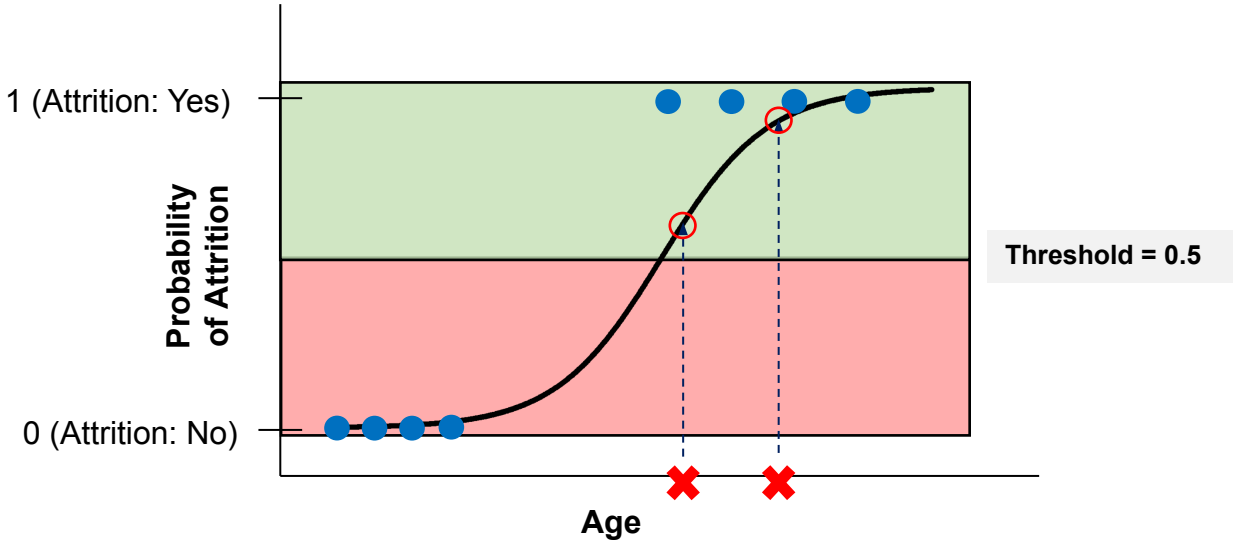
STRATEGY FOR SELECTING A THRESHOLD VALUE: WHY IS THERE A NEED FOR THIS?

- When using logistic regression to generate predictions, recall that we **produce <u>probabilities between 1 and 0</u>, rather than 1s and 0s.**
  - Hence, we need to determine a **<u>threshold value</u>** to turn the probability into a classification outcome of 1s and 0s

Age	Marital Status	Monthly Income	...	Job Satisfaction	...	Years at Company	Attrition
33	Single	4400	...	4	...	5	0.85
37	Married	3300	...	4	...	2	0.55

if 0.5 threshold

Age	Marital Status	Monthly Income	...	Job Satisfaction	...	Years at Company	Attrition
33	Single	4400	...	4	...	5	1
37	Married	3300	...	4	...	2	1

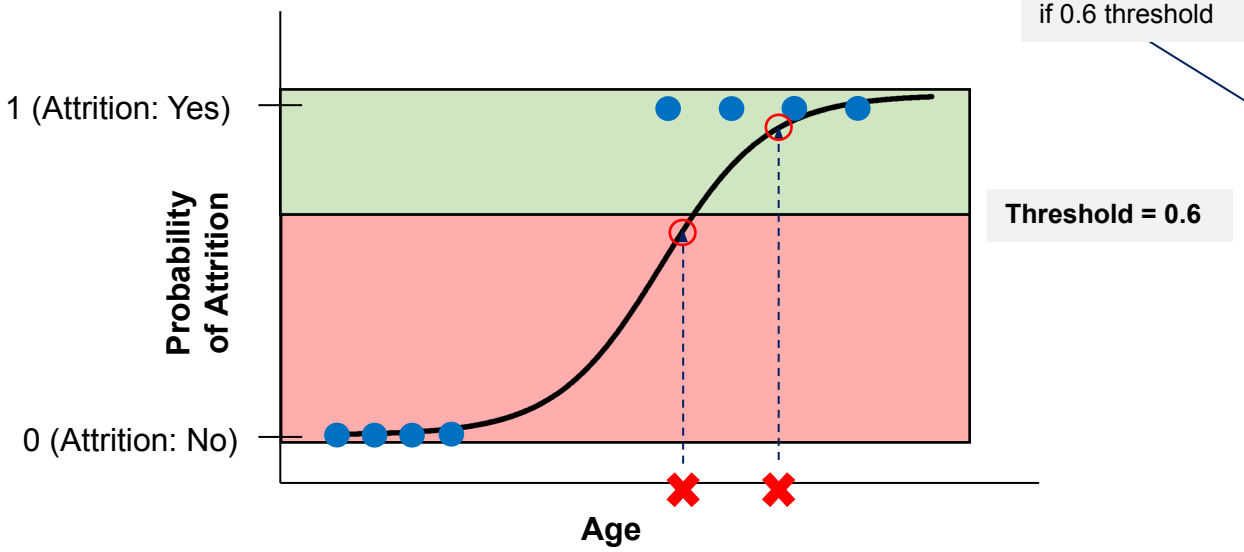


For example, if we use age as a predictor of attrition

STRATEGY FOR SELECTING A THRESHOLD VALUE: WHY IS THERE A NEED FOR THIS?

- When using logistic regression to generate predictions, recall that we **produce <u>probabilities between 1 and 0</u>, rather than 1s and 0s.**
  - Hence, we need to determine a **<u>threshold value</u>** to turn the probability into a classification outcome of 1s and 0s

Age	Marital Status	Monthly Income	...	Job Satisfaction	...	Years at Company	Attrition
33	Single	4400	...	4	...	5	0.85
37	Married	3300	...	4	...	2	0.55



if 0.6 threshold

Threshold = 0.6

Age	Marital Status	Monthly Income	...	Job Satisfaction	...	Years at Company	Attrition
33	Single	4400	...	4	...	5	1
37	Married	3300	...	4	...	2	1

Age	Marital Status	Monthly Income	...	Job Satisfaction	...	Years at Company	Attrition
33	Single	4400	...	4	...	5	1
37	Married	3300	...	4	...	2	0

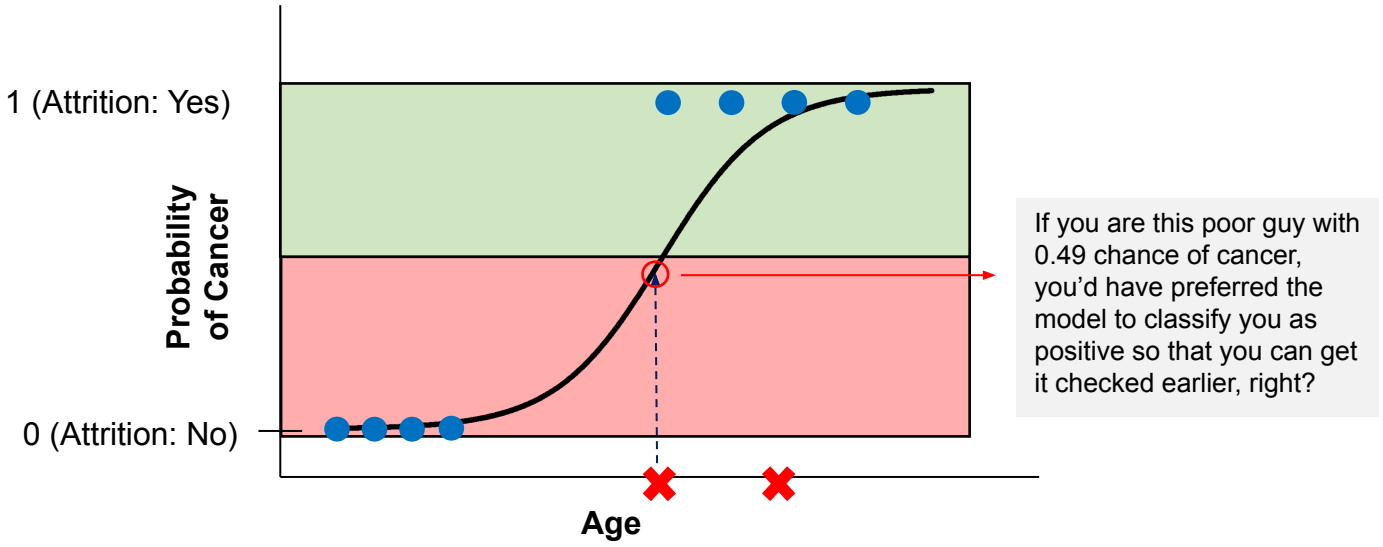
For example, if we use age as a predictor of attrition

STRATEGY FOR SELECTING A THRESHOLD VALUE: WHY IS THERE A NEED FOR THIS?

In some cases, it is alright to settle for a threshold value of 0.5. But in cases where the **cost of misclassification is asymmetrical**, you may want to be more deliberate about your threshold value.

For example, if you are building a classifier to diagnose cancer likelihood, and patients labelled “1” will be sent for further checks:

- Let’s say your **model predicts a probability score of 0.49** (for cancer likelihood)
  If you set the threshold at 0.5, this person would be deemed to be healthy.
- Now if this diagnosis was wrong, the patient would only be diagnosed and treated when the cancer has further advanced.
  The mental toll on the patient, and the financial cost on both the patient and the government would now be much higher.
- In these instances, it makes sense to lower the threshold since **the cost of wrongly diagnosing him as having cancer is lower than the cost of not diagnosing the cancer when he has it**

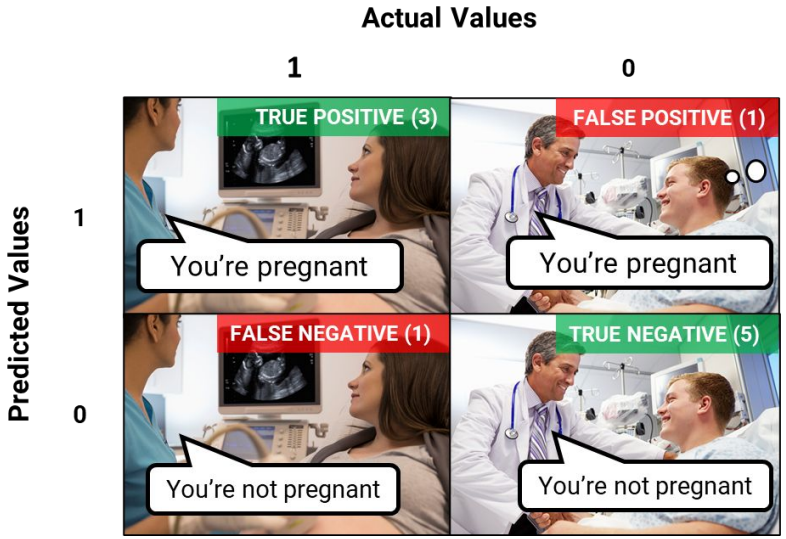


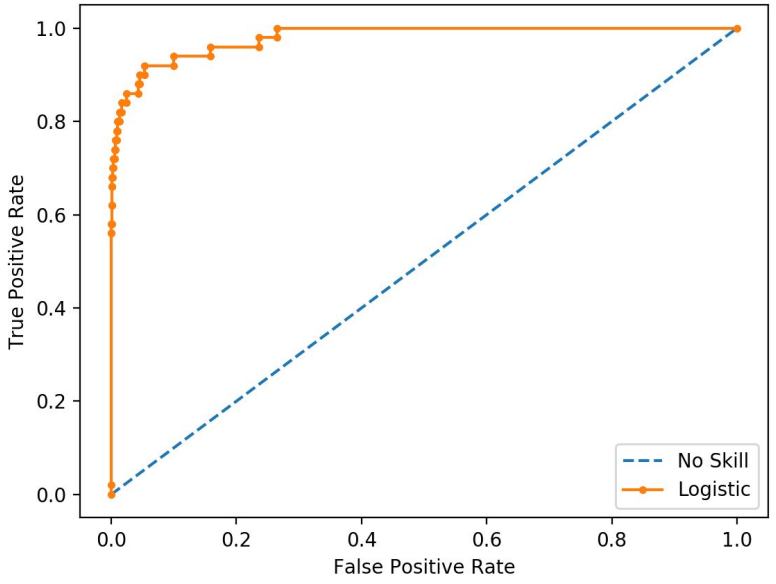
STRATEGY FOR SELECTING A THRESHOLD VALUE: ROC CURVE

- In general cases, we can also go about selecting threshold values in a methodical way, using:
  - Receiving Operating Characteristics (ROC) Curve
  - Area Under Curve

- By now, you should realize that **every time we change the threshold value, so does our predicted outcome, resulting in an entirely different Confusion Matrix**. The issue is, it would be too laborious to draw a Confusion Matrix for all threshold values and then compare each one of them to determine the optimal threshold value

- Luckily for us the ROC curve provides a way to summaries all that information in a single plot.





Notes:

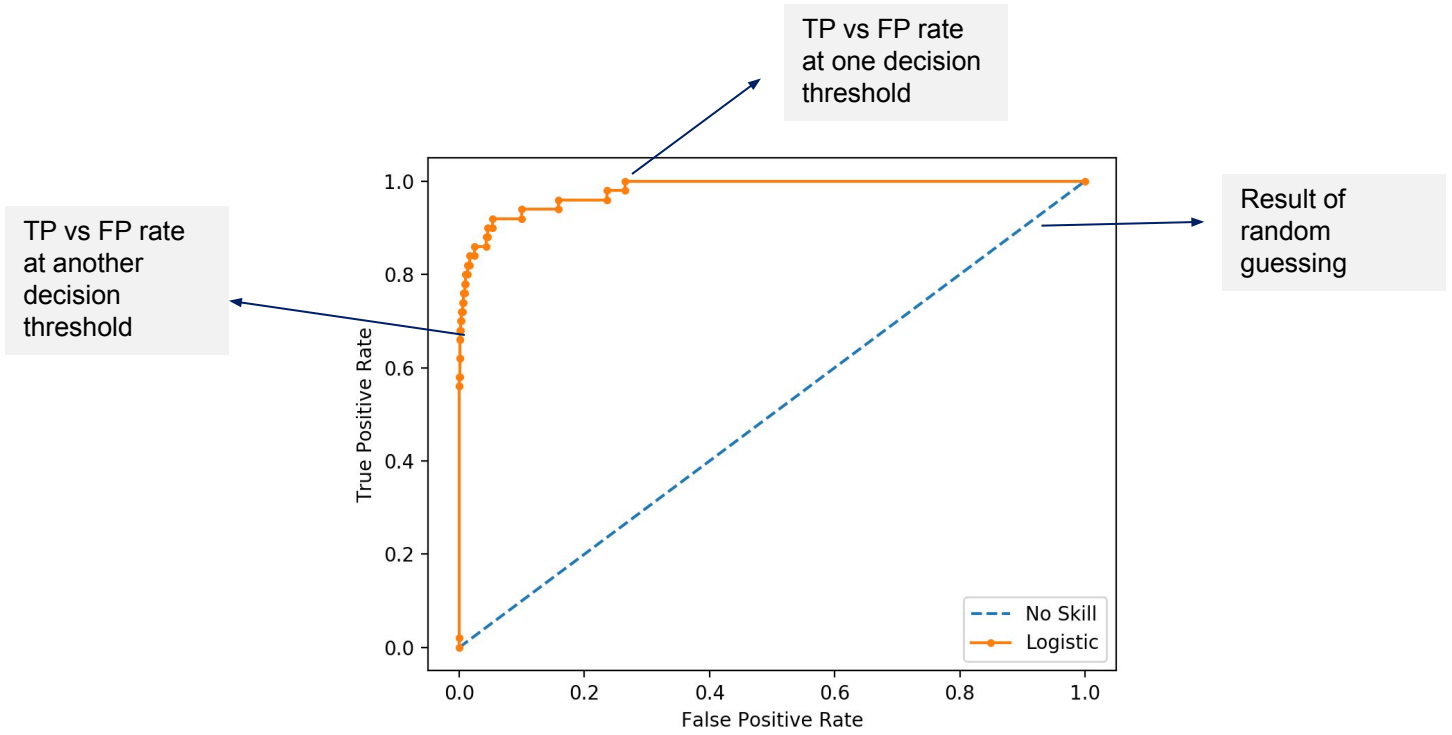
True Positive Rate (TPR) is synonymous with **Sensitivity**.

False Positive Rate (FPR) = 1 - Specificity

STRATEGY FOR SELECTING A THRESHOLD VALUE: ROC CURVE

- In general cases, we can also go about selecting threshold values in a methodical way, using:
  - Receiving Operating Characteristics (ROC) Curve
  - Area Under Curve

- Every point on the ROC curve represents the True Positive Rate & False Positive Rate given a certain decision threshold value



Notes:

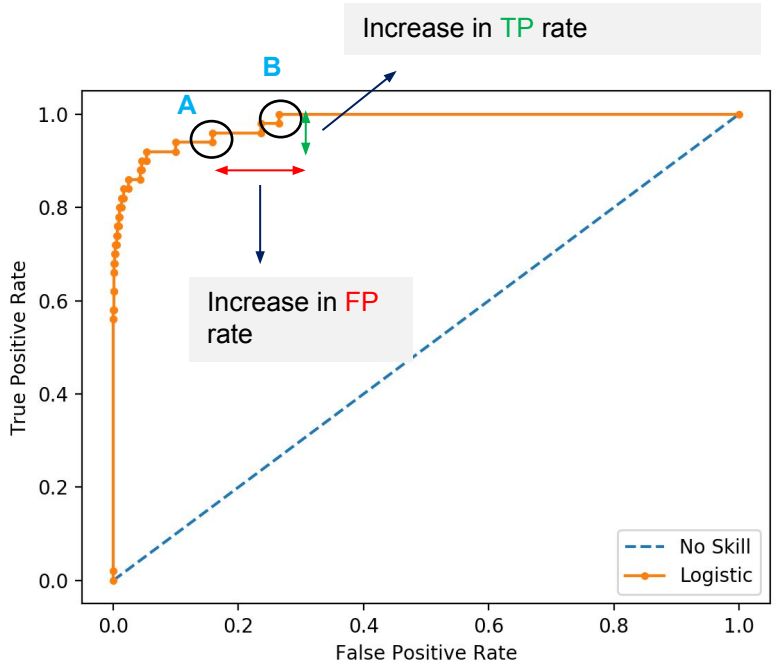
True Positive Rate (TPR) is synonymous with **Sensitivity**.

False Positive Rate (FPR) = 1 - Specificity

STRATEGY FOR SELECTING A THRESHOLD VALUE: ROC CURVE

- In general cases, we can also go about selecting threshold values in a methodical way, using:
  - Receiving Operating Characteristics (ROC) Curve
  - Area Under Curve

- Let's compare two different decision threshold to better understand the mechanism of the ROC curve



As you can tell, point B is not exactly an ideal decision threshold value compared to point A since the gain in TP rate is less than the tradeoff (increase in FP)

Notes:

True Positive Rate (TPR) is synonymous with **Sensitivity**.

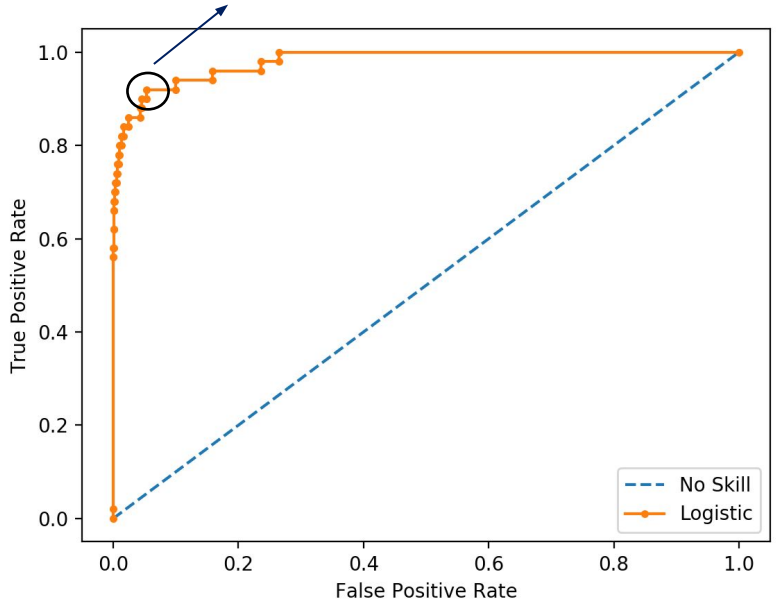
False Positive Rate (FPR) = 1 - Specificity

STRATEGY FOR SELECTING A THRESHOLD VALUE: ROC CURVE

- In general cases, we can also go about selecting threshold values in a methodical way, using:
  - Receiving Operating Characteristics (ROC) Curve
  - Area Under Curve

- A good decision threshold is one where the tradeoff between TPR (Sensitivity) and FPR (1-Specificity) are optimised

Best decision threshold given that the tradeoff between TP and FP is most optimized



Don't worry about having to visually find the decision threshold in future. During the hands-on session, we'll show you how to programmatically derive at the threshold value.

Notes:

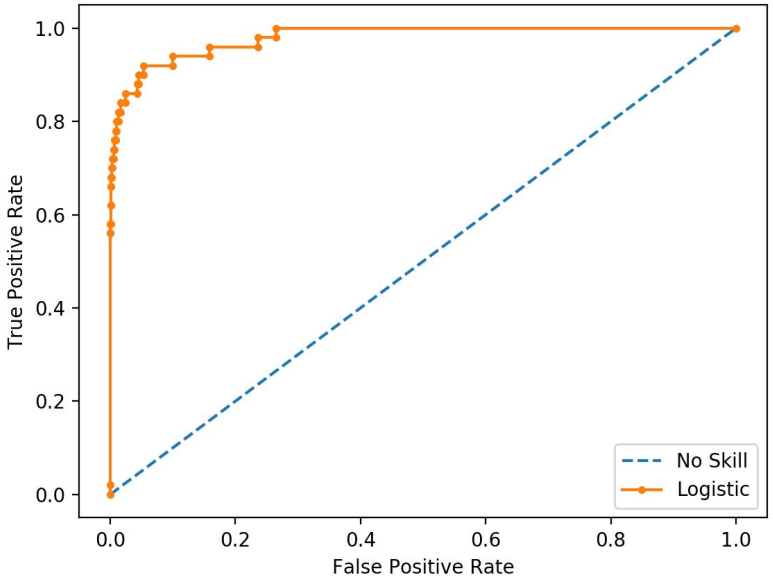
True Positive Rate (TPR) is synonymous with **Sensitivity**.

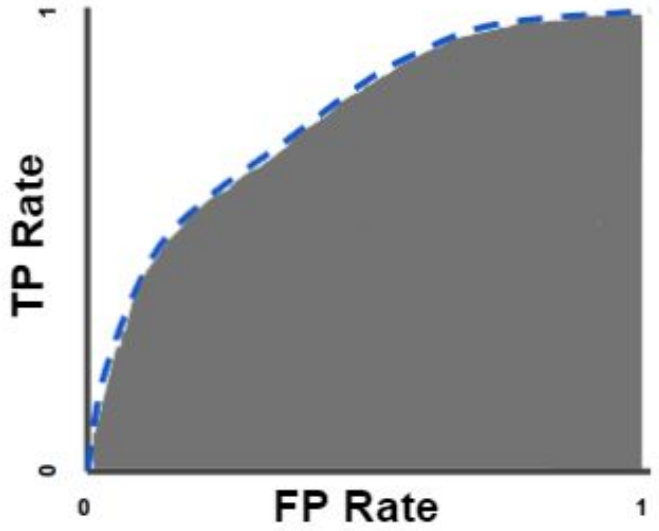
False Positive Rate (FPR) = 1 - Specificity

STRATEGY FOR SELECTING A THRESHOLD VALUE: AUC CURVE

- In general cases, we can also go about selecting threshold values in a methodical way, using:
  - Receiving Operating Characteristics (ROC) Curve
  - Area Under Curve

- AUC provides an aggregate measure of performance across all possible classification thresholds. AUC ranges in value from 0 to 1. A model whose predictions are 100% wrong has an AUC of 0.0; one whose predictions are 100% correct has an AUC of 1.0





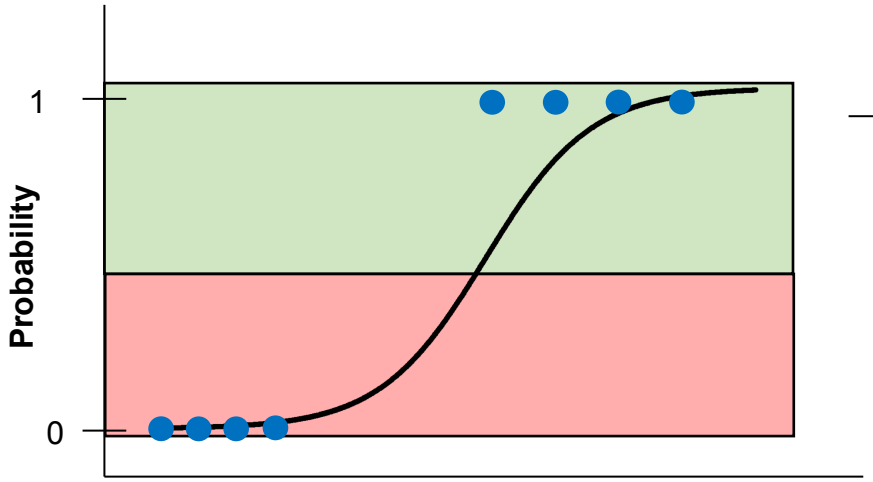
LOGISTIC REGRESSION

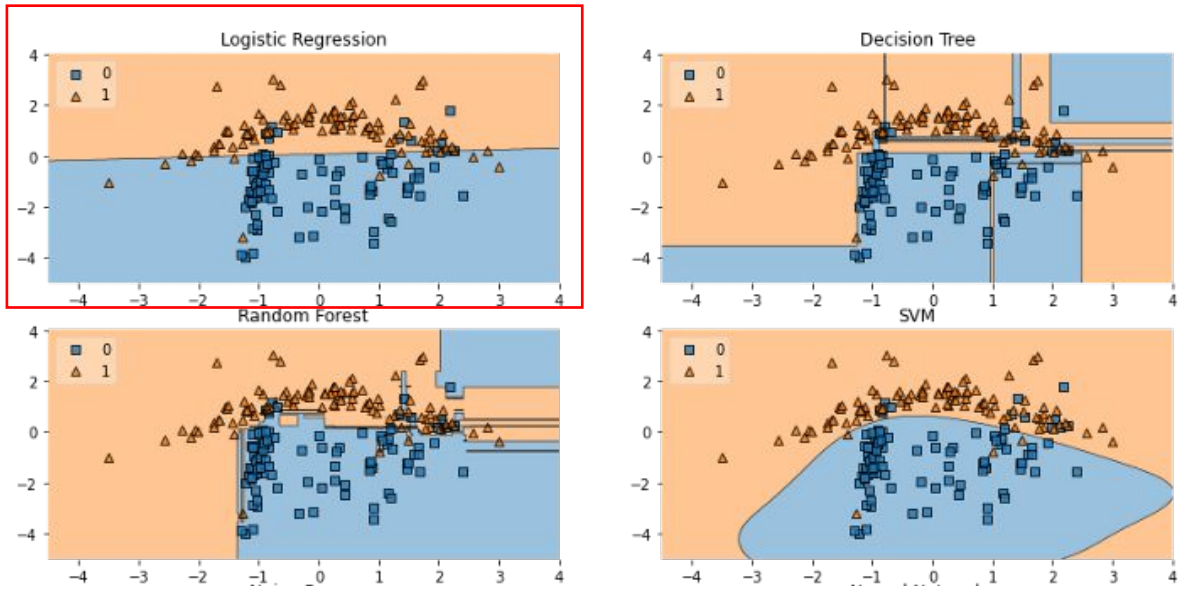
ADDITIONAL NOTES

ADDITIONAL NOTES

- **Feature Selection**: When it comes to feature selection, the tips that we taught in linear regression (using cross-validation) to measure accuracy for each permutation of features applies as well

- **Generalized Linear Model**: Logistics Regression is a generalized linear model, and by the virtue of that, regardless of what decision threshold you select, the classification would always be based on a linear boundary.





Visualization of decision boundaries of various classification models