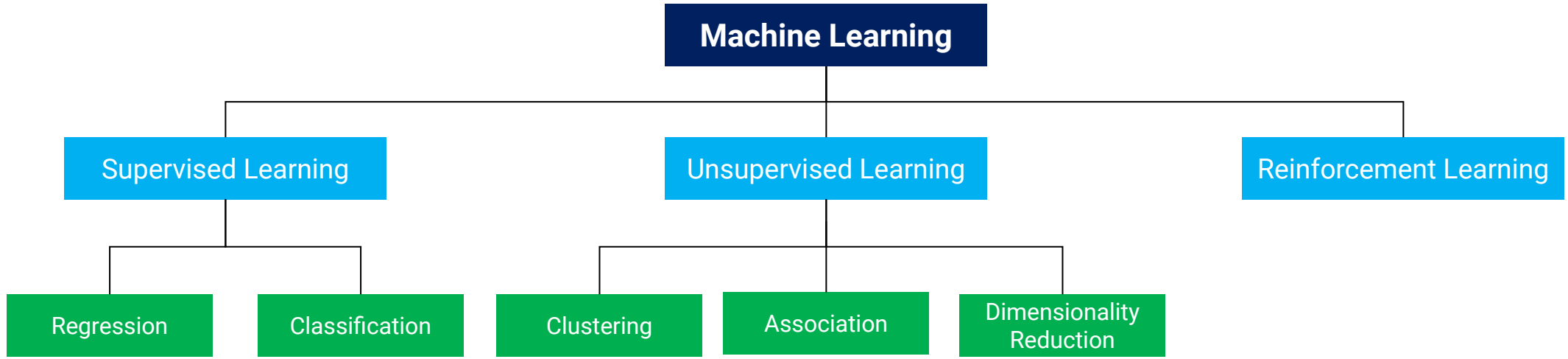
AI200: APPLIED MACHINE LEARNING

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			

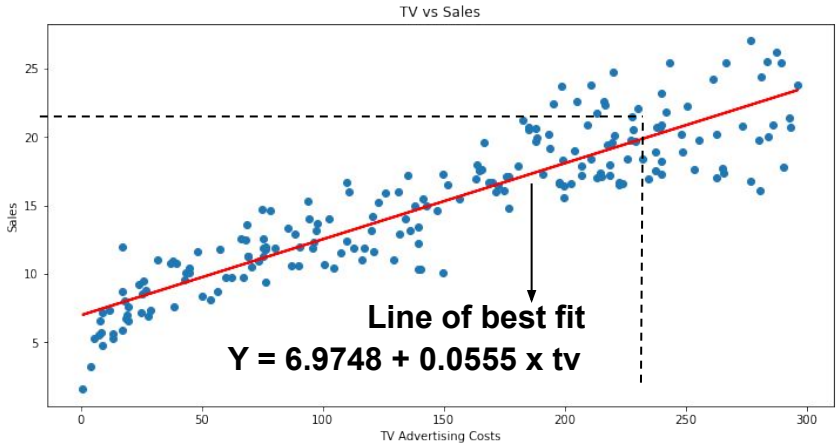
4. Types of applications

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)

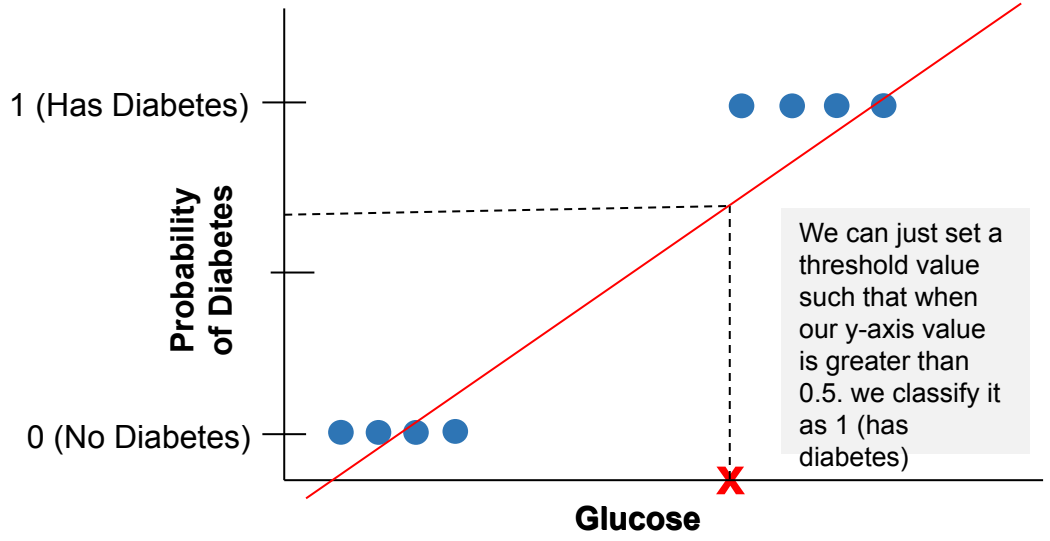


If a new data point with **TV Ad cost of $250** appears, then using the line of best fit the model predict that the Sales is likely **21**

WHY DO WE NEED LOGISTIC REGRESSION?

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

Linear Regression (Classification Problem)

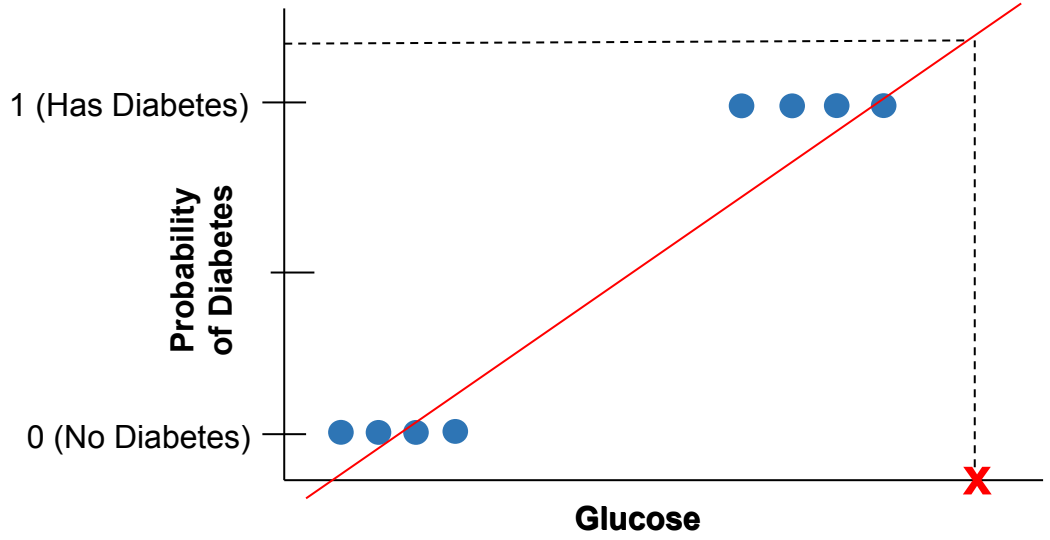


The model seems to work well enough for now. Let's test it a bit more

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 can **give outputs of beyond the range of 0 and 1**, and it makes given a suitable threshold value difficult
  - But by extending the idea of linear regression, we can make it work

Linear Regression (Classification Problem)



WHAT IS LOGISTIC REGRESSION: LAYMAN INTUITION

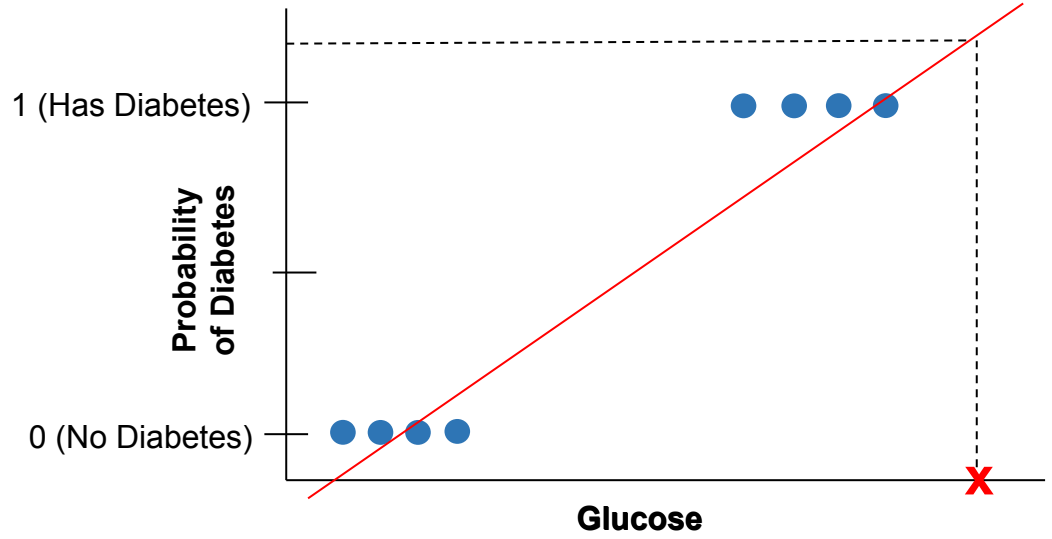
- The concept of logistic regression is similar. However, instead of fitting the data on a straight line, we fit the data on a s-curve (known as sigmoid function) that is constrained between the values of 0 to 1 on the Y-axis.

Fit / Training a model

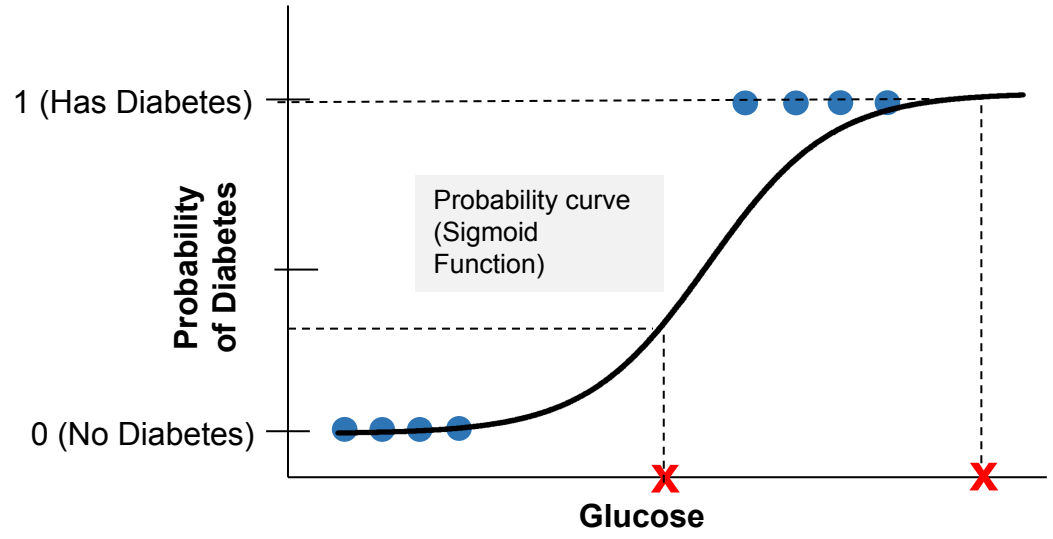
- Let's say in future a patient with a glucose level of XX comes to the hospital, wants to get checked for diabetes. We can provide his data to the model, to predict whether he has diabetes or not

Predicting an outcome with the trained model

Linear Regression (Classification Problem)



Logistic Regression (Classification Problem)



The logistic regression returns a values between 0 to 1 (which can be seen as a probability). But in classification you need a clear outcome of 0 or 1. To achieve this we can have an arbitrary threshold number. When the probability of diabetes is lesser than the threshold of 0.5, we predict the patient will not have diabetes. And notice that this logistic regression is also able to handle the data point which exceeds 1 on its linear regression counterpart. (**more on this in slide 15**)

WHAT IS LOGISTIC REGRESSION: LAYMAN INTUITION

- Summary of idea behind logistic regression
  - We **use a logistic function** to fit the data. The property of the logistic function is such that it produces a s-curve(known as sigmoid function) **constrained between the values of 0 to 1**.
  - This property is ideal given that we can use it to represent the probability distribution (whose value is also 0 to 1)
  - Let's say in future a patient with a glucose level of XX comes to the hospital, wants to get checked for diabetes. We can provide his data to the model, to predict whether he has diabetes or not

Fit / Training a model

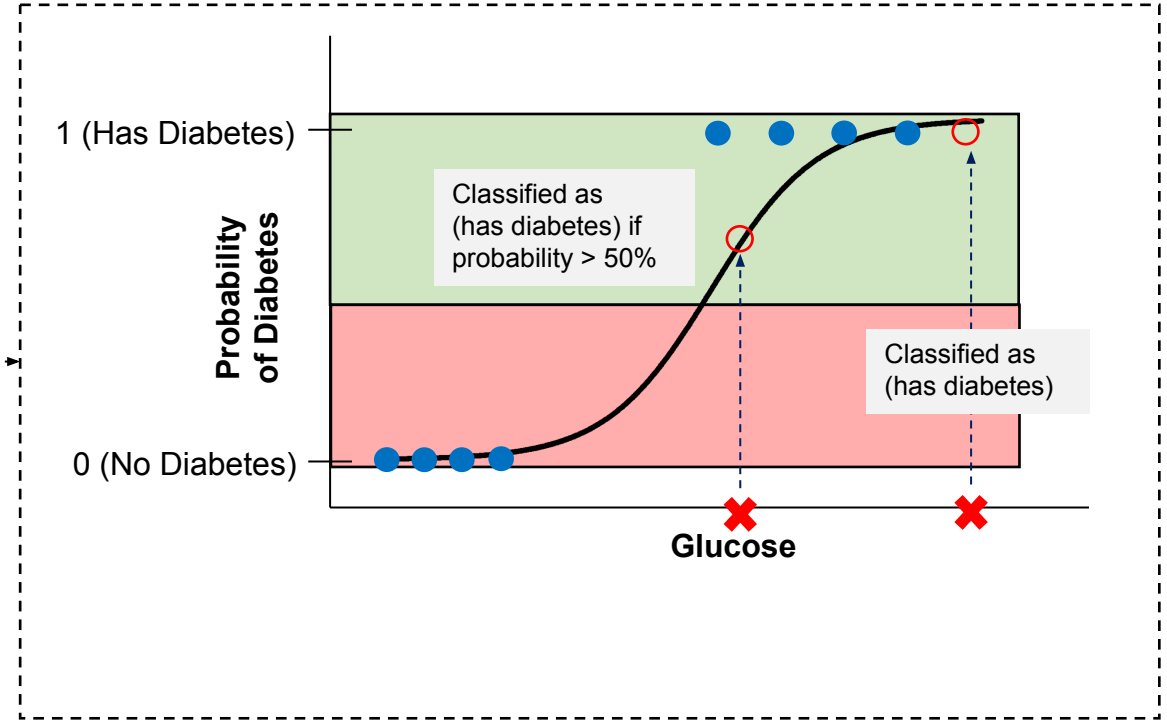
Predicting an outcome with the trained model

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

- Summary of idea behind logistic regression
  - We **use a logistic function** to fit the data. The property of the logistic function is such that it produces a s-curve(known as sigmoid function) **constrained between the values of 0 to 1**.
  - This property is ideal given that we can use it to represent the probability distribution (whose value is also 0 to 1)
  - Let's say in future a patient with a glucose level of XX comes to the hospital, wants to get checked for diabetes. We can provide his data to the model, to predict whether he has diabetes or not

Fit / Training a model

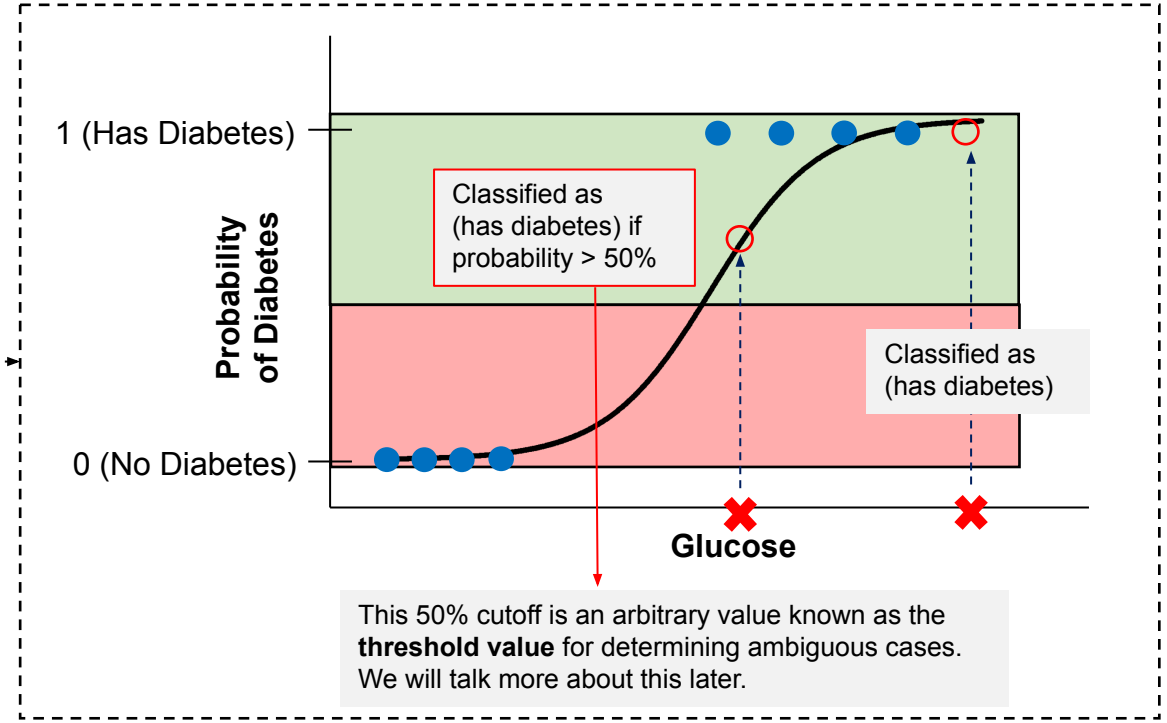
Predicting an outcome with the trained model

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

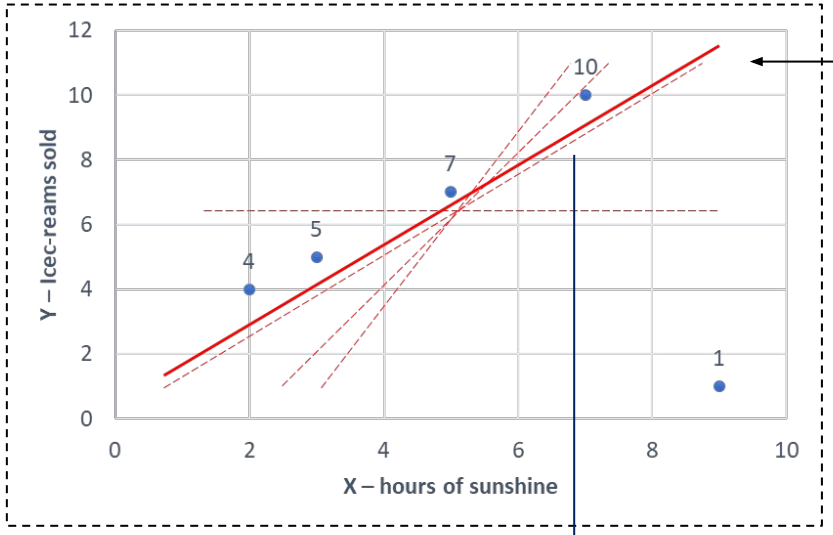
MECHANISM BEHIND MODEL

MECHANISM BEHIND MODEL: MAXIMUM LIKELIHOOD ESTIMATION

- 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

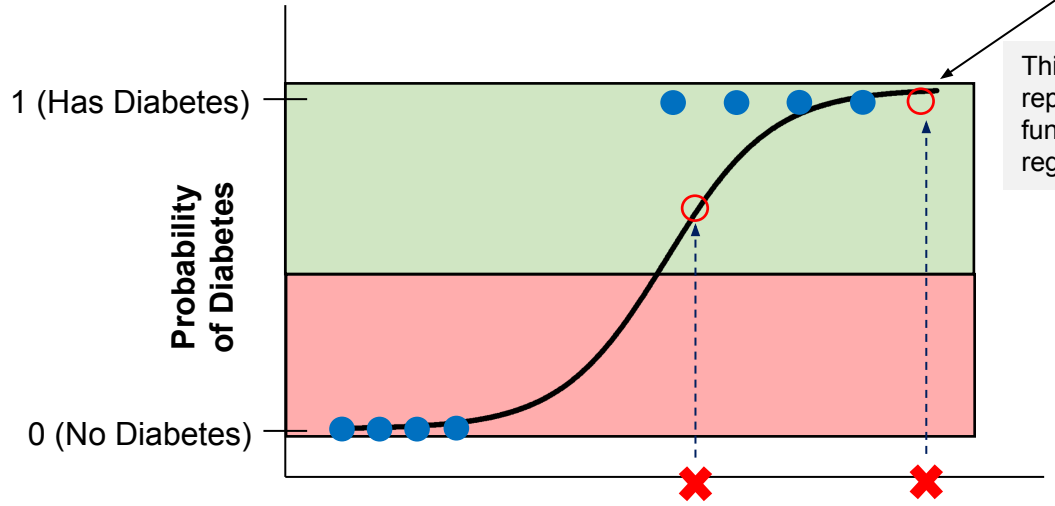
MECHANISM BEHIND MODEL: MAXIMUM LIKELIHOOD ESTIMATION

- The mechanism for logistic regression is similar. However, in this case we use a method called maximum likelihood estimation.

Logistic Regression (Sigmoid Function)

$$\log\text{Odds}(Y = 1) = \beta_0 + \beta_1 X_1 \quad \textbf{or} \quad e^{\beta_0+\beta_1(x+1)} = e^{\beta_0+\beta_1 x}$$

We prefer using this equation as it is much easier to interpret than the first equation



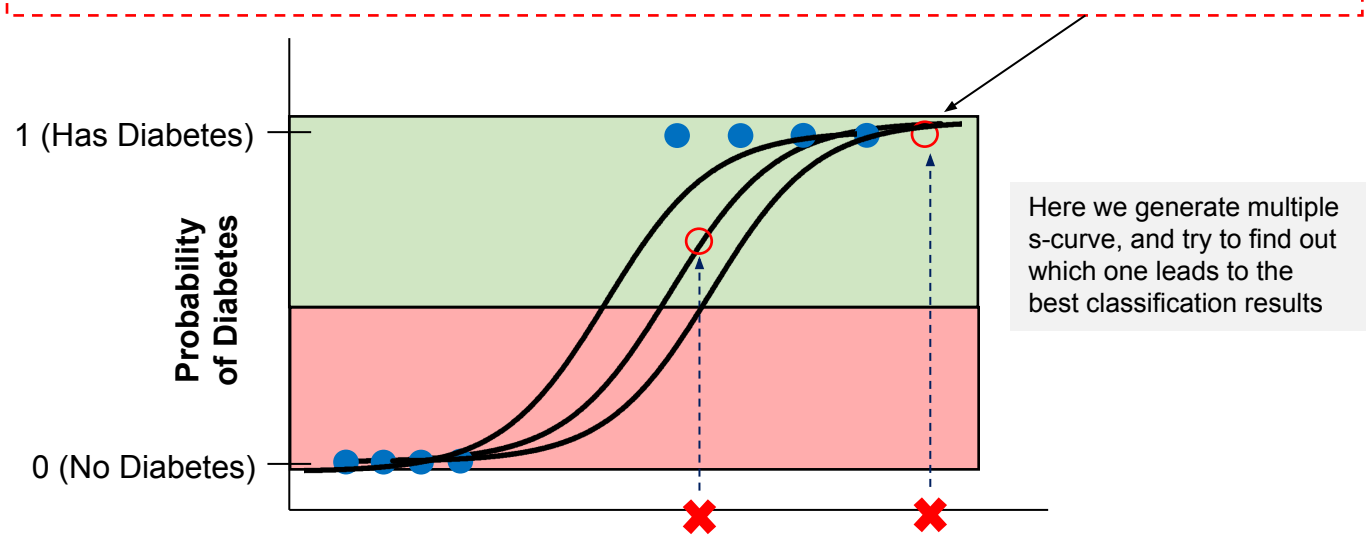
This is the formula that represents this Sigmoid function for Logistic regression

MECHANISM BEHIND MODEL: MAXIMUM LIKELIHOOD ESTIMATION

- Maximum likelihood estimation:
  - **Step 1**: Sub in random values into **β** of the general formula to generate a random sigmoid function (s-curve)
  - **Step 2**: Aggregate the likelihood (how % of outcomes was correctly classified)
  - **Step 3**: We will iteratively repeat step 1-2 to find the optimal value for **β** which has maximum probability of likelihood. We stop this process once we have a **β** which produces the lowest aggregate loss.
  - **Step 4**: Viola, with the optimal **β**, we can now generate a sigmoid function for classification!

Logistic Regression (Sigmoid Function)

$$\text{logOdds}(Y=1) = \beta_0 + \beta_1 X_1 \quad \textbf{or} \quad e^{\beta_0+\beta_1(x+1)} = e^{\beta_0+\beta_1 x}$$



LINEAR REGRESSION

MULTIPLE LOGISTIC REGRESSION

MULTIPLE LINEAR REGRESSION

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

- Essentially you are only adding additional features to the equation (we rearranged 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***

- But don't worry so much about the formula (our focus is on the intuition and not the math of the model). You will see in the in-class hands-on session later that generating multiple linear regression is as easy as just adding a few additional variables to your code

- Even though we cannot visualize anything beyond 2-dimensions, adding new features does not in anyway change the process which maximum likelihood estimation is performed to obtain the optimal parameters.

LOGISTIC REGRESSION

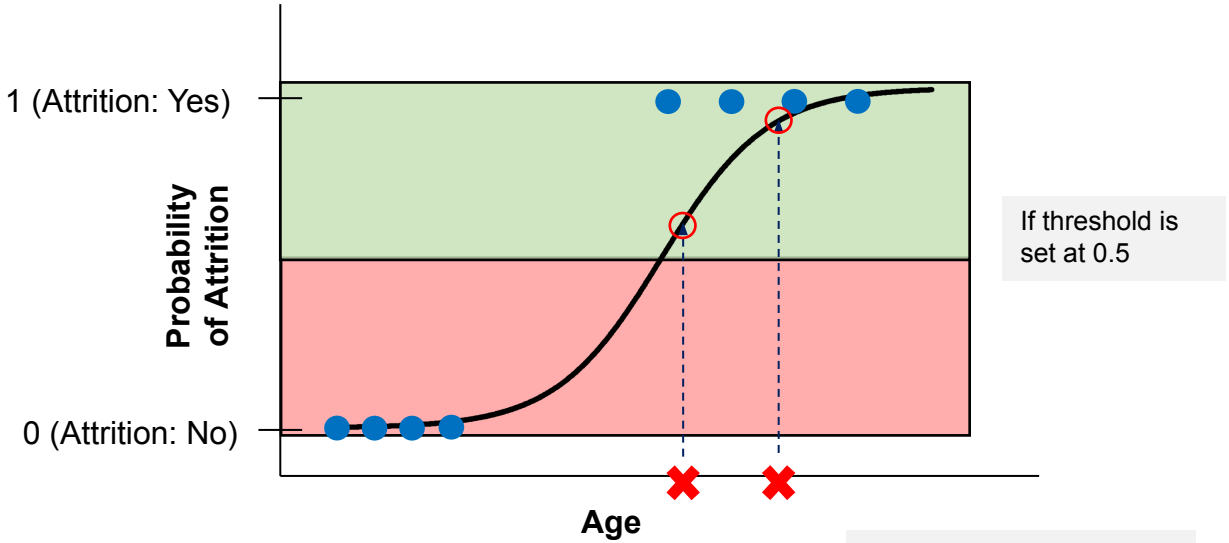
STRATEGY FOR SELECTING A THRESHOLD VALUE

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

- When you use logistic regression to generate predictions;
  - We will not straightaway obtain the predicted outcome in the form of 1s and 0s
  - What you would get instead, is a probability that the outcome is positive (in other words, has a value of 1)
  - Hence, we need to determine a threshold value 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

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



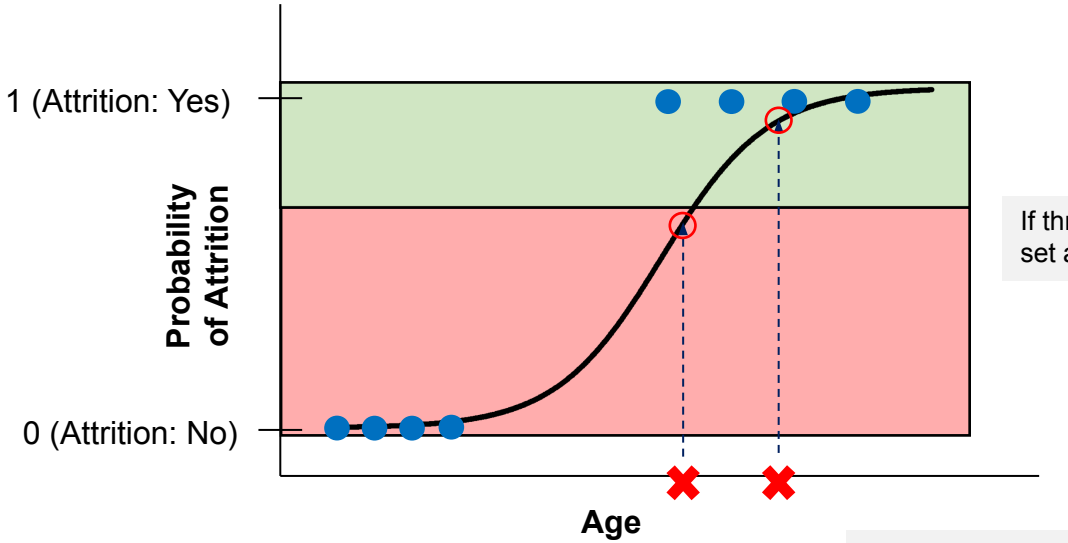
If threshold is set at 0.5

For example, if we tried using age to predict attrition

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

- When you use logistic regression to generate predictions;
  - We will not straightaway obtain the predicted outcome in the form of 1s and 0s
  - What you would get instead, is a probability that the outcome is positive (in other words, has a value of 1)
  - Hence, we need to determine a threshold value 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 threshold is set at 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

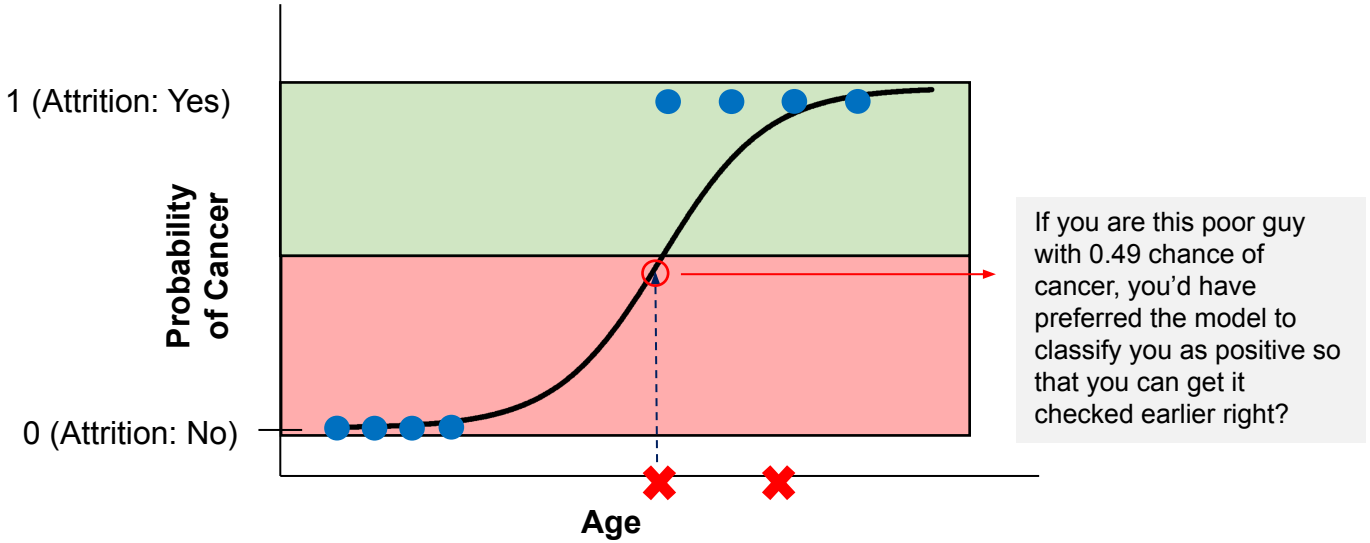
Age	Marital Status	Monthly Income	...	Job Satisfaction	...	Years at Company	Attrition
33	Single	4400	...	4	...	5	1
37	Married	3300	...	4	...	2	0

For example, if we tried using age to predict attrition

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

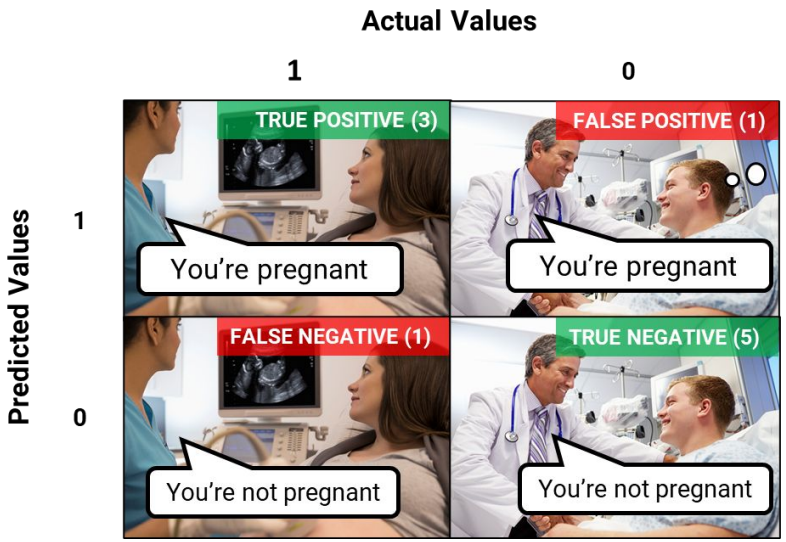
In some cases, it is alright to settle for a threshold value of 0.5. But in cases where the **cost of misclassification is asymmetrical**, then you might want to be more deliberate about your threshold value. For example, if you are building a classifier to diagnose cancer:

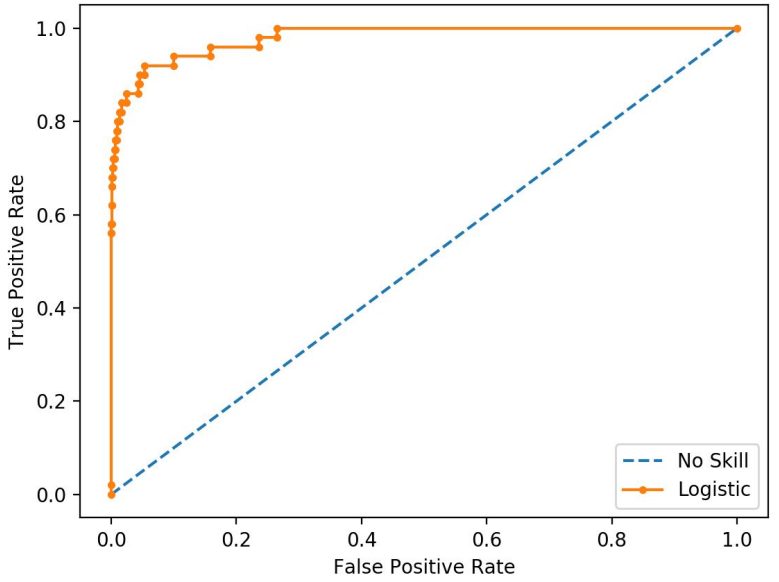
- If someone was labelled as a 1, you would send him for further checks.
- Let's say the model predicts this person has a 0.49 chance of getting cancer. if you set the threshold at 0.5, this person would be deemed to be healthy.
- But if he was wrongly diagnosed, he would be diagnosed and treated when the cancer has further advanced. The mental toll on the patient, and the financial cost on both the patient and the government is higher than if we were to treat and diagnose him at an earlier stage
- In these kind of instances, it makes sense to make the threshold value much lower, because **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

- So how do we go about selecting threshold value in a methodical way?
  - Receiving Operating Characteristics (ROC) Curve
  - Area Under Curve
  - Profit Matrix
- 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

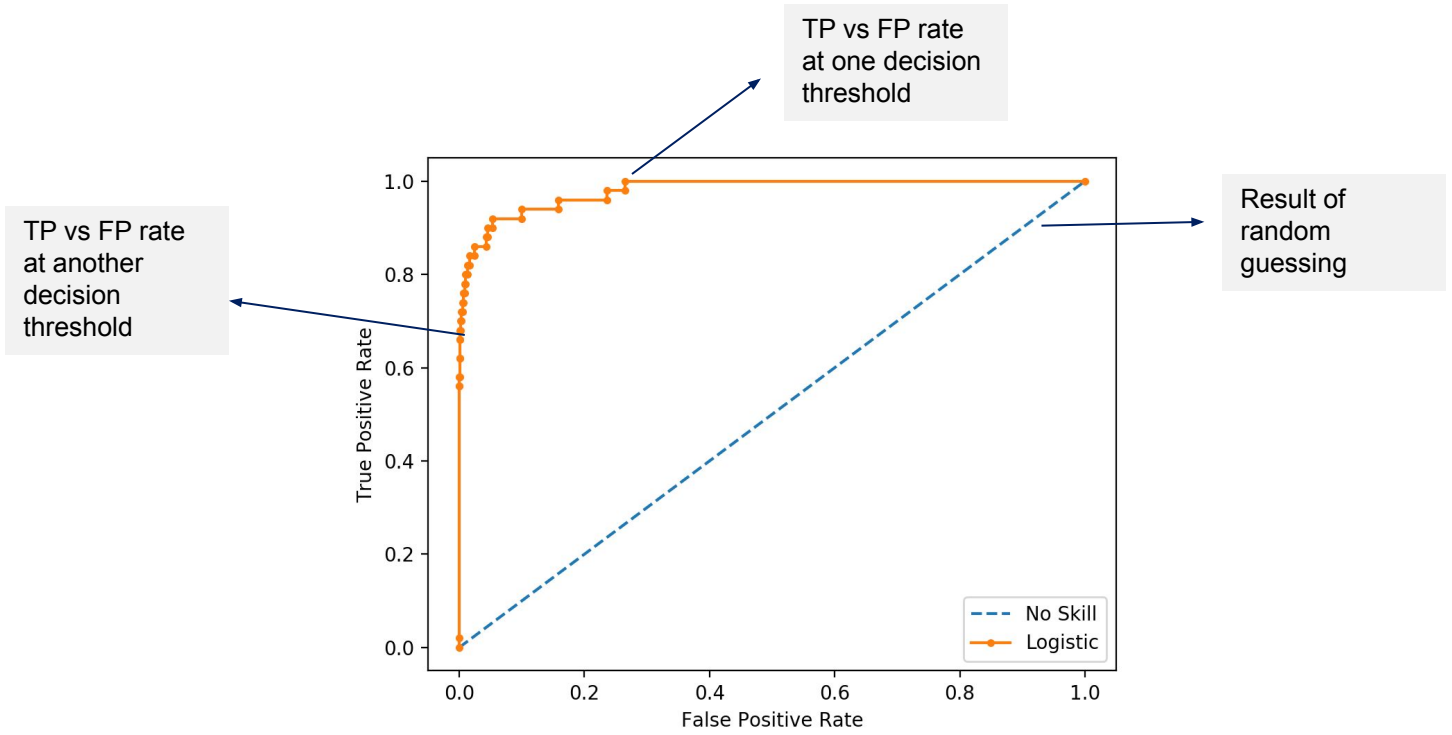




STRATEGY FOR SELECTING A THRESHOLD VALUE: ROC CURVE

- So how do we go about selecting threshold value in a methodical way?
  - Receiving Operating Characteristics (ROC) Curve
  - Area Under Curve
  - Profit Matrix

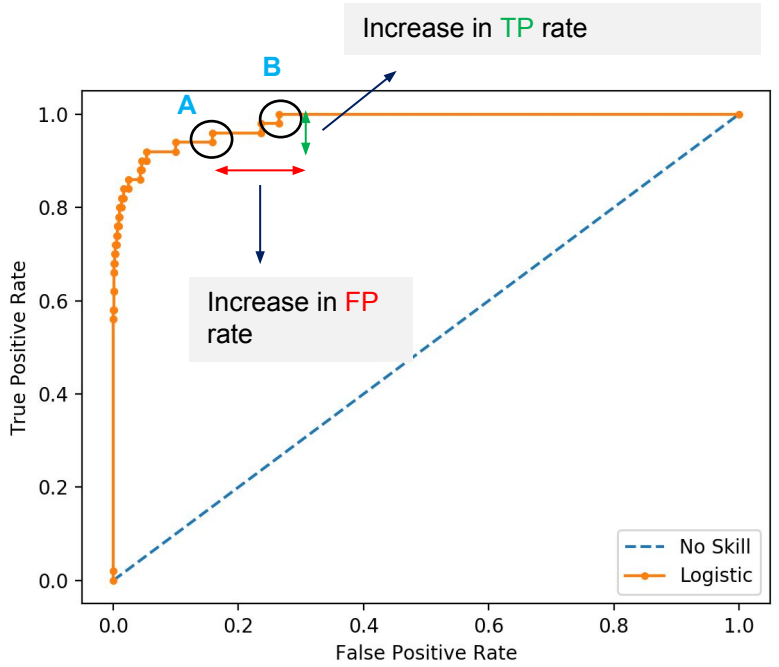
- Every point on the ROC curve represents the True Positive & False Positive rate given a certain decision threshold value



STRATEGY FOR SELECTING A THRESHOLD VALUE: ROC CURVE

- So how do we go about selecting threshold value in a methodical way?
  - Receiving Operating Characteristics (ROC) Curve
  - Area Under Curve
  - Profit Matrix

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



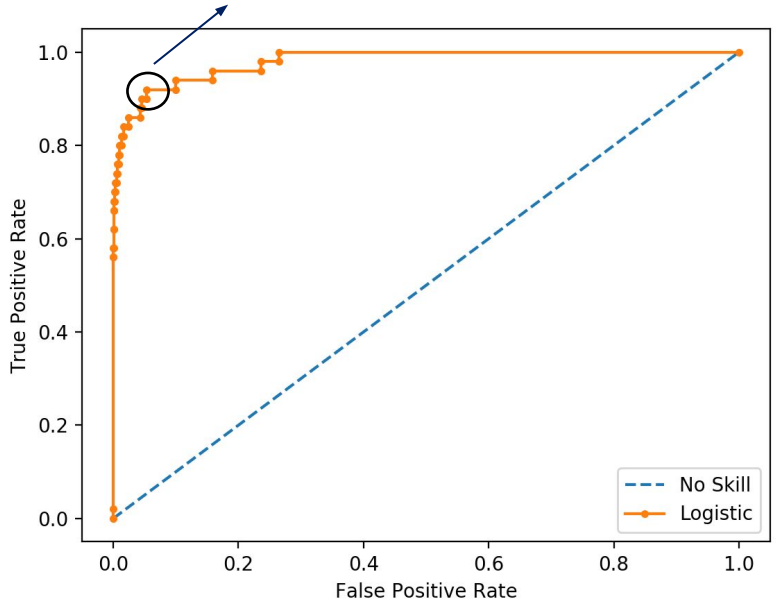
As you can probably tell, point B is not exactly an ideal decision threshold value compared to point A since the gain in TP rate is lesser than the tradeoff (increase in FP)

STRATEGY FOR SELECTING A THRESHOLD VALUE: ROC CURVE

- So how do we go about selecting threshold value in a methodical way?
  - Receiving Operating Characteristics (ROC) Curve
  - Area Under Curve
  - Profit Matrix

- Essentially, a good decision threshold is one where the tradeoff between gain in TP rate and increase in FP rate are optimised

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

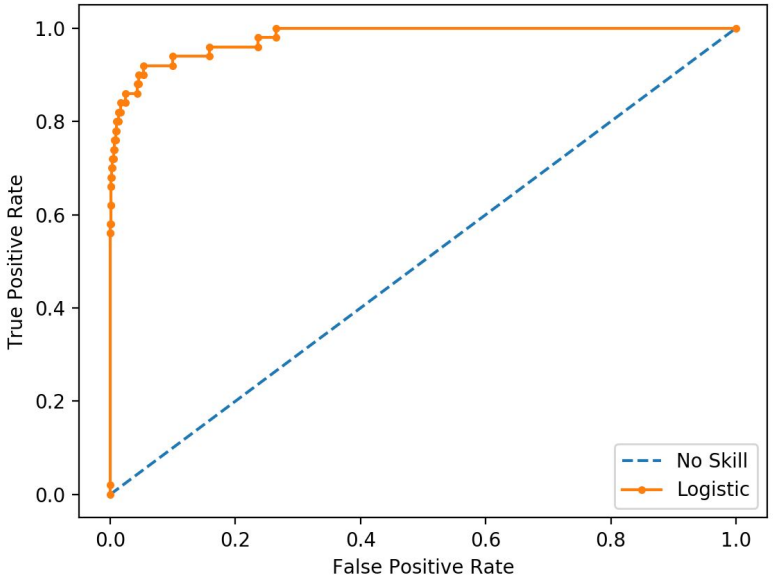


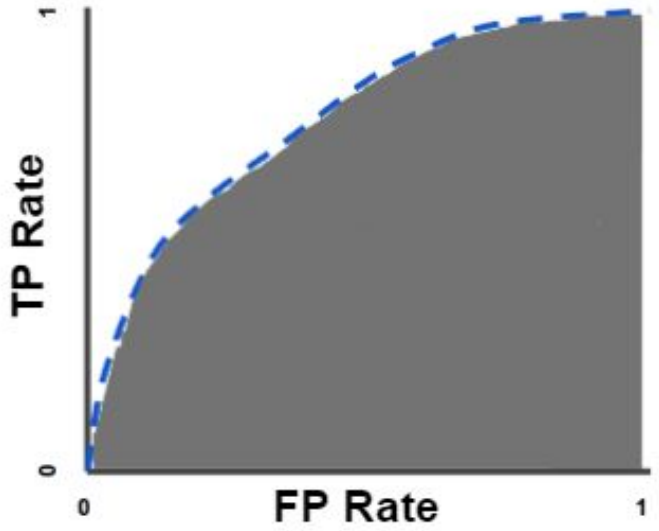
Don't worry about having to visually find the decision threshold in future. During the hands-on session, we will sure you how to programmatically derive at the threshold value

STRATEGY FOR SELECTING A THRESHOLD VALUE: AUC CURVE

- So how do we go about selecting threshold value in a methodical way?
  - Receiving Operating Characteristics (ROC) Curve
  - Area Under Curve
  - Profit Matrix

- 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





STRATEGY FOR SELECTING A THRESHOLD VALUE: PROFIT MATRIX (SELF-READING)

- So how do we go about selecting threshold value in a methodical way?
  - Receiving Operating Characteristics (ROC) Curve
  - Area Under Curve
  - Profit Matrix

- In addition to accuracy, the performance of a classification model can also be **measured by expected profit**. The profit is measured in a concrete unit defined by the final goal of the classification. Let's see an example on this:
  - Here, we assess the accuracy and expected profit of a classification model that predicts the creditworthiness of credit applicants.
  - In a credit scoring application, predicting individual customer behavior has a consequence in terms of profit (or loss). Refusing good credit can cause loss of profit margins (commercial risk). Approving credit for high-risk applicants can lead to bad debts (credit risk).
  - To evaluate misclassification in terms of expected profit, a profit matrix is requested for assigning cost to undesirable outcomes.
  - We introduce a **negative cost (-1) to the False Negatives** - risky applicants who are approved a credit - and a **positive profit (0.35)** to the True Negatives - creditworthy applicants who are approved a credit. The **profit matrix** in the table below shows the cost and profit values for these classification results.

Cost/Profit in each Scenario

	Predicted class **POSITIVE** (Risky)	Predicted class **NEGATIVE** (Credit-worthy)
Actual class **POSITIVE** (Risky)	**Profit for True Positives** 0	**Profit for False Negatives** -1(Cost)
Actual class **NEGATIVE** (Credit-worthy)	**Profit for False Positives** 0	**Profit for True Negatives** 0.35

For example if we use the model to evaluate 10 people and derive at the confusion matrix, we would multiply the cost/profit in each scenario on each instance in the confusion matrix

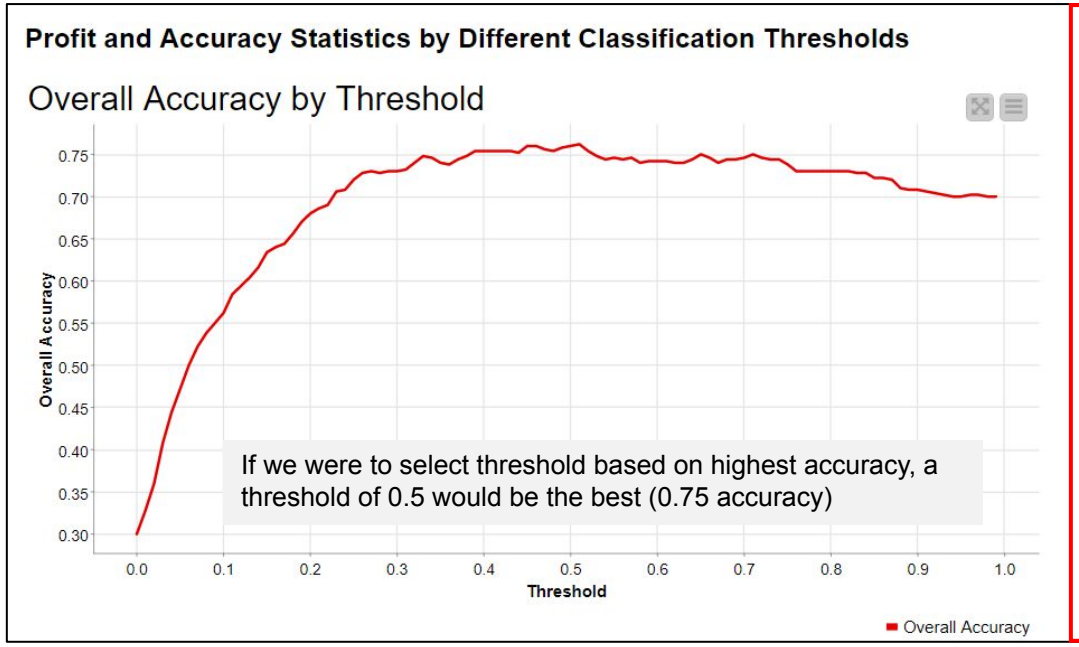
Profit Matrix

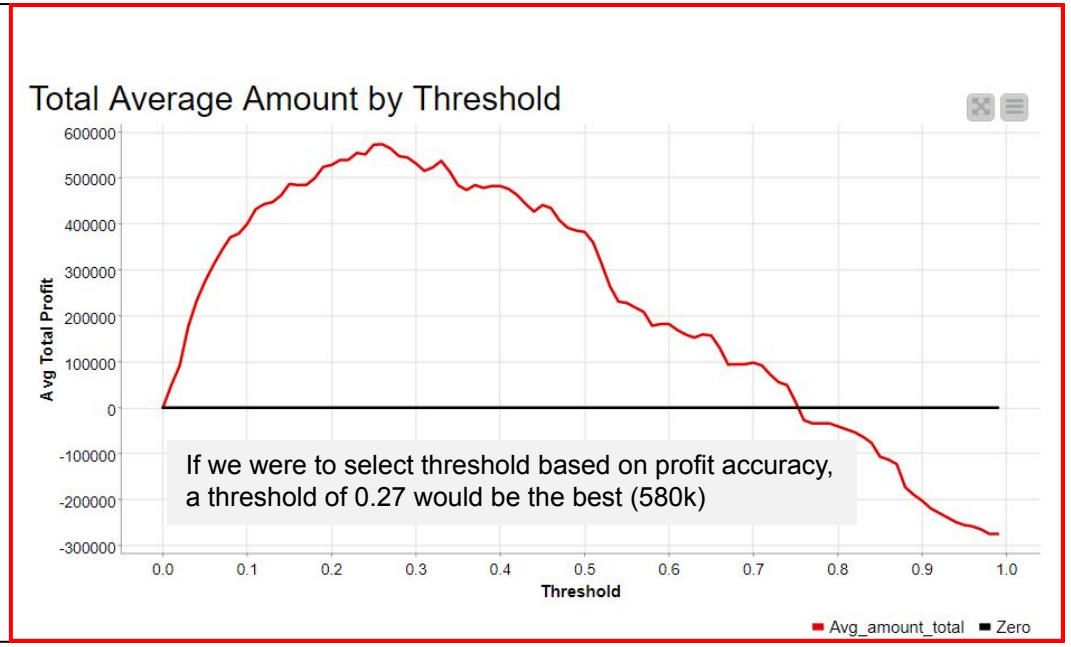
	Predicted class **POSITIVE** (Risky)	Predicted class **NEGATIVE** (Credit-worthy)
Actual class **POSITIVE** (Risky)	0 x 2 applicants = $0	-$1 x 2 applicants = **-$2**
Actual class **NEGATIVE** (Credit-worthy)	0 x 1 applicants = $0	0 x 5 applicants = **$1.75**

STRATEGY FOR SELECTING A THRESHOLD VALUE: (SELF-READING)

- So how do we go about selecting threshold value in a methodical way?
  - Receiving Operating Characteristics (ROC) Curve
  - Area Under Curve
  - Profit Matrix
- Using the profit matrix, we can then plot the profit at each threshold

	Predicted class **POSITIVE** (Risky)	Predicted class **NEGATIVE** (Credit-worthy)
Actual class **POSITIVE** (Risky)	0 x 2 applicants = $0	-$1 x 2 applicants = **-$2**
Actual class **NEGATIVE** (Credit-worthy)	0 x 1 applicants = $0	0 x 5 applicants = **$1.75**





STRATEGY FOR SELECTING A THRESHOLD VALUE: (SELF-READING)

- So how do we go about selecting threshold value in a methodical way?
  - Receiving Operating Characteristics (ROC) Curve
  - Area Under Curve
  - Profit Matrix

- Profit matrix can be applied if you can assign a profit/cost to an outcome (and in reality, you can do that for most things in life)

- However, you must discern **whether model accuracy or the profit is more important** in your case (and in the best-case scenario, the model with best accuracy is also the one with highest profits)

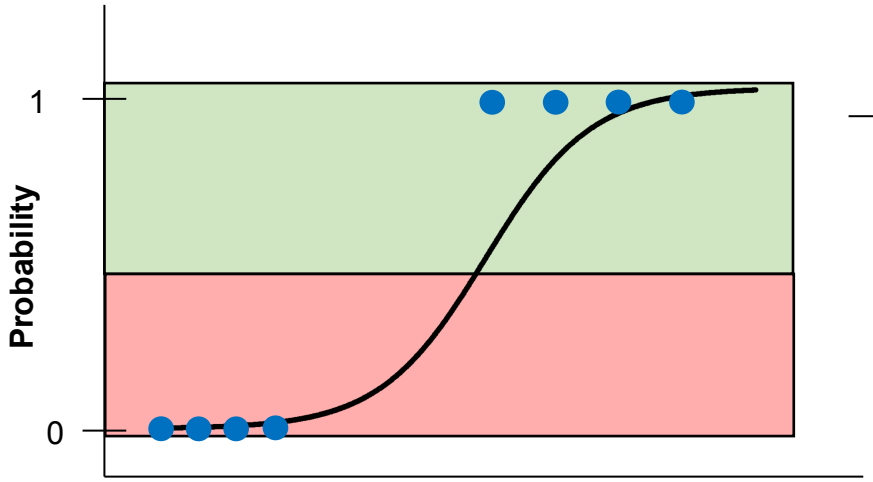
LOGISTIC REGRESSION

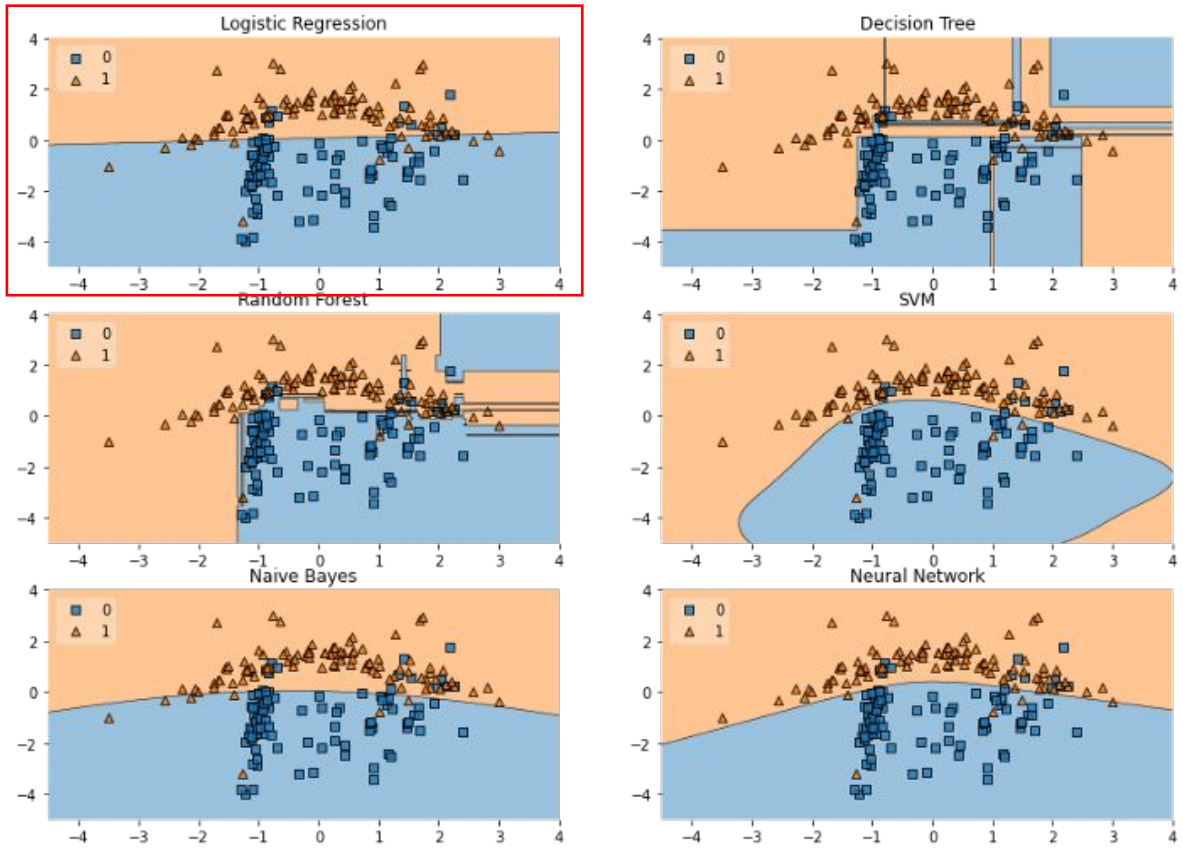
ADDITIONAL CONSIDERATIONS

ADDITIONAL CONSIDERATIONS

- **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