

Youtube - Statquest with Josh Starmer ROC and AUC

① Lojistik Regresyon ve Confusion Matrix Tekrarı

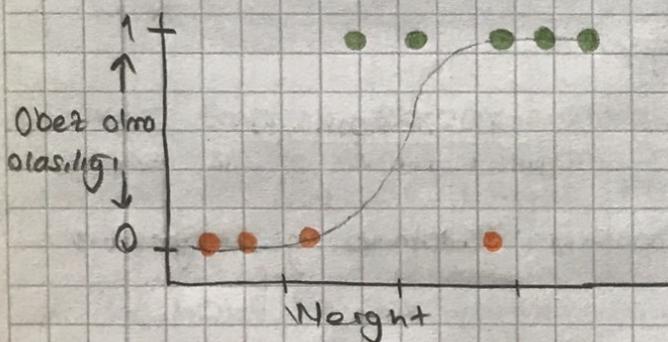
forezlerin obez olup olmadığına karar vereceğimiz bir model taramaya çalışıyoruz.

- * Dataya bakıldığında forezlerin ağırlıklarına göre obezlik durumu şu şekilde:



- * Bir forezinin ağırlığı çok farklı olmasına rağmen yine de obez.

- * Bu dataya lojistik regresyon eğrisini oturtuyoruz. Bunu yaptığımda y ekseni forezin obez olup olmayacağına döndürüyor.



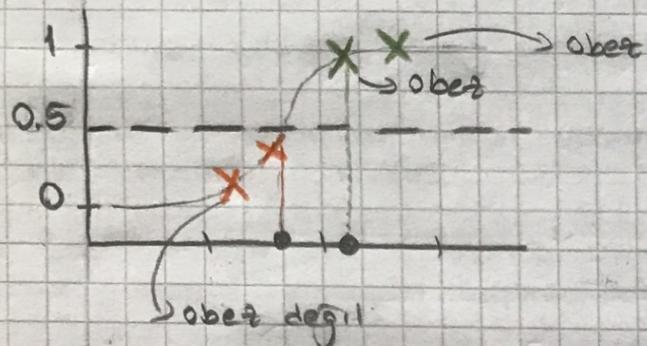
- * Bize forezin ağırlığı verildiğinde onun düşük mi yüksek mi olasılıkları obez olduğu bilgisine ulaşabiliyoruz bu grafikle



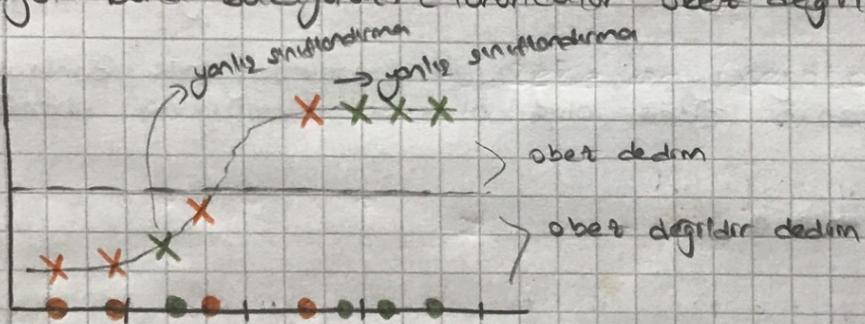
- * Bu ağırlıkta bir forezin obez alma olasılığı çok yüksek iken

- * Bu ağırlıkta bir forezin obez alma olasılığı çok düşük.

Forem: Obes olup olmadığını karar verebilmemiz için obesligi sınıflandırmaya çalıştığımızı söyleyelim. Bu nedenle threshold olarak 0.5'yi vermek. Bu şekilde obes olma olasılığı 0.5'ten büyükse obesdir, küçükse epítise obes değildir demek oluyoruz.



Modelimizin başarısını ölçeceğiz. Daha sonra elde edilen 8 forem bilgisi var. Gerçekte obes olup olmadığını biliyoruz. Bunu rica ettiğimde ne dıyar bunda bakiyoruz. (Turuncular obes değil, yeşiller obes)



Sınıflandırma sonuçlarını confusion matrix üzerinde gösteriyorum.

		Gerçekte	
Tahmin edilen	Obes	Obes değil	
	(TP) 3	(FP) 1	
	Obes değil	(FN) 1	(TN) 3

TP: Pozitif tahmin ettim, doğru çıktı.

FP: Pozitif tahmin ettim, yanlış çıktı.

FN: Negatif tahmin ettim, yanlış çıktı.

TN: Negatif tahmin ettim, doğru çıktı.

TP: 3 → 3'üne obes dedim ve doğru bıldım.

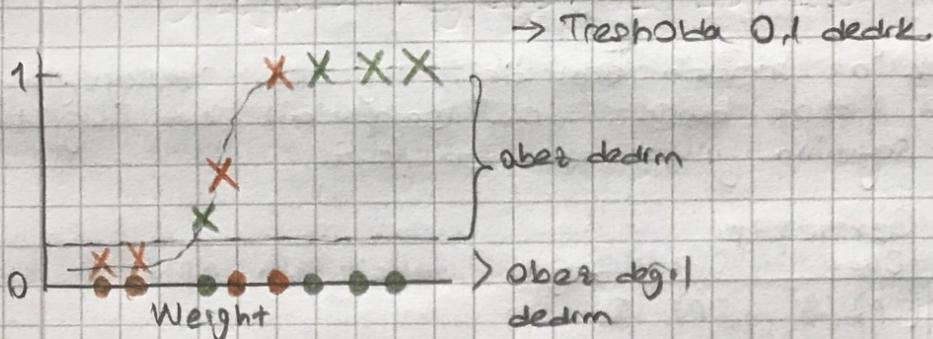
FP: 1 → 1'ine obes dedim ve yanlış bıldım.

FN: 1 → 1'ine obes değil dedim ve yanlış bıldım.

TN: 3 → 3'üne obes değil dedim ve doğru bıldım.

⑨ Thresholdin Sonuçlara Etkisi

- Samplelin obet olup olmadığını koror vermek için kullandığımız thresholdu değiştirdiğimizde ne olduğunu bakiyoruz.

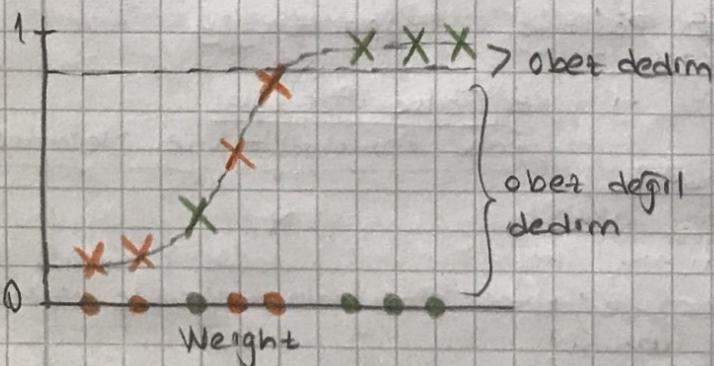


- Bu sefer 3 yerine 4 obet törayı doğru yakaladım (TP).
- Fakat egeri doğruların dağınıca da sample'on obet demek FP sayımı da artırdı. İken 2 oldu.
- Daha az sample'la obet değil dedigim için bu tabloda dahası da hata yapmış oldum. FN 1'den 0'a düştü.

		Gerçekte	
		Obet	Obet değil
Tahmin edilen	Obet	(TP) 4	(FP) 2
	Obet değil	(FN) 0	(TN) 2

Obet ömegini korona olarak düşünebilirim. Bu durumda daha fazla kişiye korona demek solğun reaksiyi ortaya getirebilir. Bu da thresholdu değiştirmeyi montikli kılıyorsa FP'yi (yani korona dedigimde gerçekte olmama sayısını) artırarak olsa bile.

- Thresholda 0,9 dedigimizde ne olduğunu bakiyoruz.



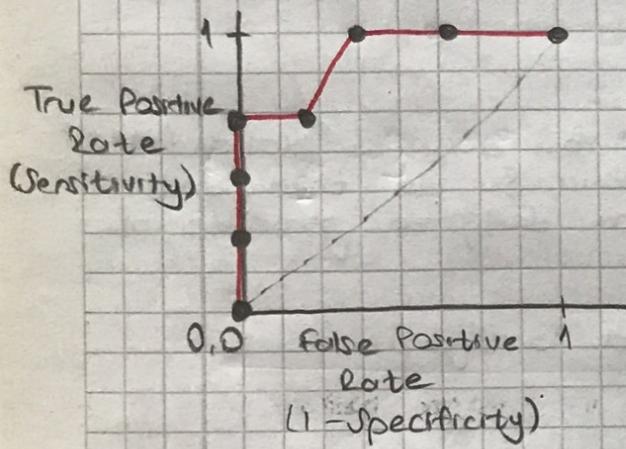
- Threshold 0,1 verdigimizde yani tam samplelerde obet dedigimizde, gerçekte obet olanların hepsi bilimetric fakat burada egeri çok yüksektik ve gerçekte obet olanların sadece bir kısmını biliyoruz. TP değeri egerin 0,5 olduğu zamanlaştırmakla aynı çıktı.
- Zaten çok az sample'la obet dedigimiz için FP sayımız da 0 oldu. Hatta burada 0 oldu.

- Gaguna Oberz degrildir dedigimiz rasi Oberz olmayanları yakalama sayımız (TN) da arttı. Hatta bu, egeri OR belirledigimizdeki sayıdan daha fazla.
- O,1 epigindeki OR aq sampleda Oberz degil durumda doğrulukla ya yanlış pozitif rasi aq. fakat burada OR sampleda Oberz degrildir dedigim rasi FN sayısının yükselmesini beklerim.

		Gerçekte	
		Oberz	Oberz degil
Tahmin	Oberz	(TP) 3	(FP) 0
	Oberz degil	(FN) 1	(TN) 4

③ Receiver Operator Characteristic (ROC)

* 3 farklı threshold'un sonucunu baktığımızda en iyi sonuc (TP+TN) 0,9'un verdiğiğini görüyoruz. Fakat threshold 0 ile 1 arasında herhangi bir değer olabilir. Hangi threshold'un en iyi olduğunu nasıl belirleyeceğiz? Bunun için her bir threshold'ı test edip confusion matrislerine bakınca genelde birincil bir bilgiyi basit bir şekilde özetlemeye yardımcı grafikleri kullanılır.



* True Positive = Sensitivity Rate

$$= \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Gerçekte Oberz olanların yüzde kaçına ben de Oberz dedim?

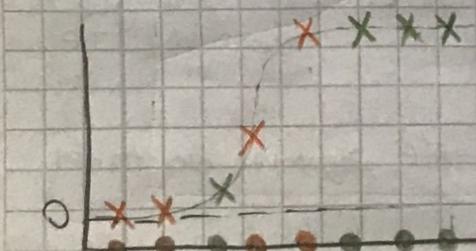
* False Positive = 1-Specificity Rate

$$= \frac{\text{False Positives}}{\text{False Positives} + \text{True Negatives}}$$

N

Gerçekte Oberz olmayanların yüzde kaçını Oberz dedim?

* Dryelim ki tüm sample'lara Oberz dedim epig'i çok düşük tutuy.



Bu durumda tüm obetleri doğru bilmig olurum ama bu durumda obet olmayanların hepsini yakalayamamız olurum.

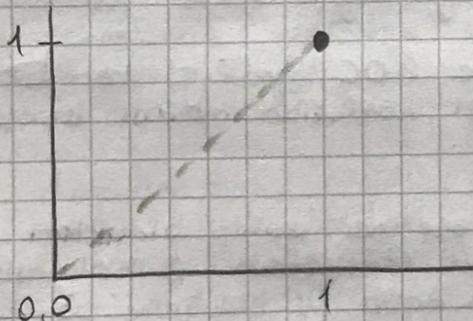
		Gerçekte	
		Obet	Obet degil
Tahmin	Obet	(TP) 4	(FP) 4
	Obet degil	(FN) 0	(TN) 0

True Positive Rate ve False Positive Rate'ı hesaplayınız.

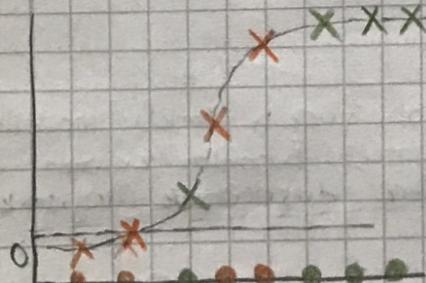
$$\text{True Positive Rate} = \frac{TP}{TP+FN} = \frac{4}{4+0} = 1$$

$$\text{False Positive Rate} = \frac{FP}{FP+TN} = \frac{4}{4+0} = 1$$

Bu durumda grafikteki 1,1 noktası tüm obetleri doğru sınıflandırılmış olsa obet olmayanların hepsini yanlış sınıflandırdığını söyleyeceğiz. Yani True Positive Rate = False Positive Rate'ı gösterir.



* Thresholdu biraz artırıyoruz. Bu şekilde daha azıda obet dedik

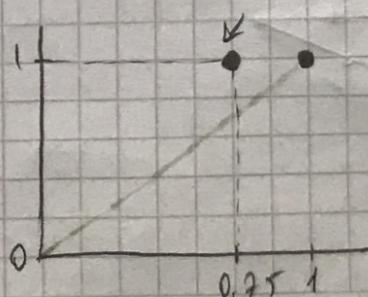


		Gerçekte	
		Obet	Obet degil
Tahmin	Obet	(TP) 4	(FP) 3
	Obet degil	(FN) 0	(TN) 1

Obet olmayanlardan da 1 tanesini yakalayamadık.

$$\text{True Positive Rate} = \frac{TP}{TP+FN} = \frac{4}{4+0} = 1$$

$$\text{False Positive Rate} = \frac{FP}{FP+TN} = \frac{3}{4} = 0.75$$



(0,25,1) noktasında TP rate'ının FP rate'inden daha büyük olduğunu söyleyorum. Yani threshold'unun sıfırda olup olmadığını konusunda daha iyis.

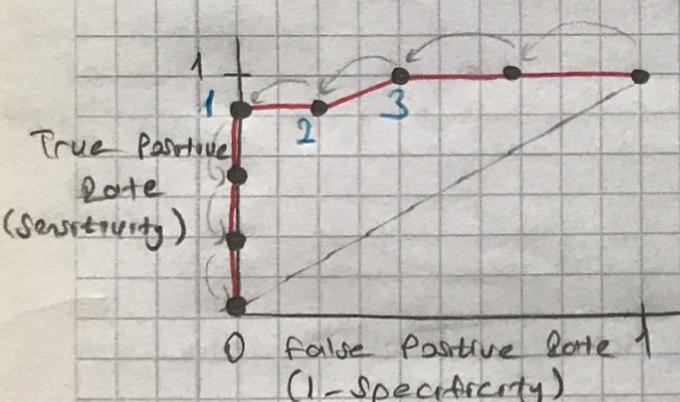
Thresholdu artırıp diğer değerleri deneğimizde FP sayıları azalıyor yani obet olmayanları obet etmemeyiz artık ve obet olmayanları daha çok tespit edebiliyoruz.

Thresholdu artırıksa (0,8 olduğunu düşün) daha da obeti tespit edebiliyoruz ama obet olmayanların sayısını yakalıyorum. Hem TP rate'si düşüyor hem FP rate'si (0,25,0,25) oluyor örneğin.

Thresholdu dahi da artırıksa artık sadece obet değerlerini doğru ve FP 0 olmaz oluyor. En az bir kalan obetlerin bildirilmemesi (0,0,0,0) oluyor yani noktası. Ve obet olmayanların da hepsi de doğru tespit etmiş oluyoruz.

Thresholdu dahi da artırınca artık TP da azalıyor sadece hepsi obet değerlerini deniyor (0,0) noktası, artık herkiler obeti tespit edememişiz (TP=0) ve herkilerin obet demedigim için FP=0 olduğunu gösterir.

Bu noktaları birleştirinden sonra ROC grafiğini elde ederiz.



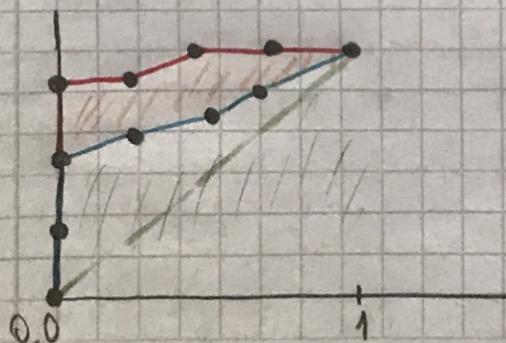
Bu şekilde ROC grafiği her bir threshold'un ürettiği confusion matrisi birebirleştirmeye yardımcı.

Aynı TP rate'ine sahip oldukları için 1 noktasındaki thresholdun 2 noktasındaki thresholdden daha iyis olduğunu söyleyebiliriz.

Ne kadar FP kabul edebildiğimizde bağlı. Olarak optimal thresholdun 1 mi 3 mü olduğunu söyleyebiliriz.

④ Area Under the Curve (AUC)

AUC, grafiğin altında kalan alanını ifade ediyor. Farklı farklı modeller deneğimizde bu işi göre karşılaştırırız yapabiliyoruz. Buradaki kırmızı eğrili lojistik regresyon, mavi eğrili random forest olarak da bilinen lojistik regresyondan farklılığı göstermektedir.



5 Hesaplama Yapanı Degistikliği

✓ Bazen false Positive Rate yerine Precision kullanmak daha mantıklı olabiliyor.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Pozitif tahminlerim

Obez söyle tahminlerinin yüzde kaçı gerçekte obet?

Obet olmayanların ağırlıkta olduğu bir dattada False Positive Rate yerine Precision kullanmak daha mantıklı oluyor. Çünkü Precision True Negative sayısını içermiyor formülde, böyleslikle dataobeli dengeye gitmektedir.

		Gerçekte	
		Obez	Obez değil
Tahmin	Obez	TP	FP
	Obez değil	FN	TN

Pratikte bu nadiren görülen hastalıklar üzerine çalışırken geçerlidir. Sü tora case'lerde hastalığa sahip olmayan çok fazla kişi varken, hasta olan çok az kişi oluyor.