

Statistics II

for Machine Learning

Chap. 5: 로지스틱 회귀분석 & SVD

Oh, Hyung Sool

로지스틱 회귀분석

- 로지스틱 회귀(Logistic Regression)의 목적은 일반적인 회귀 분석의 목표와 동일하게 종속 변수와 독립 변수간의 관계를 구체적인 함수로 나타내어 향후 예측 모델에 사용하는 것
- 로지스틱 회귀는 선형 회귀분석과는 다르게 종속 변수가 범주형 데이터를 대상으로 하며, 데이터의 결과가 특정 분류로 나뉘기 때문에 일종의 분류(classification) 기법
- 로지스틱 회귀분석은 확률을 다루는 분석법이다. 예로 나이가 증가함에 따라 당뇨 확률이 어떻게 변하는가? 허리 둘레와 당뇨 확률의 관계는 어떤가?... 등등... yes/no의 비율을 다루는 분석법
- 로지스틱 회귀분석의 용도
 - ✓ 로지스틱 회귀분석은 분류 모델링에 사용되는 기법으로 새로운 데이터에 대해 "분류를 예측"하거나 "예측 변수 프로파일링"을 할 수 있습니다.
 - ✓ 고객을 재구매 고객과 처음 구매한 고객으로 분류(분류)
 - ✓ 남자 최고 경영진과 여자 최고 경영진을 구별하는 요인 찾기(프로파일링)

로지스틱 회귀분석

- R 프로그램에서 로지스틱 회귀분석에 사용되는 함수는 glm() 함수입니다.
 - ✓ glm() 함수는 일반화 선형 모형(**G**eneralized **L**inear **M**odel)을 추정하는 함수로,
 - ✓ 로지스틱 회귀분석은 glm() 함수를 사용하여 수행
 - ✓ 일반화 선형 모형에서는 다양한 분포의 종속 변수에 적용하기 때문에, 종속 변수가 어떤 분포를 따르고 있는지 옵션을 주는 것이다.
 - ✓ 종속변수가 이항의 값을 갖는 경우, `family = binomial()`로 설정한다.
- glm() 함수를 사용하여 로지스틱 회귀분석을 수행하려면 다음과 같은 인수를 지정해야 합니다.
 - ✓ formula 인수: 종속 변수와 설명 변수의 관계를 나타내는 회귀식을 지정
 - ✓ family 인수: 종속 변수의 분포를 지정한다. 로지스틱 회귀분석의 경우 `family = binomial()`을 지정
 - ✓ data 인수: 분석에 사용할 데이터 세트를 지정

ex) Model <- glm (formular = y ~ x1+x2+x3 , data=ex , family=binomial())

로지스틱 회귀분석

- 회귀분석 모델의 성능 평가는 R^2 를 통해 성능을 평가하지만, 로지스틱 회귀분석은 **확률을 다루는 분석**이기 때문에 분류 모델이 실제로 얼마나 맞추었는가? 를 평가하는 것
- 확률 관점에서 모델의 성능을 평가 하기 위해서는 얻어진 로지스틱 회귀 모델에 의한 확률을 추정하고 그 결과를 평가 해야한다.
- 로지스틱 회귀분석의 결과 해석은 다음과 같은 단계로 진행한다.
 - ✓ 각 설명변수의 표준오차를 사용하여 계수의 유의성을 평가한다. 표준오차가 작은 경우 해당 계수는 유의하다고 할 수 있다.
 - ✓ 각 설명변수의 계수를 해석한다. 설명변수의 계수가 갖는 영향력은 승산 값(odds)으로 평가한다.
- 로지스틱 회귀분석 **모델의 적합도와 성능을 평가**하는 방법과 사용하는 값은 다음과 같다.
 - ✓ Deviance(**이탈도**) 값이 작을수록 모형이 더 적합하다고 할 수 있습니다.
 - ✓ AIC(Akaike Information Criterion) 값은 작은 모형이 더 간결하다고 할 수 있다.
 - ✓ 정확도(Accuracy), 민감도(Sensitivity), 특이도(Specificity)에 의해 평가

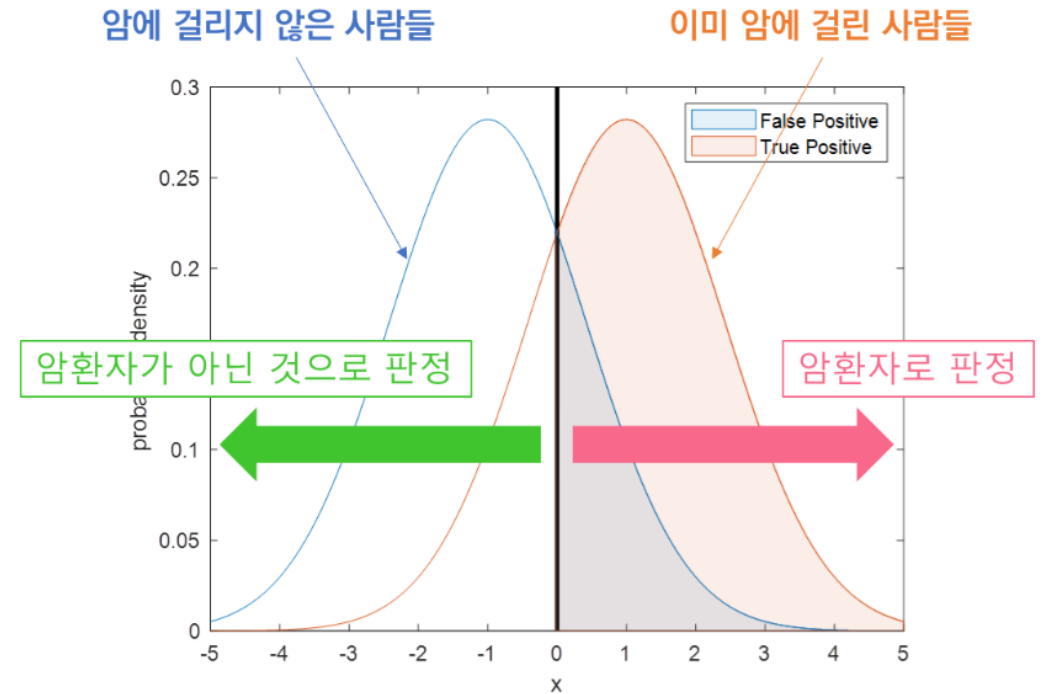
모델의 성능평가

	실제 값(Positive)	실제 값(Negative)
모델 예측값(Positive)	True Positive(TP)	False Positive(FP)
모델 예측값(Negative)	False Negative(FN)	True Negative(TN)

Sensitivity

Specificity

- Accuracy : $\frac{TP+TN}{TP+FP+FN+TN}$: 전체 정확도를 의미합니다.
- Sensitivity(민감도) : $\frac{TP}{TP+FN}$: 실제 Positive 중에서 모델이 **Positive**를 맞추었는가? 에 대한 지표
- Specificity(특이도) : $\frac{TN}{FP+TN}$: 실제 Negative 중에서 모델이 **Negative**를 맞추었는가? 에 대한 지표



로지스틱 회귀분석 예제

- 'mtcars' 데이터셋으로 로지스틱 회귀분석 실시하기
 - ✓ 예제로 사용하는 데이터셋 mtcars 는 1973~74년도에 생산된 32종류의 자동차에 대해 11개 변수를 측정
한 데이터
- 로지스틱 회귀모형은 종속변수가 범주형이기 때문에 vs를 이용해 엔진종류를 예측하기 위한 분석을 한다.

변수 이름	변수 설명	변수 유형
vs	엔진종류(0=V-shaped, 1=straight)	범주형
mpg	갤런당 마일(연비)	수치형
am	변속기 종류(0=automatic, 1=manual)	범주형
wt	차량 무게	수치형

로지스틱 회귀분석 예제

- 로지스틱 회귀분석의 변수선택법 적용하기
 - ✓ 로지스틱 회귀분석도 일반회귀분석처럼 전진대입법(forward), 후진제거법(backward), 단계적 방법(stepwise)을 통해 유의한 변수만 선택하는 방법이 있다
 - ✓ R에서는 step() 함수를 이용한다.
 - ✓ direction 인자에 forward, backward, stepwise를 적용할 수 있다.
- 오른쪽에서처럼 direction = "backward" 를 실행하면
 - ✓ formula 인자에 설정한 설명변수들을 모두 넣어 학습시키고,
 - ✓ 단계적으로 유의하지 않는 변수를 빼는 작업을 수행하게 되며
 - ✓ 최종적으로 최적의 예측변수들만 남게된다.
 - ✓ 변수가 유의한 지를 판단하는 기준은 단계별로 **이탈도(deviance)**를 이용하여 판단할 수 있으며
 - ✓ anova() 함수의 test 인자를 "Chisq"로 설정하여 확인할 수 있습니다.

```
#mpg, am, wt 예측변수를 이용한 vs 변수 예측
```

```
> glm_vs2 <- glm(vs~mpg+am+wt, data=mtcars, family=binomial)
```

```
#변수선택법 적용
```

```
> step_vs <- step(glm_vs, direction = "backward")
```

```
Start: AIC=27.34
```

```
vs ~ mpg + am + wt
```

	Df	Deviance	AIC
- wt	1	20.646	26.646
<none>		19.342	27.342
- mpg	1	21.532	27.532
- am	1	25.298	31.298

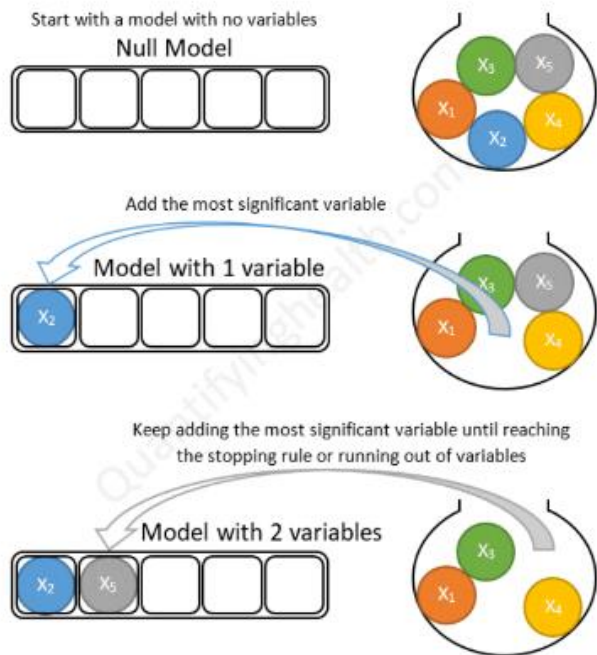
```
Step: AIC=26.65
```

```
vs ~ mpg + am
```

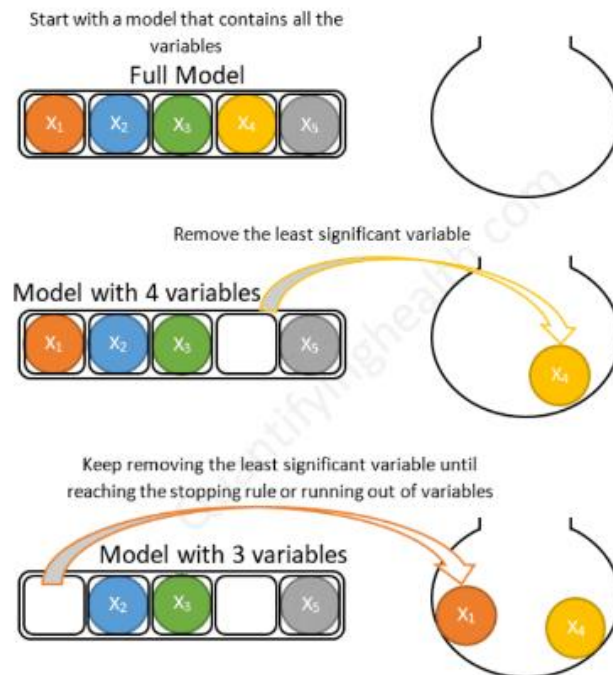
	Df	Deviance	AIC
<none>		20.646	26.646
- am	1	25.533	29.533
- mpg	1	42.953	46.953

변수선택 방법

Forward stepwise selection example with 5 variables:



Backward stepwise selection example with 5 variables:



- backward 방법은 AIC 값을 기준으로 변수를 제거하는 방법입니다.
- backward 방법을 사용하여 변수 선택을 수행하는 방법은,
 1. 모든 독립변수를 포함하는 모형을 생성한다.
 2. AIC 값이 가장 높은 변수를 제거한다.
 3. 제거된 변수를 제외한 모형을 생성한다.
 4. 2번과 3번의 과정을 반복한다.

로지스틱 회귀분석 예제

- 로지스틱 회귀분석의 이탈도(deviance)는 로그 우도값으로서,
 - ✓ 종속 변수의 성공 확률을 얼마나 잘 예측하는지 나타내는 값
 - ✓ null model과 현재 모델의 설명력의 차이를 나타내며
 - ✓ null model은 절편 만을 포함하는 모델로서, deviance 값이 작을수록 현재 모델이 null model보다 더 좋은 설명력을 갖는다는 것을 의미한다.
- 우측에서 보듯이 종속변수 vs에 대한 설명변수 mpg, am의 $\Pr(>\text{Chi})$ 가 0.05이하임으로 통계적으로 유의하다고 할 수 있다.
- wt은 0.05 보다 커서 유의하지 않는다고 판단할 수 있다. 따라서, 앞에서 수행한 변수선택법에서 wt가 가장 먼저 제거되어야 하는 변수임을 확인할 수 있다.

#이탈도 확인

```
> anova(glm_vs2, test="Chisq")
```

Analysis of Deviance Table

Model: binomial, link: logit

Response: vs

Terms added sequentially (first to last)

		Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL				31	43.860	
mpg	1	18.327	30	25.533	1.861e-05 ***	
am	1	4.887	29	20.646	0.02706 *	
wt	1	1.304	28	19.342	0.25348	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

로지스틱 회귀분석 예제

- 'Telco-Customer-Churn' 데이터셋으로 로지스틱 회귀분석 실시하기
 - ✓ 예제로 사용하는 데이터셋 Telco-Customer-Churn은 2019년 1월~12월까지 미국의 한 통신 서비스 회사의 고객 정보를 포함
 - ✓ 데이터셋에는 총 7,043개의 고객정보가 포함되어 있으며, 이 중 1,532명의 고객이 이탈했다.
- 로지스틱 회귀모델 분석에 사용되는 변수의 이름과 유형은 표와 같다.

변수 이름	변수 설명	변수 유형
Churn	고객의 이탈 여부("yes"=이탈, "No"=유지)	범주형
tenure	고객의 서비스 가입 기간	수치형
MonthlyCharges	고객의 월 요금	수치형
TotalCharges	고객이 사용하는 총서비스 요금	수치형

로지스틱 회귀분석 예제

```
Telco_data <- read.csv(file = "C:/Users/User/Desktop/R_data/Telco-Customer-Churn.csv")
str(Telco_data)
head(Telco_data)
```

```
#### 로지스틱 회귀분석을 위해 종속변수는 범주형 변수여야 함.
#### 이를 위해 현재의 char변수를 범주형 변수로 변환시켜줌
```

```
Telco_data$Churn <- factor(Telco_data$Churn)
str(Telco_data)
```

```
glm_Churn <- glm(Churn~tenure+MonthlyCharges+TotalCharges, data = Telco_data, family=binomial)
summary(glm_Churn)
```

>> .csv 엑셀 데이터 셋을 읽어 들인다.

>> 변수 'Churn'을 범주형으로 변환, 저장

>> 3개의 설명변수와 종속변수 'Churn'에 대한 로지스틱 회귀분석 실행

```
> glm_Churn <- glm(Churn~tenure+MonthlyCharges+TotalCharges, data = Telco_data, family=binomial)
Error in eval(family$initialize) :
  y 값들은 반드시 0 이상 1 이하이어야 합니다
> Telco_data$Churn <- factor(Telco_data$Churn)
> str(Telco_data)
'data.frame': 7043 obs. of 21 variables:
 $ customerID      : chr "7590-VHVEG" "5575-GNVDE" "3668-QPYBK" "7795-CFOCW" ...
 $ gender          : chr "Female" "Male" "Male" "Male" ...
 $ SeniorCitizen   : int 0 0 0 0 0 0 0 0 0 ...
 $ Partner         : chr "Yes" "No" "No" "No" ...
 $ Dependents      : chr "No" "No" "No" "No" ...
 $ tenure          : int 1 34 2 45 2 8 22 10 28 62 ...
 $ PhoneService    : chr "No" "Yes" "Yes" "No" ...
 $ MultipleLines    : chr "No phone service" "No" "No" "No phone service" ...
 $ InternetService : chr "DSL" "DSL" "DSL" "DSL" ...
 $ OnlineSecurity   : chr "No" "Yes" "Yes" "Yes" ...
 $ OnlineBackup     : chr "Yes" "No" "Yes" "No" ...
 $ DeviceProtection: chr "No" "Yes" "No" "Yes" ...
 $ Techsupport      : chr "No" "No" "No" "Yes" ...
 $ StreamingTV      : chr "No" "No" "No" "No" ...
 $ StreamingMovies  : chr "No" "No" "No" "No" ...
 $ Contract         : chr "Month-to-month" "One year" "Month-to-month" "One year" ...
 $ PaperlessBilling: chr "Yes" "No" "Yes" "No" ...
 $ PaymentMethod    : chr "Electronic check" "Mailed check" "Mailed check" "Bank transfer (automatic)" ..
 $ MonthlyCharges   : num 29.9 57 53.9 42.3 70.7 ...
 $ TotalCharges     : num 29.9 1889.5 108.2 1840.8 151.7 ...
 $ Churn            : Factor w/ 2 levels "No","Yes": 1 1 2 1 2 2 1 1 2 1 ...
```

로지스틱 회귀분석 예제

```
> glm_Churn <- glm(Churn~tenure+MonthlyCharges+TotalCharges, data =  
Telco_data, family=binomial)  
> summary(glm_Churn)
```

```
call:  
glm(formula = Churn ~ tenure + MonthlyCharges + TotalCharges,  
     family = binomial, data = Telco_data)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.8474	-0.7316	-0.4042	0.8036	3.1441

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.599e+00	1.173e-01	-13.628	<2e-16 ***
tenure	-6.711e-02	5.458e-03	-12.297	<2e-16 ***
MonthlyCharges	3.020e-02	1.717e-03	17.585	<2e-16 ***
TotalCharges	1.451e-04	6.144e-05	2.361	0.0182 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 8143.4 on 7031 degrees of freedom
Residual deviance: 6376.2 on 7028 degrees of freedom
(결측으로 인하여 11개의 관측치가 삭제되었습니다.)

AIC: 6384.2

Number of Fisher Scoring iterations: 6

- 'Null deviance'와 'Residual deviance'를 해석하는 방법은,
: 'Null deviance' 값과 'Residual deviance' 값의 차이가 크면, 독립 변수가 모형의 적합도를 개선하는 데 도움이 되었다고 판단할 수 있다.
- deviance는 모형이 관측 데이터를 얼마나 잘 설명하는지를 나타내는 지표
- deviance(이탈도) 값이 낮을수록 모형이 관측 데이터를 더 잘 설명
- Null deviance: 독립 변수가 없는 모형에서의 deviance
- Residual deviance: 독립 변수가 있는 모형에서의 deviance
- AIC(Akaike Information Criterion) 값은 모형의 적합도와 복잡도를 함께 고려한 지표로서,
- AIC 값이 낮을수록 모형이 더 적합하다고 판단할 수 있으나
- AIC 값이 절대적인 지표는 아니며, 다른 지표와 함께 고려하여 모형을 평가하는 것이 좋다.

로지스틱 회귀분석 예제

```
> glm_churn <- glm(Churn~tenure+MonthlyCharges+TotalCharges, data =  
Telco_data, family=binomial)  
> summary(glm_churn)
```

```
Call:  
glm(formula = Churn ~ tenure + MonthlyCharges + TotalCharges,  
     family = binomial, data = Telco_data)
```

```
Deviance Residuals:  
    Min       1Q   Median       3Q      Max  
-1.8474  -0.7316  -0.4042   0.8036   3.1441
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-1.599e+00	1.173e-01	-13.628	<2e-16	***
tenure	-6.711e-02	5.458e-03	-12.297	<2e-16	***
MonthlyCharges	3.020e-02	1.717e-03	17.585	<2e-16	***
TotalCharges	1.451e-04	6.144e-05	2.361	0.0182	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 8143.4 on 7031 degrees of freedom
Residual deviance: 6376.2 on 7028 degrees of freedom
(결측으로 인하여 11개의 관측치가 삭제되었습니다.)
AIC: 6384.2

Number of Fisher Scoring iterations: 6

```
> odds <- exp(coef(glm_churn)); odds
```

(Intercept)	tenure	MonthlyCharges	TotalCharges
0.2021334	0.9350881	1.0306603	1.0001451

```
> range(Telco_data$tenure);range(Telco_data$MonthlyCharges)  
[1] 0 72  
[1] 18.25 118.75
```

- 승산(odds) 값은 특정변수의 값이 1단위 증가할 때 종속 변수의 성공 확률이 몇 배가 되는지를 나타내는 값

tenure의 회귀계수: -0.0671 의미는:

tenure 가 1단위 증가하면, Churn="Yes"일 오즈 $\exp(-0.0671) \approx 0.94$ 배.

➔ 즉, 통신서비스사를 변경 할 확률이 6% 감소하며

MonthlyCharges의 회귀계수: 0.03 의미는:

MonthlyCharges 가 1단위 증가하면, Churn="Yes"일 오즈가 $\exp(0.03) \approx 1.03$ 배

➔ 즉, 통신서비스사를 변경 할 확률이 3% 증가한다는 것을 의미

- 월 사용금액(MonthlyCharges)은 통신서비스사를 변경할 확률에 유의한 영향을 미친다.
- 그러나 MonthlyCharges 의 변동 범위가 크므로, MonthlyCharges의 1단위 증가가 통신서비스사 변경 확률에 미치는 영향은 상대적으로 작다.

➤ 변수의 유의성 여부와 변수의 영향력(오즈 값의 크기) 간에는 상관이 없다.

승산(odds)값의 의미

```
> glm_churn <- glm(Churn~tenure+MonthlyCharges+TotalCharges, data =  
Telco_data, family=binomial)  
> summary(glm_churn)
```

```
Call:  
glm(formula = Churn ~ tenure + MonthlyCharges + TotalCharges,  
     family = binomial, data = Telco_data)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.8474	-0.7316	-0.4042	0.8036	3.1441

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.599e+00	1.173e-01	-13.628	<2e-16 ***
tenure	-6.711e-02	5.458e-03	-12.297	<2e-16 ***
MonthlyCharges	3.020e-02	1.717e-03	17.585	<2e-16 ***
TotalCharges	1.451e-04	6.144e-05	2.361	0.0182 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 8143.4 on 7031 degrees of freedom
Residual deviance: 6376.2 on 7028 degrees of freedom
(결측으로 인하여 11개의 관측치가 삭제되었습니다.)
AIC: 6384.2

Number of Fisher scoring iterations: 6

```
> odds <- exp(coef(glm_churn)); odds
```

(Intercept)	tenure	MonthlyCharges	TotalCharges
0.2021334	0.9350881	1.0306603	1.0001451

```
> range(Telco_data$tenure); range(Telco_data$MonthlyCharges)
```

```
[1] 0 72  
[1] 18.25 118.75
```

$$p = \beta_0 + \beta_1 x_1 + \dots + \beta_n x_n$$

$$p = \frac{\text{Odds}}{1 + \text{odds}} \quad \leftarrow \quad \text{Odds} = e^{\beta_0 + \beta_1 x + \dots + \beta_n x_n}$$

tenure의 회귀계수: -0.0671 의미는:

tenure 가 1단위 증가하면, Churn="Yes"일 오즈 $\exp(-0.0671) \approx 0.94$ 배.

→ 즉, 통신서비스사를 변경 할 확률이 6% 감소하며

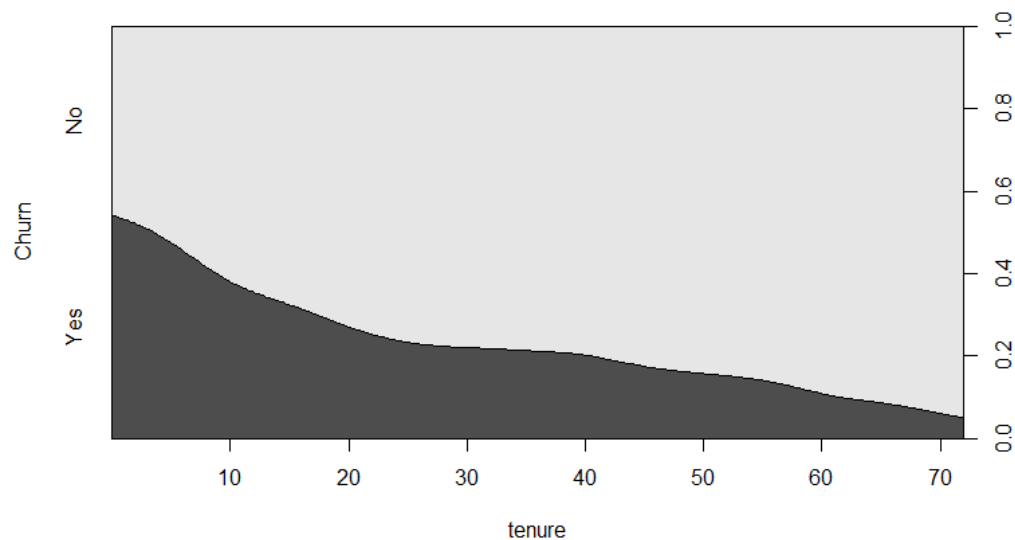
MonthlyCharges의 회귀계수: 0.03 의미는:

MonthlyCharges 가 1단위 증가하면, Churn="Yes"일 오즈가 $\exp(0.03) \approx 1.03$ 배

→ 즉, 통신서비스사를 변경 할 확률이 3% 증가한다는 것을 의미

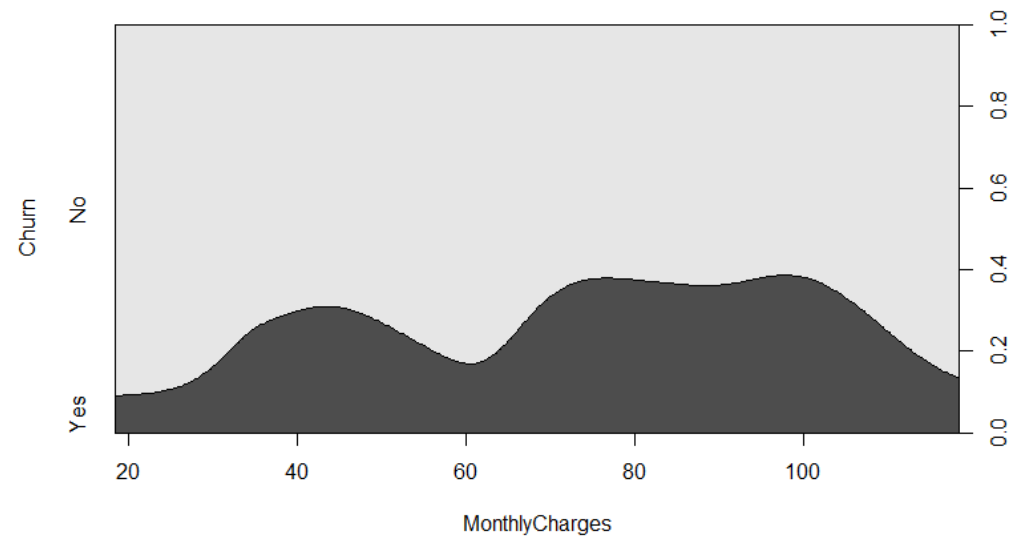
로지스틱 회귀분석 예제

- `cdplot()` 함수는 로지스틱 회귀분석에서 설명변수의 변화에 따른 종속변수의 분포를 시각화하는 함수
 - ✓ 설명변수의 변화에 따른 종속변수의 성공 확률의 변화를 시각적으로 확인할 수 있다.
 - ✓ 연속형 변수인 설명변수 별로 범주형 종속변수의 성공 확률이 어떻게 다른지 확인할 수 있다.
- 아래 그림은 계약기간(`tenure`)이 길어질수록 Churn의 확률 즉, 통신서비스 회사를 변경할 가능성이 크게 줄어든다는 것을 보여준다.



`tenure` 가 1단위 증가하면, `Churn="Yes"`일 오즈 $\exp(-0.0671) \approx 0.94$ 배.

→ 통신서비스사를 변경 할 확률이 6% 감소

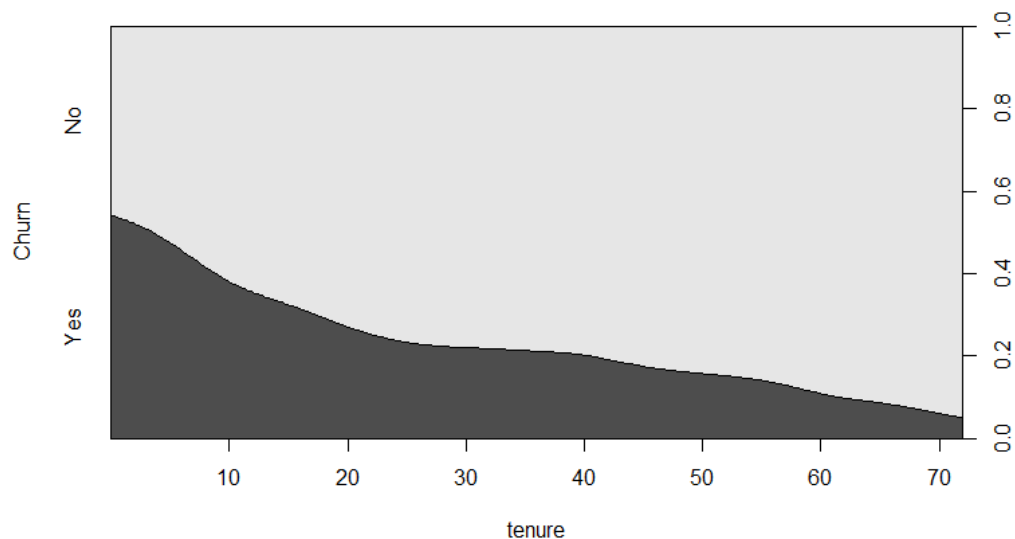


`MonthlyCharges` 가 1단위 증가하면, `Churn="Yes"`일 오즈가 $\exp(0.03) \approx 1.03$ 배

→ 통신서비스사를 변경 할 확률이 3% 증가

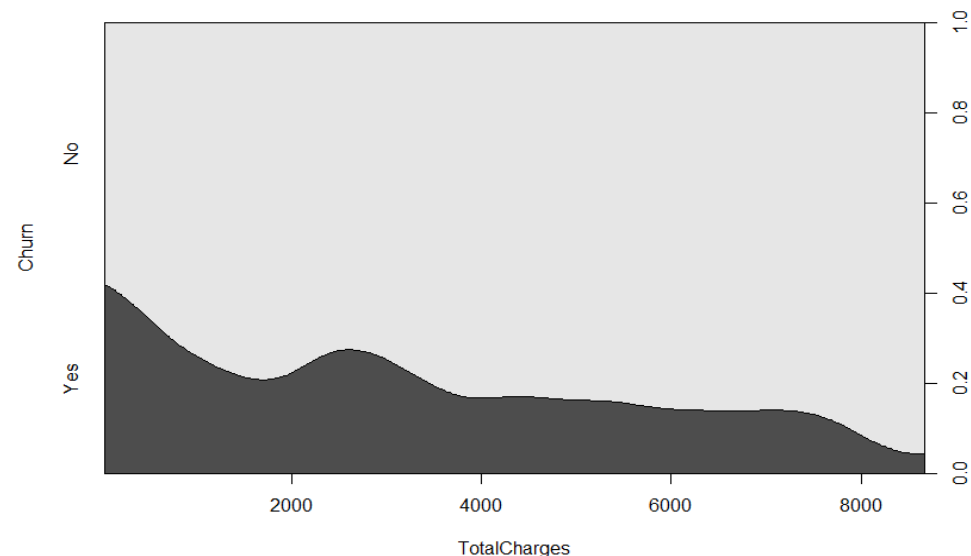
로지스틱 회귀분석 예제

- 아래 그림에서 보듯이 Churn ~ tenure 과 Churn ~ TotalChanges 관계의 특성이 거의 유사한 것을 알 수 있다.



```
> odds <- exp(coef(glm_Churn)); odds
      (Intercept)      tenure MonthlyCharges      TotalCharges
      0.2021334      0.9350881      1.0306603      1.0001451
>
> step_glm_Churn <- step(glm_Churn, direction = "backward")
Start: AIC=6384.23
Churn ~ tenure + MonthlyCharges + TotalCharges
```

	Df	Deviance	AIC
<none>		6376.2	6384.2
- TotalCharges	1	6381.9	6387.9
- tenure	1	6566.8	6572.8
- MonthlyCharges	1	6719.0	6725.0



- 'TotalCharges' 와 'tenure'의 특성이 거의 유사하게 보인다.
- 그럼에도, TotalCharges 의 odds값은 1.0001로 TotalCharges 1단위 증가하여도 Churn이 "Yes" 로 변경될 확률에 거의 영향을 미치지 못한다.

왜?????

로지스틱 회귀분석 예제

```
> vif(glm_Churn)
      tenure MonthlyCharges  TotalCharges 
13.70819      2.29811      17.79336 

> glm_Churn_2 <- glm(Churn~tenure+MonthlyCharges, data = Telco_data,
> summary(glm_Churn_2)

Call:
glm(formula = Churn ~ tenure + MonthlyCharges, family = binomial,
    data = Telco_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max 
-1.8846 -0.7146 -0.4135  0.7859  3.0066 

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.802436   0.086557  -20.82  <2e-16 ***
tenure        -0.054850   0.001689  -32.47  <2e-16 ***
MonthlyCharges  0.032954   0.001299   25.37  <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: 8150.1  on 7042  degrees of freedom
Residual deviance: 6394.4  on 7040  degrees of freedom
AIC: 6400.4

Number of Fisher Scoring iterations: 5

> vif(glm_Churn_2)
      tenure MonthlyCharges 
1.294147      1.294147
```

```
> step_glm_Churn <- step(glm_Churn, direction = "backward")
Start:  AIC=6384.23
Churn ~ tenure + MonthlyCharges + TotalCharges
```

	Df	Deviance	AIC
<none>		6376.2	6384.2
- TotalCharges	1	6381.9	6387.9
- tenure	1	6566.8	6572.8
- MonthlyCharges	1	6719.0	6725.0

```
> step_glm_Churn_2 <- step(glm_Churn_2, direction = "backward")
Start:  AIC=6400.35
Churn ~ tenure + MonthlyCharges
```

	Df	Deviance	AIC
<none>		6394.4	6400.4
- MonthlyCharges	1	7191.9	7195.9
- tenure	1	7878.2	7882.2

```
> odds_2 <- exp(coef(glm_Churn_2)); odds_2
      (Intercept)      tenure MonthlyCharges 
      0.1648967      0.9466274      1.0335029
```

```
> odds <- exp(coef(glm_Churn)); odds
      (Intercept)      tenure MonthlyCharges  TotalCharges 
      0.2021334      0.9350881      1.0306603      1.0001451
```

모델 성능평가 방법

- anova 분석은 독립변수의 영향력을 평균 차이로 측정한다. 반면, 로지스틱 회귀분석은 독립 변수의 영향력을 오즈비로 측정합니다. 따라서 anova 분석을 적용하면 독립 변수의 영향력을 정확하게 측정하기 어렵다.
- 로지스틱 회귀분석에서 처음 모델과 개선 후 모델 간의 효과를 분석하기 위해서는 anova 분석 대신 다음과 같은 방법을 사용한다.
 - ✓ ROC 커브 비교
: ROC 커브는 모델의 예측 성능을 시각적으로 비교하는 방법입니다. 처음 모델과 개선 후 모델의 ROC 커브를 비교하여 개선 후 모델이 더 좋은 성능을 보이는지 확인할 수 있습니다.
 - ✓ AUC 비교
: AUC는 ROC 커브의 아래 면적을 나타내는 지표이다. AUC 값이 높을수록 모델의 예측 성능이 좋다. 따라서 처음 모델과 개선 후 모델의 AUC 값을 비교하여 개선 후 모델이 더 좋은 성능을 보이는지 확인할 수 있다.
 - ✓ 지적적 차이 검정
: 지적적 차이 검정은 두 모형의 분류 정확도 차이가 통계적으로 유의한지 검정하는 방법이다. 처음 모델과 개선 후 모델의 분류 정확도 차이가 통계적으로 유의하다면, 개선 후 모델이 더 좋은 성능을 보이는 것으로 판단할 수 있다.

모델 성능평가 방법

```
> anova(glm_Churn, test="Chisq")
```

Analysis of Deviance Table

Model: binomial, link: logit

Response: Churn

Terms added sequentially (first to last)

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			7031	8143.4	
tenure	1	967.09	7030	7176.3	< 2e-16 ***
MonthlyCharges	1	794.37	7029	6381.9	< 2e-16 ***
TotalCharges	1	5.67	7028	6376.2	0.01728 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

- Resid. Dev 값이 많이 하락 할수록 의미있는 변수이며, 그 값이 작아질수록 모델의 성능이 좋아진다는 것을 의미한다.
- 변수가 하나씩 포함 될 때마다 Resid. Dev값으로 모델의 성능이 얼마나 나아지는지 확인 할 수 있다.
- Tenure 변수가 포함될 때 Resid. Dev 값이 가장 많이 하락한 것을 알 수 있다.

모델 성능평가 방법

- confusionMatrix는 **학습한 모델의 성능**을 평가하는 데 사용되는 지표이다.
- confusionMatrix는 실제 값과 예측 값을 비교하여 분류 모델의 정밀도, 재현율, F1-score, ROC-AUC 등을 계산한다.

```
> confusionMatrix(Telco_data_test$Churn,PREDICTED_C)
```

Confusion Matrix and Statistics

	Reference	
Prediction	No	Yes
No	2674	249
Yes	623	454

Accuracy : 0.782
95% CI : (0.7689, 0.7947)
No Information Rate : 0.8242
P-Value [Acc > NIR] : 1

Kappa : 0.3778

McNemar's Test P-Value : <2e-16

Sensitivity : 0.8110
Specificity : 0.6458
Pos Pred Value : 0.9148
Neg Pred Value : 0.4215
Prevalence : 0.8243
Detection Rate : 0.6685
Detection Prevalence : 0.7308
Balanced Accuracy : 0.7284

'Positive' Class : No

```
PREDICTED_C = ifelse(prediction > 0.5 , "Yes", "No")  
PREDICTED_C = as.factor(PREDICTED_C)
```

- "No Information Rate"는 참조변수의 수준의 비율이다. 예제의 경우, 참조 변수의 수준은 "Yes"와 "No"이고, 각각의 비율은 0.5이다.
- 따라서, 0.8242는 "No Information Rate" 기준 0.5보다 상당히 높아서 성능이 우수하다는 것을 의미
- "P-Value [Acc > NIR]"는 정확도가 No Information Rate보다 유의하게 높은지 여부를 나타내는 값이다.
- 예제의 경우, P-Value가 1로서 정확도가 No Information Rate보다 유의하게 높다는 것을 의미

즉, 예제의 "혼동 행렬"은 정확도가 0.8242이고, 정확도가 No Information Rate보다 유의하게 높다는 것을 보여준다.

- "prediction > 0.5" 에서 확률의 구분 짓는 값을 "**cut-off value**" 라고 한다

모델 성능평가 방법

```
> confusionMatrix(Telco_data_test$Churn,PREDICTED_C)
```

Confusion Matrix and Statistics

	Reference	
Prediction	No	Yes
No	2674	249
Yes	623	454

Accuracy : 0.782

95% CI : (0.7689, 0.7947)

No Information Rate : 0.8242

P-Value [Acc > NIR] : 1

Kappa : 0.3778

Mcnemar's Test P-Value : <2e-16

Sensitivity : 0.8110

Specificity : 0.6458

Pos Pred Value : 0.9148

Neg Pred Value : 0.4215

Prevalence : 0.8243

Detection Rate : 0.6685

Detection Prevalence : 0.7308

Balanced Accuracy : 0.7284

'Positive' Class : No

```
> table(prediction > 0.5, Telco_data_test$Churn)
```

	No	Yes
FALSE	2674	623
TRUE	249	454

▪ Kappa : 0.3778

Kappa는 모델의 적합도를 평가하는 지표다. Kappa 값이 0에 가까울수록 모델의 적합도가 낮고, 1에 가까울수록 모델의 적합도가 높다.

- 예제의 경우, Kappa 값이 0.3778이므로 모델의 적합도가 보통 수준임을 알 수 있습니다.

▪ McNemar's Test P-Value : <2e-16

McNemar's Test는 두 분류 모델의 성능을 비교하는 데 사용되는 검정 방법이다.

McNemar's Test P-Value가 작을수록 두 모델의 성능 차이가 유의미하다는 것을 의미합니다.

- 예제의 경우, McNemar's Test P-Value가 <2e-16이므로 두 모델의 성능 차이가 유의미하다는 것을 알 수 있다.

▪ Sensitivity : 0.8110

Sensitivity는 실제 양성 예측치를 나타내는 지표이다. Sensitivity 값이 높을수록 모델이 양성 사례를 정확하게 예측할 수 있음을 의미한다.

- 예제의 경우, Sensitivity 값이 0.8110이므로 모델이 양성 사례를 정확하게 예측할 수 있는 수준임을 알 수 있다.

▪ Specificity : 0.6458

Specificity는 실제 음성 예측치를 나타내는 지표이다. Specificity 값이 높을수록 모델이 음성 사례를 정확하게 예측할 수 있음을 의미한다.

- 예제의 경우, Specificity 값이 0.6458이므로 모델이 음성 사례를 정확하게 예측할 수 있는 수준임을 알 수 있다.

모델 성능평가 방법

```
> confusionMatrix(Telco_data_test$Churn,PREDICTED_C)
Confusion Matrix and Statistics
```

	Reference	
Prediction	No	Yes
No	2674	249
Yes	623	454

Accuracy : 0.782
95% CI : (0.7689, 0.7947)
No Information Rate : 0.8242
P-value [Acc > NIR] : 1

Kappa : 0.3778

McNemar's Test P-value : <2e-16

Sensitivity : 0.8110
Specificity : 0.6458

Pos Pred Value : 0.9148
Neg Pred Value : 0.4215
Prevalence : 0.8243

Detection Rate : 0.6685
Detection Prevalence : 0.7308
Balanced Accuracy : 0.7284

'Positive' Class : No

```
> table(prediction > 0.5, Telco_data_test$Churn)
```

	No	Yes
FALSE	2674	623
TRUE	249	454

■ Positive Predictive Value : 0.9148

Positive Predictive Value는 모델이 예측한 양성 사례 중 실제로 양성인 사례의 비율을 나타내는 지표이다. Positive Predictive Value 값이 높을수록 모델이 예측한 양성 사례가 실제로 양성 사례일 가능성이 높음을 의미한다.

- 예제의 경우, Positive Predictive Value 값이 0.9148이므로 모델이 예측한 양성 사례가 실제로 양성 사례일 가능성이 높다는 것을 알 수 있다.

■ Negative Predictive Value : 0.4215

Negative Predictive Value는 모델이 예측한 음성 사례 중 실제로 음성인 사례의 비율을 나타내는 지표다. Negative Predictive Value 값이 높을수록 모델이 예측한 음성 사례가 실제로 음성 사례일 가능성이 높음을 의미한다.

- 예제의 경우, Negative Predictive Value 값이 0.4215이므로 모델이 예측한 음성 사례가 실제로 음성 사례일 가능성은 낮다는 것을 알 수 있다.

■ Prevalence : 0.8243

Prevalence는 실제 양성 사례의 비율을 나타내는 지표다. Prevalence 값이 높을수록 데이터셋에서 양성 사례가 많다는 것을 의미한다.

- 예제의 경우, Prevalence 값이 0.8243이므로 데이터셋에서 양성 사례가 많다는 것을 알 수 있습니다.

모델 성능평가 방법

```
> confusionMatrix(Telco_data_test$Churn,PREDICTED_C)
Confusion Matrix and Statistics
```

	Reference	
Prediction	No	Yes
No	2674	249
Yes	623	454

```
      Accuracy : 0.782
    95% CI : (0.7689, 0.7947)
 No Information Rate : 0.8242
P-Value [Acc > NIR] : 1
```

```
      Kappa : 0.3778
```

```
McNemar's Test P-Value : <2e-16
```

```
      Sensitivity : 0.8110
      Specificity : 0.6458
    Pos Pred Value : 0.9148
    Neg Pred Value : 0.4215
      Prevalence : 0.8243
```

```
      Detection Rate : 0.6685
      Detection Prevalence : 0.7308
      Balanced Accuracy : 0.7284
```

```
'Positive' class : No
```

▪ Detection Rate : 0.6685

Detection Rate는 모델이 실제 양성 사례 중 얼마나 많은 사례를 양성으로 예측했는지를 나타내는 지표다. Detection Rate 값이 높을수록 모델이 실제 양성 사례를 많이 예측할 수 있음을 의미한다.

- 예제의 경우, Detection Rate 값이 0.6685이므로 모델이 실제 양성 사례를 많이 예측할 수 있는 수준임을 알 수 있다.

▪ Detection Prevalence : 0.7308

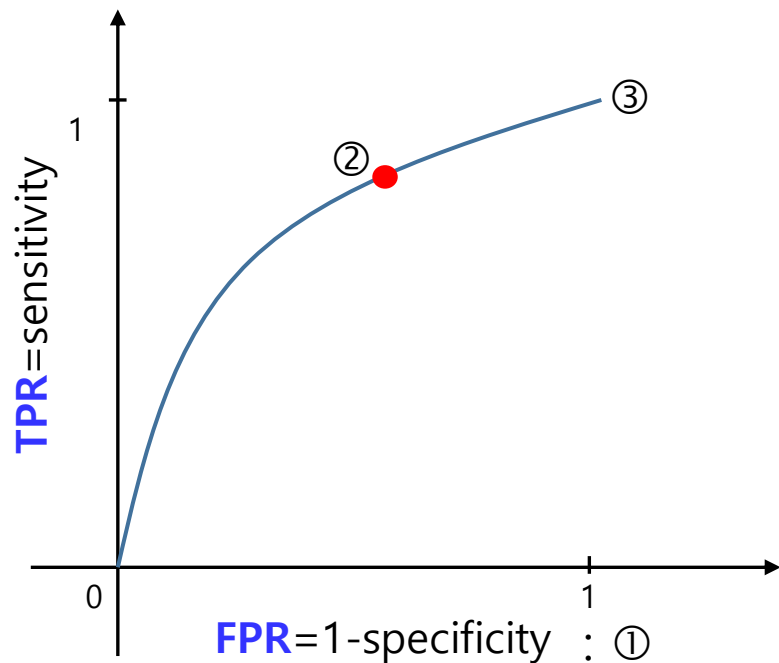
Detection Prevalence는 모델이 양성으로 예측한 사례 중 실제 양성 사례의 비율을 나타내는 지표다. Detection Prevalence 값이 높을수록 모델이 양성으로 예측한 사례가 실제로 양성 사례일 가능성이 높다는 것을 의미한다.

- 예제의 경우, Detection Prevalence 값이 0.7308이므로 모델이 양성으로 예측한 사례가 실제로 양성 사례일 가능성이 높다는 것을 알 수 있다

- **sensitivity** : 양성 데이터를 양성으로 잘 맞추는 것이 중요한 경우 평가
- **positive predictive value** : 예측이 양성으로 나온 데이터 중 실제 양성 데이터의 비율을 높이는 것이 중요한 경우 평가하는 것이 좋다.
- **prevalence** : 전체 데이터 중 양성 데이터의 비율을 파악하고자 하는 경우 평가하는 것이 좋다.

❖ **confusionMatrix**의 결과값이 양성을 기준하여 평가하는 지표가 많은 이유는 분류 모델의 목적이 양성을 정확하게 예측하는 경우가 많기 때문이다.

ROC Curve?

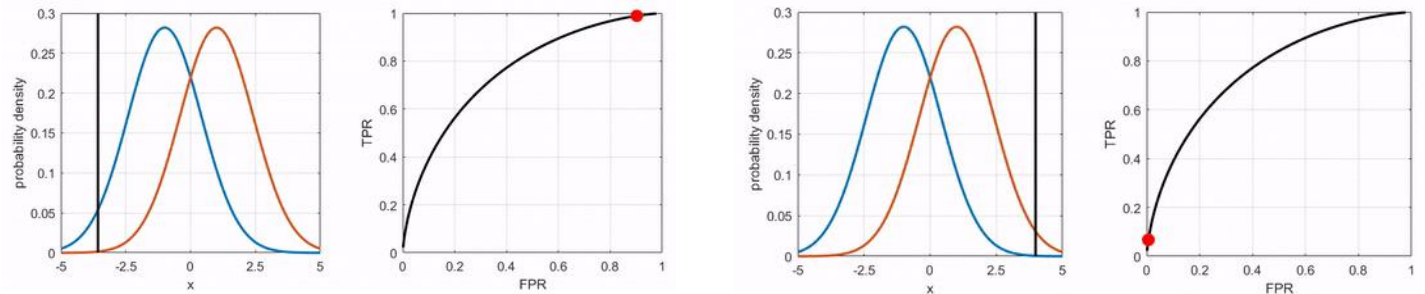


① X, Y축의 의미

- TPR(True **Positive** Rate) : True(그런 것)을 Positive(그렇다)고 판정
- Sensitivity 값으로서 높을수록 **양성 사례를 정확히 판단**할 확률이 높다.
- FPR(False **Positive** Rate) : False(아닌 것)을 Positive(그렇다)로 판정
- 음성을 음성으로 판단하는 Specificity의 반대로, **음성 사례에 대한 잘못된 판단**을 할 확률이 높아진다.

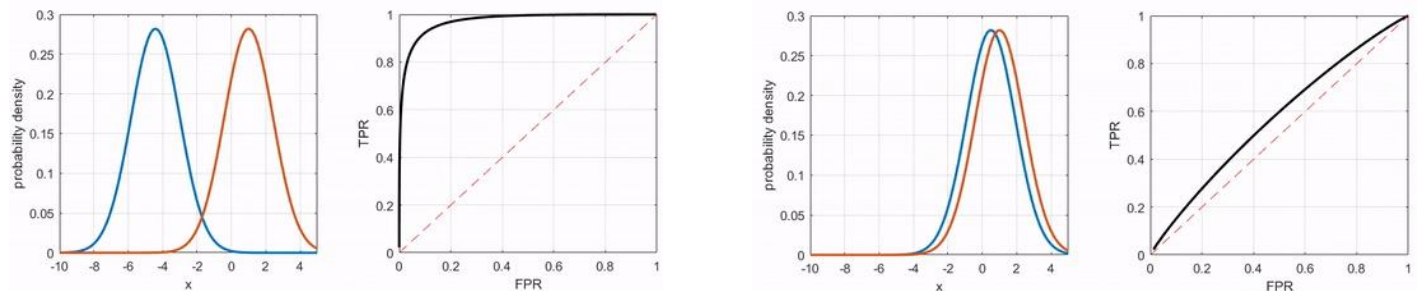
② 현 위의 점 의미 : threshold

- Threshold(판단기준)가 변함에 따라서 FPR 와 TPR 의 값이 달라지는 것을 알 수 있다.

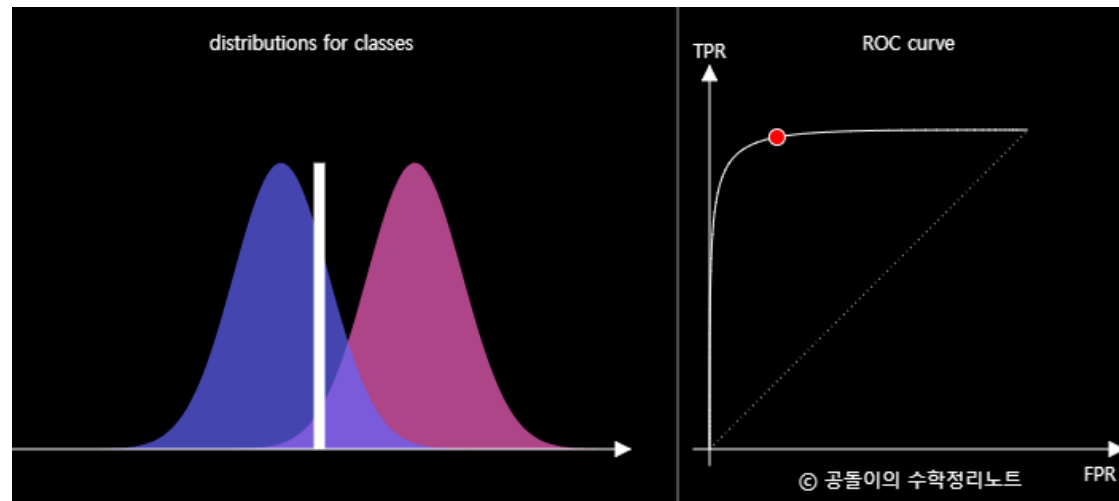
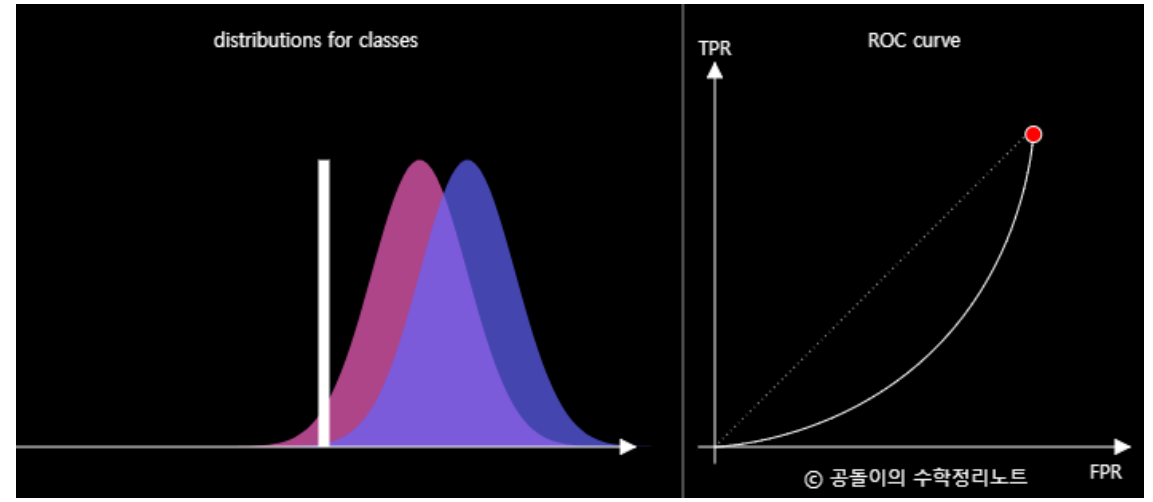
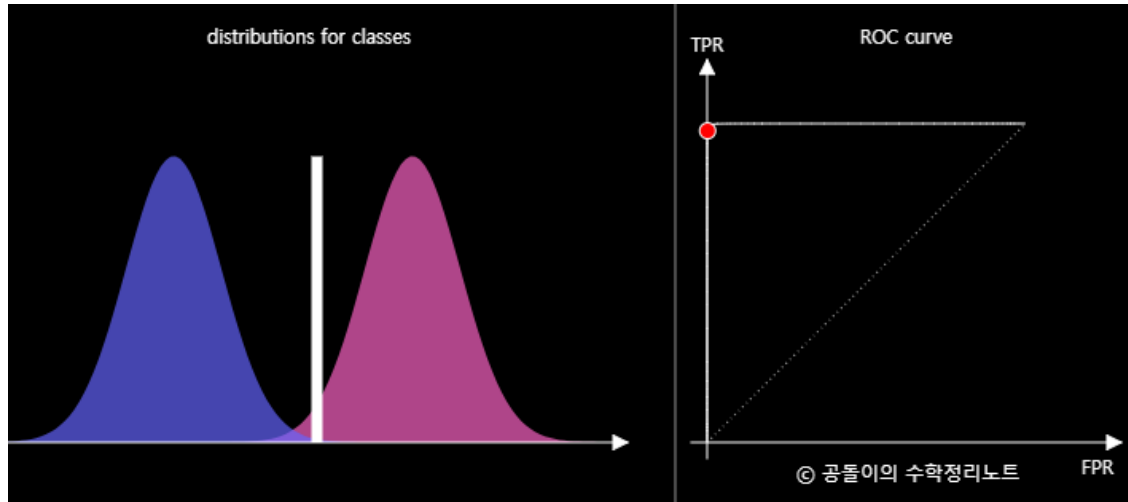


③ 현의 곡률의 의미

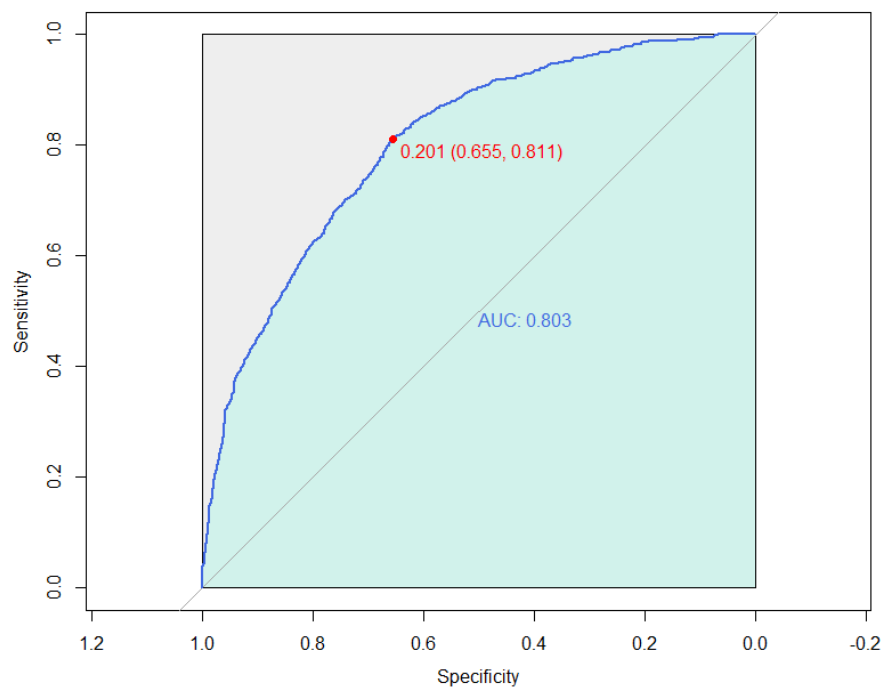
- 두 개 집단을 더 잘 구별할 수 있다면 ROC 커브는 좌 상단에 더 가까워지게 된다.



ROC Curve?



모델 성능평가 방법



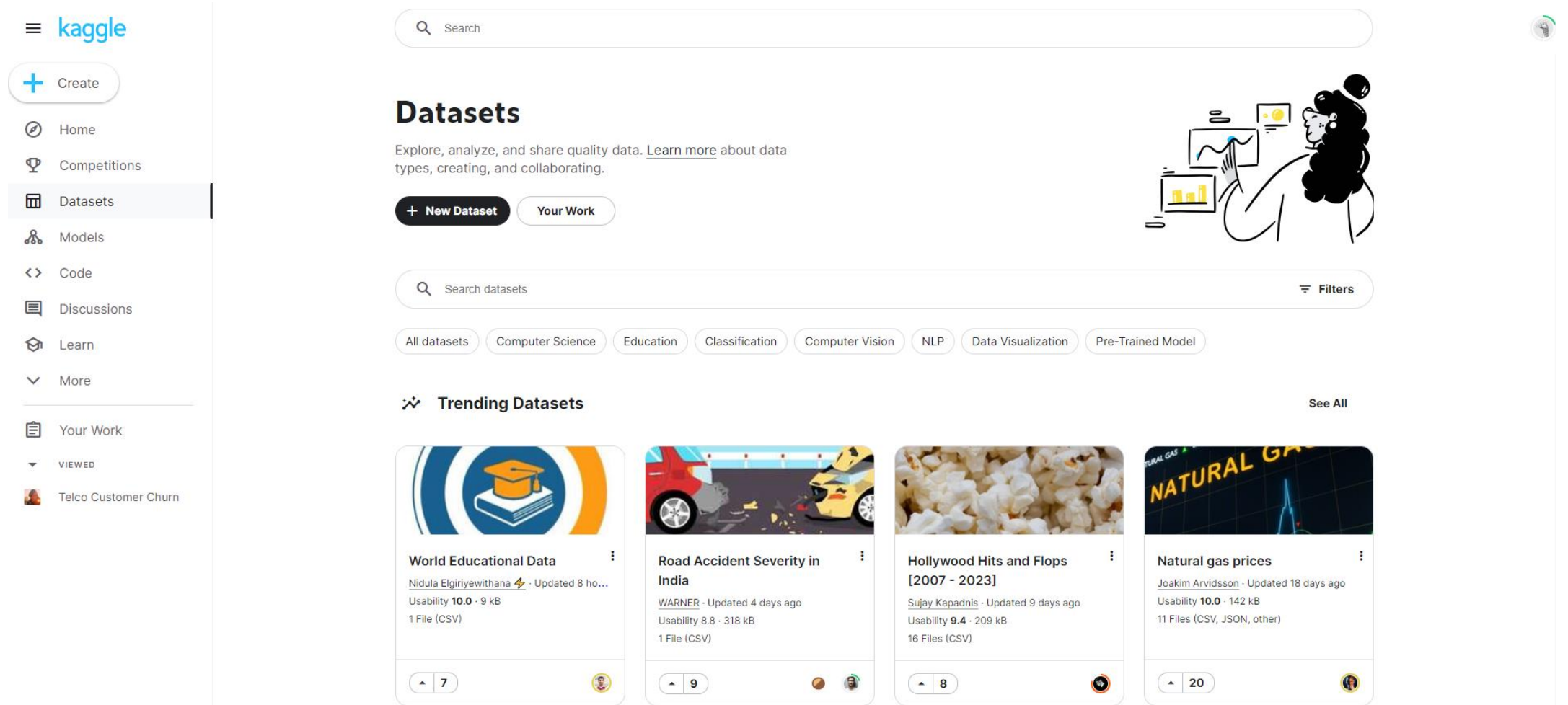
- Roc Curve는 threshold의 값에 따라 Sensitivity와 Specificity의 변화량을 나타낸 그래프
- ROC Curve은 M/L의 **분류 모델의 성능을 평가**하는 데 사용되는 그래프
- AUC(Area under the curve)는 곡선에 해당되는 면적을 나타낸다. AUC값이 높을 수록 바람직한 모델이라고 할 수 있다.

Area Under Curve(AUC)	Evaluation
$AUC \geq 0.9$	Excellent
$0.8 \leq AUC < 0.9$	Good
$0.7 \leq AUC < 0.8$	Fair
$AUC < 0.7$	Poor

- 붉은 점에 표시된 수치 **0.201(0.655,0.811)**
 - **0.201**: threshold 위치를 나타내며,
 - **(0.655,0.811)**은 threshold=0.201 점에서의 (Specificity, Sensitivity) 을 표시

데이터 뱅크

- <https://www.kaggle.com/datasets>



The screenshot shows the Kaggle Datasets homepage. On the left is a sidebar with the Kaggle logo and navigation links: Create, Home, Competitions, Datasets (highlighted), Models, Code, Discussions, Learn, More, Your Work, and a 'VIEWED' section showing 'Telco Customer Churn'. The main content area has a search bar at the top. Below it is the 'Datasets' title and a description: 'Explore, analyze, and share quality data. Learn more about data types, creating, and collaborating.' There are buttons for '+ New Dataset' and 'Your Work'. A second search bar is followed by a 'Filters' button. Below these are category tags: All datasets, Computer Science, Education, Classification, Computer Vision, NLP, Data Visualization, and Pre-Trained Model. The 'Trending Datasets' section features four dataset cards: 'World Educational Data' by Nidula Elgiriye with a usability of 10.0, 'Road Accident Severity in India' by WARNER with a usability of 8.8, 'Hollywood Hits and Flops [2007 - 2023]' by Sujay Kapadnis with a usability of 9.4, and 'Natural gas prices' by Joakim Arvidsson with a usability of 10.0. Each card includes a thumbnail image, title, author, update date, usability score, file size, and file format. A 'See All' link is at the top right of the trending section. At the bottom, a cookie notice states: 'Kaggle uses cookies from Google to deliver and enhance the quality of its services and to analyze traffic.' with 'Learn more.' and 'Ok, Got it.' buttons.