Logistička regresija - šta, kako i kad?

N.Pejovic

Korišćeni R paketi:

```
library(dplyr)
library(ggplot2)
library(broom)
library(readr)
```

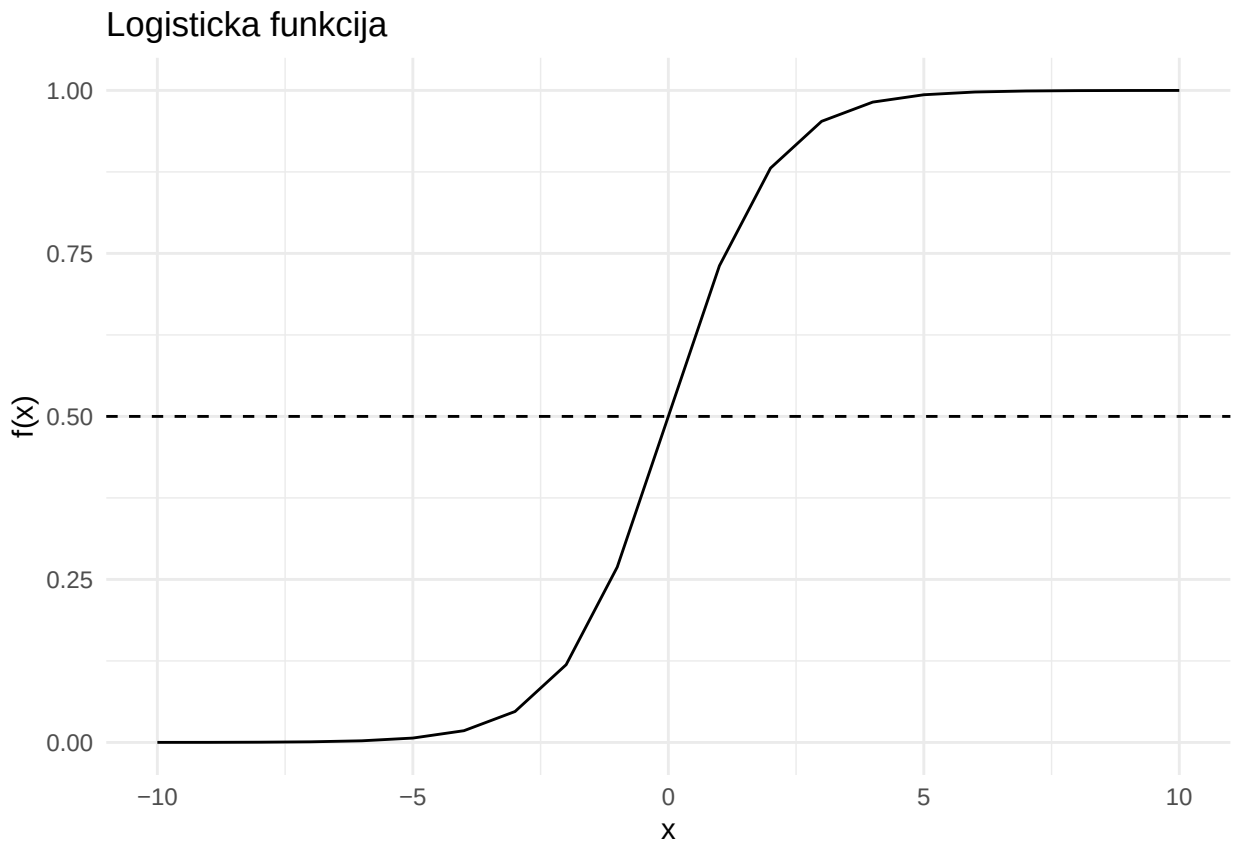
Otkud naziv 'logistička'?

Priču započinjemo sredinom 19. vijeka kada belgijski naučnik Pjer Fransoa Verhulst, baveći se istraživanjem rasta ljudske populacije, dolazi do logističke funkcije, kako je sam nazvao. Funkcija, kao i svaka druga matematička funkcija uzima brojeve, nešto radi sa njima i izbaca neki rezultat. Ono što logistička funkcija radi sa brojevima prikazano je sledećom formulom:

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

Figure 1: Pjer Fransoa Verhulst

Precizno, ovo je standardna logistička funkcija, ali je i dalje logistička funkcija. "Uzme' 1 i"izbaci" 0.7310, uzme "-1" i 'izbaci' 0.2689. Koje god vrijednosti da uzme od – besonačno, do + beskonačno "izbaciće" rezultat od 0 do 1. Tako za vrijednost x-a od -10 do 10 logistička funkcija ima sledeći izgled:



Logistička funkcija pripada sigmoidnim funkcijama, a to su funkcije koje imaju izgled latiničnog slova S, što se i može vidjeti sa grafika.

Od logističke funkcije do logističke regresije

Logistička regresija je blisko povezana sa logističkom funkcijom. Rađe nego davanje definicije na samom početku, prvo ćemo se suočiti sa problemom koji želimo da riješimo i vidjeti kako nam logistička regresija može pomoći u tome. Recimo da jedna telekomunikaciona kompanija želi da predvidi odliv svojih korisnika i sakupila je određene podatke. Podaci o klijentima dostupni su na sledećem **linku.**

Dataset ćemo nazvati 'telco', i pogledati strukturu:

```
read_csv("telco.csv") -> telco
glimpse(telco)
```

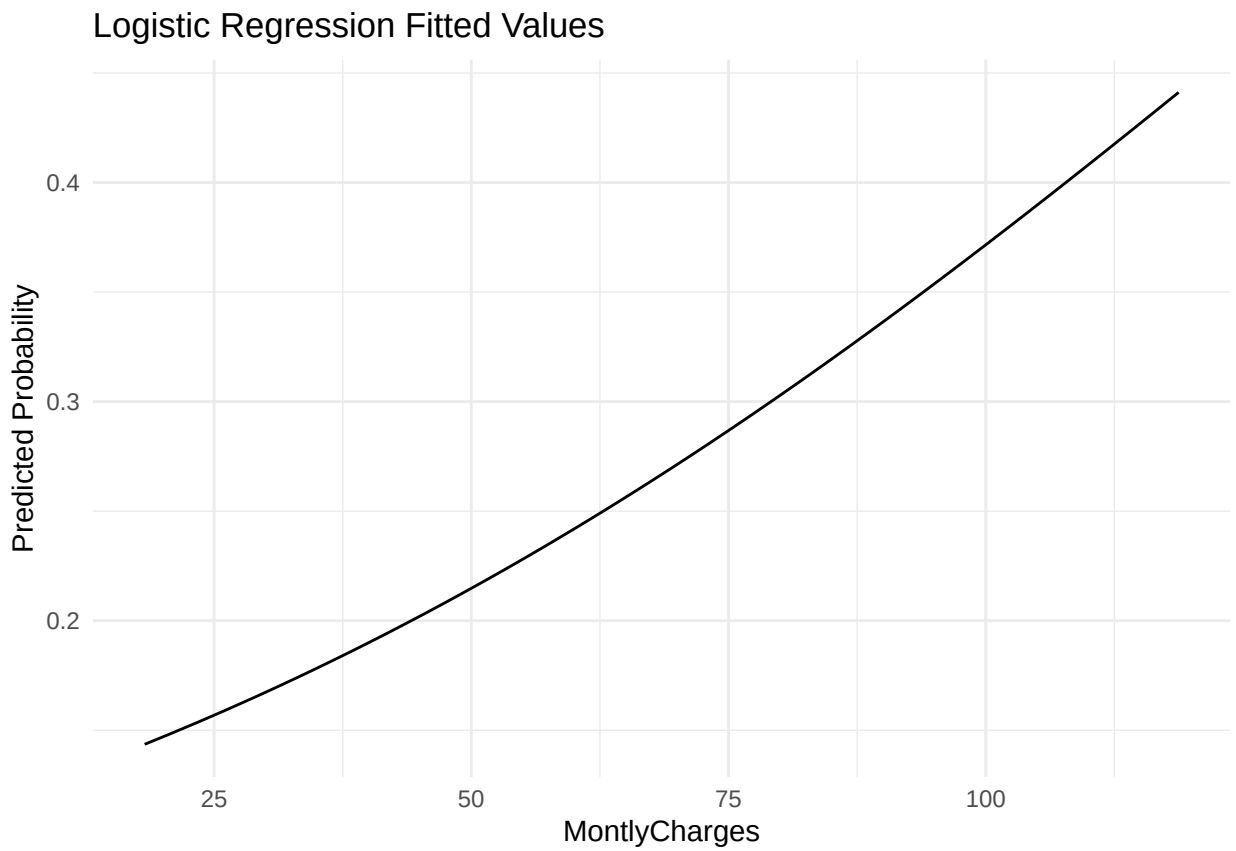
```
## Rows: 7,043
## Columns: 21
## $ customerID       <chr> "7590-VHVEG", "5575-GNVDE", "3668-QPYBK", "7795-CFOCW~
## $ gender           <chr> "Female", "Male", "Male", "Male", "Female", "Female",~
## $ SeniorCitizen    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ Partner          <chr> "Yes", "No", "No", "No", "No", "No", "No", "No", "Yes~
## $ Dependents       <chr> "No", "No", "No", "No", "No", "No", "Yes", "No", "No"~
```

```
## $ tenure           <dbl> 1, 34, 2, 45, 2, 8, 22, 10, 28, 62, 13, 16, 58, 49, 2~
## $ PhoneService     <chr> "No", "Yes", "Yes", "No", "Yes", "Yes", "Yes", "No", ~
## $ MultipleLines    <chr> "No phone service", "No", "No", "No phone service", "~
## $ InternetService  <chr> "DSL", "DSL", "DSL", "DSL", "Fiber optic", "Fiber opt~
## $ OnlineSecurity   <chr> "No", "Yes", "Yes", "Yes", "No", "No", "No", "Yes", "~
## $ OnlineBackup     <chr> "Yes", "No", "Yes", "No", "No", "No", "Yes", "No", "N~
## $ DeviceProtection <chr> "No", "Yes", "No", "Yes", "No", "Yes", "No", "No", "Y~
## $ TechSupport      <chr> "No", "No", "No", "Yes", "No", "No", "No", "No", "Yes~
## $ StreamingTV      <chr> "No", "No", "No", "No", "No", "Yes", "Yes", "No", "Ye~
## $ StreamingMovies  <chr> "No", "No", "No", "No", "No", "Yes", "No", "No", "Yes~
## $ Contract         <chr> "Month-to-month", "One year", "Month-to-month", "One ~
## $ PaperlessBilling <chr> "Yes", "No", "Yes", "No", "Yes", "Yes", "Yes", "No", ~
## $ PaymentMethod    <chr> "Electronic check", "Mailed check", "Mailed check", "~
## $ MonthlyCharges   <dbl> 29.85, 56.95, 53.85, 42.30, 70.70, 99.65, 89.10, 29.7~
## $ TotalCharges     <dbl> 29.85, 1889.50, 108.15, 1840.75, 151.65, 820.50, 1949~
## $ Churn            <chr> "No", "No", "Yes", "No", "Yes", "Yes", "No", "No", "Y~
```

Naš dataset sadrži podatke o 7043 klijenta koji su raspoređeni u 21 kolonu. Kolona od interesa nam je `Churn` kolona koja nam govorio da li je klijent prestao da koristio usluge kompanije ili nije. Kako želimo da prikažemo logiku i mehanizam logističke funkcije, prije nego da se bavimo specifikacijom modela, modeliraćemo kako varijabla mjesečne pretplate `MonthlyCharges` utiče na do da li će se klijent odliti. U našem modelu imamo zavisnu varijablu `Churn` i jednu nezavisnu varijablu `MonthlyCharges`. Modeliranje vjerovatnoće binarne zavisne varijable, kao što je varijabla `Churn` sa ishodima `Yes` i `No` je upravo situacija u kojoj koristimo logističku regresiju. Klijenta koji se odlio ćemo označiti sa 1, a onog koji nije sa 0.

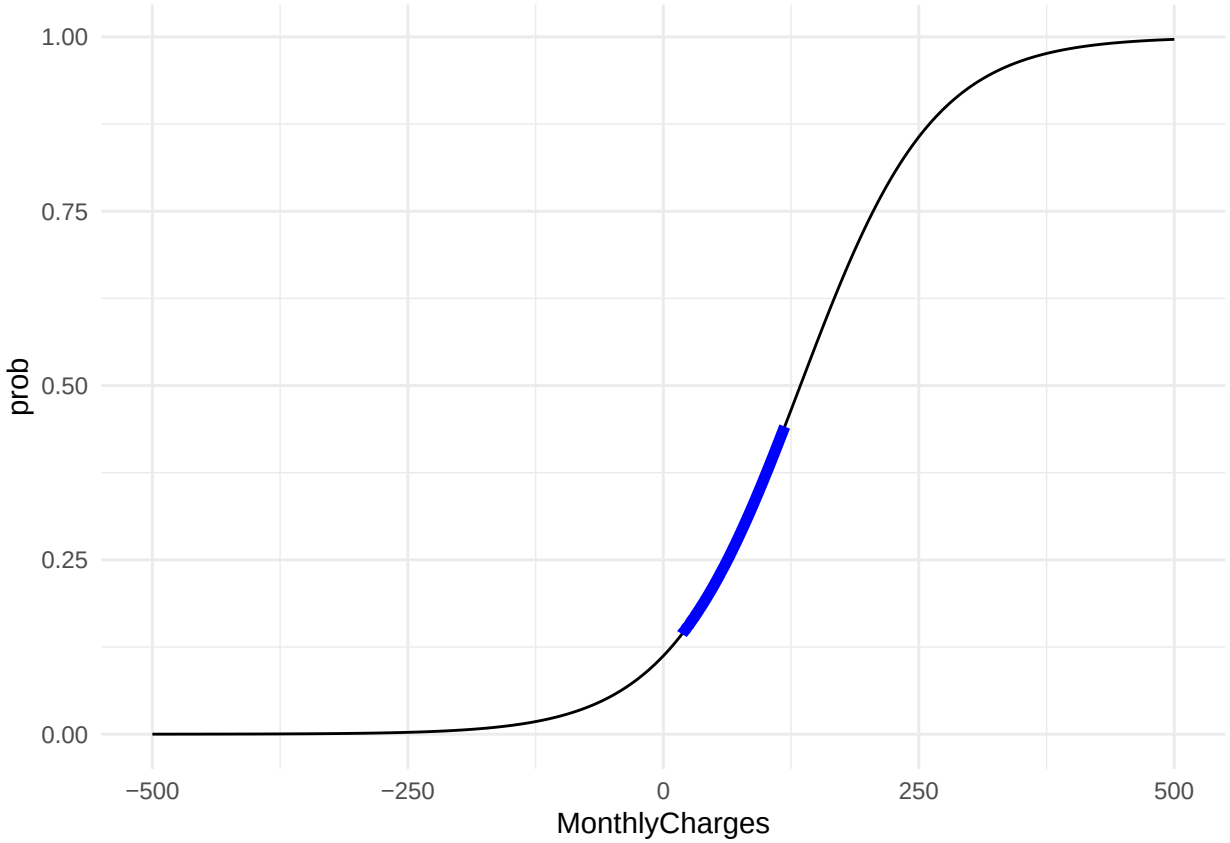
Prvo ćemo direktno sprovesti logističku regresiju, vidjeti rezultate, a potom postepeno objasniti svaki njen dio.

```
telco %>% mutate(Churn = ifelse(Churn == "Yes",1,0)) -> telco
#Kreiramo model
glm(Churn ~ MonthlyCharges, family = binomial, data = telco) -> model
#predvidjene vjerovatnoce
fitted(model) -> fitted
#graficki prikaz
ggplot(telco, aes(x = MonthlyCharges, y = fitted(model))) +
  geom_line() +
  labs(title = "Logistic Regression Fitted Values",
       x = "MontlyCharges",
       y = "Predicted Probability") +
  theme_minimal()
```



Razlog zasto vjerovatnoće nemaju oblik slova 'S' je taj sto nemamo dovoljno varijabiliteta u `MonthlyCharges`, pa navedeni grafik mozemo posmatrati samo kao dio logističke funkcije za vrijednosti `MonthlyCharges` u uzorku. Sa većim varijabilitetom, grafik bi izgledao ovako, pri čemu ja plavom bojom označen dio koji se odnosi na naš uzorak.

```
tibble(MonthlyCharges = seq(from = -500, to = 500, length.out = 7043)) -> test
predict(model, test, type = "response") -> prob
ggplot(tibble(test, prob), aes(MonthlyCharges, prob)) +
  geom_line() +
  geom_line(data = telco, aes(x = MonthlyCharges, y = fitted(model)), color = "blue", size = 2) +
  theme_minimal()
```



Generalised Linear Models

Glm funkcija kojom smo sproveli logističku regresiju na našem dataset-u skraćenica je za Generalised Linear Models. Svaki GLM model sastoji se od tri komponente:

1. Slučajne komponente koja se odnosi na raspored zavisne varijable;
2. Sistemske komponente koja predstavlja linearnu kombinaciju nezavisnih varijabli;
3. Link funkcije koja predstavlja vezu između zavisne varijable i linearne kombinacije nezavisnih varijabli.

Sigmoidna linija u našem primjeru predstavlja modeliranu vjerovatnoću odliva klijenata. Ako se prisjetimo logističke funkcije sa početka, možemo je predstaviti sledećom jednačinom:

$$\hat{p} = \frac{1}{1 + e^{-(b_0 + b_1 * x)}}$$

Izraz

$$b_0 + b_1 * x$$

predstavlja linearnu kombinaciju nezavisne varijable, tj. sistemsku komponentu. Prikazana logistička funkcija je veza nezavisne i zavisne varijable (vjerovatnoće), tj. link funkcija, dok zavisna varijabla ima binominalan raspored vjerovatnoća. Kao što vidimo, logistička regresija je vrsta Gneralised Linear modela.

Maximum Likelihood

Pitanje koje se postavlja je kako odrediti koeficijente b_0 i b_1 na optimalan način, tako da za svaku vrijednost x-a (`MonthlyCharges`) dobijemo vjerovatnoću odliva klijenta?

Razmotrimo sledeći izraz:

$y_{actual} * y_{pred} + (1 - y_{actual}) * (1 - y_{pred})$

gdje `y_actual` predstavlja stvarnu vrijednost odliva klijenta (1 ili 0), a `y_pred` predviđenu vrijednost (vjerovatnoca od 0 do 1). U situaciji kada se klijent odlio, `y_actual` = 1, pa izraz ima vrijednost:

$1 * y_{pred} + 0 * (1 - y_{pred}) = y_{pred}$

Maksimum ovog izraza je 1, ako smo predvidjeli vjerovatnoću odliva onog klijenta koji se odlio od 100%, tj. kada smo tačno predvidjeli.

U situaciji kada se klijent nije odlio, y_actual = 0, pa izraz ima vrijednost:

$0 * y_{pred} + 1 * (1 - y_{pred}) = 1 - y_{pred}$

Maksimum ovog izraza je 1, ako smo predvidjeli vjerovatnoću odliva onog klijenta koji se odlio od 0%, tj. kada smo tačno predvidjeli.

Na osnovu navedenog, vidimo da nam je cilj da maksimizirmo izraz sa početka, pa želimo da nađemo one vrijednosti b_0 i b_1 koje ga maksimiziraju. Uzimimo dvije proizvoljne vrijednosti b_0 i b_1:

```
b0 <- -1
b1 <- 0.1
```

Za ove vrijednosti inicijalnih koeficijenata, za svaku vrijednost x-a (`MonthlyChargres`) dobijamo vjerovatnoću odliva klijenta, unoseći podatke u formulu:

$$y_{pred} = \frac{1}{1 + e^{-(b_0 + b_1 * x)}}$$

```
b0 <- -1
b1 <- 0.1

y_pred <- 1/(1+exp(-(b0+b1*telco$MonthlyCharges)))
as.data.frame(cbind(y_actual = telco$Churn, y_pred)) -> data
data %>% mutate(izraz = y_actual*y_pred + (1-y_actual)*(1-y_pred)) %>% summarise(sum(izraz))
```

```
##   sum(izraz)
## 1   2212.389
```

Vrijednost navedenog 'likelihood' izraza za inicijalne vrijednost koeficijenata b_0 i b_1 je 2212.389. Zbog komputaciono- praktičnih razloga, izraz koji maksimiziramo je log likelihood.

$y_{actual} * log(y_{pred}) + (1 - y_{actual}) * log(1 - y_p red)$ - log-likelihood izraz

Napisaćemo funkciju koja za sve date vrijednosti koeficijenata računa log-likelihood vrijednost, i potom optimizovati tako da dobijemo koeficijente koji nam daju maksimalnu vrijednost fukcije. Kako po default-u optimizacija uzima minimalnu vrijednost funkcije, računaćemo minimum negtivne vrijednosti log-likelihood izraza.

```r
log_likelihood <- function(coeffs) {
  intercept <- coeffs[1]
  slope <- coeffs[2]
  y_pred <- 1/(1+exp(-(intercept+slope*telco$MonthlyCharges)))
  log_likelihoods <- log(y_pred)*telco$Churn + log(1-y_pred)*(1-telco$Churn)
    -sum(log_likelihoods)
}
```

Vrijednost log-likelihood izraza za početne koeficijente:

```r
log_likelihood(c(b0,b1))
```

```
## [1] 27028.59
```

Nalazimo koeficijente koje maksimiziraju log-likelihood izraz:

```r
optim(
  par = c(intercept = -1,slope = 0.1),
  fn = log_likelihood
)$par
```

```
##   intercept       slope
## -2.06869510  0.01543706
```

Ovim smo izračunali koeficijente logističke regresije putem Maximum Likelihood metode.

Sada ćemo vidjeti rezultate naše modela putem `glm` funkcije.

```r
summary(model)
```

```
##
## Call:
## glm(formula = Churn ~ MonthlyCharges, family = binomial, data = telco)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -2.067053   0.074203  -27.86   <2e-16 ***
## MonthlyCharges  0.015416   0.000967   15.94   <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: 7878.2  on 7041  degrees of freedom
## AIC: 7882.2
##
## Number of Fisher Scoring iterations: 4
```

Vidimo da se radi o istim koeficijentima kao što smo postupno pokazali.

Log-odds

Logističkom funkcijom, modelirali smo vjerovatnoću odliva klijenta na osnovu iznosa varijable `MonthlyCharges`. Odnos između vjerovatnoće i nezavisne varijable je nelinearan, ali postoji pokazatelj koji je linearna kombinacija nezavisnih varijabli. Da vidimo i kojem se pokazatelju rad, krenuvši od logističke funkcije:

$$
\begin{aligned}
y_{\text{pred}} &= \frac{1}{1+e^{-(b_0+b_1x)}} \\
y_{\text{pred}}(1+e^{-(b_0+b_1x)}) &= 1 \\
1+e^{-(b_0+b_1x)} &= \frac{1}{y_{\text{pred}}} \\
e^{-(b_0+b_1x)} &= \frac{1}{y_{\text{pred}}} - 1 \\
e^{-(b_0+b_1x)} &= \frac{1-y_{\text{pred}}}{y_{\text{pred}}} \\
-(b_0+b_1x) &= \log\left(\frac{1-y_{\text{pred}}}{y_{\text{pred}}}\right) \\
b_0+b_1x &= \log\left(\frac{y_{\text{pred}}}{1-y_{\text{pred}}}\right)
\end{aligned}
$$

Izraz $log(\frac{y_{pred}}{1-y_{pred}})$ naziva se log-odds i predstavlja log vrijednost odnosa vjerovatnoće da će se klijent odliti i vjerovatnoće da neće. Prilikom sagledavanja predviđenih vrijednosti logističkom regresijom, po default-u prikazuje se log-odds:

```
#Manuelno racunanje log-odds

intercept <- coef(model)[1]
slope <- coef(model)[2]
y_pred <- 1/(1+exp(-(intercept+slope*telco$MonthlyCharges)))
log(y_pred/(1-y_pred)) -> logodds

#Uporedni prikaz
head(cbind(model = augment(model)$.fitted, logodds),5)
```

```
##           model    logodds
## [1,] -1.6068921 -1.6068921
## [2,] -1.1891244 -1.1891244
## [3,] -1.2369133 -1.2369133
## [4,] -1.4149656 -1.4149656
## [5,] -0.9771573 -0.9771573
```