Lojistik Regresyon

Erdem Korhan Akçay

Giriş

Sürekli iyileştirme programları, üretim sürecinin iyileştirilmesine büyük ölçüde odaklanmıştır. Doğrusal regresyon gibi istatistiksel yöntemler, üretim süreci performansını modellemek için rutin olarak kullanılan araçlardır. Bu araçlar iş süreçlerine bu kadar yaygın olarak uygulanmamıştır. İş süreçlerinin karakterize edilmesini biraz daha zorlaştıran bir yönü, sürecin sonuç değişkenlerinin doğası ve süreç faaliyetlerinin tanımıdır. Karmaşık iş süreçlerini karakterize etmek, yeni bir ilacı test etme sürecine veya belirli bir tedavinin bir üretim süreciyle karşılaştırmak yerine bir hastalık üzerindeki etkilerini incelemeye kıyasla daha iyi olabilir. Biyomedikal dünyada, incelenen değişkenler genellikle nitel, uzun vadeli, kolayca değiştirilemez veya manipüle edilemez ve sürecin izole edilmesi zor olabilir. Bu senaryoda, sonuç değişkeni, belirli bir ilacın veya tedavinin işe yarayıp yaramadığıdır, bu da ikilemlidir. Giriş değişkenleri oldukça özneldir ve bireyin yaşadığı stres miktarının deneyin sonucunu nasıl etkilediğini belirlemek gibi faktörleri içerebilir. Klinik çalışmalarda bir biyokimyacının karşılaştığı deneysel sorunların çoğu, karmaşık iş süreçlerinin davranışını anlamaya ve modellemeye çalışan şirketin karşılaştığı sorunlara benzer. Aslında, bir iş sürecindeki varyasyon kaynaklarını inceleme yaklaşımı, sosyal ve antropolojik bilimlerde kullanılan nitel araştırma yöntemlerine benzeyebilir.

Örnek vermek gerekirse, müşteri memnuniyetini değerlendiren bir iş süreci düşünün. Bir şirket, kilit müşterilerinin ne kadar memnun olduğunu ve şirketten tekrar satın alıp almayacaklarını bilmek ister. Buna ek olarak, şirket müşterilerin şirketten fatura alma tercihlerini bilmek ister. Şirket, faturaları elektronik olarak, posta yoluyla gönderebilir veya bu müşteri için otomatik faturalandırma ve ödemeye izin veren özel bir Web portalı kurabilir. Şirket, bu verileri kilit müşteri tabanından toplamak için bir anket aracı kullanıyor. Bu öğelere ek olarak, şirket belirli müşteri türlerinin diğerlerinden

farklı yanıt verip vermeyeceğini bilmekle ilgileniyor. Anketin bir parçası olarak, belirli özelliklerin müşterilerin davranışlarını ve satın alma tercihlerini etkileyip etkilemediğini belirlemek için müşteriler hakkında demografik bilgiler toplayacaklar. Bu yaklaşımdaki sorun, doğrusal regresyonun sonuç değişkeninin doğada sürekli olmasını gerektirmesi ve bu sonuç değişkenlerinin olmamasıdır. Şirketten tekrar satın alma seçimi ikilidir(binary). Müşterinin memnuniyet düzeyiyle ilgili yanıtı ikili veya sıralı(ordinal) olabilir ve müşterinin faturalandırma tercihi nominaldir. Bu tür yanıtları modelleyebilen analiz aracı lojistik regresyondur.

Lojistik Regresyon Modeli

Lojistik regresyon öncelikle sonuç değişkeni ikili olduğunda kullanılır; ancak, doğası gereği nominal veya sıralı olan verileri işlemek için değiştirilebilir. Önce ikili verilerin analizini ele alacağız ve daha sonra nominal ve sıralı durumları tartışacağız.

İkili yanıt değişkeni durumunda, regresyon modeli aşağıdaki formu alır;

$$y_i = \boldsymbol{x_i'}\boldsymbol{\beta} + \epsilon_i$$

Burada $x_i' = [1, x_{i1}, x_{i2}, ..., x_{ik}], \beta = [\beta_0, \beta_1, \beta_2, ..., \beta_k]]$ yanıt, y_i , yalnızca 0 veya 1 değerlerini alabilir. Yanıt, olasılık dağılımı Bernoulli olan bir rassal değişken olduğu varsayılır.

$$y_i = 1, \qquad P(y_i = 1) = \pi_i$$
$$y_i = 0, \quad P(y_i = 0) = 1 - \pi_i$$

Ve $E(\epsilon_i) = 0$ aynı zamanda yanıt değişkeninin beklenen değeri,

$$E(y_i) = \boldsymbol{x_i'}\boldsymbol{\beta} = 1(\pi_i) + 0(1 - \pi_i) = \pi_i$$

Yanıt değişkeninin beklenen değeri, $E(yi) = \boldsymbol{x_i'}\boldsymbol{\beta}$, yanıtın 1 değerini alma olasılığıdır. Bu regresyon modelindeki sorunlardan biri, hata teriminin yalnızca iki değer alabilmesidir.

Sonuç olarak, bu modeldeki hatalar normal olarak dağıtılamaz. Modelle ilgili bir başka sorun, hata varyansının sabit olmamasıdır.

$$\sigma^2 = E[y_i - E(y_i)]^2$$

$$= (1-\pi_i)^2\pi_i + (0-\pi_i)^2(1-\pi_i)$$
$$= \pi_i(1-\pi_i)$$

Bu sonuç aşağıdaki formulle aynıdır.

$$\sigma^2 = E(y_i)[1-E(y_i)]$$

Bu, hata teriminin varyansının ortalamanın bir fonksiyonu olduğunu gösterir.

$$0 \leq E(y_i) = \pi_i \leq 1$$

Doğrusal bir yanıt fonksiyonu varsayımı altında, yanıtın tahmin edilen değerlerinin 0, 1 aralığının dışında olabileceği bir modele bulmak mümkündür.

Yanıt ikili olduğunda, yanıt değişkeninin şeklinin doğrusal olmadığını gösteren kanıtlar vardır. S şeklinde bir eğri sıklıkla kullanılır. Bu fonksiyona lojistik yanıt fonksiyonu veya logit denir ve aşağıdaki forma sahiptir.

$$E(y) = \pi = \frac{e^{(g(x)}}{1+e^{(g(x))}} = \frac{1}{1+e^{-g(x)}}$$

$g(x) = \boldsymbol{x'\beta}$.'dır ve g(x) lojistik regresyon modelinin doğrusal tahmincisi olarak tanımlanır. Ve doğrusal tahmicinin tahmini $\hat{g}(x) = \boldsymbol{x'\beta}$ ve değerleri fit edilmiş lojistik regresyon modeli

$$\hat{y} = \frac{1}{1+e^{-\boldsymbol{x'\beta}}}$$
$$= \hat{\pi}$$

Lojistik Regresyon ile Müşteri Memnuniyetinin Modellenmesi

Veriler 85 adet müşteri memnuniyeti anketinden toplanmıştır. Müşterilere gelecekte şirketten tekrar satın almayı düşünüp düşünmeyecekleri sorulan bu ankette. Müşteriler, müşterinin şirkete devamlı müşteri(Repeat customer) olup olmadığı ve müşterinin şirkette harcadığı hesabın büyüklüğü olmak üzere iki demografik değişkendir. Şirket, devamlı gelen müşterilerin şirketten tekrar satın almaya devam edip etmeyeceğini bilmekle ilgileniyor. Ayrıca, müşterinin şirkette sahip olduğu hesabın büyüklüğünün, müşterinin tekrar

satın alıp almayacağını belirlemede etkili olup olmadığını bilmek istiyorlar. Veri setimizde, Yanıt değişkeni, ikili(binary) yapıda olup müşterinin tekrar satın alıp almayacağıdır. Bilgi değişkeni, ikili yapıda olup müşterinin şirkette yeni olup olmadığıdır. Hesabın boyutu süreklidir ve her müşteri tarafından harcanan dolar miktarını yansıtır ve analiz için bu değişkeni kategorik olacak şekilde küçük ve büyük gruplar halinde kodlandı. Küçük hesap, değeri 100.000 doların altındaki herhangi bir hesaptır ve büyük hesap, değeri 100.000 dolardan fazla olan herhangi bir hesaptır.

Table I. Summary of customer data

Repeat customer	Account size	Would not buy again	Would buy again	Total
Yes	Small	5	21	27
Yes	Large	4	15	19
No	Small	9	13	10
No	Large	10	11	21
Total		28	57	85

Table II. Summary of estimated probabilities from linear predictor

Repeat customer	Size of account	Probability of buying again
Yes	Small	0.803
Yes	Large	0.795
No	Small	0.532
No	Large	0.518

Lojistik regresyon, verileri analiz etmek ve müşteri demografik değişkenleri ile yanıt değişkeni arasında anlamlı ilişkiler olup olmadığını belirlemek için kullanılır. Sonuçta ortaya çıkan tahmini doğrusal tahminci(linear prediction) denklemi;

$$\hat{g}(x) = 0.075 + 1.1281(RepeatCustomer) + 0.053(AccountSize)$$

Bu denklemi çözmek için kullanılan tahminci değişkenlerin değerleri 0 ve 1'dir. Müşteri, devamlı bir müşteri değilse, bu x değeri için 0 değeri kullanılır ve denklemin ikinci terimi düşer. Müşteri devamlı bir müşteriyse, $x_1 = 1$ ve ikinci terim nihai sonuca eklenir. Müşterinin küçük bir hesabı varsa, küçük için tahminci 1'e eşitlenir. Müşterinin büyük bir hesabı varsa, hesap

büyüklüğü için x değeri 0 olarak ayarlanır ve denklem üçüncü dönem olmadan değerlendirilir.

Tablodaki sonuçlar, devamlı bir müşterinin tekrar satın alma olasılığının oldukça yüksek olduğunu göstermektedir. Devamlı müşteri olmayan müşterilerin neredeyse yarısı tekrar satın alacaktır. Ham veride, ankete katılan 85 müşteriden 57'sinin şirketten tekrar satın alacaklarını belirttiklerini gösteriyor. Devamlı müşterilerin, yeni müşterilerle karşılaştırıldığında tekrar satın alma olasılıklarının daha yüksek olduğu görülmektedir. Hesap boyutunu karşılaştırırken yanıtta bir fark varmış gibi görünmüyor. Analizdeki bir sonraki adım, bu sonuçların anlamlılığını değerlendirmektir.

Model Anlamlılığının Değerlendirilmesi

Bir model uygun olduğunda, bir sonraki adım modelin önemini değerlendirmektir. Lojistik regresyonda D yani modelin sapması, uyumun iyiliğini değerlendirmek için kullanılan bir yöntemdir.

$$D(\boldsymbol{b}) = 2\sum_{i=1}^{m}[y_i ln\frac{y_i}{n_i.\hat{\pi}(\boldsymbol{x_i})} + (n_i - y_i)ln(\frac{n_i - y_i}{n_i(1-\hat{\pi}(\boldsymbol{x_i})}]$$

Eğer lojistik regresyon modeli doğru ve örneklem hacmi büyükse, modelin sapması χ^2_{n-p} dağılacaktır. Model sapmasındaki büyük değerler modelin doğru olmadığının göstergesidir. -Sapma, doğrusal regresyondaki karelerin artık toplamına benzer.- Bu model istatistiği için test kriteri:
Eğer;
$D \leq \chi^2_{\alpha,n-p}$, model yeterlidir.

$D > \chi^2_{\alpha,n-p}$, model yeterli değildir.

Lojistik regresyon modelinin genel anlamlılığını(overall significant) değerlendirmek için, modelin en çok log- olabilirlik fonksiyonu $lnL(b)$, sabit başarı olasılığı olan modelin en çok log-olabilirlik ile oranıdır. Sabit başarı olasılığı modeli:

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

Sabit başarı olasılığının en çok olabilirlik tahmin edicisi y/n'dir, burada y toplam başarı sayısıdır ve n toplam gözlem sayısıdır.

$$lnL = yln(y) + (n-y)ln(n-y) - nln(n)$$

Modelin genel anlamlılığı için test istatistiği:

$$G = 2[\sum_{i=1}^{m} y_i ln\hat{\pi}(\boldsymbol{x_i}) + \sum_{i=1}^{m}(n_i - y_i)ln(1-\hat{\pi}(\boldsymbol{x_i})] - [yln(y) + (n-y)ln(n-y) - nln(n)]$$

Test istatistiğinin yüksek bir değerinde en az bir β'nın 0 a eşit olmadığını gösterir. H_0 hipotezi her β_i değeri 0 a eşittir. Testin doğruluk kriterleri:

Eğer;
$G \geq \chi^2_{\alpha,r}$, H_0 hipotezi reddedilir.

$G > \chi^2_{\alpha,n-p}$, H_0 hipotezi reddedilemez.

Parametre anlamlılığını değerlendirmenin diğer bir yöntemi ise Wald testidir. Wald testi, her β_j'nin en çok olabilirlik tahminiyle, standart hatasının bir tahmininin oranıdır.

$$Z_0 = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$$