Lojistik Regresyon ile Bilgisayar Ağlarında Anomali Tespiti

İdris Budak

Baha Şen

Mehmet Zahid Yıldırım

İÇİNDEKİLER

AMAÇ

Bilgi çağının en önemli unsurları:

- **Bilginin üretilmesi**
- **Korunması**
- **Erişilmesi**

Bu Unsurlar İçin Bilgisayar ağlarında Gerekli Şartlar:

- Hız
- Güvenlik
- Süreklilik

Bu yüzden **ağdaki düzensizliklerin zamanında tespit edilip önlemlerin alınması** gerekmektedir.

Bu çalışmadaki amaç bilgisayar ağlarındaki anomali tespitinde **Binary Lojistik Regresyon** tekniğinin uygulanabilirliğini incelemektir.

Bu amaçla Saldırı Tespit Sistemleriyle ilgili çalışmalarda en sık kullanılan veri setlerinden olan **KDD Cup'99** veri seti kullanılarak bir matematiksel **model oluşturulup bu modelin uygunluğu test edilmiştir.**

Ağ Güvenliği ve Saldırı Tespit Sistemleri

- Bir güvenlik modeli, 3 ana amaca hizmet etmelidir.

* **Confidentiality(Gizlilik):** Bilginin doğrulanmış kişiler haricinde okunmaması.

* **Integrity(Bütünlük):** Bilginin yetkisi olmayanlarca modifiye edilmemesi.

* **Availability(Erişilebilirlik):** Bilgisayar kaynaklarının hizmet vermesinin engellenmesine karşı korunma.

CIA kısaltması, bilgisayar güvenliğindeki üç amacı simgeleyen kolay hatırlanabilir bir kelime. [20]

Güvenliği Sınıflandırmak:

Bilgi sistemleri güvenliğini üç ana bölüme ayırabiliriz:

- **Mantıksal güvenlik:** İletişim ağlarından gelecek tehlikeler.

- **Fiziksel güvenlik:** Bilgi sistemlerini barındıran fiziksel altyapının güvenliği. sunucu ve istemci donanımları, sistem odası.

- **Çevre güvenliği:** Bilgi sistemini barındıran bina veya kampüs alanının sınırlarında alınacak fiziksel güvenlik önlemleridir. [21]

Bilgisayar ağlarında alınabilecek 4 temel önlem:

1- Kurumsal politika ve bilinçlendirme çalışmaları:

BGYS, ISO 27001.

2- Kullanıcı Bilgisayarlarında Alınabilecek Temel Önlemler:

Antivirüs, yamalar, paylaşım, web.

3- Ağ Cihazlarında alınabilecek temel önlemler:

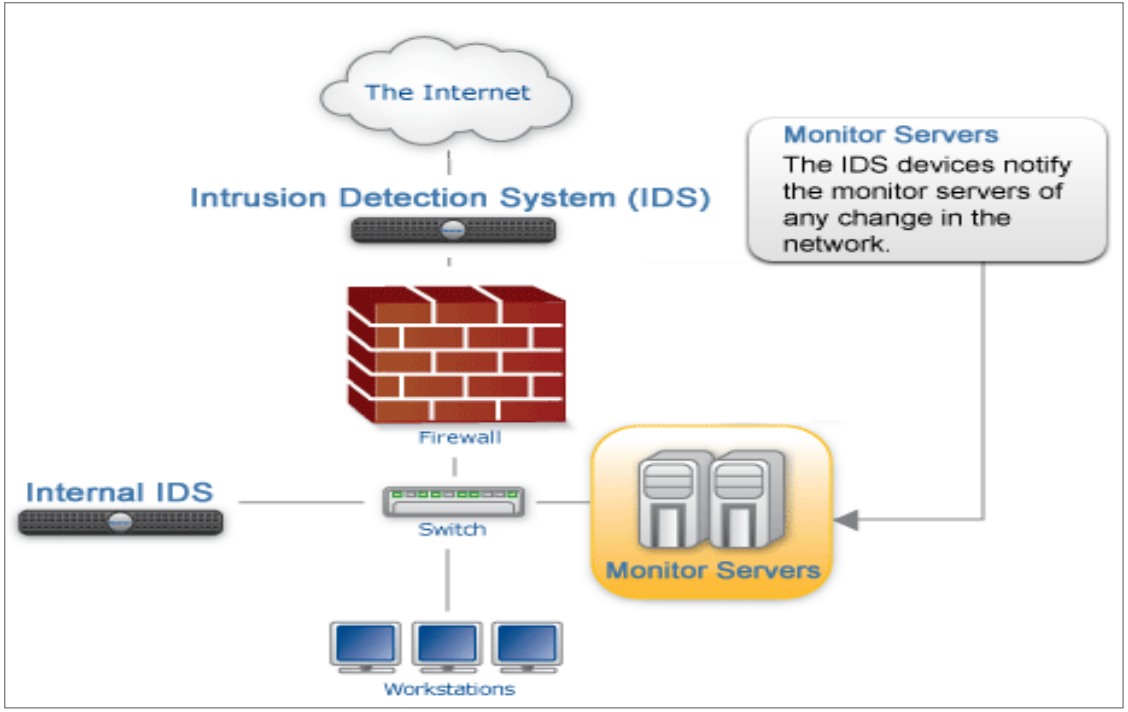
Fiziksel, Ağ Topolojisi, 802.1x.

4- Güvenlik Yazılımlarının kullanılması:

VPN , Firewall, Saldırı Tespit ve Engelleme Sistemleri

Saldırı Tespit Sistemleri(Intrusion Detection System)

- Ağ hizmetleri, **altyapı sorunları** veya **ağı kötüye kullanma** yada **ağ üzerinde düzensizlik** yaratılarak kesintiye uğratılabilir.
- Bunların önüne geçmek için **ağ trafiğini izlemek ve analiz etmek** önem kazanmaktadır.
- **STS'ler; Saldırıyı durdurma girişiminde bulunmayan** ve olası güvenlik ihlali durumlarında, ilgili yazılım veya sistem güvenlik çalışanlarına **uyarı** mesajı (alarm) veren sistemlerdir.
- **Bir STS**, olası **güvenlik açıklarını belirleyebilmek için** bilgisayar veya ağ içerisinde değişik alanlardan **bilgileri toplar ve analiz** eder.
- Güvenlik duvarının **statik** izleme kabiliyetini tamamlayan **dinamik** izleme elemanıdır.


The Internet
Intrusion Detection System (IDS)
Monitor Servers
The IDS devices notify the monitor servers of any change in the network.
Firewall
Internal IDS
Switch
Monitor Servers
Workstations

[33]

STS'lerin geliştirilmesinde istatistiksel yöntemlerin dışında:

- kural tabanlı (rule based),
- **eşik değeri belirleme (threshold value),**
- **durum geçiş diyagramları (state transition diagrams),**
- yapay sinir ağları (artificial neural networks),
- **veri madenciliği (data mining),**
- yapay bağışıklık sistemi (artificial immune system),
- **uzman sistemler,**
- **örüntü eşleme,**
- bulanık mantık (fuzzy logic)

Saldırı Tespit Yöntemi

- STS'lerde, saldırı tespit yöntemi olarak **anormallik tespiti** ve **kötüye kullanım tespiti** olmak üzere iki farklı yaklaşım kullanılır.
- **Anormallik tespitine** dayanan yaklaşım, sistemdeki **kullanıcı davranışlarını** modellerken, **kötüye kullanım (imza) tespitine** dayanan yaklaşım, **saldırganların davranışlarını** modeller.
- **Anormallik** tespitinde, bütün kötü davranışlar tespit edilmeye çalışılır.
- **Kötüye kullanım** tespiti, yöntemi kötü olarak bilinen davranışları tanımaya çalışır.
- Her iki yöntemin avantajlarını bir araya toplayan **hibrit yaklaşımlardan** faydalanmak daha uygun sonuçlar vermektedir.

İstatistik Yöntemlerle Anomali Tespiti

Ağ trafiğinde oluşan **düzensizliği** inceleyen çalışmaların çoğunda **istatistiki yöntemler ve trafik örnekleme (sampling)** kullanılarak sonuç elde edilmeye çalışılmıştır. İstatistik yöntemler genel olarak iki başlık altında toplanmıştır;

A) En Çok ve Baz Değer. **B) İz Eşleştirme.**

A. En Çok ve Baz Değer

- **Baz değer** , geçmişte kaydedilen ağ trafiğine göre oluşturulmuş **şablondur**. Bu şablon normal ağ trafiğini temsil eder. Ağ trafiğinde meydana gelen ve baz değerden farklı olan her türlü trafik düzensiz olarak kabul edilir.
- **En çok yöntemi**nde, meydana gelen düzensizlikleri tespit ederken iki ayrı yöntem izlenebilir:

En çok oturum: belirli bir zaman dilimindeki oturum sayısı (session).

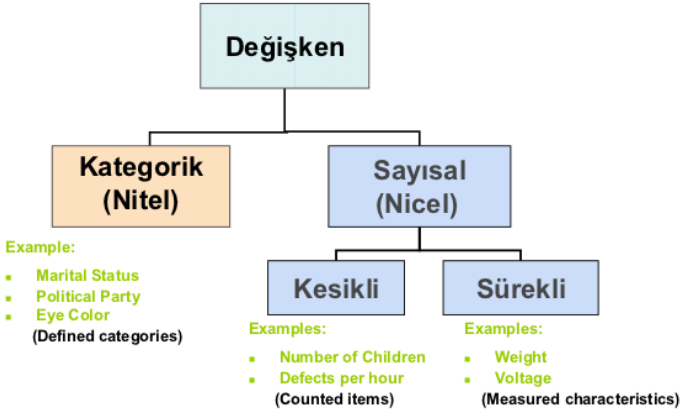
En çok veri: belirli bir zaman diliminde aktarılan veri miktarı (byte).

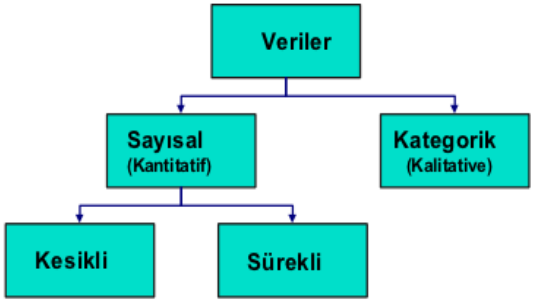
B. İz Eşleştirme

Ağda **bilinen servis ve ip bloklarının akış verileri** bir veritabanına kaydedilir. Veri tabanında **bulunmayan** port veya ip adreslerine gelen veya giden trafik **şüpheli** olarak değerlendirilebilir.

Örneğin sadece 80 portundan hizmet veren bir sunucunun 3306 portuna bir istek gönderildiğinde uyarı sistemi devreye girip şüpheli durumu bildirebilir. Bu yöntem ağ tarama tespitinde kolayca kullanılabilir.

VERİ VEYA DEĞİŞKEN ÇEŞİTLERİ





Medeni hali , Göz rengi: Kategorik
Hisse senedine yatırım yapar mısınız?: Kategorik[34]

Sınıftaki çocuk sayısı: Kesik
Boy uzunluğu: Sürekli [36]

Regresyon Analizi

– **Regresyon analizi**: değişkenler arasındaki **ilişkinin varlığı**, eğer **ilişki var ise bunun gücü**nü belirlemektir.

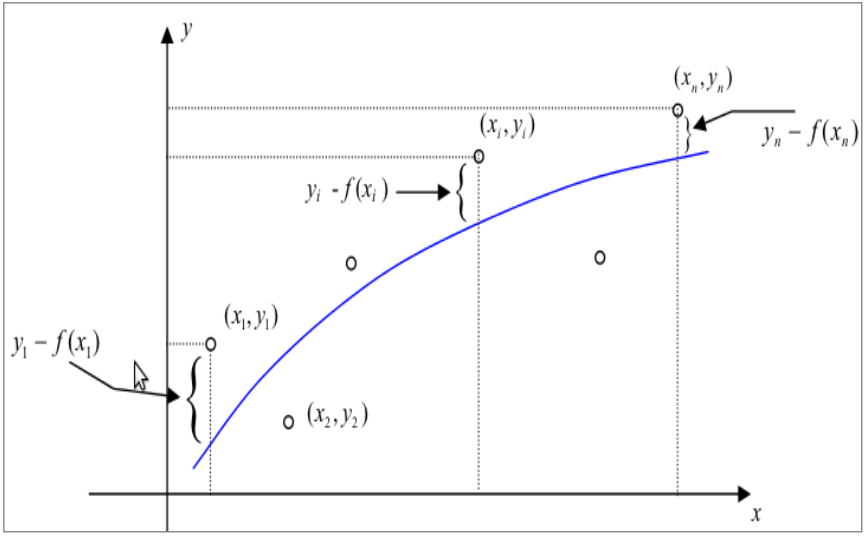
– Değişkenlerden **birinin değeri** bilindiğinde diğeri hakkında **kestirim** yapılmasını sağlar.

– Örneğin, bir **ziraatçı buğday verimi ve gübre miktarı** arasındaki ilişkiyi, bir **mühendis**, **basınç ve sıcaklık**, bir **ekonomist gelir düzeyi ve tüketim harcamaları**, bir **eğitimci öğrencilerin devamsızlık gösterdiği gün sayıları ve başarı dereceleri** arasındaki ilişkiyi bilmek isteyebilir.

– Değişkenlerden **biri bağımlı diğerleri bağımsız** değişken olmalıdır.

– Değişkenler arasındaki ilişkiyi açıklamak için kullanılan matematiksel modele **regresyon modeli** denir.

Regresyon analizinde en çok sık yöntemlerden biri **en küçük kareler**.



Tanım: $y_i - f(x_i)$ farklarından her birine bir **artık** denir.

Tanım: Bir veri tablosuna en iyi uyan doğrusal fonksiyonun grafiği olan doğruya **regresyon doğrusu** veya **en küçük kareler doğrusu** denir.

$$\sum_{i=1}^{n} (y_i - f(x_i))^2 = (y_1 - f(x_1))^2 + \cdots + (y_n - f(x_n))^2$$

Formülümüzü biraz daha açıklayacak olursak:

Gerçek değer(y) ile teorik değer(y_t) arasındaki fark Δy ile gösterilirse,

$$\Delta y = y - y_t \qquad (1)$$

olur. N tane deneysel nokta için,

$$\sum_{i=1}^{N} (\Delta y)^2 = \sum_{i=1}^{N} (y - y_t)^2 \rightarrow \text{Minumum} \qquad (2)$$

olmalıdır. Burada:

$$y_t = a + bx \qquad (3)$$

dir. Eşitlik-2 ile verilen ifadenin minumum olma şartını sağlayan a ve b sabitleri, matematiksel kural gereğince aynı eşitliğin a ve b' ye göre türevlerinin sıfıra eşitlenmesi ile bulunur. Yani;

$$\frac{\partial}{\partial a}\left(\sum_{i=1}^{N}(y-y_t)\right)^2=0 \qquad (4.a)$$

$$\frac{\partial}{\partial b}\left(\sum_{i=1}^{N}(y-y_t)\right)^2=0 \qquad (4.b)$$

$$\frac{\partial}{\partial a}\left(\sum_{i=1}^{N}\left[y-(a+bx)\right]\right)^2=2Na-2\sum_{i=1}^{N}y+2b\sum_{i=1}^{N}x=0 \qquad (5.a)$$

$$\frac{\partial}{\partial b}\left(\sum_{i=1}^{N}\left[y-(a+bx)\right]\right)^2=2b\sum_{i=1}^{N}x^2-2\sum_{i=1}^{N}xy+2a\sum_{i=1}^{N}x=0 \qquad (5.b)$$

bulunur. Bu eşitliklerden a ve b çekilerek **regresyon doğrusu** bulunur. [15]

LOJİSTİK REGRESYON

Lojistik regresyonda da bazı değişken değerlerine dayanarak **kestirim** yapılmaya çalılır, ancak **iki yöntem arasında 3 önemli fark vardır**:

1- Doğrusal regresyon analizinde tahmin edilecek olan **bağımlı değişken sürekli iken**, lojistik regresyonda bağmlı değişken **kategoriktir ve kesikli** bir değer olmalıdır.

2- Doğrusal regresyon analizinde **bağımlı değişkenin değeri**, lojistik regresyonda ise **bağımlı değişkenin alabileceği değerlerden birinin gerçekleşme olasılığı** kestirilir.

3- Doğrusal regresyon analizinde bağımsız değişkenlerin **çoklu normal dağılım** göstermesi koşulu aranırken, lojistik regresyonun uygulanabilmesi için bağımsız değişkenlerin dağılımına ilişkin **hiçbir ön koşul** yoktur.

Logistic Regresyon genel olarak üçe ayrılır:

1- İkili (Binary) lojistik regresyon:

Bağımlı değişken iki düzeyli olduğunda kullanılır(Var-Yok, Evet-Hayır).

2- Sıralı (Ordinal) lojistik regresyon:

Bağımlı değişken sıralı nitel veri tipinde (hafif-orta-şiddetli vb.) olduğunda kullanılır.

3- Multinomial lojistik regresyon:

Bağımlı değişken ikiden çok düzeyli sıralı olmayan nitel veri tipinde olduğunda kullanılır.

DEĞİŞKENLER	DOĞRUSAL REGRESYON ANALİZİ	LOJİSTİK REGRESYON ANALİZİ
BAĞIMLI	SÜREKLİ SAYISAL KESİKLİ SAYISAL	NİTELİK
BAĞIMSIZ	SÜREKLİ SAYISAL KESİKLİ SAYISAL	SÜREKLİ SAYISAL KESİKLİ SAYISAL NİTELİK (Her bağımsız değişken başka bir ölçüm biçimine de sahip olabilir)

Nitelik bağımlı değişken:

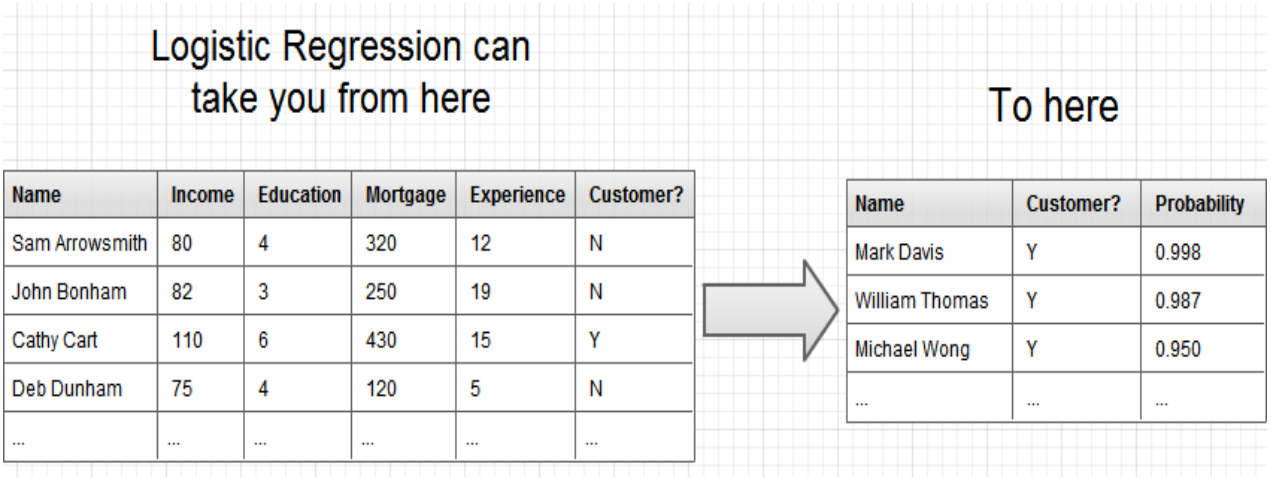
[40]

2 Kategorili olabilir (Binominal)	**: İyileşti-iyileşmedi, yaşıyor-öldü, etkili- etkisiz gibi.**
2+ Kategorili sırasız olabilir: (Multinominal)	Çalışıyor, çalışmıyor, emekli gibi
2+ Kategorili sıralı olabilir : (Ordinal)	Çok etkili-orta derecede etkili-etkisiz gibi
Her durumda lojistik regresyon analizi uygulanabilir.	

Çizelge 2.4.1 Veri tipi, gözlemlerin bağımsızlığı, grup sayısı ve testlerin ön şartları dikkate alınarak uygun analiz yöntemlerinin belirlenmesi (Çamdeviren ve ark. 2003)

Bağımlı Değişken Sayısı ve Tipi	Bağımsız Değişken Sayısı ve Tipi	Amaç	Kullanılabilecek Testin Adı
1 Tane Bağımlı Değişken ve Kategorik	1 tane bağımsız ve sürekli değişken (normal dağılımlı)	Sınıflandırma	Diskriminant Analizi veya Lojistik Regresyon Analizi
	1 tane bağımsız ve kategorik değişken	İlişki ve risk belirlemek ve sınıflandırma	Pearson Ki-Kare, Fisher Exact, Log.Likelihood, Risk İstatistikleri
	1' den fazla bağımsız ve kategorik değişken	İlişki ve risk miktarlarını belirlemek ve sınıflandırma	Lojistik Regresyon Analizi, Log.Lineer Modeller, Çok Yönlü Ki-Kare Tabloları
	1' den fazla bağımsız ve hepsi kategorik değişken	İlişki belirleme ve sınıflandırma	Lojistik Regresyon Analizi veya Diskriminant Analizi
	1' den fazla bağımsız ve hepsi sürekli değişken (normal dağılımlı)	İlişki ve risk miktarlarını belirlemek ve sınıflandırma	Lojistik Regresyon Analizi

[59]



Name	Income	Education	Mortgage	Experience	Customer?
Sam Arrowsmith	80	4	320	12	N
John Bonham	82	3	250	19	N
Cathy Cart	110	6	430	15	Y
Deb Dunham	75	4	120	5	N
...	...	...	...	...	...

Name	Customer?	Probability
Mark Davis	Y	0.998
William Thomas	Y	0.987
Michael Wong	Y	0.950
...	...	...

[55]

Bazı **nicel değişkenler** de regresyon modellerinde **nitel olarak kullanılabilmektedir**.

Örneğin **öğrenim düzeyi yıl bazında nicel** olarak ele alınabileceği gibi, ilköğretim, lise, üniversite ve üniversite üstü olmak üzere **dört şıklı bir nitel değişken** olarak da ele alınabilir.

Odds: Görülme olasılığının “p”, görülmeme olasılığına “1-p” oranıdır.
Odds ratio (OR): İki odds’un birbirine oranıdır.
Lojit: Odds ratio’nun doğal logaritmasıdır.

Risk	Hastalık		Toplam
	Var	Yok	
Var	35	16	51
Yok	25	61	86
Toplam	60	77	137

Riskli olanlarda hastalığa yakalanma odds’u: 35/16= 2.18,

Risksiz olanlarda hastalığa yakalanma odds’u: 25/61= 0.41’dir.

Bu iki odds’un birbirine oranı odds ratio’yu verir:
Odds ratio=2.18/ 0.41 = 5.3

Yorum: Risk altında olanların hastalığa yakalanma riski, risk altında olmayanlara göre **5.3 kat** daha fazladır.
[40]

Lojistik Regresyonda Dikkat Edilmesi Gerekenler:

• **Uygun** Tüm Bağımsız Değişkenler **Modele Dahil** Edilmelidir.

• **Uygun Olmayan** Tüm Bağımsız Değişkenler **Dışlanmalıdır**.

• **Aynı** birey üzerinde **bir kez gözlem** yapılmalı, tekrarlayan ölçümler olmamalıdır.

• Bağımsız Değişkenlerde **Ölçüm Hatası Küçük** olmalıdır.

• Bağımsız Değişkenler Arasında **Çoklu Bağlantı** (Multicollinearity) olmamalıdır.

• **Aşırı Değerler** olmamalıdır.

• **Örneklem Büyüklüğü** yeterli olmalıdır.

• **Beklenen** ve **Gözlenen Varyanslar** arasındaki fark büyük olmamalı:

Lojistik regresyon analizinde değişken seçimi:

- Bir regresyon eşitliğine girecek **değişken** sayısı **ne kadar çok** olursa, eşitlik **o kadar küçük hata** taşımaktadır.

- Bağımlı değişkeni **açıklayamayan** değişkenlerin denklemde tutulması lojistik regresyon denkleminin **etkinliğini ve tahmin gücünü düşürmektedir.**

Denklemde önemli etkide bulunmayan **bağımsız değişkenleri elemek** için en sık kullanılan istatistik yöntemleri:

"ileri doğru seçim" (**forward selection**), "geriye doğru eleme" (**backward elemination**) ve "tüm olası regresyon yaklaşımı" (**all possible regression**) gibi değişik yaklaşımları mevcut olan **adımsal regresyon (stepwise)**, ve en iyi regresyon modeli bulma (**best regression**) yöntemleridir.

LOJİSTİK REGRESYON MODELLERİ

İki değerli (kesikli) bağımlı değişkenleri açıklamada en çok:

Log-lineer, Logit , Probit ve Tobit Modeller kullanılır.

Bu modellerde standart regresyondan farklı olarak: Sıradan **En Küçük Kareler** tahmini yerine **Maksimum Benzerlik** (En Çok Olabilirlik) tahmini kullanılır.

* Log-lineer analizde bağımlı değişken y'nin logaritması
* Logit analizde bahis oranının (odds ratio) doğal logaritması
* Probitte ise standart normal birikimli dağılım fonksiyonunun tersi

- Lojistik regresyon analizi sonucunda elde edilen modelin uygun olup olmadığı "**model ki-kare**" testi ile,
- Her bir bağımsız değişkenin modelde varlığının anlamlı olup olmadığı ise **Wald istatistiği** ile test edilir.

Logit Model

Logit: odd değerinin doğal logaritmasıdır. Yani π olasılığı göstermek üzere, logit;

$$\log it\left(\pi(x)\right) = \log\left(\frac{\pi(x)}{1-\pi(x)}\right) \qquad (2.2)$$

Logit, **kendi parametrelerinde doğrusal** bir ölçüdür.

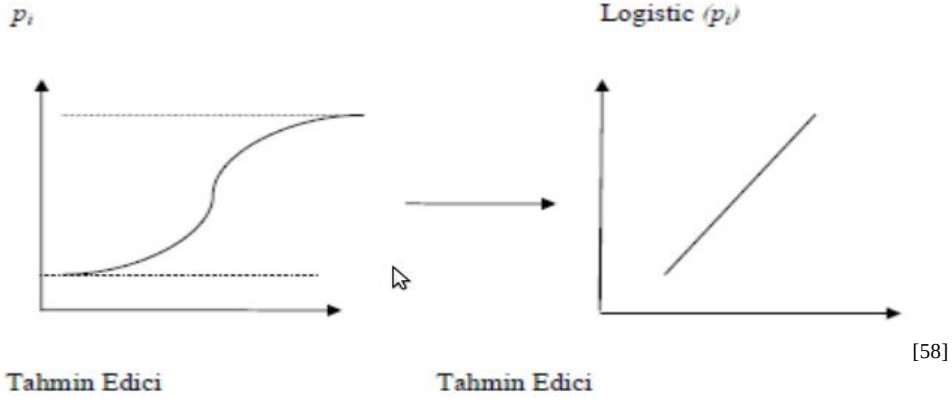
Logit model, bağımlı **değişkenin tahmini değerlerini olasılık olarak hesaplayarak** olasılık kurallarına uygun sınıflama yapma imkanı veren, tablolaştırılmış ya da ham veri setlerini analiz eden bir istatistiksel yöntemdir.

Logit model, **bağımsız değişken değeri sonsuza gittiği zaman, bağımlı değişkenin 1'e asimptot** olduğu matematiksel bir fonksiyondur.

Logit modellerinde **olasılıklar 0 ile 1 arasında** sınırlandırılmışlardır.

Olasılıklar ve tahmin edici değişken arasındaki ilişki **doğrusal değildir** ve S şeklinde bir eğridir.

Lojistik regresyon modeli olasılıklara bir dönüşüm uygulamaktadır; bu dönüşüm, tahmin edici değişkenler ile olasılıkların **doğrusal bir ilişki içerisinde sonuçlanmasını** sağlamaktadır.

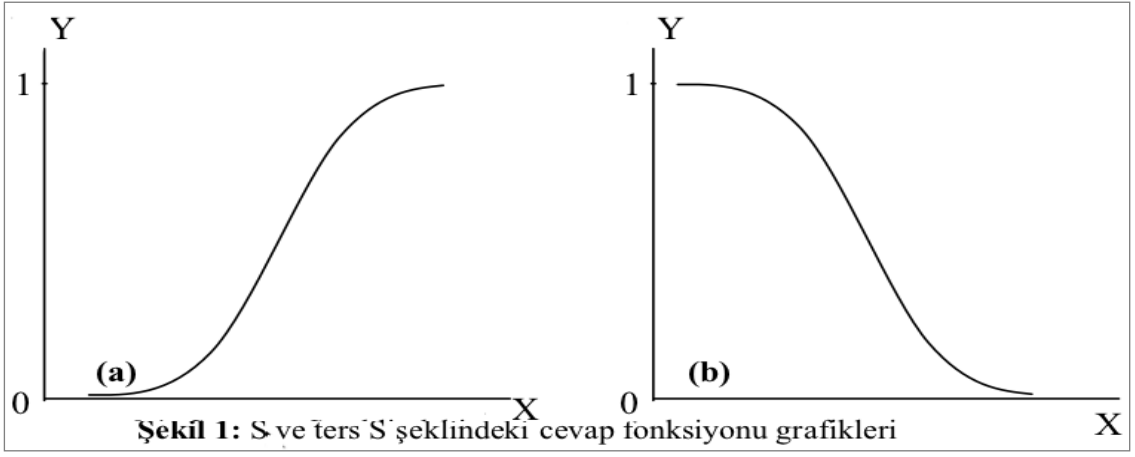


[58]

Lojistik Regresyon Formülü:

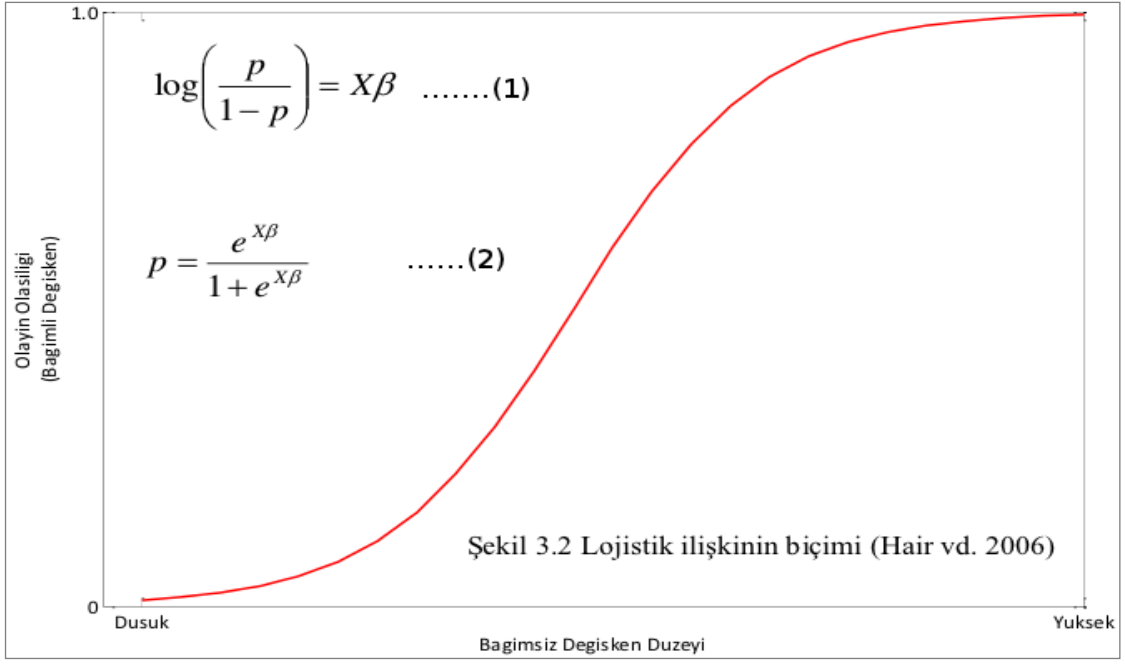
Hem **teorik** hem de **deneysel** incelemeler **bağımlı değişken iki sonuçlu** iken cevap **fonksiyonunun şeklinin S veya ters S şeklinde** olacağını göstermiştir.

Lojistik fonksiyonun **0 ile 1 arasında** bir değişim aralığına sahip olması **lojistik fonksiyonun tercih edilmesindeki ilk önemli** nedendir.[41]



Şekil 1: S ve ters S şeklindeki cevap fonksiyonu grafikleri

0-1 aralığında sınırlandırılmış ilişkiyi tanımlamak için lojistik regresyon yöntemi lojistik eğrisini kullanmaktadır(Şekil 3.2).[46]



Formülü biraz daha detaylandıracak olursak:

$$P = \frac{e^{\beta_0 + \beta_1 X_1 + \ldots\ldots + \beta_k X_k}}{1 + e^{\beta_0 + \beta_1 X_1 + \ldots\ldots + \beta_k X_k}}$$

P : İncelenen olayın gözlenme olasılığı,

e: 2.71 sayısı

β0: Bağımsız değişkenler sıfır değerini aldığında bağımlı değişkenin değerini başka bir ifadeyle sabiti,

β1 β2..... βk : Bağımsız değişkenlerin regresyon katsayılarını,

X1 X2.... Xk : Bağımsız değişkenleri,
k: Bağımsız değişken sayısını,

Bağımsız değişkenlerin en düşük düzeylerinde olasılık 0'a yaklaşıyor, fakat hiç bir zaman 0'a eşitlenmiyor.

Tersi durumda ise eğim giderek azalmaya başlıyor ve sonuçta 1'e yaklaşmasına rağmen, hiç bir zaman 1'e eşitlenmiyor.

Lojistik Regresyon Katsayı Parametre Tahmin Yöntemleri:

Doğrusal regresyonda bilinmeyen parametreleri tahmin etmek için en sık kullanılan yöntem En Küçük Kareler (**EKK**) yöntemidir.

Lojistik regresyon modelinde parametrelerin tahmininde yaygın olarak kullanılan yöntem, **en çok olabilirlik** (**Maximum Likelihood** Estimator, MLE) yöntemidir.

Parametre tahmini için "En çok olabilirlik yöntemi" haricinde:

- "**Yeniden Ağırlıklandırılmış İteratif En Küçük Kareler Yöntemi**" ,

- "**Minimum Logit Ki-Kare Yöntemi**"

ve **bunların haricinde** çok özel durumlarda kullanılan kestirim yöntemleri de bulunmaktadır.

Lojistik regresyon analizi uygulanırken izlenecek işlem sırası:

- Katsayıların en çok olabilirlik tahmin edicisi (Maximum likelihood, ML) yardımıyla tahmin edilmesi,

- Katsayıların yorumlanması,

- Katsayılara ait hipotez kontrollerinin yapılması,

- Modelin başarısının değerlendirilmesi.

Lojistik analiz için kullanılan **paket program** örnekleri:

Minitab, SPSS, SAS, Systat, NCSS ve S-Plus

Örnek:

TOTAL BYTE	PROTOKOL	PORT	ANOMALİ	NORMAL
Gönderilen veri > 150 byte ise 1 değilse 0	TCP = 1 UDP = 0	Login portu ise 1 değilse 0	Saldırı Bağlantısı Sayısı	Normal Bağlantı Sayısı
0	0	0	9	168
0	1	0	35	275
1	0	0	8	37
1	1	1	19	58

Web sunucusuna yapılan(destination ip'ler sunucunun , source'lar farklı olabilir) bağlantılar yukardaki gibiyse aşağıdaki verilerin gözlendiği bir anda saldırı olma ihtimali nedir?

Toplam oturum sayısı: 350
Protokol: TCP
PORT: 21 (FTP login)
Gönderilen veri: 200 byte

Cevap:
Denklemimizde bağımlı değişkenimiz saldırı olması. Bağımsız değişkenlerimiz:

"TOTAL BYTE" kısaca $\mathbf{X_1}$ diyelim.

"PROTOKOL" kısaca $\mathbf{X_2}$ diyelim.

"PORT" kısaca $\mathbf{X_3}$ diyelim.

Sabitimize de kısaca $\boldsymbol{\beta_0}$ diyelim.

İlk aşamada bulacağımız regresyon eşitliği aşağıdaki gibi olacaktır:

$$\mathbf{g(x) = \beta_0 + \beta_1 . X_1 + \beta_2 . X_2 + \beta_3 . X_3}$$

Şimdi ilk satırdan başlayarak denklemlerimizi yazalım:

TOTAL BYTE	PROTOKOL	PORT	ANOMALİ	NORMAL	LOGIT
Gönderilen veri > 150 byte ise 1 değilse 0	TCP = 1 UDP = 0	Login portu ise 1 değilse 0	Saldırı Bağlantı Sayısı	Normal Bağlantı Sayısı	ln(Anomali/Normal)
$X_1 = 0$	$X_2 = 0$	$X_3 = 0$	9	168	Ln(9/168) = -2.9267
$X_1 = 0$	$X_2 = 1$	$X_3 = 0$	35	275	Ln(35/275) = -2.0614
$X_1 = 1$	$X_2 = 0$	$X_3 = 0$	8	37	Ln(8/37) = -1.5315
$X_1 = 1$	$X_2 = 1$	$X_3 = 1$	19	58	Ln(19/58) = -1.1160

Yukardaki tablodan ilk satırdan başlayarak her satır için 1 denklem olmak üzere aşağıdaki denklemleri elde ederiz:

$$\beta_0 + \beta_1.0 + \beta_2.0 + \beta_3.0 = -2.9267$$
$$\beta_0 + \beta_1.0 + \beta_2.1 + \beta_3.0 = -2.0614$$
$$\beta_0 + \beta_1.1 + \beta_2.0 + \beta_3.0 = -1.5315$$
$$\beta_0 + \beta_1.1 + \beta_2.1 + \beta_3.1 = -1.1160$$

Yukardaki denklemleri ortak olarak çözdüğümüzde katsayıları aşağıdaki gibi buluruz:

$\beta_0 = -2.9267$

$\beta_1 = 1.3952$

$\beta_2 = 0.8683$

$\beta_3 = -0.4498$

Regresyon eşitliğimiz:

$g(x) = \beta_0 + \beta_1 . X_1 + \beta_2 . X_2 + \beta_3 . X_3$

$g(x) = -2.9267 + 1.3952 . X_1 + 0.8683 . X_2 + -0.4498 . X_3$

Soruda bizden olasılığı sorulan verileri aşağıdaki tabloda görebiliriz:

TOTAL BYTE	PROTOKOL	PORT
Gönderilen veri > 150 byte ise 1 değilse 0	TCP = 1 UDP = 0	Login portu = 1 değilse 0
350 > 150	TCP	21 (FTP login)
$X_1 = 1$	$X_2 = 1$	$X_3 = 1$

Bulduğumuz g(x) eşitliğinde yukardaki verileri yerine koyarsak:

g(x) = -2.9267 + 1.3952*1 + 0.8683*1 + -0.4498*1 = −1.113

İstenen olayın olasılığı (Y=1);

$$P(Y = 1 \mid X_1, X_2, ..., X_p) = \frac{e^{\beta_0+\beta_1X_1+...+\beta_pX_p}}{1+e^{\beta_0+\beta_1X_1+...+\beta_pX_p}} = \frac{1}{1+e^{-(\beta_0+\beta_1X_1+...+\beta_pX_p)}}$$

Şimdi de aşağıda olduğu gibi olasılığı hesaplayalım :

P = 1/(1 + $e^{-g(x)}$) = 1/(1 + $e^{1.113}$) = 0.2473

Yani verilen bağlantı bilgilerinin bir saldırı olma olasılığı yaklaşık olarak ¼ tür. (Bu da düşük bir ihtimali göstermektedir.)

Kullandığımız Veri Seti:

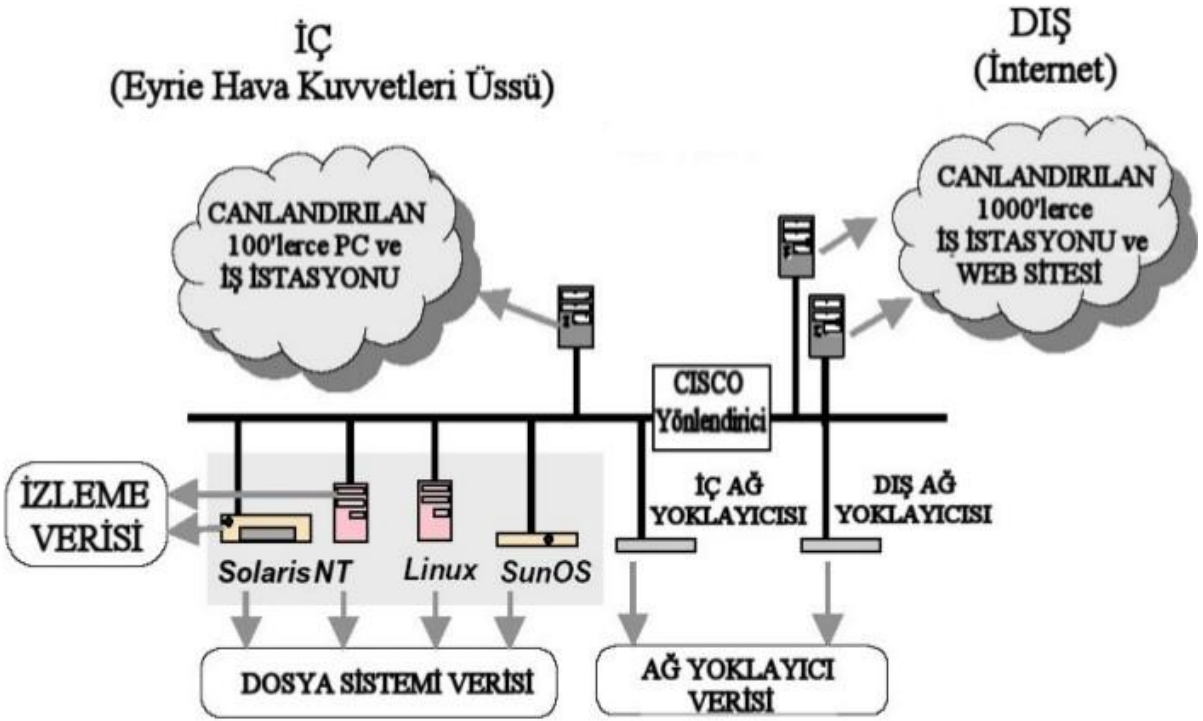
DARPA 1998 ve 1999
KDD Cup'99

Saldırının hedefi olan bir iç ağ
Saldırıyı gerçekleştiren bir dış

İç ağ içerisinde(Hava kuvvetleri) **dört "kurban" makine**
SunOS, Solaris, Linux, ve Windows NT koşmaktadır. (1998 veri setlerinde sadece UNIX makinalar kullanılmıştır.)

Trafik oluşturucular yüzlerce sunucuyu ve çeşitli uygulamaları çalıştıran **İnternet kullanıcılarını** simüle etmektedir.

2 noktadan veri toplanmıştır: **iç ve dış ağ yoklayıcısı**
Saldırı yazılımları **internetten ve hacker sitelerinden** toplanmış.

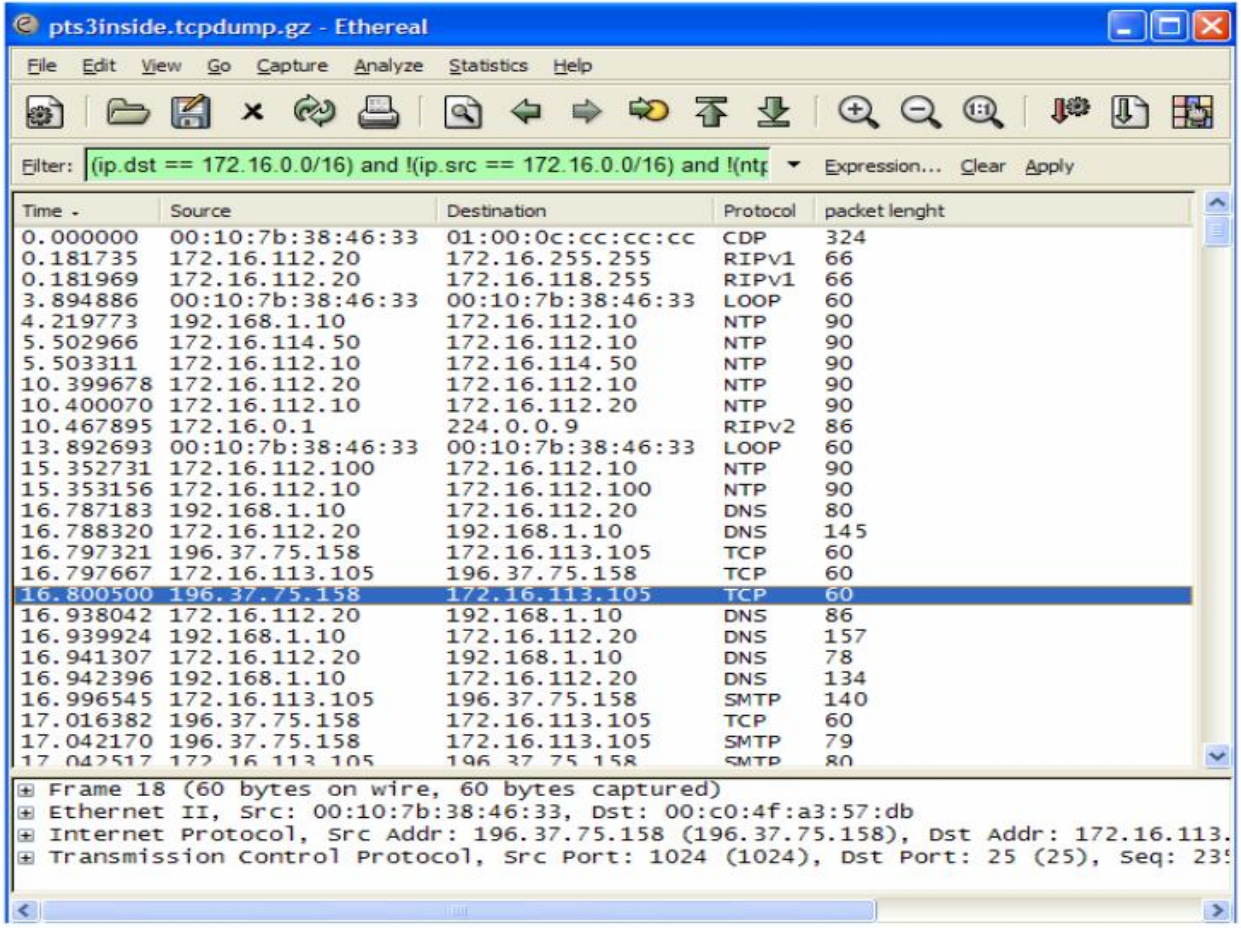

İÇ
(Eyrie Hava Kuvvetleri Üssü)
DIŞ
(İnternet)
CANLANDIRILAN
100'lerce PC ve
İŞ İSTASYONU
CANLANDIRILAN
1000'lerce
İŞ İSTASYONU ve
WEB SİTESİ
CISCO
Yönlendirici
İÇ AĞ
YOKLAYICISI
DIŞ AĞ
YOKLAYICISI
İZLEME
VERİSİ
Solaris
NT
Linux
SunOS
DOSYA SİSTEMİ VERİSİ
AĞ YOKLAYICI
VERİSİ

[68]

DARPA verileri ile çalışırken **matlab** ya da **sql** sunucularla birlikte **tcpdump** çıktılarını **wireshark**(Eski adı ethereal) programıyla da inceleyebiliriz.

Ağ dinleyicisi iki yönlü paketleri yakaladığı için kurban makinalara gelen paketler için varış ip adresi 172.16.x.x olan paketler olarak süzülmelidir. Örnek bir filtre:

(ip.dst == 172.16.0.0/16) and !(ip.src == 172.16.0.0/16) and !(ntp) and !(rip) and !(loop) and !(arp) and !(nbns) [68]


pts3inside.tcpdump.gz - Ethereal
File Edit View Go Capture Analyze Statistics Help
Filter: (ip.dst == 172.16.0.0/16) and !(ip.src == 172.16.0.0/16) and !(ntp
Expression... Clear Apply
Time Source Destination Protocol packet lenght
0.000000 00:10:7b:38:46:33 01:00:0c:cc:cc:cc CDP 324
0.181735 172.16.112.20 172.16.255.255 RIPv1 66
0.181969 172.16.112.20 172.16.118.255 RIPv1 66
3.894886 00:10:7b:38:46:33 00:10:7b:38:46:33 LOOP 60
4.219773 192.168.1.10 172.16.112.10 NTP 90
5.502966 172.16.114.50 172.16.112.10 NTP 90
5.503311 172.16.112.10 172.16.114.50 NTP 90
10.399678 172.16.112.20 172.16.112.10 NTP 90
10.400070 172.16.112.10 172.16.112.20 NTP 90
10.467895 172.16.0.1 224.0.0.9 RIPv2 86
13.892693 00:10:7b:38:46:33 00:10:7b:38:46:33 LOOP 60
15.352731 172.16.112.100 172.16.112.10 NTP 90
15.353156 172.16.112.10 172.16.112.100 NTP 90
16.787183 192.168.1.10 172.16.112.20 DNS 80
16.788320 172.16.112.20 192.168.1.10 DNS 145
16.797321 196.37.75.158 172.16.113.105 TCP 60
16.797667 172.16.113.105 196.37.75.158 TCP 60
16.800500 196.37.75.158 172.16.113.105 TCP 60
16.938042 172.16.112.20 192.168.1.10 DNS 86
16.939924 192.168.1.10 172.16.112.20 DNS 157
16.941307 172.16.112.20 192.168.1.10 DNS 78
16.942396 192.168.1.10 172.16.112.20 DNS 134
16.996545 172.16.113.105 196.37.75.158 SMTP 140
17.016382 196.37.75.158 172.16.113.105 TCP 60
17.042170 196.37.75.158 172.16.113.105 SMTP 79
Frame 18 (60 bytes on wire, 60 bytes captured)
Ethernet II, Src: 00:10:7b:38:46:33, Dst: 00:c0:4f:a3:57:db
Internet Protocol, Src Addr: 196.37.75.158 (196.37.75.158), Dst Addr: 172.16.113.
Transmission Control Protocol, Src Port: 1024 (1024), Dst Port: 25 (25), Seq: 23

Veri Setinin Hazırlanması:

"**kddcup.data_10_ percent_corrected**" dosya ismi ile internette.

kddCupp-99 veri seti'nin 10%

75Mb

500bin kayıt.

İlk 250bin kaydı **model** için **kalanı ise test** için

Veri setinde toplam **41 adet değişken** bulunmaktadır. Biz aşağıdaki prensiplere uyarak bu sayıyı 9'a indirdik:

- Paketlerin **sadece başlık bilgisine değil içeriğine** de bakılarak anlaşılacak alanlar da alınmıştır.(Örneğin bu yüzden hot, su_attempted gibi alanlar alımıştır.)

- Parametrelerin **birbirlerinden bağımsız** olanları seçilmiştir. Örneğin root_shell, su_attempted, num_root alanlarının tümü birden alınmak yerine su_attempted alanı alınmıştır.

- Parametrelerin **bağımlı değişkeni etkilemeyecek olanları** seçilmemiştir. Örneğin src_bytes ve dst_bytes alanları bu yüzden alınmamıştır.

Verileri **incelemek ve binary hale getirmek için** öncelikle verileri **sql sunucusuna alıp** aşağıdaki kurallara uygun olarak ikili hale getirdik:

protocol_type: tcp=1 ; udp veya icmp=0.
Bağlantının(connection, datasetteki her satır bir bağlantıdır.)

service: smtp, ftp, pop_3, ldap, login, imap4, auth, IRC, telnet, sql_net, exec, shell, klogin, kshell = 1 , diğerleri = 0 .
Hedefteki ağ servisini gösterir(network service on the destination).

flag: SF veya OTH = 0 ; diğerleri 1.
SF bağlantının normal bir şekilde sonlandığını, OTH ise bağlantı takip işinin bağlantının ortasında başladığını gösterir.

land: Hedef ve kaynak ip/port bilgileri aynı ise = 1; değilse 0.

wrong_fragment: sıfır ise=0; değilse=1.
Hatalı fragment sayısını gösterir.

hot: sıfırdan büyük ise 1 ; değilse 0 .
Bir bağlantıda çalıştırılan kritik komut sayısını gösterir. Örneğin sistem klasörüne girmek, programlar oluşturup çalıştırmak gibi.

num_failed_logins: sıfırsa 0 ; değilse 1 .
Yanlış login işlemleri sayısını gösterir.

su_attempted: “su root” komutu denenmişse 1 diğer durumda 0.

num_access_files: sıfırdan büyük ise 1 ; değilse 0 .
Kontrol ya da erişim izinlerini tutan kritik dosyalarda yapılan işlem sayısı.

Tablo2: Veri Seti Örnek Görüntüsü

protocol_type	service	flag	land	wrong_fragment	hot	num_failed_logins	su_attempted	num_access_files	label
1	1	0	0	1	1	1	0	1	1
1	0	0	1	0	0	0	1	1	1
0	0	1	1	1	0	0	0	1	0
0	1	1	0	0	1	1	0	0	1

Modelin Oluşturulması:

Veri setimizin SPSS binary logistic regresyon ile analiz sonucu:

Tablo3: Durum İşleme Özeti

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	250000	100,0
	Missing Cases	0	,0
	Total	250000	100,0
Unselected Cases		0	,0
Total		250000	100,0

Tablo4: Bağımlı Değişken Kodlaması

Original Value	Internal Value
0	0
1	1

İncelemek istediğimiz konu **saldırı olma durumu.**
Saldırı var = 1.
Saldırı yok = 0.
Tercihe bağlıdır, tersi de seçilebilirdi.
Sonuçların yorumunun kolay olması için genelde **asıl ilgilendiğimiz cevap için "1"** kullanmamız işimizi daha kolaylaştırır.

Tablo5: Sınıflandırma Tablosu

Observed			Predicted		
			label		Percentage Correct
			0	1	
Step	label	0	59805	11420	84
		1	736	178039	99,6
	Overall Percentage				95,1
a. The cut value is 0,5					

Tablo5'e baktığımızda Saldırı olmayan birbirinden farklı toplam 59.805+11.420=71.225 adet kayıt olduğunu ve bunların 59.805 tanesini yani 84%'ünü modelimizin doğru tahmin ettiğini görüyoruz.

Tablo6: Eşitlikteki Değişkenler

		B	S.E.	Wald	Sig.	Exp(B)
Step 1[a]	**protocol_type**	**-7,133**	,039	3,380E4	,000	,001
	service	**-1,372**	,068	401,313	,000	,254
	flag	6,882	,041	2,811E4	,000	974,877
	land	34,951	**3,446E7**	**,000**	**1,000**	1,509E15
	wrong_fragment	34,700	**6,372E6**	**,000**	**1,000**	1,175E15
	hot	6,454	,071	8,375E3	,000	635,420
	num_failed_login	3,803	,889	18,313	,000	44,854
	su_attempted	2,610	1,434	3,315	,069	13,602
	num_access_files	**-,267**	,486	,302	,583	,765
	Constant	2,843	,012	5,856E4	,000	17,175

S.E değerine göre “**land**” ve “**wrong_fragment**” stabiliteyi bozuyor.

Wald istatistiğine gore “**land**” ve “**wrong_fragment**” gereksiz.

Sig. değeri 1 olan “**land**” ve “**wrong_fragment**” anlamsız.

”**protocol_type**” en büyük mutlak katsayıyla en büyük belirleyici rolde.

” num_access_files” en küçük mutlak katsayıyla en az belirleyici rolde.

Buna göre nihai modelmiz aşağıdaki gibi olacaktır:

Regresyon eşitliği aşağıdaki gibi olmak üzere:
$\mathbf{g(x) = \beta_0 + \beta_1 . X_1 + \beta_2 . X_2 + \ldots + \beta_k . X_k}$

g(x) = 2,843 + **protocol_type***(-7,133) + **service***(-1,372) + **flag***6,882 + **hot***6,454 + **num_failed_logins***3,803 + **su_attempted***2,610 + **num_access_files***(-0,267)

$$\mathbf{P = 1/(1 + e^{-g(x)})}$$

5.3 Modelin Uygulaması:

Örnek kayıt:

protocol_type=tcp, **service**=telnet, **flag**=S0, **hot**=0, **num_failed_logins**=0, **su_attempted**=0, **num_access_files**=0 , **label**=neptune.

Kaydın label yani saldırı olup olmadığı ile ilgili bilgi alacağımız alanında "neptune" yazmaktadır. Yani bu bir neptune saldırısıdır.

Her parametreyi iki kategorili hale çevirip g(x) fonksiyonunda yerine koyarsak :

g(x) = 2,843 + **1***(-7,133) + **1***(-1,372) + **1***6,882 + **0***6,454 + **0***3,803 + **0***2,610 + **0***(-0,267) = 1,22

$$\mathbf{P = 1/(1 + e^{-g(x)}) = 1/(1 + e^{-1,22}) = 0.7721}$$

Tablo7: Test Verisi Sınıflandırma Tablosu **SQL sonuçları**

Gerçek		Tahminimiz		
		label		Doğruluk Yüzdesi
		0	1	
label	0	10912	15141	42
	1	11	217957	99,9
Toplam Yüzde				93,8
Tahminimiz >= 0,5 ise saldırı kabul ettik.				

Modelimizin uygunluğunun testi için yukarda görülen sql sonuçlarına ek olarak ayrıca **Model Ki-Kare** testi de yapılmış ve **model veri setinde** (land ve wrong_fragment) dahil **tüm değişkenlerin** Sig değeri 0,05'ten küçük çıkmış, **test veri setinde** ise **sadece land** değişkenimiz 0,5 çıkarak anlamsız olduğu görülmüştür.

Sonuç ve Öneriler:

Modelimizin Başarısı:

Gerçekte saldırı olan bir kayıtta **99%un üzerinde.**

Gerçekte saldırı olanmayan bir kayıtta **63% .**
(model veri setinde **84%**. **Test** veri setinde ise **42%**)

Modelimiz güvenlik seviyesi çok yüksek olması gereken ve yanlış alarmlarla (false-pozitif) uğraşacak yeterli elemanı olan, kritik öneme sahip ağ işletim merkezleri için uygun.

Bir sonraki çalışmamızda aynı veri seti ve parametreleri kullanarak yapay sinir ağları ile de bir model oluşturup iki modelin karşılaştırılmasını sağlayacağız.

TEŞEKKÜRLER

İLETİŞİM:

İdris Budak[1] , Baha Şen[2] , Mehmet Zahid Yıldırım[3]

[1] Karabük Üniversitesi, Fen Bilimleri Enstitüsü Bilgisayar Müh.

[2] Yıldırım Beyazıt Üniversitesi Mühendislik ve Doğa Bilimleri Fakültesi Bilgisayar Müh.

[3] Karabük Üniversitesi, Fen Bilimleri Enstitüsü Bilgisayar Müh. Bölümü

idrisbudak@karabuk.edu.tr , bsen@ybu.edu.tr , m.zahidyildirim@karabuk.edu.tr

KAYNAKLAR

[1] D.Gucarati (çev. Ümit Şenesen), Temel Ekonometri

[2] A.M. Legendre (1805), Nouvelles méthodeş pour la détermination des orbites des comètes. “Sur la Méthode des moindres quarrés” bir ek bölümde bulunur.

[3] C.F. Gauss (1809), Theoria Motus Corporum Coeleştium in Sectionibüs Conicis Şölem Ambientum.

[4] C.F. Gauss (1821/1823). Theoria çombinationis observationum erroribüs minimiş obnoxiae.

[5] Francis Galton (1877), "Typical laws of heredity", Nature 15, 492-495, 512-514, 532-533. (Galton burada bezelyelerle yaptığı kalıtım deneyi sonucunda reversion terimi kullanır.)

[6] Francis Galton (1885) Presidential address, Section H, Anthropology.(Burada insanlarin boylari uzerinde yaptigi arastirma sonucu icin "regression" terimi kullanir.)

[7] G. Udny Yule (1897) "On the Theory of Correlation", J. Royal Statist. Soç., 1897, p. 812-54.

[8] Karl Pearson, G.U.Yüle, Norman Blanchard, and Alice Lee (1903). "The Law of Ancestral Heredity", Biometrika

[9] R.A. Fisher (1922), "The goodness of fit of regression formulae, and the distribution of regression çoefficients", J. Royal Statist. Soç., 85, 597-612

[10] R.A. Fisher (1925),Statistical Methods för Research Workers

[11] http://www.fvcpsikiyatri.com/hizmetlerimiz-regresyon-terapisi

[12] http://tr.wikipedia.org/wiki/Regresyon_analizi

[13] www.fikretgultekin.com

[14] http://www.baskent.edu.tr/~afet/dersler/genel_matematik_2/dersnotlari_listesi/DERS_%207.pdf

[15] "ATOM FİZİĞİ LABORATUVARI DENEY KLAVUZU" , Prof. Dr. Mustafa TAN, Dr. Mustafa KARADAĞ , ANKARA 2004

[16] http://en.wikipedia.org/wiki/Derivative

[17] http://www.emathzone.com/tutorials/math-results-and-formulas/basic-formulas-of-derivatives.html

[18] http://istatistikanaliz.com/regresyon_analizi.asp

[19] http://www.akademikdestek.net/info/korelasyon_regresyon.doc

[20] http://hs.com.tr/tag/ag-guvenligi-ders-notlari/

[21] "BİR KURULUŞUN BİLGİ SİSTEMİ GÜVENLİĞİ İÇİN BİR YAKLAŞIM" Hakan Tan, Prof. Dr. A. Ziya Aktaş

[22] "KURUMSAL AĞLARDA ZARARLI YAZILIMLARLA MÜCADELE YÖNTEMLERİ" Enis KARAARSLAN, Gökhan AKIN ve Hüsnü DEMİR "ULAK-CSIRT"

[23] KAMPÜS AĞ YÖNETİMİ - Ar.Gör.Enis Karaarslan Ege Üniversitesi -BİTAM Kampüs Network Yönetim Grubu

[24] http://hs.com.tr/2011/11/odevler/ag-guvenligi-ders-notlari/

[25] Özel Sanal Ağ ve Servis Kalitesi(VPN- Virtual Private Network QoS- Quality of Service) Serkan GÖNEN

[26] http://www.redbilisim.com/sayfa.aspx?id=42

[27] http://tr.wikipedia.org/wiki/Virtual_Private_Network
[28] http://www.alliancedatacom.com/how-vpn-works.asp
[29]http://www.cisco.com/en/US/products/hw/vpndevc/ps2030/products_configuration_example09186a008046f307.shtml
[30] http://cehturkiye.com/fw-dmz.png
[31] http://www.avfirewalls.com/images/FortiGate/deployment-enterprise.gif
[32] http://www.mshowto.org/fortigate-110c-firewall-nasil-kurulur-ayarlari-nasil-yapilir.html
[33] http://www.interactivesys.net/intrusion-detection.html
[34] "PARAMETRİK OLMAYAN İSTATİSTİKSEL TEKNİKLER" Prof. Dr. Ali ŞEN
[35] "BİLİM FELSEFESİ" ÖĞRETİM GÖREVLİSİ NAZAN ŞAK
[36] "TEMEL İSTATİSTİK YÖNTEMLER" Dr. Mehmet AKSARAYLI
[37] "LOJİSTİK REGRESYON ANALİZİNİN İNCELENMESİ VE DİŞ HEKİMLİĞİNDE BİR UYGULAMASI" Sibel COŞKUN , Doç.Dr.Mahmut KARTAL, Yrd.Doç.Dr.Akın COŞKUN, Yrd.Doç.Dr.Hüdaverdi BİRCAN
[38] LOJİSTİK REGRESYON ANALİZİ : ÖĞRENCİLERİN SİGARA İÇME ALIŞKANLIĞI ÜZERİNE BİR UYGULAMA "Yrd. Doç. Dr. Cengiz AKTA"
[39] Tramvay Yolcu Memnuniyetinin Lojistik Regresyon Analiziyle Ölçülmesi: Estram Örneği "Yrd. Doç. Dr. Nuray GİRGİNER" , "Bülent CANKUŞ"
[40] http://78.189.53.61/-/bs/ess/k_sumbuloglu.pdf
[41] Lojistik Regresyon Analizi: Tıp Verileri Üzerine Bir Uygulama "Hüdaverdi Bircan"
[42] "Kuzey Kıbrıs Geni Bant Kullanıcılarının Davranışları" , "Devrim Seral" , "Bilişim Sistemleri Mühendisliği Bölümü, Uluslararası Kıbrıs Üniversitesi, Kıbrıs"
[43] "Saldırı Tespit Sistemleri Üzerine Bir İnceleme" Esra N. GÜVEN, Şeref SAĞIROĞLU
[44] "Pasif Ağ Verileri Üzerinden Düzensizlik Tespiti" Devrim SERAL, Beyhan ÇALISKAN
[45] "KURUMSAL AĞLARDA ZARARLI YAZILIMLARLA MÜCADELE YÖNTEMLERİ" Enis KARAARSLAN, Gökhan AKIN ve Hüsnü DEMİR "ULAK-CSIRT"
[46] YEMEKLİK YAĞ SEKTÖRÜNDE TÜKETİCİ DAVRANIŞLARINI ETKİLEYEN FAKTÖRLERİN ANALİZİ "Dr. Flora POLAT"
[47] ANALİTİK VERİLERİN DEĞERLENDİRİLMESİ "Prof. Dr. Mustafa DEMİR"
[48] "OGU TIP FAK. BİYOİSTATİSTİK VE BİLGİSAYAR" "DAĞILIM ÖLÇÜLERİ" Doç. Dr. K. Setenay ÖNER
[49] http://mimoza.marmara.edu.tr/~cahit/Yayin/belge/ista/index.html
[50] http://tr.wikipedia.org/wiki/Standart_hata_%28istatistik%29
[51] http://istatistik.yasar.edu.tr/wp-content/uploads/2011/10/TSTAT11.pdf
[52] www.akademikdestek.net/kutuphane/genel/geneldosyalar/arastirmalarda_kullanilan_ista_yont.doc
[53] "Doğrusal Olasılık ve Logit Modelleri ile Parametre Tahmini" "M. Emin İnal" , "Derviş Topuz" , "Okyay Uçan"
[54] "Araştırma Tasarımları ve İstatistiğe Giriş" Prof.Dr.Önder Ergönül
[55] http://www.simafore.com/blog/bid/99443/Understand-3-critical-steps-in-developing-logistic-regression-models
[56] YAPAY BAĞIMLI DEĞİŞKENLİ TAHMİN MODELLERİ VE BİR UYGULAMA TUĞBA ALTINTAŞ "YÜKSEK LİSANS TEZİ" "İSTATİSTİK ANABİLİM

DALI"
[57] MULTINOMIAL LOGIT MODELLER VE BİR UYGULAMA SEVİLAY KARAHAN "Biyoistatistik Programı" YÜKSEK LİSANS TEZİ
[58] http://www.acikders.org.tr/pluginfile.php/3496/mod_resource/content/2/Kredi_Riski.pdf
Dr. Göknur Büyükkara
[59] "ÇOKLU BAĞLANTI DURUMUNDA İKİLİ (BİNARY) LOJİSTİK REGRESYON MODELİNDE GERÇEKLEŞEN I. TİP HATA VE TESTİN GÜCÜ" "Yeliz KAŞKO" "ZOOTEKNİ ANABİLİM DALI"
[60] "İSTATİSTİKSEL UYGULAMALARDA LOJİSTİK REGRESYON ANALİZİ" Ersan ÜRÜK , YÜKSEK LİSANS TEZİ
[61] "Random effects logistic regression model for anomaly detection" , "MinSeokMok, SoYoungSohn ,YongHanJu" , Department of Information and Industrial Engineering, Yonsei University, 134Shinchon-dong, Seoul120-749, RepublicofKorea
[62] "Protocol-Based Classification for Intrusion Detection" "Kun-Ming Yu , Ming-Feng Wu, Wai-Tak Wong" Chung Hua University, Taiwan.
[63] "Scan Detection on Very Large Networks Using Logistic Regression Modeling" , "Carrie Gates, Joshua J. McNutt, Joseph B. Kadane, and Marc I. Kellner" Carnegie Mellon University, Pittsburgh, USA.
[64] "HoneyStat: Local Worm Detection Using Honeypots" "David Dagon, Xinzhou Qin, Guofei Gu, Wenke Lee, Julian Grizzard, john Levine, Henry Owen" Georgia Institue of Technology
[65] "Ali İhsan DARİGA" En Küçük Kareler Yöntemi
[66] http://dl.acm.org
[67] http://wenke.gtisc.gatech.edu/
[68] Saldırı Tespit Sistemlerinde İstatistiksel Anormallik Belirleme Kullanımı "Bahar 2005" Yük. Müh. Melike Erol