İSTANBUL TEKNİK ÜNİVERSİTESİ -FEN EDEBİYAT FAKÜLTESİ

MATEMATİK MÜHENDİSLİĞİ PROGRAMI

KARAR AĞACI VE LOJİSTİK REGRESYON ALGORİTMALARININ VERİ SETİ ÜZERİNE UYGULANMASI

BİTİRME ÖDEVİ

Elif UYGUN 090120406

Gamze KAPTAN 090120422

Teslim Tarihi: 29.05.2017

Tez Danışmanı: Doç.Dr.Atabey Kaygun

MAYIS 2017

ÖNSÖZ

Bu çalışmada; Karar Ağaçları (Decision Tree) ve Lojistik Regresyon (Logistic Regression) Algoritmaları kullanılarak örnek bir veri seti üzerinde tahminleme çalışması yapılmış ve doğruluk oranları hesaplanmıştır.

Bitirme tezinde, çalışmaların planlanması, yürütülmesi konusunda bilgi ve tecrübelerini benimle paylaşarak destek olan değerli tez danışmanım Doç. Dr. Atabey Kaygun'a ve eğitim hayatım süresince maddi ve manevi desteklerini esirgemeyen anneme, babama, abime, ve kardeşime teşekkür eder sevgilerimi sunarım.

İÇİNDEKİLER

ÖZET

Uygun E. , **Kaptan** G. Karar Ağaçları ve Lojistik Regresyon Algoritmalarının Veri Seti Üzerine Uygulanması. İstanbul Teknik Üniversitesi Fen Edebiyat Fakültesi, Matematik Mühendisliği, Bitirme Projesi. İstanbul. 2017.

Bilgisayar teknolojilerindeki gelişmeler, üretilen bilgi miktarlarında ve veri tabanı sistemlerinin hacminde artış meydana getirmiştir. Veri tabanlarında saklı tutulan, yararlı olma potansiyeline sahip verilerin keşfedilerek anlamlı örüntülerin ortaya çıkarılması, veri madenciliği kavramıyla ifade edilmektedir. Karar ağaçları ve Lojistik regresyon, sınıflandırma ve tahmin için sıkça kullanılan veri madenciliği yaklaşımlarından biridir. Bu çalışmada, bir bankadaki müşterilerin çeşitli bilgileri kullanılarak karar ağaçları ve lojistik regresyon algoritmaları ile bu bilgiler anlamlı hale getirilerek, hangi müşterilerin kampanyadaki ürünü alıp almayacağının tahminlemesi yapılmıştır.

Anahtar Kelimeler: Karar Ağaçları, Entropi, Lojistik Regresyon, Odds, R Studio

ABSTRACT

Uygun E. , **Kaptan** G. Application of Decision Trees and Logistic Regression Algorithms on Dataset.. İstanbul Technical University, Science&Letter Faculty, Mathematical Engineering, Graduation Project. İstanbul. 2017.

Developments in computer technologies caused increase in amount of information generated and volume of database systems. By discovering data kept stored in databases as ones having beneficial use potential and creation of meaningful patterns is expressed with the concept of data mining. Decision trees and Logistic Regression are one of the data mining approaches widely used for classification and forecasting. In this study, decision trees and logistic regression algorithms were used to make this information meaningful by using various information of customers in a bank to estimate which customers would buy the product in the campaign.

Key Words: Decision Trees, Entropy, Logistic Regression, Odds, R Stuido

Projenin Tanımı

Bu projede Karar Ağaçları (Decision Tree) ve Lojistik Regresyon (Logistic Regression) algoritmaları ile veri analizinin nasıl yapıldığı gösterilecektir. Uygulama kısmında; veri madenciliği çözümleri ile istatistik kökenli algoritmaları bir programlama ara yüzü altında sunan R programından faydalanılacaktır. R'ın kendisine ait özel kütüphaneleri sayesinde Decision Tree ve Logistic Regression algoritmaları veri kümesinde işlenerek veri kümesine ait analizler ortaya konulacaktır. Veri kümemiz bir bankanın veritabanından alınmış olup bankadaki müşterilere ait çeşitli bilgileri içermektedir. Veri kümesinde bulunan nitelikler (attribute) ve örnekler (instance) her iki algoritmada analize ulaşmak için oldukça önemlidir. Müşterilere ait bilgiler Decision Tree ve Logistic Regression algoritmaları ile işlenerek bu müşterilerden hangilerinin bankaya ait olan ürünü alırken hangilerinin bu ürünü almayacakları tahmin edilerek tahmin oranının doğruluğu hesaplanacaktır.

1.1. KARAR AĞAÇLARI VE KARAR AĞAÇLARI ALGORİTMASI

Karar Ağaçları sınıflama, özellik ve hedefe göre karar düğümleri (decision nodes) ve yaprak düğümlerinden (leaf nodes) oluşan ağaç yapısı formunda bir model oluşturan bir sınıflandırma yönetimidir. Sınıflandırma ve tahmin için sıkça kullanılan bir veri madenciliği (data mining) yaklaşımıdır. Bir karar alıcı tarafından tercihlerin, risklerin, kazançların ve hedeflerin tanımlanmasında yardımcı olmak için çeşitli karar düğümleri incelenebilir. Bu model, bir karar alıcıya karar alırken hangi faktörlerin göz önüne alınmasının belirlenmesinde yardımcı olur. Yapay Sinir Ağları(Artificial Neural Networks) gibi diğer metodolojilerin de sınıflandırma için kullanılmasına rağmen, karar ağaçları, kolay yorumu ve anlaşılabilirliği açısından karar vericiler açısından avantaj sağlamaktadır. Karar ağaçları;

- ✓ Düşük maliyetli olması,
- ✓ Anlaşılmasının, yorumlanmasının ve veritabanları ile entegrasyonunun kolay olması,
- ✓ Çoğu makine öğrenmesi algoritması ya sayısal ya da sınıflandırma verileri için kullanışlı olmasına rağmen karar ağaçlarının her iki veri tipinin işlenmesi için kullanılması,
- ✓ Güvenirliliklerinin iyi olması,
- ✓ Yüksek miktardaki veriyi kısa sürede işleyebilmesi gibi nedenlerden dolayı en yaygın kullanılan sınıflandırma tekniklerinden biridir.

Karar ağacı algoritması, veri setini küçük ve daha küçük parçalara bölerek ilerler. Karar ağacı tekniğini kullanarak verinin sınıflanması, öğrenme ve sınıflama olmak üzere iki aşamalıdır. Öğrenme aşamasında bir tane öğrenme kümesi oluşturulur. Bu öğrenme kümesindeki örnekleri (instance) en iyi ayıran nitelik (attribute) belirlenir. En iyi nitelik ağacın kök düğümü (root node) olmuş olur. Daha sonra bu kök düğümün alt düğümleri, alt düğümlerin de yine alt düğümleri veya yaprak düğümleri oluşturulur. Bu şekilde veriler küçük ve daha küçük verilere bölünmüş olur. Daha küçük verilere ayırma işlemi örneklerin hepsi aynı sınıfa ait, örnekleri bölecek nitelik kalmamış, kalan niteliklerin özelliklerini taşıyacak örnek kalmayana kadar devam eder ve yaprak düğümleriyle sonlanır. Sınıflama aşamasında ise test verisi oluşturulur ve test verisi karar ağacının doğruluğunu belirlemek amacıyla kullanılır. Veri setleri birçok nitelikten oluşmaktadır. Bu nitelikler veriyi daha iyi tanımamıza yardımcı

olmaktadır. Öğrenme aşamasında örnekleri en iyi ayıran niteliğin belirlenip o niteliğin kök düğüm olduğunu söylemiştik. Veri setindeki en iyi niteliğin belirlenmesi için kullanılan en yaygın yöntem Entropi ölçümüdür.

Entropi kimyada bilinen anlamıyla rastgeleliğin, belirsizliğin ve beklenmeyen durumun ortaya çıkma olasılığını gösterir. Veri madenciliği yaklaşımında da aynı anlama geldiğinden dolayı en iyi niteliğin belirlenmesinde entropi ölçüsü en az olan nitelikler kullanılır. Çünkü entropisi yüksek olan nitelik o oranda belirsiz ve kararsızdır. Verilen bir A_K alanının Entropi ölçüsünü bulan formüller şu şekildedir:

$$E(C|A_K) = \sum_{j=1}^{M_K} p(a_k, j) * [- \sum_{i=1}^{N} p(c_i|a_k, j) log_2 p(c_i|a_k, j)]$$

Bu formülde;

$E(C|A_K) = A_K$ alanının sınıflama özelliğinin Entropi ölçüsü,
$p(a_k, j) = a_k$ alanının j değerinde olma olasılığı,
$p(c_i|a_k, j) = a_k$ alanı j. değerindeyken sınıf değerinin c_i olma olasılığı,
$M_k = a_k$ alanının içerdiği değerlerin sayısı; $j = 1, 2, ., M_k$,
N = farklı sınıfların sayısı; $i = 1, 2, \ldots, N$,
K = niteliklerin sayısı; $k = 1, 2, \ldots, K$

Eğer bir S kümesindeki elemanlar, kategorik olarak C_1,C_2,C_3,...,C_i sınıflarına ayrıştırılırsa, S kümesindeki bir elemanın sınıfını belirlemek için gereken bilgi şu formülle hesaplanmaktadır.

$$I(s_1, s_2, \ldots . s_i) = - \sum_{i=1}^{m} p_i log_2 (p_i)$$

Bu formülde P_i, C_i sınıfına ayrılma olasılığıdır. Entropi denklemi şu şekilde de ifade edilebilir.

$$E(A) = \sum_{i=1}^{n} \left(\frac{S_i}{S} \right) I(s_i)$$

Bu durumda A niteliği kullanılarak yapılacak dallanma işleminde, Bilgi Kazancı (information gain) şu formülle hesaplanmaktadır.

$$Kazanc(A) = E(S) - E(A)$$

Burada E(S); sistemin entropi değeri olurken E(A) ise A niteliğinin entropi değeridir. Her nitelik için ayrı ayrı bilgi kazancı hesaplanır ve bilgi kazancı en yüksek olan nitelik ağacın en üstünde konumlanarak kök düğüm olmuş olur. Bu işlemler her düğüm için örneklerin hepsi aynı sınıfa ait, örnekleri bölecek özellik kalmamış, kalan özelliklerin değerini taşıyan örnek olmayana kadar devam eder. Ve bu adımlar tamamlanarak karar agacı tamamlanmış olur. Şimdi az örnekli bir veri kümesi üzerinde kök düğüm'ün nasıl bulunduğuna bakalım:

1.1.1. Örnek Uygulama

Bu örnek veri kümemizde hava, ebeveyn ve para durumuna göre o hafta ne yapılacağına karar verilecektir. Örnek veri kümesine bakıldığında;

Hafta	Hava Durumu	Ebeveyn Durumu	Para Durumu	Karar
1.Hafta	Güneşli	Var	Var	Sinema
2.Hafta	Güneşli	Yok	Var	Tenis
3.Hafta	Rüzgarlı	Var	Var	Sinema
4.Hafta	Yağmurlu	Var	Yok	Sinema
5.Hafta	Yağmurlu	Yok	Var	Ev
6.Hafta	Yağmurlu	Var	Yok	Sinema
7.Hafta	Rüzgarlı	Yok	Yok	Sinema
8.Hafta	Rüzgarlı	Yok	Var	Alışveriş
9.Hafta	Rüzgarlı	Var	Var	Sinema
10.Hafta	Güneşli	Yok	Var	Tenis

Veri setimizde toplam 10 örnek vardır. Bu 10 örnekten;

6 örnek için karar sinema (6/10),

2 örnek için karar tenis oynamak (2/10),

1 örnek için karar evde kalmak (1/10),

1 örnek için karar alışverişe gitmektir (1/10).

S kümesindeki elemanlar 4 sınıfa ayrılmıştır. Her biri için sınıflara ayrılma olasılığı parantez içinde belirtilmiştir. İlk önce sistemin entropisi bulunacaktır.

$$Entropi = -(6/10)log_2(6/10) - 2/10log_2(2/10) - 1/10log_2(1/10) - 1/10log_2(1/10);\ E(S) = 1,571\text{'dir.}$$

Kök düğümünün hangisi olduğunu bulmak için niteliklerimiz olan; hava, ebeveyn ve para özelliklerimizin bilgi kazançlarını hesaplanıp en yüksek kazanca sahip olan nitelik kök düğüm olarak seçilecektir. Hava niteliğimiz için bilgi kazancı değeri;

Kazanç (Hava Durumu)= E(S) – E(Hava Durumu)' dur. Bunun için hava durumunun entropi değerine ihtiyaç vardır.

Hava Durumu Güneşli, Rüzgarlı ve Yağmurlu şeklindedir. S kümesi Hava durumu için 3 sınıfa ayrılmıştır. **S_1** : Güneşli, **S_2**: Rüzgarlı, **S_3** : Yağmurlu

- ✓ **S_1** toplam 3 tane olup bunlardan 1'i sinemaya gitmek 2'si de tenis oynamaktır. **S_1** 2 sınıfa ayrılmış olup bu sınıfların **S_1** de bulunma olasılılarına göre bilgi değerleri

$$I(S_1) = -1/3log_2(1/3) - 2/3log_2(2/3) = 0,918$$

- ✓ **S_2** toplam 4 tane olup bunlardan 3'ü sinemeya gitmek 1'i de alışveriş yapmaktır. **S_2** 2 sınıfa ayrılmış olup bu sınıfların **S_2**'de bulunma olasılılarına göre bilgi değerleri

$$I(S_2) = -3/4log_2(3/4) - 1/4log_2(1/4) = 0,811$$

- ✓ **S_3** toplam 3 tane olup bunlardan 2'si sinemaya gitmek 1'i de evde kalmaktır. **S_3** 2 sınıfa ayrılmış olup bu sınıfların **S_3**'te bulunma olasılıklarına göre bilgi değerleri

$$I(S_3) = -2/3log_2(2/3) - 1/3log_2(1/3) = 0,918$$

Herhangi bir A niteliği için Entropi değeri

$$E(A) = \sum_{i=1}^{n} \left(\frac{S_i}{S} \right) I(s_i)$$

formülünden;

E (HavaDurumu)= 3/10∗0,918+4/10∗0,811+3/10∗0,918=0,8752

Bu durumda hava durumunun Bilgi Kazancı = 1,571-0,8752= **0,70** olarak bulunur.

Ebeveyn niteliği için bilgi kazancına bakılırsa; Ebeveyn durumu var veya yok şeklindedir. S kümesi için Ebeveyn niteliği 2 sınıfa ayrılmıştır. **S_1** : Var, **S_2** : Yok

- ✓ **S_1** toplamda 5 tane olup 5'i de sinemaya gitmektir. **S_1** 1 sınıfa ayrılmış olup bu sınıfın **S_1**'te bulunma olasılıklarına göre bilgi değerleri

$$I(S_1) = 5/5log_2(5/5) = 0$$

- ✓ **S_2** toplamda 5 tane olup 2'si tenis, 1'i ev, 1'i sinema, 1'i de alışveriştir. **S_2** 4 sınıfa ayrılmış olup bu sınıfların **S_2**'te bulunma olasılılarına göre bilgi değerleri

$$I(S_2) = -2/5log_2(2/5) - 1/5log_2(1/5) - 1/5log_2(1/5) - 1/5log_2(1/5) = 1,922$$

Herhangi bir A niteliği için Entropi değeri

$$E(A) = \sum_{i=1}^{n} \left(\frac{S_i}{S} \right) I(s_i)$$

formülünden;

E (ebevyn) = (5/10)*0+(5/10)*1,922= 0,961

Bu durumda ebevyn durumunun Bilgi Kazancı = 1,571-0,961= **0,61** olarak bulunur.

Para durumu için de aynı işlemler yapıldığında Bilgi Kazancı **0,2816** şeklinde bulunur. Hava durumu, ebevyn ve para niteliklerinden bilgi kazancı en yüksek olan 0,70 ile hava durumudur. Bu nedenle hava durumu kök düğüm olmuş olur. Böylece az örnekli bir veri setinde kök düğümün bu şekilde hesaplama ile bulunuduğunu göstermiş olduk. Bizim üzerimizde çalışacağımız veri kümesi 4521 örnekli bir veri olduğundan dolayı bu şekilde hesaplama yaparak kök düğüme ulaşmamız çok uzun zaman alacaktır. Bu nedenle R studio üzerinde R'ın kendine ait kütüphaneleri kullanılarak sonuç alınacaktır.

1.2. VERİ SETİ

Veri setimiz Portekiz'e ait bir bankanın müşteri bilgilerini içermektedir. Bankada olan kampanyalar için bu müşteriler telefon ile aranmaktadır. Eğer müşteri kampanyaya ait ürünü almayı kabul ederse birkaç defa daha iletişime geçilmektedir. Bank.csv tablomuzda 4521 tane örnek (instance) olup, 16 + karar (y) niteliği (attribute) bulunmaktadır. Bu nitelikler örneklere ait özelliklerdir ve veriyi tanımamıza yardımcı olur. Bu niteliklere bakılacak olursa;

1. age: Müşterilere ait yaş bilgisini içerir. Sayısal bir değerdir.
2. job: Müşterilerin meslek bilgisini içerir.Kategorik olup admin, unknown, unemployed, management, housemaid, entrepreneur, student, blue-collar, self-employed, retired, techinician, services şeklinde 12 çeşittir.
3. marital: Müşterilerin medeni durumlarını içermektedir. Kategorik olup married, divorced, single şeklinde 3 çeşittir.
4. education: Müşterilere ait eğitim bilgilerini içermektedir.Kategorik olup unknown, secondary, primary, tertiary şeklinde 4 çeşittir.
5. default: Müşterinin bankada kredisi olup olmadığının bilgisini içerir. Kategorik olup Yes, No şeklinde 2 çeşittir.
6. balance: Müşterinin ortalama yıllık bakiyesini euro olarak ifade etmektedir. Sayısal bir değerdir.
7. housing: Müşterinin konut kredisi olup olmadığının bilgisini içerir. Kategorik olup Yes, No şeklinde 2 çeşittir.
8. loan: Müşterinin bireysel kredisinin olup olmadığının bilgisini içerir.Kategorik olup Yes, No şeklinde 2 çeşittir.
9. contact: Müşteriye ait iletişim tipini belirtir. Kategorik olup unknown, telephone, cellular şeklinde 3 çeşittir.
10. day: Bu kampanya için müşteriyle en son iletişime geçirilen ayın gün bilgisini içerir. Sayısal bir değerdir.
11. month: Bu kampanya için müşteriyle en son iletişime geçirilen ayın bilgisini içerir. Kategorik bir değer olup jan,feb,mar,apr,may,june,july,aug,sep,oct,nov,dec şeklinde 12 çeşittir.

12. duration:Bu kampanya için müşteriyle en son iletişime geçirilen görüşmenin saniye olarak süresini içerir. Sayısal bir değerdir.
13. campaign: Bu kampanyada bir müşteri için kaç kez iletişime geçtiğinin bilgisini içerir. Sayısal bir değerdir.
14. pdays: Bir önceki kampanyada müşteri ile en son iletişime geçilmesinden sonra kaç gün geçtiğini gösterir. Sayısal bir değerdir. Burada bulunan -1 değeri daha önce iletişime geçilmediğini göstermektedir.
15. previous: Bir önceki kampanyada müşteri ile iletişime kaç defa geçirildiğinin bilgisini içerir. Sayısal bir değerdir.
16. poutcome: Müşterinin bir önceki kampanyadaki ürünü alıp almadığının sonucunu içerir. Kategorik olup unknown, other, failure, success şeklinde 4 çeşittir.
17. y: Müşteri ile ilgili kredi verilip verilmemesinin kararının verildiği alandır. Kategorik olup Yes, No şeklinde 2 çeşittir.

Uygulama kısmında veri madenciliği çözümleri ile hem istatistik kökenli uygulamaları hem de yapay zeka kökenli algoritmaları bir programlama arayüzü altında sunan R Studio'dan yararlanılmıştır. Banka verisine ait veri kümesi txt olarak düzenlenmiştir.

```
data<-read.table("Banka.txt",header=T,sep="\t")
head(data)
```

age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
30	unemployed	married	primary	no	1787	no	no	cellular	19	oct	79	1	-1	0	unknown	no
33	services	married	secondary	no	4789	yes	yes	cellular	11	may	220	1	339	4	failure	no
35	management	single	tertiary	no	1350	yes	no	cellular	16	apr	185	1	330	1	failure	no
30	management	married	tertiary	no	1476	yes	yes	unknown	3	jun	199	4	-1	0	unknown	no
59	blue-collar	married	secondary	no	0	yes	no	unknown	5	may	226	1	-1	0	unknown	no
35	management	single	tertiary	no	747	no	no	cellular	23	feb	141	2	176	3	failure	no

read.table Kullanımı

read.table fonksiyonu dosyayı tablo formatında okur ve dosyadaki alanlara ve satırlara karşılık gelen durumlarla ilgili olarak bir veri çerçevesi oluşturur.

read.table(file, header=FALSE, sep="",) şeklindedir. Burada **file parametresi;** okunacak verinin dosya ismidir. Buraya dosyanın tam yolu yazılabileceği gibi sadece adı da yazılabilir. Eğer ki dosya R'ın çalıştığı dizinde bulunuyorsa dosyanın sadece adını yazmak yeterlidir. Getwd() komutu ile R'nin çalıştığı dizinin yolu öğrenilir. Banka dosyamızı bu dizinin altına koyduğumuz için sadece dosya adını ve uzantısını yazmamız yeterli oldu**. header parametresi;** dosyadaki değişkenlerin adlarını ilk satırda içerip içermediğinin bilgisini içeren mantıksal bir değerdir. Default olarak FALSE değerindedir. Banka dosyamızda ilk satırda değişkenlerin ismi yazılmıştır. Bu nedenle TRUE değerini alır. **sep parametresi**; ayırıcı karakterdir. Satırlardaki her bir değerin birbirinden nasıl ayrıldığını belirtmek için kullanılır. Default değeri "" şeklindedir. Bizim verimizde değerler birbirinden tab ile ayrıldıklarından dolayı sep="\t" 'dir. Veri setimizi data isimli bir değişkene atadık. Veri setimiz 4521 örnekli olduğundan dolayı bütün satırları yazmak yerine head fonksiyonu ile ilk 6 satır yazılmış oldu.

Örneğin; ilk örneğimiz 30 yaşında, işsiz, evli, ilkokul mezunu, aylık geliri 1787 euro olan, en son iletişime 19 ekimde 79 sn'lik bir telefon görüşmesiyle geçilmiş, önceki kampanyadaki ürünü alıp almadığı bilinmeyen bir müşteridir.

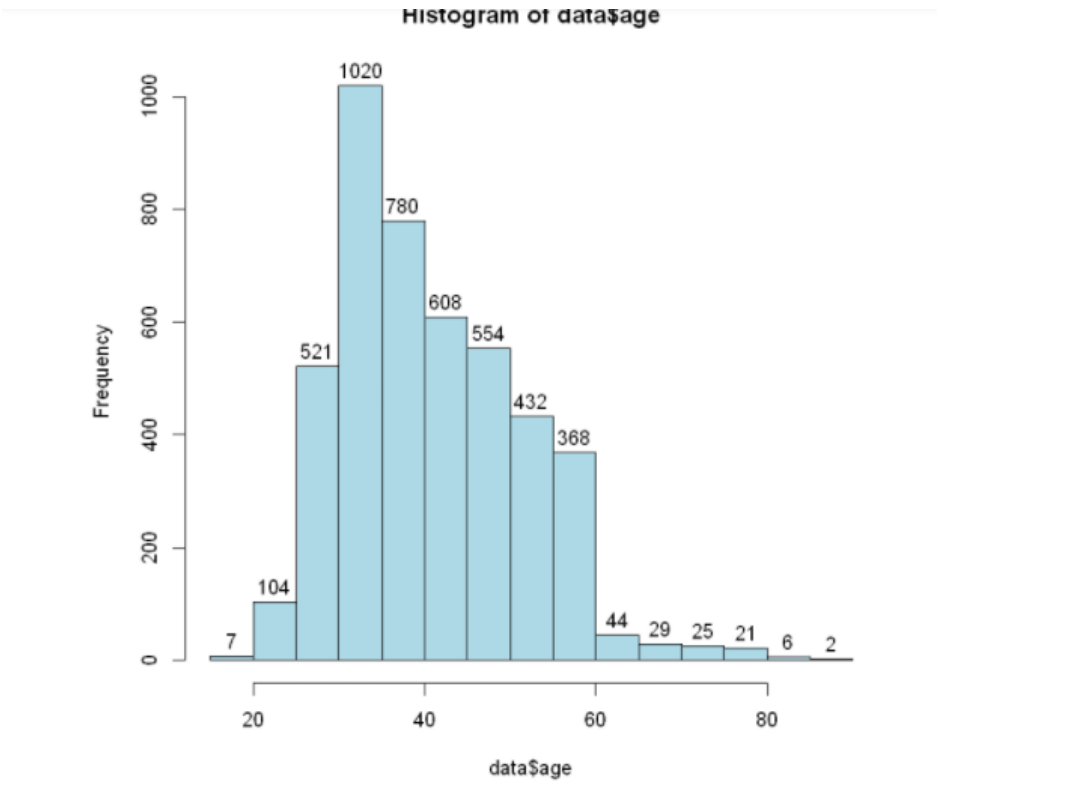
```
summary(data)
```

```
      age                job          marital         education    default
 Min.   :19.00   management :969   divorced: 528   primary  : 678   no :4445
 1st Qu.:33.00   blue-collar:946   married :2797   secondary:2306   yes:  76
 Median :39.00   technician :768   single  :1196   tertiary :1350
 Mean   :41.17   admin.     :478                   unknown  : 187
 3rd Qu.:49.00   services   :417
 Max.   :87.00   retired    :230
                 (Other)    :713
    balance       housing      loan           contact          day
 Min.   :-3313   no :1962   no :3830   cellular :2896   Min.   : 1.00
 1st Qu.:   69   yes:2559   yes: 691   telephone: 301   1st Qu.: 9.00
 Median :  444                         unknown  :1324   Median :16.00
 Mean   : 1423                                         Mean   :15.92
 3rd Qu.: 1480                                         3rd Qu.:21.00
 Max.   :71188                                         Max.   :31.00

     month         duration       campaign          pdays
 may    :1398   Min.   :   4   Min.   : 1.000   Min.   : -1.00
 jul    : 706   1st Qu.: 104   1st Qu.: 1.000   1st Qu.: -1.00
 aug    : 633   Median : 185   Median : 2.000   Median : -1.00
 jun    : 531   Mean   : 264   Mean   : 2.794   Mean   : 39.77
 nov    : 389   3rd Qu.: 329   3rd Qu.: 3.000   3rd Qu.: -1.00
 apr    : 293   Max.   :3025   Max.   :50.000   Max.   :871.00
 (Other): 571
    previous          poutcome      y
 Min.   : 0.0000   failure: 490   no :4000
 1st Qu.: 0.0000   other  : 197   yes: 521
 Median : 0.0000   success: 129
 Mean   : 0.5426   unknown:3705
 3rd Qu.: 0.0000
 Max.   :25.0000
```

Summary fonksiyonu çeşitli model oluşturma fonksiyonlarının sonuç özetlerini üretmek için kullanılan genel işlevli bir fonksiyondur. Summary fonksiyonu ile sayısal değerli niteliklerin önemli istatistiksel değerleri elde edilimiş olur. Sayısal değerlerin en küçük değeri (min), en yüksek değeri (max), ortalaması (mean), ortanca değeri (median), ortanca değerinden küçük olan değerlerin ortancası (1st Quartile), ortanca değeriden büyük olan değerlerin ortancası (3rd Quartile) gibi önemli istatistiksel değerleri elde edilir. Kategorik değerli niteliklerin ise her bir kategoriden kaç tane olduğunun özeti verilmektedir. Örneğin yaş niteliğine (sayısal) bakıldığında müşterilerden en genç kişi 19 yaşında iken, en yaşlı kişi 87 yaşındadır. 4521 müşterinin yaş ortalaması 41.17'dir. Median değeri 39 iken, 1st Quartile 33 ve 3rd Quardile değeri 49'dur. Evlilik niteliğine (kategorik) bakıldığında ise 528 tane boşanmış (divorced), 2797 tane evli (married), 1196 tane de bekar (single) müşteri olduğu görülmektedir.

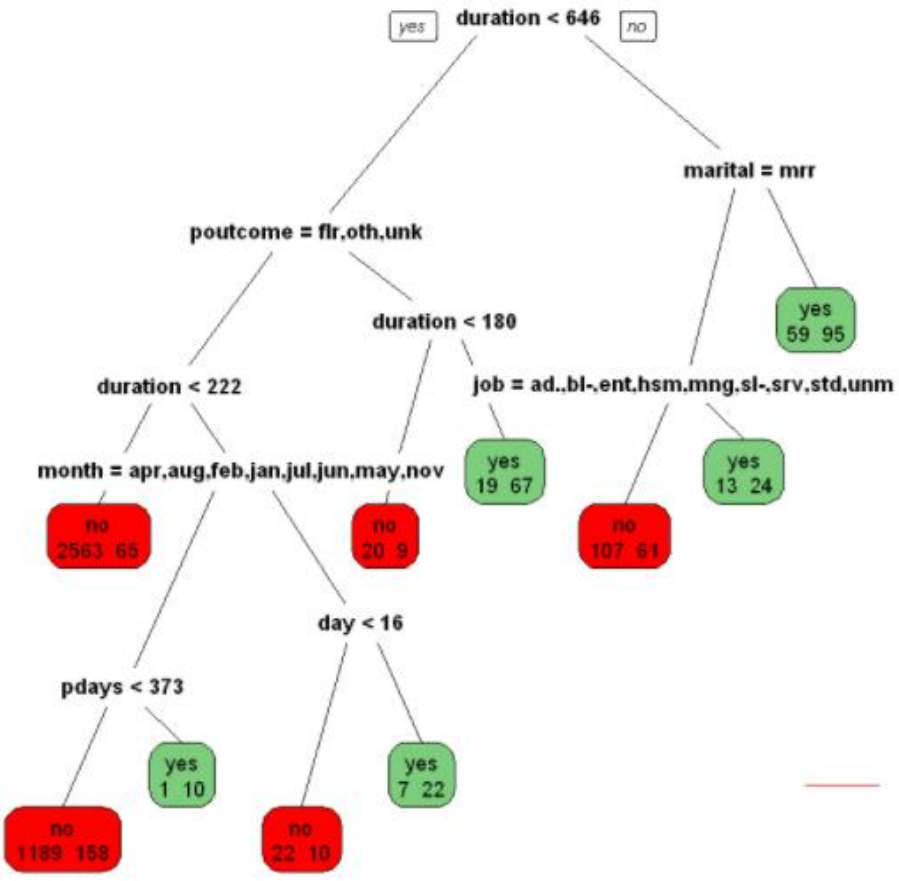
```
hist(data$age, col = "light blue", freq = TRUE,labels=TRUE)
```



hist fonksiyonu ile yaş-frekans arasındaki ilişki gösterilerek hangi yaş grubunda kaç tane müşteri olduğu görülmektedir.

Müşteri kararının tahmin edilmesi için karar ağacı algoritmalarından faydalanılmıştır. R'ın kendisine ait olan rpart ve party kütüphaneleri kullanılmıştır. Brieman, Friedman, Olshen ve Stone tarafından 1984 tarihinde geliştirilen sınıflandırma ve regresyon ağaçları (classification and regression trees(CART)) rpart paketi aracılığıyla üretilir.

```
library(rpart)
library(rpart.plot)
data<-read.table("Banka.txt",header=T,sep="\t")
elif<-rpart(y~.,data=data,method="class")
prp(elif,cex=0.9,box.palette=c("red","palegreen3"),extra=1)
legend("bottomright",legend=c("yes","no"),box.col=c("red","green"))
```



rpart ve rpart.plot paketleri R'nin çalıştığı dizine install.packages("rpart") ve install.packages("rpart.plot") kodları ile yüklenmiştir.

rpart Genel Kullanımı

rpart ile sınıflandırma ve regresyon analizleri yapılmaktadır. rpart(formula,data,method). Parametrelerin kullanımı;

formula =sonuç~nitelik1+nitelik2+..... Banka verisinde sonucumuz y alanında saklıdır. 16 tane nitelik olduğundan tek tek yazmak yerine **.** yazılmıştır. **data**=Veri kümesi, **method** ="class" (sınıflandırma ağacı için), "anova" (regresyon ağacı için) Sınıflandırma ağacı oluşturalacağı için class yazılmıştır. Yaygın kullanımı 3 parametreli şekildedir. control, weights, subset şeklinde opsiyonel parametreler de alabilir.

prp Genel Kullanımı

rpart.plot paketinin içinde kullanılan bir plot fonksiyondur. prp(x,cex,box.pallet,extra) Parametrelerin kullanımı; **x** = rpart nesnesidir. Tek zorunlu argumandır. **cex** ile yazı boyutu ayarlanır. **box.palette** ile düğüm sonucunda bulunan farklı değerler renklendirilir. **extra** ile düğüm sonucunda bulunan sayısal değerler yazılır.

Karar Ağacının Yorumlanması:

- ✓ Ağacın kök düğümünü mevduatı satmaya çalışan banka çalışanlarının müşterilerle telefonda yaptığı görüşmenin süresi(duration) oluşturmaktadır. Müşteri sayısı 4521 olan bir ortamda müşterilerin 4162'si telefonla 646 sn'den daha az konuşurken 359'u 646 sn'den daha fazla konuşmuştur.
- ✓ 646 sn'den daha az konuşan 4162 müşterinin yaklaşık 4047'si bir önceki kampanyadaki ürünü almamış (failure), ürünü alıp almadığı bilinmeyen (unknown) ve diğer durumda (other) olan müşteriler iken 115'i satın alan (success) müşterilerdir. Bu 4047 müşterinin 2628'nin telefonla yaptığı görüşme süresi 222 sn'den daha azken 1419'nun telefonla yaptığı görüşme süresi 222 sn'den daha fazladır. Telefonla yaptığı görüşme süresi 222 sn'den daha az olan 2628 müşterinin 2563'ü mevduatı satın almazken 65'i mevduatı satın almıştır (1. yaprak düğüm).
- ✓ Görüşme süresi 222 sn'den daha fazla olan 1419 müşteriden 1358'i bu telefon görüşmesini ocak, şubat, nisan, mayıs, haziran, temmuz, ağustos ve kasım aylarında gerçekleştirmiştir. Bu 1358 müşteriden 1347'sinin bir önceki kampanya çalışması için aranmasının üzerinden 373 günden daha az bir zaman geçmiştir. Bu 1347 müşteriden 1189'u mevduatı satın almazken 158'i mevduatı satın almıştır (2. yaprak düğüm). Bu 1358 müşteriden 11 kişinin bir önceki kampanya çalışması için aranmasının üstünden

373 günden fazla bir zaman geçmiştir. Bu müşterilerden 10'u mevduatı satın alırken 1'i mevduatı satın almamıştır (3. yaprak düğüm).

- ✓ Görüşme süresi 222 sn'den daha fazla konuşan 1419 müşterinin 61'i bu telefon görüşmesini mart, eylül, ekim ve aralık aylarında gerçekleştirmiştir. Bu 61 müşteriden 32'si mart, eylül, ekim ve aralık aylarının 16'sından önce aranırken 29'u bu ayların 16'sından sonra aranmıştır. 16'sından önce aranan 32 müşteriden 22'si mevduatı satın almazken 10'u mevduatı satın almıştır (4. yaprak düğüm).
- ✓ 16'sından sonra aranan 29 müşteriden 22'si mevduatı satın alırken 9'u mevduatı satın almamıştır (5. yaprak düğüm).
- ✓ Görüşme süresi 646 sn'den daha az olan ve bir önceki kampanyadaki ürünü satın aldığı bilinen 115 müşteriden 29'u telefonla 180 sn'den daha az konuşurken 86'sı 180 sn'den fazla konuşmuştur.(180 ile 646 sn arasında). 180 sn'den az daha az konuşan 29 müşteriden 20'si mevduatı satın almazken 9'u satın almıştır (6. yaprak düğüm).
- ✓ 180 sn'den fazla konuşan 86 kişiden 67'si mevduatı satın alırken 19'u satın almamıştır (7. yaprak düğüm).
- ✓ Görüşme süresi 646 sn'den daha fazla olan 359 müşteriden 205'nin medeni durumu evli değ iken 154'nün medeni durumu evlidir . Bu 205 müşterinin 168'nin meslekleri yönetici(admin), mavi yakalı(blue-collar), girişimci(entrepreneur), hizmetçi(housemaid), idareci(management), kendi işi olan(self employed), hizmet eden(services), öğrenci(student) veya işsiz(unemployed)'dir. Bu 168 müşteriden 107'si mevduatı satın almazken 61'i satın almıştır. (8. düğüm)
- ✓ Bu 359 müşteriden 37'sinin ise meslekleri bilinmeyen(unknown), emekli(retired) ve teknisyen(technician)'dir. Bu müşterilerden 24'ü mevduatı satın alırken 13'ü satın almamıştır. (9. düğüm)
- ✓ Medeni durumu evli olan 154 müşteriden 95'i mevduatı satın alırken 59'u mevduatı satın almaıştır. (10.düğüm)

Veri Kümesini Bölme ve Tahminleme

```
size<- nrow(data)*0.8
index<-sample(1:nrow(data),size=size)
training<-data[index,]
test<-data[-index,]
head(training)
head(test)
nrow(training)
```

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	pouto
1066	30	unemployed	married	tertiary	no	0	yes	no	cellular	18	nov	756	1	-1	0	unkno
1700	49	technician	married	unknown	no	323	yes	no	cellular	18	nov	41	2	-1	0	unkno
3027	39	management	divorced	tertiary	no	26	no	no	unknown	18	jun	311	5	-1	0	unkno
4194	48	services	divorced	secondary	no	3186	no	yes	cellular	9	apr	104	1	-1	0	unkno
3755	36	management	single	tertiary	no	2944	no	no	cellular	18	aug	882	8	-1	0	unkno
2796	36	technician	divorced	tertiary	no	1174	yes	no	cellular	18	nov	192	2	-1	0	unkno

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcc
6	35	management	single	tertiary	no	747	no	no	cellular	23	feb	141	2	176	3	failure
13	36	technician	married	tertiary	no	1109	no	no	cellular	13	aug	328	2	-1	0	unknow
26	41	management	married	tertiary	no	5883	no	no	cellular	20	nov	182	2	-1	0	unknow
31	68	retired	divorced	secondary	no	4189	no	no	telephone	14	jul	897	2	-1	0	unknow
34	32	management	single	tertiary	no	2536	yes	no	cellular	26	aug	958	6	-1	0	unknow
38	32	blue-collar	married	secondary	no	2089	yes	no	cellular	14	nov	132	1	-1	0	unknow

3616

Daha önce Karar Ağaçları'nda bahsedildiği üzere; Karar ağacını tekniğini kullanarak verinin sınıflanması, öğrenme ve sınıflama olmak üzere iki basamaklı bir işlemdir. Öğrenme basamağında önceden bilinen bir eğitim verisi, model oluşturmak amacıyla sınıflama algoritması tarafından test edilir. Sınıflama basamağında ise test veri seti ile karar ağacının doğruluğu test edilir. 4521 örneklik veri kümesi yapılması amacıyla %80 oranında (3616) eğitim (training) ve %20 oranında (905) da test veri kümesine bölünmüştür. Training veri seti ile model oluşturulurken, test veri seti ile de bu modelin doğruluğu test edilecektir. Sample fonksiyonu ile toplam veri üzerinden örneklem oluşturulmuştur. Training ve test veri kümelerini data veri kümesi üzerinden rastgele (random) seçilen örnekler oluşturmaktadır. Yani kodları her çalıştırdığımızda farklı örnekler gelmektedir. Örnek olması amacıyla daha önce de kullanılan head fonksiyonu ile ilk 6 örnek gösterilmiştir.

```
elif<-rpart(y~.,data=training,method="class")
prediction<- predict(elif,test,type="class")
head(prediction)
```

```
 6  no
13  no
26  no
31  yes
34  yes
38  no
```

predict Genel Kullanımı

predict fonksiyonu rpart nesnesinin tahmin edilen sonucunu bir vektör olarak döndürür. predict (object ,newdata, type=c("vector","prob","class","matrix") Parametrelerin kullanımı; **object**: rpart sınıfından rpart fonksiyonu tarafından oluşturulan nesnedir. Banka veri kümesi için rpart fonksiyonundan üretilen nesnemiz elif idi. **newdata:** Tahmin ediciler için gerekli olan değerleri içeren veri çerçevesidir. Kullanılacak newdata %80 oranında bölme işlemi yapılan test veri kümesidir. **type:** Tahmin edilen değerlerin döndürüldüğü karakter dizisidir. Vector, prob, class, matrix olarak 4 çeşidi vardır. Tercih olarak "class" kullanılmıştır.6 örnek için bakıldığında; tahmin ediciler kullanılarak ; 31 ve 34 numaralı müşterilerin ürünü satın alacağı tahmin edilirken, 6, 13, 26, 38 numaralı müşterilerin ürünü satın almayacağı tahmin edilmiştir. Peki gerçekte de durum bu şekilde mi?

```
df<- data.frame(data)
df[31,]
df[13,]
df[34,]
```

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
31	68	retired	divorced	secondary	no	4189	no	no	telephone	14	jul	897	2	-1	0	unknown	yes

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
13	36	technician	married	tertiary	no	1109	no	no	cellular	13	aug	328	2	-1	0	unknown	no

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
34	32	management	single	tertiary	no	2536	yes	no	cellular	26	aug	958	6	-1	0	unknown	yes

31, 13 ve 24 numaralı müşterilere bakıldığında tahminlerin doğru olduğu gözükmektedir.

```
confusion.matrix <- table(prediction, test$y)
confusion.matrix
```

```
prediction  no yes
       no  768  72
       yes  26  39
```

Confusion Matrix İlişkili istatistiklerle gözlemlenen ve tahmin edilen sınıfların çapraz tablolarını hesaplar.

- ✓ Test veri kümeside toplamda 840 müşterinin mevduatı satın almadığı (no) bilinirken; bunların 768 tanesi satın almaz (no), 72 tanesi de satın alır (yes) şeklinde tahmin edilmiştir. (1.satır)
- ✓ Training veri kümesinde toplamda 65 müşterinin mevduatı satın aldığı bilinirken; bunların 26 tanesi satın almaz (no), 39 tanesi de satın alır (yes) şeklinde tahmin edilmiştir. (2.satır)
- ✓ 905 örneklemli veri kümesinden 768+39=807 örnek doğru tahmin edilerek başarı oranı %89.17 olarak bulunmuştur.

1.3. LOJİSTİK REGRESYON

Lojistik regresyon; cevap değişkeninin kategorik olarak ikili, üçlü veya çoklu olduğu durumlarda açıklayıcı değişkenlerle neden sonuç ilişkisini belirlemede yararlanılan bir yöntemdir. Açıklayıcı değişkenlere göre cevap değişkeninin beklenen değerlerinin olasılık olarak elde edildiği bir regresyon yöntemidir. Lojistik regresyon analizi, sınıflama ve atama işlemi yapmaya yardımcı olan bir regresyon yöntemidir. Basit ve çoklu regresyon analizlerinde ön koşul olarak normal dağılım ve süreklilik varsayımı aranırken lojistik regreyon analizlerinde böyle bir ön koşul yoktur. Basit regresyon analizinde bağımlı değişkenler ve bağımsız değişkenler sayısal olarak belirtilir. Nitelik olarak belirtilmezler. Örneğin; yaş ile kan basıncı arasında bir ilişki aranacaksa hem yaş hem de kan basıncı sayısal olarak belirtilmelidir. Eğer ki bağımlı değişken nitelik olarak belirtilirse bağımsız değişkenlerle arasındaki ilişki lojistik resresyon yöntemiyle aranır. Lojisitk regresyon tanımını daha iyi anlayabilmek için aşağıda bazı tanımlar verilmiştir.

Nitel değişken: Ölçülemez, sadece nitelendirilebilir değişkenlerdir. Örneğin; Cinsiyet: Erkek, Kadın; Eğitim: İlk-Orta-Lise-Yüksek gibi.

Sayısal değişken: Kesikli sayısal ve Sürekli sayısal olarak iki gruba ayrılır.

Kesikli sayısal değişken: Belirli bir aralıktaki tam sayıları alabilen değişkenlerdir. Örneğin; Nabız sayısı, ölen çocuk sayısı gibi.

Sürekli sayısal değişken: Ölçümle belirtilen ve belirli bir aralıktaki bütün desimal değerleri alabilen değişkenlerdir. Örneğin; Kan basıncı ölçümü, Boy uzuluğu, ağırlık gibi

Bağımlı değişken: Diğer değişkenler tarafından etkilenen değişkendir.

Bağımsız değişken: Bağımlı değişkeni etkileyen değişkendir. Örneğin; bir öğrencinin başarısı cinsiyete, yaşa ve sosyal durumlarına bağımlı olabilir.

Lojistik regresyon analizinin amacı hem sınıflandırma yapmak hem de bağımlı ve bağımsız değişkenler arasındaki ilişkiyi araştırmaktır. Lojistik regresyon analizi daha çok sosyo-ekonomik konuların pazar araştırmalarında kullanılmaktadır. Lojistik analiz yöntemi ilk olarak nüfus artışının matematiksel bir ifadeyle açıklanmasına yönelik araştırmada ortaya sunulmuştur. Lojistik regresyon analizi bağımlı değişkenin sayısına ve kategori sayısına göre 3'e ayrılmaktadır. Eğer bağımlı değişken iki seçenekli kategorik değişkene sahip ise İkili Lojistik Regresyon Analizi (Binaty Logistic Regression) uygulanır. Eğer bağımlı değişken ikiden çok kategorili değişkene sahipse Çok Kategorili İsimsel Lojistik Regresyon Analizi

(Multinominal Lojistic Regression) uygulanır. Eğer bağımlı değişken sıralama ölçeğiyle elde edilmişse Sıralı Lojistik Regresyon Analizi (Ordinal Logistic Regression) uygulanır.

Matematiksel olarak lojistik regresyon olasılık, odds ve odds'un logaritmasına dayanır. Olasılık en temel olarak belirli bir tipteki sonuç sayısının toplam sonuçlar içerisindeki oranıdır. Örneğin; bir zar atıldığında 5 gelme olaslığı 1/6' (0,167)'dır. Çünkü zar üzerinde bir tane 5 vardır ve toplamda 6 sonuç vardır. Lojistik regresyonda odds, bir olayın olma olasılığının o olayın olmama olasılığına bölümü olarak tanımlanır.

$$Odds = \frac{p(x)}{1 - p(x)}$$

Burada p(x) bir olayın gerçekleşme olasılığını, 1-p(x) ise bir olayın gerçekleşmeme olasılığını göstermektedir. Bir zar atıldığında 5 gelme olasılığına ilişkin odds değeri 0,167/0,833=0,200'dür. Olasılıkların değeri 0 ile 1 arasında değişirken odds değeri için böyle bir kısıtlama yoktur. Odss değeri 1'den büyük olabilir. Örneğin; bir sınavda öğrencilerin başarılı olma olasılığı 0.80 olsun. Başarılı olmaya ilişkin odds 0.80/0.20=4'tür. Görüldüğü üzere odds değeri 1'den büyük olabilir. Bu örnekte odd'un anlamı başarılı olmanın olasılığı başarısız olmaya göre 4 kat fazladır. 0.50'lik bir olasılık (yani bir olayın olma ya da olmama/ ortaya çıkma ya da çıkmama şansının eşit olma olasılığı) odds'un 1 olması ile sonuçlanır. Odds'un 1'den küçük olması ilgilenilen olasılığın 0.50'den az; 1'den büyük olması ise ilgilenilen olasılığın 0.50'den fazla olduğu anlamına gelir. Odds oranı φφ ile gösterilir. Lojistik regresyon analizinde elde edilen sonuç doğrusal olmayan bir fonksiyondur. Lojistik regresyondaki en önemli kavram adını da kendisinden alan "lojit" kavramıdır. Lojit odds oranının doğal logaritmasıdır. (ln). Yukarıdaki zar örneğine geri dönülürse bir zarın 5 gelme olasılığına ilişkin odds için lojit değer ln(0,200)=-1,609'dur.

*Bazı olasılık değerlerine karşılık gelen Odds ve Log Odds Oranları

Olasılık	Odds	Log Odds
.00	.00	Hesaplanamaz
.10	.11	-2.20
.30	.43	-0.85
.50	1.00	.00
.70	2.33	0.85
.90	9.00	2.20
1.00	Hesaplanamaz	Hesaplanamaz

Olasılık, odds ve lojit kavramlarını tek bir eşitlikte bir araya getirdiğimizde;

$$Y_i = \frac{e^u}{1 + e^u}$$

Burada Yi'nci örneğin bağımlı değişkenin kategorilerinin birisinde yer almasına ilşkin kestirilen olasılıktır. e ise 2.718'e eşit bir sabittir. Burada u ise klasik regresyon eşitliğidir.

$$u = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots . \beta_k X_k$$

Doğrusal regresyon eşitliği (u) odds oranının lojitini oluşturur.

$$ln(\frac{Y}{1-Y}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots . \beta_k X$$

Uygulama

Lojistik regresyon modelinin kurulmasında SPSS, SAS, R gibi çeşitli istatistik araçları kullanılmaktadır. Bu çalışmada banka veri kümesi üzerinde lojistik regresyon analizi R'da uygulanarak müşterilerin bankaya ait satılan ürünü alıp almayacakları tahmin edilerek bu tahmini ne ölçüde gerçeklendirdiği tespit edilecektir. Bu uygulamada, bağımlı değişkenin (y= Yes, No) iki kategorili olma durumunda olmasından dolayı İkili Lojistik Regresyon Analizi (Binary Logistic Regression) kullanılmıştır. Bu uygulamada yer alan bağımsız değişkenler age, job, marital, education, default, balance, housing, loan, contact, day, month, duration, campaign, pdays, previous, poutcome'dır. Lojistik regresyon analizinde tekniklerin uygulanabilmesi için ilk önce verinin temizlenmesi işlemi yapılması gerekmektedir. Verideki bazı değerler silinmiş veya bozulmuş olabilir.

Örneğin banka veri kümesinde müşteriye ait yaş, eğitim durumu gibi değişkenlerin boş bırakılması veya birden fazla değer girilmesi sağlıklı bir analiz yapılmasını engelleyecektir.

```
training.data.raw <- read.csv('Banka.csv',header=T,na.strings=c(""))
sapply(training.data.raw,function(x) sum(is.na(x)))
sapply(training.data.raw, function(x) length(unique(x)))
```

```
age.job.marital.education.default.balance.housing.loan.contact.day.month.duration.campaign.pdays.previous.poutcome.y: 0
age.job.marital.education.default.balance.housing.loan.contact.day.month.duration.campaign.pdays.previous.poutcome.y: 4521
```

Buradan anlaşıldığı üzere herhangi silinmiş veya bozulmuş değer bulunmamaktadır. Eğer böyle bir değer bulunmuş olsa idi, o değere ait ortalama (mean) değeri yazılarak değer düzeltilebilirdi.

```
data<- read.table("banka.txt",header=T,sep="\t")
size<- nrow(data)*0.8
index<-sample(1:nrow(data),size=size)
training<-data[index,]
test<-data[-index,]
model <- glm(y~.,family=binomial(link='logit'),data=training)
summary(model)
```

```
Call:
glm(formula = y ~ ., family = binomial(link = "logit"), data = training)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.0350  -0.3721  -0.2455  -0.1526   3.0729

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)       -3.001e+00  6.997e-01  -4.288 1.80e-05 ***
age               -8.148e-03  8.128e-03  -1.002  0.31611
jobblue-collar    -4.424e-01  2.794e-01  -1.583  0.11332
jobentrepreneur   -1.629e-01  4.288e-01  -0.380  0.70408
jobhousemaid      -3.242e-02  4.606e-01  -0.070  0.94388
jobmanagement      2.566e-02  2.784e-01   0.092  0.92656
jobretired         1.029e+00  3.581e-01   2.874  0.00405 **
jobself-employed   7.650e-02  3.904e-01   0.196  0.84465
jobservices       -7.729e-02  3.096e-01  -0.250  0.80287
jobstudent         1.603e-01  4.367e-01   0.367  0.71356
jobtechnician     -1.350e-01  2.651e-01  -0.509  0.61072
jobunemployed     -5.831e-01  4.782e-01  -1.219  0.22271
jobunknown         7.649e-01  6.328e-01   1.209  0.22677
maritalmarried    -3.561e-01  2.036e-01  -1.749  0.08021 .
maritalsingle     -1.954e-01  2.349e-01  -0.832  0.40553
educationsecondary 2.267e-01  2.343e-01   0.967  0.33337
educationtertiary  3.742e-01  2.697e-01   1.387  0.16539
educationunknown  -4.066e-01  4.303e-01  -0.945  0.34473
defaultyes         7.615e-01  4.443e-01   1.714  0.08651 .
balance            2.634e-05  2.130e-05   1.237  0.21623
housingyes        -2.090e-01  1.564e-01  -1.336  0.18155
loanyes           -5.360e-01  2.262e-01  -2.369  0.01782 *
contacttelephone  -2.198e-01  2.808e-01  -0.783  0.43374
contactunknown    -1.426e+00  2.558e-01  -5.577 2.45e-08 ***
day                2.013e-02  9.326e-03   2.159  0.03087 *
monthaug          -2.565e-01  2.845e-01  -0.902  0.36727
monthdec           5.656e-01  7.421e-01   0.762  0.44599
monthfeb           4.110e-01  3.340e-01   1.230  0.21852
monthjan          -1.315e+00  4.566e-01  -2.879  0.00399 **
monthjul          -7.179e-01  2.842e-01  -2.526  0.01154 *
monthjun           6.505e-01  3.450e-01   1.886  0.05932 .
monthmar           1.315e+00  4.636e-01   2.837  0.00455 **
monthmay          -3.695e-01  2.706e-01  -1.366  0.17205
monthnov          -8.782e-01  3.123e-01  -2.812  0.00492 **
monthoct           1.566e+00  3.822e-01   4.099 4.16e-05 ***
monthsep           6.678e-01  4.701e-01   1.420  0.15547
duration           4.421e-03  2.318e-04  19.070  < 2e-16 ***
campaign          -5.426e-02  3.015e-02  -1.799  0.07197 .
pdays             -7.194e-04  1.164e-03  -0.618  0.53643
previous           1.354e-03  4.525e-02   0.030  0.97613
poutcomeother      6.620e-01  3.149e-01   2.103  0.03550 *
poutcomesuccess    2.847e+00  3.349e-01   8.502  < 2e-16 ***
poutcomeunknown    6.074e-03  3.787e-01   0.016  0.98720
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2560.9  on 3615  degrees of freedom
Residual deviance: 1687.8  on 3573  degrees of freedom
AIC: 1773.8

Number of Fisher Scoring iterations: 6
```

Burada; Karar Ağacı algoritmasında yapıldığı gibi Banka veri kümemiz %80 oranında training kümesine ve %20 oranında test kümesine bölünmüştür. Training veri kümesi lojistik regresyon modelinin oluşturulmasında kullanılırken test kümesi ise tahminleme (prediction) işlemi için kullanılacaktır.

glm Genel Kullanımı

glm doğrusal tahmin ediciler ile birlikte genelleştirilmiş lineer modelleri uygulamak için kullanılan bir fonksiyondur. glm(formula , family=familytype(link=linkfunction), data)

Parametrelerin kullanımı, formula =sonuç~nitelik1+nitelik2+..... Banka verisinde sonucumuz y alanında saklıdır. 16 tane nitelik olduğundan tek tek yazmak yerine . yazılmıştır., family= Modelde kullanılacak hata dağılımını ve link fonksiyonunu (link function) tanımlar. Binomial ailesi için link fonksiyonu logit* olmaktadır. data =Veri kümesidir. Banka veri kümesinden %80 oranında bölünen training veri kümesi kullanılmıştır. summary fonksiyonu ile oluşturulan model özetlenmektedir.

Bu tabloda;

Call olarak görünen çıktı oluşturduğumuz modeli hatırlatma niteliğindedir. Model çıktısında Deviance kelimesini iki kez görüyoruz. Türkçe karşılığı sapma olan deviance , genelleştirilmiş bir doğrusal modelin uyum iyiliğinin bir ölçüsüdür. Veya daha ziyade uyumluluğun kötü bir ölçüsüdür - daha yüksek rakamlar daha kötü uyum gösterir.

R, iki sapma şekli - boş sapma (null deviance) ve kalan sapma (residual deviance) - bildirmektedir. Boş sapma, yanıt değişkeninin sadece bağımsız değişkenlerin dahil edildiği rezidüel olarak kalan kesme noktası (büyük ortalama) içeren bir model tarafından ne kadar iyi tahmin edileceğini göstermektedir. Yukarıdaki örneğimizde 3615 derecelik serbestik değerinde 2649.9 luk bir değerimiz var. Bağımsız değişkenler dahil edildiğinde sapma 3573 serbestlik derecesinde 1742.9 a düştü ve önemli bir azalma oldu.Sapma (3615-3573) 42 serbestlik derecesi kaybıyla 907 azaldı. Fisher skorlama algoritması, maksimum olasılık problemlerini sayısal olarak çözmek için Newton'un yönteminin bir türevidir. Örneğimiz için, Fisher'ın Puan Alma Algoritmasının, uygunluğu gerçekleştirmek için altı yineleme gerektiğini görüyoruz. Fisher scoring modelin gerçekten birleştiği ve bunu yapmakta hiç zorlanmadığı dışında bir bilgi vermiyor.

```
predict<- ifelse(predict(model,test,type="response")>0.5,'YES','NO')
head(predict)
```

1	'NO'
7	'NO'
8	'NO'
17	'NO'
19	'NO'
29	'NO'

predict Genel Kullanımı

predict fonksiyonu glm nesnesinin tahmin edilen sonucunu döndürür. predict (object ,newdata, type=c ("link", "response", "terms") Parametrelerin kullanımı: object = glm sınıfından glm fonksiyonu tarafından oluşturulan nesnedir. Banka veri kümesi için glm fonksiyonundan üretilen nesnemiz model idi. newdata: Tahmin ediciler için gerekli olan değerleri içeren veri çerçevesidir. Kullanılacak newdata %20 oranında bölme işlemi yapılan test veri kümesidir. type Tahmin edilen değerlerin türünü içerir. link, response, terms olarak 3 çeşidi vardır. type="response" tahmin edilme olasılığını verir. 6 örnek için bakıldığında; tahmin ediciler kullanılarak ; 1, 7, 8, 17, 19, 29 numaralı müşterilerin ürünü satın almayacağı tahmin edilmiştir. Peki gerçekte de durum bu şekilde mi?

```
df<- data.frame(data)
df[19,]
df[29,]
```

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
19	25	blue-collar	single	primary	no	-221	yes	no	unknown	23	may	250	1	-1	0	unknown	no

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
29	56	self-employed	married	secondary	no	784	no	yes	cellular	30	jul	149	2	-1	0	unknown	no

19, 29 numaralı müşterilere bakıldığında tahminlerin doğru olduğu gözükmektedir.

Lojistik regesyon modelinin performansı

Bir lojistik regresyon modelinin performansını değerlendirmek için herhangi bir programlama dilinde (SAS,R,Phyton) bazı noktalara dikkat etmek gereklidir.

AIC (Akaike Bilgi Kriterleri): Lojistik regresyonda ayarlanmış R^2'nin benzer ölçüsü AIC'dir. AIC, model katsayılarının sayısı için modeli cezalandıran uygun ölçüttür. Bu nedenle, her zaman minimum AIC değerli modeli tercih edilmelidir.

Null Deviance: sadece bir kesme noktası olan bir model tarafından öngörülen yanıtı,Residual Deviance Artık sapma, bağımsız değişken ekleme üzerine bir model tarafından öngörülen yanıtı gösterir.Değerleri azaltmak modeli daha iyi kullanmaya yardımcı olacaktır.

Confusion Matrix: Gerçek-Öngörülen değerlerin tablolaştırma biçimidir.. Bu, modelin doğruluğunu bulmamıza ve gereğinden fazla uyulmasına yardımcı olur.

ROC Eğrisi: Alıcı İşlemsel Karakteristiği (ROC), gerçek pozitif oran (hassasiyet) ve yanlış pozitif oran (1 özgüllük) arasındaki dengeleri değerlendirerek modelin performansını özetler. ROC'yi çizmek için, p> 0.5 varsayılır, çünkü başarı oranından daha çok endişelenilir. ROC, p> 0.5 olası tüm değerler için tahmini gücü özetler. Eğri altındaki alan (AUC), doğruluk indeksi (A) veya uyum endeksi olarak anılır, ROC eğrisi için mükemmel bir performans metriğidir. Eğri altındaki alanı daha yüksek hale getirilirse , modelin tahmin gücü daha iyi olur.

```
confusion_matrix<-table(predict,test$y)
confusion_matrix
```

```
predict  no yes
    NO  771  72
    YES  24  38
```

- ✓ Test veri kümeside toplamda 843 müşterinin mevduatı satın almadığı (no) bilinirken; bunların 771 tanesi satın almaz (no), 72 tanesi de satın alır (yes) şeklinde tahmin edilmiştir. (1.satır)
- ✓ Training veri kümesinde toplamda 62 müşterinin mevduatı satın aldığı bilinirken; bunların 24 tanesi satın almaz (no), 38 tanesi de satın alır (yes) şeklinde tahmin edilmiştir. (2.satır)
- ✓ 905 örneklemli veri kümesinden 771+38=809 örnek doğru tahmin edilerek başarı oranı %89.39 olarak bulunmuştur.

TARTIŞMA VE SONUÇ

Banka veri setine Karar Ağaçları ve Lojistik Regresyon algoritmaları uygulanarak alınan sonuçları karşılaaştırılsak; her iki algoritmada da veri setinin %80'i model oluşturmada kullanılırken; %20'si test aşamasında kullanılmıştır. Confusion matrix ile çıkan tahminlerin doğruluk oranı hesaplandığında %89.17 doğruluk oranı ile karar ağaçları algoritması, %89.39 doğruluk oranı ile lojistik regresyon algoritması bulunmuştur. Görüldüğü üzere birbirine oldukça yakın bir değerdedir.

KAYNAKLAR

Tezde kaynak gösterimi yazar-tarih sistemine göre olmalıdır.

Kaynaklar metin içinde (Yazar Soyadı, yıl) şeklinde gösterilmelidir.

Kaynak listesinde ise alfabetik sırada olmalıdır. Basılı dergide makale gösterimi:

Yazar Soyadı İsim baş harfi/harfleri, (Yıl). Makale adı. *Dergi adı kısaltması (İtalik)*; **Cilt no (Koyu)** (Sayı no zorunlu değil): sayfa no-sayfa no.

Örnek olarak;

Hoogstraal H (1985). Argasid and Nuttalliellid as Parasites and Vectors. *Adv Parasit*; **24:** 135-238.

Melikoğlu G, Bitiş L, Meriçli AH (2004). Flavonoids of *Crataegus microphylla*. *Nat Prod Res* 2004; **18**: 211-213.

Diğer kaynakların kaynak listesinde gösterimi için “Bitirme Projesi Tez Yazım Klavuzu” sayfa 8’e bakılabilir.

EKLER