

Karar ağacı ile Cox karma modeli ve lastik verileri üzerine bir uygulama

Hatice ÖNCEL ÇEKİM

Hacettepe Üniversitesi,
Fen Fakültesi,
İstatistik Bölümü,
06800-Beytepe, Ankara, Türkiye
oncelhatice@hacettepe.edu.tr

Durdu KARASOY

Hacettepe Üniversitesi,
Fen Fakültesi,
İstatistik Bölümü,
06800-Beytepe, Ankara, Türkiye
durdu@hacettepe.edu.tr

Özet

Cox regresyon yönteminde temel amaç verilerinin genel durumunu yansıtacak bir model oluşturmaktır. Bu modelle, yaşam süresi üzerinde etkili olduğu düşünülen bağımsız değişkenlerin etkileri ölçülebilmektedir. Ancak orantılı tehlikeler varsayımı sağlanmadığı durumlarda Cox regresyon modeli uygun bir model değildir. Bu nedenle varsayımın sağlanmadığı durumlar için Cox regresyonu ile ağaç yapılı modelleme birleştirilerek karma bir model oluşturulmuştur. Bu karma model; karar ağacı yöntemi ile elde edilen düğümlerin, adım fonksiyonları olarak Cox regresyon modeline eklenmesiyle elde edilmektedir. Bu çalışmada karar ağacı ile Cox regresyon modeli, Kangal Kömür İşletmesi'nde kullanılan makinelere ait lastik verileri için ilk defa kullanılarak lastiklerin ömründe etkili olduğu düşünülen değişkenler belirlenmiştir. Orantılı tehlikeler varsayımını sağlamayan bu veriler için karar ağacı ile Cox karma modelin klasik Cox modelden daha iyi olduğu görülmüştür.

Anahtar sözcükler: Cox regresyon; Karar ağacı; Yaşam çözümlemesi; Veri madenciliği; Makine lastikleri.

Abstract

Cox hybrid model with decision tree and an application on data of tires

The main purpose in Cox regression method is to create a model to reflect the general status of the data. The effects of the independent variables having an effect on survival time can be measured with this model. However, when the assumption of proportional hazards is not satisfied, Cox regression model is not an appropriate model. Therefore, a hybrid model combining Cox regression with a tree-structured model is created for situations in which assumptions were not satisfied. This mixed model consists of adding to the Cox regression model as a step function of the nodes obtained by decision tree method. In this study, Cox hybrid model with decision tree is used for the first time in data of tires at Kangal Coal Plant in order to determine the variables which are thought to be effective on the survival time of tires. For this data, which do not satisfy proportional hazards assumption, the Cox hybrid model with decision tree is seen to be better than the classical Cox regression model.

Keywords: Cox regression; Decision tree; Survival analysis; Data mining; Machinery tires.

1. Giriş

Yaşam çözümlemesi tıp, sosyoloji, mühendislik, ekonomi gibi birçok alanda uygulanmaktadır. Genellikle araştırmaya konu olan birimlerin yaşam süreleri üzerinde etkili olduğu düşünülen faktörleri belirlemek

amacıyla kullanılmaktadır. Yaşam çözümlenmesinde temel modelleme yöntemlerinden biri Cox regresyon modelidir. Bu model adını, modeli öneren David R. Cox'dan (1972) almıştır [4].

Cox regresyon modeli, çalışmaya konu olan bir birimin yaşam süresi (bağımlı değişken) ile birden fazla bağımsız değişken arasındaki ilişkiyi ortaya çıkaran bir modeldir. Modelin temel varsayımı, yaşam süresi üzerinde etkili olduğu düşünülen değişkenlere ait tehlike oranının zaman boyunca sabit olmasıdır. Bu varsayım, orantılı tehlikeler varsayımı olarak ifade edilir. Orantılılık varsayımı sağlanmadığı durumda bu modeli kullanmak uygun değildir [10].

Birçok çalışmanın temel amacı, veriye eklenen yeni değişkenlerin sınıflandırılmasında kullanılabilecek güvenilir karar kuralları geliştirmektir. Bu tür sınıflama problemlerinde geleneksel istatistiksel yöntemler sınırlı fayda sağlamaktadır. Bu nedenle sınıflama ve regresyon karar ağacı (Classification and Regression Trees; C&RT) yönteminin kullanımı artmıştır. C&RT yöntemi geleneksel ağaç oluşturma yöntemlerine benzememektedir. Çünkü C&RT yöntemi, diğer yöntemlere göre daha az adım ve dallanmalar ile analizi gerçekleştirmektedir. Bu avantajlarından dolayı son yıllarda C&RT yöntemleri birçok araştırmacı tarafından yaşam çözümlenmesinde de kullanılmaktadır [3].

Bu çalışmada orantılı tehlikeler varsayımı sağlanmadığı durumlarda kullanılabilen, klasik Cox regresyon modeli ile karar ağacı modellemesi birleştirilerek karma bir modelle çalışılmıştır. Birçok açıdan incelendiğinde Cox regresyon modeli ve ağaç yapılı modelleme birbirini tamamlamaktadır. Cox regresyon modeli basit bir doğrusal etkiyi modellemek için sıkça tercih edilen bir yöntemdir. Ancak doğrusal olmayan etkiler ya da kategorik değişkenlerin düzeyleri arasında ilişkiler olduğunda problemlerle karşılaşılabilir. Diğer taraftan karar ağacı yöntemleri ise bu durumlarda oldukça kullanışlıdır. Hatta ağaç temelli yöntemlerde kategorik verilerle ilgilenildiğinde büyük bir etkinlik gösterir. Cox regresyon modeli ise birçok kategorik tahmin ediciye ya da bu tahmin ediciler çok düzeye sahip olduğunda büyük bir model formuna sahiptir. Bu nedenle, çalışmamızda karar ağacı ile Cox karma modelini kullanmak ve en iyi modeli seçmek amaçlanmıştır.

2. Cox regresyon modeli

F_i ve C_i başarısızlık zamanları olsun. $X_i = \min(F_i, C_i)$, i . gözlenen başarısızlık zamanı, $\delta_i = 1_{\{F_i \leq C_i\}}$ ise i . durumda başarısızlık ya da durdurma olup olmadığını gösterir. $z_i = (z_{i1}, z_{i2}, \dots, z_{ip}) \in \mathcal{R}^p$, i . durum için p boyutlu bağımsız değişken vektörü olmak üzere $\{(X_i, \delta_i, z_i): i = 1, 2, \dots, n\}$ gözlemlenen verileri içerir. Verilen z_i vektöründe bağımsız olan F_i ve C_i 'nin bilgi içermediğini varsayalım. $\lambda_0(t)$, temel tehlike fonksiyonu ve $g(z)$, z 'nin bir fonksiyonu olmak üzere, Cox regresyon modeli;

$$\lambda(t|z) = \lambda_0(t) \exp\{g(z)\} \quad (1)$$

biçiminde yazılabilir. $g(\cdot)$ çoğunlukla seçilmiş bileşenlerin bir doğrusal kombinasyonu olarak belirlenir. $z^0 \in \mathcal{R}^q$, z 'den seçilen q boyutlu vektör ve $\beta = (\beta_1, \beta_2, \dots, \beta_q) \in \mathcal{R}^p$ bilinmeyen regresyon parametrelerinin bir vektörü olmak üzere Cox regresyon modeli

$$\lambda(t|z) = \lambda_0(t) \exp\{\beta' z^0\} \quad (2)$$

biçimindedir [9].

3. Karar ağacı

Birçok çalışmada elde edilen problemle ilgili en iyi kararı verebilmek için bazı işlemleri yerine getirirken bir takım yöntemlere ihtiyaç duyulur. Bu yöntemler arasında kullanım açısından karar vericiye problemi anlamada kolaylık sağlayan, güvenilirliği yüksek olan ve sınıflama modelleri içinde son zamanlarda en yaygın olarak kullanılan karar ağacı yöntemidir.

Karar ağacı yöntemi pek çok test gerçekleştirerek, hedefi tahmin etmede en iyi sırayı bulmaya çalışır. Her bir test karar ağacındaki dalları oluşturur ve bu dallar da diğer testlerin gerçekleşmesine neden olur. Bu durum, test işleminin bir son düğümünde (terminal node) sonlanmasına kadar devam eder.

Karar ağacının iki temel işlemi olan bölme (splitting) ve budama (pruning) işlemlerinin sonlanması uygulanan durdurma kriterine göre olmaktadır. Bu işlemler ve durdurma kriteri kısaca şöyle açıklanabilir [2]:

- Bölme: Bu işlem, verilerin daha küçük alt kümelerine ayrılmasını sağlayan tekrarlı bir süreçtir. İlk tekrar tüm veriyi içeren kök düğüm ile başlar. Bundan sonraki tekrarlar, verinin alt kümelerini içeren üretilmiş düğümler üzerinde işlem yapmaktadır. Her bölme işleminde, değişkenler analiz edilir ve en iyi bölme seçilir.
- Budama: Bir ağaç oluşturulduktan sonra, istenmeyen alt ağaçlar veya düğümler bulunabilir. Budama işlemi ile bunlar çıkarılarak karar ağacı daha genel bir biçimde ifade edilebilmektedir.
- Durdurma kriteri: Ağaç oluşturma algoritmaları çeşitli durdurma kuralları içermektedir. Bu kurallar genellikle, maksimum ağaç derinliği, bir düğümde bölme için ele alınan minimum eleman sayısı ve yeni bir düğümde olması gereken minimum eleman sayısı gibi çeşitli faktörlere dayanır.

Karar ağacıyla oluşturulan modellerin güvenilirliğini test etmek için veriler test etme - öğrenme (testing-training) olarak ayrılır. Bu nedenle eldeki verilerin bir kısmı öğrenme amaçlı kullanılır. Karar ağacı öğrenme kısmı için kullanılan verilerden oluşur. Verilerin diğer bir kısmı ise oluşturulan karar ağacını test etmek için kullanılır [6].

Karar ağaçlarında kullanılan modellerden biri olan C&RT veya CART Breiman, Friedman, Olshen ve Stone tarafından 1984 yılında geliştirilmiştir [3]. C&RT sınıflandırma ve regresyon analizi için kullanılan bir yöntemdir. Son zamanlarda gelişen teknoloji ile uygulanabilirliği kolay hale geldiği için sınıflandırma için kullanılan diğer yöntemlerden daha çok kullanılmaya başlanmıştır.

C&RT, Gini katsayısı aracılığıyla ikili (binary) ağaç olarak büyüyen bir algoritmadır. Budama işlemi ağacın karmaşıklık ölçüsüne dayanır. Karar ağaçları içinde bağımlı ve bağımsız değişkenlerin sürekli olduğu durumlarda kullanılan bir modelidir [7].

4. Karar ağacı ile Cox karma modeli

Cox regresyon modeli literatürde sık kullanılan yöntemlerden biri iken karar ağacı ile Cox karma modelinin kullanımı oldukça azdır. Karar ağacı ile Cox karma modelini kullanan çalışmalardan ikisi Su ve Tsai (2005) ve Crocetti, Mangone, Scocco ve Carli'nin (2006) çalışmalarıdır.

Su ve Tsai (2005) çalışmasında 2047 sağlıklı erkek hastanın kanserden ölüm riskini etkileyen değişkenlerle oluşturduğu karar ağacı ile Cox karma modellerini karşılaştırarak kanserle ölüm riskine ait yeni görüşler önermiştir [9].

Crocetti, Mangone, Scocco ve Carli (2006) çalışmasında kötü huylu tümöre sahip İtalyan hastalar üzerinden bu hastalığa yakalanabilecek kişilerin özelliklerini belirlemek için Cox regresyon modelini ve C&RT yöntemini kullanmıştır. Sonuçlardan Cox regresyon modeli ve C&RT yönteminin hastaların prognostik özelliklerini benzer şekilde belirlediği görülmüştür [5].

Eşitlik (2)'nin uygun ve seçilen tekniklerle klasik Cox modelini bulmak için kullanılan en iyi model olduğunu varsayalım. Bu modelin yeterliliğini kontrol etmek ve uygunluğunu geliştirmek için, Eşitlik (2)'nin karar ağacı ile genişletilmiş Cox regresyon yapısı kullanılır. Bu durumda aşağıdaki karma model oluşturulur:

$$\lambda(t|z) = \lambda_0(t) \exp\{\beta'z^0 + \gamma'z^{(T)}\}. \quad (3)$$

Burada $z^{(T)}$, eğer i . gözlem T 'nin j . son düğümüne bağlı ise $z_{ij}^{(T)} = 1$ aksi durumda $z_{ij}^{(T)} = 0$ olmak üzere bir T ağaç yapısı tarafından tanımlanan göstermelik (dummy) vektördür. γ ise benzer parametre vektörüdür. T , Eşitlik (2) için $g(z) - \beta'z^0$ kalan kısmına göre parçalı sabit adım fonksiyonunda kullanan bir yaklaşımı sağlamayı amaçlayan karar ağacı ile genişletilmiş Cox regresyon olarak anılabilir. T ağaç yapısının genişletilmesini vermek için \tilde{T} 'nin, T 'de bütün son düğümlerin kümesini gösterdiğini varsayalım. $\gamma = (\gamma_1, \gamma_2, \dots, \gamma_{|\tilde{T}|})' \in \mathcal{R}^{|\tilde{T}|}$ olmak üzere $\sum_{j=1}^{|\tilde{T}|} \gamma_j = 0$ tanımlanabilirliğine bağlıdır. Eşitlik (3)'teki yeni modelin formu parametrik bir modelin formu kadar basit olmasına rağmen yarı parametrik bir forma sahiptir [9].

5. Uygulama

Ülkemizde madencilik ve inşaat sektörlerinde yaşam çözümlemesi uygulamaları pek yapılmamaktadır. Her alanda, herhangi bir nedenle başarısızlık ve başarısızlık süresi olduğu durumda yaşam çözümlemesi yapılabilmektedir. Bu çalışmada, Demir Export A.Ş. tarafından Kangal Kömür İşletmesi'nde kullanılan makinelere ait lastik verileri kullanılarak lastiklerin ömründe etkili olduğu düşünülen Çizelge 1'deki değişkenlerin etkileri ölçülmüştür [8].

Maden, yol ve baraj inşaatı, taş, mermer ve kum ocağı ile hafriyat gibi ağır hizmet şartlarında çalışan iş makinelerinde alışı gelinen otomobil lastiklerinden farklı tasarıma sahip yol dışı lastikler (Off The Road Tires) kullanılmaktadır. Bu çalışmadaki Kangal Kömür İşletmesi'nde kullanılan makinelere ait lastikler de yol dışı lastik özelliğine sahiptir [11].

Çizelge 1. Kullanılan Değişkenler ve Düzeyleri

Değişken	Değişken Düzeyleri	N	%
Marka	A	319	70.58
	B	80	17.70
	C	53	11.72
Pozisyon	Sol ön	78	17.25
	Sol arka iç	71	15.71
	Sol arka dış	70	15.49
	Sağ ön	84	18.58
	Sağ arka iç	79	17.48
	Sağ arka dış	70	15.49
Ortam	Dekapaj	282	62.39
	Kömür	98	21.68
	Dekapaj + Kömür	72	15.93
Diş derinliği		452	

Uygulamada kullanılan 452 lastikten 120'si çeşitli nedenlerle ömrünü tamamlamış ve hurdaya ayrılmıştır. Lastikler üç değişik markadan (A, B, C) oluşmaktadır. Lastiklerin çalışma ortamı dekapaj, kömür nakli ve dekapaj + kömür nakli olmak üzere üçe ayrılmıştır. Lastik iş makinelerine altı değişik pozisyonda (sol ön, sol arka iç, sol arka dış, sağ ön, sağ arka iç, sağ arka dış) takılabilmektedir.

Çalışmada başarısızlık, lastiğin hurdaya ayrılması olarak tanımlanmıştır. Bir lastiğin kullanımına başlandığı andan hurdaya ayrılıncaya kadar geçen süre (saat olarak) lastiğin başarısızlık süresi olarak alınmıştır. Halen kullanımı devam eden lastikler ise durdurulmuş olarak tanımlanmıştır.

Çizelge 2. Kaplan-Meier Yaşam Olasılıkları

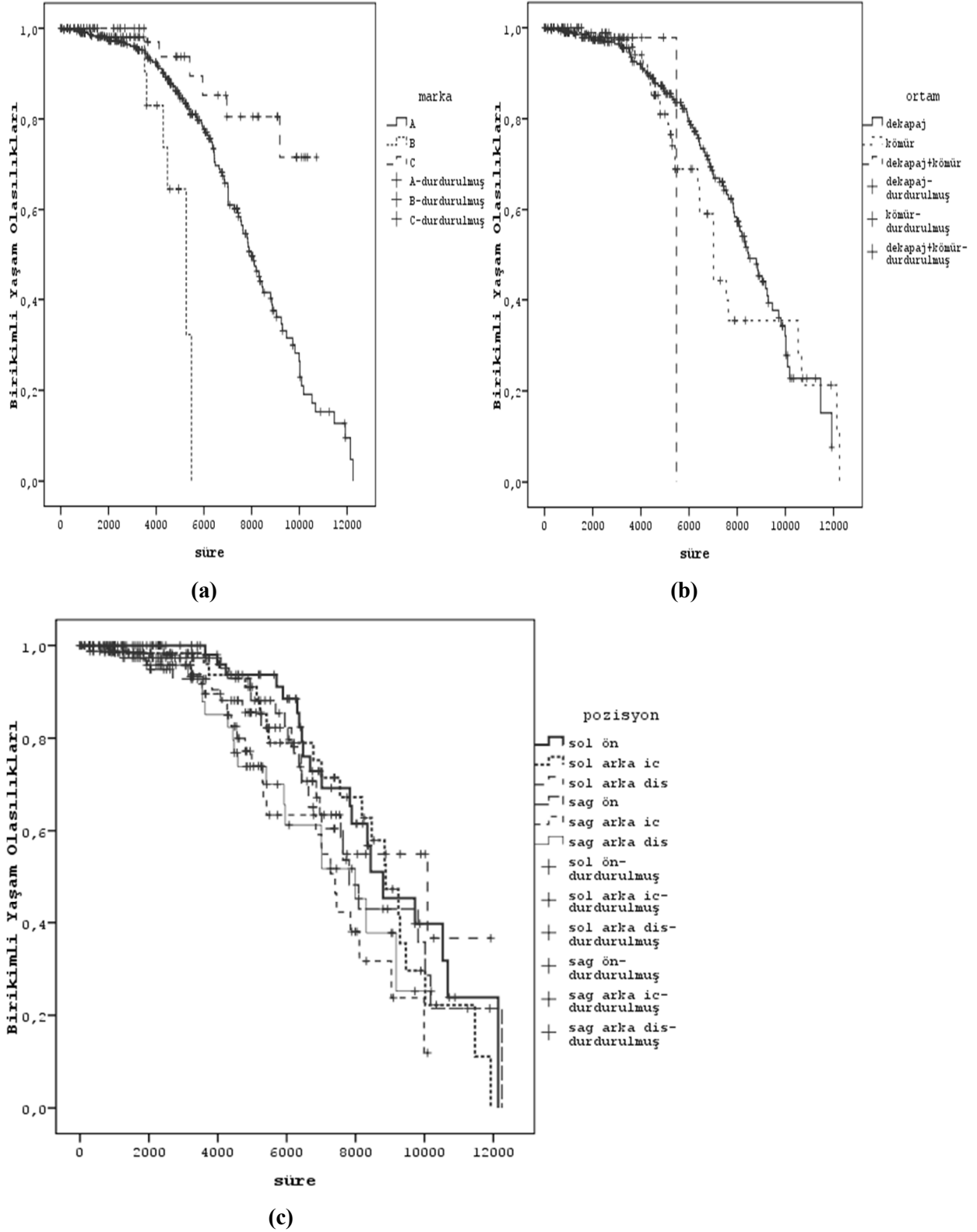
Değişken	Değişken Düzeyleri	5000. saat için Kaplan-Meier Yaşam Olasılığı	10000. saat için Kaplan-Meier Yaşam Olasılığı
Marka	A	0.845	0.264
	B	0.645	0.000
	C	0.937	0.715
Ortam	Dekapaj	0.855	0.321
	Kömür	0.810	0.354
	Dekapaj + Kömür	0.978	0.000
Pozisyon	Sol ön	0.937	0.398
	Sol arka iç	0.910	0.222
	Sol arka dış	0.855	0.550
	Sağ ön	0.881	0.286
	Sağ arka iç	0.733	0.118
	Sağ arka dış	0.739	0.252

Çalışmada kullanılan değişkenler için Şekil 1’de Kaplan-Meier yaşam olasılıkları grafikleri verilmiştir. Çizelge 2’de ise 5000. ve 10000. saatler için Kaplan-Meier yaşam olasılıkları verilmiştir. Şekil 1 ve Çizelge 2 incelendiğinde C marka lastiklerin yaşam olasılığının en fazla olduğu görülmektedir. Ayrıca 5000. saatte kömür nakli ortamında çalışan lastiklerin yaşam olasılıkları en az iken 10000. saatte yaşam olasılıkları en fazladır. Pozisyon değişkeni için yaşam olasılıkları incelendiğinde en fazla yaşam olasılığına 5000. saatte sol ön lastikler sahip iken 10000. saatte ise sol arka dış lastiklerin sahip olduğu görülmektedir.

Lastiklerin yaşam olasılıklarının değişken düzeyleri arasında farklı olup olmadığını incelemek için Log-Rank ve Tarone-Ware test istatistikleri kullanılmıştır ve sonuçları Çizelge 3’te verilmiştir. 0.05 anlamlılık düzeyinde iki testin de sonuçlarına göre marka değişkeninin düzeyleri arasında lastiklerin yaşam olasılıkları bakımından fark vardır. Tarone-Ware testi sonuçlarına göre pozisyon değişkeninin düzeyleri arasında da lastiklerin yaşam olasılıkları bakımından fark vardır. Bu fark Tarone-Ware testine göre, marka değişkeninin tüm düzeyleri arasında iken pozisyon değişkeninin de ise sağ arka iç ile sol ön, sağ arka dış ile sol ön ve sağ arka iç ile sol arka iç arasında bulunmaktadır. Ortam değişkeninin düzeyleri arasında ise her iki teste göre de lastiklerin yaşam olasılıkları bakımından fark yoktur.

Çizelge 3. Log Rank ve Tarone-Ware Test Sonuçları

Değişken	Log Rank Testi p-değeri	Tarone-Ware Testi p-değeri
Marka	0.000	0.002
Pozisyon	0.081	0.049
Ortam	0.540	0.350



Şekil 1. Marka (a), ortam (b) ve pozisyon (c) değişkenlerinin Kaplan-Meier yaşam olasılıkları grafikleri

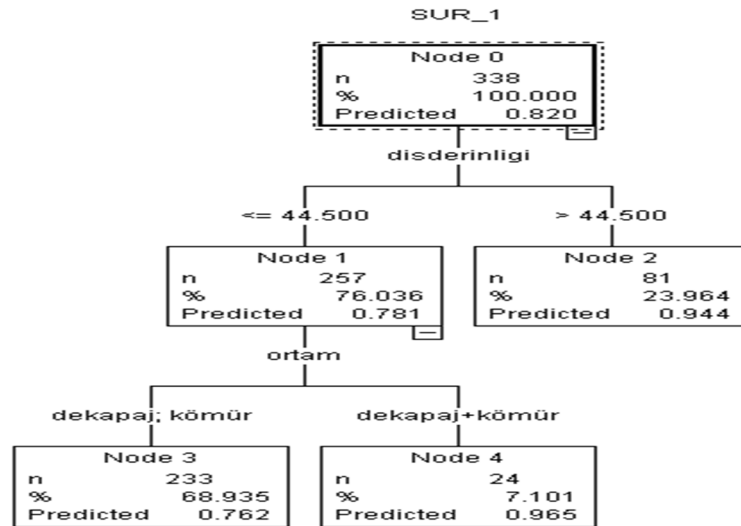
Cox regresyon modeli için orantılı tehlikeler varsayımının sağlanması gerekmektedir. Orantılı tehlikeler varsayımı yaşam süresinin rankı ile Schoenfeld artıkları arasındaki ilişki testi ile incelendiğinde pozisyon değişkeninin ($p < 0.05$) ve log-log yaşam eğrileri yöntemi kullanılarak incelendiğinde marka, pozisyon ve ortam değişkenlerinin orantılı tehlikeler varsayımını sağlamadığı görülmüştür.

Çizelge 4. Cox regresyon modelinin sonuçları

Değişken	Tahmin	Standart Hata	P-değeri	Tehlike Oranı
Marka B	2.016	0.498	0.000*	7.505
Marka C	-1.119	0.430	0.009*	0.327
Sol arka iç	1.125	0.358	0.002*	3.079
Sol arka dış	0.850	0.377	0.024*	2.339
Sağ ön	0.139	0.314	0.658	1.149
Sağ arka iç	1.649	0.346	0.000*	5.203
Sağ arka dış	1.257	0.356	0.000*	3.515
Kömür	0.374	0.244	0.125	1.454
Dekapaj + Kömür	-2.459	0.861	0.004*	0.086
Diş derinliği	0.074	0.013	0.000*	1.077
Log L = -519.057 AIC=1068.114				

* $\alpha=0.05$ 'te anlamlıdır.

Çizelge 4'teki Cox regresyon modeli sonuçlarına göre diş derinliği arttıkça lastiklerin hurdaya çıkma riski artmaktadır. A markalı lastiklere göre B markalı lastiklerin hurdaya çıkma riski 7.505 kat daha fazladır. Dekapaj+kömür ortamındaki lastiklere göre dekapaj ortamı lastiklerinin hurdaya çıkma riski $1/0.086=11.628$ kat daha fazladır. Benzer biçimde diğer değişkenler için de yorumlar yapılabilir.



Şekil 2. C&RT karar ağacı sonuçları

Clementine 12.0 paket programı ile C&RT karar ağaçları yöntemi kullanılarak lastiklerin yaşam olasılıkları üzerinde etkili olduğu düşünülen değişkenler belirlenmiştir. Analiz sonucunda ortam, marka ve diş derinliği değişkenleri önemli bulunmuştur. Modeldeki önem yüzdeleri ise sırasıyla %34, %22.5 ve %43.5 olarak elde edilmiştir.

Şekil 2'deki sonuçlara göre Düğüm 2 (Node 2) ve Düğüm 4 (Node 4) düğümleri göstermelik değişken olarak eklenebilir. Öncelikle Cox regresyon modeline, Düğüm 2 değişkeni eklenerek analiz yapılırsa

$$-2 (\log L_{Cox} - \log L_{Cox+Düğüm 2}) = 14.61 > \chi_1 = 3.84$$

olduğundan Düğüm 2 önemli bir değişkendir. Bu nedenle marka, ortam, pozisyon, diş derinliği ve Düğüm 2 değişkenlerinden oluşan model anlamlıdır. Analiz sonuçları Çizelge 5'te görülmektedir.

Çizelge 5. Karar ağacı ile Cox karma model 1'in sonuçları

Değişken	Tahmin	Standart Hata	P-değeri	Tehlike Oranı
Marka B	2.283	0.508	0.000*	9.809
Marka C	-1.201	0.431	0.005*	0.301
Sol arka iç	1.140	0.362	0.002*	3.125
Sol arka dış	0.833	0.379	0.028*	2.300
Sağ ön	0.268	0.318	0.399	1.307
Sağ arka iç	1.614	0.345	0.000*	5.024
Sağ arka dış	1.369	0.363	0.000*	3.933
Kömür	0.408	0.244	0.094	1.503
Dekapaj + Kömür	-2.975	0.890	0.001*	0.051
Dış derinliği	0.053	0.014	0.000*	1.054
Düğüm 2	1.657	0.409	0.000*	5.242
Log L = -511.752 AIC=1056.504				

* $\alpha=0.05$ 'te anlamlıdır.

Dış derinliği 44.5'ten küçük ve eşit olanlara göre 44.5'ten büyük olan lastiklerin hurdaya çıkma riski 5.242 kat daha fazladır. Ayrıca diğer değişkenler klasik Cox regresyon modeline benzer şekilde yorumlanabilmektedir.

Cox regresyon modeline, Düğüm 4 değişkeni eklendiğinde

$-2(\log L_{Cox} - \log L_{Cox+Düğüm\ 4}) = -2.254 < \chi_1 = 3.84$ olduğundan Düğüm 4'ün önemli bir değişken olmadığı görülmüştür. Ancak Cox regresyon modeline, Düğüm 2 ve Düğüm 4 değişkenleri birlikte eklendiğinde model anlamlı çıkmıştır:

$$-2(\log L_{Cox} - \log L_{Cox+Düğüm\ 2+Düğüm\ 4}) = 11.942 > \chi_1 = 3.84.$$

Çizelge 6. Karar ağacı ile Cox karma modeli 2 sonuçları

Değişken	Tahmin	Standart Hata	P-değeri	Tehlike Oranı
Marka B	2.253	0.508	0.000*	9.512
Marka C	-1.274	0.428	0.003*	0.280
Sol arka iç	0.984	0.345	0.004*	2.674
Sol arka dış	0.690	0.365	0.059	1.994
Sağ ön	0.243	0.317	0.444	1.275
Sağ arka iç	1.477	0.329	0.000*	4.379
Sağ arka dış	1.252	0.352	0.000*	3.498
Dış derinliği	0.050	0.014	0.000*	1.052
Düğüm 2	1.629	0.409	0.000*	5.098
Düğüm 4	-3.019	0.890	0.001*	0.049
Log L = -513.086 AIC=1056.172				

* $\alpha=0.05$ 'te anlamlıdır.

Karar ağacı ile Cox karma model 2 sonuçları Çizelge 6'da verilmiştir.

Yaşam çözümlenmesinde modeller arasından en uygun modeli seçmek için kullanılan yöntemlerden biri Akaike bilgi kriteridir (AIC). Bu yöntemle en küçük AIC değerine sahip model en uygun model olarak belirlenmektedir. Yarı parametrik yaşam modeli için Akaike bilgi kriteri,

$$AIC = -2\log L + \alpha p$$

biçiminde hesaplanmaktadır. Burada p modeldeki bilinmeyen parametrelerinin sayısı ve α daha önceden belirlenen bir sabittir. α genellikle 2 ile 6 arasında değerler alır. $\alpha = 3$ olarak seçilmesi, $-2\log L$ değerleri arasındaki farklılığı değerlendirmede %5 önem düzeyinin kullanılmasına eşdeğer olduğundan α 'nın genel kullanımı için bu değer önerilmektedir [1].

Kullanılan iki karma modelden en uygun olanının karar ağacı ile Cox karma model 2 olduğu AIC değerleri karşılaştırıldığında görülmektedir.

Marka, pozisyon, dış derinliği, Düğüm 2 ve Düğüm 4 değişkenlerinden oluşan modele adimsal regresyon yöntemi uygulandığında Çizelge 7'deki sonuçlar elde edilmiştir. Bu sonuçlara göre, A markalı lastiklere göre B markalı lastiklerin hurdaya çıkma riski 10.319 kat daha fazladır. Ayrıca C markalı lastiklere göre A markalı lastiklerin hurdaya çıkma riski $1/0.281=3.559$ kat daha fazladır. Pozisyonu sol ön olan lastiklere göre sol arka iç lastiklerin 1.962 kat, sağ arka iç lastiklerin 3.251 kat ve sağ arka dış lastiklerin 2.588 kat hurdaya çıkma riski fazladır. Dış derinliği 44.5'ten küçük ve eşit lastiklere göre 44.5'ten büyük olan lastiklerin hurdaya çıkma riski 5.099 kat daha fazladır. Ayrıca dekapaj+kömür ortamında kullanılan lastiklere göre ayrı ayrı dekapaj ve kömür ortamında kullanılan lastiklerin hurdaya çıkma riski $1/0.049=20.41$ kat daha fazladır.

Çizelge 7. Karar ağacı ile Cox karma modeli 2'nin adimsal yöntem sonuçları

Değişken	Tahmin	Standart Hata	P-değeri	Tehlike Oranı
Marka B	2.334	0.507	0.000*	10.319
Marka C	-1.268	0.428	0.003*	0.281
Sol arka iç	0.674	0.280	0.016*	1.962
Sağ arka iç	1.179	0.265	0.000*	3.251
Sağ arka dış	0.951	0.292	0.001*	2.588
Dış derinliği	0.044	0.013	0.001*	1.045
Düğüm 2	1.629	0.399	0.000*	5.099
Düğüm 4	-3.019	0.888	0.001*	0.049
Log L = -514.806 AIC=1053.612				

* $\alpha=0.05$ 'te anlamlıdır.

6. Sonuç

Bu çalışmada öncelikle, klasik Cox regresyon modeli ve karar ağacı hakkında bilgiler verilmiş ve karar ağacı ile Cox karma modeli tanıtılmıştır.

Karar ağacı ile Cox karma modelinde ilk defa lastik verileri kullanılarak lastiklerin hurdaya çıkmalarını etkileyen faktörler bulunmuştur.

Uygulamada ilk olarak orantılı tehlikeler varsayımının sağlanıp sağlanmadığına bakılmış ve sağlanmadığı görülmüştür. Bu varsayım sağlanmadığında kullanılacak alternatif yöntemlerden biri karar ağacı ile Cox karma modelidir. Bu model, karar ağacı yöntemi ile oluşturulan yeni değişkenlerin Cox modeline eklenmesiyle oluşturulmuştur. İncelenen karma modeller arasında en uygun modelin karar ağacı ile Cox karma model 2 olduğu görülmüştür. Bu modele adimsal regresyon yöntemi uygulandığında, dış derinliği arttıkça lastiklerin hurdaya çıkma riskinin arttığı, A markalı lastiklere göre B markalı lastiklerin hurdaya çıkma riskinin ve C markalı lastiklere göre A markalı lastiklerin hurdaya çıkma riskinin daha fazla olduğu görülmüştür. Ayrıca, pozisyonu sol ön olan lastiklere göre sol arka iç lastiklerin, sağ arka iç lastiklerin ve sağ arka dış lastiklerin hurdaya çıkma riskinin fazla olduğu, dış derinliği bakımından incelendiğinde, dış derinliği 44.5'ten büyük olan lastiklerin hurdaya çıkma riskinin daha fazla olduğu sonucuna varılmıştır. Dekapaj+kömür ortamında kullanılan lastiklere göre ayrı ayrı dekapaj ve kömür ortamında kullanılan lastiklerin hurdaya çıkma riski de daha fazla bulunmuştur.

Orantılı tehlikeler varsayımı sağlanmadığı durumda Cox regresyon modelini kullanmak doğru değildir. Bu durumda kullanılacak bir model karar ağacı ile Cox karma modelidir.

Kaynaklar

- [1] N. Ata, D. Karasoy ve M.T. Sözer, 2008, Orantısız hazardlar için parametrik ve yarı parametrik yaşam modelleri, *İstatistikçiler Dergisi*, 1 (3), 125-134.
- [2] C. Bounsaythip, R. R. Esa, 2001, Overview of data mining for customer behavior modeling, *VTT Information Technology Research Report*, Version:1, 1-53.
- [3] L. Breiman, J. Friedman, R. Olshen, C. Stone, 1984, *Classification and Regression Trees*, CRC Press, Boca Raton.
- [4] D.R. Cox, 1972, Regression models and life-tables, *Journal of the Royal Statistical Society, Series B*, 34 (2), 187-220.
- [5] E. Crocetti, L.Mangone, G. L. Scocco and P. Carli, 2006, Prognostic variables and prognostic groups for malignant melanoma. The information from Cox and Classification and Regression Trees analysis: an Italian population-based study, *Melanoma Research*, 16 (5), 429-433.
- [6] H. Margaret Dunham, 2003, *Data Mining Introductory and Advanced Topics*, Prentice Hall, New Jersey, 96.
- [7] M. R. Segal, 1988, Regression tress for censored data, *Biometrics*, 44, 35–47.
- [8] D. Sertkaya, 2001, *Yaşam Verilerinin Çözümlemesinde Yarışan Riskler (Competing Risks In The Analysis Of Survival Data)*, Doktora Tezi, Hacettepe Üniversitesi Fen Bilimleri Enstitüsü, 1-3,101.
- [9] X. Su, and C. Tsai, 2005, Tree-augmented Cox proportional hazards models, *Biostatistics*, 6 (3), 486-499.
- [10] T.M. Therneau, P.M. Grambsch, 2000, *Modelling Survival Data: Extending the Cox Model*, Springer, New York.
- [11] O. H. Tokuç, E. Bekiroğlu, K. Kökman, 1999, Yıldışı Lastiklerde Kullanım Süresini Etkileyen Faktörlerin İncelenmesi, *1. İstatistik Kongresi Bildiriler Kitabı*, Antalya, 380.