Journal of Engineering and Natural Sciences
Mühendislik ve Fen Bilimleri Dergisi

Sigma
Vol./Cilt 26
Issue/Sayı 1

Araştırma Makalesi / Research Article

SURVIVAL DATA MINING: AN APPLICATION TO CREDIT CARD HOLDERS

Nihal ATA[*1], Erengül ÖZKÖK[2], Uğur KARABEY[2]

[1]*Hacettepe Üniversitesi, Fen Fakültesi, İstatistik Bölümü, Beytepe-ANKARA*
[2]*Hacettepe Üniversitesi, Fen Fakültesi, Aktüerya Bilimleri Bölümü, Beytepe-ANKARA*

Geliş/Received: 04.09.2007 Kabul/Accepted: 24.03.2008

ABSTRACT

Survival analysis is one of the most used statistical methods in medical research. While death or recurrency of illness called as failure in medical research, in subscription based business, loss of consumer (churn) can be taken as failure. Consumer database includes data about the start and end of the usage period of products or services. This information is used to construct the model in order to gain new consumers and to determine new marketing strategies. Survival models are one of the best methods to examine the consumer database. In this study, it is shown that semi-parametric and parametric methods can be used to analyze consumer database by using information about credit card holders.
Keywords: Data mining, subscription based business, hazard function, survival function, regression models.
MSC number/numarası: 62N01, 62-07, 91B42.

VERİ MADENCİLİĞİ'NDE YAŞAM ÇÖZÜMLEMESİ: KREDİ KARTI SAHİPLERİ İLE İLGİLİ BİR UYGULAMA

ÖZET

Yaşam çözümlemesi, tıbbi araştırmalarda en çok kullanılan istatistiksel yöntemlerden biridir. Tıbbi araştırmalarda ölüm ya da hastalığın tekrar nüksetmesi başarısızlık olarak nitelendirilirken, abone tabanlı işletmelerde müşteri kaybı başarısızlık olarak nitelendirilebilir. Müşteri veri tabanı, hizmet ya da ürünün kullanımına başlanması ya da sona ermesi gibi olaylara ait verileri içermektedir. Bu veriler, yeni müşteriler kazanmak ve pazarlama stratejileri belirlemek amacıyla modeller oluşturmak için kullanılmaktadır. Yaşam çözümlemesi yöntemleri de müşteri veri tabanını incelemek için kullanılabilecek en iyi yöntemlerden biridir. Bu çalışmada kredi kartı sahiplerine ait veriler kullanılarak, müşteri veri tabanını incelemek için yarı parametrik ve parametrik yöntemlerden yararlanılabileceği gösterilmiştir.
Anahtar Sözcükler: Veri madenciliği, abone tabanlı işletmeler, hazard fonksiyonu, yaşam fonksiyonu, regresyon modelleri.

1. GİRİŞ

Bilgisayar teknolojisindeki gelişmelerle birlikte işletmelerde üretilen sayısal bilgi miktarının arttığı, veri tabanlarının daha fazla veriyi saklayabilecek boyutlara ulaştığı ve veriye ulaşmanın kolaylaştığı görülmektedir. Eldeki ham verinin bilgiye ve daha anlamlı biçime dönüştürülmesi

[*]Sorumlu Yazar/Corresponding Autor: e-mail/e-ileti: nihalata@hacettepe.edu.tr, tel: (312) 299 20 16 / 121

için "Veri Madenciliği" kullanılmaktadır. Bu çalışmada veri madenciliğinde yaşam çözümlemesi yöntemlerinin kullanımı incelenmiştir.

2. VERİ MADENCİLİĞİ

Veri madenciliği, veri ambarlarında tutulan çok çesitli verilere dayanarak daha önce keşfedilmemiş bilgileri ortaya çıkarma, bu bilgileri kullanarak karar verme ve eylem planını gerçekleştirme sürecidir[1]. Veri miktarının ve çeşitliliğinin artması, analizlerin eyleme yönelik ve anlamlı sonuçlar verecek şekilde yapılmasını gerektirmektedir. Rekabetin yoğunlaşması, değişim ve uyum sürecinin gerektirdiği hızı yakalayabilmenin, müşteri odaklı olmanın ve verimliliğin önemini her zamankinden daha çok arttırmıştır[2]. Bu süreçte istatistik, matematik, bilgisayar teknolojileri gibi birçok bilim dalından yararlanılmaktadır.

Günümüzde veri madenciliğinin başlıca ilgi alanları pazarlama, bankacılık, sigortacılık, borsa, telekomünikasyon ve endüstridir. Veri madenciliği, pazarlama sektöründe, müşteri segmentasyonunda, müşterilerin demografik özellikleri arasındaki bağlantıların kurulmasında, çeşitli pazarlama kampanyalarında, mevcut müşterilerin elde tutulması için geliştirilecek pazarlama stratejilerinin oluşturulmasında, pazar sepeti analizinde, çapraz satış analizlerinde, müşteri ilişkileri yönetiminde, satış tahminlerinde kullanılmaktadır. Bankacılık sektöründe farklı finansal göstergeler arasındaki gizli korelasyonların bulunmasında, kredi kartı dolandırıcılıklarının tespitinde, kredi taleplerinin değerlendirilmesinde, usulsüzlük tespitinde, risk analizleri ve risk yönetiminde kullanılmaktadır. Sigortacılık sektöründe yeni poliçe talep edecek müşterilerin tahmin edilmesinde, sigorta dolandırıcılıklarının tespitinde, riskli müşteri tipinin belirlenmesinde ve perakendecilik sektöründe satış noktası veri analizleri, alışveriş sepeti analizleri, tedarik ve mağaza yerleşim optimizasyonunda; borsa sektöründe hisse senedi fiyat tahmininde, genel piyasa analizlerinde, alım-satım stratejilerinin optimizasyonunda; telekomünikasyon sektöründe kalite ve iyileştirme analizlerinde, hisse tespitlerinde, hatların yoğunluk tahminlerinde; endüstri sektöründe kalite kontrol analizlerinde, üretim süreçlerinin optimizisyonunda kullanılmaktadır[3].

Veri madenciliği kendi başına bir çözüm değil çözüme ulaşmak için verilecek karar sürecini destekleyen, problemi çözmek için gerekli bilgileri sağlamaya yarayan bir araçtır. Araştırmacının veriler arasındaki ilişkileri bulmasına yardımcı olmaktadır.

2.1. Veri Madenciliği Modelleri

Veri madenciliğinde kullanılan modeller tahmin edici ve tanımlayıcı olmak üzere iki gruba ayrılabilir. Tahmin edici modellerde, sonuçları bilinen verilerden hareket edilerek bir model geliştirilmesi ve kurulan bu modelden yararlanılarak sonuçları bilinmeyen veri kümeleri için sonuç değerlerin tahmin edilmesi amaçlanmaktadır. Tanımlayıcı modellerde ise karar vermeye rehberlik etmede kullanılabilecek mevcut verilerdeki örüntülerin tanımlanması sağlanmaktadır.

Gerek tanımlayıcı gerekse tahmin edici modellerde yoğun olarak kullanılan belli başlı istatistiksel yöntemler; sınıflama ve regresyon, kümeleme, birliktelik kuralları, ardışık zamanlı örüntüler, bellek tabanlı yöntemler, yapay sinir ağları ve karar ağaçları biçiminde sıralanmaktadır. Sınıflama ve regresyon modelleri tahmin edici, kümeleme, birliktelik kuralları ve ardışık zamanlı örüntü modelleri tanımlayıcı modeller olarak sınıflandırılmaktadır[3]. Bu istatistiksel yöntemlere ek olarak yaşam çözümlemesi yöntemleri de veri madenciliğinde özellikle müşteri veri tabanı sözkonusu olduğunda kullanılabilmektedir.

3. YAŞAM ÇÖZÜMLEMESİ

Yaşam çözümlemesi başarısızlık olarak adlandırılan bir nokta olayı ile ilgilenmektedir. Yaşayan bir organizmanın ya da cansız bir nesnenin belirli bir başlangıç zamanı ile başarısızlığı arasında

geçen zamana "yaşam süresi" ya da "başarısızlık süresi" adı verilmektedir. Her bir bireye ya da birime ait yaşam süresi T, tanımı gereği sürekli ve pozitif bir değere sahiptir. Başarısızlık süresine örnek olarak, makina bileşenlerinin yaşam süreleri, işçilerin grev süreleri, ekonomide işsizlik dönemleri, psikolojik bir deneyde deneğin belirlenen görevi tamamlama süresi ya da klinik bir deneyde hastaların yaşam süreleri gösterilebilir[4]. Örneğin, Evrensel 2007 yılında yapmış olduğu çalışmada uluslararası banka krizi verilerini, parametrik regresyon modeli kullanarak incelemiş ve ekonomi alanında da yaşam çözümlemesi yöntemlerinin kullanılabileceğini göstermiştir.

4. VERİ MADENCİLİĞİ'NDE YAŞAM ÇÖZÜMLEMESİ

Literatürde yaşam çözümlemesi yöntemleri tıbbi araştırmalarda sıklıkla yer bulmaktadır. Abone tabanlı işletmeler için de ilgilenilen bir yaşam süresi olduğu takdirde bu yöntemler kullanılabilir. Veri tabanları kullanılarak bir işletmenin müşteri profili belirlenmek istendiğinde, istatistiksel yöntemlerden biri olan yaşam çözümlemesi yöntemleri kullanılabilir. Bu yöntemler, müşteriler ve müşterilerin davranışları hakkında hızlı bir geri bildirim sağlamaktadır. Ayrıca müşteri değerini (customer value) belirlemek ve müşteri sadakatini ölçmek için sağlam bir dayanaktır[6].

Veri madenciliğinde yaşam çözümlemesi yöntemleri, istatistiksel yöntemlerden biri olan yaşam çözümlemesinin müşterileri ilgilendiren veri madenciliği problemlerine uygulanması olarak düşünülülebilir.

Yaşam çözümlemesi yöntemlerinin diğer tahmin edici modellerden temel farkı durdurulmuş (censored) veri için tasarlanmış istatistiksel yöntemler bütünü olmasıdır. Gözlem süresi boyunca çalışmada yer alan birimlerin tamamı başarısızlık ile karşılaşmamış olabilir. Bu birimler yaşam çözümlemesinde durdurulmuş olarak nitelendirilmektedir.

Veri madenciliğinde yaşam çözümlemesi yöntemleri daha çok müşteri veri tabanı incelenmesinde kullanılmaktadır. İlgilenilen olay müşteri kaybıdır. Müşteri ilişkilerini iyi tanımlamış, bir başlangıç ve bitiş zamanına sahip işletmeler için yaşam çözümlemesi yöntemleri kolaylıkla uygulanabilir. Sigorta, iletişim, kablolu televizyon, gazete ve magazin yayını, bankacılık ve rekabet pazarlarını içeren geniş bir endüstride bulunabilen abone tabanlı ilişkilere sahip işletmelere örnek olarak verilebilir[7].

Müşteri veri tabanı, hizmet ya da ürünün kullanımına başlanması ya da iptal edilmesi gibi birçok önemli olaya ait verileri içermektedir. Bu veriler, müşteriyi kaybetmemek ve pazarlama stratejileri belirlemek amacıyla tahmin edici modeller oluşturmak için kullanılmaktadır. Müşteri veri tabanı incelenirken hazard fonksiyonundan, yaşam fonksiyonundan ve regresyon modellerinden yararlanılabilir.

4.1. Hazard Fonksiyonu ve Yaşam Fonksiyonu

i. müşteri için ilgilenilen olayın ortaya çıkmasına kadar geçen süre T_i ile gösterilsin. Gözlenen olay süresi $Y_i=\min(T_i, a-B_i)$ ve gösterge değişkeni ise $\delta_i=I\{T_i\leq a-B_i\}$ ile belirtilmektedir. Olayın başlangıcı, B_i, müşteriler arasında farklılık göstermektedir. Çalışmanın bitttiği tarihte (a) tüm müşteriler olayla karşılaşmamış olabilir. Bu durumda başarısızlık süresi sağdan durdurulmuş (right censored) olarak nitelendirilmektedir. Durdurmanın diğer bir nedeni ise bağımsız ve ayrık yarışan olaylar olabilir. Örneğin, ilgilenilen olay servisin iptal edilmesi ise servis alanından ayrılan bir müşteri ayrıldığı tarihte (a_i) durdurulmuş olarak düşünülebilir.

Başarısızlık süresinin dağılımı genellikle yaşam fonksiyonu, $S(t)=P(T_i\geq t)$, ya da hazard fonksiyonu ile belirtilir. Hazard, gerçekleşecek olayın koşullu olasılığıdır ve

$$h(t)=\lim_{\Delta\to 0}\left(\frac{1}{\Delta}P(t\leq T\leq t+\Delta\,|\,T\geq t)\right)=-\frac{d}{dt}\ln(S(t)) \qquad (1)$$

biçiminde gösterilebilir Hazard fonksiyonu kullanılarak yaşam fonksiyonu,

$$S(t) = \exp\left(-\int_0^t h(x)dx \right) \tag{2}$$

şeklinde gösterilebilir. Yaşam fonksiyonu S(0)=1 ve $S(\infty) = 0$ olacak biçimde değerler almaktadır[4, 6].

4.2. Regresyon Modelleri

İncelenen yaşam süresi değişik faktörlerden etkilenebilir. Bağımlı değişken olan yaşam süresi üzerinde açıklayıcı değişkenlerin etkilerinin araştırıldığı regresyon modelleri yaşam çözümlemesinde önemli bir yere sahiptir. Müşteri veri tabanı, başarısızlık süresi dağılımını etkileyebilecek açıklayıcı bilgileri, demografik özellikleri, hesap bakiyesi ve ödemeler gibi ekonomik göstergeleri içermektedir. i. müşteri için açıklayıcı değişkenler vektörü X ile gösterilmektedir ve çoğu kez zamana bağlı olabilir.

Yaşam çözümlemesinde regresyon modelleri parametrik (lojistik, Weibull, üstel) ve yarı parametrik (Cox) regresyon modelleri olarak incelenmektedir. Bu modeller aşağıdaki bölümlerde kısaca ele alınmıştır.

4.2.1. Lojistik Regresyon Modeli

Lojistik regresyon modelinde kesikli hazard fonksiyonu; temel hazard fonksiyonu ve açıklayıcı değişkenlerin uygun fonksiyonlarının doğrusal bir kombinasyonu olarak varsayılmaktadır:

$$\ln\left(\frac{h(t \mid x_i(t))}{1 - h(t \mid x_i(t))} \right) = \ln\left(\frac{h_0(t)}{1 - h_0(t)} \right) + \eta\left(x_i(t), \beta\right). \tag{3}$$

Gözlenen verinin (Y_i, δ_i) olasılık yoğunluk fonksiyonu,

$$f_Y(y_i \mid x_i(y_i)) = h\left(y_i \mid x_i(y_i)\right)^{\delta_i} \prod_{j=0}^{i-1} \left(1 - h(y_j \mid x_j(y_j))\right) \tag{4}$$

biçimindedir. j=0,1,…,i için $\delta_{ij}=\delta_i I\{i=j\}$ nin bileşik dağılımına eşittir. δ_{ij} sonsal dağılımı $h(y_i \mid x_i(y_i))$ bağımsız Bernoulli değişkeni olarak ele alınmaktadır. Parametreler lojistik regresyon modeli kullanılarak en çok olabilirlik tahmin edicisi ile bulunabilir[6, 7].

4.2.2. Weibull Regresyon Modeli

α biçim parametresi, $\exp(\eta(x_i, \beta))$ ölçek parametresi ve $\eta(x_i, \beta)$ açıklayıcı değişkenlerin uygun bir fonksiyonu olmak üzere hazard fonksiyonu,

$$h(t \mid x_i) = \alpha t^{\alpha - 1} \exp(-\alpha \eta(x_i, \beta)) \tag{5}$$

biçiminde verilir. Hazard fonksiyonu α için değişimler göstermektedir. $\alpha<1$ iken hazard fonksiyonu azalan hızda azalır. $\alpha=1$ ise hazard fonksiyonu sabittir. $1<\alpha<2$ ise hazard fonksiyonu azalan hızda artar. $\alpha=2$ (Rayleigh) iken hazard fonksiyonu doğrusal olarak artar ve $\alpha>2$ iken hazard fonksiyonu artan hızda artar. Yaşam fonksiyonu ise,

$$S(t \mid x_i) = \exp\left(- t^{\alpha} \exp(-\alpha \eta(x_i, \beta)) \right) \tag{6}$$

biçiminde verilir[7,8].

4.2.3. Üstel Regresyon Modeli

Hazard fonksiyonu sabit olduğunda üstel regresyon modelinin kullanılması uygundur. Zaman ekseni, $(b_0=0,b_1]$ $(b_1,b_2]$ … $(b_{J-1},b_j=\infty)$ olmak üzere j tane aralığa bölündüğünde üstel regresyon modeli,

$$h(t \mid x_i(t)) = h_j \exp(\eta(x_i(b_{j-1}),\beta)) \quad , \quad t \in (b_{j-1}, b_j) \tag{7}$$

biçimindedir[7, 8].

4.2.4. Cox Regresyon Modeli

Üstel, Weibull ve lojistik regresyon modelleri parametrik modellerdir. Ancak verinin hangi dağılımdan geldiği kesin olarak belirlenemediğinde parametrik regresyon modellerinin kullanılması uygun olmamaktadır. Bu gibi durumlarda parametrik olmayan dağılımlardan daha çok bilgi veren ancak herhangi bir dağılım varsayımında bulunmayan yarı parametrik modeller kullanılabilir. Cox regresyon modeli yaşam süresi dağılımına ilişkin herhangi bir varsayım içermediğinden yarı parametrik bir model olarak tanımlanmaktadır.

Cox regresyon modelinin parametreleri, olabilirlik fonksiyonu maksimize edilerek tahmin edilmektedir. Bu model hazard fonksiyonun fonksiyonel biçimini belirlemeden hazard fonksiyonu üzerindeki açıklayıcı değişkenlerin etkilerini değerlendirmek için tasarlandığından temel hazard fonksiyonunun tahmin edilmesine gerek yoktur.

T, bir birimin yaşam süresini temsil eden sürekli raslantı değişkeni ve $X=(x_1,...,x_p)'$ bu birimle ilgili bilinen açıklayıcı değişkenler vektörü olmak üzere orantılı hazard varsayımı altında hazard fonksiyonu,

$$h(t \mid x_i) = h_0(t)\exp(\beta' x_i) \tag{8}$$

biçiminde tanımlanmaktadır. Burada, β regresyon katsayıları vektörü, $h_0(t)$ ise x=0 olan bir birimin temel hazard fonksiyonu olarak tanımlanmaktadır[8].

5. UYGULAMA

Veri madenciliğinin kullanım alanına ilişkin farklı örnekler verilebilir. Bir işletme kendi müşterisiyken rakibine giden müşterilerle ilgili istatistiksel analizler yaparak rakiplerini tercih eden müşterilerinin özelliklerini elde edebilir ve bundan yola çıkarak gelecek dönemlerde kaybetme olasılığı olan müşterilerin kimler olabileceği yolunda tahminlerde bulunarak onları kaybetmemek, kaybettiklerini geri kazanmak için stratejiler geliştirebilir.

Çalışmamızın yapısına uygun gerçek bir veri kümesine ulaşılamadığından, STATA paket programı kullanılarak, kredi kartı sahiplerine ait örnek bir veri kümesi üretilmiş ve kredi kartı sahiplerinin, kredi kartlarını iptal ettirmelerini etkileyen faktörler yaşam çözümlemesi kullanılarak belirlenmeye çalışılmıştır. Veri üretme sürecinde başarısızlık süresi için Weibull dağılımı, durum değişkeni(durdurulmuş, başarısız) için ise tekbiçimili dağılım kullanılmıştır.

Çalışmada bireylerin kredi kartı sahibi olduğu tarihten, kredi kartını kullanmayı bıraktığı tarihe kadar geçen süre (yıl olarak) başarısızlık süresi olarak alınmıştır. Kredi kartını kullanmayı bırakan müşteriler başarısız, kredi kartını kullanmaya devam eden müşteriler ise durdurulmuş olarak tanımlanmıştır. Müşterilerin izlenme süresi sona erdiğinde 854 müşteriden 451'inde (%52.8) başarısızlık ve 403'ünde (%47.2) durdurma gözlenmiştir. Uygulamada cinsiyet, yaş, meslek, medeni durum, gelir, kredi kartı sayısı değişkenleri çözümlemeye alınmıştır. Bu değişkenler, değişkenlerin düzeyleri ve tanımlayıcı değerleri EK 1'de verilmiştir.

Çizelge 1. Kaplan-Meier Yaşam Olasılıkları Sonuçları

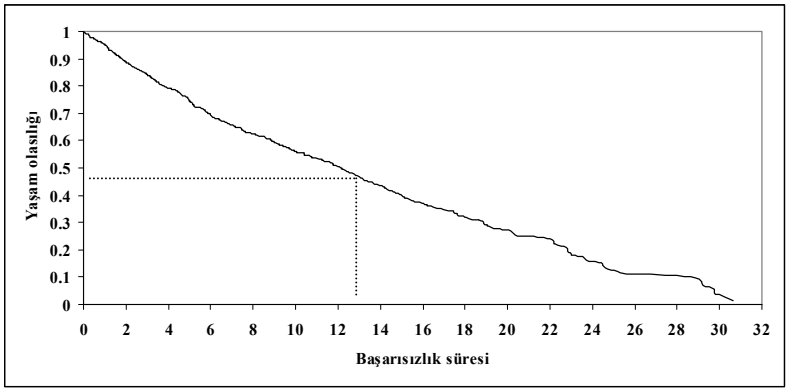
		Birikimli Yaşam Olasılığı			**Log-rank test**
		5YIL	**10 YIL**	**15 YIL**	**istatistiği (p-değeri)**
Cinsiyet	0.Kadın	0.7502	0.5691	0.3893	0.04 (0.8389)
	1.Erkek	0.7432	0.5546	0.4137	
Medeni durum	0.Bekar	0.7586	0.5939	0.4423	3.22 (0.0728)
	1.Evli	0.7338	0.5275	0.3591	
Meslek	1. İşsiz	0.7056	0.4399	0.3302	5.28 (0.2596)
	2. Memur	0.7375	0.5250	0.3205	
	3. İşçi	0.7685	0.5716	0.4511	
	4. Serbest	0.7056	0.5809	0.4112	
	5. Emekli	0.8084	0.6392	0.4713	
Gelir	1. ≤ 500	0.6756	0.4939	0.3401	5.28 (0.2594)
	2. ≤ 1000	0.7882	0.5950	0.3652	
	3. ≤ 1500	0.7116	0.5529	0.4089	
	4. ≤ 2500	0.7449	0.5375	0.3908	
	5. 2500 +	0.7791	0.6226	0.4860	
Yaş	1. 18-25	0.3066	.	.	101.27 (0.000)
	2. 26-40	0.7092	0.4784	0.2246	
	3. 41-55	0.8107	0.6228	0.4908	
	4. 56+	0.7738	0.6261	0.4714	

Çalışmada kullanılan değişkenler için 5, 10 ve 15 yıllık yaşam olasılıkları ve log-rank test istatistiği sonuçları elde edilmiş ve Çizelge 1'de verilmiştir. Değişkenlerin düzeyleri arasında yaşam olasılıkları açısından fark olup olmadığını test etmek için log-rank test istatistiği kullanılmıştır. Elde edilen sonuçlara göre yaş değişkeninin düzeyleri arasında yaşam olasılıkları açısından fark olduğu %95 güven düzeyinde söylenebilir. Yani, bankaya ait kredi kartını kullanmayı bırakma açısından yaş grupları arasında istatistiksel olarak anlamlı bir farklılık vardır.

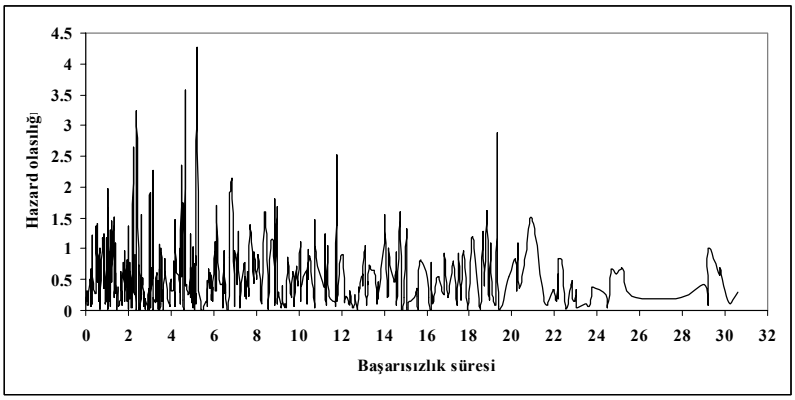
Kredi kartı sahiplerinin yaşam olasılığı ve hazard olasılığı grafikleri sırasıyla Şekil 1 ve Şekil 2'de verilmiştir. Şekil 1 incelendiğinde banka müşterilerinin %50'sinin 13. aydan itibaren bankaya ait kredi kartını kullanmayı bıraktığı söyleyebilir. Şekil 2 incelendiğinde ise 3., 5., 12. ve 20. yıllarda müşterilerin bankaya ait kredi kartı kullanımını bırakmasının daha riskli olduğu sonucuna ulaşılabilir. Buna göre banka özellikle bu yıllarda farklı çalışma stratejileri belirleyerek kaybedebileceği müşterileri yeniden kazanabilir.

Bireylerin başarısızlık sürelerini etkileyen faktörler ise üstel, Weibull ve Cox regresyon modelleri kullanılarak belirlenmeye çalışılmıştır. Yaşam çözümlemesinde modellerin anlamlığını test etmek için olabilirlik oranı (LR) test istatistiği kullanılmış ve model seçim kriteri olarak da

-2logL ve akaike bilgi kriteri (AIC) kullanılmıştır. Bu nedenle çalışmada incelenen modellere ait olabilirlik oranı test istatistikleri ve -2logL değerleri elde edilmiş ve Çizelge 2'de verilmiştir.



Şekil 1. Banka müşterilerine ait yaşam olasılığı grafiği



Şekil 2. Banka müşterilerine ait hazard olasılığı grafiği

Çizelge 2. Cox Regresyon Modeli ve Parametrik Cox Pegresyon Modelleri İçin LR ve -2logL Değerleri

	Üstel	**Weibull**	**Lojistik**	**Cox**
LR (p-değeri)	65.62 (0.00)	82.42 (0.00)	71.20 (0,00)	89.48 (0.00)
-2logL	1856.7583	1833.42348	1888.85346	5087.3808

Çizelge 2'de yer alan sonuçlar incelendiğinde üstel regresyon, Weibull regresyon, lojistik regresyon ve Cox regresyon modellerinin istatistiksel olarak anlamlı olduğu (p değeri < 0.05) görülmüştür. -2logL değerinin küçük olması modelin veri kümesine uygunluğunu gösterdiğinden, bu çalışmaya en uygun modelin Weibull regresyon modeli olduğu sonucuna ulaşılmıştır. Temel hazard fonksiyonu belirli bir dağılım ile hesaplanamadığından Cox regresyon modeli parametrik regresyon modellerine göre daha avantajlıdır. Çünkü yaşam sürelerinin olasılık dağılımının belirli bir biçimi yoktur. Ancak veri kümesi için belirli bir olasılık dağılımı varsayımı geçerli ise, bu varsayıma dayalı çıkarsamalar daha kesindir. Ayrıca parametre

tahminleri ve göreli hazard ya da ortanca yaşam süresi gibi ölçümlerin tahminleri daha küçük standart hataya sahip olur[4]. Efron (1977) and Oakes (1977) bu durumda parametrik regresyon modellerinin Cox regresyon modeline göre daha etkili parametre tahminlerine sahip olduklarını göstermişlerdir[9]. Çalışmamızda başarısızlık süresi Weibull dağılımına sahip olduğundan, Weibull regresyon modeli, diğer regresyon modellerinden daha iyi sonuç vermiştir.

Yaşam çözümlemesinde kullanılan regresyon modellerinde değişken düzeylerinden biri referans kategorisi olarak alınmakta ve değişken düzeylerinin yorumlanması buna göre yapılmaktadır. Bu çalışmada, modeldeki değişken için standart hata (S.H.), p değeri, hazard oranı (exp(β)) ile hazard oranının alt ve üst sınırları verilmiştir. Herbir değişken için ilk düzeyler referans kategorisi olarak alınmıştır. Çalışmadaki açıklayıcı değişkenlerle Weibull regresyon çözümlemesi yapıldığında elde edilen sonuçlar Çizelge 3'te verilmiştir.

Çizelge 3. Weibull Regresyon Çözümlemesinin Sonuçları

Değişken	**Exp(β)**	**S.H.**	**p-değeri**	**Alt sınır – Üst sınır**
Cinsiyet	0.9878421	0.0940363	0.898	0.8197068 - 1.190465
Medeni durum	1.1981320	0.1146596	0.059	0.9932207 - 1.445318
Meslek				
2. Memur	0.9917710	0.2112767	0.969	0.6532511 - 1.505715
3. İşçi	0.8012161	0.1638865	0.279	0.5365852 - 1.196356
4. Serbest	0.7989755	0.1644574	0.276	0.5337355 - 1.196027
5. Emekli	0.7686093	0.1829768	0.269	0.4820219 - 1.225588
Gelir				
2. ≤ 1000	0.7643591	0.1503256	0.172	0.5198674 - 1.123834
3. ≤ 1500	0.8023308	0.1709432	0.301	0.5284428 - 1.218173
4. ≤ 2500	0.8298222	0.1644252	0.346	0.5627605 - 1.22362
5. 2500 +	0.6608804	0.1437018	0.057	0.4315565 - 1.012064
Yaş				
2. 26-40	0.2962594	0.062903	0.000	0.1954077 - 0.4491615
3. 41-55	0.1770699	0.0378155	0.000	0.1165092 - 0.2691098
4. 56+	0.1711784	0.0367272	0.000	0.1124137 - 0.2606624
Kredi kartı sayısı	1.0468180	0.0263707	0.969	0.6532511 - 1.5057150

Başarısızlık süresini etkileyen faktörleri belirlemek için Çizelge 3 incelendiğinde, yaş değişkeninin önemli olduğu (p-değeri<0.05) %95 güven düzeyi ile söylenebilirken, medeni durum ve gelir değişkenlerinin önemli olduğu %90 güven düzeyi ile söylenebilmektedir. Yaş arttıkça kredi kartı kullanmayı bırakma riski de azalmaktadır. Ayrıca gelir arttıkça kredi kartı kullanmayı bırakma riskinin azaldığı ve evli müşterilerin bekar müşterilere göre yaklaşık 1.17 kat daha riskli olduğu %90 güven düzeyinde söylenebilmektedir.

6. SONUÇ

Bu çalışmada yaşam çözümlemesi yöntemleri veri madenciliği konusu çerçevesinde ele alındıktan sonra kredi kartı sahiplerine ait bir veri kümesi için yaşam olasılıkları, hazard olasılıkları ve regresyon modelleri incelenmiştir.

Uygulamada öncelikle yaşam olasılıkları 5'er yıllık 3 dönem(5,10,15) için elde edilmiştir. Daha sonra, yaşam ve hazard olasılıklarına ait grafikler verilmiş ve müşteri kaybı açısından yorumlanmıştır. Müşterilerin kredi kartını kullanmayı bırakmasını etkileyen risk faktörleri ise regresyon modelleri ile belirlenmeye çalışılmıştır. Weibull regresyon modelinin veri kümesi için en uygun regresyon modeli olduğu sonucuna ulaşılmıştır. Buna göre çalışmada yaş, gelir ve medeni durumun, müşterilerin kredi kartı kullanmayı bırakmalarını etkileyen önemli risk faktörleri olduğu görülmüştür.

İşletmeler için elde edilecek verilere göre yaşam çözümlemesinde yer alan modellerden uygun olanlar verinin analizi için kullanılabilir ve farklı modeller geliştirilebilir. İşletmeler için gerçek verilere ulaşılabilmesi halinde, ilgilenilen başarısızlık olayını etkileyen değişkenler incelenerek, bankacılık sektöründe kullanılabilecek yararlı modellerin geliştirilebileceği düşünülmektedir.

KAYNAKLAR

[1] Swift, R., "Accelerating Customer Relationship", Prentice Hall PTR, 2001.
[2] Özmen, Ş., "İş Hayatı Veri Madenciliği İle İstatistik Uygulamalarının Yeniden Keşfediyor", V. Ulusal Ekonometri ve İstatistik Sempozyumu, Adana, 2001.
[3] Akpınar, H., " Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği", İstanbul Üniversitesi İşletme Fakültesi Dergisi, 2000.
[4] Collett, D., "Modelling Survival Data in Medical Research", Chapman&Hall, UK, 1994.
[5] Evrensel, A.Y., "Banking Crisis and Financial Structure: A Survival Time Analysis", International Review of Economics and Finance, 2007.
[6] Linoff , G.S. , "Survival Data Mining for Customer Insight" , Intelligent Enterprise, 2004.
[7] Potts, W., "Survival Data Mining", SUGI Technical Report, 2004.
[8] Lee, E.T., Wang, J.W., "Statistical Methods for Survival Data Analysis", Wiley&Sons, New York, 2003.
[9] Nardi, A., Schemper, M., "Comparing Cox and Parametric Models in Clinical Studies", Statistics in Medicine, vol.22, pp.3597-3610, 2003.

EK 1. Kullanılan Değişkenler ve Düzeyleri

Değişken	**Değişken Düzeyleri ($\overline{X} \pm$ standart hata)**	**N**	**(%)**
Süre	8.2793 ± 7.0670		
Cinsiyet	0. Kadın	430	50.4
	1. Erkek	424	49.6
Medeni durum	0. Bekar	433	50.7
	1. Evli	421	49.3
Meslek	1. İşsiz	72	8.4
	2. Memur	196	23
	3. İşçi	220	25.8
	4. Serbest	228	26.7
	5. Emekli	138	16.2
Gelir	1. ≤ 500	76	8.9
	2. ≤ 1000	155	18.1
	3. ≤ 1500	108	19.7
	4. ≤ 2500	300	35.1
	5. 2500 +	155	18.1
Yaş	1. 18-25	56	6.6
	2. 26-40	227	26.6
	3. 41-55	292	34.2
	4. 56+	279	32.7
Kredi kartı sayısı	1. 1	124	14.5
	2. 2	131	14.5
	3. 3	145	17.0
	4. 4	131	15.3
	5. 5	101	11.8
	6. 6	91	10.7
	7. 7	83	9.7
	8. 8	48	5.6