MARMARA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

SAĞKALIM VERİ MODELLENMESİ: MAKİNE ÖĞRENMESİ YAKLAŞIMLARI

TUĞBA DEMİRCİOĞLU

YÜKSEK LİSANS TEZİ
İstatistik Anabilim Dalı

DANIŞMAN

Prof. Dr. Müjgan Tez

İSTANBUL, 2023

MARMARA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

SAĞKALIM VERİ MODELLENMESİ: MAKİNE ÖĞRENMESİ YAKLAŞIMLARI

TUĞBA DEMİRCİOĞLU

YÜKSEK LİSANS TEZİ
İstatistik Anabilim Dalı

DANIŞMAN

Prof. Dr. Müjgan Tez

İSTANBUL, 2023

TEŞEKKÜR

Lisansüstü öğrenimim boyunca emeği geçen, çalışmamın her aşamasında değerli fikirleri, bilgileri ve deneyimleri ile bana destek olan çok değerli danışman hocam Prof. Dr. Müjgan TEZ' e sonsuz sevgi ve teşekkürlerimi sunarım.

Tez çalışmam boyunca bilgi ve deneyimleri ile yanımda olan tavsiyelerinden yararlandığım sevgili hocam Prof. Dr. Deniz İNAN' a teşekkürlerimi sunarım.

Hayatım boyunca sevgi ve desteklerini esirgemeyen, her zaman yanımda olan değerli annem ve babama, kardeşlerim Samet DEMİRCİOĞLU ve Büşra DEMİRCİOĞLU' na, her zaman varlığıyla güç veren ve yanımda olan sevgili Fatih ŞEN'e, çok değerli arkadaşlarım Merve KÖKSOY ve Merve Nur ATAŞ' a, yoğun çalışmalarım sırasında her zaman yanımda olan ve desteğini esirgemeyen iş arkadaşlarıma en içten teşekkürlerimi sunarım.

Bu çalışma Marmara Üniversitesi Bilimsel Araştırma Projeleri Koordinasyon Birimi tarafından desteklenmiştir. Proje kodu: FYL-2021-10381

Ocak 2023 **Tuğba DEMİRCİOĞLU**

İÇİNDEKİLER

ÖZET

SAĞKALIM VERİ MODELLENMESİ: MAKİNE ÖĞRENMESİ YAKLAŞIMLARI

Bu çalışmada, Türkiye' de gıda perakende alanında faaliyet gösteren süpermarket zincirlerinden birine ait gıda e-ticaret sitesinin kampanya düzenlenen fakat kampanyalı satışı stoklarla sınırlı olan ürünlerin stok tükenme zamanları ve stok tükenme zamanlarına etki eden değişkenler sağkalım analiz yöntemleriyle incelenmiştir. İndirimli ürünün stok tükenme süresi için sağkalım olasılığının hesaplanması ve stok tükenme süresine etki eden değişkenlerin analizi için Kaplan Meier analizi, Cox Oransal Hazard Regresyon yöntemi, değişken seçimi için LASSO ve topluluk makine öğrenme yöntemi olan Sağkalım Ağaçları kullanılmıştır.

İlgilenilen olay indirimli kampanya ürünlerinin stok tükenme zamanı olduğundan değişkenlerin sağ kalma üzerindeki etkileri araştırılmıştır. Gelenekselleşmiş sağkalım analiz yöntemi olan Cox Oransal Hazard regresyon modeli ile makine öğrenmesi yöntemi olan Rastgele Sağkalım Ormanları yöntemi, modellerden elde edilen sağkalım analizi için ortak değerlendirme ölçütü C-index değerlerine göre karşılaştırılmıştır. Sonuç olarak perakende gibi farklı bir alanda hem gelenekselleşmiş sağkalım modelleri hem de Rastgele Sağkalım Ormanları modeli oluşturularak detaylı bir analiz çalışması ortaya konmuştur. Analiz sonucunda Cox oransal hazard regresyon modeline göre stok tükenme süresi içim anlamlı, Rastgele Sağkalım Ormanları yöntemi için önemli değişkenler birbirini destekler nitelikte çıkmıştır. Cox Oransal Hazard regresyon modeli ve Rastgele Sağkalım Ormanlarından elde edilen performans ölçüm metriği C-index değerleri karşılaştırılarak birbirine çok yakın sonuç verdiği görülmüştür.

Anahtar Kelimeler: Sağkalım Analizi, Cox Oransal Hazard regresyon yöntemi, Kaplan-Meier, Makine Öğrenmesi, Sağkalım Ağaçları, Rastgele Sağkalım Ağaçları

ABSTRACT

SURVIVAL DATA MODELING: MACHINE LEARNING APPROACHES

In this study, the time of depletion of stock and the variables affecting the time of depletion of the products, which are campaigned but whose sale is limited to stocks, of the food e-commerce site belonging to one of the supermarket chains operating in the field of food retail in Turkey, were examined by survival analysis methods. Kaplan Meier analysis, Cox Proportional Hazard Regression method, LASSO and ensemble machine learning method, Survival Trees, were used to calculate the survival probability for the stock depletion time of the discounted product and to analyze the variables affecting the stock depletion time.

Since the event of interest is the time of depletion of the discounted campaign products, the effects of the variables on survival were investigated. The traditional survival analysis method, the Cox Proportional Hazard regression model, and the machine learning method, the Random Survival Forests method, were compared according to the C-index values, the common evaluation criterion for the survival analysis obtained from the models. As a result, a detailed analysis study has been presented by creating both traditional survival models and Random Survival Forests model in a different field such as retail. As a result of the analysis, according to the Cox proportional hazard regression model, significant variables for stock depletion time and important variables for the Random Survival Forest method were found to be mutually supportive. The performance measurement metric C-index values obtained from the Cox Proportional Hazard regression model and Random Survival Forests were compared and it was seen that they gave very close results.

Keywords: Survival Analysis, Cox Proportional Hazard regression method, Kaplan-Meier, Machine Learning, Survival Trees, Random Survival Trees

SEMBOLLER

T : Sağkalım süresi

t : Takip süresi

F(t) : Dağılım fonksiyonu

f(t) : Olasılık yoğunluk fonksiyonu

S(t) : Sağkalım fonksiyonu

S(t,X) : X açıklayıcı değişkenlerine sahip birimin sağkalım fonksiyonu

h(t) : Tehlike fonksiyonu

$h_0(t)$: Temel tehlike fonksiyonu

h(t,X) : X açıklayıcı değişkenlerine sahip birimin tehlike fonksiyonu

H(t) : Kümülatif tehlike fonksiyonu

l_β : β' nın en çok olabilirlik fonksiyonu

$R(t_j)$: t_j anındaki risk kümesi

δ_i : i. veriye ait durdurma değişkeni

I(β) : Tahmin edilen bilgi matrisi

KISALTMALAR

NA : Nelson Aalen

YT : Yaşam tablosu

KM : Kaplan-Meier

LASSO : Least Absolute Shrinkage and Selection Operator

EN : Elastic net

CNN : Convolutional Neural Networks

YSA : Yapay sinir ağları

RNN : Reccurrent Neural Network

DVM : Destek vektör makineleri

RO : Rastgele ormanlar

RSO : Rastgele sağkalım ormanları

KHF : Kümülatif hazard fonksiyonu

OOB : Out of Bag

BS : Brier Skoru

C_{index} : Uyum indeksi

VIF : Variance inflation factors

ŞEKİL

TABLO LİSTESİ

1.GİRİŞ VE AMAÇ

Tıp araştırmacıları sağkalım analizi modelleri ile hastalığa etki eden prognostik faktörlerin ölüm ve kanser nüksü gibi sonuçlar üzerindeki önemini araştırarak yaptıkları değerlendirmeler ile hastalara tedavi seçenekleri sunar. Sağkalım analizinin amaçları; sağkalım olasılığının tahmin edilmesi, araştırmaya dahil edilen farklı grupların sağkalım sürelerinin karşılaştırılması, çeşitli konularda risk faktörlerinin belirlenmesi olarak verilebilir. Sağkalım analizi yalnızca tıbbi istatistiklerde değil diğer alanlarda da yaygın olarak uygulanmaktadır. Örneğin; bir web sitesini ziyaret eden kişilerin reklama ilk olarak ne zaman tıklayacağını tahminlemek için sağkalım analizi kullanılabilir. Tıp alanında yapılan çalışmalarda çok sayıda hasta sansürlenir. Analiz sonucu oluşacak modele önemli katkılar sağlayacak bu gözlemler göz ardı edilmemelidir. Sağkalım analizi diğer analiz yöntemlerinden farklı olarak sansürlü gözlem içeren veri kümelerine de uygulanabilir.

Sağkalım analizinde kullanılan en popüler yöntem Cox Oransal Hazard regresyon modelidir. Cox regresyon modeli; gözlemlenen ortak değişkenlerin bir olayın (ölüm gibi) meydana gelme riski üzerindeki etkilerini hesaplayan yarı parametrik istatistiksel yöntemdir. Cox regresyon analizinde; bağımsız değişkenlerin hazard fonksiyonu üzerindeki etkileri log lineer olması ve bağımsız değişkenlerin log-lineer fonksiyonu ile hazard fonksiyonu arasındaki ilişki çarpımsal olması varsayımları söz konusudur. Bu iki varsayımın yanında hazard (tehlike) oranının zamana göre değişmemesi ve gözlemlerin birbirinden bağımsız olması gerekmektedir. Cox regresyonunun bu varsayımlarını kabul etmek gerçek dünya uygulamaları için çoğu zaman mümkün değildir. Bu nedenle, sansürlü sağkalım verilerini doğrusal olmayan log risk fonksiyonlarına uydurmak için zengin bir sağkalım modelleri ailesine ihtiyaç duyulmuştur. Makine öğrenmesi yöntemleri, son dönemde yapılan araştırmalarla doğrusal olmayan ilişkileri modelleme ve tahminleme performansı açısından önemli başarılar elde ederek birçok farklı alanda popüler hale gelmişlerdir. Sağkalım analizinde kullanılan geleneksel istatistiksel yöntemleri makine öğrenmesi yöntemleriyle değiştirme fikri caziptir. Sağkalım analizi yöntemleri sansürlü sağkalım verilerine doğru şekilde işlendiğinde Cox regresyon yöntemine göre daha doğru sağkalım tahminlerine sahip olabilir.

Bu tez çalışmasının amacı, sağkalım analizi için literatürde bulunan, gelenekselleşmiş istatistiksel yöntemlerin yanı sıra makine öğrenmesi yöntemlerini araştırmak, sağkalım analizinin uygulanabilirliğini perakendecilik gibi farklı bir sektörde ortaya koymaktır.

Yapılan bu çalışmada, Türkiye' de gıda perakende alanında faaliyet gösteren süpermarket zincirlerinden birine ait gıda e-ticaret sitesinin 16.01.2022 ile 28.09.2022 tarihleri arasında kampanya düzenlenen fakat kampanyalı satışı stoklarla sınırlı olan ürünlerin sağkalım olasılığı ve sağkalım olasılıklarına etki eden değişkenler sağkalım analiz yöntemleriyle incelenmiştir. Kampanya tanımlanan her bir ürün için takip süresi olarak 15 saat alınmıştır. Takip süresi sonlandığında her bir ürün için çalışma başlangıcında belirlenen stok adeti satışı yapılamamış ise bu gözlem sansürlü gözlem olarak kabul edilmiştir. Bu kapsamda Kaplan Meier analizi, Cox Oransal Hazard Regresyon yöntemi, değişken seçimi için LASSO ve topluluk makine öğrenme yöntemi olan Sağkalım Ağaçları kullanılmıştır. İlgilenilen olay indirimli kampanya ürünlerinin stok tükenme zamanı olduğundan değişkenlerin sağ kalma üzerindeki etkileri araştırılmıştır. Gelenekselleşmiş sağkalım analiz yöntemi olan Cox Oransal Hazard regresyon modeli ile makine öğrenmesi yöntemi olan Rastgele Sağkalım Ormanları yöntemi, sağkalım analizi için ortak değerlendirme ölçütü C-index sonuçlarına göre karşılaştırılarak veri setinin detaylı bir analizi ortaya konulmuştur.

2. SAĞKALIM ANALİZİNİN KAPSAMI VE TEMEL KAVRAMLARI

2.1 Sağkalım Analizinin Kapsamı

Sağkalım analizi, sonuç değişkeni ilgilenilen olay meydana gelene kadar geçen süre olan verileri analiz etmek için kullanılan istatistiksel yöntemler topluluğudur. Sağkalım analizi; başarısızlık zamanı analizi (failure time analysis) ya da olay zaman analizi (event time analysis) olarak da ifade edilmektedir (Kleinbaum ve Klein, 2010).

Sağkalım analizinin temel amacı, araştırmacı tarafından belirlenen gözlem süresince ilgilenilen olayın meydana gelmesini etkileyen değişkenleri ve bu olay meydana gelene kadar geçen süreyi modellemektir (Lee ve Wang, 2003).

Sağkalım analizi, başta tıp bilimi olmak üzere demografi, sosyoloji, biyoloji, mühendislik, ekonomi gibi farklı disiplin dallarında yapılan bilimsel çalışmalarda kullanılmaktadır. Tıp

biliminde sağkalım analizi (survival analysis), mühendislikte güvenilirlik analizi (reliability analysis), ekonomide süre analizi (duration analysis), sosyolojide ise olay geçmişi analizi (event history analysis) olarak adlandırılmaktadır. İlgilenilen olay, hastanın; belirli bir hastalıktan ölümü, kanserli hastanın remisyondan çıkarak tekrar kötüleşmesi veya iyileşmesi, hastalık insidansı olabileceği gibi çiftlerin boşanması, makine arızalanması, şirket iflası, gibi çalışmanın amacına göre tanımlanmış herhangi bir olay olabilir (Kleinbaum ve Klein, 2010).

2.2 Sağkalım Analizinin Temel Kavramları

2.2.1 Sağkalım süresi

Sağkalım süresi, gözleme alınan her bir birim için belirli bir başlangıç zamanından ilgilenilen olayın meydana geldiği ana kadar geçen süre olarak tanımlanmaktadır (Kleinbaum ve Klein, 2010).

Sağkalım analizi araştırmalarında birimler için takip başlangıç zamanı ve ilgilenilenilen olayın meydana geldiği an net bir şekilde tanımlanmalıdır. Sağkalım süreleri, araştırmacı tarafından tanımlanan başlangıç zamanından itibaren ölçülmeye başlanmalıdır. Araştırmacı tarafından tanımlanan başlangıç zamanı her bir birim için aynı olabileceği gibi farklı da olabilir. Sağkalım sürelerinin belirlenmesinde bir zaman ölçeği kullanılmalıdır. Çalışmalarda, ilgilenilen olay gerçekleştiği ana kadar geçen zamanı ölçeklendirmek için gün, hafta, ay, yıl gibi gerçek zaman kullanılabileceği gibi olayın gerçekleştiği andaki bireyin yaşı kullanılabilir (Lee ve Wang, 2003).

2.2.2 Sansür

Sağkalım analizinin birinci temel özelliği, sonuç değişkeninin; orijin noktasından ilgilenilen olayın meydana gelmesine kadar geçen süreyi temsil eden, yani negatif olmayan iyi tanımlanmış sürekli veya kesikli rastgele değişken olmasıdır. İkinci temel özelliği ise kullanılan verilerin genellikle sansürlü gözlem içermesidir (Moore, 2016).

Sağkalım analizinde kullanılan yöntemlerin diğer istatistiksel yöntemlerden temel farkı, sansürlü gözlem içeren veri kümelerinin modellenmesinde de kullanılabilmeleridir.

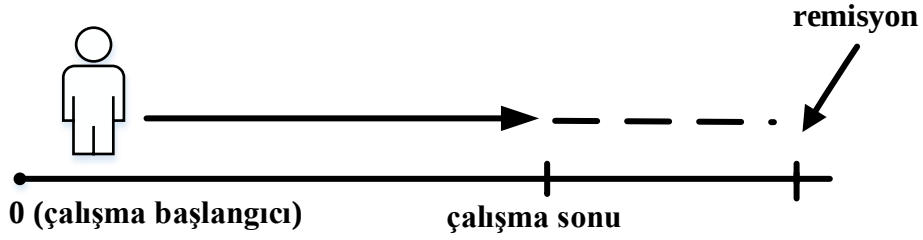
Sansürlenmiş veriler sağkalım analizinde sıklıkla rastlanan bir problemdir (Kleinbaum ve Klein, 2010).

Sansürlü veri genel olarak üç farklı nedenden dolayı meydana gelir. Bunlar aşağıdaki şekildedir:

(1) Birimde izlem süresinde ilgilenilen olay meydana gelmemiştir (administrative censoring).
(2) Birim izlem süresinde takipten çıkmış ve birimden bilgi alınamamıştır (lost to follow-up).
(3) Birim ilgilenilen olay dışında farklı bir nedenden dolayı çalışmadan çekilmiştir (withdrawing).

Bu gibi nedenlerle takibin daha fazla mümkün olmadığı birimler, sansürlü veri olarak adlandırılır. Sağkalım verisi genel olarak sansürlü verilerden meydana gelir. Bu gözlemlerin tamamı hiçbir araştırmacı tarafından kayıp veri olarak görülerek çalışma dışında bırakılmaz. Çünkü sansürlü gözlemler araştırma sonucunu etkileyerek yapılan çalışmaya büyük bir katkı sağlayabilir (Kleinbaum ve Klein, 2010).

Örneğin; lösemi hastalarının remisyondan çıkarak hastalığın tekrar nüks etmesi gözlemlenen bir araştırmada araştırma başlangıcında remisyonda olan bir hasta çalışma sonlandığı halde Şekil 2.1'deki şekilde remisyondan çıkmamış olsun. Bu hastanın en az çalışma başlangıcı ile çalışma sonu arasında geçen zaman kadar remisyonda kaldığını söyleyebiliriz. Araştırma başlangıcında belirlenen takip süresinde belirlenen olay gerçekleşmediğinden dolayı bu hastanın sağkalım süresi (remisyonda kalma süresi) sansürlenmiştir (Kleinbaum ve Klein, 2010).

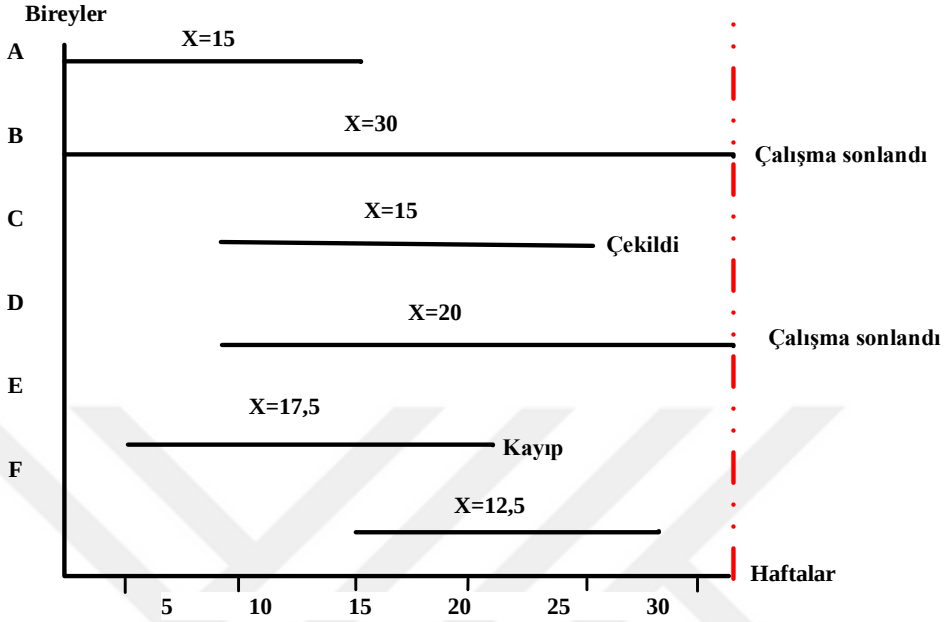


Şekil 2.1 Sansürlü veri örneği

Şekil 2.2'de 30 haftalık bir çalışma grafiklendirilmiştir. Bu çalışmada gözlemlenen A, B, C, D, E, F birimleri için takip durumlar ve sağkalım süreleri aşağıdaki şekilde belirlenebilir (Bardakcı ve Kartal, 2018).

A birimi çalışma başlangıcında çalışmaya katılmış ve bu birimde 15. haftada ilgilenilen olay meydana gelmiştir. A birimi için sağkalım süresi T=15 olarak belirlenmiş ve bu birim sansürlenmemiştir.

B birimi de çalışma başlangıcında çalışmaya katılmış fakat bu birimde çalışmanın sonuna kadar olay meydana gelmemiştir. B birimi için gerçek sağkalım süresinin 30 haftadan büyük olduğunu bilmemize rağmen tam olarak kaç hafta olduğunu söyleyemeyiz. Bu birim sansürlenmiştir. Sağkalım süresi olarak T=30 olarak alınır.



Şekil 2. 2 Sansürlemeye ilişkin grafiksel örnek

C birimi çalışmaya 10. haftada dahil olarak 15 hafta sonra ilgilenilen olay meydana gelmeden çalışmadan çekilmiştir. C birimi için de gerçek sağkalım süresinin 15 haftadan fazla olduğunu bilmemize rağmen tam olarak kaç hafta olduğunu bilemeyiz. Bu birim sansürlenmiştir. Sağkalım süresi olarak T=15 alınır.

D birimi çalışmaya 10. haftada çalışmaya katılmış fakat bu birimde çalışmanın sonuna kadar olay meydana gelmemiştir. İlgilenilen olay meydana gelmeden çalışma tamamlanmıştır. Bu birim sansürlenmiştir. Sağkalım süresi T=20 alınır.

E birimi çalışmaya 5.haftada katılmış ve 17,5 hafta boyunca izlenmiştir. İlgilenilen olayın dışında farklı bir nedenden dolayı birim takipten çıkmıştır. Bu nedenle E birimi de sansürlenmiştir. Sağkalım süresi olarak T=17,5 alınır.

F birimi çalışmaya 15.haftada katılarak 12,5 hafta sonra ilgilenilen olayı yaşamıştır. Bu birimde sansürleme meydana gelmemiştir. Sağkalım süresi T=12,5 olarak tespit edilmiştir (Bardakcı ve Kartal, 2018).

Sansürlü veriler temel olarak üç farklı başlık altında toplanmaktadırlar (Kleinbaum ve Klein, 2010).

- Sağdan Sansürlü
- Soldan Sansürlü
- Aralıklı Sansürlü

2.2.2.1 Sağdan sansürlü veriler

Sağdan sansürleme sağkalım verilerinde en sık gözlenen sansürleme tipidir. İlgilenilen olay çalışma süresi boyunca takip edilen birimlerde gerçekleşmez ise gerçek sağkalım süresi araştırma başlangıcında belirlenen çalışma süresinin sağ tarafında kalacaktır. İlgilenilen olayın gerçekleşme zamanı net olmadığından dolayı sağkalım süresi kesin olarak bilinemeyecektir. Yapılan çalışma kapsamında, bu tür veriler sansürlenecektir. Bu tip sansürlemeye sağdan sansürleme denilir (Bardakcı ve Kartal, 2018).

Birimin; gözlem süresi L_i, sağkalım süresi T_i olmak üzere; $T_i \geq L_i$ olduğunda bu birim için sağkalım süresinin sağdan sansürlenmiş olduğu söylenir. δ_i sansür değişkeni aşağıdaki şekildedir :

$$\delta_i = \begin{cases} 0, & eğer\ , T_i > L_i \\ 1, & eğer\ , T_i \leq L_i \end{cases} \quad i = 1,2, \dots, n \qquad (2.1)$$

Eğer sansür değişkeni $\delta_i = 0$ ise birim sansürlenmiş, $\delta_i = 1$ ise sağkalım süresi sansürlenmemiş yani ilgilenilen olay araştırma için belirlenen gözlem süresinde meydana gelmiştir .

Sağdan sansürleme; Tip I, II ve III olmak üzere kendi içerisinde alt gruplara ayrılır (Bardakcı ve Kartal, 2018).

- **Tip I sansürleme**

Çalışmaya alınan tüm birimler için ilgilenilen olayın gerçekleşmesi maliyet ve zaman açısından sorun olabilir. Bu gibi durumlarda çalışma süresi araştırmacı tarafından çalışma başlangıcında belirlenir. Gözlem süresi sonlandığı halde ilgilenilen olay gerçekleşmeyen veya başka bir nedenden dolayı ölen birimlere ait sağkalım süreleri I. tip sansürlü verileri meydana getirir (Bardakcı ve Kartal, 2018).

Örneğin; takip süresinin araştırmacı tarafından çalışma başlangıcında 2 yıl olarak belirlendiği karaciğer nakli yapılan hastalarla ilgili bir çalışmayı ele alalım. Belirlenen çalışma periyodu içerisinde takibe alınan tüm hastalar öldüğü takdirde sağkalım sürelerinde sansürleme olmayacaktır. Fakat takibe alınan hastaların bir kısmı herhangi bir nedenden dolayı gözlemden çıkıp bu birimlerin takibi mümkün olmaz ise sağkalım sürelerinde sansürleme meydana gelmiş olacaktır

- **Tip II sansürleme**

Çalışmaya katılan bazı birimler belirli bir süre izlendikten sonra herhangi bir nedenle takip edilemeyebilir. Söz konusu birimlerde takipsizlik meydana gelmiş olur. Bu türdeki veriler ise II. tip sansürlü verilerdir (Bardakcı ve Kartal, 2018).

- **Tip III sansürleme**

Gözlem süresinin sabit olarak önceden belirlendiği çalışmalarda, takip edilen birimler çalışmaya farklı zamanlarda dahil olabilir. Bu veriler için sağkalım süresi birimin çalışmaya katılmasından takipten çıkış zamanına kadar geçen süre olarak alınmaktadır. Çalışmaya dahil olduktan sonra çalışma süresi sonlandığı halde ilgilenilen olayın gerçekleşmediği veya herhangi bir sebepten dolayı takipsizlik durumu meydana gelmişse III. tip sansürleme söz konusudur. Bu birimler için sağkalım süresi sansürlüdür ve sağkalım süresi olarak birimin çalışmaya dahil olduğu andan çalışma bitimine kadar geçen süre alınır (Bardakcı ve Kartal, 2018).

2.2.2.2 Soldan sansürleme

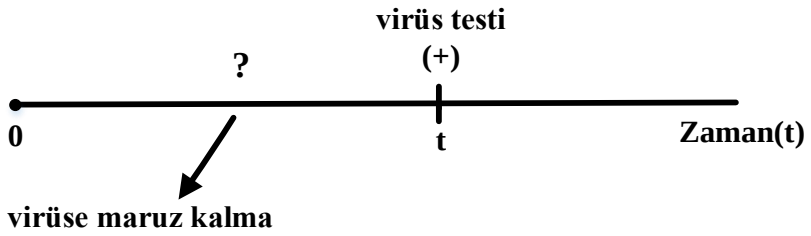
Gözleme alınan birimlerde, araştırmacı tarafından ilgilenilen olay gözlem süresi içerisinde değil gözlem süresinden önce meydana geldiğinde bu birimlerde soldan sansürleme meydana gelmiştir. Sağdan sansürlemede soldan sansürlemenin aksine birimin gözlem süresi sağkalım süresinden küçüktür.

Birimin; gözlem süresi L_i, sağkalım süresi T_i olmak üzere; $T_i < L_i$ olduğunda bu birim için sağkalım süresinin soldan sansürlenmiş olduğu söylenir. δ_i sansür değişkeni aşağıdaki şekildedir :

$$\delta_i = \begin{cases} 0, & eğer\ , T_i \leq L_i \\ 1, & eğer\ , T_i > L_i \end{cases} \quad i = 1,2, \dots, n \tag{2.3}$$

Eğer sansür değişkeni $\delta_i = 0$ ise birim sansürlenmiş, $\delta_i = 1$ ise sağkalım süresi sansürlenmemiş yani gözlemlenmiştir (Bardakcı ve Kartal, 2018) .

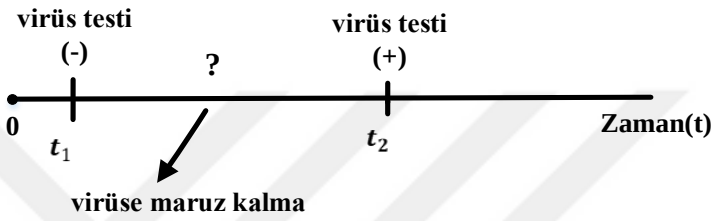
Örneğin; Şekil 2.3' de görselleştirilen virüs testini ele alalım. Bu örnek için ilgilenilen olay virüs testinin pozitif çıkması olsun. Araştırma başlangıcında virüs testi negatif olan bir kişinin belirli bir t anında virüs testi pozitif çıkarsa kişinin hastalığa tam olarak ne zaman maruz kaldığını bilemeyiz. Yalnızca 0 ile test edilen zaman arasında bilinemeyen bir anda maruz kaldığını söyleyebiliriz. Dolayısıyla gerçek sağkalım süresi virüs testinin sol tarafında kaldığından dolayı bu birim soldan sansürlenmiştir (Kleinbaum ve Klein, 2010).



Şekil 2. 3 Soldan sansürleme örneği

2.2.2.3 Aralıklı sansürleme

Sağdan ve soldan sansürlemenin genelleştirilmiş bir türü olan aralıklı sansürleme, sağkalım süresinin bir aralık içinde olduğu durumlarda meydana gelir. Aralıklı sansürleme tipinde ilgilenilen olay belirli iki zaman noktası arasında gerçekleşmiştir. İlgilenilen olayın tam olarak gerçekleşme anı hakkında kesin bir bilgi yoktur (Nelson, 1972).



Şekil 2. 4 Aralıklı sansürleme örneği

Şekil 2.4' de aralıklı sansürlemeye örnek bir çalışma görselleştirilmiştir. t_1 anında gözlemlenen birimin virüs testinin negatif çıktığı görülmektedir. Belir bir zaman sonra t_2 anında virüs testi yapıldığında test pozitif çıkmıştır. Bu senaryoda kişinin t_1 ile t_2 zamanı arasında bir noktada virüse maruz kaldığını söyleyebiliriz. Fakat tam olarak hangi anda maruz kaldığını söyleyemeyiz. Dolayısıyla birimin gerçek sağkalım süresi belirli iki nokta arasındadır. Birim aralıklı sansürlenmiştir (Kleinbaum ve Klein, 2010) .

2.3 Sağkalım Analizinde Kullanılan Fonksiyonlar

Sağkalım analizi kapsamında sağkalım sürelerinin dağılımının tanımlanabilmesi için temel olarak aşağıdaki üç fonksiyon kullanılmaktadır (Lee, 1992).

- Sağkalım Fonksiyonu
- Olasılık Yoğunluk Fonksiyonu
- Hazard (Risk) Fonksiyonu

Bu fonksiyonlar birbirleriyle matematiksel olarak ilişkilidir. Herhangi biri bilindiğinde diğer fonksiyonlara ulaşılabilir (Kleinbaum ve Klein, 2010) .

2.3.1 Sağkalım fonksiyonu

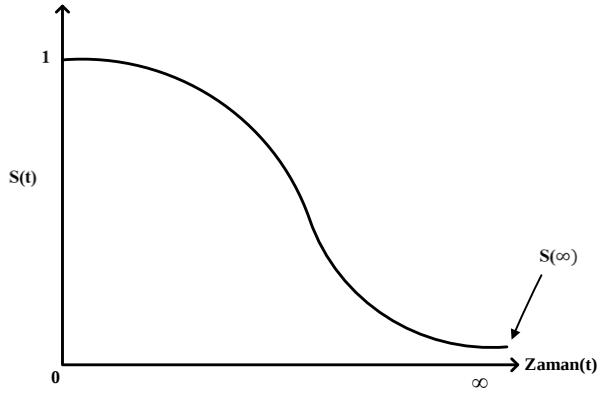
Sağkalım fonksiyonu $S(t)$ ile gösterilir. Sağkalım fonksiyonu, bir birimin belirli bir (t) zamanından daha fazla sağkalma olasılığını verir. Yani $S(t)$, gerçek sağkalım süresi T 'nin belirli bir t süresinden büyük olma olasılığıdır. Sağkalım fonksiyonunun matematiksel olarak ifadesi aşağıdaki şekildedir (Kleinbaum ve Klein, 2010):

$$S(t) = P(T > t) \tag{2.4}$$

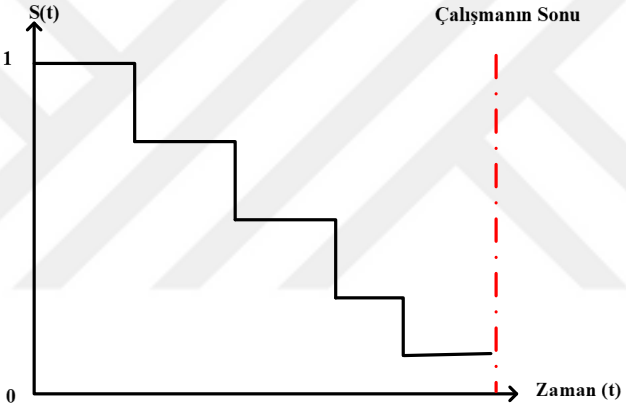
Belirlenecek farklı t değerleri için elde edilecek olan $S(t)$ sağkalım fonksiyonları, yapılacak çalışmada sağkalım verilerinden önemli bilgiler çıkarılmasını sağlayacaktır.

Çalışma başlangıcında yani $t = 0$ anında ilgilenilen olay hiçbir birimde meydana gelmeyecektir. Bu durumda çalışma başlangıcında $S(t) = S(0) = 1$'dir. Yani sağkalım olasılığı bire eşittir. $t = \infty$ anında yani araştırmacı tarafından belirlenen sağkalım süresinin sonsuz olması durumunda, birimlerin hayatta kalması mümkün değildir. Bu nedenle sağkalım süresi $S(t) = S(\infty) = 0$'dır.

Şekil 2.5' de gösterildiği şekilde, t değeri teorik olarak sıfırdan sonsuza kadar değer aldığında $S(t)$ birden başlayarak azalan düzgün bir eğri olarak grafiklendirilir. Uygulamada ise her zaman sonsuz zaman ve sürekli değişken imkanı bulunmamaktadır. Bu durumda sağkalım fonksiyonun grafiği Şekil 2.6'daki şekilde basamak fonksiyonu şeklinde çizilir . Gerçek hayat problemlerinde çalışma için belirlenen süre sonsuz değerini alamayacağından dolayı çalışma sona erdiğinde sağkalım fonksiyonu $S(t)$ y-eksenini kesmek zorunda değildir (Kleinbaum ve Klein, 2010).



Şekil 2. 5 Teorik olarak sağkalım eğrisi



Şekil 2. 6 Uygulamada sağkalım eğrisi

2.3.2 Olasılık yoğunluk fonksiyonu

Başlangıç zamanı kesin olarak belirli olan çalışmada bir birimin sağkalım süresi T olsun. Sağkalım süresi T'nin negatif olmayan sürekli bir olasılık yoğunluk fonksiyonuna sahip olduğu varsayımı altında, T mutlak süreklilik özelliğini sağlıyorsa olasılık yoğunluk fonksiyonu aşağıdaki şekilde gösterilir (Kleinbaum ve Klein, 2010):

$$\mathrm{f(t)} = \lim_{\Delta t \to 0} \frac{P\big((t, t+\Delta t)\ aralığında\ ilgilenilen\ olay\ meydana\ gelmiş\ birim\big)}{\Delta t} \quad (2.5)$$

$F(t)$ kümülatif olasılık yoğunluk fonksiyonu ise ilgilenilen olayın belirli bir t anından önce gerçekleşme olasılığı olarak tanımlanır. $F(t)$ kümülatif olasılık yoğunluk

fonksiyonunun matematiksel olarak gösterimi Denklem 2.6 kullanılarak Denklem 2.7'deki şekilde elde edilir:

$$f(t) = \frac{dF(t)}{dt} \tag{2.6}$$

$$F(t) = 1 - S(t) = P(T \leq t) = \int_0^t f(x)d_x \tag{2.7}$$

T'nin olasılık yoğunluk fonksiyonu f(t) ve sağkalım fonksiyonu $S(t)$ arasındaki ilişki Denklem 2.5 kullanarak bulunmaktadır (Kleinbaum ve Klein, 2010):

$$\text{f(t)} = \lim_{\Delta t \to 0} \frac{F(t+\Delta t) - F(t)}{\Delta t} = \frac{d}{dt} F(t) = -\frac{d}{dt}\, S(t) \tag{2.8}$$

2.3.3 Hazard (risk) fonksiyonu

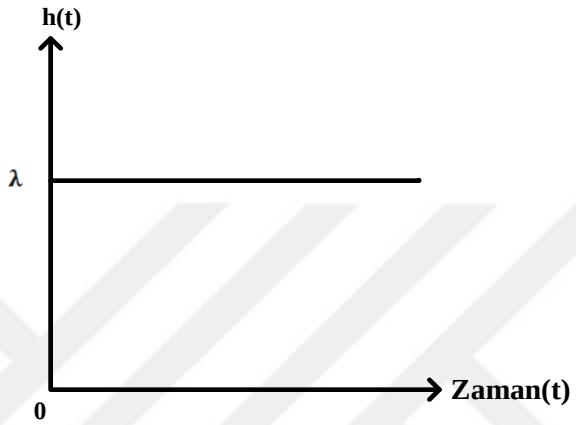
T sağkalım süresinin Hazard fonksiyonu olarak gösterilen $h(t)$, koşullu başarısızlık oranınına karşılık gelir. Birimin ilgilenilen olayın meydana gelmesi açısından başarısızlık eğilimidir. İlgilenilen olayın gözlemlenen bir birim için t anında meydana gelmemesi koşulu altında birim zaman başı Δt' ye karşılık anlık gerçekleşme potansiyelidir. Hazard (risk) fonksiyonu;

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t} \tag{2.9}$$

şeklinde ifade edilir (Kleinbaum ve Klein, 2010).

Hazard fonksiyonu bir olasılığı değil hızı temsil eder. Bu nedenle [0,1] arasında değerler almak zorunda değildir. Hazard fonksiyonu sağkalım fonksiyonu gibi 1'den başlamak zorunda olmayıp herhangi bir noktadan başlayıp zaman ilerledikçe artan ya da azalan olarak iler ve istediği değerleri alır. Ancak negatif değerler alamaz. Hazard fonksiyonunun üst sınırı yoktur. Hazard fonksiyonu 0 ile sonsuz aralığında değerler alır .

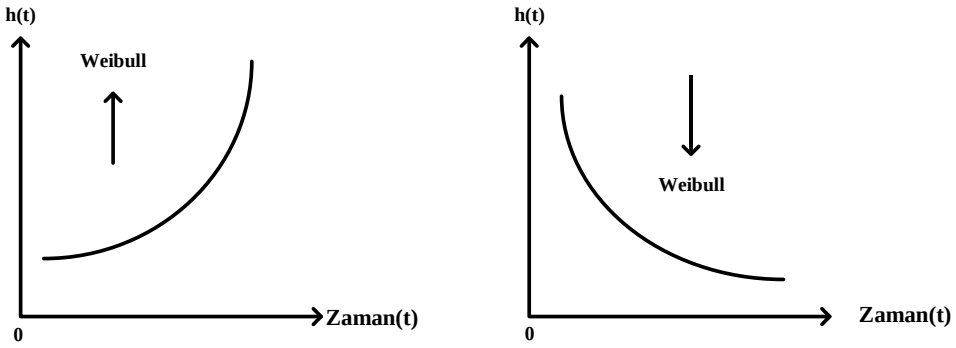
Sağkalım fonksiyonunun sahip olduğu olasılık dağılımının tipine göre Hazard fonksiyonu farklı yapıda olur. Şekil 2.7'de sağlıklı kişiler üzerinde yapılan bir çalışma için sürekli tehlike gösterilmektedir. Herhangi bir t değeri için h(t) sabit bir λ değerine sahiptir. Takip süresi boyunca sağlıklı olan birimin çalışma boyunca herhangi bir zamanda anlık hasta olma potansiyeli sabittir ve değişmez (Kleinbaum ve Klein, 2010).



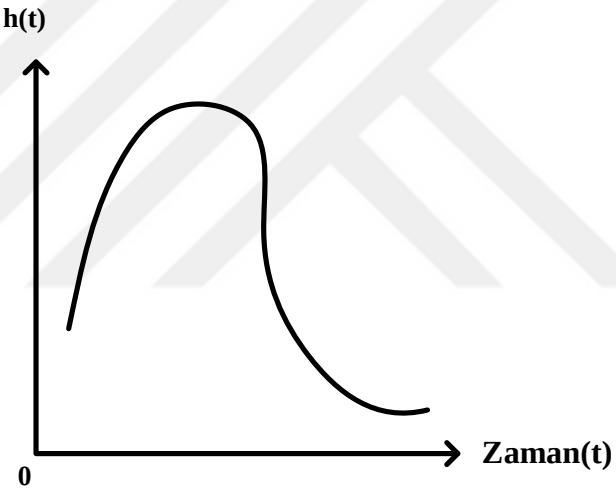
Şekil 2. 7 t üstel dağılıma sahip ise hazard fonksiyonu grafiği

Şekil 2.8'de gösterilen ilk grafik zamanla artan bir tehlike fonksiyonunu göstermektedir. Örneğin; tedaviye cevap vermeyen kan kanseri hastalarının gözlemlendiği bir araştırmada zaman ilerledikçe hastlalık prognozu kötüleşir. Buna bağlı olarak hastanın ölme olasılığı artar. Şekil 2.8' de gösterilen ikinci grafik ise zamanla azalan tehlike fonksiyonuna aittir. Örneğin hazard fonksiyonu azalan Weibull şeklinde olan kişilerin ameliyat sonrası iyileşmekte olan hastalara ait olduğunu söyleyebiliriz. Çünkü ameliyat sonrası ölme olasılığı genellikle ameliyattan sonra geçen süre arttıkça azalacaktır (Kleinbaum ve Klein, 2010).

Şekil 2.9 ise önce artan sonra azalan bir tehlike fonksiyonunu göstermektedir. Örneğin; tüberküloz hastaları için bu şekilde bir tehlike fonksiyonu grafiği beklenebilir. Çünkü tüberküloz hastalığına sahip kişilerin erken yaşlarda ölme potansiyelleri ileri yaşlara göre daha yüksektir. Yaş ilerledikçe tüberküloz hastalığı sebebiyle ölme olasılığı artacaktır (Kleinbaum ve Klein, 2010).



Şekil 2. 8 t Weibull dağılımına sahip ise hazard fonksiyonu grafikleri



Şekil 2. 9 t log-normal dağılımına sahip ise hazard fonksiyonu grafiği

Hazard fonksiyonu ve sağkalım fonksiyonu arasındaki matematiksel ilişki aşağıdaki şekildedir (Lee ve Wang, 2003):

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - \mathrm{F}(t)}{\Delta t * S(t)} \qquad (2.10)$$

Denklem 2.10'dan Denklem 2.11 ve Denklem 2.12 bulunmaktadır:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - \mathrm{F}(t)}{\Delta t * S(t)} = \frac{f(t)}{S(t)} \qquad (2.11)$$

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}\, S(t) * \frac{1}{S(t)} = -\frac{d}{dt}[logS(t)] \quad (2.12)$$

Denklem 2.12'de görüldüğü üzere, sağkalım sürelerine ait uygun dağılımı parametrik olarak ifade edebiliyorsak $S(t)$ ve $h(t)$ fonksiyonlarını da parametrik olarak belirleyebiliriz.

Ayrıca T'nin kümülatif hazard fonksiyonu aşağıdaki şekilde tanımlanır:

$$H(t) = \int_0^t h(u)du \quad (2.13)$$

Denklem 2.13 önemli bir ilişkiye yol açar. $h(t)$ ile $S(t)$ arasındaki ilişki aşağıdaki şekilde bulunmaktadır (Lee ve Wang, 2003):

$$S(t) = exp[H(t)] = \exp[-\int_0^t h(u)du] \quad (2.14)$$

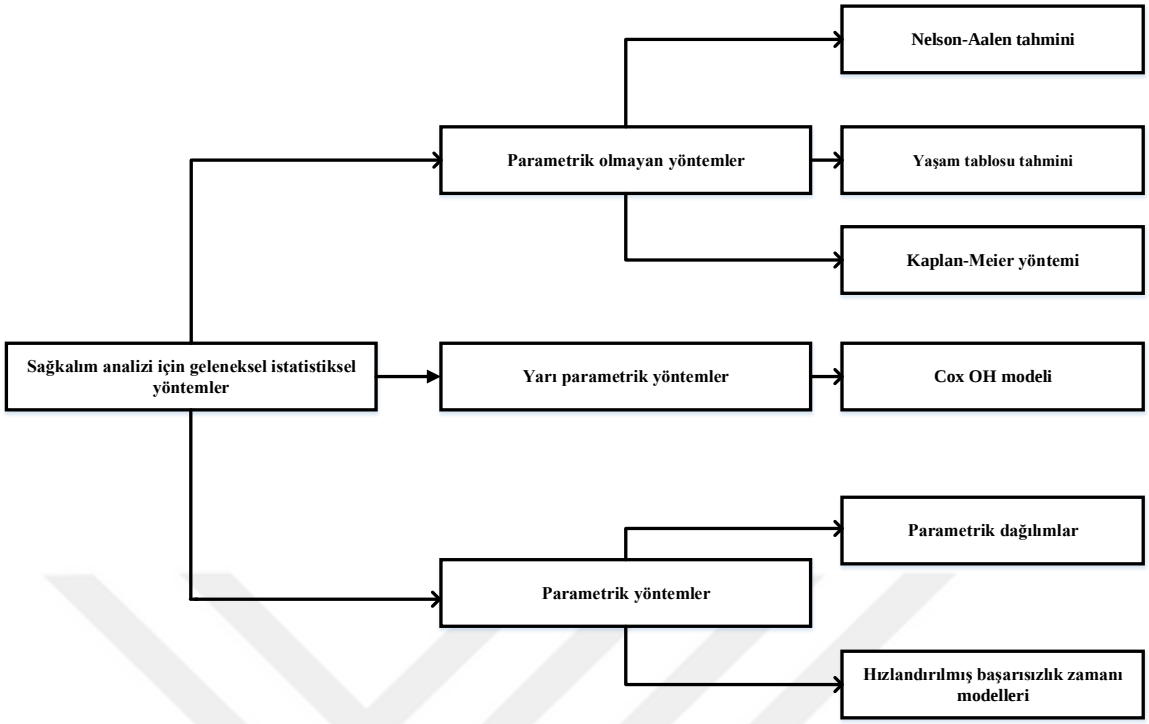
3. SAĞKALIM ANALİZİNDE KULLANILAN GELENEKSEL İSTATİSTİKSEL YÖNTEMLER

İlgilenilen problemin çözümü ve değişkenin özelliğine göre sağkalım analizinde farklı yaklaşımlar söz konusudur (Bardakcı ve Kartal, 2018).

Sağkalım ve tehlike (hazard) fonksiyonunu tahmin etmek için kullanılan geleneksel istatistiksel yöntemler; parametrik olmayan, yarı parametrik ve parametrik yöntemlerdir. Şekil 3.1'de sağkalım analizinde kullanılan geleneksel istatistiksel yöntemlerin kapsamlı bir özeti sunulmaktadır (Wang ve ark., 2019).

3.1 Parametrik Olmayan Yöntemler

Birimlerin T sağkalım sürelerinin dağılımı hakkında herhangi bir bilgi olmadığı ve oransal hazard varsayımı sağlanmadığı durumda sağkalım verilerinin analizinde parametrik olmayan yöntemler daha etkili sonuçlar ortaya koymaktadır. Parametrik olmayan yöntemler ile sağkalım fonksiyonunun amprik bir tahmini yapılır (Wang ve ark., 2019).



Şekil 3. 1 Sağkalım analizi için geleneksel istatistiksel yöntemler

3.1.1 Nelson-Aalen tahmini

Nelson-Aalen (NA) tahmin edicisi (Nelson, 1972), ilk olarak 1972'de Nelson tarafından tanıtılmıştır Daha sonra 1978'de Aelen (Aalen, 1978) tarafından yeniden keşfedilip geliştirilen NA yöntemi, modern sayma işlemi tekniklerine dayanan bir tahmin edici haline gelmiştir .

Nelson-Aelen tahmin edicisi sansürlenmiş (tamamlanmamış) verilerin kümülatif hazard fonksiyonu için parametrik olmayan bir tahmin edicidir (Anderson ve ark., 2012).

d_j, t_j anında ilgilenilen olay meydana gelen birim sayısı, r_j ise t_j anında risk altındaki birim sayısı olmak üzere kümülatif hazard fonksiyonunun Nelson-Aalen tahmin edicisi aşağıdaki şekilde tanımlanır (Wang ve ark., 2019):

$$\widehat{H}(t) = \sum_{t_j \leq t} \frac{d_j}{r_j} \tag{3.1}$$

Eşitlik (3.1)'de tanımlanan $\hat{\mathrm{H}}(t)$ kümülatif hazard fonksiyonu tahmin edicisi, Eşitlik (3.2) ile sağkalım fonksiyonunu tahmin etmek için aşağıdaki şekilde kullanılabilir (Wang ve ark., 2019):

$$\hat{\mathrm{S}}(t) = e^{-\hat{H}(t)} = \exp[-\sum_{t_j \le t} \frac{d_j}{r_j}] \tag{3.2}$$

3.1.2 Yaşam tablosu (YT) tahmini

Sağkalım verilerinin analizi ve sağkalım eğrilerinin elde edilmesinde kullanılan yaşam tablosu yöntemi (Cutler ve Ederer, 1958) çalışmaya dahil edilen birim sayısının fazla olduğu durumlar ($n \geq 100$) için tavsiye edilir. Bu yöntem ile çalışma başlangıcında %100 olan sağkalım olasılığının belirli süreç aralıkları ile yeni sağkalım olasılık düzeyleri belirlenir. Araştırmanın belirli dönemlerinde ilgilenilen olay meydana geldiğinden dolayı çalışma başlangıcında %100 olan sağkalım olasılığı giderek azalır. Belirlenen çalışma süresi sonunda sağkalım olasılık değerleri birleştirilerek çizim yapıldığında sağkalımın azalarak sürdüğü gösterilmiş olur (Cox, 1972).

Çalışmada belirlenen sağkalım süreleri her bir zaman aralığı, $[t_j, t_{j+1})$ ile gösterilmek üzere $j = 1,2, \dots, k$ için eşit aralıklara bölünür. $[t_j, t_{j+1})$ zaman aralığında; dj yaşam süreleri sansürlenmemiş birimlerin gözlemlenen toplam başarısızlık sayısını, cj sağkalım süreleri sansürlenmiş birimlerin sayısını, nj ise ilgilenilen olayın meydana gelmediği risk kümesinde bulunan birimlerin sayısını göstermek üzere sansür değişkeninin $[t_j, t_{j+1})$ zaman aralığında uniform dağıldığı varsayımı ile risk kümesinde bulunan ortalama birim sayısı (Kleinbaum ve Klein, 2010);

$$n_j' = \frac{nj - cj}{2} \tag{3.3}$$

olarak hesaplanır .

$\left[t_j, t_{j+1}\right)$ aralığında risk kümesindeki birimlerde ilgilenilen olayın meydana gelmesinin koşullu olasılığı;

$$\hat{q}_j = \frac{dj}{n_j'} \tag{3.4}$$

olarak tahmin edilir. Her bir aralık için hesaplanan $\hat{q}_j$ koşullu başarısızlık olasılığı "standart yaşam tablosu tahmini" olarak adlandırılır. $\hat{q}_j$ koşullu başarısızlık olasılığına bağlı olarak koşullu yaşam sürdürme olasılığı $\hat{p}_j = 1 - \hat{q}_j$ 'dür (Kleinbaum ve Klein, 2010).

$t \in [t_j, t_{j+1})$ için sağkalım fonksiyonunun yaşam tablosu tahmin edicisi aşağıdaki şekildedir :

$$\hat{S}(t) = \prod_{i=1}^{j} \left(\frac{n'_{j-} f_j}{n'_j}\right), \qquad \text{j} = 1,2,\dots,\text{k} \tag{3.5}$$

j. aralık boyunca bozulma hızının sabit olduğu ve bu aralıkta τ_j; j. zaman aralığının uzunluğu olmak üzere ortalama sağkalım süresinin $(n'_{j-} f_j/2)\tau_j$ olduğu varsayılsın. $t \in [t_j, t_{j+1})$, $\text{j} = 1,2,\dots,\text{k}$ için j. zaman aralığında tehlike (hazard) fonksiyonunun yaşam tablosu tahmini aşağıdaki şekilde gösterilir (Kleinbaum ve Klein, 2010):

$$\hat{h} = \frac{f_j}{(n'_{j-} f_j/2)\tau_j} \tag{3.6}$$

3.1.3 Kaplan-Meier (KM) yöntemi

Kaplan ve Meier 1958'de, sağkalım fonksiyonunu tahmin etmek için gerçek gözlem süresini kullanan Product-Limit (PL) tahmin edicisi olarak da bilinen Kaplan-Meier (KM) eğrisini geliştirmişlerdir (Kaplan ve Meier, 1958). KM tahmin edicisi, homojen birimlerden meydana gelen sağkalım verilerine ait sağkalım fonksiyonunu tahminlemede kullanılan en yaygın yöntemlerden biridir (Wang ve ark., 2019).

Sansürsüz verilerden oluşan bir çalışmalarda sağkalım fonksiyonu $S(t)$; belirli bir $t$anına eşit veya büyük zamanlar için ilgilenilen olayın meydana gelmediği birimlerde sağkalım olasılığını göstermektedir. $S(t)$ deneysel sağkalım fonksiyonundan tahmin edilmektedir. Belirli bir t anındaki deneysel sağkalım fonksiyonu; t anında yaşayan birim sayısını çalışmaya alınan birim sayısına oranlayarak bulunur (Kleinbaum & Klein, 2010).

Çalışmada; n adet birim olduğu ve $k \leq n$ olmak üzere, k adet birimde ilgilenilen olayın meydana geldiği varsayımı altında, sıralanmış sağkalım süreleri $t_1 < t_2 < \cdots <$

t_k olsun. $j = 1,2, \dots, k$ olmak üzere ilgilenilen olayın meydana geldiği bir t_j anı belirlensin. t_{j-1} ile t_j periyodu arasında, ilgilenilen olayın gerçekleşmediği birimlerin sayısı r_j, bu aralıktaki sansürlü birim sayısı ise c_j ile gösterilsin. Belirlenen t_j anında ilgilenen olay meydana gelen birim sayısı ise d_j olsun. $r_j = r_{j-1} - d_{j-1} - c_{j-1}$ olmak üzere sağkalım fonksiyonunun t anındaki Kaplan-Meier tahmin edicisi aşağıdaki şekilde bulunur (Kleinbaum ve Klein, 2010).

$$\hat{S}(t) = \prod_{j:t_j<t} (1 - \frac{d_j}{r_j}) \tag{3.7}$$

t_j anı için tahminlenen sağkalım fonksiyonunun değeri, bir sonraki olayın meydana geldiği t_{j+1} anına kadar sabittir. Bir önceki bozulma anından sonra, yeni bir olay meydana gelene kadar başka bir olay meydana gelmemesi nedeni ile olasılık değişmez. Dolayısıyla; t anındaki sağkalım fonksiyonunun, Kaplan-Meier tahmin edicisine ait çizilen grafik azalan bir basamak fonksiyonu şeklindedir. Sağkalım fonksiyonu özelliği gereği çalışma başlangıcında ilgilenilen olay meydana gelmediğinden $t < t_1$ için $\hat{S}(t) = 1$'dir. En büyük gözlem sansürlenmiş ise; $\hat{S}(t)$ değeri 0 değerini hiçbir zaman alamaz (Kleinbaum ve Klein, 2010) .

Kaplan-Meier (KM) yöntemi ile az sayıda birim ile çalışılabilir. Yaşam tablosu (YT) yönteminde aralıklara düşen az birim sayısı yapılacak tahminleri etkilemektedir. Çalışma büyük bir popülasyonu kapsıyorsa yani birim sayısı fazla ise, yapılacak analizde YT yöntemi ve KM yöntemi benzer sonuçlar vermektedir. Takip edilen birim sayısı fazla olduğunda YT yöntemi tercih edilmelidir. YT yöntemi KM yönteminin aralıklı gruplandırılmış verilere uygulanma şeklidir. KM yönteminde kayıp birimler dikkate alınmaz ve yalnızca ölüm sayıları ile sağkalım olasılığı hesaplanır. Sağkalım olasılığı ölüm olayı gerçekleşir gerçekleşmez hesaplanır. KM yönteminde kesin ölüm tarihi kullanıldığından dolayı bu yöntem sağkalım olasılığını kesin olarak verir. YT yöntemi sağkalım olasılığının yaklaşık değerini verecektir. KM ve YT yöntemlerinden farklı olarak NA tahmin edicisi, modern sayma yaklaşımı ile sansürlü gözlemler için kümülatif hazard fonksiyonunu tahmin eder (Sümbüloğlu ve Akdağ, 2009).

3.1.4 Sağkalım eğrilerinin karşılaştırılması

En az iki olmak üzere iki veya daha fazla sağkalım eğrisinin karşılaştırılması için aşağıdaki testler kullanılmaktadır (Şenocak, 1992):

- Log – Rank Testi,
- Mantel – Cox Testi,
- Breslow – Wilcoxon Testi,
- Tarone – Ware Testi

Sağkalım eğrilerinin karşılaştırılması için yaygın olarak kullanılan iki test 1965'de Gehan ve 1970'de Breslow'un önerdiği genelleştirilmiş Wilcoxon testi ve 1966'da Mantel ve 1972'de Cox tarafından önerilen log-rank testidir . Genel olarak, log-rank testi en geç zamanda dağılım farklılıklarına duyarlı olma eğilimindedir. Wilcoxon testi ise farklılıkları erkenden tespit etmede log-rank testine göre daha güçlüdür. Uygulamalarda oransal tehlike varsayımı kontrol edildikten sonra sıklıkla log-rank testi kullanılır. Wilcoxon testi ise oransal tehlike varsayımı sağlanmadığında kullanılabilecek Log-rank testinin alternatifidir (Kleinbaum ve Klein, 2010) .

3.2 Yarı Parametrik Yöntemler

Regresyon analizi, aralarında neden-sonuç ilişkisi olan iki veya daha fazla değişken arasındaki ilişkiyi bir model aracılığı ile tahmin etmek için yaygın olarak kullanılan istatistiksel analiz yöntemlerinden biridir (Wang ve ark., 2019).

Tahmin etmek istediğimiz değişkene bağımlı(yanıt) değişken, bağımlı değişkeni tahmin etmek için kullandığımız değişken veya değişkenlere ise bağımsız (açıklayıcı) değişken denilmektedir. Regresyon analizinde incelenen değişkenler kesikli veya sürekli yapıda olabilmekte ve veri yapısı gereği farklı regresyon modelleri kullanılabilmektedir. Regresyon analizi yöntemlerinden ilk akla gelen yöntem doğrusal regresyon analizidir. Doğrusal regresyon analizi, basit doğrusal regresyon analizi ve çoklu doğrusal regresyon analizi olarak iki başlık altında incelenmektedir. Bağımlı değişken ile tek bir bağımsız değişken arasındaki doğrusal ilişki basit doğrusal regresyon analizi ile açıklanır. Tek bir bağımlı değişken ve birden fazla bağımsız değişken arasındaki doğrusal veya eğrisel ilişki

modellenmek istendiğinde ise çoklu doğrusal regresyon analizine başvurulur. Hem basit hem de çoklu doğrusal regresyon analizi neticesinde elde edilen modele ait parametre kestirimlerinin güvenilir olabilmesi için modelle ilgili bazı varsayımların sağlanması gereklidir Bu varsayımlardan bağımlı ve bağımsız değişkenlerin normal dağılması en önemlilerinden biridir. Çoklu doğrusal regresyon analizinde bağımsız değişkenlerin birbirlerinden bağımsız olması, etkilerinin birbirleri üzerinde orantısal bağımlılık göstermemesi varsayımı da sağlanmalıdır (Arı ve Önder, 2013).

Sağkalım verileri ile yapılan bir araştırmada bağımlı değişken, belirli bir hastalığa yakalanan bireylerin ölümlerine kadar geçen gözlem süreleri iken bağımlı değişken üzerinde etkide bulunan değişkenler faktör değişkenlerdir. Faktör değişkenler; yaş, cinsiyet, ırk vb. olabilir. Sağkalım analizinde faktör değişken olduğu düşünülen etmenler normal dağılım göstermemekte, aralarında orantısal ilişki bulunmakta ve birbirleri ile korelasyon göstermektedir. Bu nedenlerden dolayı sağkalım verileri ile yapılan araştırmalarda çoklu doğrusal regresyon analizi kullanılamamaktadır. Sağkalım verilerinin nedensellik analizinde Cox regresyon yöntemleri olarak bilinen regresyon modellerinden yararlanılmaktadır (Lee ve Wang, 2003) .

Sağkalım verileri ile yapılan çalışmalarda çoklu doğrusal regresyon analizinin kullanılmamasının aşağıdaki şekilde iki önemli sebebi vardır:

1. Sağkalım analizinde bağımlı değişken olarak belirlenen sağkalım sürelerinin çoğunlukla weibull ya da üstel dağılmaya eğilimli olması, normal dağılım göstermemesi,
2. Sağkalım verilerinin sansürlü gözlem bulundurması yani birimlerin yaşam süresi içinde araştırmacının ilgilendiği olaya maruz kalmamış olmasıdır (Sümbüloğlu ve Akdağ, 2009).

Sağkalım verileri analizinde yarı parametrik yaklaşımlarda en çok kullanılan regresyon analizi yöntemi Cox oransal hazard regresyon modelidir (Lee ve Wang, 2003). Yarı parametrik yöntemler başlığı altında Cox regresyon yönteminin ayrıntıları verilerek bu yöntemin uzantısı olan düzenlileştirilmiş Cox regresyon yöntemi de ele alınacaktır.

3.2.1 Cox oransal hazard regresyon yöntemi

Cox tarafından 1972 (Cox, 1972) yılında ortaya konulan regresyon modeli ile sağkalım analizinde önemli adımlar atılmıştır. 1972' de Cox tarafından yapılan çalışma, 1980 yılında Kalbfleisch ve Prentice' in katkıları ile bugünkü önemi kazanmıştır (Bradburn ve ark., 2003).

Sağkalım verilerinin nedensellik analizlerinde en çok kullanılan yöntem Cox regresyon modelidir. Cox regresyon modeli yaygın kullanılan, kabul görmüş bir yöntem olup "Orantılı Risk (Hazard) Regresyon Analizi" olarak da bilinmektedir (Bardakcı ve Kartal, 2018).

Cox regresyon yöntemi, bağımlı değişken olan sağkalım süresi ile sağkalım süresine etki eden birden fazla bağımsız değişken arasındaki ilişkiyi modelleyen istatistiksel bir yöntemdir. Cox regresyon analizinin amacı sağkalım verilerinin genel durumunu yansıtacak bir modelleme yapmaktır (Kleinbaum Klein, 2010).

Örneğin tıp alanında yapılan iki tedavi türünün karşılaştırıldığı sağkalım analizi çalışmasında, hastanın; yaşı, cinsiyeti, ırkı gibi demografik değişkenler, kandaki beyaz hücre sayısı, nabız gibi fizyolojik değişkenler, hastanın sigara kullanması gibi ek bilgiler çalışmaya alınabilir. Çalışmaya alınan bu değişkenler hastanın sağkalım süresinde ve tedavi şeklinde etkili olabilir. Sağkalım süresine etkisi olduğu düşünülen bu tip değişkenlere açıklayıcı değişkenler (bağımsız değişkenler) denir. Sağkalım analizine bağımsız değişkenlerin de dahil edilmesiyle bu değişkenlerin ilgilenilen olayın meydana gelme riski üzerindeki etkileri analiz edilebilir (Kleinbaum ve Klein, 2010).

Cox regresyon modelinin iki temel varsayımı bulunmaktadır:

1. Risk (hazard) fonksiyonu üzerindeki bağımsız değişkenlerin etkileri loglineerdir.
2. Bağımsız değişkenlerin loglineer fonksiyonu ile risk fonksiyonu arasındaki ilişki çarpımsaldır (Kleinbaum ve Klein, 2010).

Yukarıdaki iki varsayıma ek olarak gözlemlerin birbirinden bağımsız olması ve risk oranının zamana göre değişmeyip sabit olması gerekir. Risk oranı ile ilgili varsayım orantısal risk varsayımı olarak bilinmektedir (Hosmer ve ark., 2002).

Cox regresyon modelinin matematiksel olarak aşağıdaki şekilde ifade edilir (Kleinbaum ve Klein, 2010).

$$h(t;X) = h_0(t)\exp\left(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p\right)$$

$$h(t;X) = h_0(t)e^{\sum_{i=1}^{p}\beta_i X_i} \tag{3.8}$$

şeklinde yazılabilir. Bu eşitlikte $X_1, X_2, \ldots, X_p$ açıklayıcı(bağımsız) değişkenlerdir. Açıklayıcı değişkenler T sağkalım süresi üzerinde etkisi bulunan yaş, sıcaklık, kan basıncı gibi sürekli değişkenler ya da hastalık evresi, cinsiyet, histolojik evreler gibi kategorik değişkenler olabilir. $\beta_1, \beta_2, \ldots, \beta_p$ regresyon katsayıları, $h_0(t)$ ise $X = 0$ yani bağımsız değişkenlerin sağkalım süresi üzerindeki etkisi sıfır olduğu durumunda temel risk (hazard) fonksiyonudur (Wang ve ark., 2019).

Tehlike oranının zamana karşı sabit olması ya da bir bireyin tehlikesinin diğer bireyin tehlikesine orantılı olması, orantılı tehlike (hazard)varsayımıdır (Kleinbaum ve Klein, 2010).

$\boldsymbol{X}^* = (X_1^*, X_2^*, \ldots, X_P^*)$ ve $\boldsymbol{X} = (X_1, X_2, \ldots, X_P)$ iki bireye ait bağımsız değişkenler vektörü olmak üzere tehlike oranı :

$$\widehat{HO} = \frac{\hat{\mathrm{h}}(t,\boldsymbol{X}^*)}{\hat{\mathrm{h}}(t,\boldsymbol{X})} = \frac{exp[\sum_{i=1}^{p}\hat{\beta}_i X_i^*]\hat{\mathrm{h}}_0(t)}{exp[\sum_{i=1}^{p}\hat{\beta}_i X_i\,]\hat{\mathrm{h}}_0(t)} = exp\left[\sum_{i=1}^{p}\hat{\beta}_i(X_{i-}^* X_i)\right] \tag{3.9}$$

şeklindedir. Denklem 3.9' da görüldüğü üzere tehlike oranı t'yi içermemektedir. Yani, model uydurularak X^* ve X değerleri belirlendiğinde, tehlike oranı tahmini için üstel ifadenin değeri sabittir, zamana bağlı değildir (Wang ve ark., 2019).

Elde edilen sabit $\hat{\theta}$ ile gösterilirse tehlike oranı;

$\widehat{\theta} = \frac{\hat{h}(t,\boldsymbol{X}^*)}{\hat{h}(t,\boldsymbol{X})}$ biçiminde yazılır. $\widehat{\theta}$, orantılılık sabiti (proportionality constant) olarak adlandırılıp zamandan bağımsızdır (Wang ve ark., 2019).

Sağkalım fonksiyonu Denklem 2.14' de;

$$S(t) = \exp\left[-\int_0^t h(u)du\right]$$

olarak verilmişti. S(t) sağkalım fonksiyonundan hareket ederek Cox regresyon modelinin sağkalım fonksiyonu S(t,X) aşağıdaki şekilde elde edilir (Wang ve ark., 2019).

$$S(t,X) = \exp\left[-\int_0^t h(u)e^{\Sigma_{i=1}^p \beta_i X_i}\,du\right] = \exp\left[e^{\Sigma_{i=1}^p \beta_i X_i}\left(-\int_0^t h(u)du\right)\right]$$

$$= [\exp\left(-\int_0^t h(u)du\right)]^{e^{\Sigma_{i=1}^p \beta_i X_i}}$$

$$S(t,X) = \left[S(t)^{exp\,\Sigma_{i=1}^p \beta_i X_i}\right] \qquad (3.10)$$

Denklem 3.10' da verildiği üzere Cox regresyon modelinde, regresyon parametresi β ve temel tehlike(hazard) fonksiyonu $h_0(t)$ olmak üzere iki adet bilinmeyen bileşen bulunmaktadır (Kleinbaum ve Klein, 2010).

Temel hazard fonksiyonu $h_0(t)$ tanımlanmamış bir fonksiyondur. Cox regresyon modelini parametrik olmayan bir model haline getirmektedir. $h_0(t)$'nin tanımlanmış belirli bir fonksiyonu olmadığından Cox yarı parametrik bir modeldir. $h_0(t)$'nin dağılım şekli üzerinde herhangi bir varsayım bulunmadığından dolayı hesaplanmasına gerekmemektedir. Önemli olan $\beta_1, \beta_2, \ldots, \beta_p$ regresyon katsayılarının hesaplanmasıdır. β bilinmeyen katsayılar vektörünün tahmininde kısmi en çok olabilirlik fonksiyonundan faydalanılır. Kısmi en çok olabilirlik fonksiyonu, sansürlü gözlemlerin olasılık değerini hesaba katmayıp yalnızca sansürlenmemiş gözlemler için olasılık değerini dikkate alır. n tane birey ve bu bireylerin r tane farklı başarısızlık zamanları olsun. Bireylerin n-r tanesi sağdan sansürlenmiş olsun. t_j j.sıradaki sağkalım fonksiyonu olmak üzere, r tane sağkalım

süresi $t_1, t_2, \dots, t_r$ olsun. j. olayın meydana gelme zamanında ilgilenilen olayı yaşamamış yani başarısızlığa uğrama risk altında olan bireylerin oluşturduğu kümeyi risk kümesi olarak isimlendirelim. Risk kümesi $R(t_j)$ ile gösterilsin. t_j anına kadar yaşamış olan birimde t_j anında ilgilenilen olayın meydana gelme olasılığı (Wang ve ark., 2019) :

$$l_j(\beta) = \frac{h_0(t_j)\exp(\beta^T X_j)}{\sum_{l\epsilon\, R(t_j)} h_0(t_j)\exp(\beta^T X_l)} = \frac{exp(\beta^T X_j)}{\sum_{l\epsilon\, R(t_j)} exp(\beta^T X_l)} \tag{3.11}$$

şeklindedir. Denklem 3.11'de verilen pay değeri açıklayıcı değişken vektörü X_j olan bir bireyin t_j anındaki ölüm riskidir. Payda ise, t_j anında başarısızlık durumu ile karşı karşıya olan tüm birimlerin ölüm risklerinin toplamıdır. Genel olarak kısmi en çok olabilirlik fonksiyonu r tane başarısızlık zamanı üzerinden olasılıkların çarpımı olarak yazılabilir. Yapılan bütün bu tanımlamalar ve açıklamalar ile Cox tarafından β'ların tahmin edilmesinde önerilen orantılı tehlike regresyon modeli için kısmi en çok olabilirlik fonksiyonu olarak aşağıdaki şekildedir (Wang ve ark., 2019):

$$l(\beta) = \prod_{j=1}^{r} \frac{exp(\beta^T X_j)}{\sum_{l\epsilon\, R(t_j)} exp(\beta^T X_l)} \tag{3.12}$$

Eşitlik 3.12 'de verilen kısmi en çok olabilirlik fonksiyonuna sansürlenmiş birimler katılmaz. Olayın meydana geldiği her başarısızlık anı için risk kümesi yeniden oluştuğundan dolayı en çok olabilirlik fonksiyonu yalnızca başarısızlık zamanlarının sıralanmasına bağlıdır. Olabilirlik fonksiyonunda sağkalım süresi gözlemlenmiş birimlerin yanı sıra sansürlenmiş gözlemler de hesaba katılırsa kısmi en çok olabilirlik fonksiyonu sansürlenmeye uygun forma gelir.

$$\delta_i = \begin{cases} 0, & sans\ddot{u}rlenme\ var\ ise \\ 1, & sans\ddot{u}rlenme\ yok\ ise \end{cases}$$

δ_i sansürlenme değişkeni olmak üzere, Cox orantılı tehlike modeli için kısmi en çok olabilirlik fonksiyonu aşağıdaki şekilde ifade edilir (Wang ve ark., 2019):

$$l(\beta) = \prod_{j=1}^{N} [\frac{exp(\beta^T X_j)}{\sum_{l\epsilon\, R(t_j)} exp(\beta^T X_l)}]^{\delta_i} \quad (3.13)$$

Kısmi en çok olabilirlik fonksiyonunun logaritması alınarak aşağıdaki şekilde log-kısmi olabilirlik fonksiyonu elde edilir (Wang ve ark., 2019).

$$\log l(\beta) = \log \left[\prod_{j=1}^{N} \frac{exp(\beta^T X_j)}{\sum_{l\epsilon\, R(t_j)} exp(\beta^T X_l)} \right]^{\delta_j}$$

$$\log l(\beta) = \sum_{j=1}^{N} \delta_j \{ \beta^T X_j - log \sum_{l\epsilon\, R(t_j)} exp(\beta^T X_l) \} \quad (3.14)$$

Orantılı tehlike regresyon modelindeki β katsayıları, log-kısmi en çok olabilirlik fonksiyonunun maksimize edilmesiyle bulunur. β katsayılarının tahmin edilmesinde Newton Rapson metodu kullanılır . Herhangi bir bağımsız değişken için elde edilen pozitif katsayı riskin yüksek, prognostik etkenlerin kötü olduğuna yorumlanırken, negatif bir katsayı gözlemler için daha iyi prognostik etkenlerin olduğu şeklinde yorumlanır (Özdemir, 2015).

3.2.1.1 Newton Rapson yöntemi

$u(\beta)$; kısmi en çok olabilirlik fonksiyonunun logaritmasının β 'ya göre birinci türevlerinden oluşan $p \times 1$'lik vektör olsun. Bu vektör "etkin skorlar vektörü" olarak isimlendirilmektedir. $I(\beta)$; kısmi en çok olabilirlik fonksiyonunun logaritmasının β 'ya göre ikinci türevlerinden oluşan $p \times p$ boyutlu matris olsun. Bu matris "tahmin edilen bilgi matrisi" olarak isimlendirilmektedir. Newton-Rapson adımsal bir yöntemdir. Bu metot $(s+1)$ döngüsünde iken, β parametre vektörünün tahmini $\hat{\beta}_{s+1}$ olur.

$$\hat{\beta}_{s+1} = \widehat{\beta}_s + I^{-1}(\widehat{\beta}_s) u(\widehat{\beta}_s) \quad (3.15)$$

$s = 1,2,\dots,n$ için $u(\widehat{\beta}_s)$ etkin skorlar vektörü olup $I^{-1}(\widehat{\beta}_s)$ bilgi matrisinin tersidir. İşleme $\widehat{\beta}_0 = 0$ değeri ile başlanabilir. Kısmi log-en çok olabilirlik fonksiyonundaki değişim yeterince küçük olduğunda döngü sona erdirilir (Collet, 1993) .

3.2.1.2 β parametrelerine ilişkin hipotez testleri ve test edilmesi

β parametresine ilişkin ;

$H_0: \beta_1 = \beta_2 = \cdots = \beta_P$

$H_1: \beta_1 \neq \beta_2 \neq \cdots \neq \beta_P$

Hipotezini test etmek için kullanılan test istatistikleri; olabilirlik oran test istatistiği, Wald test istatistiğidir (Collet, 1993).

- **Olabilirlik oran test istatistiği**

Kategorik değişken sayısının iki veya daha fazla olduğu durumlarda Cox regresyon modelinde aynı anda birden çok açıklayıcı değişkenin test edilmesinde olabilirlik oran test istatistiği kullanılır.

$i = 1,2,\dots,p$ için herhangi bir β_i parametresinin anlamlılığının testi için olabilirlik oran test istatistiği elde edilirken X_i'yi içermeyen yani indirgenmiş modelin en çok olabilirlik fonksiyonunun logaritmasının -2 katından, X_i'yi içeren yani tam modelin en çok olabilirlik fonksiyonunun logaritması çıkarılır. Sonuç olarak elde edilen test istatistiğinin p serbestlik dereceli ki-kare dağılımından olduğu kabul edilir. Aşağıdaki şekilde ifade edilir (Collet, 1993):

$$\chi^2_{LR} = 2[lnL(\widehat{\beta}) - lnL(0) \qquad (3.16)$$

- **Wald test istatistiği**

En çok olabilirlik tahmin edicilerinin normal dağıldığı varsayımına dayanır. $\hat{s}(\widehat{\beta}_i)$, $\widehat{\beta}_i$'nin tahmin edilen standart sapması olmak üzere;

$$\frac{\widehat{\beta}_i}{\hat{s}(\widehat{\beta}_i)} \approx \mathrm{Z} \tag{3.17}$$

dir.

Test istatistiğinin standart normal dağıldığı varsayılır. $\hat{V}(\widehat{\beta})$, $\widehat{\beta}'$nın varyans kovaryans matrisi olmak üzere $\widehat{\beta}'[\hat{V}(\widehat{\beta})]^{-1}\widehat{\beta}$ ifadesi p serbestlik dereceli asimptotik ki-kare dağılımına sahiptir. Wald test istatistiği aşağıdaki şekildedir:

$$\chi_w^2 = \widehat{\beta}'[\hat{V}(\widehat{\beta})]^{-1}\widehat{\beta} \tag{3.18}$$

3.2.2 Düzenlileştirilmiş Cox regresyon modelleri

Veri bulma ve toplama tekniklerinin gelişmesiyle birlikte çoğu gerçek dünya alanı yüksek boyutlu verilerle karşılaşma eğilimindedir. Bazı durumlarda, verilerdeki değişken sayısı, gözlem sayısına eşit veya büyüktür. Tüm değişkenler kullanılarak elde edilecek tahmin modeli aşırı öğrenme problemi nedeniyle hatalı sonuçlar verebilir. Değişkenlerinin tümünün önemli olmadığı varsayımı altında seyreklik normları (sparsity norms) kullanılarak hayati önem taşıyan değişkenleri seçilebilir. Doğrusal regresyon modelleri için geliştirilen değişken seçim teknikleri sağkalım modelleri bağlamına genişletilmiştir. On binlerce öznitelik arasında sonuç değişkenine en uygun değişkenleri belirlemek amacıyla lasso, grup lasso, grafik lasso gibi farklı ceza fonksiyonları da seyrek öğrenme yöntemlerini kullanarak modelin gelişimine katkı sağlar. l – norm cezalandırma fonksiyonları ailesi $l_\gamma : R^\gamma \to R, \quad l_\gamma(\beta) = ||\beta||_\gamma = (\sum_{i=1}^{P}\left||\beta_i|\right|^\gamma)^{\frac{1}{\gamma}}, \ \gamma > 0$ şeklinde yaygın olarak kullanılan ceza fonksiyonlarıdır. Yaygın olarak kullanılan düzenlileştirilmiş Cox yöntemleri aşağıdaki şekildedir (Wang ve ark., 2019):

- Lasso-Cox
- Ridge-Cox
- EN-Cox.

En Küçük Mutlak Küçülme ve Seçim Operatörü (Least Absolute Shrinkage and Selection Operator, LASSO) (Tibshirani R. , 1996) değişken seçimi yapmak ve regresyon katsayılarını tahmin etmek için kullanılan l_1-norm düzenlileştiricisidir. LASSO yöntemi

kullandığı ceza parametresinden dolayı bazı değişken katsayılarını küçültürken bazı değişken katsayılarını da sıfıra indirir. Tibshirani (Tibshirani R. , 1997) değişken seçiminde l_1-norm düzenlileştiricisinin özelliklerini esas alan Lasso-Cox algoritmasını elde etmek için l_1 cezalandırma katsayısını Denklem (3.14) 'de verilen log-kısmi olabilirlik fonksiyonuna aşağıdaki şekilde dahil etti.

$$\arg\min\{-\sum_{j=1}^{n}\delta_j[\beta^T X_j - \log(\sum_{l\epsilon\, R(t_j)} exp(\beta^T X_l)] + \lambda||\beta||_1\} \qquad (3.19)$$

Eşitlik (3.19)'da $||\beta||_1 = \sum_{p=1}^{p}|\beta_p|$'ye eşittir ve l_1-norm olarak bilinir. λ ise çapraz doğrulama ile belirlenen LASSO ayar parametresidir.

Hoerl ve Kennard (Hoerl ve Kennard, 1970) tarafından önerilen Ridge regresyonu Verweij ve ark. tarafından Cox regresyonunda başarıyla kullanılmıştır. Ridge Cox yöntemi için düzenlileştirme parametresi $\lambda\sum_{p=1}^{P}{\beta_p}^2$'dir. Ridge regresyona aynı zamanda l_2 düzenlileştirme formu da denir. Aralarında yüksek korelasyon bulunan değişkenleri seçerek katsayılarını küüçültmek için l_2-norm düzenlileştirici kullanılır. Ridge regresyonda beta parametrelerinde yapılacak düzeltme oranı çapraz doğrulama ile tespit edilen λ değeri ile belirlenir.

l_1 ve l_2 düzenlileştiricilerini birleştiren Elastik net (EN) hem değişken seçimi hem de değişkenler arasındaki korelasyon problemini aynı anda ele alma potansiyeline sahiptir (Zou ve Hastie, 2005). EN-Cox Noah Simon ve ark. tarafından önerilen çalışmada $\lambda[\alpha\sum_{p=1}^{P}|\beta_P| + \frac{1}{2}(1-\alpha)\sum_{p=1}^{P}\beta_p^2]$, $0 \le \alpha \le 1$ şeklindeki EN-Cox düzenlileştiricisi negatif log kısmi olabilirlik fonksiyonuna eklenmiştir (Wang ve ark., 2019).

3.3 Parametrik Yöntemler

Parametrik sağkalım analiz modelleri için, zamanın olasılık yoğunluk fonksiyonu olan $f(t)$'nin bilinmeyen parametreler cinsinden ifade edilebilen bir dağılımı takip ettiği varsayılır. Sağkalım süreleri için olasılık yoğunluk fonksiyonu belirtildiğinde, olasılık yoğunluk fonksiyonuna karşılık gelen $S(t)$ sağkalım fonksiyonu ve tehlike (hazard) fonksiyonu bulunabilir (Kleinbaum ve Klein, 2010).

Cox tabanlı yarı parametrik modellere önemli alternatif sunan bu modeller ilgilenilen olayın meydana geldiği zamanı tahmin etme konusunda basit, verimli ve etkili bir yol sağlar. Aynı zamanda birçok farklı uygulama alanında da yaygın olarak kullanılır. Parametrik sağkalım modelleri, teorik sağkalım dağılımıyla tutarlı olan tahminler üretme eğilimindedir. Parametrik sansürlü regresyon modellerinde sağkalım süreleri için yaygın olarak kullanılan dağılımlarından bir kaçı; üssel, weibull, log-normal ve log-lojistik dağılımlardır. Veri kümesindeki tüm birimler için zamanın olasılık yoğunluk fonksiyonu $f(t)$ bu dağılımları takip ediyorsa model doğrusal regresyon modeli olarak adlandırılır. $f(t)$' nin logaritması dağılımları takip ediyorsa araştırmaya konu olan problem hızlandırılmış başarısızlık modeli kullanılarak analiz edilir. $f(t)$ için uygun bir teorik dağılım bilinmiyorsa, parametrik olmayan yöntemler daha etkilidir Parametre tahmini için sağkalım sürelerinin tehlike (hazard) fonksiyonunun doğru belirlenebilmesi oldukça önemlidir. Kullanılan dağılımın tipine göre tehlike (hazard) fonksiyonu farklılaşmaktadır (Wang ve ark., 2019).

Parametrik sağkalım analizi modellerinin parametrelerini tahmine etmek için Maksimum Olabilirlik Tahmini (MLE) yöntemi kullanılır. Birim sayınısın N, sansürlü gözlem sayısının c, sansürlenmemiş gözlem sayısının $(N-c)$, $(\beta_1, \beta_2, \dots, \beta_p)^T$ tüm parametrelerin kümesi olsun. Sağkalım süresinin olasılık yoğunluk fonksiyonu $f(t)$ ve sağkalım fonksiyonu $S(t)$, sırasıyla $f(t,\beta)$ ve $S(t,\beta)$ ile temsil edilebilir. Belirli bir i. örnek sansürlenirse, gerçek sağkalım süresi mevcut olmayacaktır. Bununla birlikte i. birimin ilgilenilen olayı c_i sansürlenme süresinden önce yaşamadığı açıktır. Bu nedenle $S(c_i,\beta)$ sağkalım fonksiyonunun 1'e yakın olacağı sonucuna varabiliriz. Tersine, örneğin olay belirli bir T_i noktasında meydana gelirse $f(T_i,\beta)$ yüksek olasılık değerine sahip olacaktır. Sansürsüz gözlemlerin tümünün ortak olasılık yoğunluk fonksiyonu olarak $\prod_{\delta_{i=1}} f(T_i,\beta)$, sansürlü gözlemlerin ortak sağkalım fonksiyonu $\prod_{\delta_{i=0}} S(T_i,\beta)$ olarak tanımlansın. N örneğin olabilirlik fonksiyonu aşağıdaki şekilde optimize ederek β parametrelerini tahmin edebiliriz (Wang ve ark., 2019).

$$L(\beta) = \prod_{\delta_{i=1}} f(T_i,\beta) \prod_{\delta_{i=0}} S(T_i,\beta) \qquad (3.20)$$

3.3.1 Üssel model

Üssel model , λ sabit tehlike (hazard) oranı ile tek parametre kullanılarak karakterize edildiğinden dolayı sağkalım analizi için kullanılan parametrik modeller arasında en basit ve en önemlisidir . Büyük λ değerleri yüksek risk ve kısa sağkalım süresini gösterirken, küçük λ değerleri düşük risk ve yüksek sağkalımı ifade etmektedir. Üstel dağılımın olasılık yoğunluk fonksiyonu, sağkalım fonksiyonu ve hazard fonksiyonu sırasıyla (Lee, 1992);

$$f(t) = \lambda \exp(-\lambda t) \qquad (0 \leq t < \infty) \tag{3.21}$$

$$S(t) = \exp(-\lambda t) \qquad (0 \leq t < \infty) \tag{3.22}$$

$$h(t) = \lambda \qquad (0 \leq \mathrm{t} < \infty) \tag{3.23}$$

Denklem 3.22'in logaritması alınarak $\log S(t) = -\lambda t$ elde edilir. Yani üstel dağılımın sağkalım fonksiyonunun logaritması ile zaman λ eğimi ile lineer ilişki içerisindedir. O halde; $\hat{S}(t)$ sağkalım fonksiyonunun KM tahmini olmak üzere $\log \hat{S}(t)$ 'nin zamana karşı grafiğini çizdiğimizde üstel bir dağılım izleyip izlemediğini belirleyebiliriz.

3.3.2 Weibull model

Weibull dağılımı üssel dağılımın genelleştirilmiş bir halidir. $\lambda > 0$ ve $\mathrm{k} > 0$ olmak üzere iki parametre ile karakterize edilen Weibull modeli, sağkalım problemlerinde en yaygın kullanılan parametrik dağılımdır. k şekil parametresi kullanılarak tehlike (hazard) fonksiyonunun şekli belirlenir. Weibull dağılımı şekil parametresi ile üstel dağılımın sabit hazard varsayımının getirdiği kısıtlamaları ortadan kaldırdığından dolayı üstel dağılımın aksine esnek yapıdadır (Lee, 1992).

$\mathrm{k} = 1$ ise; Weibull dağılımı üstel dağılıma indirgenir. Bu durumda tehlike(hazard) fonksiyonu zamandan bağımsız olarak sabittir. $\mathrm{k} > 1$ olduğu durumda tehlike (hazard) fonksiyonu monoton artan yapıdadır. Yani t büyüdükçe birimlerde ilgilenilen olayın meydana gelme riski artar. Aksi durumda yani $\mathrm{k} < 1$ ise; tehlike (hazard) fonksiyonu monoton azalandır. t değeri büyüdükçe birimlerde ilgilenilen olayın meydana gelme riski azalmaktadır (Lee, 1992).

$$f(t) = \frac{1}{\lambda^k} kt^{k-1} \exp\left[\left(-\frac{t}{\lambda}\right)^k\right] \qquad (0 \leq t < \infty) \tag{3.24}$$

$$S(t) = \exp\left[\left(-\frac{t}{\lambda}\right)^k\right] \qquad (0 \leq t < \infty) \tag{3.25}$$

$$h(t) = \frac{1}{\lambda^k} kt^{k-1} \qquad (0 \leq \mathrm{t} < \infty) \tag{3.26}$$

$\hat{S}(t)$ sağkalım fonksiyonunun KM tahmini olmak üzere olmak $\log(-\log \hat{S}(t))$ ile $\log(t)$ fonksiyonu lineer ilişki içerisindedir. Bu fonksiyonların değerlerine ait grafikler çizdirilerek sağkalım sürelerinin Weibull dağılımına uygunluğu konusunda önemli bilgiler elde edilebilir (Lee, 1992).

3.3.3 Log normal model

Log normal dağılım en basit şekilde logaritma değerleri normal dağılım gösteren değişkenlerin dağılımı olarak tanımlanabilir. Dağılımın tehlike (hazard) oranı önce hızlı bir şekilde artan eğilim gösterir. Ortanca (medyan) değeri geçtikten sonra sıfıra doğru azalır. Bu nedenle log normal dağılım önce artan daha sonra azalan hazard oranına sahip sağkalım verilerine oldukça uygun bir sağkalım dağılımıdır (Dey ve Kundu, 2009).

Log normal dağılımın $\mu\ (>0)$ ve $\sigma\,(>0)$ olmak üzere iki adet parametresi bulunmaktadır. Log normal dağılım normal dağılımla ilişkilidir, fakat temsil edilen rasgele değişken yalnızca pozitif değeler alabileceği varsayımı bulunmaktadır. ϕ standart normal dağılımın olasılık fonksiyonu olmak üzere; log normal dağılımın olasılık yoğunluk fonksiyonu, sağkalım fonksiyonu ve tehlike (hazard) fonksiyonu aşağıda verilmiştir (Dey ve Kundu, 2009).

$\phi = \frac{1}{\sqrt{2\pi}} \exp(-\frac{t^2}{2})$ ve $a = \exp(-\mu)$ olmak üzere;

$$f(t) = \frac{1}{\sqrt{2\pi}t\sigma} \exp\left[-\frac{1}{2\sigma^2}(\log(at))^2\right] \qquad (0 < t < \infty) \tag{3.27}$$

$$S(t) = 1 - \ \phi\left(\log\frac{at}{\sigma}\ \right) \qquad (0 < t < \infty) \tag{3.28}$$

$$h(t) = \frac{\frac{1}{\sqrt{2\pi}t\sigma}\exp\left[-\frac{1}{2\sigma^2}(\log(at))^2\right]}{1 - \phi\,(\log\frac{at}{\sigma}\,)} \qquad (0 < t < \infty) \qquad (3.29)$$

3.3.4 Log lojistik model

Weibull modelin aksine log-lojistik modeller tehlike (hazard) fonksiyonunun monoton olmayan davranışına izin verir. Sağkalım süresi T ve sağkalım süresinin logaritması $\log(T)$ log-lojistik modellerde lojistik dağılımı takip eder.

Log lojistik dağılımı $\lambda\,(>0)$ ölçek ve $\mathrm{k}\,(>0)$ şekil parametresi olmak üzere bu iki parametre ile karakterize edilir. $\mathrm{k}\,(>1)$ koşuluyla $t = 0$ noktasından belirli bir t anına kadar artan ve maksimum noktaya ulaştıktan sonra azalan bir yapı gösterir. $\mathrm{k} = 1$ için tehlike (hazard) fonksiyonu $\lambda^{\frac{1}{k}}$ noktasından itibaren monoton azalır. Dolayısıyla t değeri büyüdükçe tehlike (hazard) oranı azalır. Log lojistik dağılıma ait olasılık yoğunluk fonksiyonu, sağkalım fonksiyonu ve tehlike (hazard) fonksiyonu aşağıdaki şekildedir (Dey ve Kundu, 2009).

$$f(t) = \frac{\lambda \mathrm{k} t^{k-1}}{(1 + \lambda \mathrm{t}^k)^2} \qquad (0 \le t < \infty) \qquad (3.30)$$

$$S(t) = \frac{1}{1 + \lambda \mathrm{t}^k} \qquad (0 \le t < \infty) \qquad (3.31)$$

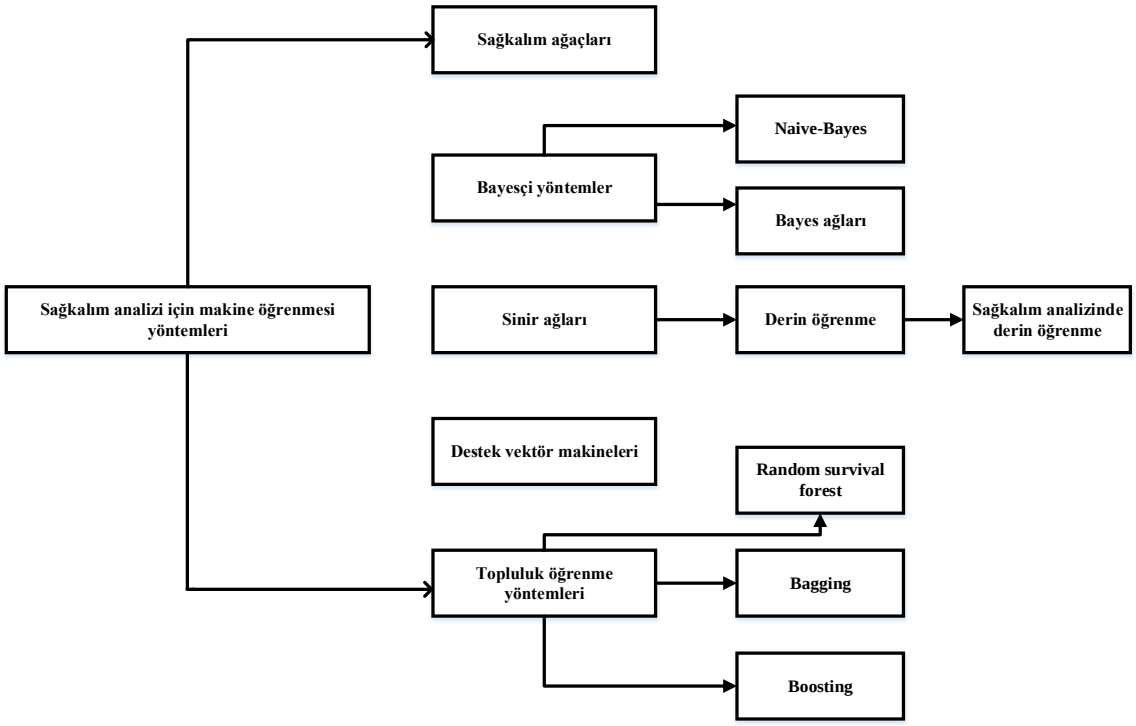
$$h(t) = \frac{\lambda \mathrm{k} t^{k-1}}{1 + \lambda \mathrm{t}^k} \qquad (0 \le t < \infty) \qquad (3.32)$$

4.SAĞKALIM ANALİZİ İÇİN MAKİNE ÖĞRENMESİ YÖNTEMLERİ

Son yıllarda farklı alanlarda yapılan araştırmalarda doğrusal olmayan ilişkileri modelleme ve tahminleme performansı açısından makine öğrenmesi yöntemleri kullanılarak önemli başarılar elde edilmiştir. Sağkalım analizinde makine öğrenmesi yöntemlerinin karşılaştığı temel problemler; sansürlü gözlemler ve sansürlü gözlemlerde sağkalım süresinin gerçek bilgiyi vermemesidir. Bu bölümde, sağkalım analizinde kullanılan makine öğrenmesi yöntemleri verilecektir. Şekil 4.1' de sağkalım analizi için makine öğrenmesi yöntemleri gösterilmektedir. (Wang ve ark., 2019).

Gordon ve Olshen tarafından 1985 yılında yayımlanan makale ile geleneksel regresyon ve sınıflandırma ağacı yöntemlerinin bir uzantısı olarak sağkalım ağaçları tanıtılmıştır. Sağkalım ağaçları, sansürlenmiş verileri modellemek için özel olarak uyarlanmış sınıflandırma ve regresyon ağaçlarıdır (Gordon ve Olshen, 1985). Geleneksel ağaç modellerinin temelinde, verilerin belirli bir bölme kriterine göre yinelemeli olarak bölünmesi ve ilgilenilen olaya bağlı olarak birbirine benzeyen nesnelerin aynı düğüme yerleştirilmesi vardır (Fadnavis, 2019). Sağkalım ağacı ağaç yapısını kullanarak sansürlenmiş verileri işleme yeteneği ile standart karar ağaçlarına göre avantaj sağlar (Wang ve ark., 2019).

Matematiksel istatistik ve olasılık teorisindeki en temel ilkelerden biri olan Bayes Teoremi sonsal olasılık ile önceki olasılık arasında bir bağlantı sağlar . Böylelikle Bayes Teoremi ile belirli bir olayı hesaba katmadan önce ve hesaba kattıktan sonra olasılık değerlerinin nasıl değiştiği görülebilir (Wang ve ark., 2019). Bayes Teoremi koşullu olasılıktan faydalanarak bir sonucun hangi olasılıkla hangi sebepten kaynaklandığını bulmaya yardımcı olur (Sırın, 2017). Naive Bayes ve Bayesian ağı (Bayesian network), yaygın olarak klinik tahmin için kullanılan ve çıktı olarak ilgili bir olayın olasılığını sağlayan modellerdir (Friedman ve ark., 1997).



Şekil 4. 1 Sağkalım analizi için makine öğrenmesi yöntemleri

Naive Bayes, makine öğrenmesi yöntemleri arasında basit ama en etkili tahmin algoritmalarından birisidir. Naive Bayes modeli bir koşullu bağımsızlık varsayımı yapar. Bu varsayıma göre her bir özellik bağımsız olarak ele alınır. Sağkalım analizinde ele alınan pek çok problem için doğru olmayabilecek tüm özellikler için bağımsızlık varsayımı Naive Bayes yönteminin bir dezavantajıdır. Özniteliklerin birbirleriyle birkaç farklı düzeyde ilişkilendirilebildiği bir Bayesian ağı, bir dizi değişken üzerinde teorik bir dağılımı grafiksel olarak ifade eder. Bayesian ağları ile değişkenler ve değişkenler arası olasılıksal ilişkilerin gösterimi grafiksel olarak sağlanır (Wang ve ark., 2019).

Frank Rosenblatt 1958'de birbirine bağlı yüz milyarlarca nöronun bilgiyi paralel olarak işlediği insan beyninin karmaşık işlevselliğinden ilham alarak Yapay Sinir Ağları (YSA) üzerine ilk makaleyi yayımladı. Bir YSA nöronların (veya düğümlerin, birimlerin) oluşturduğu bir giriş katmanından (input layer), bir veya iki gizli katmandan (hidden layers) ve bir çıktı katmanından (output layer) oluşur (Baesens ve ark., 2005). Nöronlar biyolojik bir sinir ağını simüle eden bir ağ oluşturmak ağırlıklı bir bağlantıya dayalı olarak bağlanır. Bu bağlamda bir nöron, uyarlanabilir ağırlık kümelerinden oluşan ve belirli bir aktivasyon işlevine göre çıktı üreten bir hesaplama öğesidir. YSA'lar sağkalım analizinde yaygın olarak kullanılır. Sağkalım analizi problemlerini çözmek için literatürde üç ana yapay sinir ağı yöntemi önerilmiştir (Wang ve ark., 2019).

(1) Bir nesnenin verilen girdilerden doğrudan sağkalım süresini tahminlemek için sağkalım analizinde YSA kullanılmıştır.

(2) Faraggi ve Simon 1995'de Cox oransal hazard modelindeki doğrusal tahminleyiciyi YSA'nın doğrusal olmayan çıktısıyla değiştirerek doğrusal olmayan oransal hazard modelini önerdiler (Faraggi ve Simon, 1995). Mariana ve ark. 1997'de meme kanseri nüksünün prognostik etkenlerini değerlendirmek için hem geleneksel Cox modelini hem de Faraggi ve Simon tarafından önerilen YSA yöntemini kullandılar. Yapılan deneyler ile YSA yönteminin Cox oransal hazard modelinden daha iyi performans göstermediğini ortaya koydu. Cox oransal hazard modelinin bu uzantıları, geleneksel Cox yönteminin avantajlarını korusa da modellemede optimal seviyeye ulaşamadı (Mariani ve ark., 1997).

(3) Hem Radvin ve Clark hem de Biganzoli ve ark. bir gözlemi, farklı zaman aralıkları ve sağkalım durumlarına sahip bir dizi gözlem olarak kodlamıştır. Daha sonraki zaman aralıklarında sansürlenen gözlemler hariç tutuldu. Yaptıkları çalışmalarda ek bir tahminleyici değişken olarak bir zaman aralığını ve bu zaman aralığındaki çıktı değişkeni olarak sağkalım olasılığını veya ayrık tehlike oranını benzer YSA mimarisi ile benimsemişlerdir. Radvin ve Clark tahmin sonuçlarının "sağkalım olasılığı ile kabaca orantılı" olduğunu belirtirken Biganzoli ve ark. monoton azalan sağkalım olasılığına kolayca aktarılabilen tehlike oranını doğrudan modelledi. Bu nedenle YSA ile kısmi lojistik regresyon modelleri (PLANN) tercih edilebilir. Ancak Radvin ve Clark sansürlü gözlemler nedeniyle daha sonraki zaman aralıklarında gözlemlerin büyük ölçüde ölen bireyleri temsil ettiğine dikkat çekti. Bu yanlılığı düzeltmek için, önce bu zaman aralığındaki sağkalım olasılığını belirlemek için Kaplan-Meier tahminini ve oranı dengelemek için rasgele seçilmiş sansürlü gözlemleri kullandılar (Ravdin ve Clark, 1992) (Biganzoli ve ark., 1998).

Son yıllarda yapılan çeşitli uygulama alanlarında sağkalım analizi problemlerinin üstesinden gelmek için derin öğrenme yöntemleri büyük ilgi görmüştür. Örneğin, sağlık hizmetleri alanında, hasta verilerinin çoklu modaliteleri için karmaşık etkileşimleri verimli bir şekilde öğrenmek için derin korelasyonel hayatta kalma modelleri ve Konvolüsyonel Sinir Ağları (Convolutional Neural Networks-CNN) kullanıldı. Hastalığa etki eden prognostik faktörleri tahmin ederek her bir hastaya kişiselleştirilmiş tedavi önerileri sunarak doktorlara hastalarla ilgili klinik kararlarından yardımcı olmak için çeşitli derin sağkalım analizi yaklaşımları da önerilmiştir. Ayrıca Tekrarlayan Sinir Ağı (Reccurrent Neural Network-RNN) tabanlı yaklaşımlar tekrar eden olayları incelemek için özellikle kullanıcı davranışı modelleme uygulamalarında sağkalım analizi ile başarılı bir şekilde birleştirilmiştir (Wang ve ark., 2019).

Doğrusal olmayan problemleri çözmedeki performansı diğer geleneksel öğrenme yöntemleriyle karşılaştırıldığında çok başarılı bir denetimli öğrenme yaklaşımı olan Destek Vektör Makineleri (DVM) çoğunlukla sınıflandırma problemleri için kullanılsa da regresyon problemleri için de kullanılmaktadır. DVM sağkalım analizi problemlerine de başarıyla uyarlanmıştır. DVM yöntemi yalnızca ε-duyarsız kayıp fonksiyonunu

(insensetive loss function) $f(X_i) = maks(0, |fX_i - y_i| - \varepsilon)$ düzenleyiciler ile minimize eder. Ancak bu yaklaşımın temel dezavantajı sansürlü gözlemlerin sıralama bilgisinin tamamen göz ardı edilmesidir. Sansürlenmiş verilerin üstesinden gelmenin bir diğer yolu, kısıtlandırma sınıflandırma yaklaşımını kullanarak DVM uygulamaktır. Bu yöntemde karşılaştırılabilir iki örneklem için DVM formülasyonuna kısıtlamalar getirilir. Büyük veri kümelerinde bu algoritmanın hesaplama karmaşıklığı açıktır. Ayrıca yalnızca örneklemler arasındaki sıralamayı inceler ve çıktıların gerçek değerlerini göz ardı eder.

Bu duruma karşı koymak için Khan ve Zubek 2008'de sansürlenmiş sağkalım verileri ile başa çıkmak için hem sınıflandırma hem de regresyon problemlerinde çok başarılı bir denetimli öğrenme yaklaşımı olan Destek Vektör Makineleri (DVM)'den yararlanarak sansürlü veriler için DVM kullanılmasını önerdiler. Güncellenmiş bir asimetrik kayıp fonksiyonu kullanıp DVM'den yararlandıkları için bu yönteme SVRc (Support Vector Regressionfor Censored Data-SVRc) adını verdiler. Yaklaşımında temelinde ε-duyarsız kayıp fonksiyonunu (insensetive loss function) asimetrik olarak özellikle düzenlileştirme parametreleri-γ (regularization parameters) ve hata marjı-δ ile değiştirmektir (Khan ve Zubek, 2008). Van ve ark. 2007'de yaptıkları çalışmada ortak değişkenler ile sonuç değişkeni arasında proxy işlevi gören bir sağlık indeksi sunarak sağ sansürlü sağkalım verilerinin tahminine dayalı bir öğrenme makinesi tasarladılar (Van ve ark., 2007). Aynı grup 2011'de sıralama ve regresyon yöntemlerini sağkalım analizi bağlamında birleştiren DVM tabanlı bir yaklaşım önerdi (Wang ve ark., 2019) .

Son yıllarda sansürlenmiş gözlemlerle başa çıkarak bu gözlemlere ilişkin tahminleme yapmak için bir dizi gelişmiş makine öğrenmesi yöntemi önerilmiştir. Bu yöntemler bu bölümde şu ana kadar bahsedilen diğer yöntemlere kıyasla sağkalım verilerinin analizinde benzersiz avantajlar sunar.

Topluluk (ensemble) öğrenme yöntemleri sınıflandırıcı komitesi oluştup ardından tüm bu sınıflandırıcılardan gelen tahmin sonuçları arasında ağırlıklı bir oylama yapma temeline dayanır. Sonrasında yeni veri noktaları geldikçe gelen birimlerin sınıf etiketlerini tahmin eder. Özellikle yetersiz veri varlığında başlangıç noktalarını değiştirerek iyi bir topluluk oluşturmak ve bilinmeyen fonksiyona daha iyi bir yaklaşım elde etmek genellikle

mümkündür. Torbalama yöntemi ile düşük yanlılık miktarına sahip fakat yüksek varyanslı olan değişkenler kullanılarak değişkenler daha elverişli hale getirilir.

Tek bir yöntemin kararsızlığının üstesinden gelmek için, topluluğa dayalı model oluşturmak için; 1996'da Brieman tarafından önerilen torbalama (bagging) ve 2001' de yine Brieman tarafından önerilen Rastgele Ormanlar (Random Forests-RF) yöntemi tercih edilir (Wang ve ark., 2019). Bu tür topluluk modelleri sağkalım analizine başarılı şekilde uygulanmıştır. Bagging sağkalım ağaçlarında oylama ile etiket tahmini yerine toplu (aggregated) sağkalım fonksiyonu sağkalım ağacı tarafından yapılan tahminlerin ortalaması alınarak hesaplanabilir. Bu yöntemde aşağıdaki şekilde üç ana adım vardır :

(1) Orijinal sağkalım verisinden B tane boostrap örneklemi çekilir.
(2) Her bir bootstrap örneklemi için bir sağkalım ağacı oluşturulur. Tüm terminal düğümlerde olay sayısının eşik değer d'ye eşit veya d eşiğinden büyük olması sağlanır.
(3) Yaprak (leaf) düğümlerin tahminlerinin ortalaması alınarak toplu (aggregated) sağkalım fonksiyonu hesaplanır. Her yaprak düğüm için sağkalım fonksiyonu KM tahmin edicisi kullanılarak tahmin edilir. Aynı düğümdeki tim birimlerin aynı sağkalım fonksiyonuna sahip olduğu varsayılır (Brieman, 2001).

4.1 Rastgele Sağkalım Ormanları (Random Survival Forests) Yöntemi

Breiman tarafından 2001'de topluluk öğrenme sürecine rastgelelik enjekte ederek daha da geliştirilebileceği rastgele ormanlar-RO (Random Forests- RF) adı verilen yöntem ile gösterildi (Brieman, 2001). RF yöntemi kullanılarak yapılan ilk çalışmalarda, regresyon ve sınıflandırma problemlerine odaklanılmıştır. Rastgele sağkalım ormanları-RSO (Random Survival Forests-RSF) yöntemi, Ishwaran ve ark. tarafından 2008'de sağ sansürlü sağkalım verilerinin analizinde kullanılmak üzere RF yönteminin genişletilmesiyle tanıtıldı (Wang ve ark., 2019).

RSF, topluluk öğrenme yöntemidir ve birden fazla sağkalım ağacının çıktılarını birleştirerek risk tahmin modeli meydana getirir. RSF, rastgele örneklemeyi ve topluluk öğrenme yöntemlerinin geliştirilmiş özelliklerini barındırması nedeniyle iyi genellemeler sunarak geçerli tahminlemelerde bulunur. RSF'nin uygulaması, Brieman'ın RF yöntemi ile aşağıdaki şekilde aynı genel ilkeleri takip eder (Ishwaran ve ark., 2008):

(1) Sağkalım ağaçları, bootstrap verileri kullanılarak büyütülür.

(2) Rastgele özellik seçimi, ağaç düğümlerini bölerken kullanılır.

(3) Ağaçlar genellikle derinlemesine büyür.

(4) Sağkalım ormanı topluluğu, terminal düğüm istatistiklerinin ortalaması alınarak hesaplanır (Ishwaran ve ark., 2008a).

Sansür varlığı, regresyon ve sınıflandırma için yapılan RF yöntemine kıyasla RSF uygulamasının belirli yönlerini karmaşıklaştıran; sağkalım verilerinin benzersiz özelliğidir. Sağ sansürlü sağkalım verilerinde; T olay zamanı, δ sansürleme göstergesi olmak üzere gözlemlenen veriler (T, δ) ikilisi ile gösterilir. Gerçek sağkalım süresi T^0 ve sansürlenme süresi C^0 olmak üzere gözlemlenen sağkalım süresi $\min(T^0, C^0)$ olarak tanımlanır. Dolayısıyla gerçek olay zamanı gözlemlenemeyebilir. Sansür göstergesi $\delta = I\{T^0 \leq C^0\}$ olarak tanımlanırsa; $\delta = 1$ olduğunda ilgilenilen olay meydana gelmiştir ve $T = T^0$'dır. Aksi taktirde $\delta = 0$ olduğunda, birim sansürlenir ve $T = C^0$ 'dır. Bu bölümde veriler $(T_1, X_1, \delta_1), \ldots, (T_n, X_n, \delta_n)$ olarak gösterilecektir. Burada X_i özellik vektörü, T_i gözlemlenen zaman ve δ_i i. birim için sansür göstergesidir (Ishwaran ve ark., 2008a).

4.1.1 Algoritma

RSF yönteminde kullanılan genel algoritma aşağıdaki şekildedir:

Adım (1): Verileri eğitmek için veri kümesinden B tane bootstrap örneklemi çekilir. Her bootstrap örneklemi rastgele değiştirilerek seçilen n adet birimden oluşur ve orijinal verinin %37 sini dışarıda bırakmalıdır. Dışarıda bırakılan veri out-of-bag (OOB) data olarak adlandırılır.

Adım (2): Her bootstrap örneklemi için bir ağaç büyütülür. Sağkalım ağacının her düğümünde , rastgele olarak p aday değişken seçilir. Düğüm, çocuk düğümler arasındaki sağkalım farkını maksimize eden aday değişken kullanılarak bölünür.

Adım (3): Her bir terminal düğümde, en az 1 adet ilgilenilen olay meydana gelen birim kalması gerektiği yani; $d_0 > 0$ kısıtı altında bölme işlemi devam ederek sağkalım ağacı büyütülür.

Adım (4): Büyütülen her bir sağkalım ağacı için kümülatif hazard fonksiyonu (KHF) hesaplanır. Topluluk KHF elde etmek için ortalama alınır.

Adım (5): OOB verisi kullanılarak topluluk KHF için tahmin hatası hesaplanır (Ishwaran ve ark., 2008).

4.1.2 Rastgele sağkalım ormanları ayırma kriterleri

RSF algoritmasında ayırma kriteri olarak; log-rank ayırma kriteri ve log-rank skor ayrımı kullanılır (Ishwaran ve Kogalur, 2007).

4.1.2.1 Log-rank ayırma kriteri

Geleneksel sağkalım analizinde Log-rank testi ile iki grup arasında anlamlı bir fark olup olmadığı görülebilir. Log-rank testi, takip süresinde birimleri eşit olarak ağırlıklandırır ve gruplar arasında sağkalım sürelerini karşılaştırmanının parametrik olmayan en yaygın yollarından birisidir. RSO yönteminde Log-rank bölme kriteri düğümler arasındaki sağkalım farkını en üst düzeye çıkartmak için bölme aracı olarak kullanılabilir.

Bölünecek bir ağaç düğümünü ele alalım. Bu düğümün kök düğüm (ağacın tepesi) olduğunu ve veriye boostrap yapılmadığını varsayalım. Dolayısıyla kök düğüm verileri $(T_1, X_1, \delta_1), \dots, (T_n, X_n, \delta_n)$ şeklindedir. X belirli bir değişken (özellik vektörünün koordinatlarından biri) olarak tanımlansın. X değişkeni kullanılarak c noktasına göre kök düğümde önerilen bölme, (kolaylık olması açısından X değişkeninin nominal olduğunu düşünüyoruz) $X \leq c$ ve $X > c$ biçiminde gösterilsin. Sırasıyla sağ ve sol çocuk düğümlere $L = \{X_i \leq c\}\ ve\ R = \{X_i > c\}$ olacak şekilde bölme gerçekleşir. $t_1 < t_2 < \cdots < t_m$ ayrık ölüm zamanları olsun. $d_{j,L}, d_{j,R}$ ve $Y_{j,L}, Y_{j,R}$ sırasıyla t_j zamanında olay meydana gelen ve risk altındaki birimlerin sayısı olsun. $Y_{j,L} = \#\{T_i \geq t_j,\ X_i \leq c\}$, $Y_{j,R} = \#\{T_i \geq t_j, X_i > c\}$ ise $Y_j = Y_{j,L} + Y_{j,R}$ ve $d_j = d_{j,L} + d_{j,R}$ olarak tanımlanır. X ortak değişkeninin c kesim değeri için log-rank test istatistiği aşağıdaki şekilde hesaplanır:

$$L(X,c) = \frac{\sum_{j=1}^{m}(d_{j,L} - Y_{j,L}\frac{d_j}{Y_j})}{\sqrt{\sum_{j=1}^{m}\frac{Y_{j,L}}{Y_j}(1-\frac{Y_{j,L}}{Y_j})(\frac{Y_j - d_j}{Y_j - 1})d_j}} \tag{4.1}$$

$L(X,c)$ değeri düğüm ayrımının bir ölçüsüdür. $L(X,c)$ ölçüm değeri ne kadar büyük olursa L ve R arasındaki sağkalım farkı o kadar büyük olur ve bölünme o kadar iyi gerçekleşir.

Ortak değişkenler arasında ve kesim değerleri arasında tüm X ortak değişkenleri ve c kesim noktaları için $|L(X^*, c^*)| \geq |L(X, c)|$ değerini veren X^*ortak değişkeni ve c^* kesim değeri bulunarak en iyi ayrım belirlenir (Ishwaran ve Kogalur, 2007).

4.1.2.2 Log-rank skor ayırma kriteri

Log-rank skor ayırma kriteri 2003'de Hothorn ve Lausen tarafından log-rank bölme kuralı temel alınarak geliştirilmiştir. X değişkeninin $X_1 \leq X_2 \leq \cdots \leq X_n$ şeklinde sıralandığını varsayalım. Burada X için n adet bağımsız değişken olduğunu varsayalım. Her bir T_j sağkalım süresi için ranklar aşağıdaki şekilde hesaplanır (Hothorn ve Lausen, 2003):

$$\alpha_j = \delta_j - \sum_{k=1}^{\Gamma_j} \frac{\delta_k}{n - \Gamma_k + 1} \tag{4.2}$$

Burada; $\Gamma_k = \#\{t: T_t \leq T_k\}$ ya eşittir. Log-rank skor testi aşağıdaki şekilde tanımlanır:

$$S(x, c) = \frac{\sum_{X_j \leq c} \alpha_j - n_L \bar{a}}{\sqrt{n_L(1 - \frac{n_L}{n}) s_a^2}} \tag{4.3}$$

Denklem (4.3)'de $\bar{a}$ ve s_a^2 sırasıyla rankların örneklem ortalaması ve örneklem varyansını göstermektedir. $S(x, c)$, değeri düğüm ayrımı için log-rank skor ölçüsü verir. Bu değeri maksimum yapan X ortak değişkeni ve c kesim değeri seçilir.

4.1.3 Topluluk kümülatif hazard fonksiyonu

RSF yönteminin temel unsurları; sağkalım ağacı geliştirmek ve geliştirilen her bir sağkalım ağacı için KHF hesaplayıp elde edilen bu değerleri kullanarak topluluk KHF elde etmektir. Bu bölümde topluluk KHF elde edilebilmesi için detaylar verilecektir (Ishwaran ve ark., 2008a).

4.1.3.1 İkili (binary) sağkalım ağacı

Karar ağacı oluşturmak için kullanılan CART algoritması 1984 yılında Breiman ve ark. tarafından önerilmiştir. Denetimli makine öğrenmesi yöntemi olan CART hem kategorik

hem de sürekli değişkenleri kullanan sınıflandırma ve regresyon ağacı algoritmasıdır. CART'a benzer şekilde, sağkalım ağaçları ağaç düğümlerinin yinelemeli bölünmesiyle büyüyen binary ağaçlardır. Bir ağaç tüm verileri içeren ağacın en tepe noktası olan kök düğümden başlayarak büyür. Kök düğüm belirlenmiş bir ayırma kriterine göre iki yavru düğüme ayrılır. Her yavru düğüm tekrar bölünerek sağ ve sol çocuk düğümleri doğurur. İşlem, her düğüm için özyinelemeli bir şekilde tekrarlanır.

Sağ ve sol düğümler arasındaki sağkalım farkını maksimize eden bölünme iyi bir bölünmedir. Bir düğüm için; tüm X ortak değişkenleri ve c kesim noktaları üzerinde arama yaparak düğümler arasında sağkalım farkını en üst düzeye çıkartan X^*ortak değişkeni ve c^* kesim değeri bulunarak en iyi ayrım belirlenir. Böylelikle büyütülen ağaçta düğümler arasında sağkalım farkı en üst düzeye çıkartılarak farklı durumlar birbirinden uzaklaştırılır. Nihayetinde düğüm sayısı arttıkça benzer durumlar birbirinden ayrılarak ağaçtaki her düğüm homojen hale gelir ve benzer sağkalım oranlarına sahip birimler aynı düğümlere toplanır (Ishwaran ve ark., 2008a).

4.1.3.2 Terminal düğüm tahmini

Her düğüm minimum bir $d_0 > 0$ farklı olay içerdiğinde, sağkalım ağacı bu kısıt altında doyma noktasına ulaşır ve çocuk düğümlere bölme işlemi tamamlanmış olur. Doyma noktasına ulaşmış sağkalım ağacında en uç düğümler terminal düğüm olarak adlandırılır. h sağkalım ağacının terminal düğüm ve $(T_{1,h}, \delta_{1,h}), \ldots, (T_{n(h),h}, \delta_{n(h),h})$ h terminal düğümündeki birimler için sağkalım süreleri ve sansür bilgisi olsun. $\delta_{i,h} = 0$ ise $T_{i,h}$ anında i. bireyin sağ sansürlendiği söylenir; aksi taktirde $\delta_{1,h} = 1$ ise $T_{i,h}$ anında birim ilgilenilen olayı yaşamıştır.$t_{1,h} < t_{2,h} < \ldots < t_{N(h),h}$ zamanları h terminal düğümündeki farklı ölüm zamanları $d_{j,h}$ ve $Y_{j,h}$ sırasıyla $t_{j,h}$ zamanında ölen ve risk altındaki birimlerin sayısı olsun. h için Nelson-Aalen tahmin edicisi aşağıdaki şekildedir (Ishwaran ve ark., 2008a):

$$\hat{H}_h(t) = \sum_{t_{j,h} \le t} \frac{d_{j,h}}{Y_{j,h}} \tag{4.4}$$

Bir h terminal düğümündeki tüm birimlerin KHF'si eşittir. x_i i. birimin d-boyutlu ortak değişkenler vektörüdür. Yukarıda bahsedildiği gibi x x_i'nin bir koordinatına karşılık gelir. $H(t|x_i)$ i. birim için KHF olsun. Bu değeri belirlemek için x_i sağkalım ağacında kök düğümden aşağı bırakılır. Sağkalım ağaçlarının binary doğası nedeniyle x_i benzersiz bir uç düğüm olan h'ye düşecektir. i. birim için KHF, x_i'nin terminal düğümü için NA tahmin edicisidir. $x_i \epsilon h$ olmak üzere, aşağıdaki şekilde hesaplanır (Ishwaran ve ark., 2008a):

$$H(t|x_i) = \widehat{H}_h(t) \tag{4.5}$$

Denklem (4.5) tüm birimler için ve bir sağkalım ağacı için KHF'yi tanımlar .

4.1.3.3 Bootstrap ve out of bag (OOB) için topluluk KHF

Denklem (4.5) tek bir sağkalım ağacından türetilmiştir. Topluluk KHF'sini hesaplamak için B tane sağkalım ağacının ortalamasını alırız. Hem OOB hem de bootstrap tahminini tanımlayacağız. Ormandaki her bir ağaç bağımsız bir bootstrap örneklemi kullanılarak büyütülmektedir. Eğer b'nci bootstrap örneklemi için i; OOB birimi ise $I_{i,b} = 1$, OOB birimi değil ise $I_{i,b} = 0$ olsun. Denklem (4.5)'den $H_b^*(t|x_i)$ b'nci bootstrap örnekleminden büyüyen bir ağaç için KHF'yi belirtir. Bu durumda OOB için topluluk KHF aşağıdaki şekilde hesaplanır (Ishwaran ve ark., 2008a):

$$H_e^{**}(t|x_i) = \frac{\sum_{b=1}^{B} I_{i,b} H_b^*(t|x_i)}{\sum_{b=1}^{B} I_{i,b}} \tag{4.6}$$

Denklem (4.6) i'nin OOB birimi olduğu bootstrapların ortalaması olduğu görülüyor. Eşdeğer olarak $H_e^{**}(t|x_i)$ aşağıdaki adımlar takip edilerek hesaplanabilir. OOB verileri, in-bag verilerinden elde edilen bir sağkalım ağacına bırakılır. i. biriminin terminal düğümünü ve düğümün KHF'si bulunur. Bu KHF'lerin ortalaması alınır. Bu değer Denklem (4.6)' yi verir. i birimi için topluluk bootstrap KHF;

$$H_e^*(t|x_i) = \frac{1}{B}\sum_{b=1}^{B} H_b^*(t|x_i) \tag{4.7}$$

Denklem (4.7) i'nin OOB olduğu durumları değil tüm sağkalım ağaçlarını kullanır.

4.2 Performans Değerlendirme Ölçütleri

Regresyon analizi kullanılan modellerin tahmin performansının değerlendirilmesinde; hata kareler ortalaması (MSE), hata kareler ortalaması kökü (RMSE), mutlak hata ortalaması (MAE) gibi standartlaşmış performans değerlendirme ölçütleri kullanılır. Bu ölçütler sansürlü gözlemler bulundurması nedeniyle nedeniyle, sağkalım analizinde performans değerlendirme ölçütü olarak kullanıma uygun değillerdir. Dolayısıyla sağkalım analizinde kullanılan modellerin tahmin performansının değerlendirilmesinde daha özel ölçütler kullanmak gerekir (Wang ve ark., 2019).

4.2.1 Brier skoru

Adını mucidinden alan Brier Skor (BS) 1950' de olasılıklı hava tahminlerinin yanlışlığını tahmin etmek için Glenn W. Brier tarafından geliştirildi. Yalnızca olasılıksal sonuçlara sahip tahmin modelleri değerlendirilebilir. Yani; çıktı [0,1] aralığında olmalı ve bir birim için tüm olası sonuçların toplamı 1 olmalıdır. BS, modelin tahminlediği olasılık değeri ile durum değişkeninin gerçek değeri arasındaki farkın kareli ortalamasının beklenen değeridir. N birimli bir örneklem ve her bir X_i $(i = 1,2,3, \dots, N)$ için ikili sonuç tahminini düşündüğümüzde, belirli bir t zamanında tahmin edilen değer $\widehat{y_i}(t)$ ve gerçek değer $y_i(t)$ olsun. Belirli bir t zamanında BS' nin amprik tanımı aşağıdaki şekildedir (Wang ve ark., 2019):

$$BS(t) = \frac{1}{N}\sum_{i=1}^{N}[\widehat{y_i}(t) - y_i(t)]^2 \tag{4.8}$$

Graf ve ark. tarafından 1999'da BS kavramını sansürlü bilgi bulunduran modellerin performansını değerlendirmek için genişletilmiştir. Sansürlenmiş birimleri veri kümesine dahil ederken, amprik BS'ye yapılan bireysel katkılar, sansürlenmiş bilgilere göre yeniden ağırlıklandırılır. Ardından BS aşağıdaki şekilde güncellenebilir:

$$BS(t) = \frac{1}{N}\sum_{i=1}^{N} w_i(t)[\widehat{y_i}(t) - y_i(t)]^2 \tag{4.9}$$

Denklem (4.2) de bulunan $w_i(t)$ Denklem (4.3) de verilen şekilde i. birimin ağırlığını belirtir. Belirli bir veri kümesi $(X_i, y_i, 1-\delta_i),\ i = 1,2,\dots,N$ için elde edilen G sansürleme dağılımının KM tahmin edicisi dahil edilerek tahmin edilir.

$$w_i(t) = \begin{cases} {}^{\delta_i}/_{G(y_i)} & yi \le t \\ {}^{1}/_{G(y_i)} & yi > t \end{cases} \qquad (4.10)$$

Bu ağırlık dağılımı ile t zamanından önce sansürlenen birimlerin ağırlıkları 0 olacaktır. Ancak sansürlü birimler G'nin hesaplanmasında kullanıldıklarından dolayı BS'nin hesaplanmasına dolaylı olarak katkıda bulunur. t zamanında sansürlenmemiş birimlerin ağırlıkları, tahmini sağkalım olasılıklarının BS' sinin hesaplanmasına katkıda bulunmasını sağlamak için 1'den büyüktür (Wang ve ark., 2019).

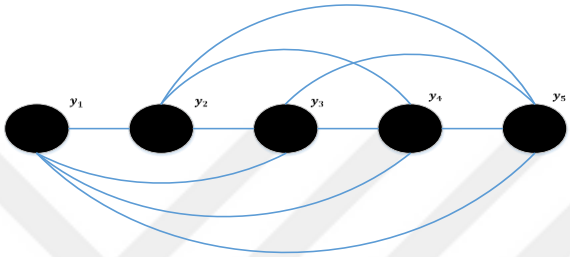
4.2.2 Harrel'in uyum indeksi (Concordance index - C-index)

Sağkalım analizinde ortak bir performans değerlendirme ölçütü Harrel'in uyum indeksi (Concordance İndex-C-index) dir. C-index rasgele seçilen herhangi iki gözlem için uyum olasılığı olarak tanımlanır. Burada uyum, daha kısa sağkalım süresine sahip gözlemin daha büyük risk puanına sahip olması anlamına gelir. Bu durumda sağkalım süresi fazla olan gözlemin ise daha düşük risk puanına sahip olması beklenir. Sansürlü gözlem, sansürlendiği süreden sonra ilgilenilen olayı yaşamış herhangi bir gözlemle karşılaştırılamaz. Çünkü sansürlü gözlemin ilgilenilen olayı, sansürsüz olarak ilgilenilen olayın meydana geldiği gözlemlerden önce veya sonra yaşadığı bilinemez. İki gözlem için sağkalım süresi aşağıdaki iki senaryo için karşılaştırılabilir:

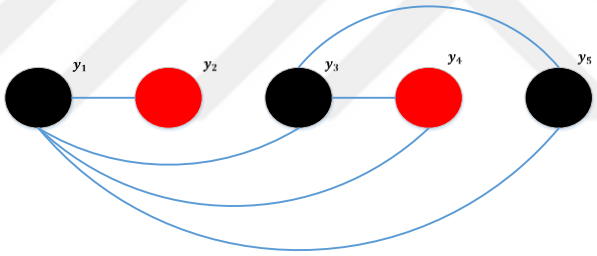
- **i.** Gözlemlerin her ikisi de sansürsüzdür.
- **ii.** Sansürsüz gözlemde ilgilenilen olayın meydana geldiği zaman, sansürlü gözlemin sansürlenme süresinden daha küçüktür. Tüm bu durumlar Şekil 4.1. ve Şekil 4.2 ile görselleştirilebilir. Siyah daireler sansürsüz gözlemleri, kırmızı daireler ise sansürlü gözlemleri temsil etmektedir.

Şekil 4.1 ve Şekil 4.2 sırasıyla sansürlü gözlemler olmadan ve sansürlü gözlemler ile sağkalım verileri için mümkün olan olası sıralama karşılaştırmalarını (gözlemler arası mavi çizgiler ile) gösterir

Şekil 4.1'de sansürlü vakalar olmayan sağkalım verilerinde beş gözlem için $\binom{5}{2} = 10$ olası ikili karşılaştırma vardır. Şekil 4.2'de kırmızı daireler ile temsil edilen sansürlü gözlemlerin varlığı nedeniyle 10 olası karşılaştırmadan yalnızca 6 tanesi karşılaştırılabilir. Sansürlü bir gözlem, yalnızca sansürlenme süresinden önce olay meydana gelmiş bir gözlemle (örneğin; $y_2 \& y_1$) karşılaştırılabilir. Ancak sansürlenen herhangi bir gözlem, sansür süresinden sonra (örneğin; $y_2 \& y_3$ ve $y_2 \& y_4$) hem sansürlü hem de sansürsüz gözlemlerle karşılaştırılamaz (Wang ve ark., 2019).



Şekil 4. 2 C-index için sansürsüz gözlemlerdeki sıralama kısıtlamalarının gösterimi



Şekil 4. 3 C-index için sansürlü gözlemlerdeki sıralama kısıtlamalarının gösterimi

İki birim için hem gözlem hem de tahmin değerleri sırasıyla $(y_1, \hat{y}_1)$ ve $(y_2, \hat{y}_2)$ ikilileri ile gösterilsin. Burada y_i gerçek gözlem süresini $\hat{y}_i$ ise tahmin edilen değeri temsil eder. Birimlerin aralarındaki uyum olasılığı şu şekilde hesaplanabilir (Ishwaran ve ark., 2008a):

$$c = \Pr(\hat{y}_1 > \hat{y}_2 | y_1 \geq y_2) \tag{4.11}$$

C-indeksini hesaplamanın birçok yolu vardır.

(1) $i, j \in \{1,2,\dots,N\}$ için num karşılaştırılabilir tüm çiftler sayısını, $I[.]$ gösterge (indicator) fonksiyon, $\hat{\beta}$ Cox tabanlı modellerde tahmin edilen parametreleri temsil etmek üzere; modelin çıktısı tehlike (hazard) oranı olduğunda, (Cox tabanlı modellerle elde edilen sonuç gibi) C-index aşağıdaki şekilde hesaplanabilir:

$$\hat{c} = \frac{1}{num} \sum_{i:\delta_i=1} \sum_{j:y_i<y_j} I\left[X_i\hat{\beta} > X_j\hat{\beta}\right] \quad (4.12)$$

(2) Sağkalım süresini doğrudan öğrenmeyi amaçlayan sağkalım yöntemleri için C-indeksi aşağıdaki şekilde hesaplanmalıdır:

$$\hat{c} = \frac{1}{num} \sum_{i:\delta_i=1} \sum_{j:y_i<y_j} I\left[S(\hat{y}_j X_j) > S(\hat{y}_i X_i\right] \quad (4.13)$$

5.UYGULAMA

5.1 Perakendecilik ve Perakende Stok Yönetimi

Perakendecilik, tüketicilere kişisel veya ailesel kullanım amacıyla, doğru hizmet veya ürünün, doğru fiyat ve zamanda sunulması yoluyla değer katan bir dizi işletme faaliyetidir. Perakendecilik alanı rekabetin yoğun olarak yaşandığı bir alandır. Yoğun rekabet ortamında işletmeler için başarının en önemli faktörü rekabet üstünlüğünü koruyarak, işletmenin karlılığının, devamlılığının ve gelişiminin sürdürülmesi olduğu bilinmektedir.

Perakende işletmeleri uygun miktarda ürün tedarikini gerçekleştirip tüketici ihtiyacını karşılayarak amaçlarına ulaşabilirler. Nihai tüketiciye satılmak için perakendecinin elinde bulundurduğu ürünlere stok denir. Perakendecilikte stok yönetiminde dinamik dengenin sağlanması önemlidir. Doğru ürün, doğru yerde, doğru zamanda, doğru şekilde ve doğru maliyetle bulunduğunda dinamik denge sağlanmış olacaktır. Stok sayısının fazla olması, ürün genişliği ve karlılığını etkilediği gibi stok sayısının az olması da satış ve müşteri kayıpları üzerinde etkilidir. Özellikle temek ürünlerde stok açığı varsa perakendeci işini iyi yapmıyor demektir. Raflardaki ürünün durumu ve stok kontrolü müşterilerin karar vermelerinde dolaylı olarak etkilidir (MEB, 2011).

Stok yapılması gereken ürün miktarının belirlenmesi perakende ürün yöneticinin temel karar problemlerinden biridir. Stokta bulunmayan bir ürün yalnızca bu ürünün satışını

engellemekle kalmayıp tamamlayıcı ürünlerin satışını da engelleyebilir. Müşteri memnun kalmadığından dolayı mağazadan satın alma işlemini gerçekleştirmeden çıkacağı gibi yaşadığı olumsuz deneyimden dolayı aynı mağazayı tercih etme olasılığı da düşecektir. Stok miktarının fazla olması durumunda ise perakendeci, mağaza alanının verimsiz kullanılmasından dolayı daha az satış yapma gereğinden fazla harcanan zaman ve emek gibi maliyetlere katlanmak durumunda kalacaktır.

Geçmiş dönemlere ait detaylı satış bilgileri (zaman, fiyat, miktar) her bir ürünün tahmini satış miktarının belirlenmesine yardımcı olur. Tahmin edilen gelecek döne satış düzeyine göre stok tedarik edilebilir. Stok yönetimini doğru şekilde geliştiren perakendeci, hangi ürünün ne zaman, hangi miktarda, hangi tüketici tarafından satın alınacağını kolayca tespit edebilir (Varley, 2006).

5.2 Gereç ve Yöntem

Uygulamada Türkiye'de gıda perakende alanında faaliyet gösteren süpermarket zincirlerinden birine ait gıda e-ticaret sitesinin 16.01.2022 ile 28.09.2022 tarihleri arasında kampanya düzenlenen fakat kampanyalı satışı stoklarla sınırlı olan ürünlerin sağkalım olasılıkları sağkalım analiz yöntemleriyle R programlama dili ile R Studio geliştirme ortamında incelenmiştir. Yapılan çalışmada her bir kampanyalı ürün 15 saat boyunca gözlemlenmiştir. 15 saat sonunda kampanya başlangıcında indirimli olarak satılmasına karar verilen ürün stok adeti satışı gerçekleşmezse bu gözlem sansürlü gözlem olarak alınmıştır.

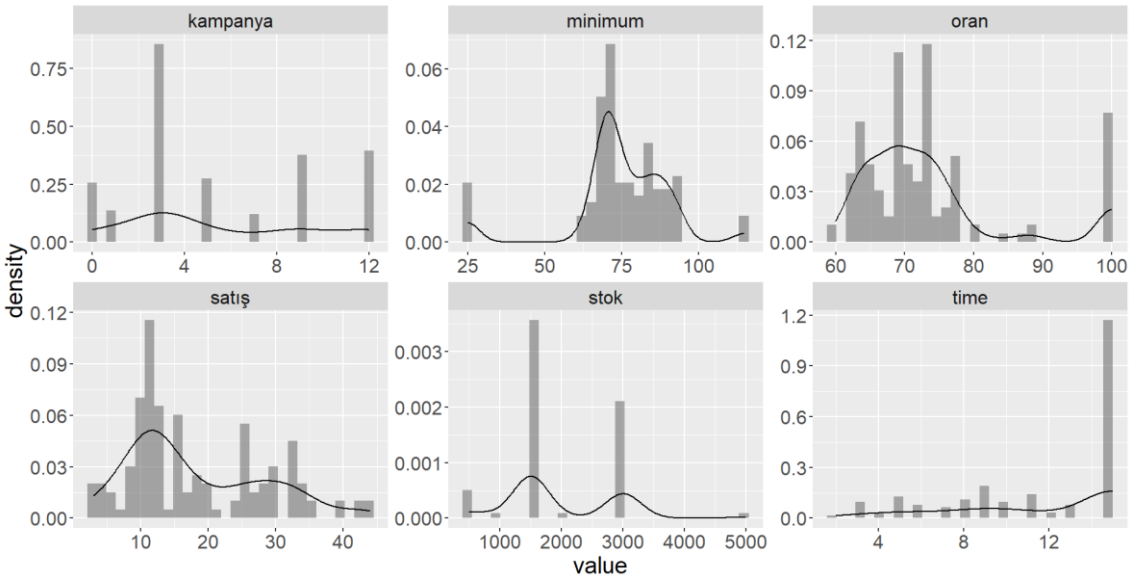
Ürünlerin; kampanyaya girmeden önceki satış fiyatı için satış, kampanyalı fiyatı için kampanya, kampanyayla birlikte sahip olduğu yüzdelik indirim oranı için oran, kampanyalı satılacak ürün miktarı için stok, kampanyadan yararlanabilmek için minimum yapılması gereken alışveriş tutarı minimum, ürünün e-ticaret sitesinde tanımlandığı kategori için kategori, kampanyanın uygulama tipi için tip, kampanyanın düzenlendiği gün için gün, her bir ürünün izlem süresi için time ve sonuç değişkeni için status (stok; tükendi=1, tükenmedi=0) değişkenleri baz alınmıştır.

Uygulamada sağdan sansürlü sağkalım verilerini analiz etmek için KM, Cox Oransal Hazar regresyon ve RSO yöntemleri kullanılarak Cox Oransal regresyon yöntemi ve RSO

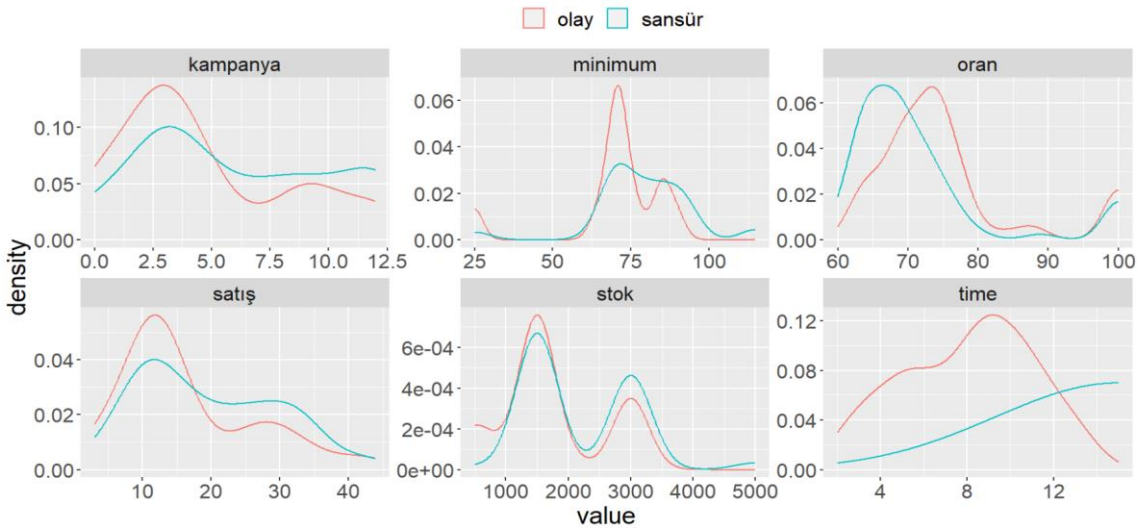
yöntemi sağkalım analizi için ortak değerlendirme ölçütü olan C-index performans değerine göre karşılaştırılmıştır. Cox Oransal Hazard regresyon yöntemi kullanılarak oluşturulan modelde çoklu bağlantı probleminin ortadan kaldırılmasında LASSO değişken seçim yöntemi kullanılmıştır.

5.3 Veri Setine Ait Bulgular

Şekil 5.1'de çalışma kapsamında ele alınan ürünler için tanımlanan sürekli değişkenlerin yoğunluk grafikleri, Şekil 5.2' de ise her bir sürekli değişkenin kendi içinde sansürlenme durumuna göre yoğunluk grafikleri gösterilmektedir.



Şekil 5. 1 Sürekli değişkenlere ait yoğunluk grafikleri



Şekil 5. 2 Sansürlü ve sansürsüz gözlemlerin yoğunluklarının karşılaştırması

Çalışma kapsamında ele alınan kampanyalı ürünlerin kategorik değişkenlerine ait özellikler aşağıdaki gibidir. İlk olarak kampanyalı ürünlerin kategori dağılımı, Tablo 5.1 ile verilmiştir;

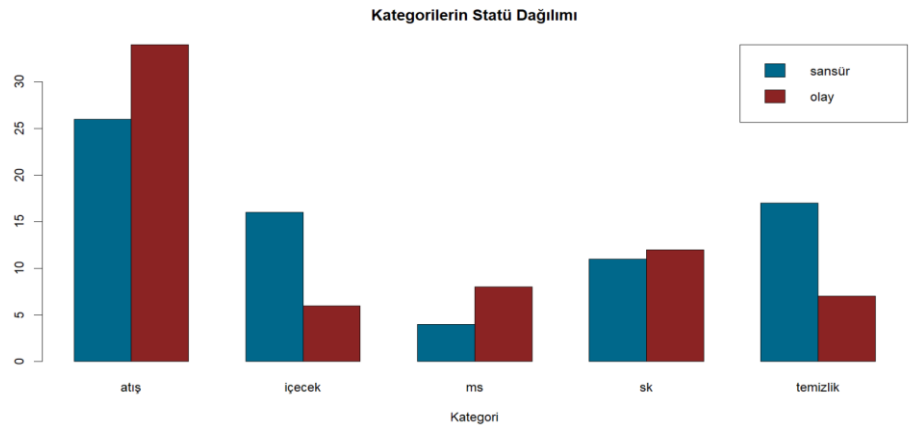
Tablo 5.1'de görüldüğü gibi kampanyalı ürünlerden seçilen örneklemin %42,55'i atıştırmalık, %15,6'sı içecek, %8,51'i meyve-sebze, %16,31'i süt ve kahvaltılık, %17,02'si temizlik kategorisinden oluşmaktadır.

Tablo 5. 1 Kampanyalı ürünlerin kategorilere göre dağılımı

Kategori	N	%
Atıştırmalık (atış)	60	42,55
İçecek	22	15,60
MeyveSebze (ms)	12	8,51
SütveKahvaltılık (sk)	23	16,31
Temizlik	24	17,02
Toplam	141	100

Şekil 5.3'de görüldüğü gibi atıştırmalık kategorisinde bulunan ürünlerin 34, içecek kategorisinde bulunan ürünlerin 6, meyve-sebze kategorisinde bulunan ürünlerin 8, süt ve kahvaltılık kategorisinde bulunan ürünlerin 12, temizlik kategorisinde bulunan ürünlerin ise 7 tanesinde, gözlem süresinde kampanyalı stok miktarında ürün satışı gerçekleşmiş ve stok tükenmiştir.

Atıştırmalık kategorisinde bulunan ürünlerin 26, içecek kategorisinde bulunan ürünlerin 16, meyve-sebze kategorisinde bulunan ürünlerin 4, süt ve kahvaltılık kategorisinde bulunan ürünlerin 11, temizlik kategorisinde bulunan ürünlerin ise 17 tanesinde ise gözlem süresinde kampanyalı stok miktarında ürün satışı gerçekleşmemiş ve bu gözlemlerin sağkalım süreleri sansürlenmiştir.



Şekil 5. 3 Kategoriler bazında statü dağılımı

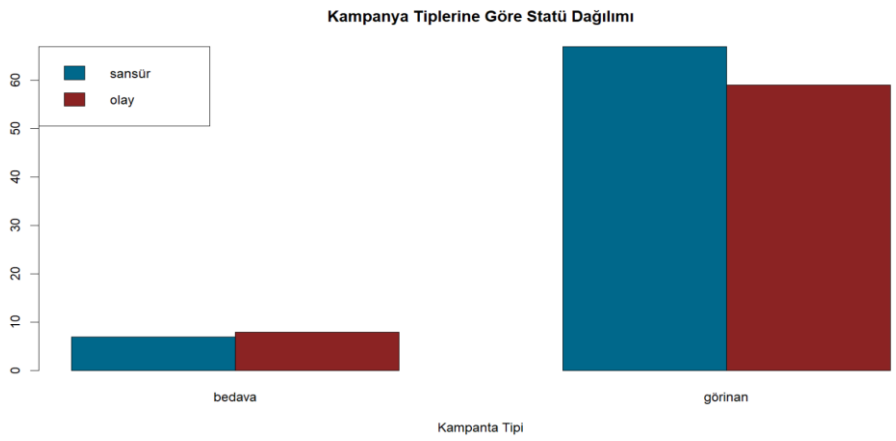
Tablo 5.2'de görüldüğü gibi kampanyalı ürünlerden seçilen örneklemin %89,36'sında görinan, %10,64'ünde bedava tipinde kampanya düzenlenmiştir.

Tablo 5. 2 Kampanyalı ürünlerin kampanya tipine göre dağılımları

Tip	N	%
Görinan	126	89,36
Bedava	15	10,64
Toplam	141	100

Şekil 5.4'de görüldüğü gibi görinan uygulama tipinde kampanyaya giren ürünlerin 59, bedava tipinde kampanyaya giren ürünlerin 8 tanesinde, gözlem süresinde kampanyalı stok miktarında ürün satışı gerçekleşmiş ve stok tükenmiştir.

Görinan uygulama tipinde kampanyaya giren ürünlerin 67, bedava tipinde kampanyaya giren ürünlerin 7 tanesinde gözlem süresinde kampanyalı stok miktarında ürün satışı gerçekleşmemiş ve bu gözlemlerin sağkalım süreleri sansürlenmiştir.



Şekil 5. 4 Kampanya tipi bazında statü dağılımı

Tablo 5.3'de görüldüğü gibi kampanyalı ürünlerden seçilen örneklemin %57,45'inde ürünlere kampanya hafta içi %42,55'ine ise hafta sonu uygulanmıştır.

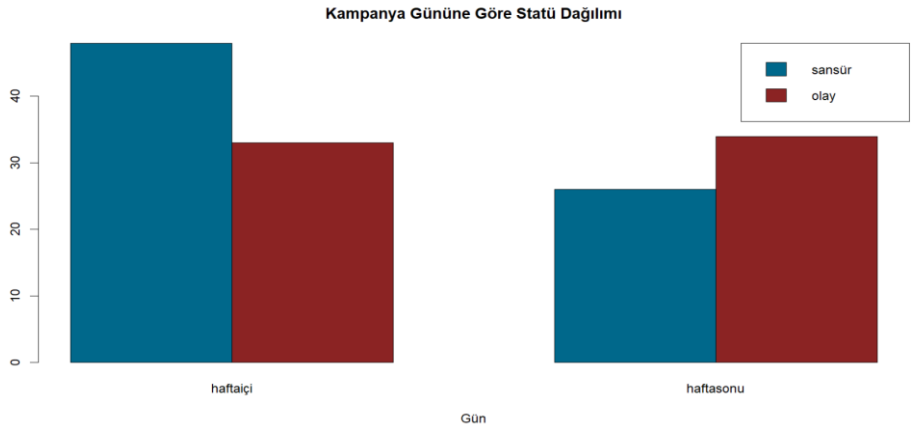
Tablo 5. 3 Kampanyalı ürünlerin kampanya gününe göre dağılımları

Gün	N	%
Haftaiçi	81	57,45
Haftasonu	60	42,55
Toplam	141	100

Şekil 5.5'de görüldüğü gibi hafta içi kampanyaya giren ürünlerin 33, hafta sonu kampanyaya giren ürünlerin 34 tanesinde gözlem süresinde kampanyalı stok miktarında ürün satışı gerçekleşmiş ve stok tükenmiştir.

Hafta içi kampanyaya giren ürünlerin 48, hafta sonu kampanyaya giren ürünlerin 26 tanesinde gözlem süresinde kampanyalı stok miktarında ürün satışı gerçekleşmemiş ve bu gözlemlerin sağkalım süreleri sansürlenmiştir.

Tablo 5.4 incelendiğinde, sansürlü değerler olayın ürünlerin ne kadarında gözlem süresi içerisinde meydana gelmediğini belirtmektedir. Kampanyalı ürünlerin %47,52'sinde olayın meydana geldiği %52,48'inde ise olayın meydana gelmedi görülmektedir. Başka bir ifadeyle kampanya tanımlanan ürünlerin %47,52'sinde gözlem süresinde kampanyaya tanımlanan stok miktarı tükenmiş, %52,48'inde ise stok tükenmemiştir.



Şekil 5. 5 Kampanya günü bazında statü dağılımı

Tablo 5. 4 Kampanyalı ürünlerin stok durumuna göre statü dağılımı

Stok	N	%
Tükendi	67	47,52
Sansür	74	52,48
Toplam	141	100

5.4 Uygulama Sonuçları

Kampanyalı ürünlerin kategori değişkeni bazında, saatlere göre sağkalım olasılıkları Şekil 5.6 ile verilmiştir. Örneğin; atıştırmalık kategorisindeki ürünler için 13. saatte sağkalım olasılığı 0,06 hatayla %43 iken, içecek kategorisindeki ürünler için sağkalım olasılığı 0,1 hatayla %73, meyve ve sebze kategorisindeki ürünler için sağkalım olasılığı 0,1 hatayla %33, süt ve kahvaltılık kategorisindeki ürünler için sağkalım olasılığı 0,1 hatayla %47, temizlik kategorisindeki ürünler için sağkalım olasılığı 0,1 hatayla %71 olarak hesaplanır.

strata(kategori)=atıştırmalık

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
2	60	1	0.983	0.0165	0.951	1.000
3	59	2	0.950	0.0281	0.896	1.000
5	57	6	0.850	0.0461	0.764	0.945
6	51	4	0.783	0.0532	0.686	0.895
7	47	2	0.750	0.0559	0.648	0.868
8	45	3	0.700	0.0592	0.593	0.826
9	42	7	0.583	0.0636	0.471	0.722
10	35	4	0.517	0.0645	0.405	0.660
11	31	1	0.500	0.0645	0.388	0.644
12	30	2	0.467	0.0644	0.356	0.612
13	28	2	0.433	0.0640	0.324	0.579

strata(kategori)=içecek

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
9	22	2	0.909	0.0613	0.797	1.000
10	20	1	0.864	0.0732	0.732	1.000
11	19	2	0.773	0.0893	0.616	0.969
13	17	1	0.727	0.0950	0.563	0.939

strata(kategori)=meyvesebze

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
3	12	1	0.917	0.0798	0.773	1.000
4	11	2	0.750	0.1250	0.541	1.000
5	9	1	0.667	0.1361	0.447	0.995
7	8	1	0.583	0.1423	0.362	0.941
11	7	2	0.417	0.1423	0.213	0.814
13	5	1	0.333	0.1361	0.150	0.742

strata(kategori)=sütvekahvaltılık

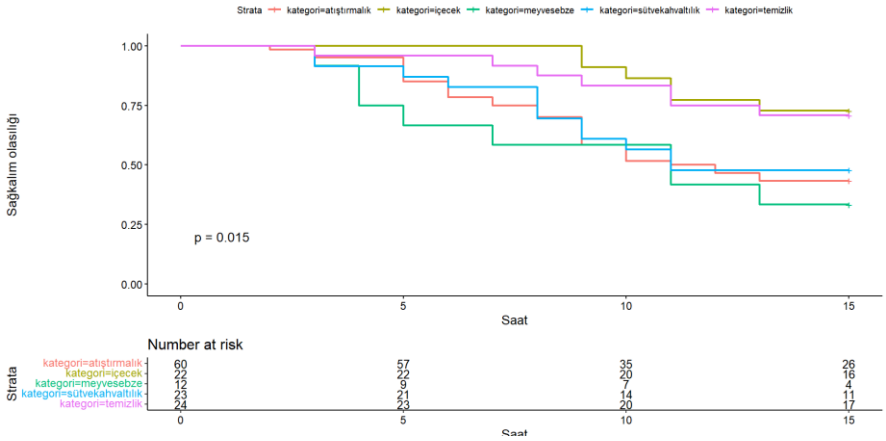
time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
3	23	2	0.913	0.0588	0.805	1.000
5	21	1	0.870	0.0702	0.742	1.000
6	20	1	0.826	0.0790	0.685	0.996
8	19	3	0.696	0.0959	0.531	0.912
9	16	2	0.609	0.1018	0.439	0.845
10	14	1	0.565	0.1034	0.395	0.809
11	13	2	0.478	0.1042	0.312	0.733

strata(kategori)=temizlik

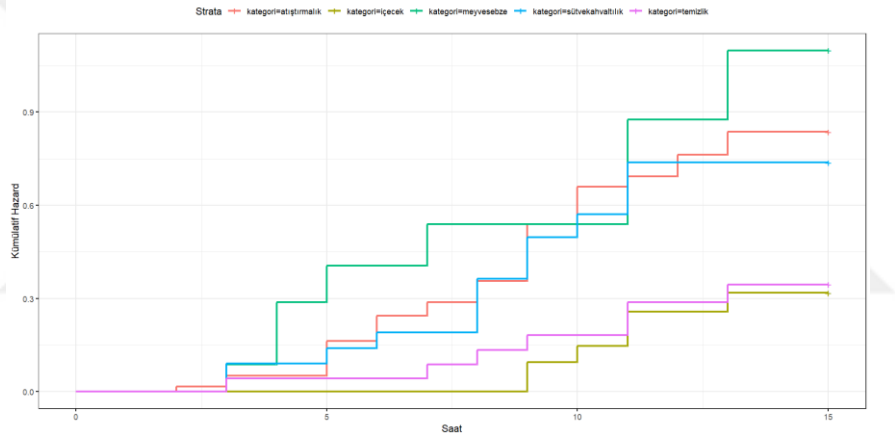
time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
3	24	1	0.958	0.0408	0.882	1.000
7	23	1	0.917	0.0564	0.813	1.000
8	22	1	0.875	0.0675	0.752	1.000
9	21	1	0.833	0.0761	0.697	0.997
11	20	2	0.750	0.0884	0.595	0.945
13	18	1	0.708	0.0928	0.548	0.916

Şekil 5.6 Kampanyalı ürünlerin kategorilerine göre sağkalım olasılıkları

Şekil 5.7'de kategori değişkeninin sağkalım süresi üzerindeki etkisini veren Kaplan-Meier analiziyle elde edilen grafiktir. Kategori değişkenine göre kampanyalı ürünlerin sağkalım sürelerinin eşitliğini kıyaslamak için Log-Rank testi kullanılmıştır. Şekil 5.7'de görüldüğü üzere kategorilerin sağkalım süreleri arasındaki fark Log-Rank test istatistiğine göre anlamlıdır ($p<0,05$).



Şekil 5. 7 Kategori değişkenine göre sağkalım olasılıkları Kaplan-Meier grafiği



Şekil 5. 8 Kategori değişkenine ilişkin hazard grafiği

Kampanyalı ürünlerin kampanya tipi değişkeni bazında saatlere göre sağkalım olasılıkları Şekil 5.9 ile verilmiştir. Şekil 5.9 incelendiğinde örneğin; bedava kampanya kurgusunda indirime giren ürünler için 9. saatte sağkalım olasılığı 0,1 hatayla %53 iken, gör inan kampanya kurgusunda indirime giren ürünler için sağkalım olasılığı 0,04 hatayla %69 olduğu görülmektedir.

Şekil 5.10 tip değişkeninin sağkalım süresi üzerindeki etkisini veren Kaplan-Meier analiziyle elde edilen grafiktir. Kategori değişkenine göre kampanyalı ürünlerin sağkalım sürelerinin eşitliğini kıyaslamak için Log-Rank testi kullanılmıştır. Şekil 5.10'da görüldüğü üzere kampanya tipinin sağkalım süreleri arasındaki fark Log-Rank test istatistiğine göre anlamlı değildir (p>0,05).

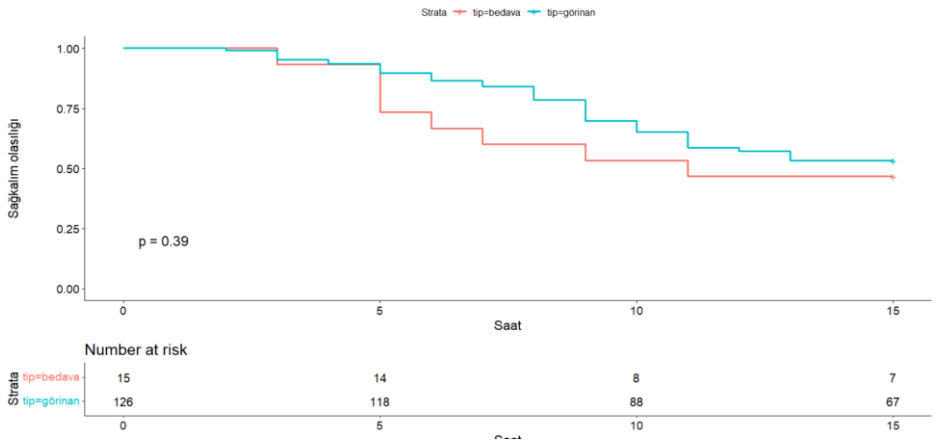
strata(tip)=bedava

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
3	15	1	0.933	0.0644	0.815	1.000
5	14	3	0.733	0.1142	0.540	0.995
6	11	1	0.667	0.1217	0.466	0.953
7	10	1	0.600	0.1265	0.397	0.907
9	9	1	0.533	0.1288	0.332	0.856
11	8	1	0.467	0.1288	0.272	0.802

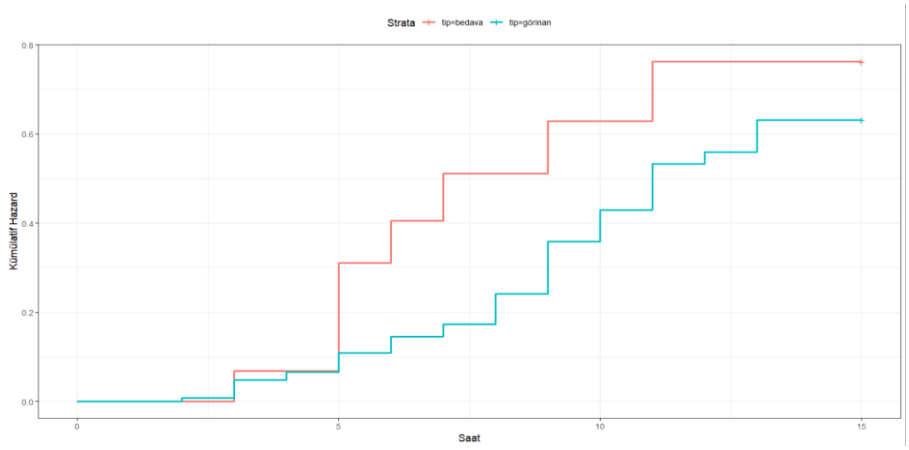
strata(tip)=görinan

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
2	126	1	0.992	0.0079	0.977	1.000
3	125	5	0.952	0.0190	0.916	0.990
4	120	2	0.937	0.0217	0.895	0.980
5	118	5	0.897	0.0271	0.845	0.952
6	113	4	0.865	0.0304	0.807	0.927
7	109	3	0.841	0.0326	0.780	0.908
8	106	7	0.786	0.0366	0.717	0.861
9	99	11	0.698	0.0409	0.623	0.783
10	88	6	0.651	0.0425	0.573	0.740
11	82	8	0.587	0.0439	0.507	0.680
12	74	2	0.571	0.0441	0.491	0.665
13	72	5	0.532	0.0445	0.451	0.626

Şekil 5. 9 Kampanyalı ürünlerin kampanya tipine göre sağkalım olasılıkları



Şekil 5. 10 Tip değişkenine göre sağkalım olasılıkları Kaplan-Meier grafiği



Şekil 5. 11 Tip değişkenine ilişkin hazard grafiği

Kampanyalı ürünlerin kampanya günü değişkeni bazında saatlere göre sağkalım olasılıkları Şekil 5.12 ile verilmiştir. Şekil 5.12 incelendiğinde örneğin; hafta içi kampanyaya giren ürünler için 13. saatte sağkalım olasılığı 0,05 hatayla %59 iken, hafta sonu kampanyaya giren ürünler için sağkalım olasılığı 0,06 hatayla %43 olduğu görülmektedir.

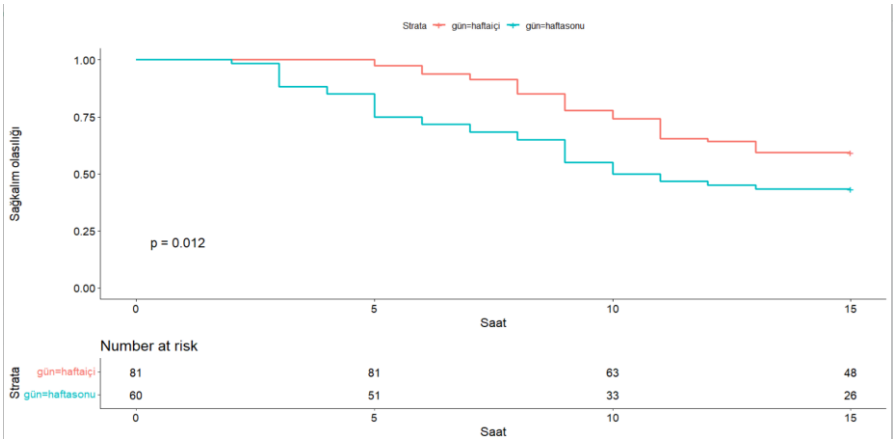
strata(gün)=haftaiçi

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
5	81	2	0.975	0.0172	0.942	1.000
6	79	3	0.938	0.0267	0.887	0.992
7	76	2	0.914	0.0312	0.854	0.977
8	74	5	0.852	0.0395	0.778	0.933
9	69	6	0.778	0.0462	0.692	0.874
10	63	3	0.741	0.0487	0.651	0.843
11	60	7	0.654	0.0528	0.559	0.767
12	53	1	0.642	0.0533	0.546	0.755
13	52	4	0.593	0.0546	0.495	0.710

strata(gün)=haftasonu

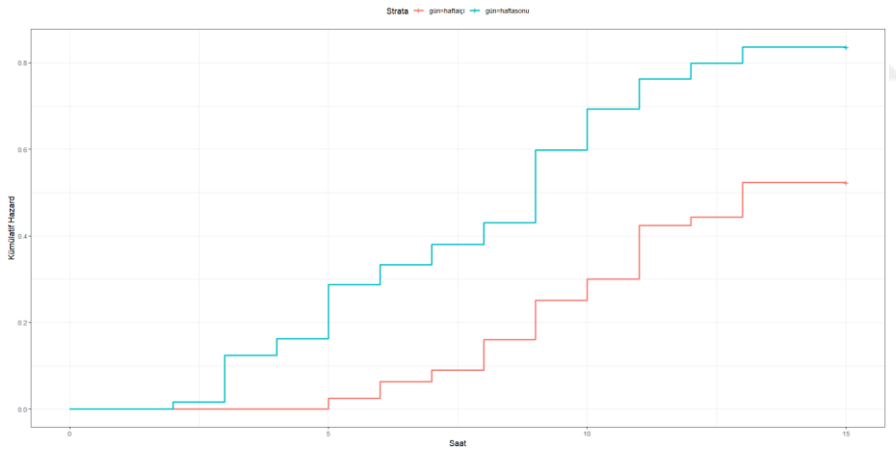
time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
2	60	1	0.983	0.0165	0.951	1.000
3	59	6	0.883	0.0414	0.806	0.968
4	53	2	0.850	0.0461	0.764	0.945
5	51	6	0.750	0.0559	0.648	0.868
6	45	2	0.717	0.0582	0.611	0.840
7	43	2	0.683	0.0601	0.575	0.812
8	41	2	0.650	0.0616	0.540	0.783
9	39	6	0.550	0.0642	0.437	0.691
10	33	3	0.500	0.0645	0.388	0.644
11	30	2	0.467	0.0644	0.356	0.612
12	28	1	0.450	0.0642	0.340	0.595
13	27	1	0.433	0.0640	0.324	0.579

Şekil 5. 12 Kampanyalı ürünlerin kampanya gününe göre sağkalım olasılıkları



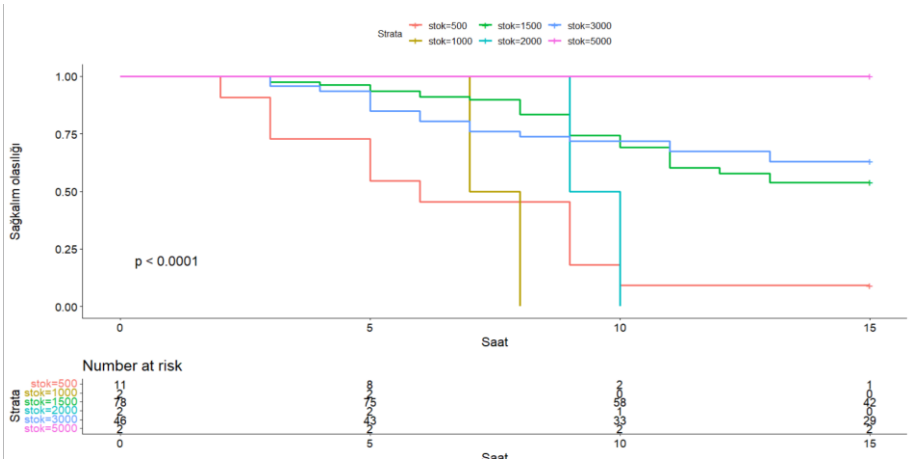
Şekil 5. 13 Gün değişkenine göre sağkalım olasılıkları Kaplan-Meier grafiği

Şekil 5.13 gün değişkeninin sağkalım süresi üzerindeki etkisini veren Kaplan-Meier analiziyle elde edilen grafiktir. Gün değişkenine göre kampanyalı ürünlerin sağkalım sürelerinin eşitliğini kıyaslamak için Log-Rank testi kullanılmıştır. Şekil 5.13'de görüldüğü üzere kampanya gününün sağkalım süreleri arasındaki fark Log-Rank test istatistiğine göre anlamlıdır ($p<0,05$).



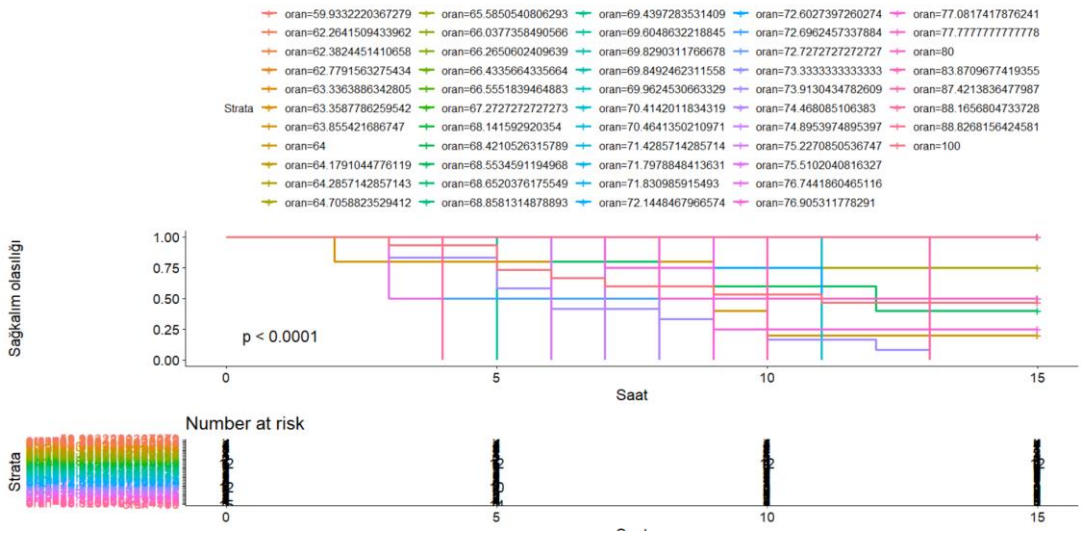
Şekil 5. 14 Gün değişkenine ilişkin hazard grafiği

Şekil 5.15 stok değişkeninin sağkalım süresi üzerindeki etkisini veren Kaplan-Meier analiziyle elde edilen grafiktir. Stok değişkenine göre kampanyalı ürünlerin sağkalım sürelerinin eşitliğini kıyaslamak için Log-Rank testi kullanılmıştır. Şekil 5.15'de görüldüğü üzere kampanya ürün stoğunun sağkalım süreleri arasındaki fark Log-Rank test istatistiğine göre anlamlıdır ($p<0,05$).



Şekil 5. 15 Stok değişkenine göre sağkalım olasılıkları Kaplan-Meier grafiği

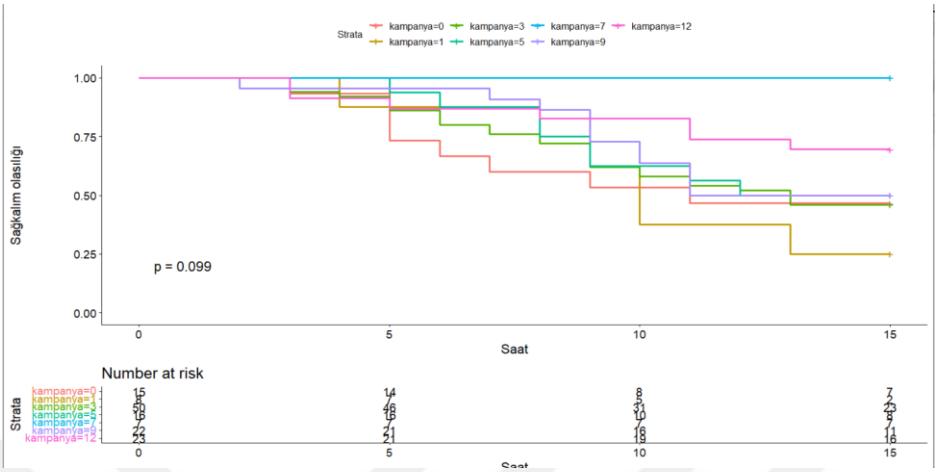
Şekil 5.16 kampanyaya tanımlanan indirim oranı için indirim değişkeninin sağkalım süresi üzerindeki etkisini veren Kaplan-Meier analiziyle elde edilen grafiktir. Oran değişkenine göre kampanyalı ürünlerin sağkalım sürelerinin eşitliğini kıyaslamak için Log-Rank testi kullanılmıştır. Şekil 5.16'da görüldüğü üzere indirim oranına göre sağkalım süreleri arasındaki fark Log-Rank test istatistiğine göre anlamlıdır (p<0,05).



Şekil 5. 16 Oran değişkenine göre sağkalım olasılıkları Kaplan-Meier grafiği

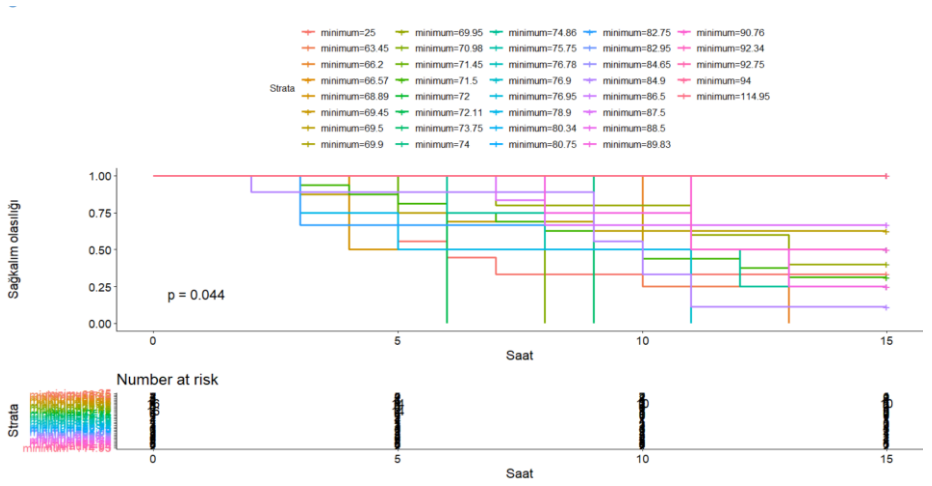
Şekil 5.17 kampanya tanımlanan ürünlerin kampanyalı satış fiyatını temsil eden kampanya değişkeninin sağkalım süresi üzerindeki etkisini veren Kaplan-Meier analiziyle elde edilen grafiktir. Kampanya değişkenine göre kampanyalı ürünlerin sağkalım sürelerinin eşitliğini kıyaslamak için Log-Rank testi kullanılmıştır. Şekil

5.17'da görüldüğü üzere kampanyalı satış fiyatına göre sağkalım süreleri arasındaki fark Log-Rank test istatistiğine göre anlamlı değildir (p>0,05).



Şekil 5. 17 Kampanya değişkeni Kaplan-Meier grafiği

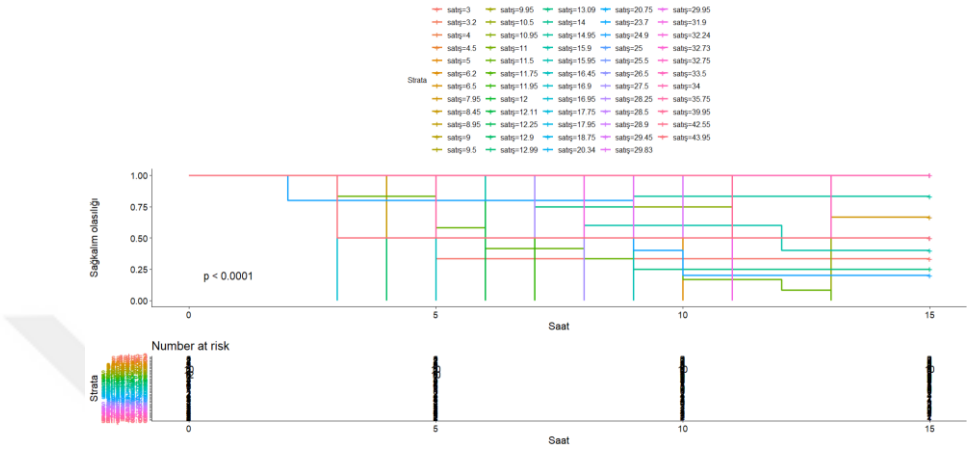
Şekil 5.18, kampanyadan yararlanabilmek için yapılması gereken minimum alışveriş tutarını temsil eden minimum değişkeninin sağkalım süresi üzerindeki etkisini veren Kaplan-Meier analiziyle elde edilen grafiktir. Minimum değişkenine göre kampanyalı ürünlerin sağkalım sürelerinin eşitliğini kıyaslamak için Log-Rank testi kullanılmıştır. Şekil 5.18'da görüldüğü üzere indirim oranına göre sağkalım süreleri arasındaki fark Log-Rank test istatistiğine göre anlamlıdır (p>0,05).



Şekil 5. 18 Minimum değişkeni Kaplan-Meier grafiği

Şekil 5.19 ürünün kampanyayla birlikte satışa sunulduğu satış değişkeninin sağkalım süresi üzerindeki etkisini veren Kaplan-Meier analiziyle elde edilen grafiktir. Satış

değişkenine göre kampanyalı ürünlerin sağkalım sürelerinin eşitliğini kıyaslamak için Log-Rank testi kullanılmıştır. Şekil 5.19'da görüldüğü üzere kampanyalı satış fiyatına göre sağkalım süreleri arasındaki fark Log-Rank test istatistiğine göre anlamlıdır (p>0,05).



Şekil 5. 19 Satış değişkeni Kaplan-Meier grafiği

```
                          coef  exp(coef)   se(coef)      z Pr(>|z|)
satış                0.1936159  1.2136301  0.0734504  2.636  0.00839 **
oran                -0.0119405  0.9881305  0.0443276 -0.269  0.78765
stok                -0.0013343  0.9986666  0.0002668 -5.002 5.69e-07 ***
minimum             -0.0459464  0.9550932  0.0177241 -2.592  0.00953 **
kategoriiçecek      -1.2526295  0.2857524  0.5548322 -2.258  0.02397 *
kategorimeyvesebze   1.8799041  6.5528764  0.5764513  3.261  0.00111 **
kategorisütvekahvaltılık  0.3743241  1.4540083  0.4708175  0.795  0.42658
kategoritemizlik    -0.5757644  0.5622749  0.4330683 -1.330  0.18368
tipgörinan          -0.0368742  0.9637974  1.0149240 -0.036  0.97102
günhaftasonu         0.2388001  1.2697247  0.2671180  0.894  0.37133
kampanya            -0.5639721  0.5689447  0.2377425 -2.372  0.01768 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                    exp(coef) exp(-coef) lower .95 upper .95
satış                  1.2136     0.8240   1.05091    1.4015
oran                   0.9881     1.0120   0.90590    1.0778
stok                   0.9987     1.0013   0.99814    0.9992
minimum                0.9551     1.0470   0.92248    0.9889
kategoriiçecek         0.2858     3.4995   0.09632    0.8477
kategorimeyvesebze     6.5529     0.1526   2.11717   20.2819
kategorisütvekahvaltılık  1.4540     0.6878   0.57784    3.6587
kategoritemizlik       0.5623     1.7785   0.24061    1.3139
tipgörinan             0.9638     1.0376   0.13185    7.0452
günhaftasonu           1.2697     0.7876   0.75221    2.1433
kampanya               0.5689     1.7576   0.35703    0.9066

Concordance= 0.785  (se = 0.028 )
Likelihood ratio test= 68.21  on 11 df,   p=3e-10
Wald test            = 53.56  on 11 df,   p=1e-07
Score (logrank) test = 56.06  on 11 df,   p=5e-08
```

Şekil 5. 20 Tüm değişkenlerle kurulan Cox regresyon analizi çıktıları

Şekil 5.20'de analiz sonuçları gösterilen Cox regresyon analizinde kullanılan bağımsız değişkenler arasındaki çoklu bağlantı(multicollinearity) probleminin kontrolü için her bir değişken için Tablo 5.5'de Varyans Enflasyon Faktörü (Variance Inflation Factors-VIF)

hesaplanmıştır. VIF' ler iki ve daha fazla çoklu bağlantının varlığını varlığını göstermede yararlı olabilecek en iyi ölçülerden biridir.

VIF değerlerinin 5 ya da 10'un üzerinde olması güçlü çoklu bağlantının bir göstergesidir ve ilgili değişkenlere ilişkin regresyon katsayılarına güvenilmemesi gerektiğini bildirir. Şekil 5.5' de görüldüğü üzere minimum, kategori ve tip değişkenlerine ait VIF'ler 5'den satış ve oran değişkenine ait VIF'ler ise 10'dan büyüktür. Çoklu bağlantı probleminin üstesinden gelebilmek için veri setine LASSO yönetimi uygulanarak değişken seçimi yapılmıştır. LASSO yöntemine göre satış, tip ve kampanya değişkenleri ilgisiz değişkenlerdir ve bu değişkenlerin katsayıları sıfıra eşit olmalıdır.

Tablo 5. 5 Tüm bağımsız değişkenler için VIF'ler

değişken	VIF
satış	39,289
oran	14,825
stok	2,725
minimum	5,102
kategori	7,094
tip	7,236
gün	1,186

LASSO değişken seçim sonuçlarına göre satış, tip, kampanya değişken katsayıları sıfır olarak alınarak kurulacak modelde çoklu bağlantı problemi ortadan kalkacaktır. Oran, stok, minimum, kategori, gün değişkenleriyle yeniden model kurulmuştur. Modele ait sonuçlar Şekil 5.21'de gösterilmiştir.

Tablo 5. 6 LASSO değişken seçimi

değişken	s0
satış	.
oran	0,0382
stok	-0,009
kategoriiçecek	-0,017
kategori meyvesebze	-0,739
kategorisütvekahvaltılık	1,272
kategoritemizlik	0,322
tipgörinan	.
günhaftasonu	-0,220
kampanya	.

Oran, stok, minimum, kategori değişkenleriyle kurulan modelde bağımsız değişkenler arasındaki çoklu bağlantı probleminin kontrolü için VIF'ler Tablo 5.7'de gösterilmiştir.

Bağımsız değişlerin VIF'leri çoklu bağlantı probleminin ortadan kalktığını göstermektedir. LASSO değişken seçimi ile birlikte

```
  n= 141, number of events= 67

                             coef  exp(coef)   se(coef)      z Pr(>|z|)
oran                     0.0532884  1.0547337  0.0195293  2.729  0.00636 **
stok                    -0.0014282  0.9985728  0.0002682 -5.324 1.01e-07 ***
minimum                 -0.0270757  0.9732876  0.0142332 -1.902  0.05713 .
kategoriiçecek          -0.9963684  0.3692179  0.4697582 -2.121  0.03392 *
kategorimeyvesebze       2.1193749  8.3259310  0.5428266  3.904 9.45e-05 ***
kategorisütvekahvaltılık 0.7369647  2.0895834  0.3805768  1.936  0.05281 .
kategoritemizlik        -0.3413990  0.7107752  0.4243086 -0.805  0.42105
günhaftasonu             0.2843354  1.3288785  0.2540930  1.119  0.26313
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                         exp(coef) exp(-coef) lower .95 upper .95
oran                        1.0547     0.9481    1.0151    1.0959
stok                        0.9986     1.0014    0.9980    0.9991
minimum                     0.9733     1.0274    0.9465    1.0008
kategoriiçecek              0.3692     2.7084    0.1470    0.9271
kategorimeyvesebze          8.3259     0.1201    2.8733   24.1261
kategorisütvekahvaltılık    2.0896     0.4786    0.9911    4.4056
kategoritemizlik            0.7108     1.4069    0.3094    1.6327
günhaftasonu                1.3289     0.7525    0.8076    2.1866

Concordance= 0.774  (se = 0.03 )
Likelihood ratio test= 61.12  on 8 df,   p=3e-10
Wald test            = 46.15  on 8 df,   p=2e-07
Score (logrank) test = 47.07  on 8 df,   p=1e-07
```

Şekil 5. 21 LASSO-Cox model sonuçları

Tablo 5. 7 LASSO-Cox yöntemi bağımsız değişkenlere ait VIF'ler

değişken	VIF
oran	3,331
stok	2,878
minimum	2,929
kategori	2,443
gün	1,074

Oran, stok, minimum, kategori, gün değişkenleriyle oluşturulan Cox regresyon analizi sonuçlarına göre gün değişkeninin stok tükenme süresi üzerinde anlamlı bir etkisi bulunmamaktadır. Gün değişkenini çıkartarak oran stok minimum ve kategori değişkenleriyle kurulan Cox regresyon analiz sonuçları Şekil 5.22'deki şekildedir.

```
  n= 141, number of events= 67

                             coef  exp(coef)   se(coef)      z Pr(>|z|)
oran                     0.0563691  1.0579881  0.0192556  2.927  0.00342 **
stok                    -0.0014548  0.9985462  0.0002663 -5.462 4.70e-08 ***
minimum                 -0.0265618  0.9737879  0.0139249 -1.907  0.05646 .
kategoriiçecek          -0.9752535  0.3770967  0.4689940 -2.079  0.03758 *
kategorimeyvesebze       2.2173359  9.1828344  0.5367892  4.131 3.62e-05 ***
kategorisütvekahvaltılık 0.7240170  2.0627025  0.3799297  1.906  0.05669 .
kategoritemizlik        -0.3810709  0.6831294  0.4226320 -0.902  0.36724
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                        exp(coef) exp(-coef) lower .95 upper .95
oran                       1.0580     0.9452    1.0188    1.0987
stok                       0.9985     1.0015    0.9980    0.9991
minimum                    0.9738     1.0269    0.9476    1.0007
kategoriiçecek             0.3771     2.6518    0.1504    0.9455
kategorimeyvesebze         9.1828     0.1089    3.2067   26.2962
kategorisütvekahvaltılık   2.0627     0.4848    0.9796    4.3434
kategoritemizlik           0.6831     1.4639    0.2984    1.5640

Concordance= 0.762  (se = 0.031 )
Likelihood ratio test= 59.88  on 7 df,   p=2e-10
Wald test            = 45.42  on 7 df,   p=1e-07
Score (logrank) test = 45.34  on 7 df,   p=1e-07
```

Şekil 5. 22 Oran, stok, minimum ve kategori değişkenleri Cox regresyon modeli

Tablo 5.8’de oran, stok, minimum değişkenlerine ait VIF’ler verilmiştir. Tablo 5.8 incelendiğinde tüm değişkenler için VIF değerleri 5’den küçük görülmektedir. Şekil 5.22’de kurulan Cox regresyon modelinde çoklu bağlantı problemi yoktur ve kurulan model Likelihood ratio, Wald ve Score testlerine göre anlamlıdır ($p<0,05$) Sağkalım analiz yöntemleri ortak değerlendirme ölçütü olan C-index değeri ise 0,76 olarak bulunmuştur.

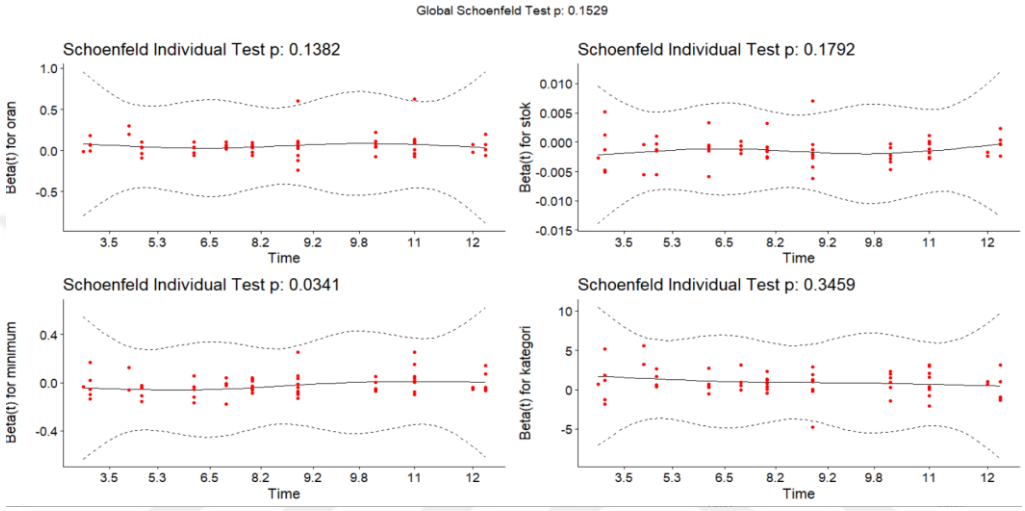
Tablo 5. 8 Oran, stok, minimum ve kategori bağımsız değişkenlerine ait VIF’ler

değişken	VIF
oran	3,158
stok	2,777
minimum	3,697
kategori	2,336

Cox oransal hazard modelinin sağlaması gereken en önemli varsayım oransallığın sağlanmasıdır. Tablo 5.9 incelendiğinde GLOBAL değer 0,153 olarak görülmektedir. Bulunan bu değere göre oransallığın sağlanmadığı varsayımı reddedilir.($p>0,05$)

Tablo 5. 9 Cox regresyon modeli oransallık varsayımı için değerler

değişken	chisq	df	p
oran	2,20	1	0,138
stok	1,80	1	0,179
minimum	4,49	1	0,034
kategori	4,47	4	0,346
GLOBAL	10,69	7	0,153



Şekil 5. 23 Schoenfeld artıklarına ait grafikler

Rasgele sağkalım ormanları-RSO yöntemi için ranger paketi kullanılmıştır. Ranger paketi sağkalım ormanlarını desteklemektedir. Sağkalım ormanları Ishwaran ve ark. tarafından geliştirilen rasgele sağkalım ormanları yöntemindeki gibi uygulanır.

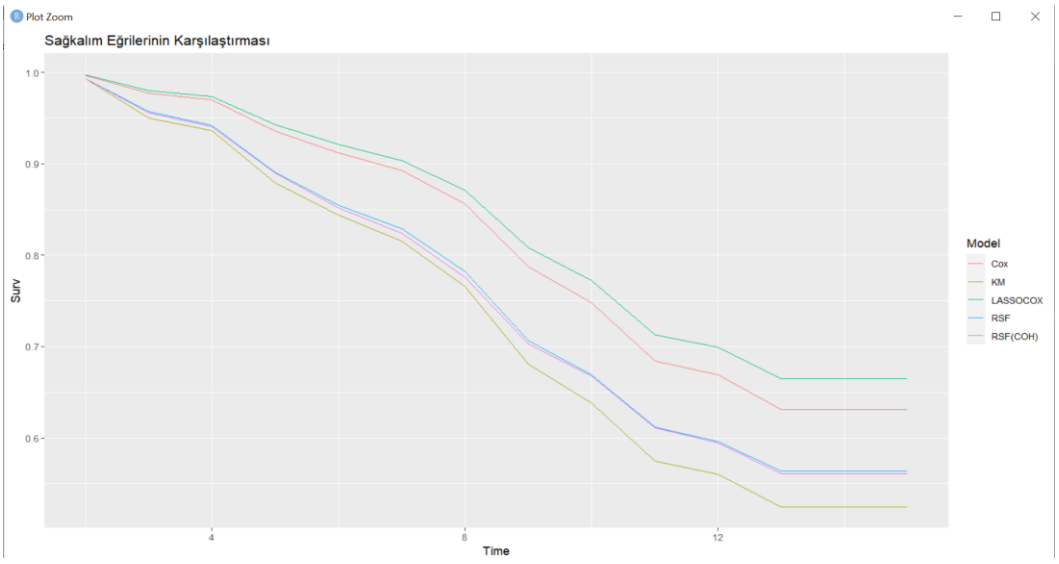
RSO yöntemi kullanılarak yapılan ilk modelde tüm bağımsız değişkenler kullanılmıştır. Log-rank ayırma kriteri kullanılarak yapılan bu modelde değişkenlerin önem değerleri Tablo 5.10' da verilmiştir. Tablo 5.10 incelendiğinde RSO yöntemine göre önemli olan ilk üç değişken oran, minimum ve stok değişkenleri olarak sıralanabilir. Şekil 5.22'de sonuçları verilen nihai Cox oransal regresyon modeline göre oran, stok, minimum değişkenleri stok tükenmesi üzerinde anlamlı değişkenler olarak bulunmuştu. RSO yöntemi de Cox oransal regresyon yöntemini destekler niteliktedir.

Tablo 5. 10 RSO yöntemi değişken önem değerleri

değişken	önem
oran	0,0917
minimum	0,0636
stok	0,0576
satış	0,0337
kategori	0,0198
gün	0,0134

Ranger paketi sağkalım analizinde ortak bir performans değerlendirme ölçütü olan C-index kullanır. Tüm değişkenlerin modele dahil edilerek kurulduğu ilk modelde log-rank ayırma kriteri kullanılmıştır. Bu model için modelin test verisindeki hata oranı %24 olup, C-index değeri 0,76 olarak bulunmuştur.

Oransal hazard varsayımın sağlandığı değişkenlerle kurulan Şekil' 5.22'de sonuçları verilen Cox oransal hazard regresyon yöntemindeki değişkenler oluşturulan RSO yönteminde ayırma kriteri olarak log-rank ayırma kriteri kullanılmıştır. Bu model için C-index değeri ise 0,75 olarak bulunmuştur. Aynı değişkenlerle kurulan Cox oransal hazard regresyon analizi için bu değer 0,76 olarak görülmektedir. Şekil 5.24'de yapılan uygulama çalışmasında kullanılan tüm modellere ait ortalama sağkalım eğrileri gösterilmektedir.



Şekil 5. 24 Yöntemlere göre sağkalım eğrilerinin karşılaştırılması

7.TARTIŞMA VE SONUÇ

Yoğun rekabet ortamı bulunan perakende alanında işletmeler için başarının en önemli faktörü rekabet üstünlüğünü koruyarak, işletmenin karlılığının, devamlılığının ve gelişiminin sürdürülmesi olduğu bilinmektedir. Perakende işletmeleri uygun miktarda ürün tedarikini gerçekleştirip tüketici ihtiyacını karşılayarak amaçlarına ulaşabilirler. Nihai tüketiciye satılmak için perakendecinin elinde bulundurduğu ürünlere stok denir. Perakendecilikte stok yönetiminde dinamik dengenin sağlanması önemlidir.

Bu noktadan hareketle Türkiye' de gıda perakende alanında faaliyet gösteren süpermarket zincirlerinden birine ait gıda e-ticaret sitesinin 16.01.2022 ile 28.09.2022 tarihleri arasında kampanya düzenlenen fakat kampanyalı satışı stoklarla sınırlı olan ürünlerin sağkalım olasılıkları ve sağkalım olasılıklarına etki eden değişkenler sağkalım analiz yöntemleriyle incelenmiştir. Kampanya tanımlanan her bir ürün için takip süresi olarak 15 saat alınmıştır. Takip süresi sonlandığında her bir ürün için çalışma başlangıcında belirlenen stok adeti satışı yapılamamışsa bu gözlem sansürlü gözlem olarak kabul edilmiştir. Bu kapsamda Kaplan Meier, Cox Oransal Hazard Regresyon yöntemi, değişken seçimi için LASSO ve topluluk makine öğrenmesi yöntemi olan sağkalım ağaç yöntemlerinden Rasgele Sağkalım Ağaçları kullanılmıştır. İlgilenilen olay indirimli kampanya ürünlerinin stok tükenme zamanı olduğundan değişkenlerin sağ kalma üzerindeki etkileri araştırılmıştır.

Parametrik olmayan sağkalım analiz yöntemi olan Kaplan-Meier kullanılarak her bir değişken için sağkalım eğrileri çizdirilmiştir. Kategori, gün, stok, oran, kampanya, minimum, satış değişkenleri için Kaplan-Meier eğrilerinin Log-rank test istatistiğine göre anlamlı olarak farklılık gösterdiği ortaya konulmuştur. Tüm değişkenler kullanılarak yapılan Cox regresyon analizinde Kaplan-Meier analizini destekler nitelikte satış, stok, minimum, kampanya değişkenleri anlamlı bulunmuştur. Ancak değişkenlere ait VIF'ler hesaplandığında değişkenlerin birbirleriyle ilişkili olduğu gösterilmiştir. Modelde meydana gelen çoklu bağlantı problemi LASSO değişken seçim yöntemiyle ortadan kaldırılmıştır. Analizler sonucunda; modelde kullanılan değişkenler arasında çoklu bağlantı problemi bulunmayan ve oransal hazard varsayımının sağlandığı Cox Oransal Hazard regresyon modeline göre; oran, stok, minimum, kategori süt ve kahvaltılık,

kategori meyve ve sebze, kategori atıştırmalık değişkenleri anlamlı bulunmuştur ($p<0,05$). Cox Oransal Hazard regresyon modelinden elde edilen β katsayıları ise kampanyayla birlikte ürüne uygulanan indirim oranını temsil eden oran değişkeni için indirim oranındaki artış, kampanyadan yararlanmak için yapılması gereken minimum alışveriş tutarındaki azalış ve kampanyaya tanımlanan stok miktarındaki azalışın sonuç değişkeni olan stok tükenme süresini hızlandıracağını göstermektedir. Bu model için ortak sağkalım değerlendirme ölçütü olan C-index değeri 0,76 olarak bulunmuştur. Parametrik olmayan sağkalım analiz yöntemi olan Kaplan-Meier kullanılarak her bir değişken için sağkalım eğrileri çizdirilmiştir.

Kaplan-Meier ve Cox Oransal Hazard regresyon analizi sağkalım verilerini analiz etmek için kullanılan gelenekselleşmiş istatistiksel yöntemlerdir. Ancak Cox regresyonunun varsayımlarını kabul etmek gerçek dünya uygulamaları için çoğu zaman mümkün değildir. Bu nedenle, sansürlü sağkalım verilerini doğrusal olmayan log risk fonksiyonlarına uydurmak için zengin bir sağkalım modelleri ailesine ihtiyaç duyulmuştur. Makine öğrenmesi yöntemleri, son dönemde yapılan araştırmalarla doğrusal olmayan ilişkileri modelleme ve tahminleme performansı açısından önemli başarılar elde ederek birçok farklı alanda popüler hale gelmişlerdir. Son yıllarda sansürlenmiş gözlemlerle başa çıkarak bu gözlemlere ilişkin tahminleme yapmak için bir dizi gelişmiş makine öğrenmesi yöntemi önerilmiştir.. Bu nedenle sağkalım analizinde kullanılan geleneksel istatistiksel yöntemleri makine öğrenmesi yöntemleriyle değiştirme fikri caziptir.

Breiman 2001'de topluluk öğrenme sürecine rastgelelik enjekte ederek RO adı verilen yöntemi tanıtmıştır (Brieman, 2001). RO yöntemi kullanılarak yapılan ilk çalışmalarda, regresyon ve sınıflandırma problemlerine odaklanılmıştır. RSO yöntemi, Ishwaran ve ark. tarafından 2008'de sağ sansürlü sağkalım verilerinin analizinde kullanılmak üzere RO yönteminin genişletilmesiyle tanıtılmıştır (Wang ve ark., 2019). Yapılan bu tez çalışmasında RSO yöntemi kullanılarak yapılan ilk modelde tüm bağımsız değişkenler kullanılmıştır. Log-rank ayırma kriteri kullanılarak yapılan bu modelde oran minimum ve stok değerlerinin en önemli değişkenler olduğu görülmektedir. Cox oransal regresyon modeline göre oran, stok, minimum değişkenleri stok tükenmesi üzerinde anlamlı değişkenler olarak bulunmuştu. RSO yöntemi de Cox oransal regresyon yöntemini

destekler niteliktedir. Oransal hazard varsayımın sağlandığı Cox oransal hazard regresyon yöntemindeki değişkenler kullanılarak oluşturulan RSO yönteminde ayırma kriteri olarak log-rank ayırma kriteri kullanılmıştır. Bu model için C-index değeri ise 0,75 olarak bulunmuştur. Aynı değişkenlerle kurulan Cox oransal hazard regresyon analizi için bu değer 0,76 olarak görülmektedir. Son olarak Cox Oransal Hazar regresyon, LASSOCOX, tüm değişkenler kullanılarak oluşturulan RSO, oransal hazard varsayımının sağlandığı değişkenlerle oluşturulan RSO ve KM yöntemi için sağkalım eğrileri oluşturularak karşılaştırılmıştır.

8.KAYNAKÇA

Aalen, O. (1978). Nonparametric inference for a family of counting processes. *The Annals of Statistics, 6*(4), 701-726.

Anderson, P. K., Borgan, O., Gill, R. D., & Keiding, N. (2012). *Statistical models based on counting process.* Springer Science & Business Media.

Arı, A., & Önder, H. (2013). Farklı Veri Yapılarında Kullanılabilecek Regresyon Yöntemleri. *DergiPark, 28*(3), 168-174.

Baesens , B., Van, G. T., Stepanova, M., Van den Poel, D., & Vanthienen, J. (2005). Neural network survival analysis for personal Loan data. *Journal od teh Operational Research Society, 56*(9), 1089-1098.

Bardakcı, S., & Kartal, M. (2018). *Sağkalım analizi.* Akademisyen.

Biganzoli, E., Boracchi, P., Mariani, L., & Marubini, E. (1998). Feed forward neural networks for the analysis of censored survival data: A partial logistic regression approach. *Statistics in medicine, 17*(10), 1169-1186.

Bradburn, M. J., Clark, T. G., Love, S. B., & Altman, D. G. (2003). Survival analysis part II: multivariate data analysis–an introduction to concepts and methods. *British Journal of Cancer, 89*(3), 431-436.

Brieman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32.

Collet, D. (1993). *Modeling Survival Data in Medical Research.* Chapman&Hall.

Cox, R. D. (1972). Regression models and life tables. *Journal of the Royal Statistical Society, 2*(34), 187-220.

Cutler, S. J., & Ederer, F. (1958). Maximum utilization of the life table method in analyzing survival. *Journal of Choronic Disease, 8*(6), 699-712.

Dey, K., & Kundu, D. (2009). Discriminating Between the Log-Normal and. *Communications in Statistics-Theory and Methods, 39*(2), 280-292.

Fadnavis, R. A. (2019). Application of Machine Learning For Survival Analysis- A Review. *IOSR Journal of Engineering, 9*(5), 56-60.

Faraggi, D., & Simon, R. (1995). A Neural Network Model for Survival Data. *Statistics in Medicine, 14*(1), 73-82.

Friedman, N., Geiger, D., & Moises, G. (1997). Bayesian network classifiers. *Machine Learning, 29*(2), 131-163.

Gordon, L., & Olshen, R. (1985). *Tree-structured survival analysis.* Cancer Treatment.

Hoerl, A. E., & Kennard, W. R. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55-57.

Hosmer, D. W., Lemeshow, S., & May, S. (2002). *Applied survival analysis: regression modeling of time to data.* NY: Wiley New York.

Hothorn, T., & Lausen, B. (2003). On the exact distribution of maxximally selected rank ststistics. *Computational Statistics & Data Analaysis, 43*, 121-137.

Hui, Z., & Trevor, H. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*(2), 301-320.

Ishwaran, H., & Kogalur, U. B. (2007). Random Survival Forests for R. *R News ISSN 1609-3631, 7*(1), 25-31.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008a). Random survival forests. *The Annals of Applied Statistics, 2*(3), 841-860.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008b). *RandomForestSRC:Random Forests for Survival, Regression ans Classification.* R package version 2.4.1: https://cran.r-project.org/ adresinden alındı

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 457-481.

Khan, F. M., & Zubek, V. B. (2008). Support vector regression for censored data (SVRc): A novel tool for survival analaysis. *In Proceedings of the IEEE International Conference on Data Mining (ICDM)* (s. 863-868). Pisa: IEEE.

Kleinbaum, D. K., & Klein, M. (2010). *Survival analysis* (Third b.). Springer.

Lee, E. T. (1992). *Statistical Methods for Survival Data Analysis* (3 b.). New Jersey: John Wiley&Sons.

Lee, E. T., & Wang, J. (2003). *Statistical Methods for Survival Data Analysis* (Cilt 476). John Wiley & Sons.

Mariani, L., Coradini, D., Biganzoli, E., Boracchi, P., Marubini, E., Pilotti, S., . . . Zucali, R. (1997). *Breast cancer research and treatment, 44*(2), 167-178.

MEB. (2011). *Pazarlama ve Perakende –Perakendecilik.* 08 14, 2016 tarihinde Eğitim Modülü: http://hbogm.meb.gov.tr adresinden alındı

Moore, F. D. (2016). *Applied survival analysis using R.* Springer.

Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics, 14*(4), 945-966.

Nelson, W. (1982). Applied Life Data Analysis. Kanada: John Wiley&Sons.

Özdemir, O. (2015). Sağkalım Analizi Yöntemleri-1. *İyi Klinik Uygulamalar Dergisi*(3), 21-33.

Ravdin, P. M., & Clark, G. M. (1992). A practical application of neural network analysis for predicting outcome of individual breast cancer patients. *Breast cancer research and treatment, 22*(3), 285-293.

Sırın, E. (2017). *Bayes Teoremi Giriş*. https://www.veribilimiokulu.com/bayes-teoremi-giris/. adresinden alındı

Sümbüloğlu, K., & Akdağ, B. (2009). *İleri Biyoistatiksel Yöntemler* (1 b.). Ankara: Alp Ofset Matbaacılık.

Şenocak, M. (1992). *Özel Biyoistatistik Epidemiyolojide Sayısal Çözümleme.* İstanbul: Çağlayan Kitapevi.

Tibshirani, R. (1996). Regression Shirinkage and Selection via the Lasso. *Journal of the Royal Statistical Society, 58*(1), 267-288.

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in medicine, 16*(4), 385-395.

Van, B., Pelckmans, K., Suykens, J. A., Van, H. S., & Huffel, S. V. (2007). Support vector machines for survival analysis. *In Proceedings of the 3rd International Conference on Computational Intelligence in Medicine and Healtcare (CIMED'07)*, (s. 1-8). Plymouth.

Varley, R. (2006). *Retail Product Management Buying and Mershandising* (2 b.). Routledge.

Wang, P., Li, Y., & Chandan, R. K. (2019). Machine Learning for Survival Analysis: A Survey. *ACM Computing Surveys, 51*(6), 1-36.

Wang, S. (2003). Artifical Neural Network. S. Wang içinde, *Interdisciplinary Computing In Java Programming Language* (s. 81-100). Springer.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*(2), 301-320.

ÖZGEÇMİŞ

Bilgiler

Adı, soyadı : Tuğba DEMİRCİOĞLU

Yabancı dil : İngilizce

Eğitim

Derece	**Eğitim Birimi**
Yüksek Lisans	Marmara Üniversitesi/İstatistik
Lisans	İstanbul Üniversitesi/Matematik