Müşteri Ayrılma Analizi ve Sınıflandırma Karşılaştırması

Merve POSLU
N20137450

Anahtar Kelimeler
Müşteri ayrılma analizi, Veri madenciliği, SVM, Telekomünikasyon

Özet: Veri madenciliği, büyük veri kümeleri içindeki anlamlı bilgiyi ortaya çıkarma sürecidir. Veri madenciliğinin yaygın olarak kullanıldığı uygulama alanlarından biri, ayrılma eğilimi gösteren müşterilerin tahmin edilmesidir. Churn adı verilen bu analiz, şirketlerin kaybetme potansiyeli olan müşterilerine özel pazarlama kampanyalarını geliştirmelerini sağlamaya yöneliktir.Bu çalışmada telekom sektörüne ait müşterilerden ayrılma eğilimi gösteren müşteriler analiz edilerek; ayrılma eğilimi gösteren müşteriler tahmin edilmiştir. Bu müşterilere özel pazarlama stratejileri geliştirilmesini hedeflemektedir. Ayrılacak müşteri profilini belirlemek için SVM algoritması çalışılarak diğer algoritmalarla doğruluk tahmini açısından karşılaştırma yapılmıştır.

Churn Analysis and Classification Comparison

Keywords
Churn analysis, Data mining, SVM, Telecommunication

Abstract: Data mining is used to analyze mass databases for having meaningful output. One of the most common applications of the data mining, which is called as Churn Analysis is used to predict behavior of customers who are most likely to change provided service, and to create special marketing tools for them. The aim of this paper is to determine customers who want to churn, and to create specific campaigns to them by using a customer data of a major telecommunication firm. In order to determine the customer profile to leave, the SVM algorithm was studied and compared with other algorithms in terms of accuracy estimation.

1. Giriş

Birçok şirket için müşteri kaybetme nedenlerini bulmak, müşteri sadakatini ölçmek ve müşteriyi geri kazanmak çok önemli kavramlar haline gelmiştir. Firmalar yeni müşteri kazanmak yerine müşterilerini kaybetmemek için çeşitli çalışmalar ve kampanyalar düzenlemektedir. Telekomünikasyon sektörü, hızla yenilenebilir teknolojiler, abone sayısındaki artış ve katma değerli hizmetler nedeniyle büyük miktarda veri elde etmektedir. Bu alanın kontrolsüz ve çok hızlı genişlemesi, dolandırıcılık ve teknik zorluklara bağlı olarak artan kayıplara neden olmaktadır. Bu nedenle yeni analiz yöntemlerinin geliştirilmesi bir zorunluluk haline gelmiştir.

Veri madenciliği algoritmaları ve bilgi keşfi çerçevesi, ticaret, astronomi, jeolojik araştırma, güvenlik ve telekomünikasyon dahil olmak üzere bir dizi uygulama alanında başarıyla uygulanmıştır [1]. Ren, Zheng ve Wu [2] yaptıkları çalışmada telekomünikasyon müşteri alt bölümü için genetik algoritmaya dayalı bir kümeleme yöntemi sunmuşlardır. İlk olarak, telekomünikasyon müşterilerinin özellikleri (arama davranışı ve tüketim davranışı gibi) ayıklanır. Daha sonra, telekomünikasyon müşterilerinin çok boyutlu öznitelik vektörleri arasındaki benzerlikler, iki boyutlu bir düzlemde örnekler arasındaki mesafe olarak hesaplanır ve haritalanır. Son olarak, genetik algoritma ile kademeli olarak benzerliklere yaklaşmak için mesafeler ayarlanır.

Wei ve Chiu [3] bir tasarım önermiş ve abone sözleşme bilgilerinden müşteri kaybını ve arama detaylarından çıkarılan arama düzeni değişikliklerini tahmin eden bir kayıp tahmin tekniğini deneysel olarak değerlendirmiştir. Önerilen bu teknik, belirli bir tahmin zaman periyodu için sözleşme düzeyinde potansiyel müşteri kayıplarını belirleme yeteneğine sahiptir. Başvuru için Tayvan'ın 21 milyon abonesi olan en büyük telekom şirketi seçilmiştir. Ne kadar çok çağrı kaydına sahip olurlarsa Churn analizinden o kadar doğru sonuçlar alabileceklerini ifade etmişlerdir.

Bugün Türkiye'de tüm telekomünikasyon şirketlerinin veri madenciliği kullandığı doğrudur. Telsim'i satın alan Vodafone, satış, pazarlama, finansal yönetim, gelecek tahmini ve birçok farklı ihtiyaç için veri madenciliği uyguluyor. Vodafone, veritabanlarını kullanarak yoğun saatleri tespit eder ve iletişimde herhangi bir kesinti olmaması için daha fazla iş gücünü hazır hale getirir. Ayrıca Vodafone, satın alınan ön ödemeli dakikaların ortalamasını belirler ve abonelik bırakma olasılığı olan aboneleri bulur [4].

Kuruluşlar, büyük olasılıkla hizmet sağlayıcıyı değiştirecek müşterileri tahmin ederek, müşteri sadakatini artırmayı amaçlayan kampanyalar oluşturabilir ve daha yüksek müşteri elde tutmak için pazarlama stratejileri geliştirebilir. Bu çalışmanın amacı, telekom firmasının müşterilerini kaybetme nedenlerini belirlemektir. Nedenlerin belirlenmesi gibi, ne tür müşterilerin kaybedildiğinin belirlenmesi de araştırılır.

2. Materyal ve Metot

Churn analizinde hangi müşteriler rakiplere gitmeye daha çok eğilim gösteriyor, bu soruya cevap aranır. Birçok sektör bu risk ile karşı karşıya kalmaktadır. Churn analizi şirketlere, müşterilerinin neden rakiplere yöneldiğini anlamalarını sağlar. Bu şekilde şirket, müşteri ilişkilerini yeniden düzene sokar ve müşteri bağlılığını arttırır [5].

Bir Churn analizi uygulamalarında ilk iş müşteri verilerine ulaşmaktır. Daha sonra, müşteri kaybı kararını hangi faktör veya faktörlerin etkilediğine karar vermek için faktörler sınıflandırılır. Hangi müşterilerin ayrılma ihtimalinin bulunduğunu belirledikten sonra, belirli bir zaman diliminde hedef müşterilere farklı ve spesifik pazarlama ve elde tutma stratejileri uygulanabilir.

2.1. Veriyi Anlama

Veri anlama aşaması, ilk veri toplama ile başlar ve verilere aşina olmak için veriyi işleme süreci ile devam eder. Veri kalitesi problemlerini belirlemek, verilere ilişkin ilk bilgileri keşfetmek ve gizli bilgilerden hipotezler oluşturmak için ilginç alt kümeleri tespit etmek bu adımın faaliyetleridir [6].

İnternet hizmetleri satan bir telekomünikasyon şirketi (TelCo), yaklaşık %27'lik büyük bir müşteri kaybı oranı yaşıyor. Bu düzeyde bir kayıp, gerçek hayattaki bir şirketi iflas ettirmek için yeterli olabilir. Kamuya açık müşteri verilerinin olmaması nedeniyle, 7.043 müşteri için etiketli kayıp içeren IBM Cognos Telco Customer Churn benzetilmiş veri seti bu çalışmada kullanılmıştır.

Müşteri kaybı analizi için 2 farklı abone verisine ihtiyaç vardır. Bunlardan birincisi aboneliğini iptal etmiş olan abonelere ait veriler, bir diğeri ise aktif olarak hizmeti kullanmaya devam eden abonelere ait verilerdir. Biz veri setimizde abonelik durumu ifade eden bu veriyi “churn” niteliği ile isimlendirecek ve bu niteliği eğitim, doğrulama ve test aşamalarında sınıf etiketi olarak kullanacağız.

Bu araştırmada veriler Guido Van rossum tarafından geliştirilmeye başlanan bir programlama dili olan Python ile analiz edilmiştir. Müşteri kaybı riskini sınıflandırmak için tahmine dayalı bir model geliştirilerek her bir tahmin edicinin modelin tahminleri üzerindeki göreli etkisini açıklamak hedeflenmiştir. Bu hedefin çıktıları doğrultusunda da müşteri kaybını azaltmak için potansiyel yaklaşımlar önerilebilir.

2.2. Veriseti

Telco Veriseti’nde her satır bir müşteriyi temsil eder, her sütun ise müşterinin meta veriler sütununda açıklanan özelliklerini içerir. Ham veriler, 7043 satır (müşteriler) ve 21 sütun (özellikler) içerir.

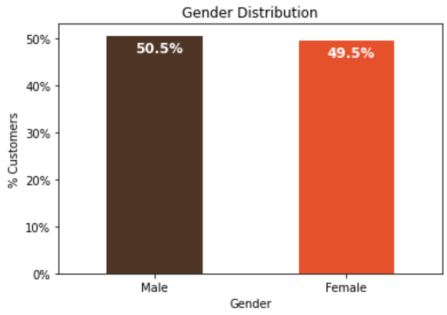
21 sütun yani değişkenler ise şu bilgileri içermektedir:

- Müşteriler hakkında demografik bilgiler - cinsiyet, yaş aralığı ve ortakları ve bakmakla yükümlü oldukları kişiler olup olmadığı
- Geçen ay içinde ayrılan müşteriler - sütunun adı Churn
- Her müşterinin kaydolduğu hizmetler – telefon, birden çok hat, internet, çevrimiçi güvenlik, çevrimiçi yedekleme, cihaz koruması, teknik destek ve TV ve film akışı
- Müşteri hesap bilgileri – ne kadar süredir müşteri oldukları, sözleşme, ödeme yöntemi, kağıtsız faturalandırma, aylık ücretler ve toplam ücretler

2.3. Analiz

İlk olarak veriseti için tanımlayıcı istatistikler analiz edilebilir. Bu ilk adım verisetindeki sapmaları ve anormalileri görmek, veri manipülasyonuna gidip gitmemek için verimli bir yoldur. Ayrıca verilerin kullanımı adına öngörü sağlaması için faydalıdır.

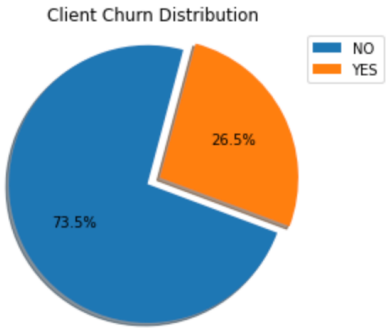
Müşterilerin cinsiyet dağılımına bakıldığında Şekil 1’de kadınlar toplamın %49.5’ini oluştururken; erkek müşteriler toplamın %50.5’ini oluşturmaktadır. Cinsiyet dağılımı oldukça yakındır.



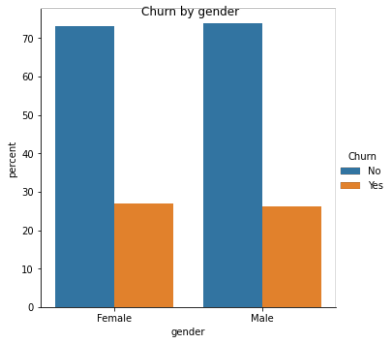
Şekil 1. Müşterilerin cinsiyet dağılım grafiği

Bir verideki değişkenlerin numerik olması hedef değişkeni tahmin etmede faydalıdır. Sadece numerik komutların çalıştığı kodlar için değişkenlerin numerik hale getirilmesi veriyi analiz etmede hata almamak için hazırlamış olur. Python'da bunun için "factorize" komutu kullanılmıştır. Bu komut verideki değişkenleri gözden geçirerek farklı değer tanımlamaları olduğunda, bir dizinin sayısal bir temsilini elde etmek için kullanışlıdır.

Verisetinde churn olan müşteri "yes"; churn olmayan müşteri ise "no" olarak tanımlanmıştır. Bu değişkenleri numerik hale getirmemiz analiz için daha sağlıklıdır. Churn müşteriler için "1"; churn olmayan müşteriler için "0" tanımlanır. Şekil 2'de churn eden ve churn etmeyen müşterilerin dağılımına bakılacak olursa Şekil 2'de müşterilerin %26.5'i churn olurken %73.5'i hala churn olmamış yani aktif kullanıcıdır. Şekil 3'te ise cinsiyete göre churn durumuna bakıldığında benzer davranışlar görülse de churn olan kadın müşteriler erkeklere kıyasla daha daha fazladır.



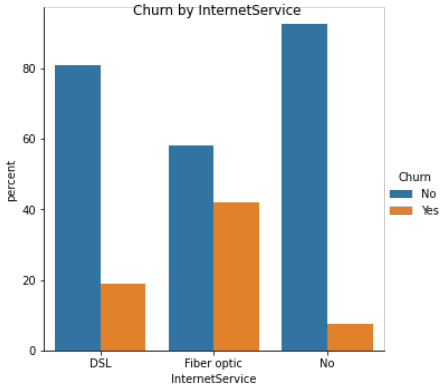
Şekil 2. Müşterilerin churn dağılım grafiği



Şekil 3. Müşterilerin cinsiyete göre churn dağılım grafiği

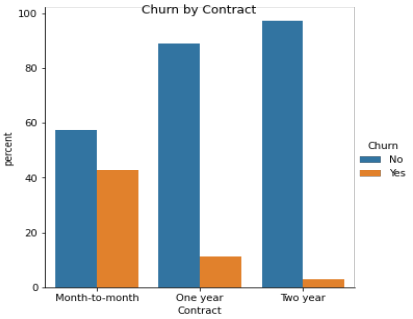
Koronavirüs süreci Çin'de başladıktan kısa bir süre sonra tüm dünyayı tehdit eder hale gelmiştir. Pandemi önleyici olarak birçok tedbir alınmıştır. Eğitim, çalışma ve birçok faaliyet uzaktan yürütülmek zorunda kalmıştır. Bu dönemde internet hızı, internete erişim ve internet kalitesi de firmalar için önemli hale gelmiş bunun yanında gözden geçirilmesi gereken bazı problemleri beraberinde getirmiştir. Bu gibi kriz dönemlerinde internet servis hizmetinin de churn durumları için önemli olduğunu ortaya çıkarır.

Şekil 4'e bakılırsa internet servis türlerine göre churn dağılımlarını görebiliriz. İnternet servis tipi hızı etkileyen bir değişkendir. Fiber'de daha hızlı internet erişimi olurken DSL'de ise çıkacağı hız daha düşüktür. Bu değişkenler müşteri deneyimini etkileyen temel faktörlerdendir. İnternet servisi olmayanların bu sebepten churn etme durumu düşük olasılıklıdır. Fiber optik internet servisinde churn eden müşteriler %40 üzerinde olup en fazladır. Bunun sebebi ise internet hız beklentisi yüksek müşterinin beklediği hizmeti alamaması olabilir.



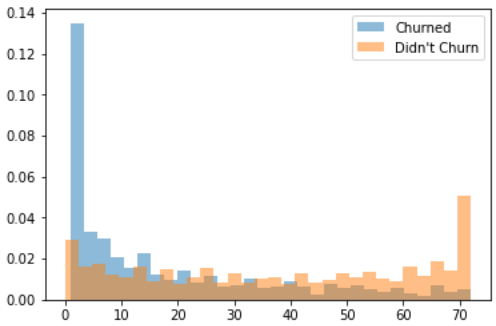
Şekil 4. Müşterilerin internet hizmetine göre churn dağılım grafiği

Müşteriler alacakları hizmet karşılığında belli bir abonelik sözleşmeleri ile abone olmaya başlarlar. Müşterilen abonelik dönemleri bittiğinde ise tekrar sözleşme yenilenebilir. Bu dönemlerde müşteri deneyimden memnun kalmadıysa abonelikten ayrılabilir bu durumda da churn olur. Eğer müşterinin sözleşmesi bitmeden aboneliği sonlandırmak isterse belli bir taahhüt ücreti ödemesi gereklidir. Aydan aya sözleşmesi bulunan müşterilerin daha fazla churn yüzdesine sahip olduğunu görmekteyiz. Sözleşme yılı arttıkça churn de azalmaktadır.



Şekil 5. Müşterilerin abonelik sözleşmesine göre churn dağılım grafiği

Genelde uzun süreli hizmet alan müşterilerin churn etme yüzdesinin düşük olduğu görülür. Yeni katılan müşterilerde ise ilk 10 ay hizmet deneyimi açısından çok önemlidir. Bu sebeple yeni gelen müşterilerin ilk 10 ay hizmet deneyimi firmalar için oldukça kıymetlidir. Şekil 6'da hizmet süresi ay cinsinden gösterilmiştir. Bu grafikte hizmet süresi ile churn dağılımına bakılabilir. İlk aylar kritik ve churn etme olasılığı yüksek olan zaman dilimidir. Bu aylarda hizmet deneyimi takibi ve iyileştirmesi sürekli hale getirilmelidir.



Şekil 6. Hizmet süresi dağılım grafiği

3. Bulgular

Demografik bilgiler grafiklerden çıkarım yapacak olursak aydan aya sözleşmesi olan, fiber optik internet hizmeti olan müşteriler daha fazla churn eğiliminde iken, iki yıllık sözleşmesi olan ve internet hizmeti olmayan müşteriler churn olmama eğilimindedir.

Korelasyon matrisi çoklu değişkenler arasındaki korelasyon katsayılarının tablosudur. Bu tabloda bir değişkenin diğer her değişken ile arasındaki korelasyon görülebilir. Korelasyon -1 ile 1 arasında değerler alır. 0'dan büyük bir korelasyon varsa pozitif yönlü ve 1'e yaklaştıkça güçlü bir korelasyon söz konusu iken 0'dan küçük negatif ve -1'e yaklaştıkça güçlü negatif korelasyon vardır denebilir. Korelasyon 0 iken değişkenler arasında hiç ilişki yoktur. Korelasyonun 1 olduğu durum ise değişkenin kendisiyle ilişkisini temsil eder. Verilerin birbiriyle fazla ilişkili olması, analizimizi saptırabilir. Bu istenmeyen ve yanlılık yaratabilecek bir durumdur. Dolayısıyla korelasyon analiz için oldukça önemli yer tutar.

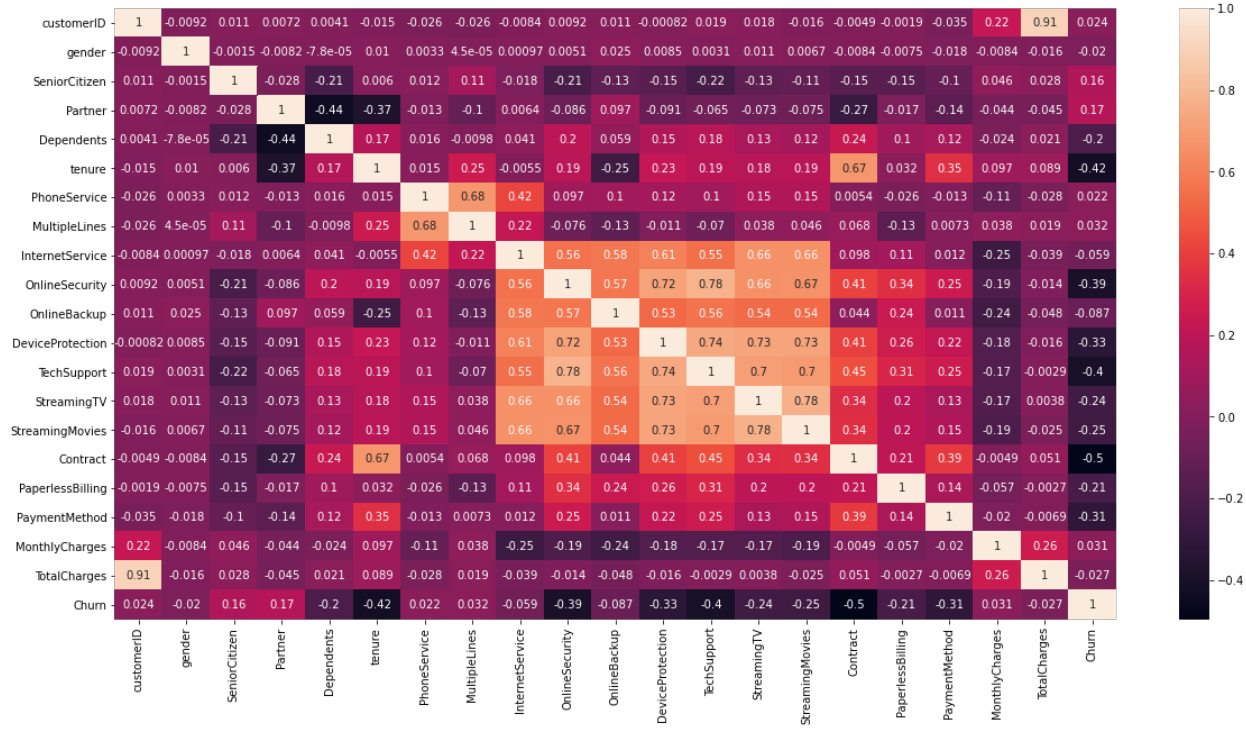
Aşağıda Tablo 1'de değişkenlerin birbiri ile ilişkisini gösteren ısı grafiğini görmekteyiz. Burada 1 gelen sonuçlar değişkenin kendisiyle ilişkisi sebebiyle 1'dir. Anlamlı bir sonuç sayılamaz. Renkler açıldıkça pozitif yönlü güçlü bir korelasyon söz konusu iken renkler koyulaştıkça negatif yönlü güçlü korelasyon vardır yorumunda bulunulabilir.

- Teknik destek ile çevrimiçi güvenlik değişkenleri arasında 0.78 ile pozitif yönlü ve güçlü bir ilişki vardır. (Techsupport & Onlinesecurity)

- Müşterinin toplam ödediği ücreti ile müşteri id arasındaki ilişki 0.91 ile pozitif yönde ve oldukça güçlüdür bunun sebebi ise müşterinin kendisiyle direkt ilişkili olmasıdır. (Customer id & TotalCharges)

- Film ve dizi izleme servisleri ile kanal servisleri arasında 0.78 ile pozitif yönlü ve güçlü bir ilişki vardır. Müşteri kanal servisinden memnun ise film ya da dizi izleme servisinden de memnun kalacaktır. (Streamingmovies & StreamingTV)

Tablo 1. Değişkenlerin birbiri ile ilişkisini gösteren ısı haritası grafiği



Bağımlı değişken verisetinden çıkarılarak churn tahmini farklı algoritma ve modellemelerle tahmin edilerek tahmin performansı karşılaştırılabilir.

Analizde verisetini train ve test olarak %20-%80 olarak bölünmüş olup SVM(Support Vector Machine) sınıflandırma KNN, lojistik regresyon, naive bayes, karar ağacı ve rasgele orman algoritmalarıyla karşılaştırılmıştır. KNN için K = 11 alındığında en yüksek accuracy %78.7 olarak bulunmuştur. Lojistik regresyonda %79.8; bizim üzerine çalıştığımız SVM sınıflandırma için accuracy %79.5; naive bayes accuracy %72.1; karar ağacı %71.7 ve rasgele orman ise 5 tree için %76.6 doğrulukla sınıflandırmıştır.

Tablo 2. Sınıflandırıların doğruluk karşılaştırması

SINIFLANDIRICI	ACCURACY (%)
KNN	78.7
LOJİSTİK REG.	79.8
SVM	79.5
NAİVE BAYES	72.1
KARAR AĞACI	71.7
RASGELE ORMAN	76.6

4. Tartışma ve Sonuç

Bu çalışmada, telekomünikasyon sektöründe faaliyet gösteren bir şirketin verileri, kayıp müşteri davranışlarını tahmin etmeye yönelik modeller ortaya koymak, müşteri ilişkileri yönetimini geliştirmek, müşteriyi elde tutma ve sadakatine yönelik çeşitli kampanyalar ve pazarlama stratejileri geliştirmek amacıyla veri madenciliği teknikleri ile analiz edilmektedir. Demografik çıkarımlar, ilişkili değişkenlerin belirlenmesi ve hazırlık aşamalarının ardından müşteri kaybı tahmini için çeşitli algoritmalar uygulanmıştır.

Lojistik Regresyon en yüksek doğruluğu verirken bizim üzerine pythonda çalıştığımız SVM ise oldukça yakın bir sonuç vermiştir.
Karar ağacı en düşük doğruluk oranını vermiştir.

Aydan aya sözleşmesi olan kişiler, uzun vadeli sözleşmeleri olan kişilere göre daha fazla kayıp verme eğilimindedir.
Hizmet süresi arttıkça, kayıp olasılığı azalır.
Aylık ücretler arttıkça, kayıp olasılığı da artar.
Cinsiyetin ise müşteri kaybında çok da ayırıcı etkisi olduğu söylenemez.

Kaynakça

[1] P. Fule, Exploratory Medical Knowledge Discovery: Experiences and Issues. ACM SIGKDD Explorations Newsletter. 5(1), 94-99 (2003).

[2] H. Ren, Y. Zheng, Y. Wu, Clustering Analysis of Telecommunication Customers. The Journal of China Universities of Post and Telecommunications. 16(2), 114-116 (2009).

[3] C.P. Wei, I.T. Chiu, Turning Telecommunications Call Details to Churn Prediction: A Data Mining Approach. Expert Systems with Applications. 23, 103-112 (2002).

[4] N. Akkaş, Kahin Şirketlerin Kehanetleri. http://www.sas.com/offices/europe/turkey/news/basindasas/inthenews_new_0109 08.htm (2010), (Erişim: 30.05.2021).

[5] Şimşek, G., Umman, T., (2010). Telekomünikasyon sektöründe müşteri ayrılma analizi. İstanbul Üniversitesi İşletme Fakültesi Dergisi, 39, 35-49.

[6] Cross Industry Standard Process for Data Mining, www.crisp-dm.org (2010), (Erişim: 31.05.2021).

[7] Churn Prediction on Telco Customers, www.kaggle.com/meldadede/churn-prediction-of-telco-customers (2019), (Erişim 08.06.2021)