T.C.
BEYKENT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
BİLGİSAYAR MÜHENDİSLİĞİ BİLİM DALI

DOĞRUDAN SATIŞ SEKTÖRÜNDE VERİ MADENCİLİĞİ TEKNİKLERİ İLE MÜŞTERİ KAYIP ANALİZİ

(Yüksek Lisans Tezi)

Tezi Hazırlayan:
Ceyhun ALEMDAR

İstanbul, 2019

T.C.
BEYKENT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
BİLGİSAYAR MÜHENDİSLİĞİ BİLİM DALI

DOĞRUDAN SATIŞ SEKTÖRÜNDE VERİ MADENCİLİĞİ TEKNİKLERİ İLE MÜŞTERİ KAYIP ANALİZİ

(Yüksek Lisans Tezi)

Tezi Hazırlayan:
Ceyhun ALEMDAR

Öğrenci No:
110820038

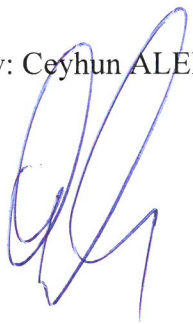
Danışman:
Dr. Öğr. Üyesi Ege KİPMAN

İstanbul, 2019

YEMİN METNİ

Yüksek lisans tezi olarak sunduğum "Doğrudan Satış Sektöründe Veri Madenciliği Teknikleri ile Müşteri Kayıp Analizi" başlıklı bu çalışmanın, bilimsel ahlak ve geleneklere uygun şekilde tarafımdan yazıldığını, yararlandığım eserlerin tamamının kaynaklarda gösterildiği ve çalışmamın içinde kullanıldıkları her yerde bunlara atıf yapıldığını belirtir ve bunu onurumla doğrularım. 24/05/2019

Aday: Ceyhun ALEMDAR



T.C.
BEYKENT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

YÜKSEK LİSANS TEZ SAVUNMA SINAVI SONUÇ TUTANAĞI

Beykent Üniversitesi
Fen Bilimleri Enstitüsü Müdürlüğü'ne,

Aşağıda tez adı belirtilen yüksek lisans öğrencisi. 11.0820038 .no'lu Ceylan ALEMDAR.'in 13/06/2019 tarihinde yapılan tez savunma sınavı[1] sonucunda... 45 dakika süreyle sunduğu ve savunduğu tezi hakkında[2] oybirliğiyle, BAŞARILI kararı verilmiştir.

Bilgilerinize saygılarımızla arz ederiz.

Anabilim Dalı : BİLGİSAYAR MÜHENDİSLİĞİ
Programı : BİLGİSAYAR MÜHENDİSLİĞİ
Tez Başlığı[3] : DOĞRUDAN SATIŞ SEKTÖRÜNDE VERİ MADENCİLİĞİ TEKNİKLERİ İLE MÜŞTERİ KAYIP ANALİZİ

Tez Sınav Jürisi	Öğretim Üyesi	İmza
Danışman	: Dr. Öğr. Üyesi Ege KIPMAN	
Üye	: Prof. Dr. Gökhan SİLAHTAROĞLU	
Üye	: Dr. Öğr. Üyesi Atınç YILMAZ	

---

[1] Jüri üyeleri, söz konusu tezin kendilerine teslim edildiği tarihten itibaren en geç bir ay içinde toplanarak öğrenciyi tez sınavına alır. Tez savunma sınav süresi en az 45, en çok 90 dakikadır. Jüri üyeleri, sınav öncesi yapılacak toplantıda, kendi aralarından danışman dışında bir üyeyi başkan seçer. Tez sınavı, tez çalışmasının sunulması ve bunu izleyen soru-cevap bölümünden oluşur. Tez sınavı, öğretim elemanları, lisansüstü öğrenciler ve alanın uzmanlarından oluşan dinleyicilerin katılımına açık ortamlarda gerçekleştirilir. Belirlenen günde yapılamayan jüri toplantısı, katılanların hazırladığı bir tutanakla enstitü yönetimine bildirilir. Bu durumda, jüri en geç on beş gün içinde toplanarak adayı tez savunma sınavına alır. (05 Ağustos 2017 tarihli 30145 sayılı Resmi Gazetede Yayınlanan Değişiklik-Madde 29-3)

[2] Tez sınavının tamamlanmasından sonra jüri, tez hakkında salt çoğunlukla "kabul", "düzeltme" veya "ret" kararı verir. Jüri başkanı, jüri üyelerince imzalanmış karar tutanağını, tez sınavını izleyen üç gün içinde ilgili enstitü yönetimine teslim eder. Tezi hakkında düzeltme kararı verilen öğrenci en geç üç ay içinde gerekli düzeltmeleri yaparak ve birinci fıkradaki usule göre tezini aynı jüri önünde yeniden savunur. Süresi içerisinde "düzeltme" savunmasına girmeyen öğrencinin enstitü ile ilişiği kesilir. (Beykent Üniversitesi Lisansüstü Eğitim ve Öğretim Yönetmeliği-Madde 29-4)

[3] İleride doğabilecek aksaklıkların engellenmesi için tezin başlığının yazılması gerekmektedir.

Adı ve Soyadı: Ceyhun ALEMDAR
Danışmanı: Dr. Öğr. Üyesi Ege KİPMAN
Türü ve Tarihi: Yüksek Lisans, 2019
Alanı : Bilgisayar Mühendisliği
Anahtar Kelimeler: Sınıflandırma algoritmaları, Karar ağacı, Veri madenciliği, doğrudan satış, R

ÖZ

DOĞRUDAN SATIŞ SEKTÖRÜNDE VERİ MADENCİLİĞİ TEKNİKLERİ İLE MÜŞTERİ KAYIP ANALİZİ

Bu çalışmada, bir doğrudan satış firmasından temin edilen Temsilci verileri kullanılmıştır. Temin edilen veriler üzerinde veri madenciliği yöntemlerinden olan sınıflandırma algoritmaları ile müşteri kayıp analizi yapılmıştır. Analiz yapılırken Veri Madenciliği için Çapraz Endüstri Standard Süreç Modeli (CRISP) adımları takip edilmiş ve uygulanmıştır. Çalışmada sınıflandırma algoritmalarından karar ağaçları tercih edilmiş, C4.5 karar ağacı ve Gini karar ağacı algoritmaları kullanılmıştır.

C4.5 karar ağacı ve Gini karar ağacı algoritması ile elde edilen modellerin performansları hold-out performans yöntemi ile ölçülmüş ve değerlendirilmiştir. Hold-out yöntemi ile veri seti sırasıyla eğitim ve test veri seti olmak üzere %90-%10, %80-%20, %70-%30, %60-%40 oranlarıyla ayrılmıştır. Yapılan analiz çalışmasında R programlama dili kullanılmıştır.

Name and Surname: Ceyhun ALEMDAR
Supervisor: Dr. Ege KİPMAN
Degree and Date: Master, 2019
Major: Computer Engineering
Key Words: Classification algorithms, Decision tree, Data mining, Direct sales, R

ABSTRACT

CUSTOMER CHURN ANALYSIS WITH DATA MINING TECHNIQUES IN DIRECT SALES SECTOR

All representative data is obtained from a direct sales company for this study. Customer loss has been calculated, by classification algorithms which are data mining methods, with the data provided. Cross Industry Standard Process Model (CRISP) steps is followed for the data mining while analyzing. Decision trees have been chosen from classification algorithms. In addition to this C4.5 decision tree and Gini decision tree algorithms were used in this study.

The performance of the models obtained by C4.5 decision tree and Gini decision tree algorithms were measured and evaluated by hold-out performance method. The data set is separated with hold-out method by %90-%10, %80-%20, %70-%30, %60-%40 respectively. R programming language has been used for the analysis conducted.

İÇİNDEKİLER

TABLOLAR LİSTESİ

Sayfa No.

ŞEKİLLER LİSTESİ

Sayfa No.

KISALTMALAR

ABD **:** Amerika Birleşik Devletleri

ACC **:** Accuracy

BSGM **:** Bölge Satış Grup Müdürü

BSM **:** Bölge Satış Müdürü

BSS **:** Bölge Satış Sorumlusu

CART **:** Classification and Regression Tree

CRISP : Cross Industry Standard Process

CRM **:** Customer Relationship Management

DOR **:** Diagnostic Odds Ratio

DSD **:** Doğrudan Satış Derneği

ERR **:** Error Rate

FPR **:** False Positive Rate

FNR **:** False Negative Rate

LR+ **:** Positive Likelihood Ratio

LR- **:** Negative Likelihood Ratio

MİY **:** Müşteri İlişkileri Yönetimi

PPV **:** Positive Predictive value

NPV **:** Negative Predictive Value

SL **:** Satış Lideri

TPR **:** True Positive Rate

1. GİRİŞ

Hiç durmaksızın büyüyen ve küreselleşen dünyada her gün binlerce şirket kuruluyor ve binlerce şirket iflas bildiriyor. Veriyi yönetmek, veriden anlam çıkarmak, bilgi edinmek, bir şirket için hayatta kalmanın vazgeçilmez yöntemlerinde biri. Çoğu zaman yeni müşteri edinmek var olan müşteri elde tutmaktan çok daha maliyetli olabiliyor. Durum böyleyken günümüz dünyasında şirketlerin kâr savaşlarında veri madenciliği, müşteri kayıp analiz yöntemleri en önemli silahlardan biri.

Doğrudan satış; bir ürünün veya hizmetin kullanıcılara/tüketicilere doğrudan sunulduğu bir pazarlama tekniğidir [14]. Doğrudan satış şirketleri pazarlama ve satış sürecini Temsilciler üzerinden yürütmektedirler. Temsilcilerin bir kısmı sadece kendileri için alışveriş yaparken bir diğer kısmı müşterilere satış yapmaktadır. Bu sebeple doğrudan satış sektöründe bir Temsilcinin kaybı onlarca müşteri kaybı anlamına gelebilmektedir. Bu nedenle müşteri kayıp analiz çalışması doğrudan satış sektörü için bir kat daha önem kazanmaktadır

Bu çalışma kapsamında veri madenciliği teknikleri kullanılarak doğrudan satış sektörü veri setinde müşteri kayıp analizi yapılması amaçlanmıştır. Analizlerde bir doğrudan satış şirketinden alınan gerçek veriler kullanılmıştır. Çalışmada 347.001 Temsilci verisi kullanılmıştır. Analiz çalışmaları R programlama dili ile yapılmıştır.

Tez çalışmasında uygulanan teknik ve yöntemler sırasıyla adım adım anlatılmıştır. Bölüm 2.'de doğrudan satış sektörü genel terimleri ile birlikte tanıtılmış. Bölüm 3.'te müşteri kayıp analizi tanımlanmış ve amaçlarından bahsedilmiştir. Bu alanda yapılan çalışmalar hakkında bilgi verilmiştir. Doğrudan satış sektöründe müşteri kayıp analiz çalışmalarının yetersizliğinden bahsedilmiştir. Bölüm 4.'te R programla dili hakkında genel bilgiler verilmiştir. Bölüm 5.'te veri madenciliği hakkında genel bilgi verilmiş, çalışma boyunca takip edilen veri madenciliği süreçleri detaylı şekilde anlatılmıştır. Bölüm 6.'da CRISP modeli anlatılmış ve çalışmada nasıl uygulandığı anlatılmıştır. Bölüm 7.'de Gini Karar ağacı ve C4.5 karar ağacı ile oluşturulan modeller karşılaştırılmıştır, sonuçlar değerlendirilmiştir.

2. DOĞRUDAN SATIŞ

Doğrudan satış; bir ürünün veya hizmetin kullanıcılara/tüketicilere doğrudan sunulduğu bir pazarlama tekniğidir [14]. Bir başka ifadeyle doğrudan satış, bir bağımsız satıcının tüketim malının veya hizmetinin satıcının evinde, ofisinde veya işyerinde ya da ürün ve hizmete talip olan tüketicinin evinde, bir tanıdığının evinde, ofisinde, iş yerinde ya da satış noktası olmayan başka bir yerde doğrudan satışının gerçekleştirlmesi olarak da tanımlanabilir [15].

2.1. Doğrudan Satış'ın Tarihçesi

Doğrudan satış oldukça eski bir kavramdır, öyle ki katalogların 1948 yılında Venedik'te kitapların tanıtılması ve pazarlanmasında kullanıldığı görülmüştür. 1700'lü yıllarda benzer şekilde Avrupa'da bitki ve tohum katalogları kullanılmıştır. Bu yöntemin yani katalog ile ürünlerin tanıtılması ve satışı özellikle kırsal bölgede yaşayan ve kent merkezine sıklıkla uğrama olanağı olanayan köylü tüketiciler için kullanıldığı görülmüştür. 1800'lü yılların sonlarına doğru, doğrudan satış artık ABD'de de görünmeye başlanmıştır [16].

Doğrudan satış pazarı özellikle 2.Dünya savaşından sonra hızlı bir büyüme katedmiştir. Teknolojik gelişmeler, kredi kartları ile tanışmamız ve her sektörde olduğu gibi bilgisayarların hayatımızda yer almaya başlamasıyla 1980'li yıllar ve sonrasında doğrudan satış hızla büyümeye devam etmiştir [16].

Türkiye'de doğrudan satış 1970'li yıllarda yerel bazı firmaların kitap ve ansiklopedi satışı ile başlamıştır. 1970'li yılların sonlarına doğru doğrudan satış sektöründe dünya lideri olan Avon, Amway ve Oriflame gibi firmaların Türkiye pazarına girmesiyle yaygınlaşmıştır [15]. Türkiye'de şu anda kozmetikten küçük ev aletleri ve gıda takviye ürünlerine kadar bir çok ürün gamında doğrudan satış pazarlama tekniği kullanılmaktadır.

2.2. Diğer Satış Modelleri ve Doğrudan Satış

En temelde tüm satış modelleri tüketicilere ürün ve hizmet sunmayı amaçlar. Doğrudan satış yöntemi, müşteri ile satıcıyı oldukça farklı bir kanal kullanarak buluşturur. Diğer tüm satış modellerinde satış şirketleri kendi çalışanları ile satış yaparken doğrudan satış şirketleri bağımsız satıcılar ve onların sosyal cevreleri ile satış gerçekleştirir [14].

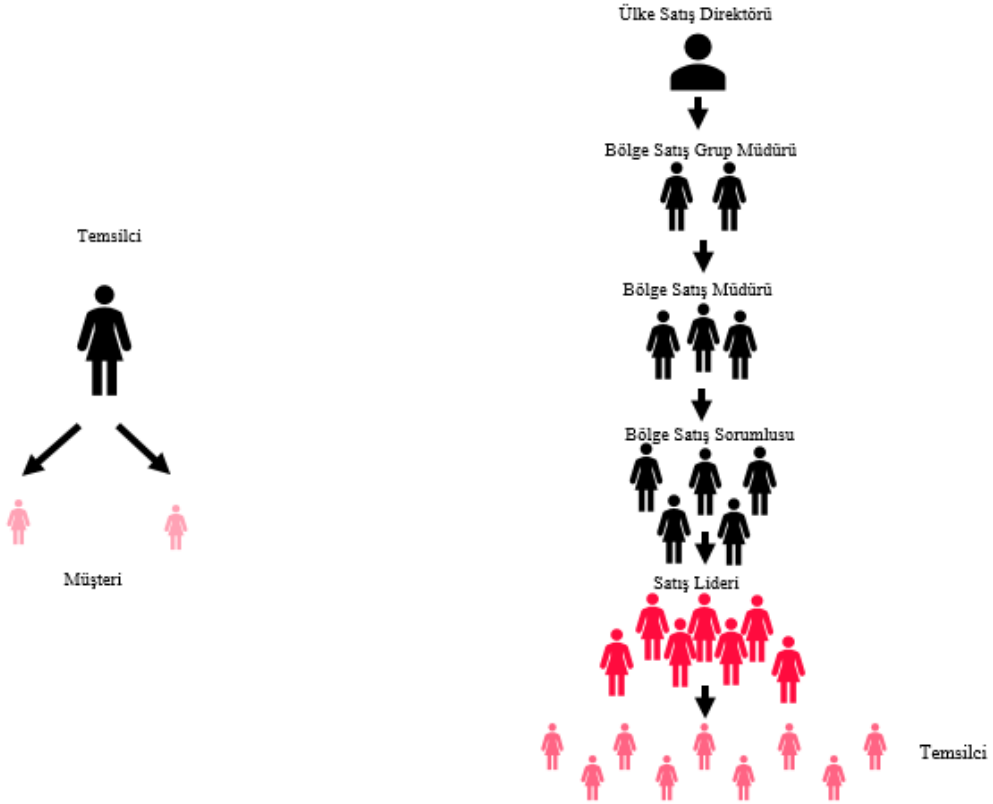
2.3. Doğrudan Satış ve Piramit Düzenler

Doğrudan satış Ülkemizde ve bir çok ülkede yasal bir pazarlama ymntemi ve tekniğidir. Piramit düzenler ise kötü niyetli kişilerin tüketicileri dolandırmaya yönelik kurguladıkları yapılardır. Doğrudan Satış Derneği (DSD) aşağıdaki üç madde/ilke ile Piramit düzeni doğrudan satıştan ayırmaktadır [15].

- Doğrudan satışın amacı pazarlama maliyetleri düşürmek ve Tüketiciye uygun fiyata katileli ve hızlı, ürün ve hizmet sunmaktır. Bunun için başlama ücreti talep edilmez.
- Ürünler değeri ve kalitesi şüphe götürecek şekilde göstermelik olmamalıdır. Temil kriter tüketim ürünü kazanç sağlanmayacak olsa bile alınabilir nitelikte bir tüketim ürünü olmalıdır.
- Doğrudan satışta ürünler tüketici için olmalı ve tüketiciye satılmalıdır. Temel amaçlardan biride tüketicilere satış yaparak bir pazar oluşturmaktır. Piramit düzenlerde ürünü kullanacak tüketici/kişilerle ilgilenilmez yeni katılanlara işe başlamak için satın almaları gerektiği söylenir.

2.4. Doğrudan Satış Sistemi, İşleyişi

Bu çalışmada kullanılan değişkenlerin daha iyi anlaşılabilmesi ve yorumlanabilmesi için doğrudan satış işleyişinin bilinmesi gerekir. Türkiyedeki doğrudan satış şirketleri müşterilerine farklı isimler ile hitap etmekte, bu çok katlı ağ pazarlama (network marketing) yapılarını farklı terimlerle tanıtmaktadırlar. Örneğin bir firma müşterilerini Temsilci, diğeri üye, bir diğeri girişimci olarak isimlendirmektedir. Bu çalışmada genel kabul görmüş Temsilci ve Satış Lideri (SL) terimleri kullanılmıştır. Sekil1.'de örnek bir doğrudan satış işleyiş şeması görünmektedir.



Şekil 1. Doğrudan Satış İşleyiş Seması

Şekil 1'de görülen Ülke Satış Direktörü, Bölge Satış Grup Müdürü (BSGM), Bölge Satış Müdürü (BSM) ve Bölge Satış Sorumlusu (BSS) doğrudan satış şirketi personeli iken Satış Lideri (SL) ve Temsilciler bağımsız satıcılardan oluşmaktadır.

2.4.1. Temsilci

Temsilciler doğrudan satış sistemine birden farklı amaç ile kayıt olmuş olabilirler. Bu amaçlar ise; kendileri için indirimli ürün almak, kendileri ve aile fertleri için indirimli ürün almak, ürünleri müşterilere pazarlayarak gelir elde etmek ve sosyal bir çevre edinmek olarak sıralanabilir. Temsilcilerin sisteme hangi nedenle kayıt oldukları anlamak oldukça zor olduğu gibi zamanla Temsilcilerin sistemden beklentileri de değişkenlik gösterebilmektedir. Örneğin kendine indirimli ürün almak için kayıt olmuş bir Temsilci bir müddet sonra ürünleri pazarlayarak gelir etme amacı edinmiş olabilir. Bu çalışma özelinde detaylı olarak anlattığımız Temsilci grubunun sistemden ayrılma yani müsteri kayp analiz (customer churn analysıs) çalışması yapılacaktır. Temsilciler sattıkları ürünlerden belirli bir kazanç elde ettikleri gibi teşvik

programlarından da çeşitli ürünler, tatiller, araç kullanım hakkı ve bir çok hediye kazanabilmektedirler.

2.4.2. Satış Lideri (SL)

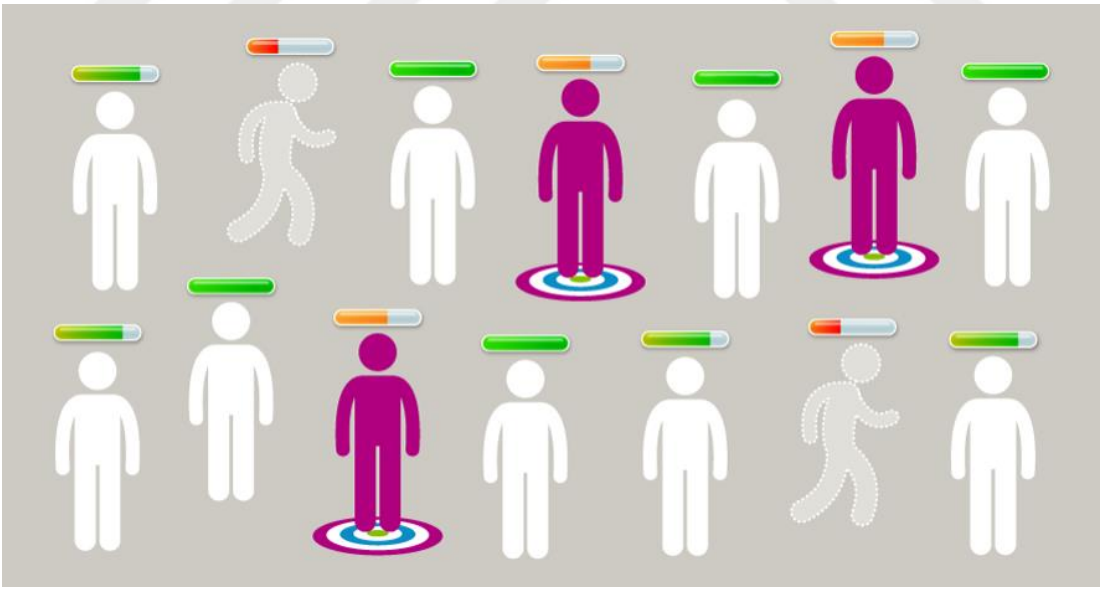
Doğrudan satışta Satış Lideri (SL)'nin rolü ve kazanç sistemi Temsilcilciden daha farklıdır. SL, bir Temsilci gibi ürün pazarlayabileceği gibi sisteme kayıt ettiği Temsilcilerin satış hacminden de komisyon kazanmaktadır. SL, katlı bir yapıyla büyük bir ekip oluşturup profesyonel bir çalışma sistemi kurabilir ve komisyonunu bu sayede arttırabilir.

3. MÜŞTERİ KAYIP ANALİZİ

Müşteri kayıp analizi genellikle telekom, bankacılık ve sigortacılık sektörlerinde kullanılan mevcut müşterilerin alışveriş alışkanlıklarını ve kaybını önceden tahmin etmeye dayanan analiz yöntemidir. Yapılan bu analiz çalışması sonrasında müşteri ilişkileri yönetimi (MİY) ingilizce terimi ile customer relationship management (CRM) kapsamında müşteri kaybını önleyecek veya en aza indirecek yöntem ve çözümler geliştirilmektedir [1].

Telekom, bankacılık, sigortacılık gibi müşteri sürekliliği/devamlılığı bulunan sektörlerde müşteri kaybı oldukça kritik bir öneme sahiptir. Çünkü bu sektörlerde mevcut müşterilerin sistemde tutulması, yeni müşteri kazanmaya göre daha düşük maliyetli operasyonlar gerektirir [2].

Müşteri kayıp analizleri için litaratürde, müşteri kaçması (customer defection), müşteri cayması (customer turnover), müşteri yıpranması (customer attrition) ve müşteri sallanması (customer churn) gibi terimler de kullanılmaktadır.

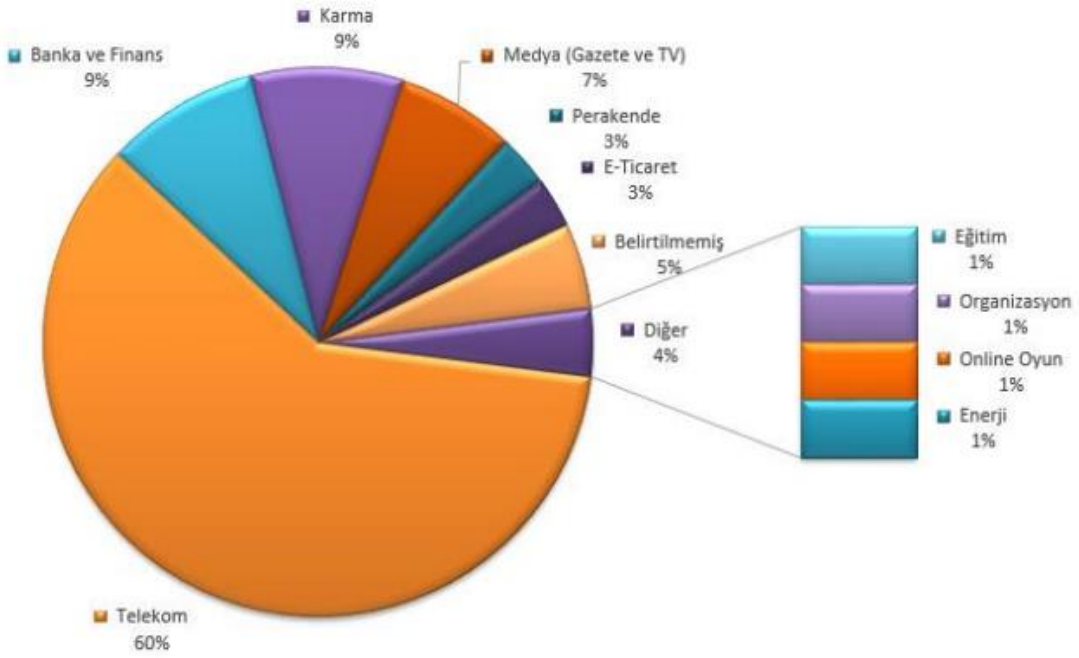


Şekil 2. Müşteri kayıp analizi (customer churn analysıs) ile müşteri durumu grafik gösterimi [17]

Kaynak: Kılıç, D., **Makine Öğrenimi ve Derin Öğrenme ile Müşteri Kayıp (Churn) Analizi-1,** https://medium.com/deep-learning-turkiye/makine-%C3%B6%C4%9Frenimi-ve-derin-%C3%B6%C4%9Frenme-ile-m%C3%BC%C5%9Fteri-kay%C4%B1p-churn-analizi-1-63a4513b8a6f (Erişim Tarihi: 05.03.2019)

3.1. Müşteri Kayıp Analiz Çalışmaları Literatür Taraması

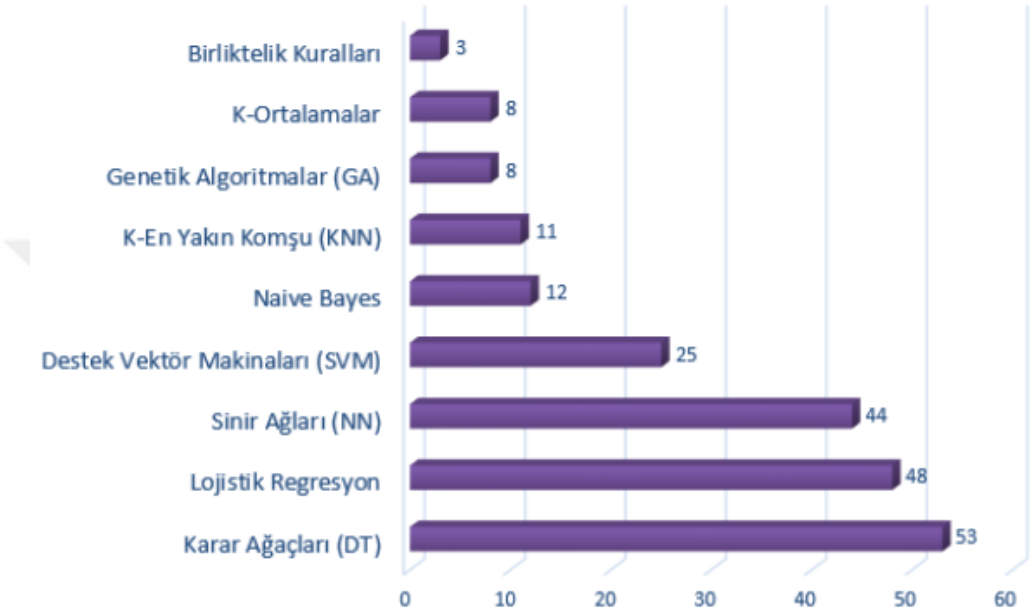
Değerlendirilen 100 makalede sektörlere göre dağılım Şekil 3.'te görülmektedir. Özetle, %60 oranında Telekominikasyon, %9 Banka ve Finans, %9 Karma (Telekom, Banka, Perakende vb.), %7 Medya, %1 Eğitim, %1 Online, %1 Oyun ve %1 Enerji ve Organizasyon sektörlerinde çalışma yapılmıştır [3].



Şekil 3. Makalelerde Yer Alan Uygulamaların Sektörlere Göre Dağılımı

Kaynak: Önay Koçoğlu, F., Özcan, T., Baray, Ş. A., **Veri Madenciliğinde Ayrılan müşteri Analizi Problemi Üzerine Bir Literatür Araştırması,** 2016

İncelenen 100 makalede müşteri kayıp analizi çalışmalarında kullanılan veri madenciliği yentemleri Sekil 4.'te gözterilmiştir. Özetle, 53 makalede Karar Ağacı algoritmaları, 48'inde Lojistik Regresyon, 44'ünde Sinir Ağları ve 25'inde ise Destek Vektör Makineleri kullanılmışıtr. Her çalışmada birdern fazla yöntem kullanıldığı hatırdan çıkarılmamalıdır [3].



Şekil 4. Makalelerde Kullanılan Veri Madenciliği Yöntemlerinin Frekans Dağılımı

Kaynak: Şeker, Ş. E., **Müşteri Kayıp Analizi (Customer Churn Analysis),** YBS Ansiklopedi, Cilt 3, Sayı 1, 2016, ss. 26-27.

Yapılan literatür çalışmasında da görüldüğü üzere niş bir sektör olan doğrudan satış sektöründe müşteri kayıp analiz çalışması neredeyse hiç yapılmamıştır. Oysaki müşteri kaybının en fazla önemsenmesi gereken sektörlerin başında dağrudan satış gelmektedir. Öyle ki doğrudan satış sistemindeki bir Temsilci doğrudan satış şirketinin ürünlerini onlarca müşteriye ulaştırıyor olabilir. Ya da sistemden çıkan bir SL onlarca Temsilciyi sistemden çıkarabilir bu temsilciler ile de doğrudan satış şirketi binlerce müşterisini kaybedebilir.

4. R PROGRAMLAMA DİLİ

Bu çalışmada kullanılan algoritmalar, veri işleme ve manipülasyon işlemlerinin bir kısmı ve yine veri görselleştirme işlemlerinin bir kısmı R programlama dili kullanılarak yapılmıştır. Bu maksatla bu bölümde genel ve kısaca tezi okuyacak araştırmacıya R programlama dili hakkında bilgi verilmek istenmiştir [4].

R, istatistiksel ve ekonometrik hesaplamalar gibi özelliklerinin yanında veri manipülasyonu, programlama ve grafiksel gösterimine olanak sağlayan bütünleşmiş bir yazılım ortamıdır [4]. C, C++ ve Fortran yazılan kodların R' a kolay ve hızlı bir şekilde eklenebilmesi, dilin gelişimini hızlandırmıştır. Yazım şekli bakımından R ALGOL dil ailesinden oldukça fazla özelliği miras almıştır [9]. Bu sebeple genel kabul görmüş diller ile kodlama yapan geliştiricilerin dile adapte olması biraz vakit alabilmektedir.

R, daha önce geliştirlmiş olan S programlama dilinden türetilmiştir. S, 1976 yılında Bell Laboratuvarlarında John Chamber ve ekibi tarafından geliştirlmişitr. R dili S ve daha sonra geliştirilen S Plus dillerinin bir türevidir. 1991 yılında Auckland Üniversitesi istatistik bölümünde Ross Ihaka ve Robert Gentlement tarafından geliştirilmiş ve 1993 yılında duyurulmuştur. 1995 yılında açık yazılım haline gelmiş, 1997 yılında R çalışma grubu kurulmuştur [4].

R'in ilk sürümü 1.0.0. 2000 yılında yayınlanmıştır. Bu çalışmada ise 3.5.3 sürümü kullanılmıştır.

5. VERİ MADENCİLİĞİ

Veri madenciliği en temel tanımıyla büyük veriler arasından bilgiye ulaşma işidir [5]. Bir başka şekilde veri madenciliğini, büyük miktarlardaki veri içerisinden, gizli kalmış kıymetli ve kullanılabilir bilgileri açığa çıkarmak şeklinde de tanımlayabiliriz [6].

Bilgiyi toplamanın, toplanan bu büyük veriye ulaşmanın ve bilgisayarların performansları ile veriyi işlemenin kolaylaştığı günümüzde veri madenciliği hızla gelişmektedir. Bu gelişimi üniversite proje ve tezleriyle birlikte başarılı açık kaynak kodlu analiz ve programlama dilleri de desteklemektedir.

5.1. Veri Madenciliği Tarihçesi

1950'li yıllarda bilgisayarlar sayımlar için kullanılırken 1960'lı yıllarda veri tabanı ve verilerin depalanması/saklanması kavramı teknoloji dünyasında yerini almıştır. 1960'lı yılların sonunda bilim insanları basit öğrenmeli bilgisayarlar geliştirmişlerdir. Minsky ve Papert, günümüzde sinir ağları olarak bilinen perseptron'ların sadece çok basit olan kuralları öğrenebileceğini göstermişlerdir. 1970 yıllara gelindiğinde ise ilişkisel veri tabanı yönetim sistemleri kullanılmaya başlanmıştır. Bu gelişmeler ile birlikte bilim insanları basit kurallara dayanan sistemler geliştirmişler ve basit anlamda makine öğrenimini sağlamışlardır. 1980'lere gelindiğinde veri tabanı yönetim sistemleri yaygınlaşmış böylelikle bilimsel alanlarda uygulanmaya başlanmıştır. Veri tabanı yönetim sistemlerindeki bu gelişmelerle birlikte şirketler rakipleri, müşteriler ve ürünler ile ilgili bilgilerden oluşan veri tabanları oluşturmuşlardır. 1990'lı yıllara gelindiğinde hızla ve katlanarak artan veriden faydalı bilgilerin nasıl elde edilebileceği düşünülmeye başlanmıştır. 1989, KDD (IJCAI)-89 Veri Tabanlarında Bilgi Keşfi Çalışma Grubu toplantısı ve 1991, KDD (IJCAI)-89'un sonuç bildirgesi sayılabilecek "Knowledge Discovery in Real Databases: A Report on the IJCAI-89 Workshop" makalesinin KDD (Knowledge Discovery and Data Mining) ile ilgili temel tanım ve kavramları ortaya koyması ile süreç daha da hızlanmış. Bu gelişmeler sonrasında 1992 yılında veri madenciliği için

ilk yazılım gerçekleştirlmiştir. 2000'li yıllarda artık veri madenciliği inanılmaz bir gelişim sürecine girmiş ve neredeyse tüm alanlarda uygulanmaya başlamıştır [7].

5.2. Veri Madenciliği Uygulama Alanları

Veri madenciliği bankacılık, pazarlama, sigortacılık, sağlık, özellikle müşteri tabanlı alanlarla birlikte farklı alan ve sektörlerde uygulanmaktadır. Veri madenciliğinin kullanılmasında alan ve sektör farkı gözetilmemekle birlikte geniş veri ambarı oluşturulmasına olanak sağlayacak parakende satış, sigortacılık ve sağlık gibi sektörlerde kullanılması daha yaygın ve doğrudur [8]. Veri madenciliği bahsedilen tüm bu uygulama alanlarında kullanılırken Şekil 5.'te de görüleceği gibi birçok farklı disiplinle ilişki içindedir.

Şekil 5. Veri Madenciliğinin Diğer Disiplinler ile İlişkisi

Kaynak: Savaş, S., Topaloğlu, N., Yılmaz M., **Veri Madenciliği ve Türkiye'deki Uygulama örnekleri,** İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, Cilt 11, Sayı 21, 2012, ss. 4-5.

5.3. Veri Madenciliği Süreci

Veri madenciliği sürecini veya bir başka ifadeyler veri madenciliği aşamaları aşağıdaki şekilde sıralanabilir. Bu adımların ne manaya geldikleri, nasıl uygulandıkları yine bu bölümde özet olarak anlatılmıştır.

- Veri Temizleme
- Veri Bütünleştirme
- Veri İndirgeme
- Veri Dönüştürme
- Veri Madenciliği Algoritmalarının Uygulanması
- Sonuçları Sunum ve Değerlendirme

5.3.1. Veri Temizleme

Veri tabanında bulunan tutarsız ve hatalı veriler gürültü olarak isimlendirilir. Bu veriler çalışmaya dahil edilmeyebilir, yerlerine sabit bir değer atanabilir veya diğer verilerin ortalaması hesaplanarak bu ortalama değer kullanılabilir. Ya da karar ağacı veya regresyon gibi yöntemlerle tahmin edilerek çalışmaya dahil edilebilinir. Tüm bu işlemler ve gürültü olarak isimlendirilen verinin nasıl kullanılacağı ve yönetileceği kararı bu adımda verilir.

5.3.2. Veri Bütünleştirme

Birden fazla veri tabanından toplanan verilerin birlikte değerlendirmeye alınabilmesi için verinin tek türe dönüştürülmesi işlemidir. Örneğin cinsiyet bilgisi bir veri tabanında 1, 0 olarak, diğerinde E, K olarak, bir başka veri tabanında ise M, F olarak tutuluyor olabilir. İşte bu farklı işaretlenmiş verinin sadece E,K olarak gösterilemsi işlemine veri bütünleştirme denir.

5.3.3. Veri İndirgeme

Büyük boyuttaki veri ile çalışmak oldukça zordur. Bu sebeple analiz için kullanılan veriler azaltıldığında sonuç değişmiyorsa veri indirgeme işlemi gerçekleştirilir. Bu yöntemle daha hızlı sonuç elde edilebilinir.

5.3.4. Veri Dönüştürme

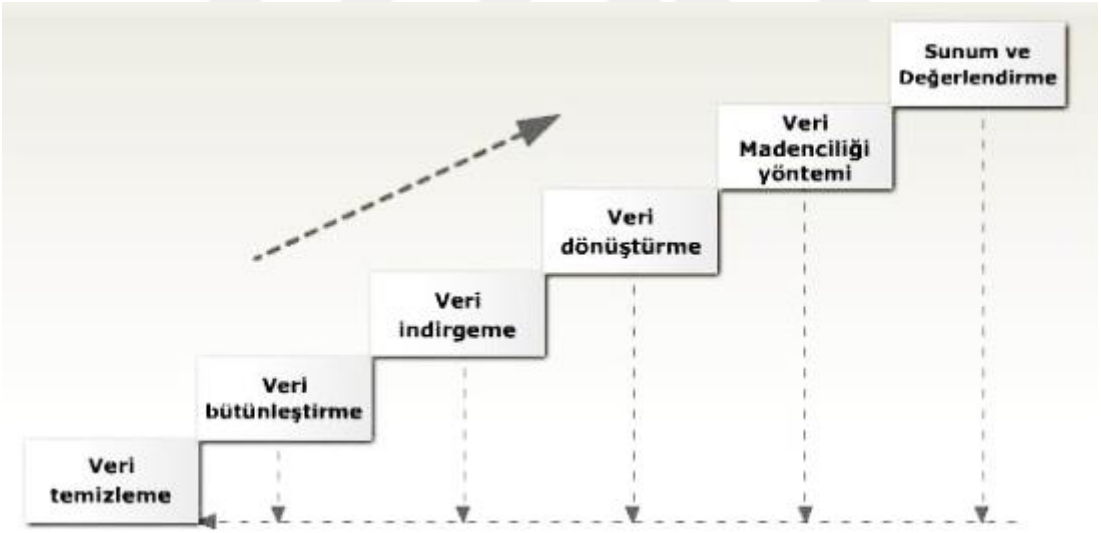
Verinin içeriğini korumak şartıyla, verinin kullanılacak modele uygun şekile dönüştürülme işlemidir. Örneğin bazı algoritmelar sadece nümerik veri ile çalışırken bazı algoritmalar sadece kategorik veri ile çalışmaktadır. Böyle bir durumda veri dönüştürme işlemi uygulanmaktadır.

5.3.5. Veri Madenciliği Algoritmalarının Uygulanması

Veri hazırlandıktan sonra yani gerekli görülmesi halinde veri temizleme, veri bütünleştirme, veri indirgeme ve veri dönüştürme işlemleri uygulandıktan sonra elde edilen veri setine veri madenciliği algoritmaları ile modellerin uygulandığı aşamadır.

5.3.6. Sonuçları Sunum ve Değerlendirme

Veri madenciliği algoritmaları ile modellerin uygulanmsı sonrasında elde edilen sonuçların karşılaştırılması ve değerlendirilmesidir.



Şekil 6. Veri Madenciliği Süreci

5.4. Veri Madenciliği Yöntemleri

Veri madenciliği yöntemleri; verileri kullanarak analiz edilmesi, sınıflandırılması, kategorize edilmesi, özetlenmesi ve birliktelik kurallarının tespit edilmesi için kullanılırlar. Bu çalışmada veri madenciliği yöntemleri; sınıflandırma, birliktelik kuralları ve ilişki analizi, kümeleme olmak üzere üç ana grupta anlatılmıştır.

5.4.1. Sınıflandırma Yöntemi

Veri madenciliği yöntemlerinden olan sınıflandırma yöntemi en çok bilinen veri madenciliği tekniklerinden birisidir. Sınıflandırma teknikleri resim, örüntü tanıma, hastalık tanıları, dolandırıcılık tespiti, kalite kontrol çalışmaları ve pazarlama konuları sınıflandırma tekniklerinin sıklıkla tercih edildiği alanlardır. Sınıflandırma teknikleri tahminleyici bir modeldir. Örneğin; hava sıcaklığının haftalık tahmini bir sınıflandırma işlemidir.

Veri madenciliğinde sınıflandırma işleminde; karar ağaçları (örneğin; C4.5, CART), istatistiğe dayalı algoritmalar (bayesyen sınıflandırma algoritmaları, regresyon), mesafeye dayalı sınıflandırma algoritmaları (k-en yakın komşu algoritması) ve yapay sinir ağları temelli algoritmalar gibi birçok teknik ve algoritma geliştirilmiştir [8].

5.4.2. Birliktelik Kuralları ve İlişki Analizi Yöntemi

İlişki analizi veri tabanındaki bir dizi bilgi, veri veya kaydın diğer kayıtlarla olan bağlantısını açıklayan işlemler dizisidir. İlişki analizi, bir kayıt varken herhangi bir başka kaydın var olma olasılığının cevabını verir. Aslında bir başka ifadeyle verilerin birlikte olan kurallarını ortaya çıkarır. Birliktelik kuralları ve ilişki analizi yöntemi sıklıkla pazar sepeti analizlerinde (Market Basket Analysis) tercih edilmekte ve kullanılmaktadır.

Birliktelik kuralları ve ilişki analizi yönteminde AIS algoritması, SETM algoritması ve AprioriTid algoritmaıs sıklıkla tercih edilen algoritmalardandır [8].

5.4.3. Kümeleme Yöntemi

Kümeleme yöntemi aynı sınıflandırma yönteminde olduğu gibi mevcut verileri gruplara ayırma işlemidir. Bu iki yöntemin aralarındaki fark ise sınıflandırma yöntemi ile yapılan sınıflandırma işleminde sınıflar önceden belirliyken kümeleme yönteminde ise sınıflar önceden belirli değildir. Verilerin hangi gruplara, kaç değişik gruba ayrılacağı verilerin birbirine olan benzerliği gere belirlenir [8]. Belirlenen bu gruplara da küme ismi verilir. Veriler sıralıysa yakın komşu, uzak komşu algoritmaları

kullanılırken belirli bir sıra olmayan verilerde K-ortalamalar yöntemi tercih edilir. Kümeleme yöntemi tıp, biyoloji, antropoloji, ekonomi, pazarlama ve telekomünikasyon gibi birçok sektör ve bilim dalında kullanılmaktadır.

6. ANALİZ UYGULAMASI VE YÖNTEM

Tezin bu bölümünde makine öğrenmesi teknikleri kullanılarak doğrudan satış sektör verisi üzerinden müşteri kayıp analizi çalışması anlatılacaktır. Gerçek veriler üzerinden yapılan bu müşteri kayıp analiz çalışması için CRISP modeli adımları izlenmiş/uygulanmıştır. Giriş bölümünde de anlatıldığı üzere 347.001 Temsilci verisi üzerinde çalışılmıştır.

6.1. Problemin Tanımlanması

Tartışmasız tüm sektörlerde müşteri kaybı en önemli problemlerin başında gelmektedir. Hatta en önemli problemdir. Yeni bir müşteri edinme maliyeti çok yüksek olduğu gibi mevcut müşterileri sınıflandırmadan müşteri kaybını önlemek için yapılan teşvik programları da oldukça yüksek maliyetlidir. Doğrudan satış sektöründe ise müşteri yani sektör deyimiyle Temsilci kayıpları bir kat daha önem kazanmaktadır. Bunu sebebi ise bir Temsilcinin birden fazla müşteriye aracılık etmesidir.

Verisini incelediğimiz şirket dört haftalık dönemler halinde kampanyalar oluşturmakta, bu kampanyalarda fiyatlar değiştirilmeden ağırlıkla katalog üzerinden Temsilcileri aracılığı ile satış yapmaktadır. Temsilcilerin performansına göre onları sistemden çıkarmakta veya sistemde tutmaya devam etmektedir. Bu performans ölçütleri de dönemlik olarak değişebilmektedir. Örneğin; altı kampanya sipariş vermeyen bir Temsilci sistemden çıkarılmakta ve sipariş girişi engellenmektedir. Bu durum Temsilcinin isteği ve iradesi dışında gerçekleşmektedir. Anlaşılacağı üzere sektörün dinamikleri oldukça farklıdır. Bu sebeple bu çalışmada Temsilcilerin bir sonraki kampanya sipariş verip vermeyecekleri tahmini yapılmıştır. Tahmin için sınıflandırma algoritmaları kullanılarak modeller oluşturulmuş ve sonuçlar karşılaştırılmıştır.

6.2. Veriyi Anlama

Analizlerde kullanılan veri setinde herhangi bir Temsilciye ait ad, soyadı, T.C. kimlik numarası, Temsilci numarası veya Temsilcinin kimliğini ortaya çıkaracak

herhangi bir bilgi mevcut değildir. Veri seti, anonimliğin bozulması engellenecek önlemler alınarak anonimleştirilmiş ve bu şekilde kullanılmıştır.

Tahmin için 24 bağımsız değişken 1 bağımlı değişken kullanılmıştır. Bağımsız değişkenlerden LOA, AGE, SL_LLOA nümeriktir, ancak 6.3.4. Veri Dönüştürme bölümünde detayları anlatıldığı şekilde kategorik hale dönüştürülmüştür. Bir başka ifadeyle bu sürekli değişkenler kesikli hale getirilmiştir. Veri setinde bulunan değişkenlere ait özellikler Tablo 1'de gösterilmiştir.

Tablo 1. Veri Setindeki Değişkenlere İlişkin Özellikler

TAHMİN İÇİN KULLANILAN BAĞIMSIZ DEĞİŞKENLER			
Sıra No	**Değişken Adı**	**Açıklama**	**Değişken Tipi**
1	LOA	Kampanya bazında Temsilcilik süresi	Nümerik
2	AGE	Yaş	Nümerik
3	AWARD_SALES	Sipariş tutarı	Nümerik
4	DEBT	Borç tutarı	Nümerik
5	PAYMENT_MODE	Ödeme Yöntemi	Kategorik
6	RETURNS	İade Tutarı	Nümerik
7	ORDER_COUNT	Sipariş sayısı	Nümerik
8	TOTAL_UNITS	Ürün adedi	Nümerik
9	INCENTIVE_BROCHURE_COUNT	Ücretli alınan katalog sayısı	Nümerik
10	OPPORTUNITY_GIFT	Kazanılan hediye	Nümerik
11	LOA_GIFT	İlk üç kampanya içerisinde kazanılan hediye	Kategorik
12	PREVIOUS_CAMPAIGN_AWARD_SALES	Bir önceki kampanya sipariş tutarı	Nümerik
13	SL_LLOA	Temsilcinin SL'nin kampanya bazında SL'liği Süresi	Nümerik
14	SL_GENERATION	Temsilcinin SL'nin BSS'na nesil uzaklığı	Nümerik
15	SL_G1_REPS	Temsilcinin SL'nin Temsilci Sayısı	Nümerik
16	SL_G1_ACTIVITY	Temsilcinin SL'nin Temsilcilerinin kampanya içerisinde sipariş verme oranı	Nümerik
17	SL_GEOGRAPHICAL_ZONE_FLAG	Temsilcinin SL ile aynı coğrafi bölgede olduğunu gösterir	Kategorik

Tablo 1. (devamı) Veri Setindeki Değişkenlere İlişkin Özellikler

18	SL_STATE_FLAG	Temsilcinin SL ile aynı ilde olduğunu gösterir	Kategorik
19	SL_COUNTY_FLAG	Temsilcinin SL ile aynı ilçede olduğunu gösterir	Kategorik
20	STARTING_FAMILY	CRM Segmenti	Kategorik
21	STATE	İl bilgisi	Kategorik
22	GENDER	Cinsiyet	Kategorik
23	SL_PAID_TITLE	Temsilcinin SL'nin seviye bilgisi	Kategorik
24	ACTIVIZATION	Son iki kampanya sipariş durumu	Kategorik
HEDEF DEĞİŞKEN (Bağımlı Değişken)			
1	CHURN_FLAG	Temsilci ayrılma durum bilgisi	Kategorik

Tablo 1.'de veri seti tanıtılmıştır. Açıklama kolonunda değişkenler hakkında genel bilgi verilmiştir. Bu bilgiler doğrultusunda SL_PAID_TITLE değişkeni hakkında daha detaylı bilgi verilmesi gereği görülmüştür. SL hakkında genel bilgi bölüm 2.4.2. Satış Lideri (SL)'de verilmişti. SL'nin seviyesi yönettiği Temsilci sayısına, ekibinin satış hacmine ve verilen hedefleri gerçekleştirme durumuna göre farklılık gösterir. Bu seviye ile SL'leri satış hacimleri üzerinden farklı oranlarda komisyon alırlar.

Şekil 7.'de Tablo 1.'de verilen değişkenler hakkında daha detaylı bilgi edinilebilir. Bu özet bilgi incelendiğinde nümerik değişkenleri için minimum, maksimum, medyan, ortalama, 1. ve 3. kartil (dörttebirlik) değerleri görülmektedir. Kategorik değerler için en çok örneğe sahip altı değer gösterilmektedir.

Temsilci grubunu tanımak için Şekil 7. incelendiğinde temsilcilerin büyük bir bölümünün kadınlardan oluştuğu görünmektedir. Temsilci grubunun ~%93,5'i kadın, ~%6,5'i erkektir. Temsilcilerin yaş ortalaması ~34.6'dır. En çok Temsilci İstanbul'da yaşarken sırasıyla diğer iller İzmir, Ankara, Bursa, Adana ve Hatay'dır. Temsilciler kampanyada 1.08 sipariş vererek dokuz ürün satın almaktadırlar. Temsilcilerin ~%71,7'si Satış Liderleri ile aynı ilde yaşarken aynı ilçede yaşama oranı ise ~%44,5'dir.

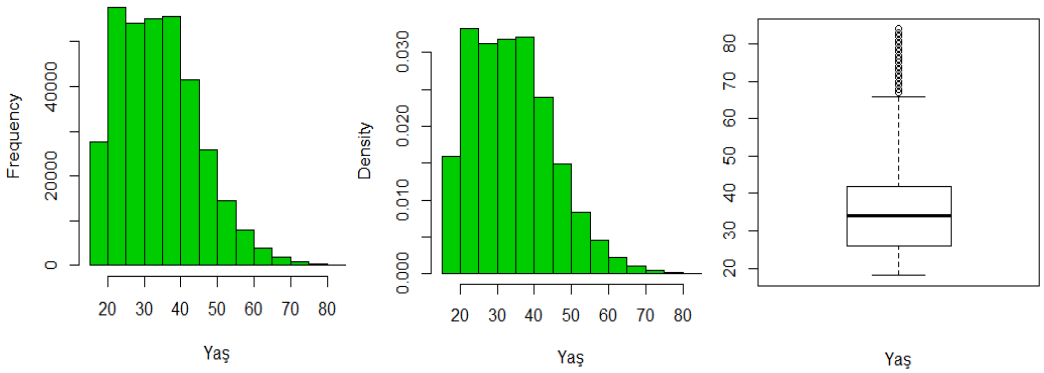
```
> summary(raw_data)
      LOA               AGE         AWARD_SALES              DEBT          PAYMENT_MODE
 Min.   :  1.00   Min.   :18.0   Min.   : -1253.0   Min.   :    0.00   HAV  :272187
 1st Qu.:  5.00   1st Qu.:26.0   1st Qu.:     0.0   1st Qu.:    0.00   KKHAV: 22191
 Median : 16.00   Median :34.0   Median :   121.5   Median :    0.00   KKOOS:  5175
 Mean   : 37.33   Mean   :34.6   Mean   :   220.8   Mean   :   35.03   OOS  : 47448
 3rd Qu.: 50.00   3rd Qu.:42.0   3rd Qu.:   248.2   3rd Qu.:    0.00
 Max.   :337.00   Max.   :84.0   Max.   :1663904.4  Max.   :64989.75

    RETURNS           ORDER_COUNT       TOTAL_UNITS       INCENTIVE_BROCHURE_COUNT
 Min.   :-43738.00   Min.   : -3.000   Min.   :  -73.00   Min.   :   0.000
 1st Qu.:     0.00   1st Qu.:  0.000   1st Qu.:    0.00   1st Qu.:   0.000
 Median :     0.00   Median :  1.000   Median :    9.00   Median :   0.000
 Mean   :    -3.68   Mean   :  1.078   Mean   :   15.53   Mean   :   1.651
 3rd Qu.:     0.00   3rd Qu.:  1.000   3rd Qu.:   19.00   3rd Qu.:   4.000
 Max.   :     0.00   Max.   :221.000   Max.   :74665.00   Max.   :2008.000

 OPPORTUNITY_GIFT LOA_GIFT   PREVIOUS_CAMPAIGN_AWARD_SALES    SL_LLOA
 Min.   :0.0000   E: 34049   Min.   : -1763.3              Min.   : 1.00
 1st Qu.:0.0000   H:312952   1st Qu.:     0.0              1st Qu.:26.00
 Median :0.0000              Median :   100.0              Median :73.00
 Mean   :0.2301              Mean   :   140.7              Mean   :52.36
 3rd Qu.:0.0000              3rd Qu.:   169.3              3rd Qu.:73.00
 Max.   :3.0000              Max.   :958813.3              Max.   :73.00

 SL_GENERATION      SL_G1_REPS      SL_G1_ACTIVITY   SL_GEOGRAPHICAL_ZONE_FLAG
 Min.   : 1.000   Min.   :  1.00   Min.   :0.0000   E:178718
 1st Qu.: 1.000   1st Qu.: 21.00   1st Qu.:0.6441   H:168283
 Median : 2.000   Median : 55.00   Median :0.7200
 Mean   : 2.279   Mean   : 82.14   Mean   :0.7108
 3rd Qu.: 3.000   3rd Qu.:115.00   3rd Qu.:0.7755
 Max.   :10.000   Max.   :681.00   Max.   :1.0000

 SL_STATE_FLAG SL_COUNTY_FLAG       STARTING_FAMILY        STATE         GENDER
 E:245186      E:154544       A.Classic      :106661   İSTANBUL: 56554   F:324421
 H:101815      H:192457       A.Classic Plus: 89885   İZMİR   : 29635   M: 22580
                              A.Club         : 66820   ANKARA  : 23321
                              A.Elite        :  8420   BURSA   : 14686
                              A.Inci         : 75215   ADANA   : 13109
                                                       HATAY   : 11038
                                                       (Other) :198658

 SL_PAID_TITLE   ACTIVIZATION CHURN_FLAG
 L03    :62236   A :  29256   E:174899
 L04    :58035   AA:155348    H:172102
 L02    :47955   AI: 57219
 L05    :30526   I :     8
 L09    :27637   IA: 62004
 L06    :25974   II: 43166
 (Other):94638
```

Şekil 7. Veri Önişleme Öncesi Veri Seti Özet Bilgisi

```
> str(raw_data)
'data.frame':   347001 obs. of  25 variables:
 $ LOA                           : int  251 124 247 4 336 155 154 155 14 337 ...
 $ AGE                           : int  57 32 64 40 56 54 45 42 32 80 ...
 $ AWARD_SALES                   : num  0 124.4 296.2 340.5 -23.2 ...
 $ DEBT                          : num  0 0 0 0 0 ...
 $ PAYMENT_MODE                  : Factor w/ 4 levels "HAV","KKHAV",..: 3 4 2 2 4 3 4 4 2 3 ...
 $ RETURNS                       : num  0 0 -76 0 -23.2 ...
 $ ORDER_COUNT                   : int  0 1 4 2 0 1 2 0 0 5 ...
 $ TOTAL_UNITS                   : int  0 9 25 33 -1 22 18 0 0 60 ...
 $ INCENTIVE_BROCHURE_COUNT      : int  0 0 1 0 0 1 0 0 0 12 ...
 $ OPPORTUNITY_GIFT              : int  0 0 0 0 0 0 0 0 0 2 ...
 $ LOA_GIFT                      : Factor w/ 2 levels "E","H": 2 2 2 2 2 2 2 2 2 2 ...
 $ PREVIOUS_CAMPAIGN_AWARD_SALES: num  104.7 100 41.5 151.2 114.2 ...
 $ SL_LLOA                       : int  73 73 55 73 73 73 73 73 18 73 ...
 $ SL_GENERATION                 : int  1 1 3 2 1 1 2 1 3 1 ...
 $ SL_G1_REPS                    : int  98 173 59 57 98 74 97 225 28 180 ...
 $ SL_G1_ACTIVITY                : num  0.755 0.711 0.678 0.667 0.755 ...
 $ SL_GEOGRAPHICAL_ZONE_FLAG     : Factor w/ 2 levels "E","H": 1 1 2 2 2 1 1 1 1 2 ...
 $ SL_STATE_FLAG                 : Factor w/ 2 levels "E","H": 1 1 1 2 1 1 1 2 1 1 ...
 $ SL_COUNTY_FLAG                : Factor w/ 2 levels "E","H": 2 1 2 2 2 1 2 2 1 2 ...
 $ STARTING_FAMILY               : Factor w/ 5 levels "A.Classic","A.Classic Plus",..: 1 2 2 3 1 2 3 3 1 4 ...
 $ STATE                         : Factor w/ 84 levels "\002","0","ADANA",..: 43 44 43 15 43 40 40 21 43 43 ...
 $ GENDER                        : Factor w/ 2 levels "F","M": 1 1 1 1 1 1 1 1 1 1 ...
 $ SL_PAID_TITLE                 : Factor w/ 14 levels "0_REP","2_CAN",..: 9 11 6 6 9 10 9 11 4 11 ...
 $ ACTIVIZATION                  : Factor w/ 6 levels "A","AA","AI",..: 3 2 2 2 3 5 2 3 3 2 ...
 $ CHURN_FLAG                    : Factor w/ 2 levels "E","H": 1 2 2 1 1 2 2 1 1 2 ...
```

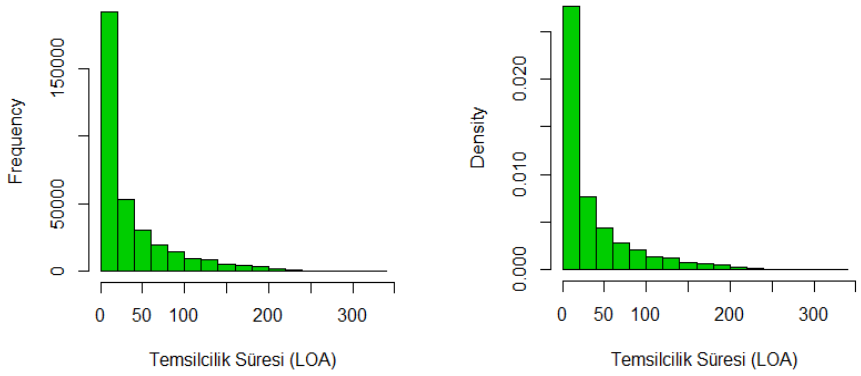
Şekil 8. Veri Önişleme Öncesi Veri Seti Değişkenlerinin Gösterim Biçimleri ve Türleri

Veri setindeki sayısal değişkenler özet veri üzerinden incelenmişti. Bu değişkenlerin bir kısmını grafikler yardımıyla aşağıda görselleştirilmiştir. Görselleştirmede histogram ve boxplot (kutu) grafikleri kullanılmıştır.



Şekil 9. Yaş Değişkeni Histogram ve Boxplot (kutu) Grafikleri

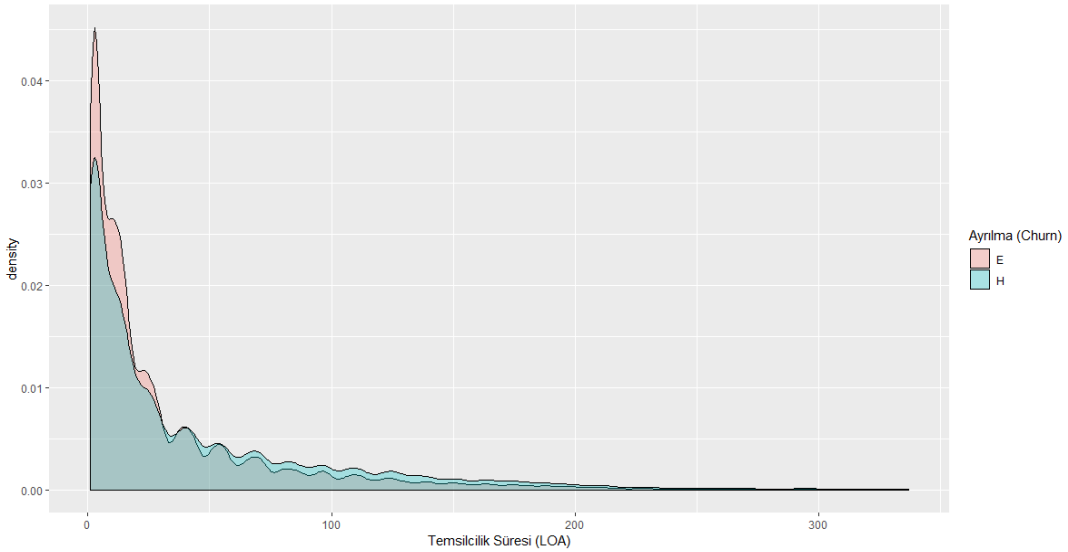
Şekil 9'da ki histogram grafiklerinde belirli yaş gruplarında kaç Temsilci olduğu (frekans (frequency)'a göre) yine belirli yaş gruplarında Temsilcilerin yüzdelik dağılımları (yoğunluk (density)'a göre) görünmektedir.



Şekil 10. Temsilcilik Süresi (LOA) Değişkeni Histogram Grafikleri

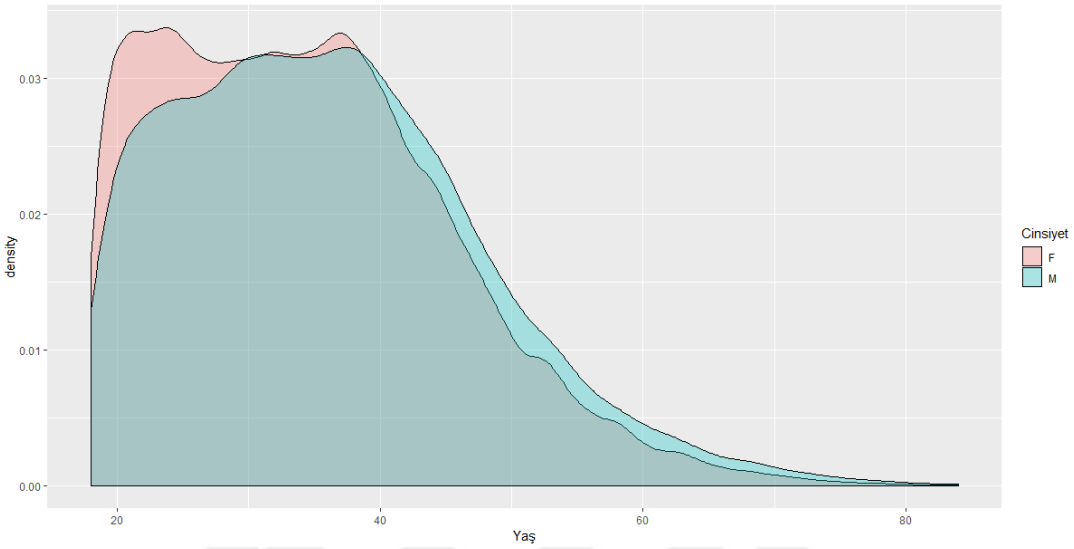
Şekil 10.'da ki histogram grafiklerinde belirli Temsilcilik süresi (LOA) gruplarında kaç Temsilci olduğu (frekans (frequency)'a göre) yine belirli Temsilcilik süresi (LOA) gruplarında Temsilcilerin yüzdelik dağılımları (yoğunluk (density)'a göre) görünmektedir.

Veri setinin anlaşılması için nümerik (sayısal) değişkenlerin görselleştirilip incelendiği gibi kategorik değişkenler de bu şekilde incelenmiştir. Şekil 11.'de Temsilcilik süresi (LOA) ile ayrılma (churn) değişkenlerinin yoğunluk grafiği görünmektedir. Grafikte görüleceği üzere Temsilcilik süresi (LOA) arttıkça ayrılma (churn) durumu azalmaktadır.



Şekil 11. Temsilcilik Süresi (LOA) ve Ayrılma (Churn) yoğunluk Grafiği

Şekil 12.'de Temsilci cinsiyet ve yaş yoğunluk grafiği görünmektedir. Kadın Temsilcilerin 18-30 yaş aralığında ki yoğunluğu kendi gruplarında daha yüksekken bu durum erkek Temsilcilerde 28-38 yaş grubundadır.



Şekil 12. Temsilci Cinsiyet ve Yaş Yoğunluk Grafiği

Değişkenler ile hedef değişkenin ya da değişkenlerin birbiri ile ilişkisi grafikler ile gösterilebilir. Ancak bu yöntem oldukça uzun zaman alabileceği gibi değişkenlerin birbirleri ile olan ilişkilerini gözden kaçırmaya sebebiyet verebilir. Bu nedenle çalışmada değişkenlerin birbiri ile olan korelasyonları incelenmiştir. Korelasyon hesabı için bazı değişkenler kategorik veriden nümerik veriye dönüştürülmüşlerdir. Örneğin; CHURN_FALG değişkeni 'E' ve 'H' değerlerini alırken veri tipi değiştirilerek 1 ve 0 değerleri ile güncellenmiştir.

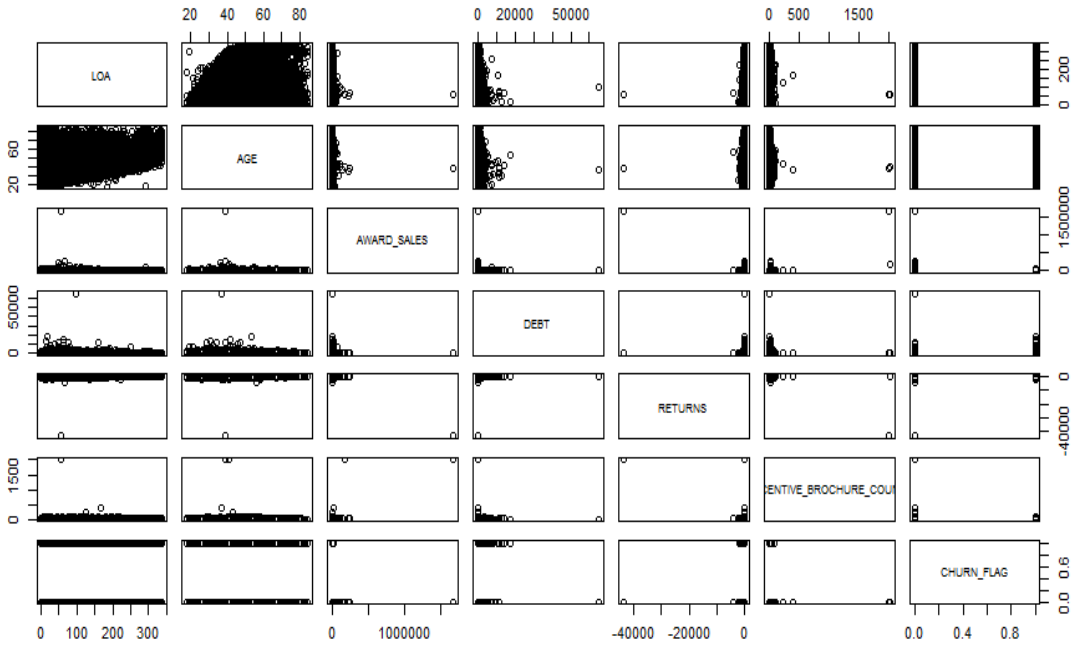
Şekil 13.'te görüleceği üzere Hedef değişken (CHURN_FLAG) ile Temsilcilik süresi (LOA), yaş (AGE), sipariş tutarı (AWARD_SALES), sipariş adedi (ORDER_COUNT), ücretli alınan katalog (INCENTIVE_BROCHURE_COUNT), Kazanılan fırsat hediyesi (OPPORTUNITY_GIFT), yeni Temsilcilik hediyesi (LOA_GIFT), SL temsilci sayısı (SL_G1_REPS), SL ile aynı coğrafi bölgede olma (SL_GEOGRAPHICAL_ZONE_FLAG), SL ile aynı ilçede olma (SL_COUNTY_FLAG) değişkenleri arasında negatif korelasyon görülmektedir.

Hedef değişken (CHURN_FLAG) ile borç tutarı (DEBT), iade tutarı (RETURNS), ve Temsilcinin SL'nin BSS'na nesil uzaklığı (SL_GENERATION) arasında pozitif korelasyon görünmektedir.

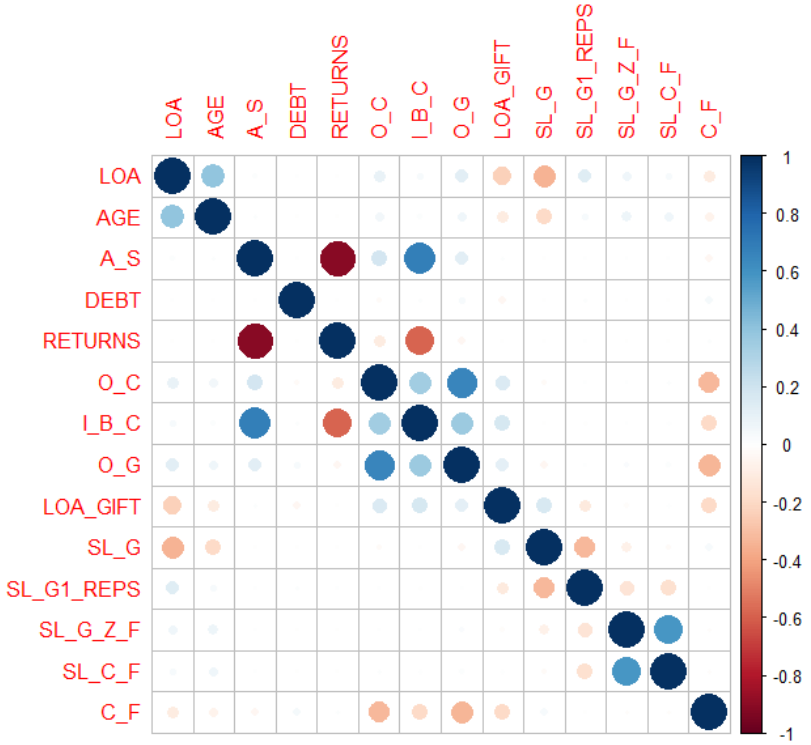
	LOA	AGE	AWARD_SALES	DEBT	RETURNS	ORDER_COUNT	INCENTIVE_BROCHURE_COUNT	OPPORTUNITY_GIFT	LOA_GIFT	SL_GENERATION	SL_G1_REPS	SL_GEOGRAPHICAL_ZONE_FLAG	SL_COUNTY_FLAG	CHURN_FLAG
LOA	1	0.39562182	0.01843613	0.00664331	-0.00303046	0.09436041	0.03705555	0.12170490	-0.23714059	-0.34486745	0.13358310	0.06425001	0.03348722	-0.10747821
AGE	0.39562182	1	0.01064719	-0.00065851	-0.00291517	0.05444131	0.01851170	0.06649597	-0.10887268	-0.19319303	0.03470501	0.07270331	0.06863899	-0.06633389
AWARD_SALES	0.01843613	0.01064719	1	0.00747561	-0.90090248	0.18949238	0.68897988	0.12182910	0.01356769	-0.00598940	0.00086126	0.00436995	0.00509422	-0.04291034
DEBT	0.00664331	-0.00065851	0.00747561	1	-0.00382521	-0.02102171	0.01036421	0.03684750	-0.04146495	-0.00393393	0.01442698	-0.00165976	-0.00520317	0.04135362
RETURNS	-0.00303046	-0.00291517	-0.90090248	-0.00382521	1	-0.10825412	-0.58668268	-0.04133381	-0.00087788	-0.00196689	0.00130842	-0.00154611	-0.00229256	0.01042153
ORDER_COUNT	0.09436041	0.05444131	0.18949238	-0.02102171	-0.10825412	1	0.34422609	0.65483040	0.15866518	-0.02310533	0.00448748	0.00608765	0.01103867	-0.32563935
INCENTIVE_BROCHURE_COUNT	0.03705555	0.01851170	0.68897988	0.01036421	-0.58668268	0.34422609	1	0.36574670	0.17141476	-0.00151791	-0.00882051	0.00757537	0.00910190	-0.19276278
OPPORTUNITY_GIFT	0.12170492	0.06649597	0.12182914	0.03684750	-0.04133381	0.65483040	0.36574667	1	0.11197028	-0.04195501	0.00007364	0.02044673	0.01960482	-0.33900743
LOA_GIFT	-0.23714059	-0.10887268	0.01356769	-0.04146495	-0.00087788	0.15866518	0.17141476	0.11197030	1	0.16834091	-0.11014360	-0.01852030	0.00954270	-0.19301348
SL_GENERATION	-0.34486745	-0.19319303	-0.00598940	-0.00393393	-0.00196688	-0.02310533	-0.00151791	-0.04195501	0.16834091	1	-0.32695060	-0.07373848	-0.02468857	0.03767316
SL_G1_REPS	0.13358307	0.03470501	0.00086126	0.01442698	0.00130842	0.00448748	-0.00882051	0.00007364	-0.11014365	-0.32695059	1	-0.14218971	-0.16406491	-0.00712152
SL_GEOGRAPHICAL_ZONE_FLAG	0.06425001	0.07270331	0.00436995	-0.00165976	-0.00154611	0.00608765	0.00757537	0.02044673	-0.01852030	-0.07373848	-0.14218970	1	0.58923492	-0.01123628
SL_COUNTY_FLAG	0.03348722	0.06863899	0.00509421	-0.00520317	-0.00229256	0.01103867	0.00910190	0.01960481	0.00954270	-0.02468857	-0.16406490	0.58923492	1	-0.01245360
CHURN_FLAG	-0.10747821	-0.06633389	-0.04291034	0.04135362	0.01042153	-0.32563935	-0.19276278	-0.33900740	-0.19301348	0.03767316	-0.00712152	-0.01123628	-0.01245360	1

Şekil 13. Veri Setindeki Sayısal Değişkenler Arası Korelasyon Değerleri

Şekil 14.'te nümerik değişkenlerin bir kısmına ait kutu grafiği ve Şekil 15.'te korelasyonun şekilsel gösterimi bulunmaktadır.



Şekil 14. Nümerik Değişkenlere Ait Kutu Grafiği



Şekil 15. Korelasyonun Şekilsel Gösterimi

6.3. Veriyi Hazırlama

Veri seti yapılacak olan analiz çalışmasına hazır hale getirilmesi için sırasıyla aşağıdaki adımlar uygulanmıştır.

6.3.1. Veri Setindeki Eksik, Kayıp Değerlerin Tespiti

Veri seti R programına aktarılmadan önce, Oracle SQL'de hazırlanması aşamasında 3 Temsilcinin yaş bilgisinin eksik 2 Temsilcinin de segment bilgisinin eksik olduğu görülmüştür. Nümerik değer olan Temsilci yaş bilgisi için hedef değişkendeki sınıfa göre ortalama alınarak, kategorik değer olan segment bilgisi için de en çok tekrar eden kategori ile tamamlanması düşünülmüştür. Ancak eksik değerlerin veri seti içindeki payı çok küçük olduğu için bu Temsilciler veri setinden çıkarılmıştır.

6.3.2. Aykırı Verilerin (Outliers) Tespiti

Veri setinde 7 Temsilcinin doğum tarihinin 01.01.1900 olduğu ve bu sebeple yaşlarının 118 geldiği görülmüş. Bu Temsilcilerin 6'sının doğum tarihinin güncel müşteri tablosunda düzeltildiği görülmüş ve veri setindeki yaş bilgisi de düzeltilmiştir. Güncel müşteri tablosundan diğer Temsilcilerin de doğum tarihleri kontrol edilmiş ancak uyumsuz başka kayıt görülmemiştir. 1 Temsilcinin ise doğum tarihine ulaşılamadığı için veri setinden çıkarılmıştır.

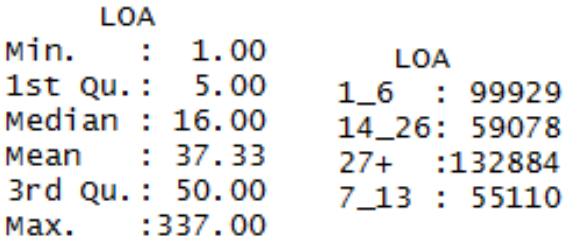
6.3.3. Tekrar Eden Kayıtların Tespiti

Veri setindeki 3 Temsilcinin tüm değişkenleri aynı ikişer kayıtları olduğu görülmüştür. Bu kayıtlar Oracle SQL de distinct() fonksiyonu ile veri setinden çıkarılmıştır.

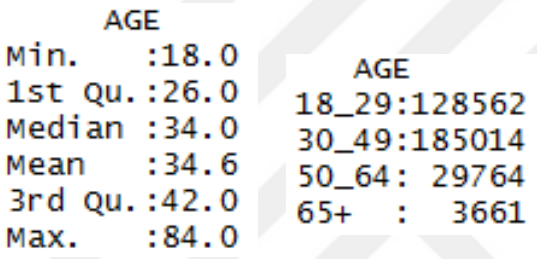
6.3.4. Veri Dönüştürme

Bu çalışmada veri dönüştürme işlemi Temsilcilik süresi (LOA), yaş (AGE) ve Temsilcinin SL'nin SL'liği süresi (SL_LLOA) için yapılmıştır. Bu değişkenler sürekli veri tipinden kesikli veri tipine dönüştürülmüşleridir. Diğer bir ifadeyle nümerik veri tipinden kategorik veri tipine dönüştürülmüşlerdir. Şekil 16.'da Temsilcilik süresi (LOA) değişkeninin sürekli değişken hali ile kesikli değişken hali, Şekil 17.'da yaş

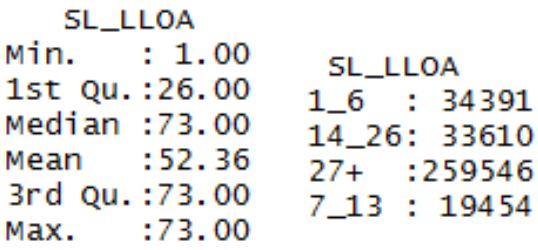
(AGE) değişkeninin sürekli değişken hali ile kesikli değişken hali ve Şekil 18.'de de Temsilcinin SL'nin SL'liği süresi (SL_LLOA) değişkeninin sürekli değişken hali ile kesikli değişken hali görülmektedir.



Şekil 16. Temsilcilik Süresi (LOA) Değişkeninin Sürekli Değişken Hali ile Kesikli Değişken Hali



Şekil 17. Yaş (AGE) Değişkeninin Sürekli Değişken Hali ile Kesikli Değişken Hali



Şekil 18. Temsilcinin SL'nin SL'liği Süresi (SL_LLOA) Değişkeninin Sürekli Değişken Hali ile Kesikli Değişken Hali

6.3.5. Normalizasyon

Çalışmada kullanılan veri setinde nümerik değişkenlerin değişim aralıkları yüksek olduğu için nümerik değişkenler normalize edilmiştir. Bu işlem için min-max, z-score ve ondalık ölçekleme gibi birçok normalizasyon yönteminden min-max seçilerek normalizasyon işlemi yapılmıştır. Min-max normalizasyon yöntemi denklemi Denklem 6.1'de gösterilmiştir.

$$x' = \frac{x_i - x_{min}}{x_{max} - x_{min}} \quad (6.1)$$

Denklem 2.1.'de ki eşitlikte;

$X^{'}$: Normalize edilmiş veriyi

X^{i} : Girdiğ değerini,

X_{min} : Girdi seti içerisinde yer alan en küçük sayısı,

X_{max} : Girdi seti içerisinde yer alan en büyük sayısı,

ifade etmektedir.

Yapılan normalizasyon çalışması sonrasında model tekrar çalıştırılmış (C4.5 karar ağacı algoritması Hold-out yöntemi ile veri seti sırasıyla eğitim ve test veri seti olmak üzere %90-%10 oranlarıyla ayrılmıştır) ve doğru tahmin oranın ~%0,36 puan daha kötü sonuç vermesi nedeniyle normalize edilmiş veri modellerde kullanılmamıştır.

6.4. Modelleme

Modelleme kullanılacak veri seti hakkında tam bilgi sağlandıktan ve veri ön işleme adımları tamamlandık sonra yapılmalıdır. Bu bölümde, analiz çalışmasında kullanılan karar ağaçları (C4.5 karar ağacı ve Gini karar ağacı) hakkında teorik bilgi verilecek ve model performans karşılaştırma yöntemi anlatılacaktır.

6.4.1. Karar Ağaçları

Karar ağaçları gözetimli öğrenme yöntemidir ve bir hedef değişkene ihtiyaç duyar. Karar ağacını oluşturan algoritma hedef değişkeni en iyi tahmin edecek kurallar silsilesini oluşturmaya çalışır. Karar ağaçları sınıflandırma ve tahmin için oldukça sık kullanılan bir veri madenciliği yaklaşımıdır. Sinir ağları gibi diğer yöntembilimler de sınıflandırma için kullanılırken, karar ağaçları kolay yorumlanabilmesi ve anlaşılabilirliği sebebiyle tercih edilmektedir [10] [11].

Karar ağacı tekniğinde kullanılan verinin sınıflanması, öğrenme ve sınıflama olmak üzere iki basamaklı bir işlemdir. Öğrenme basamağında hedef değişkeni önceden bilinen eğitim verisi bir model oluşturmak için algoritma tarafından analiz

edilir. Bu işlemden sonra öğrenilen model kara ağacı olarak gösterilir. Sınıflama basamağında, test verisi sınıflama kurallarının doğruluğunu ölçmek amacıyla kullanılır. Eğitim verisindeki hangi alanların hangi sıra ile kullanılacağı entropi ölçümüyle belirlenir. Entropi ölçüsü ne kadar fazla ise o alan kullanılarak ortaya konulan sonuçlar o oranda belirsiz ve kararsızdır. Bu sebeple karar ağacının kökünde entropi ölçüsü en az olan alanlar kullanılır. Entropi ölçüsü aşağıdaki formüller ile hesaplanabilir.

$$E(C|A_k) = \sum_{J=1}^{M_k} p(a_k, j) x \left[- \sum_{i=1}^{N} p(c_i, |a_k, j) log_2 P(c_i|a_k, j) \right] \quad (6.2)$$

Denklem 6.2.'de ki eşitlikte;

$E(C|A_k) = A_K$ Alanının sınıflama özelliğinin entropi ölçüsü,

$p(a_k, j) = a_k$ Alanının j değerinde olma olasılığı,

$P(c_i|a_k, j)$=a_k alanı j. Değerindeyken sınıf değerinin c_i olma olasılığı

M_k = a_k alanının içerdiği değerlerin sayısı; j=1,2,...,M_k,

N = Farklı sınıfların sayısı; i = 1, 2,…, N,

K= alanların sayısı; k = 1, 2, …, K.

Eğer bir S kümesindeki elemanlar, kategorik olarak C1, C2, C3, …., C_i sınıflarına ayrılmışlarsa, S kümesindeki bir elemanın sınıfını belirlemek için gereken bilgi aşağıdaki formül ile hesaplanabilir.

$$I(S) = -(P_1 log_2(P_1) + (P_2 log_2(P_2) + \cdots + (P_i log_2(P_i)) \quad (6.3)$$

Bu formülde P_i, C_i sınıfına ayrılma olasılığıdır.

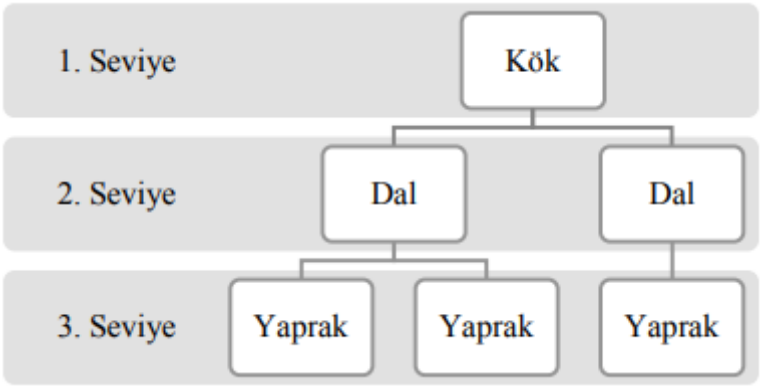
Entrolpi denklemi şu şekilde de ifade edilebilir:

$$E(A) = \sum_{i=1}^{n} \frac{|S_i|}{|S|} x l(S_i) \quad (6.4)$$

Bu durumda A alanı kullanılarak yapılacak dallanma işleminde bilgi kazancı denklem 6.5 ile hesaplanmaktadır.

Kazanç (A) = I (S) – E (A) (6.5)

Başka bir ifadeyle Kazanç (A), A alanının değerini bilmekten kaynaklanan entropideki azalmadır [11]. Şekil 19.'da örnek bir karar ağacı yapısı gösterilmektedir.



Şekil 19. Bir Karar Ağacının Yapısı

Kaynak: Daş, B., Türkoğlu, İ., **DNA Dizilimlerinin Sınıflandırılmasında Karar Ağacı Algoritmalarının Karşılaştırılması,** Eleco 2014 Elektrik – Elektronik – Bilgisayar ve Biyomedikal Mühendisliği Sempozyumu, 2014

6.4.1.1. C4.5 Karar Ağacı

En yaygın kullanılan ID3 algoritmasından geliştirilmiştir. ID3 algoritmasına üstünlük sağladığı alanlar işe kısaca şu şekilde özetlenebilir. ID3 algoritmasında karar ağacı oluşturulurken kayıp veriler hesaba katılmaz. Kazanım oranı hesaplanırken sadece verileri eksik olmayan diğer kayıtlar kullanılır. C4.5 algoritması ise kayıp verileri diğer veri ve değişkenler yardımıyla öngörerek kazanım oranının hesaplamasında kullanır. Bu sayede daha duyarlı ve anlamlı kurallar çıkartabilen bir ağaç üretebilir [8].

C4.5 karar ağacı her düğümden çıkan çoklu dallar ile karar ağacını oluşturur. Dal sayısı oluşturulmak istenen sınıf sayısına eşittir. Tek bir sınıflayıcıda birden çok karar ağacı birleştirir. Ayırma işlemi bilgi kazancı esasına göre yapılır [11] [12].

6.4.1.2. Gini Karar Ağacı

Gini karar ağacı CART algoritmasından geliştirilmiştir. Bu sebeple CART algoritmasının genetik özellikleri taşır. İkili bölümler şeklinde gerçekleşen bir sınıflandırma yöntemidir. Algoritma değişken değerlerinin sol ve sağda olmak üzere ikili bölünmeler şeklinde ayrılması temeline dayanır. Her bir düğümde ilgili sol ve sağ bölünmeleri için ayı hesaplamalar yapılır [8] [11] [12].

6.4.2. Model Değerlendirme ve Performans Ölçütleri

Bu çalışmada oluşturulan modellerin performansı ölçümlemek için dışarıda tutma yöntemi (Hold-Out) kullanılmıştır. Bu yöntem de veri seti eğitim ve test kümesi olmak üzere ikiye ayrılmaktadır. Eğitim veri seti eğitim sürecinde model oluşturulmak için seçilen sınıflandırıcının eğitiminde kullanılmaktadır. Modelin parametreleri için en iyi değerler ve uygun performans ölçüsü bu adımda belirlenir. Veri setinin ikinci parçası olan test veri seti geliştirilen modelin performansını ölçümlemede kullanılır [13].

Çalışmada Hold-out yöntemi ile veri seti sırasıyla eğitim ve test veri seti olmak üzere %90-%10, %80-%20, %70-%30, %60-%40 oranlarıyla ayrılmıştır. Bu işkem sonrasında test ve eğitim veri seti üzerinde sınıflandırma algoritmaları çalıştırılmış. Modeller uygulandıktan sonra hangi modelin daha iyi sonuç ürettiği kontenjans tablosu yöntemiyle ölçülmüştür.

Tablo 2. Kontenjans tablosu

		Gerçek		
		Pozitif	**Negatif**	**Toplam**
Tahmin	**Pozitif**	Doğru Pozitif (dp)	Yanlış Pozitif (yp)	tPoz
	Negatif	Yanlış Negatif (yn)	Doğru Negatif (dn)	tNeg
	Toplam	poz	Neg	m

Tablo 2'ye göre sınıflandırma algoritmaları ile elde edilen modelin doğruluk ve hata oranı denklemleri aşağıda gösterilmiştir [13].

$$Doğruluk\ (ACC) = \frac{dp + dn}{m} \tag{6.6}$$

$$Hata\ Oranı\ (ERR) = 1 - ACC \tag{6.7}$$

Sınıflandırma algoritmasının pozitif örnekleri tahmin etmedeki etkinliğine duyarlılık denmektedir. Bir başka ifadeyle duyarlılık, doğru tahmin edilen (sınıflandırılan) pozitif örneklerin toplam pozitif örnek sayısına oranıdır.

$$Duyarlılık\ (TPR) = \frac{dp}{poz} + \frac{dp}{dp + yn} \tag{6.8}$$

Sınıflandırma algoritmasının negatif örnekleri tahmin etmedeki etkinliğine belirleyicilik denmektedir. Bir başka ifadeyle belirleyicilik, doğru tahmin edilen (sınıflandırılan) negatif örneklerin toplam negatif örnek sayısına oranıdır.

$$Belirleyicilik\ (SPC) = \frac{dn}{neg} = \frac{dn}{dn + yp} \tag{6.9}$$

Aslında negatif olan ancak pozitif olarak tahmin edilen örneklerin, tüm negatif örneklere oranına yanlış pozitif oranı denir. Aslında pozitif olan ancak negatif olarak tahmin edilen örneklerin, tüm pozitif örneklere oranına ise yanlış negatif oranı denir.

$$Yanlış\ Pozitif\ Oranı\ (FPR) = 1 - SCP = \frac{yp}{neg} = \frac{yp}{yp + dn} \tag{6.10}$$

$$Yanlış\ Negatif\ Oranı\ (FNR) = 1 - TPR = \frac{yn}{poz} = \frac{yn}{yn + dp} \tag{6.11}$$

Doğru sınıflandırılan pozitif örneklerin toplam pozitif tahmin edilen örneklere oranına pozitif öngörü değeri denir. Doğru sınıflandırılan negatif örneklerin toplan negatif tahmin edilen örneklere oranına ise negatif öngörü değeri denir.

$$Pozitif\ Öngörü\ Değeri\ (PPV) = \frac{dp}{tPoz} = \frac{dp}{dp + yp} \tag{6.12}$$

$$Negatif\ Öngörü\ Değeri\ (NPV) = \frac{dn}{tNeg} = \frac{dn}{dn + yn} \tag{6.13}$$

F-ölçüsü (F-measure), kesinlik ve duyarlılık performans değerlendirme ölçülerinin harmonik ortalamasıdır.

$$F - \text{Ölçüsü}\ (F) = \frac{2 * PPV * TPR}{PPV + TPR} \tag{6.13}$$

Tahmin sonucunun aslında pozitif olan örneklerin pozitif çıkma olasılığının, negatif örneklerin pozitif çıkma olasılığına oranına pozitif olabilirlik oranı denir. Tahmin sonucunun aslında pozitif olan örneklerin negatif çıkma olasılığının, negatif örneklerin negatif çıkma olasılığı oranına ise negatif olabilirlik oranı denmektedir.

$$Pozitif\ Olabilirlik\ Oranı\ (LR+) = \frac{TPR}{FPR} = \frac{TPR}{1 - SCP} \tag{6.14}$$

$$Negatif\ Olabilirlik\ Oranı\ (LR-) = \frac{FNR}{TNR} = \frac{1 - TPR}{SCP} \tag{6.15}$$

Tahmin edilen pozitif sınıfın üstünlüğünün, negatif sınıfın üstünlüğüne oranına Tanısal üstünlük oranı denir.

$$Tanısal\ \text{Üstünlük}\ Oranı\ (DOR) = \frac{LR+}{LR-} \tag{6.16}$$

7. BULGULAR

Bu bölümde C4.5 ve Gini karar ile oluşturulan modellerin veri seti üzerinde uygulanması ile elde edilen sonuçlar ve bu analizler sonucunda elde edilen bulgulara yer verilmiştir. İki algoritma ile oluşturulan sekiz model bölüm 6.4.2. Model Değerlendirme ve Performans Ölçütleri' de detaylı şekilde anlatılan yöntemler ile karşılaştırılacaktır.

7.1. Gini Karar Ağacı ile Oluşturulan Model

Gini karar ağacı algoritması ile oluşturulan modelin hedef değişkeni, performans değerlendirme yöntemi ve kullanılan R kütüphaneleri Tablo 3.'te gösterilmiştir.

Tablo 3. Gini Karar Ağacı Algoritması Analiz Özeti

Hedef Değişken	CHURN_FLAG (E/H)
Performans Değerlendirme yöntemi	Sırasıyla eğitim ve test veri seti olmak üzere %90-%10, %80-%20, %70-%30, %60-%40 oranlarıyla Hold-out
Kullanılan R kütüphaneleri	• CSV dosyadan veri okuma • Caret: Veri setini, eğitim ve test olarak ayırabilmek için kullanılmıştır. Bu paketi aktif etmek için lattice ve ggplot2 paketleri de yüklenmelidir. • Rpart: Gini karar ağacı algoritmasını çalıştırır. • Rpart.plot: Gini karar ağacı algoritması ile oluşturulan karar ağacının grafiğini çizmek için kullanılır.

Modelin tahmin sonuçları ve C4.5 karar ağacı ile oluşturulan modelin sonuçları ile karşılaştırılması 7.3. model performans karşılaştırılması bölümünde anlatılmıştır. Modele ait kodlar EK-1'de gösterilmiştir.

7.2. C4.5 Karar Ağacı ile Oluşturulan Model

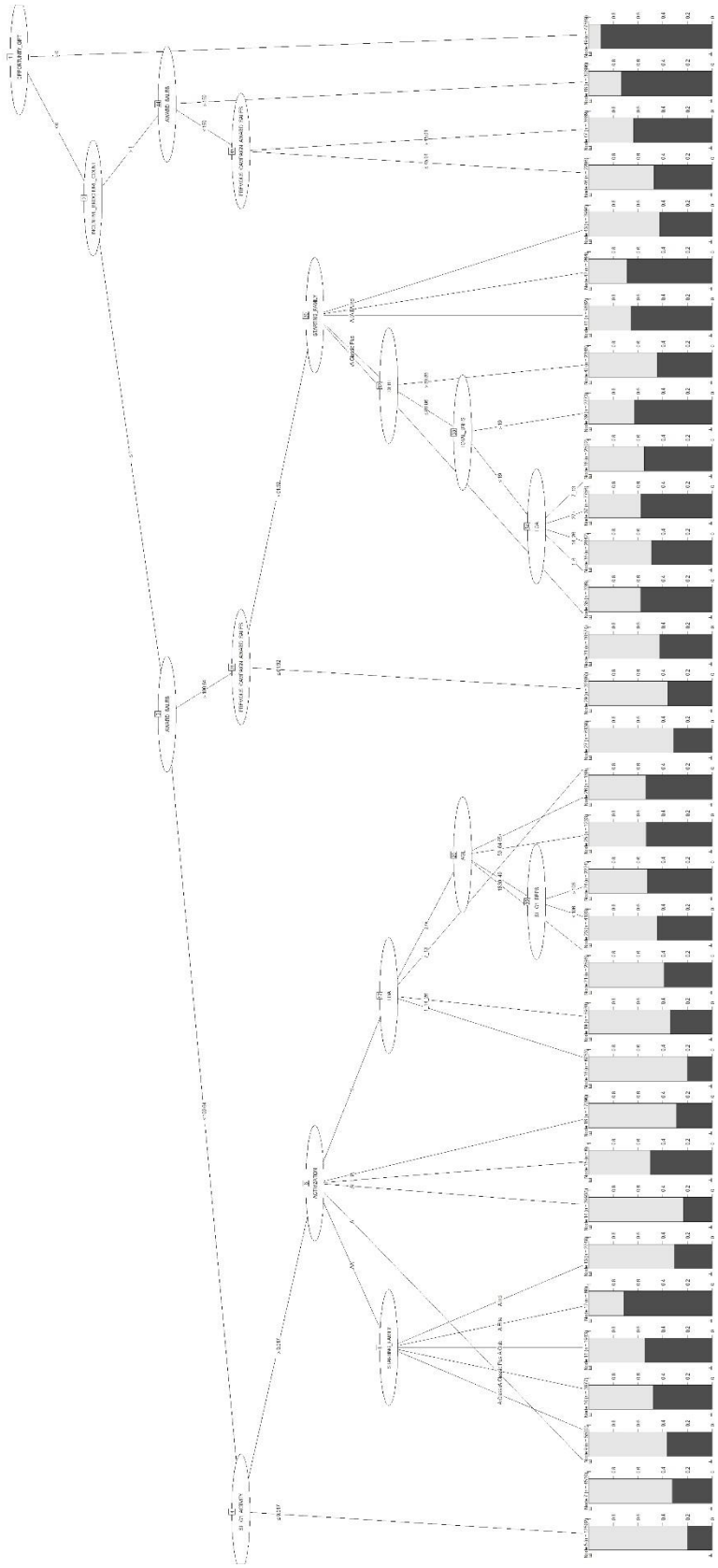
C4.5 karar ağacı algoritması ile oluşturulan modelin hedef değişkeni, performans değerlendirme yöntemi ve kullanılan R kütüphaneleri Tablo 3.'te gösterilmiştir.

Tablo 4. C4.5 Karar Ağacı Algoritması Analiz Özeti

Hedef Değişken	CHURN_FLAG (E/H)
Performans Değerlendirme yöntemi	Sırasıyla eğitim ve test veri seti olmak üzere %90-%10, %80-%20, %70-%30, %60-%40 oranlarıyla Hold-out
Kullanılan R kütüphaneleri	• CSV dosyadan veri okuma • Caret: Veri setini, eğitim ve test olarak ayırabilmek için kullanılmıştır. Bu paketi aktif etmek için lattice ve ggplot2 paketleri de yüklenmelidir. • Rweka: Gini karar ağacı algoritmasını çalıştırır. (Modelin çalıştırıldığı bilgisayarda java kurulmuş olmalıdır) • Rpart.plot: Gini karar ağacı algoritması ile oluşturulan karar ağacının grafiğini çizmek için kullanılır.

Modelin tahmin sonuçları ve Gini karar ağacı ile oluşturulan modelin sonuçları ile karşılaştırılması 7.3. model performans karşılaştırılması bölümünde anlatılmıştır. Modele ait kodlar EK-2'de gösterilmiştir.

Gini karar ağacı ve C4.5 karar ağacı algoritmaları ile oluşturulan modellerin veri seti üzerindeki performansları Tablo 5. ve Tablo 6.'da gösterilmiştir. Hold-out performans değerlendirme yöntemi ile veri seti sırasıyla eğitim ve test veri seti olmak üzere %90-%10, %80-%20, %70-%30, %60-%40 oranlarıyla ayrılmıştır. Yapılan performans ölçümünde görüldüğü üzere C4.5 karar ağacı algoritması Gini karar ağacı algoritmasına göre daha iyi sonuç vermiştir. C4.5 karar ağacı en iyi sonucu %80-%20 eğitim, test verisi hold-out ayrımında vermiştir. Örnek olarak elde edilen karar ağacı Şekil 20.'de gösterilmiştir.



Şekil 20. C4.5 Karar Ağacı Görüntüsü (%80-%20 eğitim, test verisi hold-out)

Şekil 20'de ki karar ağacı görüntüsü incelenerek kurallar silsilesinin çıkarılması oldukça zordur. Şekil 21.'de ağacın yazdırılmış kurallarının örnek ekran görüntüsü görünmektedir. Ağaç okunurken kural silsilesinin sonunda hedef değişken ve tahmin sayıları, doğru / yanlış şeklinde gösterilmektedir.

Örneğin OPPORTUNITY_GIFT > 0: H (31720.0/3202.0) kuralında Churn_flag hedef değişkeni "H" için 31720 doğru sınıflandırılmış gözlem 3202 yanlış sınıflandırılmış gözlemdir.

```
OPPORTUNITY_GIFT <= 0
|   INCENTIVE_BROCHURE_COUNT <= 1
|   |   AWARD_SALES <= 100.64
|   |   |   SL_G1_ACTIVITY <= 0.51715: E (8445.0/1649.0)
|   |   |   SL_G1_ACTIVITY > 0.51715
|   |   |   |   ACTIVIZATION = A: E (3021.0/978.0)
|   |   |   |   ACTIVIZATION = AA
|   |   |   |   |   STARTING_FAMILY = A.Classic: E (3631.0/1314.0)
|   |   |   |   |   STARTING_FAMILY = A.Classic Plus: E (2582.0/1250.0)
|   |   |   |   |   STARTING_FAMILY = A.Club: H (1264.0/572.0)
|   |   |   |   |   STARTING_FAMILY = A.Elite: H (37.0/9.0)
|   |   |   |   |   STARTING_FAMILY = A.Inci: E (1799.0/562.0)
|   |   |   |   ACTIVIZATION = AI: E (26489.0/6123.0)
|   |   |   |   ACTIVIZATION = I: H (5.0/2.0)
|   |   |   |   ACTIVIZATION = IA: E (11423.0/3396.0)
|   |   |   |   ACTIVIZATION = II
|   |   |   |   |   LOA = 1_6: E (4508.0/888.0)
|   |   |   |   |   LOA = 14_26: E (3978.0/1350.0)
|   |   |   |   |   LOA = 27+
|   |   |   |   |   |   AGE = 18_29: E (1756.0/672.0)
|   |   |   |   |   |   AGE = 30_49
|   |   |   |   |   |   |   SL_G1_REPS <= 106: E (2730.0/1226.0)
|   |   |   |   |   |   |   SL_G1_REPS > 106: H (1504.0/728.0)
|   |   |   |   |   |   AGE = 50_64: H (794.0/354.0)
|   |   |   |   |   |   AGE = 65+: H (103.0/48.0)
|   |   |   |   |   LOA = 7_13: E (4139.0/1301.0)
|   |   AWARD_SALES > 100.64
|   |   |   PREVIOUS_CAMPAIGN_AWARD_SALES <= 61.92: E (21885.0/7833.0)
|   |   |   PREVIOUS_CAMPAIGN_AWARD_SALES > 61.92
|   |   |   |   STARTING_FAMILY = A.Classic: E (7066.0/3017.0)
|   |   |   |   STARTING_FAMILY = A.Classic Plus
|   |   |   |   |   DEBT <= 89.86
|   |   |   |   |   |   TOTAL_UNITS <= 19
|   |   |   |   |   |   |   LOA = 1_6: H (221.0/93.0)
|   |   |   |   |   |   |   LOA = 14_26: E (1905.0/938.0)
|   |   |   |   |   |   |   LOA = 27+: H (5192.0/2190.0)
|   |   |   |   |   |   |   LOA = 7_13: H (1651.0/746.0)
|   |   |   |   |   |   TOTAL_UNITS > 19: H (1858.0/686.0)
|   |   |   |   |   DEBT > 89.86: E (1562.0/697.0)
|   |   |   |   STARTING_FAMILY = A.Club: H (6514.0/2249.0)
|   |   |   |   STARTING_FAMILY = A.Elite: H (160.0/59.0)
|   |   |   |   STARTING_FAMILY = A.Inci: E (2650.0/1114.0)
|   INCENTIVE_BROCHURE_COUNT > 1
|   |   AWARD_SALES <= 150
|   |   |   PREVIOUS_CAMPAIGN_AWARD_SALES <= 45.01: E (1554.0/716.0)
|   |   |   PREVIOUS_CAMPAIGN_AWARD_SALES > 45.01: H (2437.0/877.0)
|   |   AWARD_SALES > 150: H (20485.0/5366.0)
OPPORTUNITY_GIFT > 0: H (31720.0/3202.0)
```

Şekil 21. C4.5 Karar Ağacı Kuralları (%80-%20 eğitim, test verisi hold-out)

Şekil 21. İncelenerek kuralların bir kısmı aşağıda örnek olarak verilmiştir.

Kural 1 - Eğer OPPORTUNITY_GIFT > **0 Ayrılma yok**

Kural 2 - Eğer OPPORTUNITY_GIFT <= **0** ve AWARD_SALES <= **100.64** ve SL_G1_ACTIVITY <= **0.51715** ise **Ayrılma var**

Kural 3 - Eğer OPPORTUNITY_GIFT <= **0** ve AWARD_SALES <= **100.64** ve SL_G1_ACTIVITY > **0.51715** ve ACTIVIZATION =**A** ise **Ayrılma var**

Kural 4 - Eğer OPPORTUNITY_GIFT <= **0** ve AWARD_SALES <= **100.64** ve SL_G1_ACTIVITY > **0.51715** ve ACTIVIZATION =**AA** ve STARTING_FAMILY = **A.Classic** ise **Ayrılma var**

Kural 3 - Eğer OPPORTUNITY_GIFT <= **0** ve AWARD_SALES <= **100.64** ve SL_G1_ACTIVITY > **0.51715** ve ACTIVIZATION =**AA** ve STARTING_FAMILY = **A.Classic** ise **Ayrılma var**

Kural 4 - Eğer OPPORTUNITY_GIFT <= **0** ve AWARD_SALES <= **100.64** ve SL_G1_ACTIVITY > **0.51715** ve ACTIVIZATION =**AA** ve STARTING_FAMILY = **A.Classic Plus** ise **Ayrılma var**

Kural 5 - Eğer OPPORTUNITY_GIFT <= **0** ve AWARD_SALES <= **100.64** ve SL_G1_ACTIVITY > **0.51715** ve ACTIVIZATION =**AA** ve STARTING_FAMILY = **A.Club** ise **Ayrılma yok**

Kural 6 Eğer OPPORTUNITY_GIFT <= **0** ve AWARD_SALES <= **100.64** ve SL_G1_ACTIVITY > **0.51715** ve ACTIVIZATION =**AA** ve STARTING_FAMILY = **A.Elite** ise **Ayrılma yok**

Kural 7 - Eğer OPPORTUNITY_GIFT <= **0** ve AWARD_SALES <= **100.64** ve SL_G1_ACTIVITY > **0.51715** ve ACTIVIZATION =**AA** ve STARTING_FAMILY = **A.Inci** ise **Ayrılma var**

Kural 8 - Eğer OPPORTUNITY_GIFT <= **0** ve AWARD_SALES <= **100.64** ve SL_G1_ACTIVITY > **0.51715** ve ACTIVIZATION =**AI** ve ise **Ayrılma var**

Kural 9 - Eğer OPPORTUNITY_GIFT <= **0** ve AWARD_SALES <= **100.64** ve SL_G1_ACTIVITY > **0.51715** ve ACTIVIZATION =**IA** ve ise **Ayrılma var**

7.3. Model Performans Karşılaştırması

Gini karar ağacı ve C4.5 karar ağacı ile elde edilen modeller hold-out performans yöntemi ile değerlendirilmiştir. Veri seti sırasıyla eğitim ve test veri seti olmak üzere %90-%10, %80-%20, %70-%30, %60-%40 oranlarıyla ayrılmıştır. Sonrasında model kurulmuş ve sonuçlar, 6.4.2. Model Değerlendirme ve Performans Ölçütleri bölümünde detayları anlatılan kontenjans tablosu yöntemiyle ölçülmüştür. Gini karar ağacı ile oluşturulan modellerin performans sonuçları Tablo 5.'te C4.5 karar ağacı ile oluşturulan modellerin performans sonuçları Tablo 6'da gösterilmiştir.

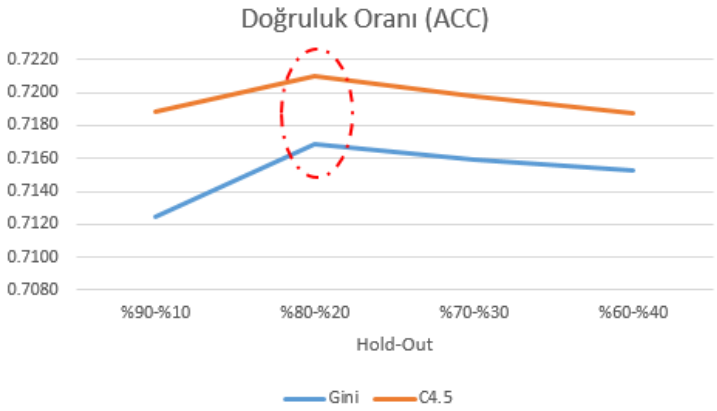
Tablo 5. Gini Karar Ağacı ile Oluşturulan Modellerin Performans Sonuçları

	Gini Karar Ağacı Algoritması			
	%90-%10	%80-%20	%70-%30	%60-%40
Doğruluk Oranı (ACC)	0.7125	0.7169	0.7159	0.7153
Hata Oranı (ERR)	0.2875	0.2831	0.2841	0.2847
Duyarlılık (TPR)	0.7878	0.7870	0.7853	0.7865
Belirleyicilik (SPC)	0.6359	0.6457	0.6454	0.6430
Pozitif Öngörü Değeri (PPV)	0.6874	0.6930	0.6924	0.6912
Negatif Öngörü Değeri (NPV)	0.7468	0.7489	0.7473	0.7477
Yanlış Pozitif Oranı (FPR)	0.3641	0.3543	0.3546	0.3570
Yanlış Negatif Oranı (FNR)	0.2122	0.2130	0.2147	0.2135
Pozitif Olabilirlik Oranı (LR +)	2.1638	2.2211	2.2145	2.2027
Negatif Olabilirlik Oranı (LR -)	0.3337	0.3299	0.3327	0.3321
Tanısal Üstünlük Oranı (DOR)	6.4846	6.7323	6.6567	6.6322
F-ölçüsü	0.7342	0.7370	0.7359	0.7358

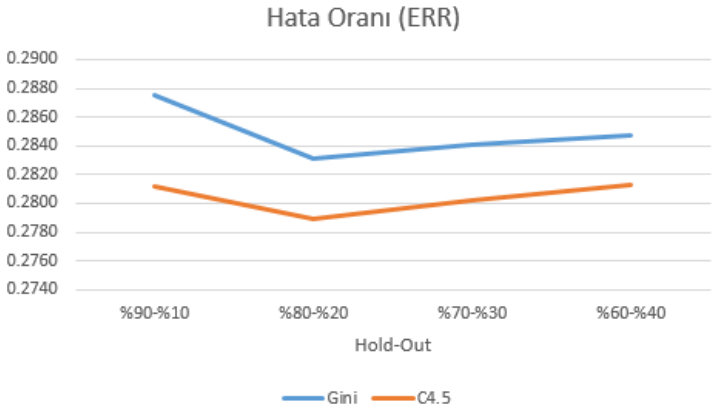
Tablo 6. C4.5 Karar Ağacı ile Oluşturulan Modellerin Performans Sonuçları

	C4.5 Karar Ağacı Algoritması			
	%90-%10	%80-%20	%70-%30	%60-%40
Doğruluk Oranı (ACC)	0.7188	0.7210	0.7198	0.7188
Hata Oranı (ERR)	0.2812	0.2790	0.2802	0.2812
Duyarlılık (TPR)	0.8096	0.8067	0.8069	0.8094
Belirleyicilik (SPC)	0.6266	0.6340	0.6313	0.6267
Pozitif Öngörü Değeri (PPV)	0.6878	0.6913	0.6898	0.6878
Negatif Öngörü Değeri (NPV)	0.7641	0.7635	0.7629	0.7639
Yanlış Pozitif Oranı (FPR)	0.3734	0.3660	0.3687	0.3733
Yanlış Negatif Oranı (FNR)	0.1904	0.1933	0.1931	0.1906
Pozitif Olabilirlik Oranı (LR +)	2.1682	2.2041	2.1884	2.1680
Negatif Olabilirlik Oranı (LR -)	0.3039	0.3049	0.3059	0.3042
Tanısal Üstünlük Oranı (DOR)	7.1356	7.2295	7.1540	7.1269
F-ölçüsü	0.7438	0.7446	0.7438	0.7437

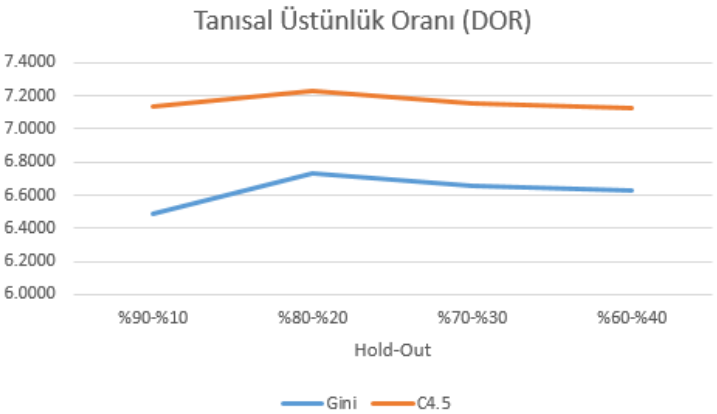
C4.5 Karar ağacı algoritması tüm hold-out ayrım değerlerinde birçok performans ölçütüne göre Gini karar ağacı algoritmasına göre daha başarılı sonuçlar üretmiştir. Her iki algoritma için de doğruluk oranı (ACC) kriterine göre en optimum ayrım değeri %80 eğitim %20 test veri seti olarak gözlemlenmiştir. Şekil 22.'de Doğruluk oranı (ACC), Şekil 23.'de hata oranı (ERR), Şekil 24.'de tanısal üstünlük oranı (DOR) ve Şekil 25.'te F-ölçüsü değerleri grafik olarak gösterilmiştir. Diğer performans kriterleri için Tablo 5. Ve Tablo 6. İncelenebilir.



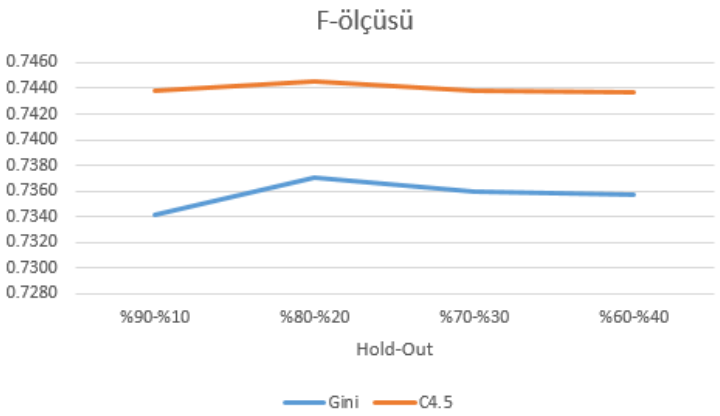
Şekil 22. Doğruluk Oranı (ACC)



Şekil 23. Hata Oranı (ERR)



Şekil 24. Tanısal Üstünlük Oranı (DOR)



Şekil 25. F-ölçüsü

SONUÇ

Çalışmada veri madenciliği teknikleri kullanılarak oldukça niş bir sektör olan doğrudan satış sektöründe müşteri kayıp analizi yapılması amaçlanmıştır. Diğer sektörlerden farklı olarak birden fazla müşteri adına alışveriş yapan Temsilcinin bir sonraki kampanya alışverişe devam edip etmeyeceği tahmin edilmeye çalışılmıştır.

Veri bir doğrudan satış şirketinden temin edilmiştir. Analizlerde 347.001 Temsilci verisi kullanılmıştır. Veri analiz öncesinde detaylıca incelenerek değişkenler anlaşılmaya çalışılmıştır. Sonrasında veri, analiz için hazırlanmıştır. Kayıp, aykırı (outliers) ve tekrar eden veriler tespit edilmiş ve veri setinden çıkarılmıştır. Yapılan analizlerde pozitif etkisi görülen değişkenlere veri dönüştürme işlemi uygulanmıştır. Nümerik değişkenler normalize edilmiş ancak bu veriler ile daha kötü sonuçlar elde edildiğinden normalize edilmiş veri modellerde kullanılmamıştır.

Analiz çalışması için oluşturulan modellerde sınıflandırma algoritmalarından karar ağaçları tercih edilmiştir. Gini karar ağacı ve C4.5 karar ağacı ile modeller oluşturulmuş ve modellerin sonuçları hold-out yöntemi ile ölçülmüştür. Bu yöntem ile veri seti sırasıyla eğitim ve test veri seti olmak üzere %90-%10, %80-%20, %70-%30, %60-%40 oranlarıyla ayrılmıştır.

Yapılan analiz sonucunda C4.5 karar ağacı algoritması Gini karar ağacı algoritmasına göre daha iyi sonuçlar üretmiştir. Her iki algoritma için doğruluk oranı (ACC) kriterine göre en optimum ayrım değeri %80 eğitim %20 test veri seti olarak gözlemlenmiştir. %80 eğitim %20 test veri seti üzerinde çalışan modellerde C4.5 karar ağacı doğruluk oranı (ACC) ~0.721 iken Gini karar ağacında bu oran ~0.717 olarak gerçekleşmiştir. Aynı sırayla tanısal üstünlük oranı (DOR) ~7.230 ile ~6.732, F-ölçüsü ~0.745 ile ~0.737 olarak gerçekleşmiştir.

Karar ağacının kurallarının listelendiği Şekil 21'de görüldüğü üzere karar ağacının kök değişkeni OPPORTUNITY_GIFT (kazanılan fırsat hediyesi) olmuştur. Yine Tablo 1'de görüleceği üzere hedef değişken ile en yüksek korelasyon yine OPPORTUNITY_GIFT değişkenidir. Görüldüğü gibi bir hediye kazanan Temsilciler

sistemden çıkmamakta ve bir sonraki kampanya sipariş vermektedirler. Müşteri kaybını önlemek için hediye bütçesi bir miktar arttırılarak daha düşük tutardaki siparişler için de hediye verilmesi müşteri kaybını önleyebilir.

Karar ağacının ilk dal değişkeni ise INCENTIVE_BROCHURE_COUNT (ücretli alınan katalog) olmuştur. En az bir adet ücretli katalog alan Temsilciler hiç ücretli katalog almayan Temsilcilere göre müşteri kaydı daha düşüktür. İkinci ve üçüncü dal ise AWARD_SALES (sipariş tutarı) olmuştur. 150 TL üzerinde sipariş tutarına sahip olan Temsilcilerin diğerlerine göre müşteri kaybı çok daha düşüktür. Tablo 1'de de görüldüğü üzere INCENTIVE_BROCHURE_COUNT ile AWARD_SALES değişkenleri arasında yüksek korelasyon görünmektedir. Bu bağlamda katalog fiyatlarında bir miktar indirim yapılarak veya dijital kanallar aktif şekilde kullanılarak Temsilcilerin kataloğa erişmeleri ve müşterileriyle paylaşmaları kolaylaştırılarak müşteri kaybı düşürülebilir.

Karar ağacında görüldüğü üzere hedef değişken ile ilintili diğer en önemli değişken PREVIOUS_CAMPAIGN_AWARD_SALES (Bir önceki kampanya sipariş tutarı) ve SL_G1_ACTIVITY (Temsilcinin SL'nin Temsilcilerinin kampanya içerisinde sipariş verme oranı) olarak görülmektedir. PREVIOUS_CAMPAIGN_ AWARD_SALES için önceki paragrafta AWARD_SALES hakkındaki yorumlar değerlendirilebilir. SL_G1_ACTIVITY oranını arttırmak için SL'lerine ekip yönetimi ve iletişim eğitimleri verilerek aktivite oranlarını yükseltmeleri desteklenebilir.

Tablo 1'incelendiğinde ORDER_COUNT (Sipariş sayısı) ile hedef değişken arasında korelasyon olduğu görünmektedir. Sipariş sayısı arttıkça müşteri kaybı da azalmaktadır. Sipariş sayısını arttırmak için minimum sipariş tutarı miktar düşürülerek Temsilcilerin bir kampanya içerisinde daha fazla adette sipariş vermeleri sağlanabilir. Bu sayede Temsilciler müşterilerinden yeterli talep toplamayı beklemeden siparişlerini girerek müşterilerine daha hızlı ürünlerini ulaştırabilirler. Hedef değişken ile bir diğer korelasyon LOA_GIFT (İlk üç kampanya içerisinde kazanılan hediye) arasında görülmektedir. Yeni temsilcilere hediye kazanan Temsilci oranını arttırmak

için iletişim frekansı arttırılarak hediye kazanma yöntemleri anlatılabilir. İlk üç kampanya ile sınırlı olan hediye kazanma fırsatı süresi uzatılabilir.

Çalışılan veri setinde GENDER (cinsiyet) değişkeninin müşteri kaybına etkisi olmadığı görülmüştür. Bunun sebebinin veri setindeki Temsilcilerin yaklaşık %93.5'inin kadın olmasından kaynaklandığı düşünülmektedir.

Veri seti zenginleştirilerek modellerin tahmin yetenekleri güçlendirilebilir. Çalışmada Temsilcilerin daha çok kişisel bilgileri (yaş, cinsiyet, Temsilcilik süresi vb.) ile sipariş tutar bilgileri incelenmiştir. Veri setine, ürün ve kategori bilgisi, Temsilcinin alışveriş sıklıkları, Temsilci ile yapılan iletişimler (müşteri iletişim merkezi, SMS, mail vb.), Temsilcinin en son şikâyette bulunduğu konular ve şirketin çözüm sonuçları gibi ek değişkenler eklenerek tahmin oranı yükseltilebilir.

KAYNAKÇA

[1] Şeker, Ş. E., **Müşteri Kayıp Analizi (Customer Churn Analysis),** YBS Ansiklopedi, Cilt 3, Sayı 1, 2016, ss. 26-27.

[2] Rechinhheld, F., Sasser, W., **Zero Defections: Quality comes to service**, Harvard Business review, 1990

[3] Önay Koçoğlu, F., Özcan, T., Baray, Ş. A., **Veri Madenciliğinde Ayrılan müşteri Analizi Problemi Üzerine Bir Literatür Araştırması,** 2016

[4] Arslan, İ., **R ile İstatistiksel Programlama** (1. Baskı), Pusula Yayınevi, İstanbul, 2015

[5] Yıldız, M., Şeker, Ş. E., **Veri Madenciliği Araçları (Data Mining Tools),** YBS Ansiklopedi, Cilt 3, Sayı 4, 2016, ss. 10.

[6] Albayrak, A. S., Koltan Yılmaz, Ş., **Veri Madenciliği: Karar Ağacı Algoritmaları ve İMKB Verileri Üzerine Bir Uygulama,** Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, Cilt 14, Sayı 1, 2009, ss. 35.

[7] Savaş, S., Topaloğlu, N., Yılmaz M., **Veri Madenciliği ve Türkiye'deki Uygulama örnekleri,** İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, Cilt 11, Sayı 21, 2012, ss. 4-5.

[8] Silahtaroğlu, G., **Kavram ve Algoritmalarıyla Temel Veri Madenciliği** (1. Basım), Papatya Yayınevi, İstanbul, 2008

[9] Satman, M. H., **Genetik Algoritmalar** (1. Basım), Papatya Kitapevi, İstanbul, 2016

[10] Akküçük, U., **Veri Madenciliği Kümeleme ve Sınıflama Algoritmaları** (1. Basım), Yalın Yayıncılık, İstanbul, 2011

[11] Çalış, A., Kayapınar, S., Çetinyokuş, T., **Veri Madenciliğinde Karar Ağacı Algoritmaları ile Bilgisayar ve İnternet Güvenliği Üzerine Bir Uygulama,** Endüstri Mühendisliği Dergisi, Cilt 25, Sayı 3-4, 2014, ss. 2-19.

[12] Aytekin, Ç., Sütçü, C., S., Özfidan, U., **Karar Ağacı Algoritması ile Metin Aınıflandırma: Müşteri Yorumları Örneği,** Uluslararası Sosyal Araştırmalar Dergisi, Cilt 11, Sayı 55, 2018, ss. 782-792.

[13] Balaban, M., E., Kartal, E., **R ile Veri Madenciliği Uygulamaları** (1. Basım), Çağlayan Kitapevi, İstanbul, 2016

İNTERNET KAYNAKLARI

[14] Doğrudan Satış Derneği, **Doğrudan Satış Nedir,** http://dsd.org.tr/dogrudan-satis-nedir/ (Erişim Tarihi: 05.01.2019)

[15] Doğrudan Satış Derneği, **Sıkça Sorulan Sorular,** http://dsd.org.tr/sikca-sorulan-sorular/ (Erişim Tarihi: 05.01.2019)

[16] Doğrudan Satışın Tarihçesi, https://www.disells.com/dogrudan-satisin-temelleri-education.html (Erişim Tarihi: 05.03.2019)

[17] Kılıç, D., **Makine Öğrenimi ve Derin Öğrenme ile Müşteri Kayıp (Churn) Analizi-1,** https://medium.com/deep-learning-turkiye/makine-%C3%B6%C4%9Frenimi-ve-derin-%C3%B6%C4%9Frenme-ile-m%C3%BC%C5%9Fteri-kay%C4%B1p-churn-analizi-1-63a4513b8a6f (Erişim Tarihi: 05.03.2019)

Ek-1: Gini Karar Ağacı Algoritması R kodu

```r
# Gini Karar Ağacı Algoritması Modeli
#Kullanılacak paketlerin aktif edilemsi
library(caret)
library(rpart)
library(rpart.plot)
#Rasgele eğitim ve test veri seti ayrımı
set.seed(2018)
oran_yuzde_90_gini<-createDataPartition(y=raw_data$CHURN_FLAG, p=.90, list=FALSE)
#%90 eğitim ve %10 test verisi ayrımı. Diğer Hold-out oranları için "P" parametresi değiştirlmelidir
egitim_yuzde_90_gini <- raw_data[oran_yuzde_90_gini,]
test_yuzde_90_gini <- raw_data[-oran_yuzde_90_gini,]
#Eğitim veri seti ile gini karar ağacı modelinin oluşturulması
gini_model<-rpart(CHURN_FLAG~.,egitim_yuzde_90_gini,method="class",minsplit=50,parms=list(split="gini"))
#Karar ağacı sınıflarının gösterilemsi
show(gini_model)
#Gini karar ağacı grafiğinin çizdirilmesi
prp(gini_model)
#Test veri seti üzerinden tahmin işleminin yapılması
gini_tahmin<-predict(gini_model,test_yuzde_90_gini,type="class")
#confusion (karışıklık) matrisinin oluşturulması
TGini<-table(gini_tahmin, test_yuzde_90_gini$CHURN_FLAG, dnn=c("Tahmin", "Gerçek"))
TGini
#Performans Değerlendirmesi
(tp <- TGini[1])
(fp <- TGini[3])
(fn <- TGini[2])
(tn <- TGini[4])
paste0("Dogruluk = ",(dogruluk <- (tp+tn)/sum(TGini)))
paste0("Hata = ",(hata <- 1-dogruluk))
paste0("TPR = ",(TPR <- tp/(tp+fn)))
paste0("SPC = ",(SPC <- tn/(fp+tn)))
paste0("PPV = ",(PPV <- tp/(tp+fp)))
paste0("NPV = ",(NPV <- tn/(tn+fn)))
paste0("FPR = ",(FPR <- fp/(fp+tn)))
paste0("FNR = ",(FNR <- fn/(fn+tp)))
paste0("LR_p = ",(LR_p <- TPR/FPR))
paste0("LR_n = ",(LR_n <- FNR/SPC))
paste0("DOR = ",(DOR <- LR_p/LR_n))
paste0("F_measure = ",(F_measure <- (2*PPV*TPR)/(PPV+TPR)))
```

Ek-2: C4.5 Karar Ağacı Algoritması R kodu

```r
# C4.5 Karar Ağacı Algoritması Modeli
#Kullanılacak paketlerin aktif edilemsi
library(caret)
library(RWeka)
#Rasgele eğitim ve test veri seti ayrımı
set.seed(2018)
oran_yuzde_90 <- createDataPartition(y = raw_data$CHURN_FLAG, p = .90, list = FALSE)
#%90 eğitim ve %10 test verisi ayrımı. Diğer Hold-out oranları için "P" parametresi değiştirlmelidir
egitim_yuzde_90 <- raw_data[oran_yuzde_90,]
test_yuzde_90 <- raw_data[-oran_yuzde_90,]
#Eğitim veri seti ile C4.5 karar ağacı modelinin oluşturulması
c_4_5_model <-J48(CHURN_FLAG~., data=egitim_yuzde_90)
#Karar ağacı sınıflarının gösterilemsi
print(c_4_5_model)
#Karar ağacı özet bilgisinin gösterilemsi
summary(c_4_5_model)
#C4.% karar ağacı grafiğinin çizdirilmesi
plot(c_4_5_model)
#Test veri seti üzerinden tahmin işleminin yapılması
tahmin_egitim_yuzde_90<-predict(c_4_5_model, newdata = test_yuzde_90[,-27])
#confusion (karışıklık) matrisinin oluşturulması
km_yuzde_90 <-table(tahmin_egitim_yuzde_90, test_yuzde_90$CHURN_FLAG, dnn = c("Tahminler","Gerçekler"))
km_yuzde_90
#performans değerlendirme
(tp<-km_yuzde_90[1])
(fp<-km_yuzde_90[3])
(fn<-km_yuzde_90[2])
(tn<-km_yuzde_90[4])
paste0("Dogruluk = ",(dogruluk <- (tp+tn)/sum(km_yuzde_90)))
paste0("Hata = ",(hata <- 1-dogruluk))
paste0("TPR = ",(TPR <- tp/(tp+fn)))
paste0("SPC = ",(SPC <- tn/(fp+tn)))
paste0("PPV = ",(PPV <- tp/(tp+fp)))
paste0("NPV = ",(NPV <- tn/(tn+fn)))
paste0("FPR = ",(FPR <- fp/(fp+tn)))
paste0("FNR = ",(FNR <- fn/(fn+tp)))
paste0("LR_p = ",(LR_p <- TPR/FPR))
paste0("LR_n = ",(LR_n <- FNR/SPC))
paste0("DOR = ",(DOR <- LR_p/LR_n))
paste0("F_measure = ",(F_measure <- (2*PPV*TPR)/(PPV+TPR)))
```

ÖZGEÇMİŞ

30 Mayıs 1987 tarihi, İstanbul ili Şile ilçesi doğumluyum. Şişli Teknik Lisesi Bilgisayar Yazılım bölümünden 2005 yılında, Marmara Üniversitesi Teknik Bilimler Meslek Yüksekokulu Bilgisayar Teknolojisi ve Programlama bölümünden 2007 yılında, Anadolu Üniversitesi İşletme Fakültesi'nden 2010 yılında, Anadolu Üniversitesi Adalet bölümünden 2018 yılında mezun oldum. Şu anda Beykent Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği yüksek lisans eğitimimi sürdürmekteyim.