MARMARA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

SİGORTACILIK SEKTÖRÜNDE MÜŞTERİ ÜRÜN EĞİLİM MODELLEMESİ

İBRAHİM FURKAN AKYÜZ

YÜKSEK LİSANS TEZİ
İstatistik Anabilim Dalı
İstatistik Programı

DANIŞMAN
Prof. Dr. Birsen EYGİ ERDOĞAN

İSTANBUL, 2023

MARMARA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

SİGORTACILIK SEKTÖRÜNDE MÜŞTERİ ÜRÜN EĞİLİM MODELLEMESİ

İBRAHİM FURKAN AKYÜZ

521320004

YÜKSEK LİSANS TEZİ
İstatistik Anabilim Dalı
İstatistik Programı

DANIŞMAN
Prof. Dr. Birsen EYGİ ERDOĞAN

İSTANBUL, 2023

ÖNSÖZ VE TEŞEKKÜR

"Sigortacılık Sektöründe Müşteri Ürün Eğilim Modellemesi" isimli bu tez Marmara Üniversitesi Fen Bilimleri Enstitüsü, İstatistik Anabilim Dalı, Yüksek Lisans Programı'nda hazırlanmıştır ve bu tezde sigorta sektöründe yer alan bir ürün için önemli olan müşteri özellikleri araştırılmıştır.

Tez çalışma konusunun belirlenmesinde ve çalışmanın planlanmasında, araştırılmasında, yürütülmesinde ilgi ve desteğini esirgemeyen, değerli bilgi ve tecrübelerinden yararlandığım, her fırsatta yardımcı olan değerli hocam Prof. Dr. Birsen EYGİ ERDOĞAN'a, Marmara Üniversitesi'nde geçirdiğim yüksek lisans hayatım boyunca bana kazandırdıkları her şey için değerli bölüm hocalarıma, hayatım boyunca beni her zaman destekleyen ve arkamda duran aileme ve sevdiklerime sonsuz teşekkürlerimi sunarım. Bu tezin, bundan sonra yapılacak çalışmalara katkı sağlamasını temenni ederim.

Bu çalışma, Marmara Üniversitesi Bilimsel Araştırma Projeleri Koordinatörlüğü tarafından Yüksek Lisans Tez Projeleri (FYL-2022-10626) kapsamında desteklenmiştir. Bu destek için Marmara Üniversitesi Bilimsel Araştırma Projeleri Koordinatörlüğü'ne ayrıca teşekkürlerimi sunarım.

İÇİNDEKİLER

ÖNSÖZ VE TEŞEKKÜR ..... **İ**

İÇİNDEKİLER ..... **İİ**

ÖZET ..... **İV**

ABSTRACT ..... **Vİ**

SEMBOLLER ..... **Viİİ**

ŞEKİL LİSTESİ ..... **İX**

TABLO LİSTESİ ..... **X**

1.BÖLÜM: GİRİŞ **1**

1.1.KONU VE KAPSAM 1

1.2.LİTERATÜR ÖZETİ 2

1.3.AMAÇ VE YÖNTEM 4

2.BÖLÜM: İSTATİSTİKSEL ANALİZ **7**

2.1.İSTATİSTİKSEL ANALİZ NEDİR? 7

2.1.1.KORELASYON ANALİZİ *7*

2.1.2.ÇOKLU REGRESYON ANALİZİ *8*

2.1.3.LOJİSTİK REGRESYON *8*

2.1.4.FAKTÖR ANALİZİ *9*

2.1.5.TEMEL BİLEŞEN ANALİZİ *10*

2.1.6.LASSO REGRESYON *10*

3.BÖLÜM: MAKİNE ÖĞRENMESİ **12**

3.1.MAKİNE ÖĞRENMESİ NEDİR? 12

3.1.1.DENETİMLİ MAKİNE ÖĞRENMESİ *12*

3.1.2.DENETİMSİZ MAKİNE ÖĞRENMESİ *12*

3.1.3.PEKİŞTİRMELİ MAKİNE ÖĞRENMESİ *13*

3.2.EN YAYGIN MAKİNE ÖĞRENMESİ ALGORİTMALARI 14

3.2.1.DESTEK VEKTÖR MAKİNESİ (SUPPORT VECTOR MACHINE) *15*

3.2.2.RASTGELE ORMAN (RANDOM FOREST) *15*

3.2.3.XGBOOST *16*

3.3.MODEL PERFORMANS METRİKLERİ 17

3.3.1.DOĞRULUK (ACCURACY) *18*

3.3.2.KESİNLİK (PRECISION) *18*

3.3.3.HATIRLAMA (RECALL) *19*
3.3.4.F1 SKOR (F1 SCORE) *19*
3.3.5.AUC-ROC EĞRİSİ (AUC-ROC CURVE) *19*
4.BÖLÜM: UYGULAMA **21**
4.1.VERİ SETİ TANIMI 21
4.2.VERİ SETİ ANALİZİ 22
4.3.MODEL KURMA ADIMLARI 26
4.3.1.VERİ ÖN İŞLEME (DATA PREPROCESSING) *26*
4.3.2.ÖZELLİK MÜHENDİSLİĞİ (FEATURE ENGINEERING) *29*
4.3.3.ÖZELLİK SEÇİMİ (FEATURE SELECTION) *31*
4.3.4.VERİ ÖLÇEKLENDİRME *32*
4.3.5.MODEL PERFORMANSLARI VE MODEL SEÇİMİ *33*
4.3.6.MODEL İYİLEŞTİRME *35*
5.BÖLÜM: SONUÇ VE TARTIŞMA **39**
KAYNAKLAR **40**
ÖZGEÇMİŞ **43**

ÖZET

SİGORTACILIK SEKTÖRÜNDE MÜŞTERİ ÜRÜN EĞİLİM MODELLEMESİ

İbrahim Furkan AKYÜZ

Fen Bilimleri Enstitüsü
İstatistik Programı
Yüksek Lisans Tezi

Danışman: Prof. Dr. Birsen EYGİ ERDOĞAN

Günümüz dünyasında değişen tüketim alışkanlıkları ile beraber müşteri hareketlerini anlamak ve pazarlama faaliyetlerine yön vermek oldukça önemli hale geldi. Teknolojinin gelişmesi, yeni sektörlerin doğması, online alışverişin kolaylaşması gibi etmenler ile birlikte müşterilerin piyasadaki ürünlere ulaşma imkanları ve satın alma eğilimleri arttı. Artan hacimle birlikte var olan müşteri potansiyelinin büyüklüğü göz önüne alındığında bir şirket veya kuruluş, mevcutta yer alan müşteri potansiyelini doğru yöneterek şirketin veya kuruluşun büyümesine ve gelişmesine katkı sağlayabilirler. Şirketler veya kuruluşlar, müşterilerini veri analizi yardımıyla sınıflandırarak mevcutta yer alan ürünlerini doğru müşteri kitlelerine pazarlayarak satış oranlarını arttırabilir, zamandan tasarruf kazanabilir, iş gücünü doğru kullanabilir. Bunu sağlamak için ilk yöntemlerden birisi müşterilerin eğilimli oldukları ürünleri bulmak adına yapılan modellemelerdir.

Bir müşteri davranışının modellemesi yapılırken veri setlerinde yer alan müşteri bilgileri, müşterilerin davranışlarını anlamak ve ürüne ait eğilim skorlarını bulmak için çok önemlidir. Bir ürüne ait yapılan başarılı satışlar ve başarısız satış denemeleri belirlenerek kalan müşteri havuzundaki müşterilerin o ürüne eğilimi bilimsel yöntemlerle tahmin edilebilir. Bu bilimsel yöntemler bilgisayar ortamında birçok algoritma kullanılarak yapılabilir ve bir makine öğrenmesi sistemi kurulabilir. Girdi verilerinden tahmini bir model oluşturan bu sistem ile öğrenilen model, daha önce hiç görülmemiş müşteri verileri için yararlı tahminler yapabilir.

Müşteri eğilim modellemesi yapılırken kullanılacak model için denetimli öğrenme türlerinden olan sınıflandırma modelleri kullanılabilir. Sınıflandırma işlemi, bir veri seti üzerinde tanımlı olan çeşitli sınıflar arasında veriyi dağıtmaktır. Sınıflandırma modelleri, verilen eğitim kümesinden bu dağılım şeklini öğrenir ve sınıfın belirli olmadığı bir test kümesini doğru şekilde sınıflamaya çalışır. Sınıflandırma modellerinde tahmin edilen bağımlı değişken kategoriktir. Bağımlı değişken müşterinin ürünü satın alıp almaması -hedef değişkeni- olurken, bağımsız değişkenler ise müşteriye ait bilgileri içeren değişkenleridir.

Şirketler veya kuruluşlar, var olan kaynaklarını doğru yönlendirerek, hem ürünleriyle topluma faydalı bir şekilde hizmet etmek hem de kendi varlıklarını sürdürebilir kılmak için müşteri davranışlarını doğru tanımlamaya çalışmalıdır. Bunun için de müşterilere sunulacak ürün ve hizmetler için pazarlama faaliyetleri belirlenirken müşterilerine özel eğilim ve segmentasyon çalışmaları yapılmalıdır. Oluşturulan modeller ile ürün ve hizmetlerin sunumu için rastgele kitlelere efor harcamak yerine, modelleme sayesinde özellikleri belirlenmiş kitlelere ulaşarak optimizasyon sağlanmalıdır. Sigortacılık sektöründe satışlar genellikle çağrı merkezleri üzerinden yapılmaktadır. Aranacak kitlenin sunulacak ürüne eğilimi yüksek kişilerden oluşması bu optimizasyonu sağlayacaktır.

Anahtar Kelimeler: Modelleme, eğilim, sınıflandırma

ABSTRACT

CUSTOMER PRODUCT PROPENSITY MODELING IN THE INSURANCE INDUSTRY

İbrahim Furkan AKYÜZ

Institute of Science
Statistics Program
Master's Thesis

Supervisor: Prof. Dr. Birsen EYGİ ERDOĞAN

With the changing consumption habits in today's world, it has become very important to understand customer movements and give direction marketing activities. With the development of technology, the emergence of new sectors and the facilitation of online shopping, the opportunities of customers to reach the products in the market and their purchasing tendencies have increased. Considering the size of the existing customer potential with the increasing volume, a company or organization can contribute to the growth and development of the company or organization by managing the existing customer potential correctly. By classifying their customers with the help of data analysis, companies or organizations can increase their sales rates, save time, and use their workforce correctly by marketing their existing products to the right customer groups. One of the first methods to achieve this is modeling to find the products that customers are inclined to.

When modeling a customer behavior, the customer information in the datasets is very important to understand the behavior of the customers and to find the customers' propensity scores of the product. By determining the successful sales and unsuccessful sales attempts of a product, the tendency of the customers in the remaining customer pool to that product can be predicted using scientific methods. These scientific methods can be done by using many algorithms in the computer environment and a machine learning system can be established. Creating a predictive model from input data, the learned model can make useful predictions for never-before-seen

customer data. Classification models, one of the types of supervised learning, can be used for the model to be used while performing customer propensity modeling. The classification process is to distribute data among various classes defined on a data set. Classification models learn this distribution pattern from the given training set and try to correctly classify a test set for which the class is not specific. The predicted dependent variable in classification models is categorical. The dependent variable is whether the customer buys the product -target value- while the independent variables are the variables containing the customer's information.

Companies or organizations should try to define customer behaviors correctly in order to both serve the society in a beneficial way with their products and to make their own existence sustainable by directing their existing resources. For this, while determining the marketing activities for the products and services to be offered to the customers, specific trend and segmentation studies should be carried out. Instead of spending effort on random audiences for the presentation of products and services with the models created, optimization should be ensured by reaching the masses whose characteristics are determined by modeling. In the insurance sector, sales are generally made through call centers. This optimization will ensure that the audience to be searched consists of people with a high propensity towards the product to be presented.

Keywords: Modeling, propensity, classification

SEMBOLLER

λ **:** Küçültme Sabiti

∑ **:** Toplam

√ **:** Karekök

| | **:** Mutlak Değer

e **:** Euler Sayısı

ŞEKİL LİSTESİ

Şekil 1: Model Kurma Süreç Akışı 6
Şekil 2: Makine Öğrenmesi Teknikleri 14
Şekil 3: Hedef Değişken Değerinin Dağılımı 22
Şekil 4: Cinsiyet Dağılımı 23
Şekil 5: Medeni Durum Dağılımı 23
Şekil 6: Eğitim Düzeyi Dağılımı 24
Şekil 7: Şehir Düzeyi Dağılımı 24
Şekil 8: Gelir Düzeyi Dağılımı 25
Şekil 9: Korelasyon Matrisi 25
Şekil 10: Veri Dağılımı 27
Şekil 11: Aykırı Değer Bulunan Değişkenler 28
Şekil 12: Verilerin Yeni Dağılımı 29
Şekil 13: Lasso Regresyon Değişken Azaltma 32
Şekil 14: Lasso Regresyon Sonrası Kalan Değişkenler 32
Şekil 15: Lojistik Regresyon Sınıflandırma Raporu 33
Şekil 16: Destek Vektör Makinesi Sınıflandırma Raporu 33
Şekil 17: Rastgele Orman Sınıflandırma Raporu 33
Şekil 18: XGBoost Sınıflandırma Raporu 34
Şekil 19: AUC-ROC Eğrisi Karşılaştırması 34
Şekil 20: XGBoost Seçilen Hiperparametreler 36
Şekil 21: XGBoost Model İyileştirme 36
Şekil 22: XGBoost Karışıklık Matrisi 37

TABLO LİSTESİ

Tablo 1: Değişken Listesi 21
Tablo 2: Hedef Değişken Değerinin Adetsel Dağılımı 22
Tablo 3: Model Karşılaştırması 35
Tablo 4: Değişken Önem Derecesi 38

1. GİRİŞ

1.1. Konu ve Kapsam

Günümüzde ve çağımızda, müşteri davranışlarının itici güçlerini anlamak için verilerden yararlanan bir işletme, gerçek bir rekabet avantajına sahiptir. İşletmeler, müşteri seviyesindeki verileri efektif bir şekilde analiz ederek ve eforlarını etkileşim kurma olasılığı daha yüksek olan müşterilere odaklayarak pazardaki performanslarını önemli ölçüde arttırabilirler.

İşletmenin müşteriler üzerinde uygulayacağı pazarlama faaliyetleri için kaynakları sınırsız değildir. İşletmeler ve işletmelerin pazarlama ekiplerinin, belirli müşteri gruplarının belirli koşullar altındaki davranışlarını tahmin etmeye çalışması yaygın bir uygulama biçimidir. Geleneksel istatistiksel yöntemleri veya insan sezgileri kullanılarak oluşturulan tahminlerin doğruluğu her zaman yüksek değildir. Bu temel nedenlerden dolayı işletmeler alacağı aksiyonlarda kişiselleştirilmiş pazarlamanın potansiyelini doğru kullanmak adına makine öğrenmesinin gücünü kullanmalıdır.

Büyük veri stratejilerini uygulayan birçok e-ticaret şirketi için çözülmesi gereken en önemli konu, kullanıcıların tercihlerini öğrenmek, ihtiyaçlarını anlamak, satın alma davranışlarını satın alma niyetlerine göre analiz etmek için bu verileri etkin ve verimli bir şekilde nasıl kullanacaklarıdır [1]. Potansiyel müşterileri doğru bir şekilde belirlemek, teknik düzeyde hassas pazarlamaya ulaşmanın anahtarıdır [2].

Verilerden bir içgörü çıkarmak için yaklaşımlardan biri; müşterilerin demografik bilgileri (yaş, ırk, din, cinsiyet, aile büyüklüğü, gelir düzeyi, eğitim düzeyi vb.), müşterinin psikografik bilgileri (kişilik özellikleri, yaşam tarzı, sosyal sınıf vb.), müşteri etkileşimleri (açılan e-postalar, tıklanan e-postalar, mobil uygulamada yapılan aramalar, web sayfasında kalma süresi vb.), müşteri deneyimi (ortalama hizmet süresi, müşteri hizmetlerinde bekleme süresi, geri ödeme sayısı vb.) ve müşteri davranışları (farklı zamanlarda hizmet satın alma, son hizmet alımından geçen süre, son teklif ve etkileşimden geçen süre vb.) gibi bilgileri belirli bir müşteri profilinin belirli bir davranış türünü gerçekleştirme olasılığını tahmin etmek için birleştiren eğilim modellemesidir. Eğilim modellemesi, hedef kitlenin geçmiş davranışlarını analiz ederek davranışını tahmin etmek için tahmine dayalı modeller oluşturmaya yönelik bir dizi yaklaşımdır.

Tahmini analitik, gelecekteki olayları veya davranışları tahmin eden modeller geliştirmek için kullanılan çeşitli istatistiksel ve analitik teknikleri tanımlayan geniş bir terimdir. Bu tahmine dayalı modellerin biçimi, tahmin ettikleri davranışa veya olaya bağlı olarak değişir. Tahmine dayalı modellerin çoğu, belirli bir davranış veya olayın meydana gelme olasılığının daha yüksek olduğunu gösteren bir puan üretir [3].

Tahmine dayalı modelleme, sonuçları tahmin etmek için istatistikleri kullanır [4]. Tahmine dayalı modelleme bir eğitim veri seti ile başlar. Bir eğitim veri setindeki gözlemler, eğitim örnekleri veya kayıtları olarak bilinir. Değişkenler girdiler (tahmin ediciler, özellikler, açıklayıcı değişkenler veya bağımsız değişkenler) ve hedefler (tepki, sonuç veya bağımlı değişken) olarak adlandırılır. Belirli bir durum için girdiler, hedefi ölçmeden önceki bilgi durumunuzu yansıtır [5].

Eğilim modelleri, hedeflenen müşterilere ve çözülmesi gereken sorunlara göre çeşitli konular üzerinde ele alınabilir. Bu konular, müşteri kazanmayı amaçlayan satın alma eğilim modelleri veya müşteri kaybetmemeyi amaçlayan müşteri kaybetme eğilim modelleri olabilir. Satın alma eğilim modelleri, hangi müşterilerin ürün veya hizmetleri satın alma eylemini gerçekleştirme olasılığının daha yüksek veya daha düşük olduğunu gösterir. Müşteri kaybetme eğilim modelleri, hangi müşterilerin ürün veya hizmetlerden memnun olmadığını, risk altında olduğunu, ürünü veya hizmetleri bırakma olasılığının daha yüksek veya daha düşük olduğunu gösterir.

Müşterilerin bir ürünü satın alma eğilimi, birçok sektörde istatistiksel modeller kullanılarak başarılı bir şekilde tahmin edilmiştir. Bu kapsamda, eğilim modelleri, satış ve pazarlama faaliyeti bulunan bütün sektörlerde, bankacılık, sigortacılık, sağlık, ulaşım, seyahat, ticaret, e-ticaret gibi, kullanım potansiyeline sahiptir.

Bu tezde, sigortacılık sektöründen alınan gerçek bir veri üzerinden bir modelleme kurgulanarak bir sigorta ürününe ait müşterinin ürün eğilimini hangi değişkenlerin etkilediği ve aynı zamanda satın alan ve almayan müşteri ayrımını yapabilecek modeli tahmin etmek konu edilecektir.

1.2. Literatür Özeti

Tahmine dayalı modeller, tüm insanlık tarihi boyunca kullanılmıştır. Örneğin, eski Mısırlılar,

Nil'in taşmasını tahmin etmek için Sirius'un yükselişinin gözlemlerini kullandılar.

1689'da, Lloyd şirketi tarafından deniz yolculukları için sigorta yapmak üzere tahmine dayalı modeller kullanıldı. Şirket, verileri kullanarak, bir prim karşılığında deniz yolculuklarının riskini kabul edecekti. Lloyd, bu yolculukların riskini değerlendirmek ve sorumluluk kalıplarını tahmin etmek için geçmiş yolculukların veri setlerini kullandı.

Model inşasına daha titiz bir yaklaşım ise iki yüzyıldan uzun bir süre önce Legendre tarafından 1805'te ve Gauss tarafından 1809'da yayınlanan en küçük kareler yöntemine atfedilebilir. Zamanla ekonomi, tıp, biyoloji ve tarımdaki uygulama sayısı arttı.

Regresyon terimi, 1886'da Francis Galton tarafından icat edildi. Bu terim, başlangıçta biyolojik uygulamalara atıfta bulunurken, günümüzde sürekli değişkenlerin tahminine izin veren çeşitli modeller için kullanılmaktadır.

Sınıflandırma olarak adlandırılan nominal değişkenlerin tahmininin başlangıcı, 1936'da Ronald Fisher'ın çalışmalarına atfedilebilir.

Fair Isaac Corporation'a dönüşen danışmanlığın kurulduğu 1950'lerin başında istatistiksel kredi puanlamanın tanıtılmasından bu yana müşteri yönetimindeki sorunlara istatistiksel modelleme uygulandı. Bunu, özellikle talebi teşvik etme ve müşteriyi elde tutma alanlarında, müşteri hedefleme için tahmine dayalı modellemenin giderek daha sofistike kullanımı izledi (Thomas, 2000).

İflası tahmin etmeye yönelik Z-skoru formülü, Edward Altman tarafından 1968'de yayınlandı. Formül, bir firmanın iki yıl içinde iflas etme olasılığını tahmin etmek için kullanıldı. Z-skoru, kurumsal temerrütleri tahmin etmek için ve akademik çalışmalarda şirketlerin mali sıkıntı durumu için hesaplaması kolay bir kontrol ölçüsü olarak kullanılır. 1980'de James Ohlson tarafından finansal sıkıntıyı tahmin etmek için Altman Z-skoruna alternatif olarak öne sürülen çok faktörlü bir finansal formül olan Ohlson O-skoru yayınlandı.

Eğilim modellemesinin kökenleri 1983'e kadar izlenebilse de makine öğrenmesinin tüm potansiyelini ortaya çıkarması ancak yakın zamanda mümkün oldu. 1983'te istatistikçiler Rosenbaum ve Rubin, davranışların nedenlerini ve sonuçlarını anlamanın bir yolu olarak eğilim

modellemesi fikrini ortaya attılar.

Müşteri satın alma davranışının analizi, e-ticaretin başlangıcına kadar uzanır (Bellman, Lohse ve Johnson, 1999) ve günümüzde birçok uygulamaya sahiptir. Müşteriler için ürün önerileri (Y. Tan, Xu ve Liu, 2016; Hidasi, Quadrana, Karatzoglou ve Tikk, 2016), müşterilerin alıcılar, ziyaretçiler vb. gibi belirli kategorilere göre sınıflandırılması (Moe, 2003; Fajta, 2014), satın alma olasılığı daha yüksek olan müşterilere daha kaliteli hizmet sunmak için satın alma olasılığını tahmin etme (Lo, Frankowsik ve Leskovec, 2014; Korpusik, Sakaki, Chen ve Chen, 2016; Suchacka ve Templewski , 2017; Lang & Rettenmeier, 2017) ve müşteri kaybını önlemek için müşteri kaybının zamanında tespit edilmesi (Xie, Li, Ngai, & Ying, 2008; Castanedo, Valverde, Zaratiegui ve Vazquez, 2014) gibi çalışmalar günümüzdeki uygulamalara örnek çalışmalardır.

1.3. Amaç ve Yöntem

Tahmine dayalı bir analitik görevi olan sınıflandırmanın amacı, bir dizi girdi değişkeninden kategorik bir hedef değişkeni tahmin etmektir. Bu hedef değişken, çeşitli kategoriler aracılığıyla ifade edilebilir veya ikili nitelikte olabilir (Kotu ve Deshpande, 2014). Eldeki görev ikili bir sınıflandırma görevidir, çünkü hedef değişkenin iki kategorisi vardır: satın alma ve satın almama. Hedef değişkeni tahmin etmek için, etiketli bir veri kümesinden girdi ve hedef değişken arasındaki genelleştirilmiş bir ilişki öğrenilir. Daha sonra sınıflandırma için yeni verilere uygulanır. Öğrenme algoritmaları hem eğitim verilerine uymalı hem de yeni veriler üzerinde genelleme yapmalıdır (P. Tan, Steinbach ve Kumar, 2005). Bu ilişkiyi ayıklamak için farklı yöntemlere sahip çeşitli makine öğrenmesi algoritmaları mevcuttur. Bu tez, sigortacılık sektöründen alınan gerçek bir veri üzerinden başarılı bir sınıflandırma modeli kurgusu oluşturularak, sigorta sektöründe yer alan bir ürün için hangi değişkenlerin önemli olduğunu bulmayı ve işletmelerin hedef müşteri seçimlerinde oluşturulan model kurgusunu kullanarak verimlilik elde etmesini amaçlamaktadır.

Verilerin yapısına, dağılımına ve çıktının biçimine bağlı olarak bir eğilim modeli oluşturmak için istatistiksel analiz ya da makine öğrenmesi algoritmalarından konuya uygun yaklaşımlar uygulanabilir. Bu tezde yer alan, bir müşterinin bir sigorta ürünü üzerinde hangi seçimi yapabileceği, satın alma ve satın almama, konusu için sınıflandırma modeli yaklaşımı daha uygun olacaktır. Bu model, Python dilinde Jupyter Notebook programında yazılacaktır. Jupyter

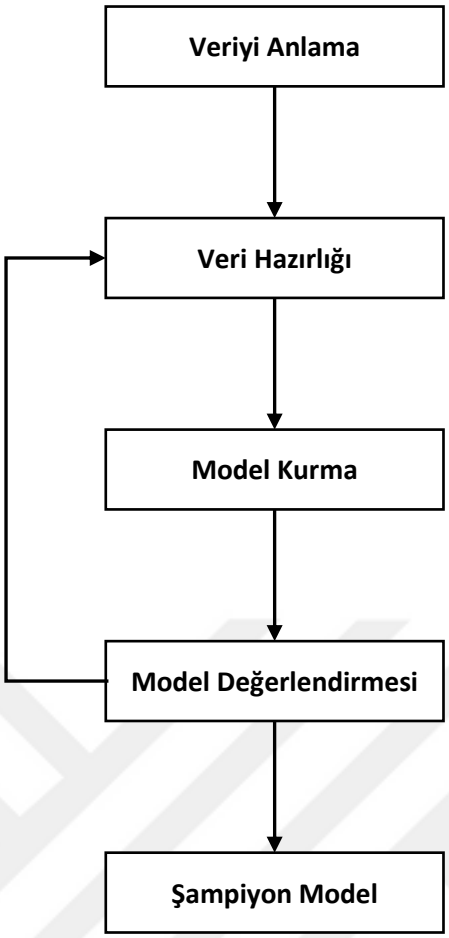
Notebook, çeşitli programlama dilleri için etkileşimli bir ortam sağlayan açık kaynak kodlu bir programdır.

Bir makine öğrenmesi algoritması kullanarak bir eğilim modeli oluşturmak için en önemli kısımlardan biri veriyi anlamak ve veri hazırlığını doğru yapmaktır. Bir makine öğrenmesi algoritması, büyük veri kümelerinde desenleri ve korelasyonları bulmak ve en iyi kararları ve tahminleri yapmak için eğitilir. Algoritmalar için veri, sayısal değerlerden ibarettir. Bu yüzden, model eğitilirken iş ihtiyacına doğru cevap verecek veri eğitilmelidir. Veri hazırlığı adımına geçmeden önce bu aşamada verinin temel özelliklerini özetlemek için keşifsel veri analizi süreci tamamlanarak veri seti net bir şekilde anlaşılacaktır.

Veri seti elde edildikten sonra ve model kurma aşamasına geçilmeden önce veri hazırlığı yapılmalıdır. Bu aşamada veri setindeki değişkenler üzerinde veri ön işlemesi (data preprocessing) yapılabilir. Bunlar, eksik değerlerin incelenmesi, aykırı değerlerin incelenmesi, değişken dönüşümünün değerlendirilmesi, değişken ölçeklendirmesinin değerlendirilmesi aşamaları olabilir. Bu tezde, veri ön işlemesi haricinde, modelin sonuçlarını iyileştirmek için özellik mühendisliği (feature engineering) uygulanarak var olan değişkenlerden yeni değişkenler yaratılacak, yeni değişkenlerin dönüşümleri yapılacak, yeni değişkenlerden uygun olanlarının seçimi yapılacaktır. Veri hazırlığının son aşamasında, model basitliği ve doğruluğu arasında bir denge bulmak amacıyla, eski ve yeni bütün değişkenler Lasso Regresyon yöntemi ile değerlendirilerek değişken elemesi yapılarak bu adım tamamlanacaktır.

Model kurma aşamasında tez konusu için belirlenen makine öğrenmesi algoritmaları, Lojistik Regresyon, Destek Vektör Makinesi (Support Vector Machine), Rastgele Orman (Random Forest), XGBoost algoritmaları, kullanılarak model çıktıları değerlendirilecektir. Model çıktılarının değerlendirme aşaması, kurulan modelin başarısı ve doğruluğu, model performansı değerlendirme metrikleri ile değerlendirilecektir.

Model değerlendirme aşamasında model performans değerlendirme metriklerine göre en iyi sonucu veren model şampiyon model seçilecektir.



Şekil 1: Model Kurma Süreç Akışı

Şekil 1'de yer alan akış şemasında, uygulanacak yönteme ait tüm süreç özetlenmektedir.

2. İSTATİSTİKSEL ANALİZ

2.1. İstatistiksel Analiz Nedir?

İstatistiksel analiz, işletmelerin verileri anlamlandırmak ve karar verme süreçlerine rehberlik etmek için kullandıkları bir araçtır. İstatistiksel analiz, kalıpları ve eğilimleri belirlemek için yerleşik ilkelere dayalı olarak verilerin toplanmasını, düzenlenmesini ve analiz edilmesini içerir. Akademi, işletme, sosyal bilimler, genetik, nüfus çalışmaları, mühendislik ve diğer birçok alanda uygulamaları olan geniş bir disiplindir.

Verilerin çeşitlerine ve uygulamalara göre kullanılabilecek farklı istatistiksel analiz yöntemleri bulunmaktadır. İstatistiksel analiz yöntemlerini kullanmak, verileri anlamlandırmaya, çözüm için modeller bulmaya, işletmelerin bulunduğu pazar için eğilimleri keşfetmeye yardımcı olabilir.

2.1.1. Korelasyon Analizi

Korelasyon analizi, iki değişken arasında bir ilişki olup olmadığını ve bu ilişkinin ne kadar güçlü olabileceğini keşfetmek için kullanılan istatistiksel bir yöntemdir. Korelasyon analizinin temel faydası, işletmelerin hangi değişkenleri daha fazla araştırmak istediklerini belirlemelerine yardımcı olması ve hızlı hipotez testine izin vermesidir.

Korelasyon analizinde, istatistiksel kavramlardan biri, korelasyon katsayısıdır. Korelasyon katsayısı, bir korelasyon analizinde yer alan değişkenler arasındaki doğrusal ilişkideki yoğunluğu hesaplamak için kullanılan ölçüm birimidir. Korelasyon katsayıları -1 ile 1 arasında bir değere sahiptir. 0 değeri, değişkenler arasında hiçbir ilişki olmadığı anlamına gelir. -1 ve 1, mükemmel bir negatif ve pozitif korelasyon olduğu anlamına gelir.

Pearson Korelasyon Katsayısı'nın hesaplanması aşağıdaki gibi formülize edilmektedir.

$$r = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2} * \sqrt{\sum(Y - \bar{Y})^2}}$$

r : Korelasyon Katsayısı
$\bar{X}$: X Değişkeninin Ortalaması
$\bar{Y}$: Y Değişkeninin Ortalaması

2.1.2. Çoklu Regresyon Analizi

Çoklu regresyon analizi, tek bir bağımlı değişken ile birden fazla bağımsız değişken arasındaki ilişkiyi analiz etmek için kullanılan istatistiksel bir yöntemdir. Çoklu regresyon analizinin amacı, tek bağımlı değişkenin değerini tahmin etmek için değerleri bilinen bağımsız değişkenleri kullanmaktır. Çoklu regresyon analizi aşağıdaki gibi formülize edilmektedir.

$$Y = b_0 + b_1 * X_1 + b_2 * X_2 + \cdots + b_n * X_n$$

Y : Bağımlı Değişken
b_0, b_n : Katsayılar
X_1, X_n : Bağımsız Değişkenler

2.1.3. Lojistik Regresyon

Lojistik Regresyon, bir girdinin belirli bir gruba ait olup olmadığını belirlemek için kullanılan istatistiksel bir analiz yöntemidir. Lojistik Regresyon, eğitilen verilere dayanarak bir olayın meydana gelme olasılığını tahmin etmeye odaklanır.

Lojistik Regresyon, hedef değişken ikili bir değişken olduğunda ve genelleştirilmiş doğrusal modeller adlı bir model sınıfına ait olduğunda uygulanan bir regresyon türüdür. Lojistik Regresyon'un amacı, bir olayın olasılığını bir dizi girdi değişkenine koşullu olarak tahmin etmektir [6]. Bir örneğin olasılığını tahmin ettikten sonra, bunun olay ya da olay olmayan olarak sınıflandırılması yapılabilir.

Lojistik Regresyon, belirli bir bağımsız değişken veri kümesine dayalı olarak, genellikle iki

değerden birinin, doğru veya yanlış, geçti veya kaldı, aldı veya almadı, sonuçlandığı durumlarda ikili sınıflandırma yapmak için kullanışlı bir yöntemdir. Girdi olarak nitel veya ordinal tahmin değişkenlerini alır ve ardından sigmoid işlevini kullanarak çıktı değerinin olasılığını ölçer. Lojistik Regresyon, hedef değişken ile açıklayıcı değişkenler arasındaki ilişkiyi inceler.

$$y = \frac{e^{b_0+b_1x}}{1 - e^{b_0+b_1x}}$$

x : Bağımsız Değişken
b_0, b_1 : Parametre veya Ağırlıklar
y : Bağımlı Değişken

Lojistik Regresyon, basit ve yapılandırılmış verilerle çalışıldığı zaman hızlı ve kesin veri analizi yapma konusunda avantajlı bir yöntemdir. Veri karmaşık bir hale geldiği zaman sonuçların doğruları azalabilir.

Lojistik Regresyon, tıp bilimi, sosyal bilimler ve ekonomi gibi ikili sonuçları tahmin etmek için çeşitli alanlarda başarıyla kullanılan yerleşik bir yöntemdir.

2.1.4. Faktör Analizi

Faktör analizi, çok sayıda değişkeni daha az sayıda faktöre indirgemek için kullanılan istatistiksel bir yöntemdir. Temel olarak iki tür faktör analizi vardır. Bunlar açıklayıcı faktör analizi ve doğrulayıcı faktör analizidir. Açıklayıcı faktör analizi, bir dizi yanıtı etkileyen yapıların doğasını keşfetmeye çalışır. Doğrulayıcı faktör analizi, belirli bir yapı grubunun yanıtları tahmin edilen şekilde etkileyip etkilemediğini test eder. Faktör analizinin amacı, aralarında ilişki olabilecek çok sayıdaki değişkenin aralarındaki ilişkileri anlamak ve daha az sayıda temel boyuta indirgemektir.

Faktör analizi, gözlemlenen ölçümler arasındaki korelasyonların veya kovaryansların modelini inceleyerek gerçekleştirilir. Yüksek korelasyonlu, pozitif ya da negatif, ölçümler muhtemelen aynı faktörlerden etkilenirken, göreceli olarak ilişkisiz olanlar muhtemelen farklı faktörlerden etkilenir.

Faktör analizinin uygulanabilmesi için verilerin bazı koşullara uygun olarak toplanmış olması gerekir. Veri hatalı ölçülmemiş olmalıdır ve aykırı değerlerden arındırılmış olmalıdır. Değişkenlerin ölçümleri en az eşit aralıklı ölçek düzeyinde yapılmış olmalıdır. Değişkenler arasındaki ilişki doğrusal olmalıdır ve birbirleriyle orta ya da yüksek düzeyde ilişkili olması gerekmektedir.

2.1.5. Temel Bileşen Analizi

Bir veri kümesindeki değişkenlerin, özelliklerin veya boyutların sayısı arttıkça, istatistiksel olarak anlamlı bir sonuç elde etmek için gereken veri miktarı artar. Bu, hesaplama süresinin artması ve makine öğrenmesi modellerinin doğruluğunun azalması gibi sorunlara yol açabilir. Bazı makine öğrenmesi algoritmaları boyutların sayısına duyarlı olabilir ve daha düşük boyutlu verilerle aynı doğruluk düzeyine ulaşmak için daha fazla veri gerektirebilir.

Temel bileşen analizi, çok boyutlu uzaydaki bir verinin daha düşük boyutlu bir uzaya izdüşümünü, varyansı maksimize edecek şekilde bulma yöntemidir [7]. Temel bileşen analizi, değişkenler arasındaki karşılıklı ilişkileri incelemek için kullanılan denetimsiz bir öğrenme algoritması yöntemidir. Temel bileşen analizinin temel amacı, hedef değişken hakkında herhangi bir ön bilgi olmaksızın değişkenler arasındaki en önemli kalıpları veya ilişkileri korurken bir veri kümesinin boyutsallığını azaltmaktır.

2.1.6. Lasso Regresyon

Değişken seçimi, özellikle değişken sayısının fazla olduğu ve bağımsız olmadığı problemler için ele alınan bir konudur. Tibshirani, değişken azaltma ve değişken seçimi için yeni bir yöntem olarak bir Lasso'yu önerdi. Lasso, amaç fonksiyonuna L1 cezası ekler [8]. L1 cezası seyrek çözümlere yol açar ve sıfır olmayan ağırlıklara sahip daha az tahmin edici kalır.

$$Loss = Error(y, y) + \lambda * \sum |w_i|$$

Lasso (Least Absolute Shrinkage and Selection Operator) Regresyon, istatistik ve makine öğrenmesinde, ortaya çıkan istatistiksel modelin tahmin doğruluğunu ve yorumlanabilirliğini geliştirmek için hem değişken seçimini hem de regularizasyonu gerçekleştiren bir regresyon

analizi yöntemidir.

Lasso Regresyon'un ana amacı özellik seçimi ve veri modellerinin düzenlenmesidir. Lasso Regresyon, gereksiz veya işe yaramayan değişkenleri eleyerek yararlı olan değişkenleri otomatik olarak seçecektir. Lasso Regresyon'da bir değişkenin elenmesi, o değişkenin katsayısını sıfıra eşitleyecektir.

Lasso Regresyon yönteminde, Ridge Regresyon'da da olduğu gibi, bir ceza terimi kullanılarak katsayılar sıfır olmaya doğru zorlanır. Ridge Regresyon ile en temel fark, Ridge Regresyon L2 düzenlemesi kullanırken, Lasso Regresyon L1 düzenlemesi kullanmaktadır. L1 düzenlemesi maliyet fonksiyonunda ağırlık parametrelerinin mutlak değerini toplayarak ceza terimini eklerken, L2 düzenlemesi maliyet fonksiyonunda ağırlıkların karesini toplar.

3. MAKİNE ÖĞRENMESİ

3.1. Makine Öğrenmesi Nedir?

Makine öğrenmesi, eldeki verilere göre öğrenen ya da performansı iyileştiren sistemler oluşturmaya odaklanan bir yapay zeka alt kümesidir. Yapay zeka, insan zekasını taklit eden sistemler veya makineler anlamına gelen kapsamlı bir terimdir. Makine öğrenmesi ve yapay zeka genellikle bir arada değerlendirilir. Kimi durumlarda birbirinin yerine kullanılır ancak aynı anlama gelmezler. Tüm makine öğrenmesi çözümleri yapay zeka iken tüm yapay zeka çözümlerinin makine öğrenmesi olmaması önemli bir ayrımdır.

Makine öğrenmesi teknikleri, geleneksel modellere göre daha kesin ve daha yüksek doğrulukta çözümler sağlayabilir (West, 2000, Tsai & Wu, 2008, Atiya, 2001, Khashman, 2010, Bellotti & Crook, 2009, Kim & Ahn, 2012, Chen & Ali, 2014). Büyük ve çok sayıda veri kaynağını yönetme ve insan önyargısı olmadan tahminler yapma kapasitesine sahiptirler ve tam otomatik uygulamanın bir parçası olarak konumlandırılabilirler.

Makine öğrenmesi teknikleri üç temel gruba ayrılmaktadır. Bunlar, denetimli makine öğrenmesi, denetimsiz makine öğrenmesi ve pekiştirmeli makine öğrenmesidir.

3.1.1. Denetimli Makine Öğrenmesi

Denetimli öğrenme, sisteme eğitim veri seti ve test veri setinin yüklenmesi, veri setinde her bir veri için gerekli etiketlenmenin yapılması ve bu sayede girdi veri seti ile çıktı veri seti arasında ilişki kurulması mantığına dayanır. Temel amaç, sonuçları bilinen veri setinden yapılan sınıflandırmadan hareketle sonuçları bilinmeyen veri setine dair etkili tahminler yapabilmektir [9].

Denetimli öğrenmede hedef değişkeni bellidir. Hedef değişken sayısal ise regresyon modeli; kategorik ise sınıflandırma modeli kullanılır.

2.1.2. Denetimsiz Makine Öğrenmesi

Denetimsiz öğrenme, herhangi bir etiket olmadan verilerdeki kalıpları bulmaya çalışan bir

makine öğrenmesi tekniğidir.

Denetimsiz öğrenmede kullanıcının sisteme herhangi bir müdahalesi yoktur. Sadece girdi verileri sisteme verilir ancak herhangi bir işaretleme yapılmaz. Sistem otomatik olarak keşifler yapar, ilişki ağını ortaya koymaya çalışır [10].

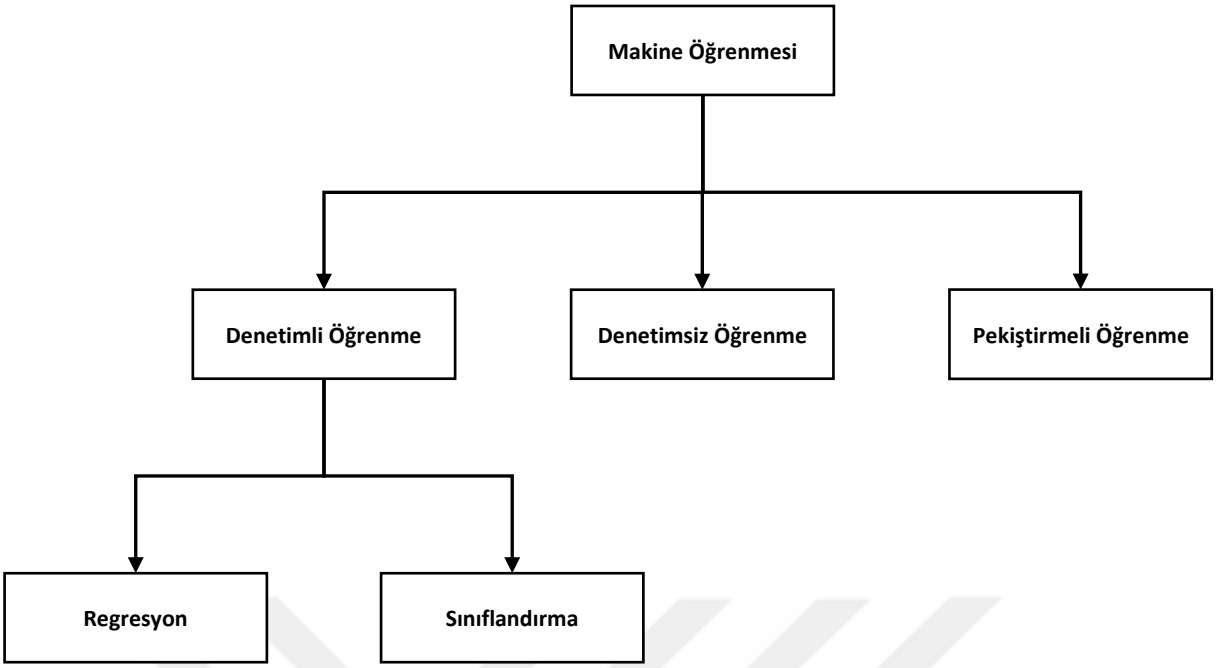
Denetimli öğrenmeden farklı olarak denetimsiz öğrenmede hedef değişken yoktur. Denetimsiz öğrenmede kullanılan iki ana yöntem ise kümeleme ve boyut azaltmaktır. Kümeleme, örneklerin homojen olarak farklı gruplar veya kümeler halinde gruplandırılmasını sağlar. Boyut azaltma, bir veri kümesindeki önemli bilgilerin iletilmesini sağlarken değişkenlerin sayısını azaltmak için kullanılır.

3.1.3. Pekiştirmeli Makine Öğrenmesi

Pekiştirmeli öğrenme, eğitim verisi olmadan makinenin çevreden alacağı tepkiler ile eğitilmesini amaçlayan bir makine öğrenmesi tekniğidir.

Temelinde denetimli öğrenmenin yer aldığı bu teknik, modelde hedef parametre çıktılarının ne derecede doğru olduğunu belirten yeni bir hedef parametresi oluşturma mantığına dayanmaktadır.

Pekiştirmeli öğrenmede yer alan fikir, makinenin çevreye bağlı olarak kendisini sürekli olarak eğitmesi ve problemleri çözmek için zenginleştirilmiş bilgilerini uygulamasıdır. Bu sürekli öğrenme süreci, insan uzmanlığının daha az katılımını sağlayarak daha fazla zaman kazandıran bir süreçtir. Denetimli öğrenmeden farklı olarak pekiştirmeli öğrenme bir çevre ile etkileşime girerek öğrenmeyi amaçlar.



Şekil 2: Makine Öğrenmesi Teknikleri

Şekil 2'de yer alan şemada, makine öğrenmesi teknikleri gösterilmektedir.

3.2. En Yaygın Makine Öğrenmesi Algoritmaları

Makine öğrenmesi algoritmaları, bir veri kümesindeki desenleri, kalıpları bulmak için uygulanan matematiksel bir yöntemlerdir.

Makine öğrenmesi yardımı ile çok çeşitli ihtiyaçlar için çözümler yaratılabilir. Bu çok çeşitli ihtiyaçları karşılamak için birbirinden farklı dayanaklara sahip makine öğrenmesi algoritmaları tasarlanmıştır. Örneğin, verileri bir kategoriye atamak için tahmine dayalı hesaplamalar kullanmak denetimli öğrenme tekniğine ait bir sınıflandırma algoritması çözümü iken veriler arasındaki benzerlik düzeyini belirleyerek veriyi birden fazla gruba bölmek denetimsiz öğrenme tekniğine ait bir kümeleme algoritması çözümüdür.

Verilerin yapısına, dağılımına ve çıktının biçimine bağlı olarak bir eğilim modeli oluşturmak için makine öğrenmesi algoritmalarından konuya uygun yaklaşımlar uygulanmalıdır. Müşteri davranışlarını, satın alma veya almama, müşteriyi kaybetme veya kaybetmeme, tahmin etmek için kurulacak eğilim modelleri için makine öğrenmesi tekniklerinden denetimli öğrenme

tekniği daha uygundur. Müşteri davranışı tahmini yapılırken hedef değişken kategorik olacağı için sınıflandırma modeli yaklaşımı tercih edilir. Bu tezde yer alan, bir müşterinin bir sigorta ürünü üzerinde hangi seçimi yapabileceği, satın alma ve satın almama, konusu için sınıflandırma modeli yaklaşımı kullanılacaktır.

Makine öğrenmesi algoritmaları içerisinde oldukça sık kullanılan ve iyi sonuçlar veren sınıflandırma modeli çözümleri bulunmaktadır. Bunlar, Lojistik Regresyon, Destek Vektör Makinesi (Support Vector Machine), Rastgele Orman (Random Forest) ve XGBoost algoritmalarıdır. Bu tezde yer alan model denemeleri ve karar verilen model, Python dilinde, Jupyter Notebook programında yazılacaktır. Jupyter Notebook, çeşitli programlama dilleri için etkileşimli bir ortam sağlayan açık kaynak kodlu bir programdır.

3.2.1. Destek Vektör Makinesi (Support Vector Machine)

Destek Vektör Makinesi (Support Vector Machine), eğitim verilerindeki herhangi bir noktadan en uzak olan iki sınıf arasında bir karar sınırı bulunan vektör uzayı tabanlı makine öğrenmesi yöntemi olarak tanımlanabilir [11].

Destek Vektör Makinesi veya Destek Vektör Ağı, denetimli öğrenme kapsamına girer. Sınıflandırma ve regresyon amaçları için kullanılırlar. Destek vektörleri, hiper düzleme yakın olan veri noktalarıdır. Veriler algoritmaya beslendiğinde, algoritma bir sınıfa veya diğerine yeni örnekler atamak için kullanılabilecek bir sınıflandırıcı oluşturur [12]. Bir Destek Vektör Makinesi, uzayda mümkün olduğu kadar geniş bir boşlukla ayrılmış noktalardan oluşur. Yeni bir örnekle karşılaşıldığında, onu karşılık gelen kategoriyle eşler.

Destek Vektör Makinesi, bir düzlem üzerine yerleştirilmiş noktaları ayırmak için bir doğru çizer. Bu doğrunun, iki sınıfının noktaları için de maksimum uzaklıkta olmasını amaçlar. Destek Vektör Makinesi, karmaşık ama küçük ve orta ölçekteki veriler için uygun bir yöntemdir.

3.2.2. Rastgele Orman (Random Forest)

Ağaç tabanlı modeller, yapılandırılmış verilerden tahminler oluşturmada en popüler olanlardan biridir. Özellikle Gradient Boosted Trees algoritmaları, örneğin XGBoost, ve Rastgele Orman

(Random Forest) algoritması birçok Kaggle yarışmasında başarılı olmuştur [13].

Rastgele Orman algoritması, iyi tahmin performansı, düşük dereceli değişken etkileşimlerini kavrama yeteneği ve kararlılığıyla bilinir [14].

Rastgele Orman, eğitim zamanında çok sayıda karar ağacı oluşturarak çalışan sınıflandırma, regresyon ve diğer görevler için bir toplu öğrenme yöntemidir. Sınıflandırma görevleri için Rastgele Orman algoritmasının çıktısı, çoğu ağaç tarafından seçilen sınıftır. Regresyon görevleri için, tek tek ağaçların ortalama veya ortalama tahmini döndürülür [15][16]. Rastgele karar ormanları, karar ağaçlarının eğitim setlerine gereğinden fazla uyma alışkanlığını düzeltir [17].

Tek bir karar ağacı genellikle yüksek bir varyans gösterir. Bu sorunu ele almak için Rastgele Orman, n adet karar ağacından oluşan bir komite kurar. Eğitim sırasında çoklu karar ağaçları oluşturan ve bireysel ağaçların tahmin ettiği sınıfların modu olan sınıfı çıkaran, sınıflandırma için bir toplu öğrenme yöntemidir. Rastgele Orman, eğitim verilerinden rastgele n örnek seçerek bir karar ağacı oluşturur. Burada n, eğitim verilerindeki toplam örnek sayısından küçüktür. Ormanda n ağaç oluşturmak için bu işlemi n kez tekrarlar.

3.2.3. XGBoost

XGBoost, ilk olarak 2016 yılında Tianqi Chen ve Carlos Guestrin tarafından makalesi yayınlanan yenilikçi bir makine öğrenmesi algoritmasıdır. Yayınlanan makale, veri bilimciler tarafından ilgiyle karşılanmış, Kaggle yarışmalarında kullanımı gün geçtikçe artmıştır. XGBoost makalesinde yer alan bilgiye göre, 2015 yılında 29 Kaggle yarışmasının 17 tanesi XGBoost ile kazanılmıştır [13]. XGBoost'un geliştirilmesi, veri bilimciler tarafından halen devam etmektedir.

XGBoost, eXtreme Gradient Boosting, C++, Java, Python, R, Julia, Perl ve Scala için düzenli bir gradyan artırma çerçevesi sağlayan açık kaynaklı bir yazılım kitaplığıdır. Proje açıklamasına göre "Ölçeklenebilir, Taşınabilir ve Dağıtılmış Gradient Boosting (GBM, GBRT, GBDT) Kitaplığı" sağlamayı amaçlamaktadır.

XGBoost algoritması, Gradient Boosting algoritmasının optimize edilmiş bir türüdür. Önceki

versiyonlara göre sağladığı avantajları, kullanımının yaygınlaşmasındaki en önemli nedendir. XGBoost algoritması, ağacı oluştururken maksimum derinlik değerini kullanır. Oluşturulan ağaç aşağı yönde aşırı ilerleme gösterirse budama gerçekleştirilir ve aşırı öğrenmenin önüne geçilir. Gradient Boosting algoritması, kayıp fonksiyonun hesaplanmasında birinci dereceden fonksiyon kullanırken, XGBoost algoritması bu hesaplamaları ikinci dereceden fonksiyonlar kullanarak gerçekleştirir. Paralel çalışma özelliği diğer algoritmalara göre sonuca daha kısa sürede ulaşılmasını sağlar.

3.3. Model Performans Metrikleri

Performans metrikleri, istatistiksel veya makine öğrenmesi modelinin performansını ve etkililiğini değerlendirmek için kullanılan nicel ölçülerdir. Bu metrikler, modelin ne kadar iyi performans gösterdiğine dair bir içgörü sağlar ve farklı model veya algoritmalar ile karşılaştırılmasına yardımcı olur.

Regresyon ve sınıflandırma modellerinde kullanılan performans metrikleri birbirinden farklıdır. Bu tezde ele alınan konu için sınıflandırma modeli yaklaşımı esas alınacaktır.

Sınıflandırma modellerinde ulaşılacak çıktının türüne bağlı olarak iki tür çıktı elde edilir. Bunlardan birincisi sınıf çıktısıdır. Sınıf çıktısı, Destek Vektör Makinesi (Support Vector Machine) ve K-En Yakın Komşu (K-Nearest Neighbors) gibi algoritmaları kapsar. İkili bir sınıflandırma modelinde çıktı 0 ya da 1 olacaktır. İkinci tür çıktı ise olasılık çıktısıdır. Günümüzde bu sınıf çıktılarını olasılık olarak veren algoritmalar bulunmaktadır. Lojistik Regresyon, Rastgele Orman (Random Forest), XGBoost gibi algoritmalar olasılık çıktıları verir. Olasılık çıktılarını sınıf çıktısına dönüştürmek için bir eşik değeri belirlenir. Elde edilen çıktılar üzerinden modelin performans ölçümü yapılır.

Karışıklık Matrisi (Confusion Matrix), tahmin edilen ve gerçek değerlerin 4 farklı kombinasyonundan oluşan bir tablodur. Doğruluk (Accuracy), Kesinlik (Precision), Hatırlama (Recall), F1 Skor (F1 Score) ve AUC-ROC Eğrisi'ni (AUC-ROC Curve) ölçmek için kullanışlıdır. Karışıklık matrisi ile ilişkili terimlerin açıklaması aşağıdaki gibidir.

- Gerçek Pozitifler (TP): Hem gerçek sınıf hem de tahmin edilen veri sınıfının 1 olduğu durumdur.

- Gerçek Negatifler (TN): Hem gerçek sınıf hem de tahmin edilen veri sınıfının 0 olduğu durumdur.
- Yanlış Pozitifler (FP): Gerçek veri noktası sınıfının 0 ve tahmin edilen veri sınıfının 1 olduğu durumdur.
- Yanlış Negatifler (FN): Gerçek veri noktası sınıfının 1 ve tahmin edilen veri sınıfının 0 olduğu durumdur.

Sınıflandırma modellerinde kullanılan performans metrikleri şunlardır:

- Doğruluk (Accuracy)
- Kesinlik (Precision)
- Hatırlama (Recall)
- F1 Skor (F1 Score)
- AUC-ROC Eğrisi (AUC-ROC Curve)

3.3.1. Doğruluk (Accuracy)

Doğruluk (Accuracy), doğru sınıflandırılan noktaların sayısı ile toplam nokta sayısı arasındaki basit orandır.

$$Doğruluk = \frac{TP + TN}{TP + FP + TN + FN}$$

3.3.2. Kesinlik (Precision)

Kesinlik (Precision), modelin doğru bir şekilde pozitif tahmin ettiği noktaların tüm pozitif tahmin ettiği noktalara oranıdır ve aşağıdaki gibi formülize edilir.

$$Kesinlik = \frac{TP}{TP + FP}$$

Kesinlik, daha çok, farklı sınıflandırma modellerini karşılaştırırken, yanlış pozitif değerlerden

kaçınmak istendiği zamanlarda bakılması iyi bir metriktir.

3.3.3. Hatırlama (Recall)

Hatırlama (Recall), modelin doğru bir şekilde pozitif tahmin ettiği noktaların gerçekten pozitif olarak ne kadar iyi sınıflandırdığını belirtir.

$$Recall = \frac{TP}{TP + FN}$$

Hatırlama, daha çok, farklı sınıflandırma modellerini karşılaştırırken, yanlış negatif değerlerden kaçınmak istendiği zamanlarda bakılması iyi bir metriktir.

3.3.4. F1 Skor (F1 Score)

F1 Skor (F1 Score), kesinlik ve hatırlama değerlerinin harmonik ortalamasıdır ve aşağıdaki gibi formülize edilir.

$$F1\ Skor = 2 * (\frac{Kesinlik * Recall}{Kesinlik + Recall})$$

F1 Skor aralığı [0, 1] olarak tanımlanır ve bir anlamda modelin ne kadar kusursuz olduğunu ifade eder. Bir model, yüksek kesinlik değerine sahip olup, aynı zamanda, düşük hatırlama değerine sahip olabilir. Bu durum doğruluğu artırsa da sınıflandırması zor olan örneklerin sınıflandırılmadığı anlaşılır. Bu yüzden, F1 Skor, kesinlik ve hatırlama değerleri arasında bir denge kurar. Bu dengenin göz önünde bulundurulduğu vakalarda bakılması iyi bir metriktir.

3.3.5. AUC-ROC Eğrisi (AUC-ROC Curve)

AUC (Eğri Altındaki Alan) ve ROC (Alıcı İşletim Karakteristiği), sınıflandırma problemleri için değişen eşik değerlerine dayanan bir performans metriğidir. ROC bir olasılık eğrisidir ve AUC ayrılabilirliği ölçer. Basit bir ifadeyle, AUC-ROC metriği modelin sınıfları ayırt etme

kabiliyeti hakkında bilgi vermektedir.

AUC-ROC'un önemi, olası tüm sınıflandırma eşiklerinde bir modelin performansının kapsamlı bir görünümünü sağlama becerisinde yatmaktadır. Gerçek pozitif oran ile yanlış pozitif oran arasındaki değiş tokuşu dikkate alır ve sınıflandırıcının iki sınıf arasında ayrım yapma yeteneğini ölçer.

AUC [0, 1] aralığındadır. Bu değer ne kadar yüksek olursa, modelin o kadar iyi olduğu anlamı ortaya çıkar.

4. UYGULAMA

4.1. Veri Seti Tanımı

Bu tezde, uygulama aşamasında kurulacak istatistiksel ve makine öğrenmesi modeli için, kullanılan veri seti, anonim bir şekilde sigortacılık sektöründen elde edilmiş olup gerçek müşteri verisidir. Veri seti içerisinde yer alan değişkenler aşağıdaki tabloda gösterilmektedir.

Tablo 1. Değişken Listesi

Değişken Adı	Açıklama	Veri Tipi
URUN_ADET_TOPLAM	Toplam ürün adeti	Nümerik
URUN_ADET_FK	Ferdi kaza ürün adeti	Nümerik
URUN_ADET_HAYAT	Hayat ürün adeti	Nümerik
URUN_ADET_EMEKLILIK	Emeklilik ürün adeti	Nümerik
MUSTERILILIK_SURE	İlk müşteri olmadan geçen süre	Nümerik
INAKTIF_SURE	İnaktif kalınan süre	Nümerik
KISA_KONUSMA	Kısa süreli konuşma sayısı	Nümerik
ORTA_KONUSMA	Orta süreli konuşma sayısı	Nümerik
UZUN_KONUSMA	Uzun süreli konuşma sayısı	Nümerik
MAX_SURE	Konuşulan maksimum süre	Nümerik
MIN_SURE	Konuşulan minimum süre	Nümerik
SON1YIL_ARANMA	Son 1 yıl içerisinde aranma sayısı	Nümerik
SON1YIL_ULASIM	Son 1 yıl içerisinde ulaşım sayısı	Nümerik
SON1YIL_SATIS	Son 1 yıl içerisinde satış sayısı	Nümerik
YAS	Müşterinin yaşı	Nümerik
CINSIYET	Müşterinin cinsiyeti	Kategorik
MEDENI_DURUM	Müşterinin medeni durumu	Kategorik
GELIR_DUZEY	Müşterinin gelir düzeyi	Kategorik
EGITIM_DUZEY	Müşterinin eğitim düzeyi	Kategorik
SEHIR_DUZEY	Müşterinin şehir düzeyi	Kategorik
TOPLAM_AKTIF_PRIM	Müşterinin aktif toplam primi	Nümerik
TOPLAM_PRIM	Müşterinin toplam primi	Nümerik
TAZMINAT_FLAG	Tazminat ödemesi alma almama	Nümerik
TARGET	Ürün satın alma almama	Kategorik

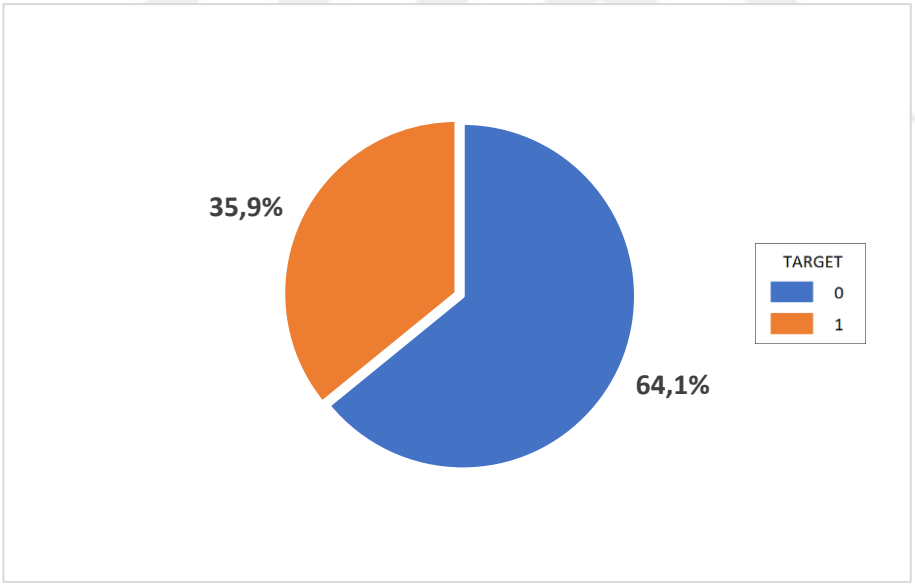
4.2. Veri Seti Analizi

Veri seti içerisinde yer alan değişkenlere ait istatistikler aşağıdaki gibi özetlenmektedir. 0 değeri müşterinin ürünü satın almamış olduğunu, 1 değeri müşterinin ürünü satın almış olduğunu göstermektedir.

Tablo 2. Hedef Değişken Değerinin Adetsel Dağılımı

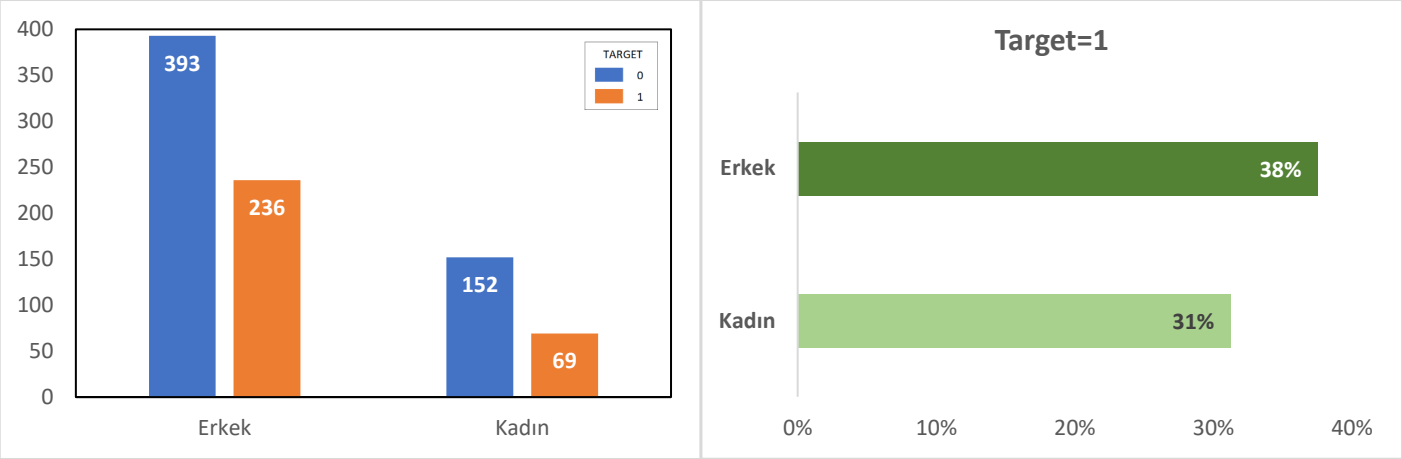
Target	Adet
0	545
1	305
Toplam	850

Veri setinde toplam 850 adet veri bulunmaktadır. Bu verilerden 305 tanesi satın alım ile sonuçlanmıştır.



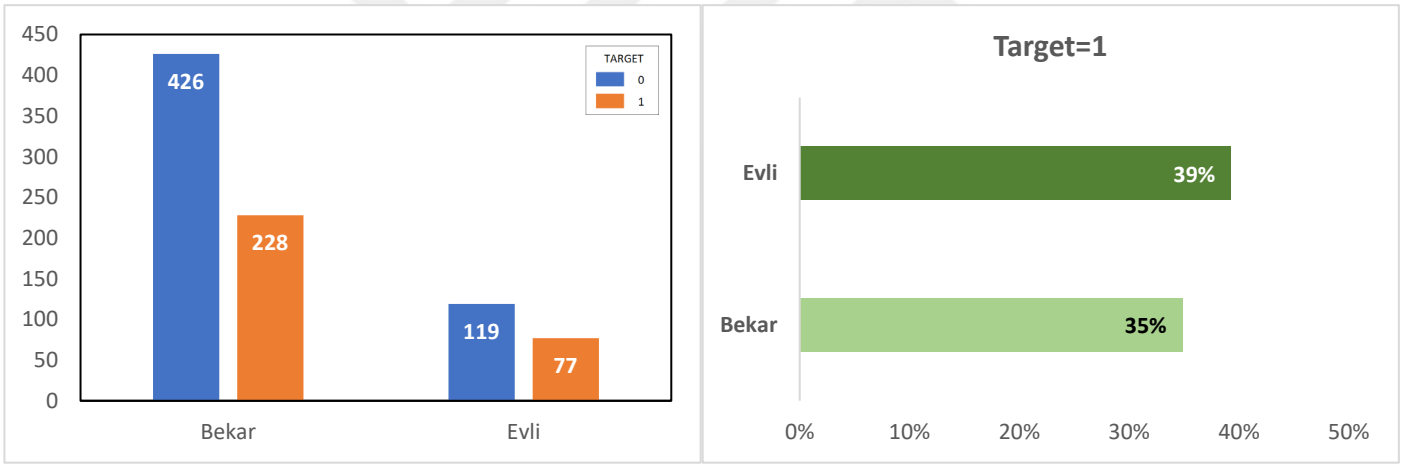
Şekil 3: Hedef Değişken Değerinin Dağılımı

Veri setinde yer alan müşterilerin %35.9'u ürünü satın almış; %64.1'i ürünü satın almamış olduğu görülmektedir.



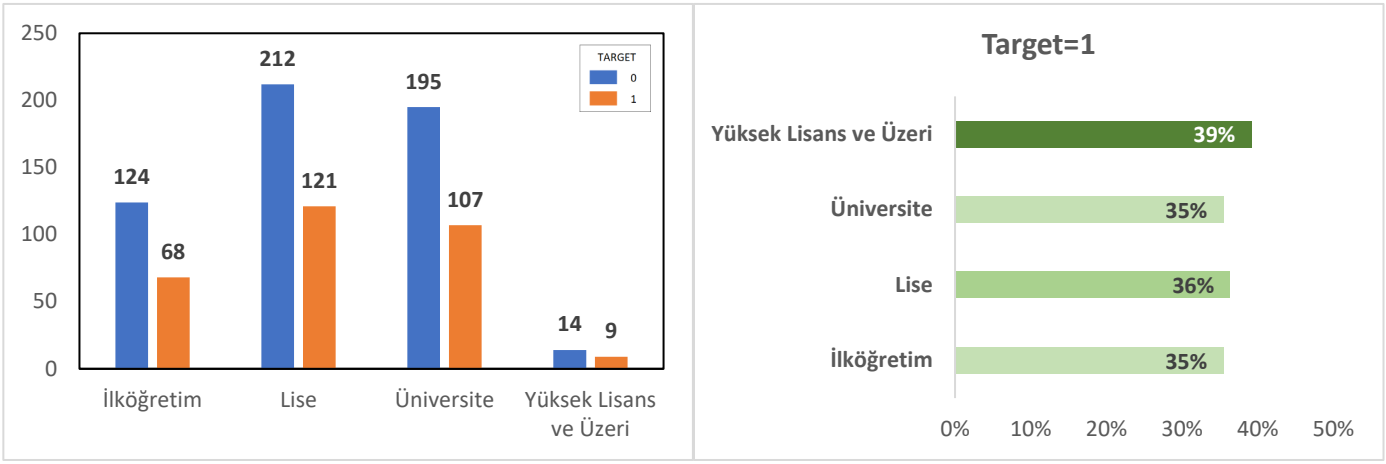
Şekil 4: Cinsiyet Dağılımı

Erkek müşterilerin kendi grubu içerisinde satın alma oranı %38; kadın müşterilerin kendi grubu içerisinde satın alma oranı %31 olarak görülmektedir.



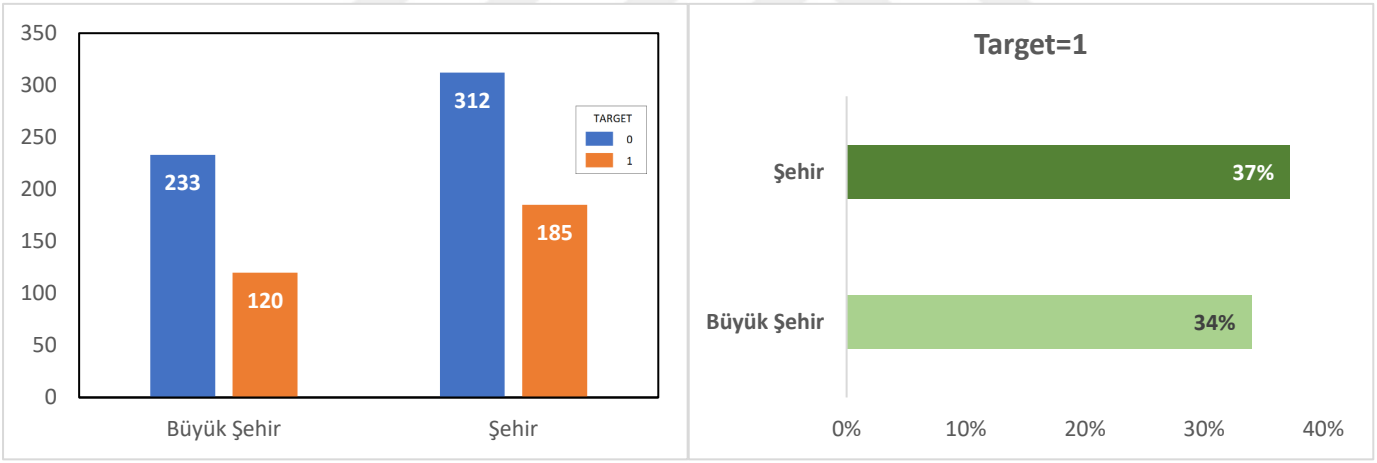
Şekil 5: Medeni Durum Dağılımı

Evli müşterilerin kendi grubu içerisinde satın alma oranı %39; bekar müşterilerin kendi grubu içerisinde satın alma oranı %35 olarak görülmektedir.



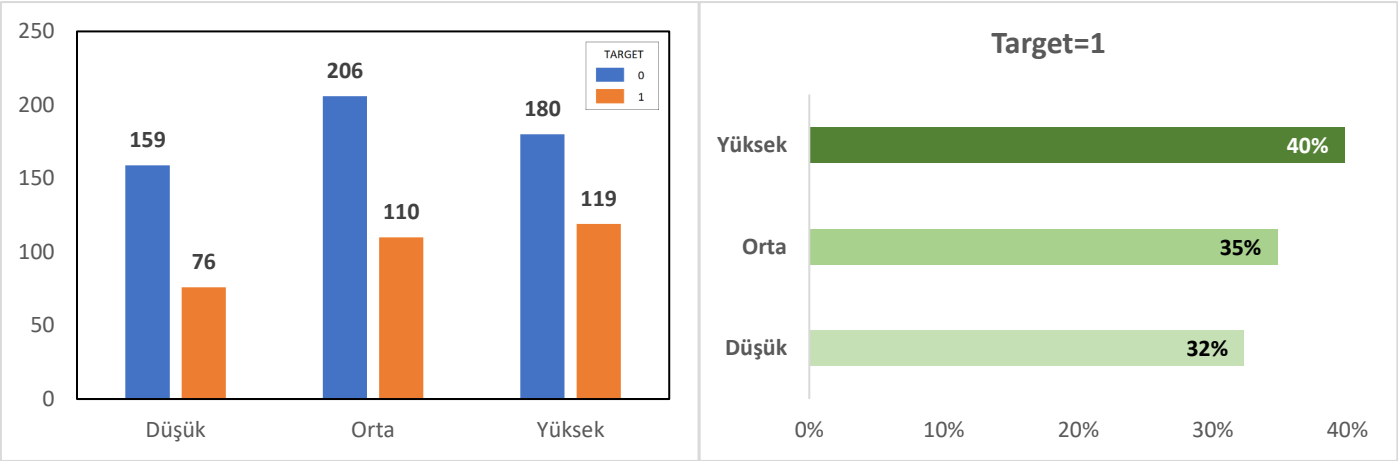
Şekil 6: Eğitim Düzeyi Dağılımı

Yüksek lisans ve üzeri mezun müşterilerin kendi grubu içerisinde satın alma oranı %39; ilköğretim mezunu müşterilerin kendi grubu içerisinde satın alma oranı %35 olarak görülmektedir.



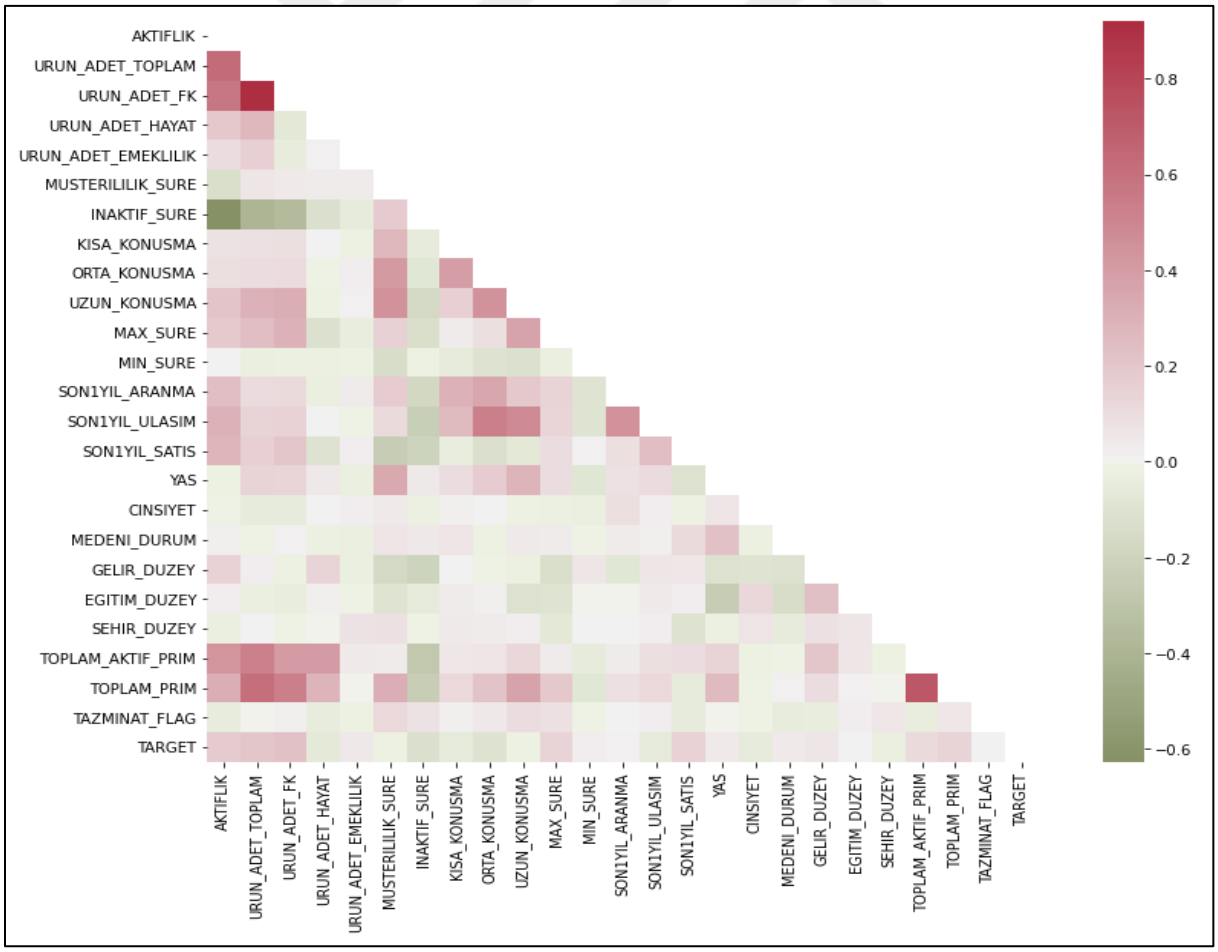
Şekil 7: Şehir Düzeyi Dağılımı

Şehirde yaşayan müşterilerin kendi grubu içerisinde satın alma oranı %37; büyük şehirde yaşayan müşterilerin kendi grubu içerisinde satın alma oranı %34 olarak görülmektedir.



Şekil 8: Gelir Düzeyi Dağılımı

Yüksek gelir düzeyine sahip müşterilerin kendi grubu içerisinde satın alma oranı %40; düşük gelir düzeyine sahip müşterilerin kendi grubu içerisinde satın alma oranı %32 olarak görülmektedir.



Şekil 9: Korelasyon Matrisi

Değişkenler arasındaki ilişkinin gözlemlenebildiği korelasyon matrisi Şekil 9'da gösterilmektedir. Buna göre en güçlü ilişki, pozitif yönde 0.92 ile urun_adet_toplam ile urun_adet_fk arasındaki ilişkidir. Değişkenlerin büyük bölümünde güçlü ilişkiler göze çarpmamaktadır.

4.3. Model Kurma Adımları

4.3.1. Veri Ön İşleme (Data Preprocessing)

Veri ön işleme, ham verilerin bilgisayarlar tarafından anlaşılabilecek ve analiz edilebilecek bir formata dönüştüren veri madenciliği ve veri modellemesi sürecinde yer alan bir adımdır.

Veri ön işleme, performansı sağlamak veya artırmak için verilerin kullanılmadan önce manipüle edilmesi veya düşürülmesi anlamına gelebilir [18]. "Çöp içeri, çöp dışarı" ifadesi özellikle veri madenciliği ve makine öğrenmesi projeleri için geçerlidir. Veri toplama veya oluşturma yöntemleri sırasında eksik veya kusurlu değerlere rastlanabilir. Bu tür problemler için dikkatli bir şekilde taranmamış verilerin analiz edilmesi yanıltıcı sonuçlar verebilir. Bu nedenle, herhangi bir analiz yapmadan önce verilerin temsili ve kalitesi her şeyden önce gelir [19]. Veri ön işleme, nihai veri işlemenin sonuçlarının yorumlanabilme biçimini etkileyebilir [20].

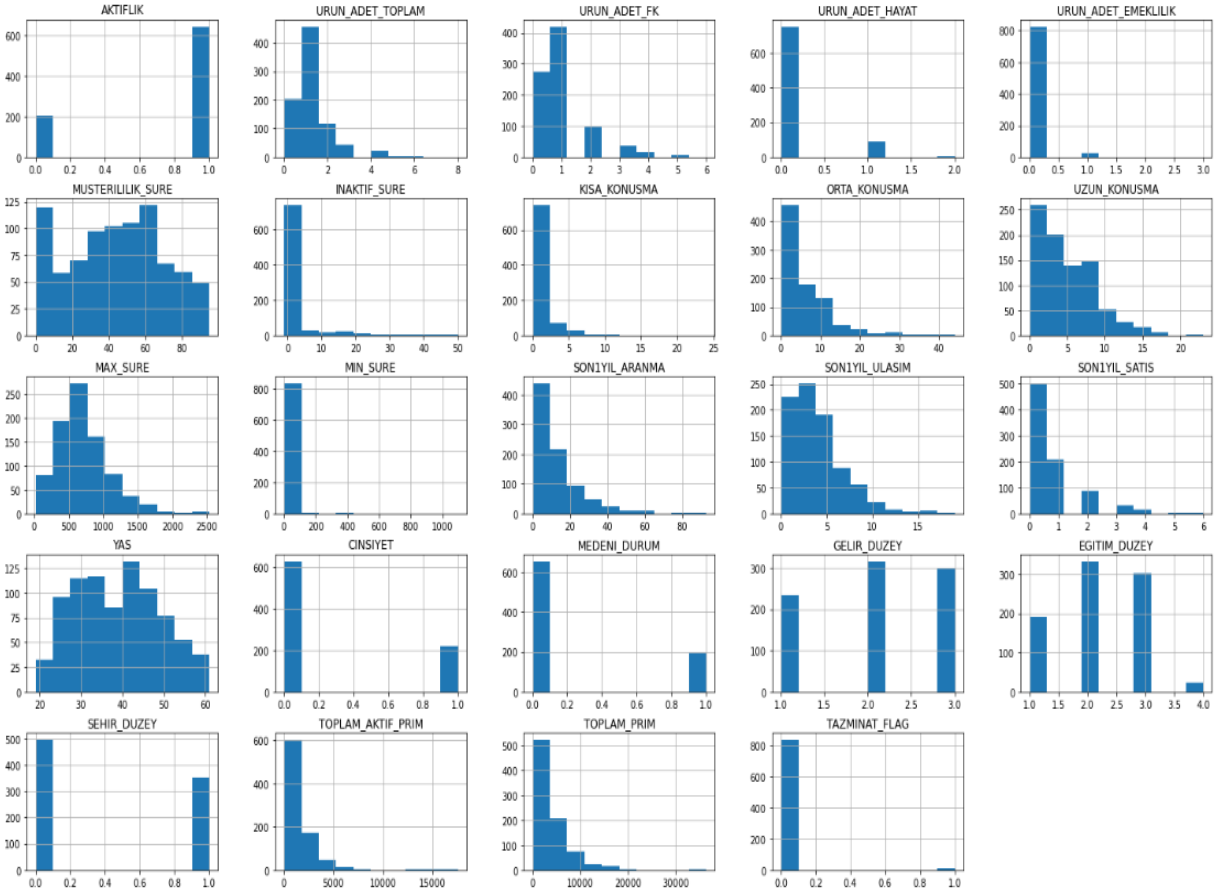
Gerçek dünyadaki veriler dağınıktır ve genellikle çeşitli insanlar, iş süreçleri ve uygulamalar tarafından oluşturulur, işlenir ve depolanır. Sonuç olarak, bir veri kümesinde ayrı ayrı alanlar eksik olabilir, manuel giriş hataları içerebilir veya aynı şeyi açıklamak için yinelenen veriler veya farklı adlar olabilir. İnsanlar genellikle iş kolunda kullandıkları verilerdeki bu sorunları tanımlayabilir ve düzeltebilir, ancak makine öğrenimini veya derin öğrenme algoritmalarını eğitmek için kullanılan verilerin otomatik olarak önceden işlenmesi gerekir. Veri ön işleme, verileri makine öğrenmesi ve diğer veri bilimi görevlerinde daha kolay ve etkili bir şekilde işlenecek bir formata dönüştürür. Teknikler, doğru sonuçlar elde etmek için genellikle makine öğrenmesi ve yapay zeka geliştirme zaman çizelgesinin en erken aşamalarında kullanılmalıdır.

Veri ön işleme adımları olarak literatürde kullanılan yöntemler, veri temizleme, veri azaltma veri dönüştürme olarak sıralanabilir.

Bu tezde kullanılan veri seti incelendiğinde, ilk göze çarpan detay olarak kategorik veri tipine

sahip değişkenlerin bulunmasıdır. Bunlar sırasıyla, cinsiyet, medeni_durum, gelir_duzey, eğitim_duzey, sehir_duzey değişkenleridir. Bu değişkenlerin makine öğrenmesi algoritmaları tarafından okunabilmesi için bu değişkenler sayısal değerlere dönüştürülmelidir. Etiket Kodlama (Label Encoding), kategorik sütunları yalnızca sayısal verileri alan makine öğrenmesi modellerine uyacak şekilde sayısal sütunlara dönüştürmek için kullanılan bir yöntemdir. Bu yöntem ile ilgili beş değişken sayısal sütunlara dönüştürülmüştür.

Bu değişkenler sayısal sütunlara dönüştürüldükten sonra veri dağılımı Şekil 10'da gösterilmektedir.

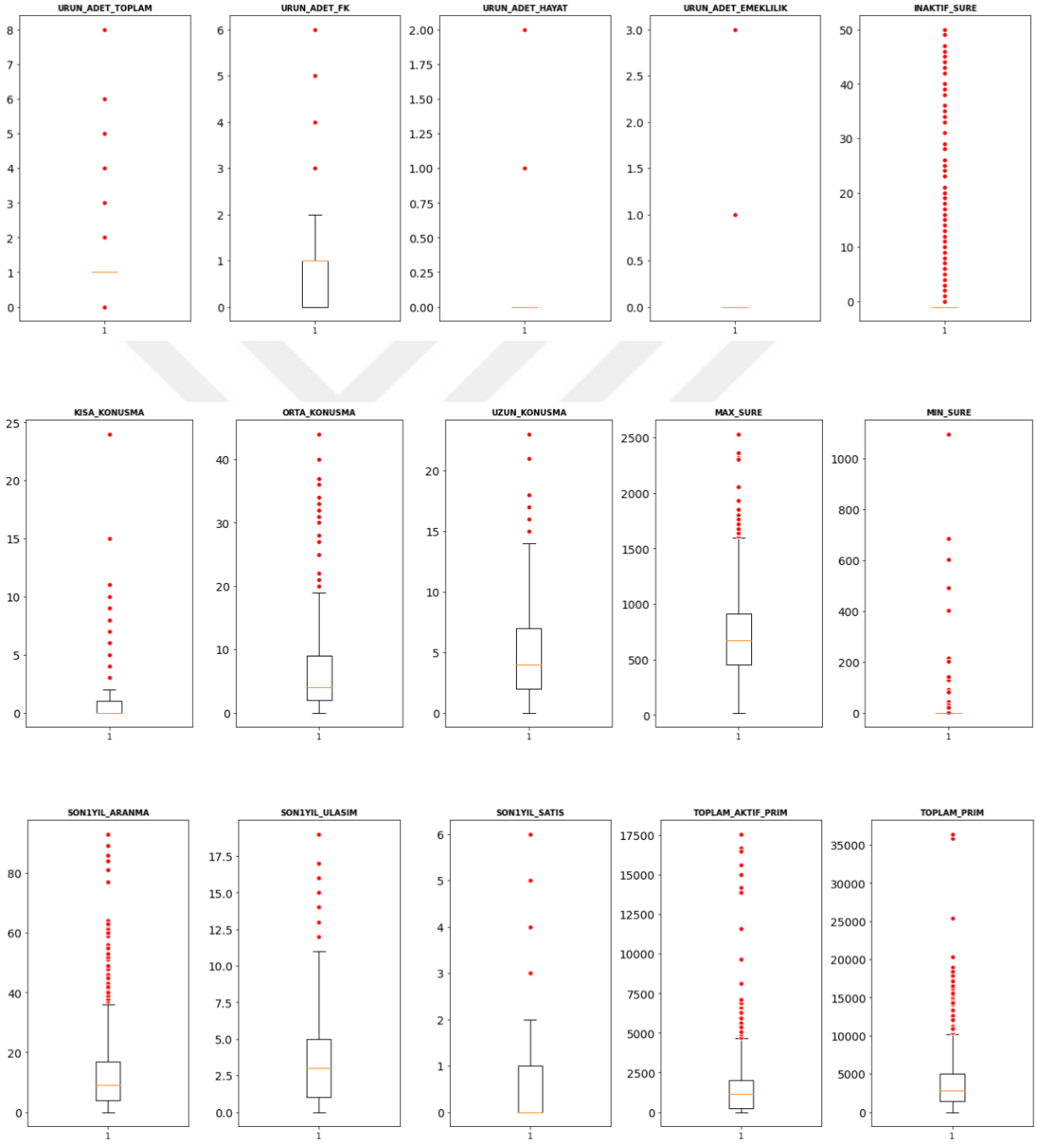


Şekil 10: Veri Dağılımı

Şekil 10'da verilen, bütün değişkenlerin yer aldığı özet bir histogram dağılım grafiklerine bakılarak, veri dağılımı incelendiğinde bazı değişkenlerin aykırı değerlere sahip olduğu görülmektedir. Bu değişkenler, urun_adet_toplam, urun_adet_fk, urun_adet_hayat, urun_adet_emeklilik, inaktif_sure, kisa_konusma, orta_konusma, uzun_konusma, max_sure, min_sure, son1yil_aranma, son1yil_ulasim, son1yıl_satis, toplam_aktif_prim, toplam_prim

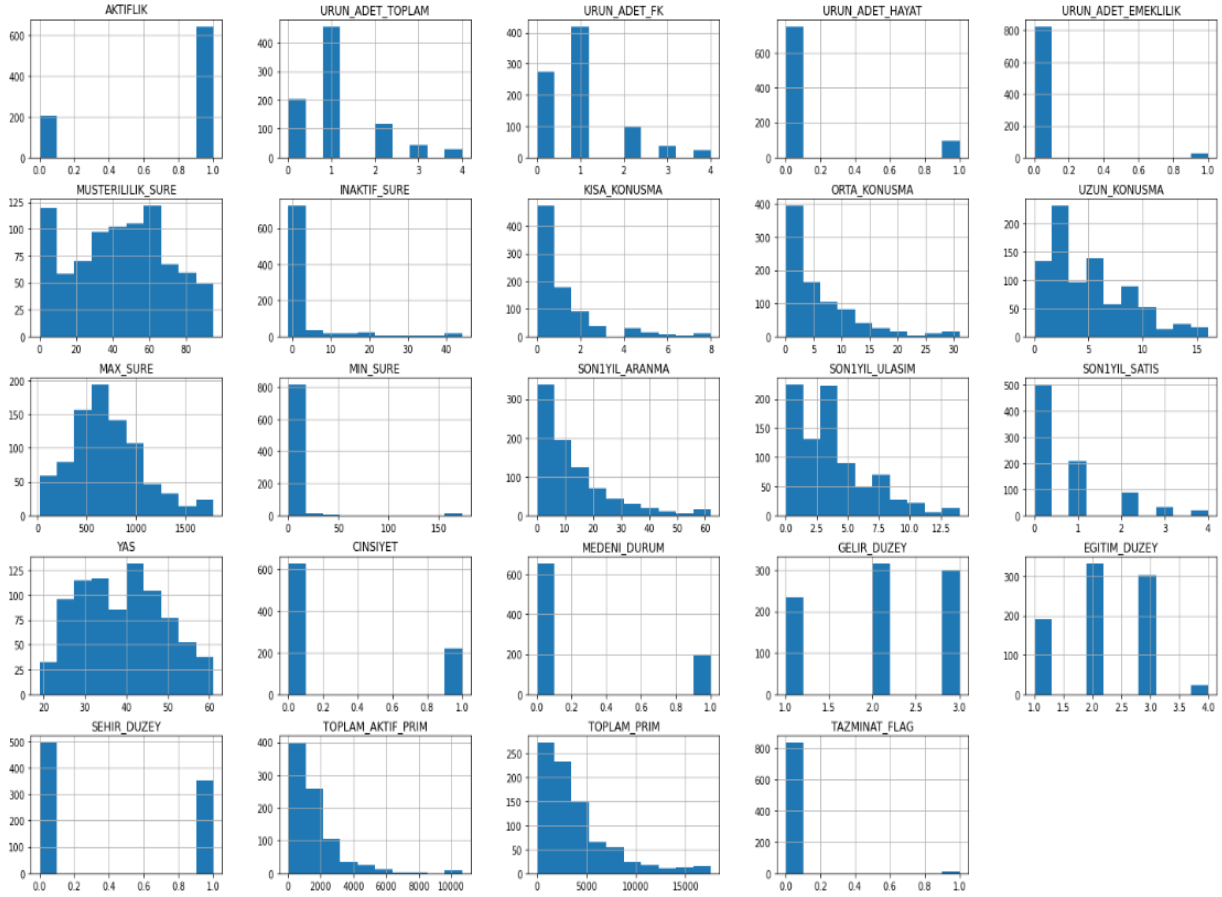
değişkenleridir. Bu değişkenler hakkında karar vermeden önce detaylı olarak tekrar incelenmesi sağlanmalıdır.

Tekrar incelenmesi gereken ilgili değişkenler için detaylı analiz Şekil 11'de gösterilmektedir.



Şekil 11: Aykırı Değer Bulunan Değişkenler

Aykırı değerler, ilgili değişkenin 99. yüzdelik dilimdeki değer ile değiştirilerek veri güncellenmiştir. Bunun sonucunda ortaya çıkan veri dağılımı Şekil 12'de gösterilmektedir.



Şekil 12: Verilerin Yeni Dağılımı

Veri ön işleme sürecinde veri üzerinde yapılan değişiklikler aşağıdaki gibi özetlenmektedir.

- Kategorik sütunların sayısal sütunlara dönüştürülmesi
- Aykırı değerlerin tespit edilmesi ve baskılanması

4.3.2. Özellik Mühendisliği (Feature Engineering)

Özellik mühendisliği veya özellik çıkarma veya özellik keşfi, ham verilerden özellikleri (karakterleri, öznitelikleri) çıkarmak için etki alanı bilgisini kullanma sürecidir [21]. Özellik mühendisliğinde temel motivasyon, bir makine öğrenmesi sürecinde yalnızca ham veri kullanmaya kıyasla, bir makine öğrenme sürecinden elde edilen sonuçların kalitesini artırmak için, ekstra özellikler kullanmaktır.

Özellik mühendisliği, makine öğrenmesi sürecini kolaylaştırmaya yardımcı olan yeni özellikler oluşturarak makine öğrenmesi algoritmalarının tahmin gücünü artırmayı ve daha iyi sonuçlar elde edilmesini amaçlar.

Bu tezde var olan değişkenler üzerinde aşağıdaki uygulamalar yapılmıştır.

- Son 1 yıl ulaşım ile son 1 yıl aranma değişkenlerinin oranı üzerinden son1yil_ulasim_oran değişkeni oluşturulmuştur.
- Son 1 yıl satış ile son 1 yıl ulaşım değişkenlerinin oranı üzerinden son1yil_satis_oran değişkeni oluşturulmuştur.
- Kısa konuşma sayısının kısa, orta ve uzun konuşma sayısına oranı üzerinden kisa_konusma_oran değişkeni oluşturulmuştur. Bu gruptaki tüm değişkenler için bu oran hesaplanarak orta_konusma_oran ve uzun_konusma_oran değişkenleri de oluşturulmuştur.
- Toplam aktif prim ile toplam prim değişkenlerinin oranı üzerinden toplam_aktif_prim_oran değişkeni oluşturulmuştur.
- Toplam aktif prim değerinin tüm veri seti içindeki ortalaması tespit edilerek bu ortalamadan yüksek ve düşük olan veriler için 1 ve 0 değerleri atanarak toplam_aktif_prim_grup değişkeni oluşturulmuştur.
- Toplam prim değerinin tüm veri seti içindeki ortalaması tespit edilerek bu ortalamadan yüksek ve düşük olan veriler için 1 ve 0 değerleri atanarak toplam_prim_grup değişkeni oluşturulmuştur.
- Ürün adetlerini gösteren değişkenler üzerinden, var ve yok ifadesi tanımlanacak şekilde, 1 ve 0 değerleri kullanılarak urun_adet_fk_flag, urun_adet_hayat_flag ve urun_adet_emeklilik_flag değişkenleri oluşturulmuştur.
- Müşterililik süresi ve inaktif süresi değişkenleri aylık değerlerden yıllık değerlere dönüştürülerek musterililik_sure_grup ve inaktif_sure_grup değişkenleri oluşturulmuştur.
- Yaş değişkeni üzerinden yaş gruplaması yapılarak yeni değişken oluşturulmuştur. İş ihtiyacını karşılayacak sezgisel bir yöntem izlenerek 18-25, 26-35, 36-45, 46-55, 56-65 ve 65+ yaş grupları oluşturulmuştur. Bu gruplar sırasıyla yas_grup_1, yas_grup_2, yas_grup_3, yas_grup_4, yas_grup_5, yas_grup_6 olarak tanımlanmıştır.
- Ham değişkenler arasından gelir_duzey ve egitim_duzey değişkenleri dummy

değişkene dönüştürülmüştür. Bu değişkenler arasından gelir_duzey_1 düşük, gelir_duzey_2 orta, gelir_duzey_3 yüksek, eğitim_duzey_1 ilköğretim, eğitim_duzey_2 lise, eğitim_duzey_3 üniversite, eğitim_duzey_4 yüksek lisans ve üzeri olarak tanımlanmıştır.

- Yeni oluşturulan değişkenler arasından musterilik_sure_grup, inaktif_sure_grup, yas_grup değişkenleri dummy değişkene dönüştürülmüştür.

Bu uygulamalar sonrasında veri setindeki toplam sütun sayısı 25'ten 60'a, toplam değişken sayısı 23'ten 58'e çıkmıştır.

4.3.3. Özellik Seçimi (Feature Selection)

Özellik seçimi, tahmine dayalı bir model geliştirirken girdi değişkenlerinin sayısını azaltma işlemidir. Özellik seçimi, makine öğrenmesinde kurulan modelin performansını etkileyen temel kavramlardan biridir. Makine öğrenmesi modellerini eğitmek için kullanılan veri özelliklerinin model performansı üzerinde etkisi vardır. Regresyon veya sınıflandırma gibi makine öğrenmesi modellerinde genellikle üzerinde çalışılacak çok fazla değişken veya özellik vardır. Öznitelik sayısı arttıkça, onları modellemenin zorluğu artabilir. Alakasız veya kısmen alakalı özellikler, model performansını olumsuz etkileyebilir.

Hem modellemenin modelin performansını iyileştirmek hem de bazı durumlarda hesaplama maliyetini azaltmak için girdi değişkenlerinin sayısını azaltmak arzu edilir. Sınıflandırma problemine pek yardımcı olmayan, alakasız özellikleri kaldırmak için özellik seçimi gerçekleştirin [22].

Bu tezde, özellik seçimi sürecinde Lasso Regresyon yöntemi kullanılmıştır. Lasso Regresyon'un ana amacı özellik seçimi ve veri modellerinin düzenlenmesidir. Özellik mühendisliği adımında artan değişken sayısı ile bütün değişkenler arasından anlamlı olan değişkenler model girdisi olanlar kullanılmak istenmiştir. Yeni değişkenlerin yanında eski değişkenlerden de anlamsız olanlar kaldırılmak istenmiştir.

Veri seti üzerinde id ve target sütunları dışında 58 adet sütun, yani modele girdi olarak sokulabilecek değişken, bulunmaktadır. $\alpha = 0.005$ anlamlılık düzeyinde, optimum sayıda değişken ile, uygulanan Lasso Regresyon yöntemi sonrasında değişken sayısı 22'ye

indirilmiştir.

Şekil 13'te Lasso Regresyon sonrası kalan adetler gösterilmektedir. Toplam değişken sayısı 58 iken seçilen değişken sayısı 22 olmuştur. Katsayıları sıfıra yakın olup sıfıra küçültülen değişken sayısı 36 olmuştur.

```
Total features: 58
Selected features: 22
```

Şekil 13: Lasso Regresyon ile Değişken Azaltma

Şekil 14'te Lasso Regresyon sonrası kalan 22 adet değişkenin isimleri gösterilmektedir.

```
Index(['URUN_ADET_FK', 'URUN_ADET_HAYAT', 'MUSTERILILIK_SURE', 'INAKTIF_SURE',
       'KISA_KONUSMA', 'ORTA_KONUSMA', 'UZUN_KONUSMA', 'MAX_SURE', 'MIN_SURE',
       'SON1YIL_ARANMA', 'SON1YIL_ULASIM', 'SON1YIL_SATIS',
       'SON1YIL_ULASIM_ORAN', 'YAS', 'YAS_GRUP_3', 'CINSIYET', 'MEDENI_DURUM',
       'GELIR_DUZEY_3', 'EGITIM_DUZEY_1', 'EGITIM_DUZEY_2',
       'TOPLAM_AKTIF_PRIM_GRUP', 'TOPLAM_PRIM_GRUP'],
      dtype='object')
```

Şekil 14: Lasso Regresyon Sonrası Kalan Değişkenler

4.3.4. Veri Ölçeklendirme

Farklı ölçeklerde ölçülen değişkenler, modeli uyumlama ve modeli öğrenme işlevine eşit derecede katkıda bulunmaz ve sonuçta bir yanlılık oluşturabilirler. Bu nedenle, bu potansiyel problemle başa çıkmak için değişkenler üzerinde ölçeklendirme yöntemi uygulanır.

Standardizasyon veya Standart Ölçeklendirme (Standard Scaler), girdi veri kümesinin işlevsellik aralığını standart hale getirmek için birçok makine öğrenmesi modeli kurulmadan önce gerçekleştirilen bir yöntemdir. Gözlenen değerlerin ortalaması 0 ve standart sapması 1 olacak şekilde değerlerin dağılımını yeniden boyutlandırmak için bu yöntem kullanılır.

Makine öğrenmesi modeli kurmadan önce veri setinde Lasso Regresyon işlemi sonrası kalan tüm değişkenler standardize edilmiştir.

4.3.5. Model Performansları ve Model Seçimi

Tezin bu aşamasında, Lojistik Regresyon, Destek Vektör Makinesi, Rastgele Orman ve XGBoost algoritmaları ile makine öğrenmesi modelleri kurulmuştur ve sonuçları karşılaştırılmıştır. Her model aynı eğitim ve test seti üzerinden kurulmuştur. Eğitim ve test seti, veri seti üzerinden alınmıştır. Eğitim seti oranı %80, test seti oranı %20'dir.

Şekil 15-16-17-18'de her bir model için sınıflandırma raporu ayrıntılı bir şekilde gösterilmektedir. Sınıflandırma raporları içerisinde modellere ait doğruluk (accuracy), precision (kesinlik), recall (hatırlama) ve F1 skor (F1 score) performans metrikleri yer almaktadır.

Lojistik Regresyon:
Model Accuracy: 73.0

	precision	recall	f1-score	support
0	0.78	0.85	0.81	117
1	0.59	0.45	0.51	53
accuracy			0.73	170
macro avg	0.68	0.65	0.66	170
weighted avg	0.72	0.73	0.72	170

Şekil 15: Lojistik Regresyon Sınıflandırma Raporu

SVM:
Model Accuracy: 75.0

	precision	recall	f1-score	support
0	0.76	0.94	0.84	117
1	0.72	0.34	0.46	53
accuracy			0.75	170
macro avg	0.74	0.64	0.65	170
weighted avg	0.75	0.75	0.72	170

Şekil 16: Destek Vektör Makinesi Sınıflandırma Raporu

Random Forest:
Model Accuracy: 77.0

	precision	recall	f1-score	support
0	0.78	0.93	0.85	117
1	0.73	0.42	0.53	53
accuracy			0.77	170
macro avg	0.76	0.67	0.69	170
weighted avg	0.76	0.77	0.75	170

Şekil 17: Rastgele Orman Sınıflandırma Raporu

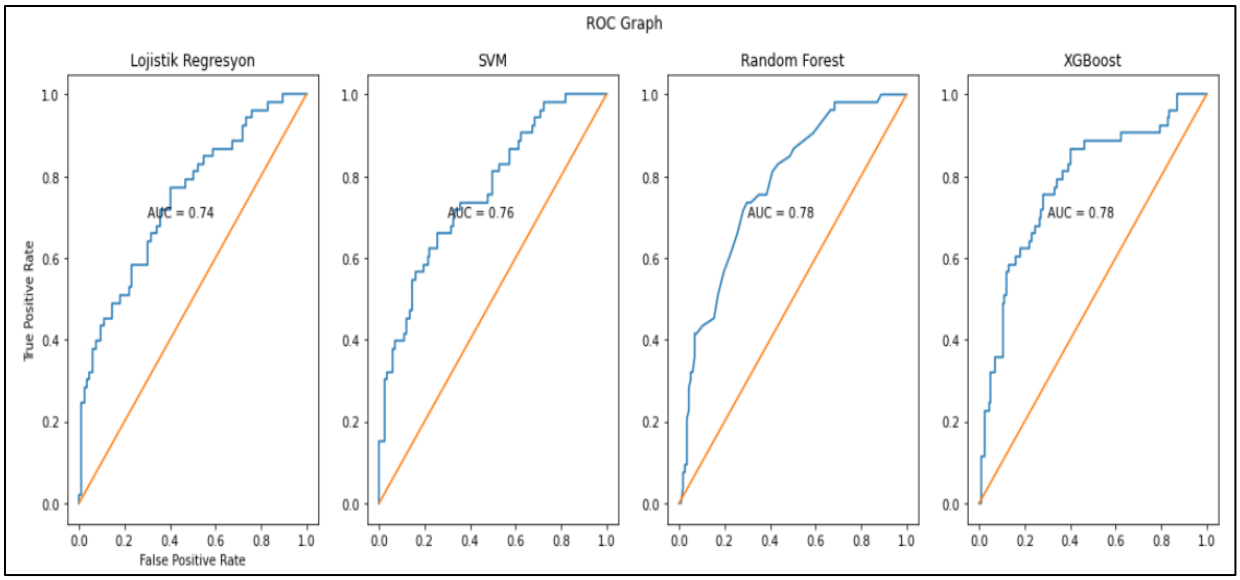
```
XGBoost:
Model Accuracy: 77.0
              precision    recall  f1-score   support

           0       0.82      0.85      0.84       117
           1       0.65      0.58      0.61        53

    accuracy                           0.77       170
   macro avg       0.73      0.72      0.73       170
weighted avg       0.77      0.77      0.77       170
```

Şekil 18: XGBoost Sınıflandırma Raporu

Sınıflandırma raporu sonuçlarına göre;

- Doğruluk oranları her modelde 70 puanın üstünde olup en iyi algoritmalar Rastgele Orman ve XGBoost algoritmaları olmuştur.
- XGBoost algoritması 1 tahmininde en iyi F1 skora ulaşmıştır. 0 tahmininde ise F1 skoru en iyi ikinci model olup geride kalmamıştır.

Modellere ait AUC-ROC eğrisi karşılaştırması Şekil 19'da gösterilmektedir.



Şekil 19: AUC-ROC Eğrisi Karşılaştırması

AUC-ROC eğrisi sonuçlarına göre;

- Lojistik Regresyon modeline ait AUC skoru 0.74; Destek Vektör Makinesi modeline ait AUC skoru 0.76; Rastgele Orman algoritmasına ait AUC skoru 0.78; XGBoost algoritmasına ait AUC skoru 0.78 olmuştur.

Sınıflandırma raporları ve AUC-ROC eğrisi incelendiğinde en başarılı model XGBoost algoritmasına ait model seçilmiştir.

Orijinal değişkenler ile elde edilen sonuçlar ile özellik mühendisliği ve özellik seçimi aşamaları sonrasında el edilen sonuçların karşılaştırılması Tablo 3'te gösterilmektedir.

Tablo 3. Model Karşılaştırması

Değişken Adı	Doğruluk	AUC-ROC	F1 Skor
Orijinal Değişkenlerle Lojistik Regresyon	0.69	0.71	0.39
Orijinal Değişkenlerle Destek Vektör Makinesi	0.68	0.72	0.18
Orijinal Değişkenlerle Rastgele Orman	0.76	0.76	0.54
Orijinal Değişkenlerle XGBoost	0.72	0.74	0.49
Özniteliklerle Lasso Lojistik Regresyon	0.73	0.74	0.51
Özniteliklerle Lasso Destek Vektör Makinesi	0.75	0.76	0.46
Özniteliklerle Lasso Rastgele Orman	0.77	0.78	0.53
Özniteliklerle Lasso XGBoost	0.77	0.78	0.61

4.3.6. Model İyileştirme

Model iyileştirme, model performansını en üst düzeye çıkarmak için hiperparametrelerin optimal değerlerini bulmaya yönelik deneysel süreçtir. Hiperparametreler, değerleri model

tarafından eğitim verilerinden tahmin edilemeyen değişkenler kümesidir. Bu değerler eğitim sürecini kontrol eder. Model iyileştirme, hiperparametre optimizasyonu olarak da bilinir.

Grid Search, hiperparametre değerlerinin farklı kombinasyonlarını dener ve en iyi sonucu veren kombinasyon, nihai optimal hiperparametre seti olarak seçilir. Seçilen modelin geliştirip geliştirilemeyeceğini görmek için GridSearchCV fonksiyonu kullanılarak modelin hiperparametreleri kontrol edilir.

Model performans metriklerine göre seçilen algoritma olan XGBoost algoritmasına ait modelin seçilen hiperparametreleri Şekil 20'de gösterilmektedir.

```
{'learning_rate': 0.1, 'n_estimators': 75}
```

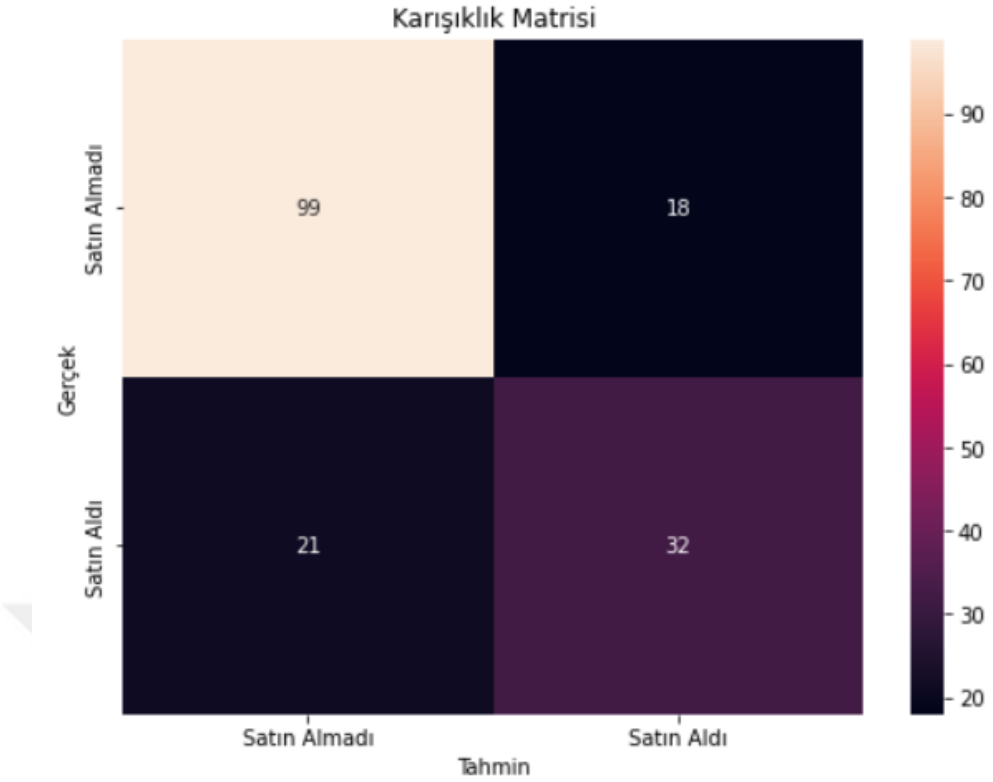
Şekil 20: XGBoost Seçilen Hiperparametreler

İyileştirilmiş modele ait sınıflandırma raporu tekrar çalıştırılmıştır. Bu model ile önceki model arasındaki fark model fonksiyonuna seçilen hiperparametrelerin eklenmesidir. İyileştirilmiş modele ait sınıflandırma raporu Şekil 21'de gösterilmektedir.

```
XGBoost Tuned:
Model Accuracy: 77.06
              precision    recall  f1-score   support

           0       0.82      0.85      0.84       117
           1       0.64      0.60      0.62        53

    accuracy                           0.77       170
   macro avg       0.73      0.72      0.73       170
weighted avg       0.77      0.77      0.77       170
```

Şekil 21: XGBoost Model İyileştirme

Elde edilen son model olan XGBoost Tuned modeline ait karışıklık matrisi çıktısı Şekil 22'de gösterilmektedir.



Şekil 22: XGBoost Karışıklık Matrisi

Makine öğrenmesinde, tahmine dayalı bir model oluştururken bir veri kümesindeki her bir özelliğin göreli önemini belirlemek için özellik önem puanları kullanılır. Özelliğin önemini hesaplamak, özellikleri tahmine katkılarına göre sıralamak için bir yöntemdir. Özellik önemi, belirli bir model için tüm girdi özellikleri için bir puan hesaplayan teknikleri ifade eder.

Elde edilen son model olan XGBoost Tuned modeli için seçilen değişkenlerin önem derecesi Tablo 4'te gösterilmektedir. Modelde yer alan 22 değişkene ait önem derecesi, feature_importances fonksiyonu ile hesaplanmıştır.

Tablo 4. Değişken Önem Derecesi

Değişken Adı	Önem Derecesi
URUN_ADET_FK	0.104066
SON1YIL_ULASIM_ORAN	0.056268
SON1YIL_SATIS	0.056192
MAX_SURE	0.054433
YAS	0.051663
MEDENI_DURUM	0.049584
YAS_GRUP_3	0.048799
ORTA_KONUSMA	0.047688
UZUN_KONUSMA	0.047350
TOPLAM_PRIM_GRUP	0.045580
MUSTERILILIK_SURE	0.043117
SON1YIL_ARANMA	0.041984
EGITIM_DUZEY_1	0.041684
KISA_KONUSMA	0.040309
TOPLAM_AKTIF_PRIM_GRUP	0.039079
INAKTIF_SURE	0.038413
GELIR_DUZEY_3	0.037984
SON1YIL_ULASIM	0.037656
CINSIYET	0.032812
URUN_ADET_HAYAT	0.030648
EGITIM_DUZEY_2	0.028587
MIN_SURE	0.026104

5. SONUÇ VE TARTIŞMA

İşletmeler için kaynakları verimli kullanmak her zaman önemli bir yer tutmaktadır. İşletmelerin istatistiksel modeller yardımıyla müşteri kitlelerinin eğilimlerini bulması ve bulduğu sonuçlar doğrultusunda doğru müşteri kitlesini hedefine alarak pazarlama faaliyetleri uygulaması bu verimliliği sağlayabilecek bir yöntemdir. Bu fikirden yola çıkarak, bu tezde, sigortacılık sektöründen toplanan gerçek bir veri ile müşterilerin bir sigorta ürününü satın alma eğiliminde etkili olan faktörler araştırılmıştır. Toplanan veri, müşteri kimliğini ifade eden 1 sütun, 1 bağımlı değişken ve 24 adet bağımsız değişkenden oluşmaktadır. Veri, detaylı bir şekilde incelenmiş, tanıtılmış, analiz edilmiştir. Makine öğrenmesi algoritmaları yardımıyla bir model kurulmadan önce veri hazırlığı aşamaları yapılmıştır. Bu aşamalarda, veri ön işlemesi yapılmış, veri içerisindeki değişkenlerden yeni değişkenler oluşturulmuş, istatistiksel yöntemler kullanılarak değişken seçimi yapılmış, veri ölçeklendirmesi yapılmıştır. Son aşamada, makine öğrenmesi algoritmaları ile bir sınıflandırma modelleri kurulmuş, şampiyon model algoritması olarak XGBoost algoritması seçilmiş ve sonuçları paylaşılmıştır.

KAYNAKLAR

[1] Ebner, K., Buhnen, T., Urbach, N., (2014). “Think Big with Big Data: Identifying Suitable Big Data Strategies in Corporate Environments”, Hawaii International Conference on System Sciences, 3748-3757.

[2] Yao-Te, T., Shu-Ching, W., Kuo-Qin, Y., (2017). “Precise Positioning of Marketing and Behavior Intentions of Location-Based Mobile Commerce in the Internet of Things”, Symmetry, Vol: 9, No: 8, 139.

[3] Nyce, C., (2007). Predictive Analytics White Paper, 1-24.

[4] Geisser, S, (1993). Predictive Inference: An Introduction. Chapman & Hall, ISBN 978-0-412-03471-8.

[5] Christie, P., (2011). Applied Analytics Using SAS Enterprise Miner. Course Notes, 113-412.

[6] Hosmer, D.W., Lemeshow, S., (1989). Applied Logistic Regression, The Multiple Logistic Regression Model, 25-37.

[7] Alpaydın, E., (2010). Introduction to Machine Learning Second Edition, MIT Press, Cambridge, 113-120.

[8] Tibshirani, R., (1996). Regression Shrinkage and Selection via the LASSO, J. R. Statist. Soc. (B) 58, 267-288.

[9] Aydın, S., Özkul, E.A., (2015). “Veri Madenciliği ve Anadolu Üniversitesi Açıköğretim Sisteminde Bir Uygulama”, Eğitim ve Öğretim Araştırmaları Dergisi, Ankara, Cilt: 4, Sayı: 3, 38.

[10] Alpaydın, E., (2010). Introduction to Machine Learning Second Edition, MIT Press, Cambridge, 2010, 4.

[11] Schlkopf, B., (2018). Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond, MIT Press, ISBN 0-262-53657-9. OCLC 1039411838.

[12] Corinna, C., Vapnik V.N., (1995). Support Vector Networks, Machine Learning, 20(3): 273-297.

[13] Chen, T., Guestrin, C., (2016). "Xgboost: A scalable tree boosting system", Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 785-794.

[14] Breiman, L., (2001). "Random Forests" Machine Learning, 45: 5–32.

[15] Ho, T.K., (1995). Random Decision Forests, Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, 278–282.

[16] Ho, T.K., (1998). "The Random Subspace Method for Constructing Decision Forests", IEEE Transactions on Pattern Analysis and Machine Intelligence, 20 (8): 832–844.

[17] Hastie, T., Tibshirani, R., Friedman, J., (2008). The Elements of Statistical Learning (2nd ed.), Springer, ISBN 0-387-95284-5.

[18] "Guide To Data Cleaning: Definition, Benefits, Components, And How To Clean Your Data". Tableau. Retrieved 2021-10-17.

[19] Pyle, D., (1999). Data Preparation for Data Mining. Morgan Kaufmann Publishers, Los Altos, California.

[20] Oliveri, P., Malegori, C., Simonetti, R., Casale, M., (2019). "The impact of signal preprocessing on the final interpretation of analytical outcomes – A tutorial". Analytica Chimica Acta, 1058: 9–17.

[21] "Machine Learning and AI via Brain Simulations". Stanford University. Retrieved 2019-08-01.

[22] Murphy, K.P., (2012). Machine learning: a probabilistic perspective, MIT Press, 86.

ÖZGEÇMİŞ

İbrahim Furkan AKYÜZ

Eğitim Geçmişi

Lisans : Yıldız Teknik Üniversitesi
Makine Fakültesi, Endüstri Mühendisliği, (2014-2020)
Yüksek Lisans : Marmara Üniversitesi
Fen Bilimleri Enstitüsü, İstatistik, (2020-)

Mesleki Geçmişi

Veri Bilimci : QNB Sigorta, (2021-)