



**T.C. İSTANBUL T CARET
 NİVERSİTESİ
FEN B LİMLERİ ENSTİT S **

**A A  TEMELL  MAKİNE   RENMES  Y NTEMLER  İLE E-T CARET
S TES  M   TER  HARCAMALARINI TAHM N EDECEK MODEL N
BEL RLENMES **

Mehmet YAL IN

**Dan şman
Do . Dr. Seda BA DATLI KALKAN**

**Y KSEK L SANS TEZ 
İSTATİSTİK ANAB L M DALI
İSTANBUL – 2021**

KABUL VE ONAY SAYFASI

Mehmet YALÇIN tarafından hazırlanan “**Ağaç Temelli Makine Öğrenmesi Yöntemleri ile E-Ticaret Sitesi Müşteri Harcamalarını Tahmin Edecek Modelin Belirlenmesi**” adlı tez çalışması 5 / 10 / 2021 tarihinde aşağıdaki jüri üyeleri önünde başarı ile savunularak İstanbul Ticaret Üniversitesi Fen Bilimleri Enstitüsü **İstatistik Anabilim Dalı**’nda **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Danışman	Doç. Dr. Seda BAĞDATLI KALKAN
	İstanbul Ticaret Üniversitesi
Jüri Üyesi	Prof. Dr. Münevver TURANLI
	İstanbul Ticaret Üniversitesi
Jüri Üyesi	Prof. Dr. Filiz KARAMAN
	Yıldız Teknik Üniversitesi

Onay Tarihi: 28.10.2021

İstanbul Ticaret Üniversitesi, Fen Bilimleri Enstitüsünün 28.10.2021 tarih ve 2021/326 numaralı Yönetim Kurulu Kararının 4. Maddesi gereğince, ders yüklerini ve tez yükümlülüğünü yerine getirdiği belirlenen “Mehmet YALÇIN” adlı öğrencinin mezun olmasına oy birliğiyle karar verilmiştir.

Prof. Dr. Necip ŞİMŞEK
Enstitü Müdürü

AKADEMİK VE ETİK KURALLARA UYGUNLUK BEYANI

İstanbul Ticaret Üniversitesi, Fen Bilimleri Enstitüsü, tez yazım kurallarına uygun olarak hazırladığım bu tez çalışmada,

- tez içindeki bütün bilgi ve belgeleri akademik kurallar çerçevesinde elde ettiğimi,
- görsel, işitsel ve yazılı tüm bilgi ve sonuçları bilimsel ahlak kurallarına uygun olarak sunduğumu,
- başkalarının eserlerinden yararlanılması durumunda ilgili eserlere bilimsel normlara uygun olarak atıfta bulunduğumu,
- atıfta bulunduğum eserlerin tümünü kaynak olarak gösterdiğimi,
- kullanılan verilerde herhangi bir tahrifat yapmadığımı,
- ve bu tezin herhangi bir bölümünü bu üniversitede veya başka bir üniversitede başka bir tez çalışması olarak sunmadığımı

beyan ederim.

28 / 10 / 2021

Mehmet YALÇIN

İÇİNDEKİLER

	Sayfa
İÇİNDEKİLER	i
ÖZET.....	iii
ABSTRACT.....	iv
TEŞEKKÜRLER	v
ŞEKİLLER.....	vi
ÇİZELGELER.....	vii
SİMGELER VE KISALTMALAR DİZİNİ	viii
1. GİRİŞ	1
2. LİTERATÜR ÖZETİ.....	4
3. E - TİCARET.....	6
3.1. E-Ticaret Kavramı	6
3.2. E-Ticaret Tarihsel Gelişimi	8
3.3. E-Ticaretin Türleri.....	10
3.3.1 Şirketler arası (B2B, Business to Business) e-ticaret	11
3.3.2 Şirket - Tüketici arası (B2C, Business to Consumer) e-ticaret.....	12
3.3.3. Tüketici - Tüketici arası (C2C, Consumer to Consumer) e-ticaret.....	12
3.3.4. İşletmeden Devlete (B2G, Business to Government) e-ticaret.....	13
4. MAKİNE ÖĞRENMESİ	14
4.1. Makine Öğrenmesinin Tarihsel Gelişimi	15
4.2. Makine Öğrenmesi Türleri	18
4.2.1. Denetimli öğrenme.....	18
4.2.2. Denetimsiz öğrenme.....	20
4.2.3. Takviyeli öğrenme	20
4.3.Makine Öğrenmesi Uygulama Aşamaları	21
4.3.1. Problemin belirlenmesi	21
4.3.2. Verilerin toplanması.....	22
4.3.3. Veri ön işleme	23
4.3.4. Model oluşturulması	24
4.3.5. Model sonuçlarının değerlendirilmesi.....	25
4.3.6. Model performansları değerlendirme ölçütleri	26
4.3.6.1. Ortalama Kare Hata (Mean Squared Error, MSE).....	27
4.3.6.2. Kök Ortalama Kare Hata (Root Mean Squared Error, RMSE).....	27
4.3.6.3. Ortalama Mutlak Hata (Mean Absolute Error, MAE)	28
5. AĞAÇ TEMELLİ MAKİNE ÖĞRENMESİ MODELLERİ.....	29
5.1. Karar Ağaçları	29
5.1.1. Karar Ağacı algoritmaları	33
5.1.1.1. CART algoritması	34
5.2. Ağaç Temelli Topluluk Öğrenme Yöntemleri (Ensemble Learning Methods)....	38
5.2.1. Rassal Ormanlar (Random Forests).....	40
5.2.2. Gradyan Artırma Makineleri (Gradient Boosting Machine)	41
5.2.3. Ekstrem Gradyan Artırma (Extreme Gradient Boost, XGBoost).....	43
5.2.4. Light Gradyan Artırma (Light Gradient Boosting, LightGBM).....	44
5.2.5. Kategorik Artırma (Category Boosting, CatBoost)	44
6. UYGULAMA	46
6.1. Problemin Tanımlanması	46
6.2. Verinin Toplanması	47
6.3. Veri Ön İşleme	48

6.4. Modelin Oluřturulması ve Sonuların Deęerlendirilmesi.....	56
7. SONU VE NERİLER.....	64
KAYNAKLAR	67
ZGEMİř	72

ÖZET

Yüksek Lisans Tezi

AĞAÇ TEMELLİ MAKİNE ÖĞRENMESİ YÖNTEMLERİ İLE E-TİCARET SİTESİ MÜŞTERİ HARCAMALARINI TAHMİN EDECEK MODELİN BELİRLENMESİ

Mehmet YALÇIN

**İstanbul Ticaret Üniversitesi
Fen Bilimleri Enstitüsü
İstatistik Anabilim Dalı**

Danışman: Doç. Dr. Seda BAĞDATLI KALKAN

2021, 72 sayfa

Teknoloji, günümüz dünyasında en hızlı gelişen ve yaygınlaşan bir kavramdır. Teknolojinin yaygınlaşması ile dijitalleşme de doğru orantılı olarak sürekli artış göstermiş olup bireylerin satın alma davranışları da bu değişime ayak uydurmuştur.

İnternet ortamına kolayca erişebilen bireyler, fiziksel mağazalardan alışveriş yapmak yerine online alışveriş yöntemine yönelmiştir. Bireylerin online alışverişe yönelmesi ile e-ticaret sektörü gelişmiştir. Genel olarak bakıldığında; küresel ölçekte satış yapabilme, kira ve çalışan ücretleri gibi şirket içindeki sabit giderlerde azalma, 7/24 satış yapabilme, online ve düşük maliyetli stok takibi gibi faktörler e-ticaretin önemini giderek arttırmaktadır. E-ticaretin en avantajlı yanlarından biri de online platformda müşteri hareketliliğinin kolayca takip edilebilmesidir. Şirketler, müşterinin her “tıklamasını” analiz ederek müşteriye tanımaya çalışmaktadır.

Bu çalışma kapsamında, makine öğrenmesi kavramı, süreçleri anlatılmakta ve uygulama olarak Google Merchandise Store (googlemerchandisestore.com) web sitesi verileri kullanılmaktadır. Makine öğrenmesi yöntemlerinden olan Ağaç Temelli Modellerden (Tree Based Methods) faydalanılarak online platforma giren müşterilerin alışverişte ne kadar harcama yapacağını tahmin edecek modeller üzerine çalışılmaktadır. Çalışmada, makine öğrenmesi uygulama adımları tamamlanarak Karar Ağaçları, Rassal Ormanlar, GBM, LightGBM, XGBoost ve CatBoost modelleri kullanılmaktadır. Modellerin başarı performansları kıyaslanarak en iyi tahmin sonuçlarını veren modeli bulmak amaçlanmış ve sonuçta en iyi tahmin performansı veren modelin hangisi olduğu belirlenmiştir. Ayrıca en iyi performansı veren model ile müşterilerin harcamalarında hangi faktörlerin daha etkili olduğu sıralanmıştır.

Anahtar Kelimeler: Ağaç temelli modeller, e-ticaret, makine öğrenmesi, topluluk öğrenmesi.

ABSTRACT

M.Sc. Thesis

DETERMINING THE MODEL TO ESTIMATE E-COMMERCE SITE CUSTOMERS' SPENDING WITH TREE-BASED MACHINE LEARNING

Mehmet YALÇIN

**Istanbul Commerce University
Graduate School of Applied and Natural Sciences
Department of Statistics**

Supervisor: Assoc. Prof. Seda BAĞDATLI KALKAN

2021, 72 pages

Technology is rapidly developing and spreading in today's world. With the spread of technology, digitalization has also increased in direct proportion, and the purchasing behavior of individuals has also kept up with this change.

Individuals who can easily access the internet have turned to online shopping instead of shopping from physical stores. With the orientation of individuals to online shopping, the e-commerce sector has developed. In addition, factors such as the ability to sell on a global scale, decrease in company fixed expenses such as rent and employee wages, 24/7 sales, online and low-cost inventory tracking are increasing the importance of e-commerce. One of the most advantageous aspects of e-commerce is that customer activity can be easily followed on the online platform. Companies try to get to know the customer by analyzing every "click" of the customer.

Within the scope of this study, the concept of machine learning and its processes are explained and Google Merchandise Store ([googlemerchandisestore.com](https://www.googlemerchandisestore.com)) website data is used as an application. Using Tree Based Methods, one of the machine learning methods, models that will estimate how much customers who enter the online platform will spend on shopping are studied. In the study, after the data was pre-processed, the amount of expenditure was estimated with Decision Tree, Random Forest, GBM, LightGBM, XGBoost ve CatBoost. By comparing the success performances of the models, it was aimed to find the model that gave the best prediction results, and as a result, it was determined which model gave the best prediction performance. In addition, the model with the best performance listed which factors are more effective in customers' spending.

Keywords: E-Commerce, ensemble learning, machine learning, tree-based models.

TEŞEKKÜR

Yüksek lisans öğrenimim boyunca desteğini ve bilgisini benden esirgemeyen, bu alanda kariyer yapma konusunda bana ilham olan değerli danışman hocam Doç. Dr. Seda BAĞDATLI KALKAN hocama teşekkür ve saygılarımı sunarım.

Yüksek lisans öğrenimim boyunca benden desteklerini esirgemeyen, Prof. Dr. Ünal Halit ÖZDEN ve Arş. Gör. Kubilay ERİŞLİK hocalarıma teşekkür ederim.

Hayatımdaki önemli kararlarımda her daim yanımda yer alıp destek olan Prof. Dr. Safa Bozkurt Coşkun hocama ve değerli dostum Öğretim Görevlisi Murat Göçer'e teşekkürlerimi sunarım.

Tüm hayatım boyunca desteklerini esirgemeyen kıymetli babam İbrahim Halil Yalçın ve değerli annem Zehra Yalçın başta olmak üzere tüm aileme teşekkürlerimi sunarım.

Mehmet YALÇIN

İSTANBUL, 2021

ŞEKİLLER

	Sayfa
Şekil 3.1. E-Ticaret ekosistemi	8
Şekil 4.1. Makine öğrenmesi yöntemleri	18
Şekil 4.2. Makine öğrenmesi uygulama adımları.....	21
Şekil 5.1. Karar Ağaçları örnek gösterim.....	30
Şekil 5.2. Rassal Ormanlar algoritması örneği.....	40
Şekil 6.1. Harcama Miktarı Histogramı	52
Şekil 6.2. Aylara göre toplam Harcama Miktarı	53
Şekil 6.3. Günlere göre toplam Harcama Miktarı	53
Şekil 6.4. Harcama Miktarı ile Kanal kırılımı.....	53
Şekil 6.5. Harcama Miktarı ile Cihaz kırılımı.....	54
Şekil 6.6. Harcama Miktarı ile İşletim Sistemi kırılımı	54
Şekil 6.7. Harcama Miktarı ile Tarayıcı kırılımı.....	55
Şekil 6.8. Korelasyon matrisi	55
Şekil 6.9. Modelde kullanılan değişkenler	56
Şekil 6.10. Değişken önem düzeyleri.....	62

ÇİZELGELER

	Sayfa
Çizelge 3.1. Türkiye’deki perakende e-ticaret rakamları ve tahmini	10
Çizelge 5.1. Sınıflandırma ve regresyon ağaçları algoritmaları	33
Çizelge 6.1. Kümelenmiş değişkenler.....	47
Çizelge 6.2. Sürekli değişkenler	50
Çizelge 6.3. Kategorik değişkenler	50
Çizelge 6.4. Kullanılan modeller, kütüphaneler ve kod blokları	58
Çizelge 6.5. Model hata metriklerinin karşılaştırılması	59
Çizelge 6.6. Parametreler ve tanımları	60
Çizelge 6.7. Parametreler ve sınanan değerleri.....	61
Çizelge 6.8. Parametreler ve bulunan değerleri	61

SİMGELER VE KISALTMALAR

B2B	Business to Business
B2C	Business to Consumer
B2G	Business to Government
C2C	Consumer to Consumer
CV	Cross Validation
DSL	Digital Subscriber Line
DTÖ	Dünya Ticaret Örgütü
EDI	Electronic Data Interchange
E-Ticaret	Elektronik Ticaret
ETKK	Türkiye Elektronik Ticaret Koordinasyon Kurulu
MAE	Mean Absolute Error
MSE	Mean Squared Error
OECD	Organisation for Economic Co-operation and Development
RMSE	Root Mean Squared Error
SSL	Secure Socket Layer

1. GİRİŞ

Teknolojinin hızlı bir şekilde gelişmesi, değişmesi ve iletişim alanının yaygınlaşması ile birlikte ülkeler arasındaki ticaret sınırları da etkisini kaybetmiştir. Sınırların ortadan kalkması ile birlikte ticaret küreselleşmiş ve toplumların tüm bireylerinin kolayca yapabileceği, kullanabileceği, faydalanabileceği bir faaliyet haline gelmiştir. Bu küreselleşmeyi sağlayan en önemli araç ise yer, zaman ve mekân kavramlarının etkisini yok eden internet olmuştur. İnternetin ortaya çıkmasından sonra ticaret alanında uluslararası rekabet artmış ve online pazarlar kurulmaya başlanmıştır.

İnternete erişimin kolaylaşması, mobil hizmetlerdeki hızlı gelişmeler ve dijital yenilikler tüketicilerin hayat tarzında ve tercihlerinde çok büyük değişimlere sebep olmuş, fiziki pazarlar ortadan kalkmıştır. İşletmeler internetin satış üzerindeki etkisini inceleyerek markalarını bu platformlara taşımışlardır. Bu durum neticesinde e-ticaretin temelleri atılmış, perakende sektörü online alışveriş sitelerinde yer almaya başlamıştır. Böylece tüm dünyada online alışverişe olan ilgi hızla artmıştır. Şüphesiz bu artışın en önemli sebepleri hem satıcıya hem de alıcıya sağladığı faydalardan kaynaklanmaktadır. Satıcıya 7/24 ulaşılabilirlik, her an her yerde alışveriş imkânı, anlık fiyat karşılaştırma fırsatı ve eve teslimat gibi birçok başlık tüketicinin hayatını kolaylaştırmaktadır. Üretici ya da satıcılar açısından ise müşterilerini izleme, onlara yönelik kampanyalar düzenleme ve pazarlama stratejileri gibi alanlarda e-ticaret (elektronik ticaret) oldukça önemli hale gelmiştir.

E-ticaret platformlarının hızlı bir şekilde büyüme göstermesiyle birlikte sektörde yer alan firmaların müşterilerini daha iyi tanıma ve analiz etme ihtiyacı oluşmuştur. Web sitesi veya uygulamalardan gelen müşterilerin davranışlarını inceleyerek buna göre aksiyon almaları sektörde var olabilmek açısından önemli bir unsur haline gelmiştir. Aynı zamanda artan rekabet, satıcı ile alıcı arasındaki etkileşimin hızlı olması bu alanda stratejilerin oluşturulmasını gerekli kılmıştır. Örneğin; müşteri segmentasyonu, müşteriye özel pazarlama fikirleri, ürünün doğru konumlandırılması, web sitesi üzerindeki reklam stratejileri, ürün-fiyat etki analizi, ürün tavsiye sistemleri, birliktelik analizleri ve ürün satın alma tahminleri gibi analizler bu alanda müşteri davranışlarını incelemek adına sıkça başvurulmuş yöntemlerin başında gelmektedir.

E-ticaret firmalarının müşterilerini izleme ve buna göre aksiyon alma ihtiyaçları belirgin bir gerçek olsa da bu stratejileri faaliyete geçirmek kolay değildir. Müşteri, ilgili e-ticaret firmasının web sitesine veya mobil uygulamasına giriş yaptığı andan itibaren veri girişi sağlanmaktadır. Her müşteri için bir kimlik oluşturulması ve müşteri bazında davranış hareketlerine dair verilerinin tutulması büyük hacimli veri akışı oluşturmaktadır. Ancak bu veri akışının takibi ve analizi güçtür. Bu nedenle verilerin tutulması, işlenmesi ve uygun analizler çıkarılması e-ticaret alanında güncel teknolojilerin kullanılmasını gerektirmektedir.

E-ticaret alanındaki verilerin hızla artması sonucu kullanılan teknolojilerin başında matematik ve istatistik temelli algoritmalarından oluşan, yazılımlar yoluyla kullanılan makine öğrenmesi yöntemleri gelmektedir. Makine öğrenmesi, bilgisayarlı sistemlerin insanlar gibi öğrenmesi fikrine dayanan matematiksel ve istatistiksel temelli algoritmaların bütünüdür. Verilerin işlenmesi, analiz edilmesi, akıllı sistemler oluşturulması, müşteri davranışlarının incelenmesi, trendlerin takibi ve veriye dayalı öngörülerin sağlanması makine öğrenmesi yöntemleri ile mümkün olmaktadır. Makine öğrenmesi yöntemleri geçmiş verilerle eğitilmekte, probleme ve veri tipine göre farklılık göstermektedir. Bu nedenle problemlere yaklaşımda uygun makine öğrenmesi yöntemlerini uygulamak oldukça önemlidir.

Bu çalışmada, müşterilerin harcama davranışları üzerine makine öğrenmesi yöntemleri ile öngörü oluşturulmaya çalışılmıştır. Makine öğrenmesi yöntemlerinden olan ağaç temelli modeller ile müşterilerin yapacakları harcamaları gerçeğe en yakın hata oranıyla bulacak modelin belirlenmesi amaçlanmıştır. Belirlenecek bu model ve kullanılacak girdiler ile yeni gelen bir müşterinin harcama miktarı tahmin edilebilecektir. Ayrıca harcama miktarı etkili olan değişkenler model yardımıyla tespit edilmiştir. Daha önce literatürde yer almayan bu çalışma ile harcamaların tahmin edilmesi şirketlerin gelecek dönemlerde öngöründe bulunarak aksiyon almalarına yardımcı olacaktır. Böylece şirket müşterilere özel kampanyalar sunabilir, müşteri davranışlarını analiz edebilir, harcama üzerindeki etkili değişkenleri kullanarak segmentasyon yapabilir ve veriye dayalı karar verme aşamalarını gerçekleştirebilir.

Çalışma kapsamında kullanılan veriler, müşterilerin web sitesi üzerindeki her davranışını izleyerek kayıtlarını tutan Google Analytics aracılığı ile global bir marka olan Google Merchandise Store e-ticaret sitesinden alınmıştır. 2 Eylül 2016 ile 04 Ocak 2017 tarihleri arasında yer alan veriler, 903653 gözlem ve 55 değişken içererek kaggle.com üzerinde Google tarafından erişim izni verilerek yayınlanmıştır. Google Merchandise Store, Google tarafından resmi olarak kullanılan bir e-ticaret sitesidir. Bu web sitesinde, giyim ve kırtasiye başta olmak üzere Google ürünleri satılmaktadır.

Çalışmada, e-ticaret ile makine öğrenmesi kavramları, ağaç temelli modeller ve modellerin uygulama adımları açıklanmıştır. Uygulama olarak e-ticaret verileri üzerinden, son zamanlarda hız ve performans bakımından oldukça sık uygulamaları yapılan ağaç temelli makine öğrenmesi modelleri kullanılmıştır. Karar Ağaçları ile bagging ve boosting modelleri olarak adlandırılan topluluk öğrenme modellerinden olan Rassal Ormanlar, GBM, LightGBM, Xgboost ve Catboost modelleri uygulanmıştır. Modeller eğitim seti üzerinde eğitilmiş ve test verisi üzerinde hata metrikleri kullanılarak tahmin skorları üzerinden kıyaslanmıştır. En başarılı tahmin sonuçlarını veren model ile müşterilerin harcamaları üzerinde etkili olan değişkenlere ulaşılmıştır. Bu alanda yapılacak çalışmalara da tez kapsamında oluşturulan modeller ve analizler ile fikir oluşturması amaçlanmıştır.

2. LİTERATÜR ÖZETİ

Çalışma kapsamında araştırılan; e-ticaret üzerine olan makine öğrenmesi çalışmaları ve ağaç temelli makine öğrenmesi modelleri ile ilgili literatür, çeşitli kaynaklar incelenerek hazırlanmıştır.

Gürsoy (2010), “Telekomünikasyon Sektöründe Müşteri Ayrılma Analizi” çalışmasında, veri madenciliğinin kullanıldığı Churn analizi ile şirket aboneliğinden ayrılma eğilimi gösteren müşteriler üzerine tahmin modelleri kurmuştur. Türkiye’de telekomünikasyon sektöründe faaliyet gösteren firmanın, ayrılma eğilimi gösteren müşterilerini belirleyerek, bu müşterilere özel pazarlama stratejileri geliştirilmesi üzerine çalışılmıştır. Lojistik Regresyon Analizi ve sınıflandırma tekniklerinden Karar Ağaçları kullanılmıştır.

Fidan (2018), “E-ticaret Müşteri Bağlılığı Gri İlişkisel Kümeleme Analizi” çalışmasında, bir eticaret sitesinden temin edilen, toplam satın alma işlem sayısı, toplam işlem tutarı, ortalama işlem tutarı, siteye giriş sayısı, şikâyet sayısı ve ürün geri iade sayısı bilgilerini içeren gerçek işlem verileri temel alınarak müşteri bağlılığı kümeleme analizi gerçekleştirmiştir. Kümeleme işlemi, Gri İlişkisel Analiz ile gerçekleştirilmiş; sonuçlara göre, analiz öncesi küme sayısı ve küme merkezleri belirlenmeksizin kümelenmenin gerçekleştirilebileceği ortaya konulmuştur. Gri İlişkisel Kümeleme analizi ile e-ticaret müşterilerinin bağlılık kümelenmeleri gerçekleştirilmiştir.

Kazan ve Karakoca (2019), “Makine Öğrenmesi ile Ürün Kategorisi Sınıflandırma” adlı çalışmalarında, bir e-ticaret sitesinden alınan ürün bilgileri etiketlenerek veri seti elde edilmiş ve bu veri seti ile çalışmayı gerçekleştirmişlerdir. Toplanan veri seti üzerine makine öğrenmesi algoritmalarıyla sınıflama tahmin modelleri kurulmuştur. Çalışma neticesinde Random Forest, Karar Ağacı, Multinomial Naive Bayes (Multinomial NB), Lojistik Regresyon, Destek Vektör Makineleri (DVM) ve Yapay Sinir Ağları (YSA) sınıflandırıcıları kullanılmış ve çıkan sonuçlar hata matrisleri gösterilerek tablolarla karşılaştırılmıştır. En başarılı sonucu, %97 başarı oranı ile Yapay Sinir Ağları modeli vermiştir.

Gürcanok (2020), “E-ticaret Satış Verileri Üzerinde Bir Veri Bilimi Vaka Çalışması” tezinde, bir e-ticaret sayfasında tüketicilerin satın aldığı ürünler üzerine veri madenciliği yöntemleri ile birliktelik analizleri yapılmıştır. Birliktelik analizi algoritmaları ile müşterilerin alışkanlıklar gibi durumlarda ürünlere olan eğilimi, ürünlere ne şekilde ihtiyaç duyabileceği incelenmiştir.

Behgounia ve Zohuri (2020), “Machine Learning Driven An E-Commerce” adlı çalışmalarında; makine öğrenmesinde yaşanan gelişmelerin e-ticaret alanındaki etkisini değerlendirmişlerdir. Çalışma kapsamında, trendlerin takip edilmesinde, promosyon fiyatlarının belirlenmesinde, yeni müşterilere ulaşmada, web sitelerinin görünümünde ve e-ticaret alanında hızlı aksiyom alma adına makine öğrenmesi tekniklerinin etkileri değerlendirilmiştir.

Özdemir (2021), “Makine Öğrenmesi Sınıflandırma Yöntemlerinin Karşılaştırılması ve E-Ticaret Üzerinde Bir Uygulama” adlı tezinde, e-ticaret web sitesi verileri kullanılarak Lojistik Regresyon, Naïve Bayes, Destek Vektör Makineleri ve Rastgele Ormanlar sınıflandırma yöntemleri ile satın alma yapan ve yapmayan müşterilere yönelik tahminleme gerçekleştirmiştir.

Karadeniz vd. (2021), “Ensemble Methods for Heart Disease Prediction” adlı çalışmalarında, kalp hastalığı riski üzerine topluluk öğrenme modelleri ile sınıflama tahmini gerçekleştirmişlerdir. Çalışma kapsamında, Spectf veri seti kullanılarak yapılan topluluk öğrenme modellerin kurulumu neticesinde yüksek başarı oranları yakalanmıştır.

Çelik ve Yelkikalan (2021), Makine Öğrenme Yöntemlerinin Depo Yönetim Süreçlerinde Uygulanması: Azure ML Studio Örneği adlı çalışmalarında; e-ticaret faaliyetlerinde önemli yer tutan, depo süreçlerinin iyileştirilmesini ele almışlardır. Depo süreçlerinde tutulan firma verilerinin, bulut teknolojileri üzerinde yer alan makine öğrenme platformlarından Azure MLStudio ortamında işlenmesi yapılmıştır. Öngörü ve talep tahmin uygulamaları; makine öğrenme yöntemleri ve yapay sinir ağları modelleriyle hesaplanmış ve model performansları birbiriyle karşılaştırılmıştır.

3. E - TİCARET

3.1. E-Ticaret Kavramı

Çeşitli mal ve hizmetlerin alım satım işlemi olarak tanımlanan geleneksel ticaret, küreselleşme ve internetin ortaya çıkmasıyla beraber yeni bir kavramı meydana getirmiştir. Ticaretin elektronik ortamda yapılmasını ifade eden bu kavram ‘e-ticaret’ olarak nitelendirilmiştir. Zaman içinde değişen ve farklılaşan e-ticaret kavramının literatürde birçok tanımlaması vardır.

Dünya Ticaret Örgütü (DTÖ) nün tanımına göre e-ticaret; “Firmaların ürünlerinin üretimini, tanıtımlarını, satış ve hizmetleri ile beraber dağıtımlarını internet ortamında yapmasıdır” (Küçükylmazlar, 2006).

Ekonomik Kalkınma ve İşbirliği Örgütü (OECD)’nün e-ticaret tanımı: “Bir e-ticaret işlemi, siparişlerin alınması veya verilmesi amacıyla özel olarak tasarlanmış sistemlerle internet ortamı üzerinden yapılan mal veya hizmetlerin satışı veya satın alınmasıdır. Mallar veya hizmetler bu sisteme göre sipariş edilir, ancak malların ve hizmetlerin ödenmesi ve sonuç olarak teslim edilmesinin online olmasına gerek yoktur.” Bir e-ticaret işlemi firmalar, hanehalkları, kişiler, hükümetler ve diğer kamu ya da özel kuruluşlar arasında olabilir (OECD, 2011).

Türkiye Elektronik Ticaret Koordinasyon Kurulu (ETKK) ise e-ticareti; “Bireyler ve firmaların açık ağ şartlarında (internet), belirli seviyede erişimci tarafından ulaşım sağlanabilen kapalı ağ koşullarında (intranet) ses, görüntü ve yazı biçimindeki sayısal bilgilerin iletilmesi, saklanması ve işlenmesi faaliyetlerine dayanan ve bir ticari faaliyet oluşturmayı hedefleyen ticari uygulamaların hepsi” olarak ifade etmektedir (Kayahan, 2016)

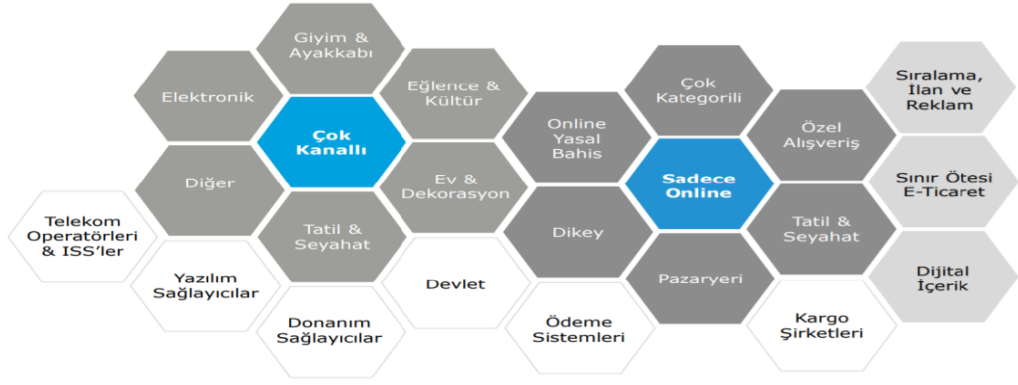
E-ticaret genel anlamda İnternet’i kullanarak satın alma ve satma manasına gelmektedir. E-ticaret, bir kuruluş ve onunla ilgilenen üçüncü taraflar arasındaki tüm elektronik olarak gerçekleşen işlemler olarak düşünülmelidir. Bu tanımla, müşteri

istekleri gibi mali olmayan işlemler de e-ticaretin bir bölümü olarak düşünülebilmektedir (Chaffey, 2002).

Elektronik ticareti geniş anlamıyla ele almak gerekirse, bilgisayar, telefon ve çeşitli iletişim ağları vasıtasıyla üretilen mal veya hizmetler için reklam, satış, sigorta, ödeme işlemlerinin hepsini içine alan süreçlerin tamamıdır. Küreselleşen dünyada ticaretin serbestleşmesi ile bilginin hızla yayılması, elektronik ticareti dünya çapında hemen herkes tarafından kullanılır hale getirmiştir (Günaydın ve Benk, 2003)

E-ticaret, iş piyasasını farklılaştırarak firmaların elindeki gücü bir anlamda müşteriye vermektedir. E-ticaret kullanan müşteriler açısından; e-ticaret kolaylık, geniş seçim yelpazesi, çeşitlilik, maliyet ve zaman tasarrufu imkânı sunmaktadır. Bunun yanında, işletmeler tarafından ise, her müşterinin ürün, hizmet talebini tek tek yerine getirme fırsatını kolaylaştırmaktadır. İşletmeler e-ticaret sayesinde tüm dünyadaki müşterilerine, satıcılarına, tedarikçilerine ve çalışanlarına ulaşma fırsatı yakalamaktadır (Pitt vd., 2002).

E-ticaret, Şekil 3.1’de gösterildiği gibi belirli kategoriler altında tanımlanmaktadır. Bu kategoriler çok kanallı, sadece online, pazaryeri, özel alışveriş çok kategorili ve dikey kavramları altında açıklanmaktadır. Hem internet hem de fiziki bir mağazanın olduğu e-ticaret modeline “Çok Kanallı” denmektedir. Yalnızca internetten satış hizmeti veren e-ticaret modeli ise “Sadece Online” olarak ifade edilmektedir. Müşterilerin ve satıcıların e-ticaret sitesi içerisinde iletişimde bulunarak satış ve mağaza açmalarına imkân veren e-ticaret modeli ise literatürde “Pazaryeri” olarak geçmektedir. Özel üyelik sistemi ile herkese açık olmayan ve sadece özel üyelere teklif sunulan e-ticaret modeli “Özel Alışveriş” olarak ifade edilmektedir. Birden fazla sayıda ürün ve hizmet veren e-ticaret modeline “Çok Kategorili”, belirli ürün veya hizmet veren sitelere de “Dikey” denmektedir.



Şekil 3.1. E-Ticaret ekosistemi (Tübisad vd., 2018)

3.2. E-Ticaretin Tarihsel Gelişimi

E-ticaret, ilk olarak 70'li yılların sonunda Amerika Birleşik Devletleri'nde kuruluş maliyeti fazla ve küçük firmalar için kurulabilmesi zor olan Elektronik Veri Değişimi (EDI) sisteminin iyileştirilmesi amacı ile ortaya çıkmıştır. 1987 yılında ise ticari verilere elektronik yoldan ulaşmanın standardı üzerinde çalışmalar yapılmıştır. Elektronik veri alışverişi, firmaların bilgisayar vasıtasıyla ile haberleşerek, ticari faaliyet yapmasını ve standart belgelerin iletilmesini sağlayan herkese açık bir ticari faaliyet sistemi olarak tanımlanmaktadır (Çetin, 2014).

1979 yılında ilk online alışveriş İngiltere'de televizyonun telefon hattı üzerinden bilgisayara bağlanmasıyla Michael Aldrich tarafından yapılmıştır. 1980 yılından itibaren piyasaya çıkan bu sistem İngiltere, İrlanda ve İspanya gibi ülkelerde kullanılmıştır. Bununla birlikte en eski online mağazalardan birisi 1992 yılında Charles M. Stack tarafından oluşturulan Book Stacks Unlimited isminde bir kitap mağazasıdır. Bu mağaza sonraki yıllarda internet ortamına boks.com adresi ile taşınmış ve satılmıştır. Ayrıca internet üzerinden ilk satış işlemi 1994 yılında gerçekleşmiştir. Güçlü bir veri şifreleme programı neticesinde gerçekleştirilen bu işlemde Sting'e ait bir müzik CD'si satılmıştır. Bu satış neticesinde internet, insanlar arasında yayılmaya başlamış ve sabit bağlantılar oluşturulmaya çalışılmıştır. İnternetin yaygınlaşmaya başlaması ile birlikte güvenlik protokolleri (HTTP) ve DSL (Digital Subscriber Line - Sayısal Abone Hattı) geliştirilme süreçleri de başlamıştır (Hermogeno, 2019).

EDI (Electronic Data Interchange - Elektronik Veri Değişimi) protokolü, ortaya çıktıktan sonra işletmeler, diğer işletmelerle dosyaları paylaşmaya başlamışlardır. Aynı zamanda, ABD ordusu, nükleer saldırıları keşfedebilmek için ARPAnet'i geliştirmişlerdir. Bu gelişmeler özellikle ARPAnet, TCP/IP'e dönüştürüldükten sonra, e-ticaret kavramının temelini oluşturmuştur. Bu doğrultuda önemli bir sorunla karşılaşılmıştır. Bu sorun; karmaşık, açık ve büyük ortama sahip olan internette güvenliğin nasıl gerçekleştirileceğidir. Çünkü internette dosyaların ve verilerin paylaşımı sağlandıktan sonra artık, gönderici, alıcı gizlilikleri ve kişisel bilgilerinin korunması gerektiği ihtiyaç haline gelmiştir. Bu ihtiyaca reaksiyon olarak da 1994 yılında Netscape 1.0 piyasaya sürülmüş, SSL (Secure Socket Layer - Güvenli Soket Katmanı) Protokolü ortaya çıkmıştır. Bu protokol ile gönderici ve alıcı verileri şifrelenmiş, bu verilere erişim imkanı kısıtlanmıştır. Bu olay, üçüncü taraf olarak kredi kartı işlem şirketlerinin bulunmalarına yol açmıştır. 11 Ağustos 1994'te New York Times gazetesi, Phil Brandenberger adlı Philadelphia'lı kişinin bilgisayarının üzerinde Sting albümünü satın aldığını raporunda söylemiştir. Bu gelişme, dünyada ilk online alışveriş olarak gerçekleşmiştir. Yenilikçi girişimciler içinde hızla yükselen trend içinde kendilerini bulundurup işlerini geliştirip, satışlarını çoğaltmaları için yeni ve çok büyük fırsat kapısı açmıştır. Bunlarla birlikte, Amazon ve eBay ortaya çıktığında e-ticarete gerçek devrim başlamıştır (Hussung, 2016).

2000'li yıllara geldiğimizde ise başta Amerika olmak üzere birçok Avrupa ülkesindeki firmalar, hizmetlerini World Wide Web (www) ortamına taşımaya başlamıştır. Bu süreçte, e-ticaret ifadesi daha fazla kullanılmaya başlanmış ve yaygınlaşmıştır. Ayrıca 2002 yılında Paypal'ın eBay tarafından satın alınması, 2004 yılında DHgate.com sitesinin kurulması, 2007 yılında business.com sitesinin 345 milyon dolara satılması, 2015 yılında ise amazon.com adresinin kurularak dünya üzerindeki e-ticaret Pazar payının yarısını oluşturmaya başlaması ve perakende sektöründe 2019 yılında 1.9 milyon dolarlık satışın olması e-ticaretin önemini göz ardı edilemez hale getirmiştir.

E-ticaretin ülkemizde kullanılmaya başlaması ise 1993 yılında gerçekleşmiştir. 12 Nisan 1993'te ODTÜ Bilgi İşlem Bölümleri'nde ilk internet bağlantısı gerçekleşmiştir. 1997 yılında ise Yüksek Bilim ve Teknoloji Konseyi tarafından e-ticaret konusu tartışılmıştır. Burada alınan en önemli kararlar, e-ticaret ağının kurulması ve Türkiye'de e-ticareti yaygınlaştıracak bir çalışma grubunun

oluşturulmasıdır. Bu karardan sonra Elektronik Ticaret Koordinasyon Komitesi (ETTK) kurularak faaliyetlerine başlamıştır (Demirdöğmez vd., 2018).

1998 yılında Kamu Net Yüksek Kurul ve Kamu-Net Teknik Konseyi'nin kurulması, 2001 yılında başbakanlık tarafından e-Türkiye çalışmalarının başlatılması, 2003 yılında Devlet Kalkınma Teşkilatı Bilgi Toplumu Departmanı'nın oluşturulması, 2007 yılında e-Dönüşüm Danışma Komitesi'nin kurulması ve 2009 yılı başında denizaşırı e-ticaret işlemlerinin Gümrük Müsteşarlığı ve Dış Ticaret Müsteşarlığı tarafından yürütülmeye başlanması e-ticaret işlemlerinin artmasına ve büyümesine katkı sağlamıştır.

Çizelge 3.1. Türkiye'deki perakende e-ticaret rakamları ve tahmini (Statista, 2020))

Yıllar	Perakende E-Ticaret Satışları (Milyon \$)	Bir Önceki Yıla Göre Değişim (%)
2017	5.702	
2018	6.458	%13
2019	8.050	%25
2020	10.657	%32
2021 (T)	13.948	%31

Yukarıda yer alan Çizelge 3.1'de, 2017-2019 yılları arasındaki Türkiye'de perakende e-ticaret hacmi gerçekleşen rakamları ve 2021 tahmini verilmiştir. Buna göre 2017 yılında 5.702 milyon dolar olan e-ticaret hacmi, 2020 yılında 10.657 milyon dolara ulaşmıştır. Özellikle 2020 yılında ortaya çıkan Covid-19 salgını ile e-ticarete yönelimler daha da artmıştır. Birçok işletme salgının kısıtlama yönünü fırsata çevirerek online alışveriş siteleri oluşturmuş ve kısıtlama altında yaşayan insanlarla yoğun talep görmüştür. Bu nedenle, hem dünyada hem Türkiye'de e-ticaret hacminin gidecek artacağı beklenmektedir.

3.3. E-Ticaretin Türleri

Temel olarak baktığımızda e-ticaret literatürde iki farklı biçimde ele alınmaktadır. Bunlar işletmeden işletmeye (B2B) ve işletmeden tüketiciye (B2C) olarak belirtilmektedir. Bunun yanında tüketiciden tüketiciye (C2C), şirketten devlete (B2G), devletten tüketiciye (G2C) gibi farklı başlıklarda literatürde yerini almıştır. E-ticaret

hacminin giderek artması türlerinde da artış meydana getirmiş, çeşitlenmiş ve farklı yeni alanların oluşmasına sebep olmuştur.

E-ticaret çeşitleri taraflarına ve faaliyetlerine göre ikiye ayrılmaktadır. Taraflarına göre e-ticarete; ticaretin kimler arasında gerçekleştiğine bakılırken, faaliyetlerine göre e-ticarete ise yapılan ticaretin hangi yollar ile gerçekleştiği önemlidir.

3.3.1 Şirketler arası (B2B, Business to Business) e-ticaret

Açılımı "Business to Business " olan B2B ve "İşletmeden işletmeye" manasına gelen bir e-ticaret şeklidir. B2B modelinde, bayilik sistemi oluşturulmakta ve müşteriler kendilerine özel giriş bilgileri ile sistem üzerinden siparişlerini vermektelerdir. Bu e-ticaret modelinde tüketiciye herhangi bir hizmet ya da ürün satışı olmamaktadır (Kavaklı, 2018).

B2B e-ticaret; doğrudan veya bir aracı vasıtasıyla, iki veya daha fazla firma arasında elektronik olarak iş faaliyetlerini kolaylaştırmakla ilgilidir. B2B e-ticaret, firmalar arasındaki her türlü etkileşimi içermektedir. Sipariş yönetimi, fatura, ödemeler ve envanter yönetimi de bu süreç içerisinde (Perera vd., 2017). B2C ile karşılaştırıldığında, çoğu durumda, B2B işlemlerinin genel hacmi, B2C işlemlerinin hacminden çok daha yüksektir (Shelly ve Rosenblatt, 2011).

Firmalar arası elektronik ticaret, kurumlar arası ilişkiyi ve işlemi kolaylaştırır. Bu tür elektronik ticaret, birbirleriyle doğrudan veya bir aracı vasıtasıyla etkileşime giren iki veya daha fazla firma gerektirir. İşletmeler arası elektronik ticaretteki araçlar, alıcıları ve satıcıları eşleştirmeye, bir anlaşmaya varabilmeye yardımcı olan, piyasa yapıcı ve rehber hizmet sağlayıcıları olabilir. B2B elektronik ticaretin ticari uygulaması; envanter yönetimi, kanal yönetimi, dağıtım yönetimi, sipariş karşılama ve teslimat ve ödeme yönetimi gibi kuruluşlar arasındaki etkileşimin neredeyse tüm yönlerini kolaylaştırmak için kullanılabilir. B2B e-ticaret; tedarikçi merkezli, alıcı merkezli veya aracı merkezli olabilir (Bhasker, 2013).

3.3.2 Şirket - Tüketici arası (B2C, Business to Consumer) e-ticaret

Açılımı "Business to Consumer" olan B2C ve "İşletmeden Tüketiciye" manasına gelen bir e-ticaret şeklidir. B2C modeli en fazla kullanılan e-ticaret kavramıdır (Kavaklı, 2018).

B2C e-ticaret, işletmeler ve tüketiciler arasındaki ticari işlemleri kolaylaştırmaktadır. B2C'de işletmeler, araçlara ihtiyaç duymadan doğrudan tüketicilere mal ve hizmet satmakta ve tüketiciler, kaliteyi ve fiyatlandırmayı esas alarak işletmeleri seçme becerilerine sahip olmaktadır. Tüketiciler daha düşük fiyatlardan ve zamanında teslimattan yararlanırken, şirketler ise doğrudan pazarlama imkânına sahip olarak daha fazla kar elde etme imkânına kavuşmaktadır (Perera vd., 2017).

B2C e-ticaret; tüketicilere daha geniş bir satıcı portföyünden daha iyi fiyatlarla çevrimiçi ürünleri inceleme, seçme ve satın alma fırsatı sağlamaktadır. Bu tür işlemlerde, birbiriyle etkileşimde bulunan iki veya daha fazla birim, bir satış yapan işletme ve bir tüketiciyi içermektedir. Satış yapan işletmeler, verilen fiyatlarla bir dizi ürün, indirimler ve nakliye ve teslimat seçenekleri sunmaktadırlar. Bu tür elektronik ticarete; satıcılar ile tüketiciler, doğrudan etkili pazarlama, özelleştirmeler ve çevrimiçi müşteri hizmetleri için artan fırsatlarla, dünyanın herhangi bir yerinden 7 gün 24 saat boyunca alışveriş erişilebilirliğinden faydalanmaktadırlar (Bhasker, 2013). Farklı birçok alanda kendini gösteren bu sitelere örnek olarak; Amazon.com, n11.com, Trendyol.com, Hepsiburada.com ve Yemeksepeti.com gösterilebilir.

İşletmeden tüketiciye e-ticaret siteleri, günümüzde internet platformunda en çok var olan e-ticaret türüdür. Birçok sektörle ilgili e-ticaret siteleri internet ortamında varlığını sürdürmektedir.

3.3.3. Tüketici - Tüketici arası (C2C, Consumer to Consumer) e-ticaret

C2C, tüketiciden tüketiciye işlem yapılan bir e-ticaret şeklidir. Bir tüketicinin başka bir tüketiciye çevrimiçi olarak mal veya hizmet satmasıdır (Turban vd., 2015). Bu satış, genellikle, işlemin detaylarının oluşturulmasına yardımcı olan üçüncü taraf bir

site tarafından düzenlenir. Gittigidiyor.com, letgo.com, eBay.com vb. siteler buna örnek olarak gösterilebilmektedir.

Tüketiciden tüketiciye e-ticaretin temel amacı, herhangi bir aracı kuruluşu devreye sokmadan, tüketicilerin direkt olarak diğer tüketicilere satış yapmalarına imkân sağlamaktır. Bu sayede, satıcı karını arttırırken, alıcı ise piyasa şartlarının altında daha iyi bir fiyata ürün sahibi olacaktır. Kısaca, satıcılar; online firmaların yarattığı sanal pazarda ürünlerini ve hizmetlerini satışa sunmakta, alıcılar da satışa sunulan ürün ve hizmetler arasında seçim yaparak elektronik ortamda iş yapmaktadırlar.

3.3.4. İşletmeden Devlete (B2G, Business to Government) e-ticaret

İşletmeden devlete (B2G); devlet veya devlet kurumlarına ürün, hizmet veya bilgi satan işletmeler anlamına gelen bir işletme şeklidir.

Devletler, ekonomiyi yönlendirmede, yönetmede ve düzenlemede etkin bir rol oynamaktadır. E-ticaretin ortaya çıkışı, oluşacak yeni talebi hükümetlerin özgün işlevlerine yönlendirmektedir. Hükümetler bir yandan e-pazarı etkili bir biçimde yönetmeli, e-devlet tarafından işletmelere ve halka daha iyi hizmet vermelidir. Diğer taraftan, ekonomideki “büyük müşteriler” olarak hükümetler, e-ticareti benimsemeli ve devlet ihalesine elektronik ihale daveti yoluyla etkili bir yol sunmalıdır (Qin, 2009).

İnternet üzerinde B2G olarak adlandırılan işletmeden devlete satış kavramında; işletmeler ile devlet kuruluşları bilgi alışverişi ve karşılıklı olarak web dışı ortamlardakinden daha verimli bir yapıda iş yapabilmeleri için merkezi bir web sitesini kullanmaktadırlar. Örneğin, B2G hizmeti sunan bir web sitesi, devletin bir ya da daha çok kademesine başvuruların ve vergi formlarının tek bir noktadan gönderilmesi; doldurulmuş formların ve ödemelerin teslim edilmesi, kurumsal bilgilerin güncellenmesi ve özel sorulara yanıt alınması gibi konularda kolaylık sağlamaktadır (Yamamoto, 2013).

B2G e-ticaret siteleri ile birçok devlet; hem işlem maliyetlerini azaltmak hem de işlemlerin daha hızlı ve hatasız yapılabilmesini sağlamak için birçok uygulamayı elektronik ortama hızlı bir şekilde taşımaya çalışmaktadır (Ekici, 2013).

4. MAKİNE ÖĞRENMESİ

Günümüz teknoloji dünyasında veriye dayalı karar süreçlerin güncellenmesi ile artan hesaplama ihtiyacı, algoritma ve yazılımlar üzerinden sağlanmaktadır. Artan hesaplama ihtiyacının geleneksel metotlar ile karşılanmaması üzerine günümüz teknoloji dünyasında, hesaplama ve analiz ihtiyacı akıllı sistemler üzerinden gerçekleştirilmektedir. Bu sistemlerin başında da son dönemlerde sıkça kullanılan makine öğrenmesi algoritmaları gelmektedir.

Makine öğrenmesi kavramı, günümüzde sıkça kullanılsa da esasında kavram olarak ortaya atılması 1950’li yıllara dayanmaktadır. Makinelerin bilgiyi işleyerek hesaplama gücünden faydalanma fikri, o yıllarda ortaya atılmıştır.

Makine öğrenmesi kavramı sıkça yapay zeka kavramı ile karıştırılmaktadır. Ancak makine öğrenmesi, özelinde makinelerin insani öğrenme süreçlerini temel alan yapay zeka kavramının alt koludur. Yapay zeka, insanlardaki zeka işleyişini temel alarak öğrenme süreçlerini bilgisayar modelleri yardımıyla inceleyip bunları formal hale getirdikten sonra yapay sistemlere uygulamayı amaçlayan bir araştırma alanıdır (Russel ve Norvig, 1995).

Yapay zeka, akıllı sistemler oluşturmak için kullanılan bilgisayar sistemleridir. Yapay zeka, insani düşünme yeteneğini ve davranışını taklit edebilen akıllı makine sistemlerin tümüne verilen genel addır. Yüz tanıma, görüntü tanıma ve okuma, chatbotlar, sesli çevirmenler, otomatik dil işleme süreçleri, spam filtrelemelerinin yapılması, kişisel sanal asistanlar ve veriye dayalı öngörülerin yapılması gibi birçok alanda kullanılan yapıların bütünüdür. Makine öğrenmesi ise bahsedilen örneklerin uygulanması, uygun algoritmalar kullanılması ve algoritmaların performansı ile ilgilidir.

Makine öğrenmesi çeşitli algoritma, modelleme, istatistik ve teknikleri geliştirmek için bilgisayarlı sistemlerin insanlar gibi öğrenmesi fikrine dayanan bilimsel bir çalışma alanıdır (Alpaydın, 2010). Makineler denilen; yazılım ve istatistiksel algoritmalarından oluşan bu makine öğrenmesi sistemi, verilerden öğrenmeyi gerçekleştirmektedir. En temel mantık ile açıklamak gerekirse, geçmiş verilerden yola çıkarak hem tahmin

modelleri oluşturmak hem de veri seti içindeki desenleri, kırılımları ve farklılıkları tespit etmek için kullanılmaktadır. Veriler üzerine çalışılırken istatistiksel ve matematiksel modeller kullanarak uygulama aşamasında yazılımlardan faydalanılarak veriye dayalı öğrenme süreçleri gerçekleştirilmektedir.

Günümüzde, verilerin hacim büyüklüğünün artması ile donanımsal gerekliliğin sonucunda, 2010'lu yıllarda makine öğrenmesinin bir alt dalı olarak derin öğrenme kavramı da geliştirilmiştir. Yapay sinir ağları tabanlı bu öğrenme çeşidi; yüksek performans ve işlem gücü gerektiren görüntü, ses ve dil işleme alanlarında yoğunlukla kullanılmaktadır.

4.1. Makine Öğrenmesinin Tarihsel Gelişimi

1949 yılında Edmunda Berkeley, Dev Brains (Düşünen Makineler) adlı eserinde ilk olarak makinelerin düşünebileceğini, beyne benzer kompleks yapılar ile hesaplama yapabileceği ifadelerinde bulunmuştur.

1950 yılında, İngiliz matematikçi Alan Turing, Computing Machinery and Intelligence adlı makalesiyle makinelerin düşünebilmesine dair fikirlerini paylaşmıştır. Literatüre Turing Testi olarak geçen, “makinelerin düşünüp düşünemeyeceğini” test eden sorgusuyla bu alanda önemli bir yer almıştır.

1952 yılında Arthur Samuel, bilgisayarda ilk satranç oyununu ve ilk kendi kendine öğrenebilen bilgisayar programını geliştirmiştir.

1957 yılında, Frank Rosenblatt, Cornell Havacılık Laboratuvarı için 1957'de “Perceptron: Algılayan ve Tanıyan bir Otomaton (The Perceptron: A Perceiving and Recognizing Automaton)” başlıklı yazısını yayınlamıştır. Burada “optik bir beynin algısal süreçlerine yakın bir şekilde, optik, elektriksel veya sese dair bilgi kalıpları arasındaki benzerlikleri veya kimlikleri tanımayı öğrenecek elektronik veya elektromekanik bir sistem inşa edeceğini” belirtmiştir.

1958 yılında Cahit Arf, “Makineler Düşünebilir mi? ve Nasıl Düşünebilir?” adlı makalesini yayınlamıştır. Bu makale günümüzde halen popülerliğini korumaktadır.

1959 yılında, Arthur Samuel, “makine öğrenmesi” terimini ortaya atmıştır, bu terimi açıklamak için “Program yazan kişinin oynayabileceğinden daha iyi bir dama oyunu oynamayı öğrenecek.” ifadelerinde bulunmuştur.

1965'de Matematikçi Alexey Ivakhnenko, V.G. Lapa ve arkadaşları ilk derin öğrenme ağlarını kurmuştur. Yine 1965 yılında basit anlamda ilk defa İngilizce diyaloglara girilebilen program olan ELIZA'yı geliştirmiştir.

Arthur Bryson and Yu-Chi Ho 1969 yılında, çok katlı bir dinamik sistem optimizasyon yöntemi olarak geri yayılımı tanımlamışlardır. Çok tabakalı yapay sinir ağları üzerine olan bu öğrenme algoritması, bilgi işlem gücü büyük ağların eğitimini karşılamak için 2000'li ve 2010'lı yıllarda derin öğrenmenin başarısına önemli ölçüde katkıda bulunmuştur.

Fukushima (1970), görsel kayıtları tanıyabilen yapay sinir ağı geliştirerek Neocognitron buluşuna imza atmıştır. Bu buluş sonraki yıllarda özellikle Doğal Dil İşleme algoritmalarının gelişmesinde büyük rol oynamıştır.

1971 ile 1980 yılları arasında yapay zeka alanında literatüre Yapay Zeka Kışı olarak geçerek bu alandaki çalışmalarda ilerleme sağlanamamıştır.

Gerald Dejong (1981), verileri analiz edebilen ve veri ön işleme olarak izlenebilecek genel bir kural oluşturan Açıklama Tabanlı Öğrenme (Explanation Based Learning-EBL) modelini tanıtmıştır.

Ernst Dickmanns (1985), Münih'teki Bundeswehr Üniversitesi'nde kamera ve sensörlerle donatılmış ilk sürücüsüz arabayı tanıtmıştır.

1990 yılından itibaren yapay zeka alanındaki çalışmalar hızlı bir şekilde artmış ve başarılı sonuçlar alınmaya başlanmıştır.

1997 yılında IBM tarafından geliştirilen, geçmiş satranç karşılaşmaları üzerinden eğitilen, Deep Blue adlı bilgisayar ile dünya satranç şampiyonu Garry Kasparov'u yenerek bu alanda önemli bir eşğin geçilmesinde rol oynamıştır.

Sepp Hochreiter ve Jürgen (1997), günümüzde el yazısı ve konuşma tanımda kullanılan tekrarlayan sinir ağının bir türü olan Uzun ve Kısa Vadeli Hafıza (LSTM) fikrini ortaya atmışlardır.

LeCun (1998), “Gradyan Tabanlı Öğrenimin Belge Tanımaya Uygulanması (Gradient-based Learning Applied to Document Recognition)” başlıklı yazısı ile stokastik gradyan iniş algoritmasının geri yayılım algoritmasıyla birleştirilmesi üzerine derin öğrenmede performans bakımından önemli bir yaklaşım olarak bu alanda önemli bir yer almıştır.

Fei Fei Li ve Princeton Üniversitesi’ndeki meslektaşları (2009), görsel nesne tanıma yazılımı araştırmasına yardımcı olmak için tasarlanmış etiketlenmiş verilerin olduğu büyük bir veri tabanı olan ImageNet’i tanıtmışlardır. Günümüzde 14 milyondan fazla etiketlenmiş görselin bulunduğu bu sistem, bu alandaki çalışmalar için önemli bir eşik olmuştur.

Rajat Raina, Anand Madhavan ve Andrew Ng (2009), “modern grafik işlemcilerinin çok çekirdekli işlemcilerin yetersiz kaldığını ve denetimsiz öğrenme yöntemlerinin uygulanabilirliğinin devrim potansiyeline sahip olduğunu” savunarak “Grafik İşlemcilerini Kullanarak Büyük Ölçekli Denetimsiz Öğrenme”yi yayınlamışlardır.

2012 yılında Apple, Siri’yi tanıtmış ve cihazlarında kullanmaya başlamıştır. Siri, kullanıcı ile iletişim kurarak söylediklerini anlayan ve cevap veren başarılı bir uygulama olarak günümüzde de kullanılmaktadır.

2014 yılında Facebook, insan yüzlerini %97 ile tanıyabilen ve etiketleyebilen Deep Face yazılımını tanıtmıştır.

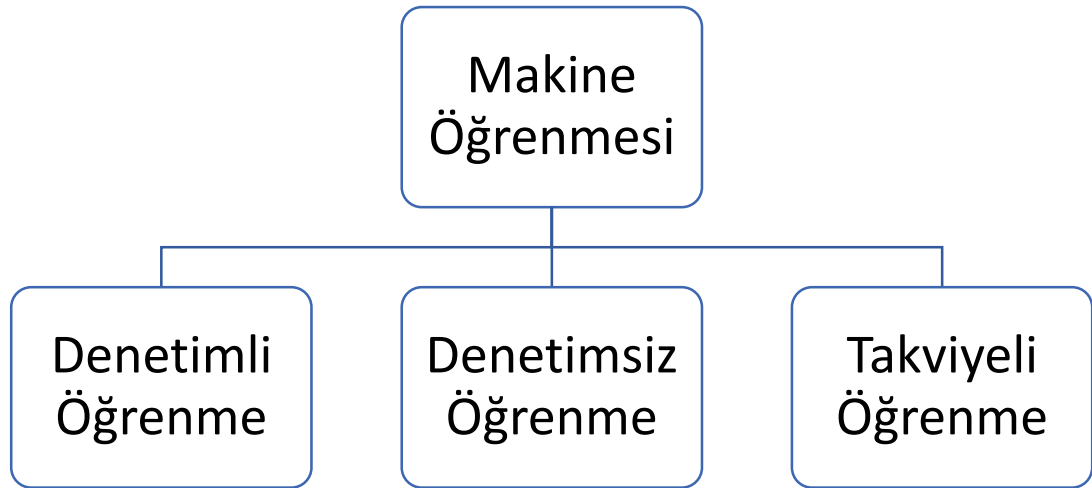
Ian Goodfellow liderlik ettiği bir araştırma ekibi tarafından 2014 yılında Üretken Çekişmeli Ağlar (Generative Adversarial Networks) olarak tanıtılan yüksek başarı oranı ile sentetik veri üretilmeye imkan tanıyan kavramı bu alana kazandırmıştır.

Güncel verilerin çok büyük boyutlara varması ve bunların işlenmesinin zorluğu performans bakımından gelişme gösterme zorunluğu 2016 yılından sonra yapay zeka

alanındaki çalışmaları çok ileri seviyelere getirmiştir. Son olarak 2020 yılında OpenAI tarafından tanıtılan GPT-3 ile 175 milyar parametre kullanarak insanların yazdığından ayırt edilemeyen metinler üretilmesine imkan tanınmıştır. 2021 yılında ise OpenAI tarafından, metinleri daha önce olmayan görsellere çeviren DALL – E yazılımı tanıtılarak bu alanda çok önemli bir gelişime katkı sağlamıştır.

4.2. Makine Öğrenmesi Türleri

Makine öğrenmesi; birçok alanda, birçok problemin çözümü için aktif olarak kullanılmakta ve sürekli olarak gelişen ve ilerleyen algoritmaların kullanımı ile ilgili kavramsal olarak genişlemektedir. Verilere dayalı işlemleri gerçekleştirebilmek ve makine öğrenmesi algoritmalarını kullanmak için belirli yöntemler geliştirilmiştir. Problemlerin yanıtları aranırken, belirlenen bu yöntemlere göre hareket etmek hem zaman kazanmak açısından hem de doğruluk yönünden önem teşkil etmektedir. Bu nedenle aranan çözümlere göre makine öğrenmesi türleri olarak Şekil 4.1’de gösterildiği gibi üç temel sınıf altında yer almaktadır.



Şekil 4.1. Makine öğrenmesi yöntemleri

4.2.1. Denetimli öğrenme

Denetimli öğrenme, her bir veri satırına (girdi) karşılık gelen bir bağımlı değişken (çıkıtı) olduğu durumlarda kullanılan algoritmaların genel adıdır. Veri setindeki her girdiye karşılık gelen çıkıtı bilindiği için algoritmalar yardımıyla çıkıtı ile girdiler

arasındaki ilişki, modeller tarafından öğrenilmeye çalışılmaktadır. Makine öğrenmesi algoritmaları ile tekrar tekrar tahminler üretip eldeki bağımlı değişkene ait değerler ile karşılaştırılarak en iyi tahmin fonksiyonu oluşturulmaya çalışılmaktadır. Böylece bu fonksiyon yardımıyla veri setine eklenen yeni girdi ile tahmin değerleri elde edilmektedir.

Matematiksel olarak ifade edilirse, girdi değişkenleri (X) ve bir çıktı değişkeni (Y) olan algoritmalar aracılığıyla girdilerden çıktı arasında eşleme yaparak bir fonksiyon oluşturulmaktadır (Brownlee, 2016).

$$Y = f(X) \quad (4.1)$$

Buradaki fonksiyonun amacı, her satır için en iyi şekilde sonuca yaklaşacak şekilde algoritmalar eşliğinde öğrenme sürecini gerçekleştirmesidir. Böylece algoritma girdiler ile çıktılar arasında sürekli olarak öğrenme işlemi yaparak en uygun modeli ve sonucu gerçekleştirmeye çalışmaktadır. Modelin ne kadar zamanda bu öğrenmeyi gerçekleştirmesi modelin performansı ile ilgilidir. Öğrenme süreci, modelin performansı optimum seviyeye geldiğinde durdurulmaktadır. Böylece bu fonksiyon yardımıyla yeni girdi olarak eklenen X'e karşılık bir Y değeri tahmin edilmektedir.

Denetimli öğrenme problemlerinde, bağımlı değişkenin (çıktı) nümerik veya kategorik yapıda olmasına göre sınıflandırma ve regresyon olarak iki grup altında incelenmektedir.

Regresyon modelleri olarak nümerik tahminler üreten; Lineer Regresyon, Ridge-Lasso Regresyon ve Rassal Ormanlar gibi modeller kullanılmaktadır. Örneğin, bir şirketin gelir tahminini yapmak için regresyon modelleri kullanılmaktadır. Sınıflandırma modelleri ise kategorik bir bağımlı değişkene (çıktı) sahip olan problemlerde kullanılmaktadır. Bu modeller içinde; Lojistik Regresyon, Destek Vektör Makineleri ve sınıflandırma problemleri için de kullanılan Karar Ağaçları, Rassal Ormanlar modelleri yer almaktadır. Örneğin, bir kişinin bir ürünü satın alıp almayacağı yönünde bir problemin araştırılmasında bu modellerden faydalanılmaktadır.

4.2.2. Denetimsiz öğrenme

Denetimsiz öğrenmede, denetimli öğrenmeden farklı olarak her veri satırına karşılık gelen bir bağımlı değişken (çıktı) bulunmaması olarak tanımlanmaktadır. Sadece girdilerin olduğu veri setlerinde genellikle kullanılmaktadır. Burada, her girdiye karşılık gelen bir çıktı olmadığı için girdi ile çıktı arasında bir tahmin fonksiyonu oluşturulamaz. Girdilerin kendi aralarındaki ilişkiler ve desenler genel olarak incelenmektedir (Brownlee, 2016). Denetimsiz öğrenmede; amaç, değişkenlerin birbiri ile ilişkilerini analiz etme ve dağılımlarını inceleme üzerinedir (Bruce ve Bruce, 2017).

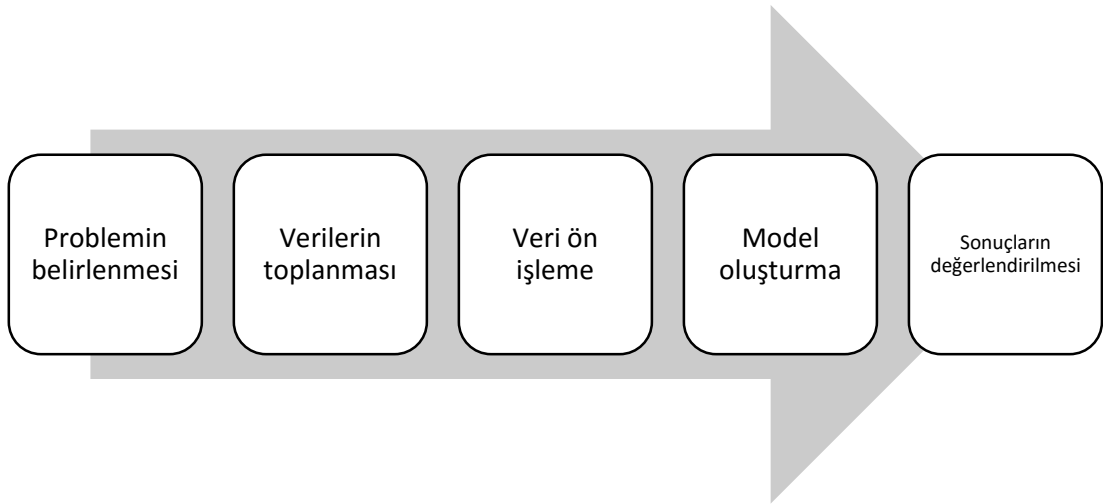
Denetimsiz öğrenme altında; kümeleme, boyut indirgeme ve anomali tespiti gibi modeller kullanılmaktadır. Örneğin, şirketten alışveriş yapan müşterilerin ortak özelliklerini bulmak adına gruplama, segmentasyon ve kümeleme modelleri kullanılabilir. Böylece müşteriler belirli sınıflar altında gruplanarak daha homojen sınıflar altında temsil edilmektedir. Buna göre kampanyalar veya uygun fırsatlar müşterilere tanımlanabilmektedir.

4.2.3. Takviyeli öğrenme

Takviyeli Öğrenme, olası durumları kendi kendine öğrenerek buna dayalı kararlar alan öğrenme yöntemidir. Aracı olarak adlandırılan öğrenme sistemi, veri setinden yola çıkarak olası durumları gözlemleyerek, eylemleri seçip gerçekleştirerek, karşılığında ödülleri veya cezalar vererek öğrenme tahminlerini geliştirmeye çalışmaktadır. Genellikle oyun programlarında ve robotik alanlarda kullanılmaktadır. Takviyeli öğrenme için en iyi örnek olarak Deep Mind şirketinin AlphaGo programı verilebilir. Takviyeli öğrenme yapısı ile oluşturulan AlphaGo, Go oyununda dünya şampiyonu Ke Jie'yi mağlup etmiştir. AlphaGo, milyonlarca Go oyunu analiz ederek, kendi kendine milyonlarca kez oynayarak, sonuçlara göre kendini düzelterek üstel olarak öğrenme sağlayan AlphaGo, kazanma politikası uygulayarak maçı kazanmıştır (Geron, 2019).

4.3.Makine Öğrenmesi Uygulama Aşamaları

Makine öğrenmesi modelleri uygulanırken, makine öğrenmesi modellerinin ve model performanslarının doğru analiz edilmesi için belirli bir akış izlemek ve bu akış çerçevesinde ilerlemek büyük önem taşımaktadır. Çalışmaya başlamadan önce çalışmaya dair bir planlama veya akış oluşturulması ve bu akışın takip edilmesi araştırmacıya hem zaman hem de maliyet açısından tasarruf yapmasını sağlamaktadır. Akış içerisindeki her bir adım, bir sonraki adımı kolaylaştırmakta ve hata yapma oranını azaltmaktadır. Veriyi analize uygun hale getirmek, modelin oluşması için en önemli aşamalardandır. Problemin belirlenmesinden başlayarak model oluşturma ve sonuçlara ulaşılmasına dek olan bu akışa, makine öğrenmesini çalışma döngüsü denmektedir.



Şekil 4.2. Makine öğrenmesi uygulama adımları

4.3.1. Problemin belirlenmesi

Makine öğrenmesi uygulama adımlarının birinci koşulu bir problemin belirlenmesidir. Belirlenen probleme ilişkin verilerin toplanması, devamında makine öğrenmesi uygulama adımları ve uygulanacak model belirlenmesi gerekmektedir. Problemin doğru tanımlanması modellemeye geçip geçmeme adına önemli bir adımdır. Problemin çözümüne yönelik ön analizler, model kurulumu için gerekli araçlar ve literatür araştırmaları gerekmektedir. Böylece makine öğrenmesi modelleri kurmadan

önce problem iyice analiz edilerek tespit edilmesi ve doğru modellerin kurulumu için bilgi edinimi sağlanmaktadır.

4.3.2. Verilerin toplanması

Problemin belirlenmesinden sonraki diğer önemli aşama probleme yönelik verilerin toplanması aşamasıdır. Kullanılacak veri tipleri, probleme yönelik örneklem boyutu, toplanması gereken veri hacmi, verinin toplanma yöntemleri, verinin nereden toplanması gerektiği gibi detayların bu aşamada araştırmacı tarafından belirlenmesi gerekmektedir. Probleme yönelik veri boyutunun problemin çözümüne ulaşmak için yeterli miktarda olması ve problemin cevabını verebiliyor özellikte olması gerekmektedir. Günümüzde verinin arttığı durumlarda, artan veriyle birlikte, veriden çıkarılacak anlamın da artacağı yönünde bir önyargı oluşmaktadır. Ancak veri ile veriden çıkarılacak anlam, doğrusal bir şekilde artmayabilir. Problemin çözümüne yönelik uygun olmayan veriler; zaman kaybetmeye, problemin çözümüne ulaşılmasını zor hale getirebilmeye, veride sapmalar veya yanlış sonuçlar elde etmeye neden olabilmektedir. Bu nedenle problemin çözümü için gerekli sonuca ulaştırmaya yarayacak yeterli büyüklükte veriler kullanmak daha doğru olmaktadır.

Veriler toplanırken veri setinde bulunan tahmincilerin yansızlık, tutarlık, etkinlik varsayımlarının istatistiksel olarak sağlanması gerekmektedir. Verilerin toplanması aşamasında yapılan yanlış seçimler, ana kütleyi tamamen temsil etmeyen, hata oranları yüksek modelleri beraberinde getirmektedir. Verilerde olan yanlışlık; varyansı arttıracak, modellerin gerçek hayata dair tahmin başarı performanslarını düşürecektir. Nihayetinde makine öğrenmesi modelleri, seçilen örneklem üzerinden ana kütle hakkında tahmin ve çıkarım yapmayı sağlayan istatistiksel modellerden oluşmaktadır. Bu sebepten ötürü, istatistiksel tahmin edicilerin varsayımları büyük derecede önem taşımaktadır. Tahmin ve çıkarım gücünü etkileyen en önemli nokta verilerin kalitesidir.

4.3.3 Veri ön işleme

Problemin çözümüne uygun olan doğru verilerin toplanması aşamasının ardından, veri ön işleme süreci gelmektedir. Makine öğrenmesinde kullanılan modellerinin çoğu önemli ölçüde veri ön işleme sürecine ihtiyaç duymaktadır.

Makine öğrenme adımları arasında, genel olarak en çok zaman harcanan adım veri ön işleme olmaktadır ve veri seti üzerine ön işleme yaparak model öncesi bir dizi işlem uygulamak gerekmektedir. Bu işlemlerden ilki, veri setinde boş veya problemin amacına uygun olmayan anlamsız değişkenler var ise bu değişkenlerin çıkarılmasıdır. Bunun sebebi, veri setindeki her değişkenin model için anlamlı bilgiler içermemesi olabilmektedir. Modelin performansını olumsuz yönde etkilememesi açısından gereksiz değişkenlerin çıkartılması, daha sağlıklı bir veri seti elde etmeyi sağlamaktadır. Veri setinde yer alan tüm değişkenler birer girdi olduğundan, bu tarz değişkenlerin olması modele yük olarak modelin performansını olumsuz yönde etkileyebilmektedir. Bazı değişkenler de tüm satır boyunca aynı gözlem değerini içermektedir. Bu değişkenlerde varyans sıfır olmakta, aynı bilgiyi içerdiğinden satır bazında bir fark yaratmamaktadır ve bu değişkenlerin de tespit edilerek veri setinden çıkarılması gerekmektedir.

Veri setinde eksik verilerin yer alması, modelin oluşturulmadan önceki süreçlerinde sık sık karşılaşılan durumlardan biridir. Bu nedenle değişken bazında eksik gözlemlerin incelenmesi gerekmektedir. Değişken bazında eksik verilerin olması durumunda ne yapılacağına karar verilmesi, eksik gözlemlerin çıkarılması veya bir dizi işlemle eksik gözlemlere değer atanması gerekmektedir. Günümüzde, eksik gözlemlere değerler atanması üzerine sıklıkla lineer regresyon veya KNN (En Yakın Komşuluk) temelli değer atama yöntemleri kullanılmaktadır. Bu yöntem, eksik gözlemlere değişken bazında ortalama veya mod ile değerler atanmasına göre daha güvenilir bir yöntem olarak sıklıkla uygulanmaktadır.

Veri ön işleme süreçlerinde bir diğer adım olarak, karakter içeren kategorik değişkenlerin sayısallaştırılması gelmektedir. Bölüm 5.2.5.'de anlatılan CatBoost modeli haricinde diğer Makine Öğrenmesi modelleri ile çalışırken tüm girdilerin sayısal olma zorunluluğunun mevcut olmasıdır. Bu nedenle kategorik değişkenlere,

ordinal veya nominal deęişken olmasına göre sayısallaştırma uygulanması gerekmektedir. Bu sayısallaştırma işlemi Python programlama dili yardımıyla hızlı bir şekilde yapılabilmektedir.

Veri ön işleme sürecinde yapılan işlemler, model kurulumu sürecinden algoritmaların doğru ve hızlı çalışması adına oldukça önemlidir. Eğitim ve test süreçlerinde daha az hata ile karşılaşmak ve doğru modeller kurmak için veri ön işleme adımlarının eksiksiz yapılması gerekmektedir.

4.3.4. Model oluşturulması

Makine öğrenmesi uygulama adımlarının sorunsuz bir şekilde tamamlandıktan sonra model oluşturma aşamasına geçilmektedir. Ancak model oluşturma aşamasından önce problemin hangi modeller ile çözüleceęi analiz edilmesi gerekmektedir. Öncelikle kurulacak modelin denetimli mi denetimsiz öğrenme mi olduęu tespit edilmelidir. Eğer hedef (bağımlı) deęişken mevcutsa denetimli öğrenme yöntemleri uygulanmaktadır. Bağımlı deęişkenin sürekli ya da kategorik olmasına göre de regresyon veya sınıflandırma modelleri uygulanmaktadır. Denetimli öğrenme modelleri kullanılacaksa model kurulum öncesinde; veri setinin, eğitim ve test verisi olmak üzere ikiye ayrılması gerekmektedir. Eğitim seti, modelin eğitildięi verilerden oluşmaktadır. Test seti ise model tarafından görülmeyen verilerden oluşmaktadır. Makine öğrenmesi modelleri, eğitim verisi üzerinde uygulanarak elde edilir ve test verisi üzerinde test edilmektedir. Veri setinin bu şekilde ayrılmasında, eğitilmeyen veriler üzerinde modelin ne kadar iyi performans gösterdięi fikri yatmaktadır.

Veri setinin, eğitim ve test verisi olarak ayrılmasında kullanılan birkaç yöntem bulunmaktadır, bu yöntemlerden birisi de Holdout yöntemidir. Eğitim ve test verisinin yüzdesel olarak ne oranda ayrılacağı verinin boyutuna göre farklılık göstermektedir. Genelde verilerin üçte ikisi (%70) eğitim veri seti olarak ayrılırken üçte biri (%30) test veri seti olarak ayrılmaktadır. Bu ayırım veri setine ve farklı oranlarda ayrılarak alınan sonucun doğruluęuna göre deęişebilmektedir. Fakat bu ayırımı yaparken örneklemin rassal dağıtılmasının sağlanması gerekmektedir (Lanz, 2013). Verilerin rassal olarak eğitim ve test verisi olarak ayrılması, modelin yanlılıęını düşürerek gerçek hayat

verilerine daha uygun olarak, ana kitle tahmin doğruluğunun artmasına yardımcı olmaktadır.

Holdout yöntemi model performansı açısından iyi sonuçlar elde edilmesine yardımcı olsa da sadece eğitim seti üzerinde bir model kurmaktadır. Sadece bir modelin kurulması, zaman yönünden az maliyetli olsa da modelin daha az sınanmasını beraberinde getirmektedir. Bu nedenle, eğitim setini de kendi içinde birden fazla kere eğitim - test verisi olarak bölmeye yarayan ve modelin birden fazla sayıda sınanmasını sağlayan K-Fold Cross Validation (K-Katlı Çapraz Doğrulama) yöntemi kullanılmaktadır.

K- Fold Cross Validation yöntemi, modelin görmediği veriler üzerinde çalışma imkanı sağlayarak objektif ve doğru bir şekilde değerlendirme yapmak için kullanılan istatistiksel bir yeniden örnekleme yöntemidir. K-kat çapraz doğrulama yönteminde veri seti rassal olarak “k” eşit parçaya ayrılır. Her bir parça test seti olarak belirlenir ve kalan parçalar ile model eğitilir. Bu döngü bütün parçalar test seti olana kadar devam ettirilir. Sonuç olarak elde edilen bütün tahmin hatalarının ortalaması alınarak modelin hata ortalaması elde edilmiş olur. Bu ortalama göre model değerlendirilir (Lanz, 2013). K değeri olarak herhangi bir sayı alınabilir ancak genel olarak 5 veya 10 kullanılır. Veri setinin çok büyük olduğu durumlarda k değerinin 10 alınması modelin çalışma zamanını oldukça arttırmaktadır.

4.3.5. Model sonuçlarının değerlendirilmesi

Eğitim verisi üzerinde uygulanarak elde edilen model, test verisine de uygulandıktan sonra elde edilen sonuçların doğruluğunu ölçmek gerekmektedir. Bu nedenle, tahmin değerleri ve gerçek değerlerin karşılaştırılması gerekmektedir. Bu karşılaştırma sonucunda kurulan modelin tipine göre hata değerleri veya doğruluk oranları gibi başarı oranını gösteren sonuçlar elde edilmektedir.

Karşılaştırma sonrasında çıkan ölçümlere göre modelin başarılı olup olmadığına yönelik net bulgular elde edilmektedir. Bu ölçümler sonucunda modelin veri setine ne derece uygun olup olmadığı sonucuna da varılmaktadır. Çünkü genellikle bir veri setine uygun olan birden fazla model olabilir. Böylece daha az hataya sahip olan model

tespit edilebilmekte ve en başarılı model ile çalışılarak daha önce modelin görmediği yeni bir veri geldiğinde başarılı tahminler yapılabilmektedir. Sonuçta elde edilen model ile bağımlı değişken üzerinde etkili olan değişkenlerin belirlenmesi de mevcut olmaktadır. Model sonucundan yola çıkarak en etkili değişkenlerin gözlenmesi ve bu değişkenlerin değerlendirilmesi söz konusu olmaktadır. Etkili değişkenlerin belirlenmesi, bağımlı değişkeni anlamada yardımcı olarak problemi çözümüne ulaşma açısından önemli bir faktör olmaktadır.

4.3.6. Model performansları değerlendirme ölçütleri

Uygulanan denetimli makine öğrenmesi yöntemlerinin başarılı sonuçlar verip verdiğini tespit edebilmek açısından bazı performans değerlendirme ölçüt değerlerine bakılması gerekmektedir. Bu değerler, uygulanan yöntemin başarısına dair değerlendirme yapılması adına önemlidir.

Başarı ölçütleri, modelin sınıflandırma veya regresyon problemi olmasına göre farklılık göstermektedir. Sınıflandırma problemlerinde; doğru sınıf tahmini yapılmasına dayandığı için ölçütler, modelin başarı oranını yüzdesel olarak ifade etmektedir. Tahmin edilen sınıflar ile gerçekte veride yer alan sınıfların değerlerini kıyaslayarak başarı metriklerini oluşturmaktadır. Doğru tahmin edilen sınıfların, toplam sınıf sayısına bölünmesi ile modelin ne kadar doğru tahmin yaptığını gösteren accuracy(doğruluk) ölçütü genellikle kullanılan önemli ölçütlerden biridir.

Çalışmada ele alınan regresyon problemlerinde ise modelin başarısını ölçmek için gerçek ile tahmin edilen değerler arasındaki farka bakılmaktadır. Aradaki farkın minimal olması istenmektedir. Bu nedenle, tahmin edilen değer ile gerçek değer arasındaki farklara bakmak için çalışma kapsamında, yaygın olarak kullanılan Ortalama Kare Hata (Mean Squared Error, MSE), Ortalama Mutlak Hata (Mean Absolute Error, MAE) ve Kök Ortalama Kare Hata (Root Mean Squared Error, RMSE) değerlendirme kriterlerine bakılmıştır. Bu kriterlere göre kıyaslama yapılırken; hangi modelin gerçek ile tahmin arasındaki hata değeri az ise o modelin diğerine daha başarılı olduğu anlaşılmaktadır.

4.3.6.1. Ortalama Kare Hata (Mean Squared Error, MSE)

Ortalama Kare Hata metriği, regresyon eğrisinin gerçek verilere ne kadar yakın olduğunu söylemektedir. Regresyon doğrusu (tahmin değerler) ile gerçek değerler arasındaki farkın karesi alınarak örneklem sayısına bölünmesiyle elde edilir (James vd., 2013).

Formül 4.3'te; gerçek değer ile tahmin edilen değerler arasındaki fark, hata (kalıntı) "e" olarak ve örneklem sayısı "n" olmak üzere matematiksel olarak Ortalama Kare Hata tanımlanmaktadır.

$$MSE = \frac{1}{n} \sum_{j=1}^n e_j^2 \quad (4.2)$$

Ortalama Kare Hata, bir makine öğrenmesi modelinin, tahminleyicinin performansını ölçmektedir. Formül gereği karesi alındığı için her zaman pozitif değerler almaktadır ve sıfıra yakın olan tahminleyicilerin daha iyi bir performans gösterdiği bu durumdan dolayı söylenebilmektedir.

4.3.6.2. Kök Ortalama Kare Hata (Root Mean Squared Error, RMSE)

RMSE, Ortalama Kare Hata değerinin kökü alınarak elde edilen değere karşılık gelmektedir. Tahminleyicinin tahmin ettiği değerler ile gerçek değerleri arasındaki uzaklığın bulunmasında sıklıkla kullanılan, hatanın büyüklüğünü ölçen kuadratik bir metriktir.

RMSE, tahmin hatalarının (kalıntıların) standart sapması olarak ifade edilmektedir. Kalıntılar, regresyon hattının veri noktalarından ne kadar uzakta olduğunun bir ölçüsü ise, RMSE de bu kalıntıların ne kadar yayıldığıнын bir ölçüsü olarak ifade edilmektedir (James vd., 2013).

RMSE, büyük hataları daha fazla cezalandırmanın avantajına sahiptir, bu yüzden bazı problemlerde kullanmak daha uygun olabilmektedir. Ayrıca MSE değerlerinin bazen

yeterince büyük çıkması modeller arasında kıyaslamayı zorlaştırdığından, RMSE metriğini baz almak kıyaslanabilirliği kolaylaştırmaktadır.

4.3.6.3. Ortalama Mutlak Hata (Mean Absolute Error, MAE)

Ortalama Mutlak Hata, gerçek ile tahmin değerleri arasındaki farkların mutlak değeri olarak ifade edilmektedir. MSE metriğinde hata terimlerinin karesi alınırken bu metrikte hata (kalıntı) değerlerinin mutlak değeri alınmaktadır.

Formül 4.4'te; gerçek değer ile tahmin edilen değerler arasındaki fark, hata (kalıntı) "e" olarak ve örneklem sayısı "n" olmak üzere matematiksel olarak Ortalama Mutlak Hata tanımlanmaktadır.

$$MAE = \frac{1}{n} \sum_{j=1}^n |e_j| \quad (4.3)$$

MAE değeri kolay yorumlanabilir ve kıyaslanabilir olduğu için regresyon ve zaman serisi problemlerinde sıkça kullanılmaktadır. MAE, her gerçek değer ile veriye en iyi uyan çizgi arasındaki ortalama dikey ve yatay mesafe olarak tanımlanmaktadır. MAE, yönlerini dikkate almadan bir dizi tahmindeki hataların ortalama büyüklüğünü ölçen, tüm tekil hataların ortalama eşit olarak ağırlıklandırıldığı doğrusal bir skordur.

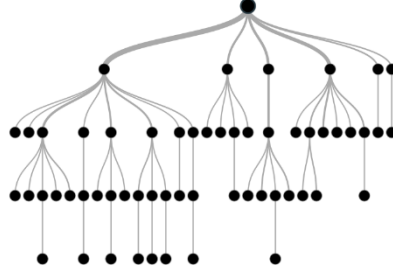
5. AĞAÇ TEMELLİ MAKİNE ÖĞRENMESİ MODELLERİ

Makine öğrenmesi modelleri genel olarak matematik ve istatistik temelli yapılardan oluşmaktadır. Bu nedenle, bazı modellerin oluşumunda birtakım varsayımları sağlamak gerekmektedir. Ancak günümüzde kullanımı artan, denetimli öğrenme modellerinden olan, ağaç yapılarına dayanan, ağaç temelli algoritmalarda ise genel olarak varsayımlar bulunmamaktadır. Aykırı gözlemlerden, eksik değerlerden ve çoklu bağlantı problemlerinden etkilenmemektedir. Bazı modellerinde eksik verilerle çalışma model kurma imkanı mevcuttur. Bağımlı ve bağımsız değişkenler için dağılım varsayımları söz konusu değildir. Bağımlı ve bağımsız değişkenlerin hem kategorik hem de sürekli olması durumunda da kullanılmaktadır. Anlaşılması kolaydır ve performans olarak da hızlı yapılara sahiptirler. Bu nedenle, günümüzde verilerin hacminin büyüklüğünün artması göz önüne alındığında ağaç temelli modellerin kullanımı da artış göstermektedir. Özellikle bu kullanım artışını, gerçek veya sentetik veriler üzerine, makine öğrenmesi modelleri kurma yarışmaları düzenleyen Kaggle gibi platformlardan gözlemlemek mümkündür. Son dönemlerde yarışmalarda başarılı performans gösteren modeller genellikle ağaç temelli modeller olmaktadır.

Çalışmada, ağaç temelli modeller seçilerek, Karar Ağaçları ve ağaç temelli topluluk (ensemble) öğrenme yöntemlerinin kavramsal çerçeveleri anlatılarak uygulanmıştır. Bağımlı (hedef) değişkenin sürekli bir yapıda olması nedeniyle ağaç temelli modeller kullanılarak regresyon ağaçları kurulmuştur. Bu nedenle modeller anlatılırken matematiksel olarak daha çok regresyon ağaçlarına dair formülasyonlar gösterilmiştir.

5.1. Karar Ağaçları

Karar ağaçları, veri seti içerisindeki karmaşık yapıları basit karar yapılarına dönüştürmekte kullanılan denetimli öğrenme modelidir. Karar ağaçları, karışık ve bilinmeyen verilerin üstesinden gelinmesini, kolayca yorumlanmasını sağlayan modeller oluşturmaktadır. Hem sınıflandırma ve hem de regresyon problemleri için kullanılabilir (James vd., 2017)



Şekil 5.1. Karar ağaçları örnek gösterim

Karar ağaçları ile heterojen veri setleri; bir bağımlı(hedef) değişkene göre, yukarıdan aşağıya doğru, bir dizi karar kuralları uygulayarak homojen alt gruplara ayrılmaktadır (Breiman vd., 1984).

Karar ağaçları modelinin önemli avantajlarından biri, veri setinin çok karmaşık olduğu durumlarda bile, bağımlı(hedef) değişkeni etkileyen değişkenleri ve bu değişkenlerin modeldeki önemini sırasıyla basit bir ağaç yapısı ile görsel olarak sunmaktadır (Kayri vd., 2014).

Karar ağaçlarının daha az veri ön hazırlığı gerektirmesi, açıklanabilirliğinin kolay olması ve sonuçların daha kolay anlaşılması modelin avantajları arasında yer almaktadır (James vd., 2013).

Karar ağaçlarının dezavantajı olarak; eğitim seti performansı yüksek iken test seti performansı oldukça düşük çıkması olarak tanımlanan “ezbere öğrenme (overfitting)” gösterilebilir. Karar ağacı veriyi çok fazla sayıda takıma bölmektedir. Bu takımlar spesifik bir hal aldıkça küçülmeye başlamaktadır. İncelenmesi gereken farklı durumların sayısı arttıkça, eğitim kümelerinin her biri daha da küçülmektedir (Altaş ve Gülpınar, 2012). Bu durumda karmaşık ve anlaşılması güç bir ağaç yapısı meydana gelebilir. Bu problemin çözümü için modelin parametrelerine kısıtlamalar getirilmekte ve budama gibi yöntemler kullanılmaktadır.

Karar Ağaçları yardımı ile öncelikle veri seti, bir ağaç üzerinden modellenerek bağımsız değişkenlerin hem kendi arasındaki ilişkisi hem de bağımlı değişken üzerindeki etkisine ağaç yardımı ile ulaşılmaktadır. Oluşturulan model ile sonradan eklenecek gözlemlenmeyen verinin tahmini de gerçekleştirilmiş olmaktadır.

Karar ağaçları yönteminde, ağaç görseline benzer olarak, veri setinde bir değişken kök olacak ağaçlandırma başlanır, dallanma ise düğüm ile karar verilen noktalarla sağlanır. Her bir gözlem, düğümler yardımıyla sınıflandırılır. Yapraklar ise sonucu verir ve yapraklara ulaşana kadar işlemler devam eder (Bell, 2014).

Karar ağacı, örnek uzayının yinelemeli olarak bölümlere ayrılması olarak ifade edilen bir sınıflandırma tekniğidir. Karar ağacı, bir kök ve ağacı oluşturan düğümlerden oluşur. Kök, herhangi bir gelen uca sahip olmamasına rağmen, diğer düğümler bir adet gelen uca sahiptir. Giden uca sahip düğümler; dâhili ya da test düğümü olarak, diğer düğümler ise terminal ya da karar düğümleri olarak bilinen yapraklar olarak adlandırılır. Bir karar ağacında, her dâhili düğüm, girdi özelliklerinin belli bir kesikli (ayrık) fonksiyonuna göre örnek uzayını iki ya da daha çok alt uzaya böler (Rokach ve Maimon, 2005). Karar ağacı başlatıcıları, verilen veri kümesinden bir karar ağacını otomatik olarak oluşturan algoritmalarlardır. Genel olarak amaç, genelleme hatasını en aza indirerek en uygun karar ağacını bulmaktır. Bununla birlikte; diğer hedef fonksiyonlar, düğümlerin sayısını veya ortalama derinliği en aza indirmek olarak açıklanmaktadır.

Karar ağaçları oluşturmada, ortak bir yaklaşım olarak böl ve elde et yöntemi kullanılmaktadır. En iyi özelliğin seçilmesi için, yukarıdan aşağıya doğru, ağaçların iç düğümüne bir etiket verilmesiyle yinelemeli bir şekilde oluşan süreçten meydana gelmektedir. İlk olarak ağacın kökünü temsil eden bir özellik seçilir. İlk özelliğin seçiminden sonra, özelliğin olası her bir değeri için bir dal oluşturulur ve veri kümesi, seçili özelliğin örnek değerlerine göre alt kümelerine ayrılır. Seçim süreci, her dalına yinelemeli olarak, örneklerin uyumlu alt kümeleri kullanılarak uygulanır. Örneğin; alt kümede, dal ile ilgili özellik değerlerine sahip tüm örnekler yer alır. Aynı sınıf etiketine sahip tüm alt örnekler bir sınıfta olduğunda veya başka bir durdurma kriteri ile karşılaşıldığında, tahmin edilecek bir sınıf etiketini temsil etmek için bir yaprak düğümü oluşturulur (Otero vd., 2012).

Bölünmelerin yani dallanmaların hangi kriterlere göre nasıl gerçekleştiği model için önemli bir yer teşkil etmektedir. Ayrıca karar ağaçlarının oluşturulması sırasında dallanmaya hangi değişken ile başlanacağı önemlidir. Çünkü olası tüm ağaç yapılarını

ortaya çıkararak içlerinden en uygun olanı ile başlamak mümkün değildir. Bu nedenle algoritmalarının çoğu daha başlangıçta bir takım değerleri hesaplayarak ona göre ağaç oluşturma yoluna gitmektedir (Tapkan vd., 2011). Bölünmenin hangi kriterlere göre belirleneceği sınıflandırma ve regresyon ağaçlarına göre farklılık göstermektedir. Bölünmenin nasıl gerçekleşeceği ağacın doğruluğunu etkileyen faktörlerdendir.

Bölünmenin belirlendiği kriterler; bölme uyum kriterleri olarak da adlandırılıp entropiye ve safsızlık fonksiyonuna dayalı olarak incelenir. Entropi, veri seti ile ilgili belirsizliğin bir ölçüsüdür. Veri setindeki bir verinin yalnızca bir etiketi varsa daha düşük bir entropiye sahip olduğunu söylenebilir. Bu durumda veri setinin entropiyi en aza indirecek bir şekilde bölünmesi gerekmektedir.

Entropi, veri setindeki örneklerin homojenliğini hesaplayan bir ölçüttür. Eğer, entropi 0 ise örnekler tamamen homojen, entropi 1 olduğunda ise örnekler homojen değildir anlamına gelmektedir (Sastry vd., 2010). Entropiye dayalı kriterler ve onun türevleri olan kazanç oranı ile kazanç ölçütü, bağımlı değişkenin kategorik olduğu problemlerde bölme yöntemi olarak kullanılmaktadır.

Bir diğer kriter olan safsızlık fonksiyonuna dayalı kriterler, sınıflama ağacı problemlerinde kullanılan gini kriteri ve twoing bölme kriteridir. Regresyon ağacı problemlerinde kullanılan ve safsızlık fonksiyonu dayalı bölme kriteri ise en küçük kareli sapma (least squared deviation: LSD) ve onun alternatif yöntemi olan en küçük mutlak sapma (least absolute deviation: LAD) yöntemidir (Akpınar, 2000).

Karar ağaçları oluşturulduktan sonra oluşturulan yapının karmaşıklığını, sınıflandırma hatasını ve ezbere öğrenmeyi engellemek adına budama işlemi yapmak gereklidir. Budamayla, ağaçta bulunan, sonucu etkilemeyen ve sınıflamaya herhangi bir katkısı olmayan dalların ağaçtan alınması sağlanır. Eğer ağaçta birçok düğüm ve dal oluşursa, ağacın alt dalları ve yapraklarına ulaşan veri sayısı da azalacaktır. Bu da ağacın hassasiyetinin azalması anlamına gelmektedir (Silahtaroglu, 2013).

Budama işleminde, karar ağacında bir alt ağacı atarak yerine, bir yaprağın yani uç düğümün yerleştirilmesi sağlanmaktadır. Böylece ağacın derinliği yani dallanması

azaltılmış olmaktadır. Bu nedenle, daha az karmaşık ve açıklanabilir yapılara ulaşılması söz konusu olmaktadır.

5.1.1. Karar Ağacı algoritmaları

Karar ağaçlarına dayalı olarak geliştirilen birçok algoritma vardır. Bu algoritmalar temel anlamda bağımlı değişkenin tipine göre ayrılrsa da aslında kök, düğüm ve bölme kriteri seçimlerinde izledikleri yola göre birbirinden farklılık göstermektedir. Kullanılan algoritmalara göre ağaçların yapıları da birbirinden farklılık göstermektedir (Silahtaroglu, 2013).

Karar ağaçları algoritmalarına karar verirken üzerine çalışılacak veri setinin yapısı ve probleme göre seçim yapılmaktadır. Genel olarak kullanılan algoritmalar aşağıdaki tabloda gösterilmektedir.

Çizelge 5.1. Sınıflandırma ve regresyon ağaçları algoritmaları (Saitoğlu, 2015; Yücel, 2017)

Algoritma Adı	Veri Tipi	Özellikleri	Ağaç Dallanma Kriteri
AID	Bağımlı değişken sürekli, bağımsız değişkenler kategorik (nominal /ordinal) olmalıdır.	Ağaçta ikili ağaçlar oluşturulur. Bağımlı değişken kategorik ise Sınıflandırma, sürekli ise Regresyon Ağacı kurulur.	Gruplar arası kareler toplamı kriterini kullanır.
CHAID	Bağımlı değişken kategorik (nominal, ordinal) veya sürekli olabilir. Bağımsız değişkenler ise kategorik olabilir.	Ağaç çoklu alt gruplara ayrılır. Yani çoklu ağaçlar oluşturulur. Bağımlı değişken kategorik ise Sınıflandırma, sürekli ise Regresyon Ağacı kurulur.	Bağımlı değişken sürekli ise F testini- Bağımlı değişken nominal ise pearson ki-kare testini, bağımlı değişken ordinal ise en çok olabilirlik testini kullanır.
CART	Bağımlı ve bağımsız değişkenler sürekli veya kategorik olabilir.	Ağaçta sadece ikili bölünmeler olur. Yani ikili ağaçlar oluşturulur. Karmaşıklık durumuna göre budama işlemi yapılır. Bağımlı değişken kategorik ise Sınıflandırma, sürekli ise Regresyon Ağacı kurulur.	Bağımlı değişken kategorik ise, gini ya da twoing kriteri, bağımlı değişken sürekli ise EKK yöntemi kullanır.
ID3	Bağımlı ve bağımsız değişkenler kategorik olmalıdır (ordinal / nominal).	Her düğümden çıkan çoklu dallar ile ağaç oluşturulur. Dalların sayısı tahmin edicinin kategori sayısına eşit olur.	Kazanç ölçütü kriterini kullanır.

MARS	Bağımlı değişken ikili (binary) veya sürekli olabilir. Bağımsız değişkenler sürekli veya kategorik olabilir.	Ağaçta sadece ikili bölünmeler olur. Yani ikili ağaçlar oluşturulur.	İleriye ve geriye doğru adım algoritmalarını kullanır.
Exhaustive CHAID	Bağımlı değişken kategorik veya sürekli, bağımsız değişkenler kategorik olmalıdır.	Ağaç çoklu alt gruplara ayrılır.	Bağımlı değişken; sürekli ise F testini, nominal ise pearson ki-kare testini, ordinal ise en çok olasılık testini kullanır.
C4.5 - C5.0 Algoritması	Bağımlı değişken kategorik olmalıdır. Bağımsız değişkenler ise kategorik veya sürekli olabilir.	Her düğümden çıkan çoklu dallar ile ağaç oluşturulur. Dalların sayısı tahmin edicinin kategori sayısına eşit olur	Kazanç ölçütü kriterini kullanır.
SLIQ	Bağımlı değişken kategorik (nominal, ordinal), bağımsız değişkenler kategorik veya sürekli olabilir.	Ağaç çoklu alt gruplara ayrılır, çoklu ağaçlar oluşturulmuş olur. Ağaç budama işlemleri hızlı bir şekilde gerçekleştirilir.	Gini kriterini kullanır.
SPRINT	Bağımlı değişken kategorik olmalıdır. Bağımsız değişkenler kategorik veya sürekli olabilir.	Ağaç çoklu alt gruplara ayrılır, çoklu ağaçlar oluşturulmuş olur. Büyük veri kümeleri üzerine çalışıldığında ideal bir algoritmadır.	Gini kriterini kullanır.
QUEST	Bağımlı değişken kategorik olmalıdır. Bağımsız değişkenler kategorik veya sürekli olabilir.	Bağımlı değişkeni sadece ikiye bölerek ikili ağaçlar oluşturur.	Kuadratik diskriminant analizini kullanır.

5.1.1.1. CART algoritması

Sınıflandırma ve Regresyon Ağaçları (CART) yöntemi, hem regresyon hem de sınıflandırma problemleri için kullanılan özyinelemeli bir yöntemdir. Amaç, bağımlı (hedef) değişken ile ilgili mümkün olan genişlikte, homojen alt gruplar oluşturabilmektir.

Bağımlı değişkenin kategorik olduğu durumlarda, Sınıflandırma Ağacı (Classification Tree - CT); sürekli olduğu durumlarda ise Regresyon Ağacı (Regression Tree - RT) olarak tanımlanmaktadır (Chang ve Wang, 2006).

1984 yılında Breiman, Friedman, Olshen ve Stone tarafından CART algoritması geliştirilmiştir. CART algoritması; parametrik regresyon modellerinin sağlamak zorunda olduğu varsayımları gerektirmeyen, kategorik ya da sürekli bağımlı değişkenin sınıf atamasını tahmin eden ters ağaç görünümündeki modellerdir (Temel, 2004).

CART algoritması; her adımda ilgili grubun, kendisinden daha fazla homojen 2 alt gruba ayrılmasına dayanmaktadır. Yani her dal ikili alt dallara bölünerek büyümektedir. Eğer bağımlı değişken türü kategorik ise twoing, gini indeksi; sürekli değişken ise en küçük kareler sapması veya en küçük mutlak sapma (LAD) hesaplamasına göre büyüme yapılmaktadır (Akpınar, 2000).

CART yöntemi; ağaç oluşturmada üç temel kural içermektedir. Bunlar; “örnekleme bölme kuralı (sample-splitting rule)”, “bölme uyum kriteri (goodness of split criteria)” ve “optimum ağacı seçme” kriterleridir. CART önceden belirlenmiş bölme kuralına ve düğüm bölme sürecinin her aşamasında uygulanan bölme uyum kriterine göre ağaç oluşturmaktadır. Ağaç oluşum sürecinde CART başlangıç veri setini, bağımsız değişkene sorduğu sorulardan alacağı evet/hayır cevaplarına göre ikili alt parçaya (alt örnekleme) bölmektedir. Her soru, sadece bir bağımsız değişken için sorulmaktadır. Alınan cevaba bağlı olarak veri seti sağ ve sol alt parçaya ayrılmaktadır (Yohannes ve Webb, 1999).

CART algoritması, parametrik yöntemleri etkileyen aykırı değerler, çoklu bağlantı problemi, değişen varyans ve dağılımsal hata şekli gibi durumlardan etkilenmezler. Aykırı gözlemler, düğüm içinde izole edilirler ve bölünmeye etki etmezler (Baysal, 2011).

Bağımlı değişkenin sürekli olduğu durumlarda CART algoritması Regresyon Ağacı olarak ifade edilmektedir. Regresyon Ağacı, bağımlı değişkenler ile bağımsız değişkenler arasında etki ve ilişkiyi tespit eden, veriyi homojen alt sınıflara ayıran parametrik olmayan bir yöntemdir (Kayri ve Boysan, 2008).

Regresyon Ağacı; parametrik regresyon varsayımlarını gerektirmeyen; aykırı gözlemlerden, çoklu bağlantı probleminden, eksik gözlemlerden etkilenmeyen bir yöntemdir. Bu gibi avantajlarından dolayı tercih edilmektedir (Akkartal ve Mendes, 2009).

Regresyon Ağacı üç aşama üzerine oluşmaktadır. İlk olarak ağaç oluşturulmakta, oluşturulan ağaç üzerine budama işlemi yapılmakta ve son olarak budama işlemi sonrası en uygun ağaç seçimi yapılmış olmaktadır.

Regresyon Ağacı yapısı oluşturulurken en fazla sayıda homojen alt gruplar oluşturulmaktadır. Ancak bu ağacın karmaşıklığını ve ezbere öğrenmesini beraberinde getirebilmektedir. Bu nedenle, budama işlemi ile alt gruplar (ağaçlar) arasında bağımlı değişken ile önemli düzeyde ilişki içeren ağaçlar seçilmektedir (Baysal, 2011). Böylece ağaca katkısı olmayan dalların ağaçtan alınması ile gereksiz ayrıntılardan model arındırılmış olmaktadır.

Regresyon Ağacı yapıları oluşturulurken bölünme (dallanma) kısmında safsızlık kriteri olarak en küçük kareli sapma hesabı kullanılmaktadır. Yine aynı şekilde budama kısmında da en küçük kareli sapma (LSD) hesabı kullanılmaktadır (Breiman vd., 1984).

Ağacın bölünmesi için kullanılan LSD yöntemi, aşağıdaki Formül 5.1 ve 5.2’de ifade gösterilmektedir. N (gözlem sayısı) ve m (yaprak sayısı), olmak üzere, Hata Ortalamalarının Karesi (MSE):

$$\bar{y}_m = \frac{1}{N_m} \sum_{y \in Q_m} y \quad (5.1)$$

$H(Q_m)$ LSD ölçüsü ise:

$$H(Q_m) = \frac{1}{N_m} \sum_{y \in Q_m} (y - \bar{y}_m)^2 \quad (5.2)$$

olarak formüle edilmektedir. Ağaç dallanırken ve budanırken her bir yaprak için hata karelerini baz almaktadır. Her bir yaprak için bir hata değeri oluşmaktadır. LSD ölçüsü ise her yaprak için oluşturulan hata karelerinin ortalamasını vermektedir. Regresyon ağaçları bu ölçüye göre dallanıp budanmaktadır.

CART algoritmasının altında yer alan Sınıflandırma Ağacı’nın kullanılması için ise bağımlı değişkenin kategorik olması gerekmektedir. Sınıflandırma Ağacı çalışmalarının amacı, doğru bir sınıflayıcı üretmek ve problemin altında yatan tahmin edici yapıyı ortaya çıkarmaktır. Bir gözlemin hangi sınıfta yer alacağının belirlenmesinde; hangi değişkenlerin ya da değişkenler arası etkileşimin problemi çözmede faydalı olacağının anlaşılmasına çalışılması gerekmektedir. Sınıflandırma

Ağacı oluştururken kriter sadece veri setine yönelik doğru bir sınıflayıcı üretmek değil, aynı zamanda gözlenmeyen verinin de hangi sınıfa atanacağını tahmin etmek ve anlamaktır (Breiman vd., 1993).

Sınıflandırma Ağacı oluşturulurken Regresyon Ağacı yapısındaki gibi öncelikle bağımlı değişken üzerine olası tüm homojen alt gruplar oluşturulmakta, budama işlemi yapılmakta ve budama sonrası ağaç seçimi yapılmaktadır. Ancak bölünme yaparken yani ağaç oluşturulurken safsızlık kriteri olarak Regresyon Ağacı'ndan farklı olarak yaygın olarak Gini ve Twoing ölçüleri kullanılmaktadır.

Sınıflandırma Ağacı'nda bölünme oluşturulduktan sonra budama işlemi ise maliyet – karmaşıklık (cost-complexity) parametresine göre yapılmaktadır. Ancak sınıflandırma ağacı için bu parametreyi baz olarak budama için gerçekleştirilen yaygın kullanımlarından biri Minimal Maliyet – Karmaşıklık Budaması (Minimal Cost-Complexity Pruning) terimidir.

Minimal Maliyet – Karmaşıklık Budaması, modelin ezberlemesini engelleyen bir algortima ve budama tekniğidir. (Breiman vd.,1984). Budamadan önceki başlangıç karar ağacı T_0 ve budama sonundaki karar ağacı T_k olmak üzere, bu dizi $T_0, T_1, \dots, T_k$ şeklinde karar ağaçlarında oluşmaktadır (Quinlan, 1987).

Budama aşamasında, T_{i+1} ağacı, son ağaç olan T_k son bir yaprak olana kadar T 'nin bir ya da daha fazla yaprağı yer değiştirmektedir. Sonraki aşamada ise bu ağaçların tamamı veya bu ağaçlardan biri seçilerek budanmış ağaç olarak değerlendirilmektedir (Quinlan, 1987).

T_{i+1} ağacı, kendinden önceki ağacın (T_i) uygun yapraklar ile bir veya daha fazla ağacın yer değiştirmesi ile elde edilmektedir. Budanmış ağaçlar, yaprak başına hata oranındaki en düşük artışı sağlayan ağaçlardır (Rokach ve Maimon, 2005).

$$\alpha = \frac{\varepsilon(\text{pruned}(T,t),S) - \varepsilon(T,S)}{|\text{leaves}(T)| - |\text{leaves}(\text{pruned}(T,t))|} \quad (5.3)$$

Formül 5.3'te, $\varepsilon(T, S)$ gösterilen terimde S, T ağacının örnekleri üzerindeki hata oranını göstermektedir. $leaves(T)$ terimi, T ağacındaki yaprak sayısını ve $pruned(T, t)$ terimi de T ağacındaki t düğümünün uygun bir yaprak ile yer değiştirmesi ile elde edilen ağacı göstermektedir (Rokach ve Maimon, 2005).

Bundan sonraki aşamada, $T_0, T_1, \dots, T_k$ dizisine ait genelleme hataları tahmin edilerek, budanmış ağaçlar arasından en iyisi seçilir. Veri setindeki gözlem sayıları yeterince büyükse tavsiye olarak veri seti eğitim ve budama seti olarak ayrılmalıdır. Ağaçlar eğitim seti üzerinden oluşturulup budama seti üzerinden değerlendirilir. Veri setinin büyüklüğü yetersiz ise çapraz doğrulama yöntemi tavsiye edilir (Rokach ve Maimon, 2005).

Budama metodunda, farklı sayılarda bir dizi ağaç budanır ve sonrasında veri setinde oluşan sınıflandırma hatalarının sayısı incelenerek seçim yapılır. Bu yöntem, hem ağacın karmaşıklık boyutunu hem de hata sayısını ele almaktadır (Mingers, 1989). Bu işlemlerin sonunda bölünme ve budama gerçekleştirilerek optimum ağaç seçimi yapılmış olmaktadır.

5.2. Ağaç Temelli Topluluk Öğrenme Yöntemleri (Ensemble Learning Methods)

Topluluk öğrenme yöntemleri temel olarak birden fazla modelin tahmin gücünün ve performansının, bir modelin tahmin performansından daha iyi olduğu fikrine dayanmaktadır.

Temel olarak topluluk modeli, birçok farklı modelin tahmin sonuçlarını birleştirmektedir. Bireysel modeller, tek modele sahip oldukları için “zayıf öğrenci” olarak adlandırılmaktadır. Topluluk öğrenme yöntemi ile zayıf öğrenciler bir araya gelerek yüksek tahmin performansına sahip bir “güçlü öğrenci” oluşturabilmektedir (James vd., 2013).

Kümeleme ve sınıflama yapan çok sayıda ilgi çekici topluluk öğrenme yöntemleri kullanılmaktadır. Literatürde karar ağaçlarını temel alarak geliştirilen ve daha güçlü tahmin sonuçları veren üç farklı topluluk (ensemble) öğrenme algoritması

bulunmaktadır. Bunlar; Torbalama (Bagging), Rassal Ormanlar (Random Forest) ve Artırma (Boosting) algoritmalarıdır (James vd., 2013).

Torbalama (bagging) yöntemi ile veri setinde modelin eğitilmesi için oluşturulan eğitim seti, bootstrap (yeniden rastgele seçilen örneklem) yöntemi ile daha küçük parçalara bölünür. Oluşturulan bu farklı eğitim setleri, tek bir algoritmanın eğitilmesi için farklı modeller (ağaçlar) oluşturur. Modellerin tahmin sonuçlarına, oylama (voting) ile elenerek ya da ortalama alınarak ulaşılır. Bu yolla, farklı eğitim seti ile eğitilen modelin, tek bir eğitim seti ile eğitilen modele göre daha başarılı tahmin sonuçları üreteceği amaçlanmaktadır (Breiman vd., 1993; James vd., 2013).

Boosting (Artırma) yöntemi, torbalama yöntemine çok benzemekle beraber ağaçların dallanması açısından farklılık göstermektedir. Torbalama yönteminde yeniden seçim yapılarak veri seti üzerinde birbirinden etkilenmeyen farklı ağaçlar oluşturulur ve bu ağaçların bir araya gelmesiyle bir tahmin modeli ortaya çıkar. Anca artırma (boosting) yönteminde, ağaçlar daha önce büyüyen ağaçlardan elde ettikleri bilgileri kullanarak etkileşim (sequential) halinde büyürler. Böylece ağaç büyürken yapmış olduğu tahmin hatalarını tekrar etmemeye çalışır. Artırma yönteminde bootstrap örnekleme kullanılmaz, bunun yerine her bir ağaç sonradan değiştirilmiş veri seti üzerine köklerini salar (James vd., 2013).

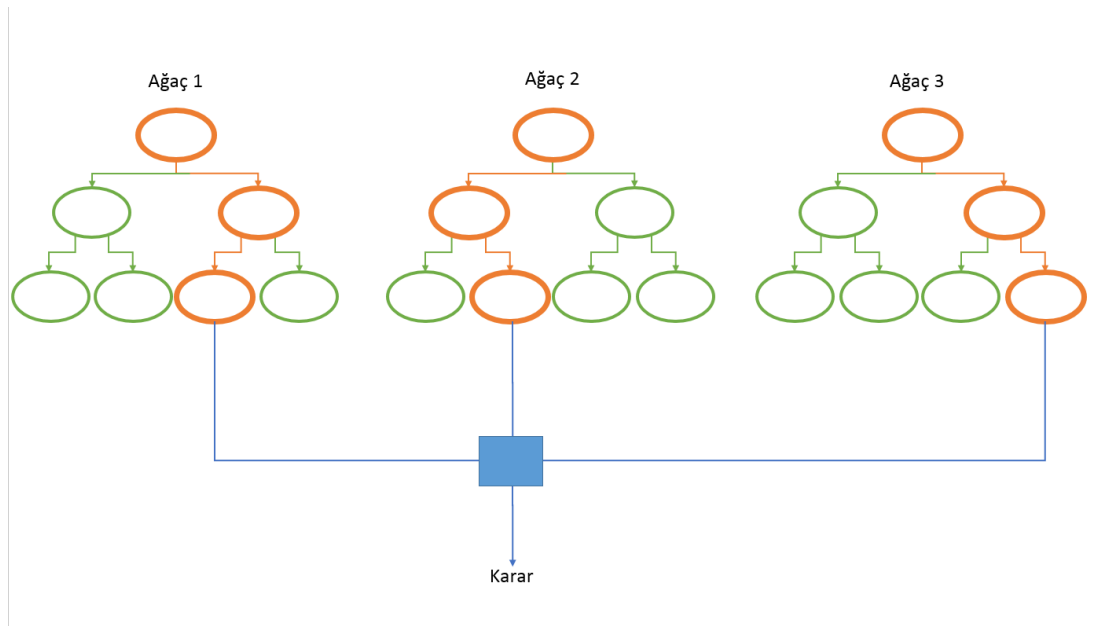
Boosting yöntemi ile her bir adımda, önceki adımda yanlış sınıflandırılan durumlara ağırlık artışı verilerek, başarısız tahminler cezalandırılmış olmaktadır. Sonuç tahminleri, yinelemeli olarak üretilen tahmin sonuçlarının ağırlıklandırılması üzerine oluşmaktadır. Bu yöntem sınıflama için oluşturulmasına rağmen, zayıf öğrencilerin çalıştırılması için de uygulanabilmektedir (Sutton, 2005).

Özetlemek gerekirse; Bagging yöntemi, genel tahmini elde etmek için tahmin sonuçlarının ortalamasını alırken; Boosting yöntemi, değişik örneklemelere bir tahmin yöntemi uygulayarak elde edilen sonuçların ağırlandırılmış ortalamasını kullanmaktadır (Sutton, 2005).

5.2.1. Rassal Ormanlar (Random Forests)

Rassal Ormanlar algoritması temeli, Şekil 5.2’de gösterildiği üzere, birden çok karar ağacının bir araya gelmesiyle bir orman yapısının oluşturulması ve ağaçların ürettiği tahmin sonuçlarının oylama (voting) ile elenmesi ya da ortalamalarının alınması fikrine dayanmaktadır (Breiman, 2001)

Bagging (Breiman, 1996) yöntemi ile Random Supspace yöntemlerinin birleşiminden oluşmaktadır.



Şekil 5.2. Rassal Ormanlar algoritması örneği (Odabaşı, 2019)

Rassal Ormanlar algoritmasında, Bagging yönteminde olduğu gibi ağaçlar için gözlemler bootstrap ile rastgele örnek seçim yöntemi kullanılır. Ancak farklı olarak değişkenler arasından rastgele olarak Random Supspace (Rastgele Alt Uzaylar) yöntemi ile rastgele bir altküme seçilir. Seçilen bu alt küme, ilgili düğümün dallanması için kullanılır. Bu rassal seçim orman içindeki ağaçların birbirine benzeme ihtimali olasılığını düşürür. Bu açıdan rastgele orman yöntemi torbalama yönteminin geliştirilmiş halidir (Breiman, 2001).

Rassal Ormanlar çalışma adımlarında, öncelikle bootstrap örnekleme ile alt uzaylar oluşturulur. Ancak gözlemlerin tümü kullanılmaz. Gözlemlerin yaklaşık %63’ü eğitim verisi olarak kullanılır. Seçilmeyen orijinal veri kümesindeki gözlemler, torba dışı

gözlemler (Out of Bag) olarak adlandırılır. Bu torba dışı gözlemler, hata oranını tahmin etmek için kullanılırlar. Hata oranı genellikle torba dışı hata oranı olarak adlandırılır ve özelliğin (değişkenin) önemini tahmin etmek için kullanılırlar. Örneklerin rastgele seçim süreci, tek bir karar ağacı üreten, her bir alt örnekte birçok kez tekrarlanır ve bunun sonucunda karar ağaçlarının bir ormanı meydana gelir. Budanmadan ve mümkün olan en geniş ölçüde ağaçlar büyümüş olur. Bireysel ağaçlar, bir sınıflandırma veya tahmin kararı yaparlar. Bir gözlemin en son tahmin edilen sınıfı, sınıflandırma için çoğunluk oyu ile veya regresyon için ağırlıklı ortalamalar yoluyla yapılır (Lewis, 2017; Şahinarslan, 2019).

Her bir ağaç için OOB (Out of Bag) hata oranı tespit edildiği gibi ormanın da hata oranı tespit edilmiş olur. Bunun sonucunda seçilen alt kümelerin veri setini ne derece iyi düzeyde açıkladığı ölçülür. Ayrıca her bir ağacın genel orman için bulunan OOB hata verisine olan katkısı hesaplanarak ağaçlar arasında bir değerlendirme yapılır. En az hata sonuçlarını veren ağaç, en çok değişkeni açıklayan ağaç olarak ormana en çok katkıyı veren ağaç olur. Böylece model sonucu olarak aynı zamanda değişken önem sıralamasının da ulaşılmaya olanak sağlanmış olur (Lewis, 2017; Şahinarslan, 2019).

Rassal Ormanlar algoritmasını oluşturmanın ve çalıştırmanın oldukça basit olması en önemli avantajlarından biridir. Son dönemlerde yapılan çalışmalar veri yapıları ve problemlerin birçoğunda ortaya çıkan regresyon ve sınıflandırma hatasını Rassal Ormanlar yönteminin kullanılmasıyla azaltılabileceğini göstermektedir. Ayrıca daha az parametreye sahip oldukları için eğitilmeleri daha hızlıdır. Minimum veri hazırlığına ihtiyaç duyarak ve büyük veri tabanlarında bile etkin şekilde verimli çalışabilmektedirler. Veri setinde yer alabilecek aykırı gözlemlere ve aşırı öğrenmeye karşı dayanıklıdırlar (Lewis, 2017).

5.2.2. Gradyan Artırma Makineleri (Gradient Boosting Machine)

Gradyan Artırma modeli, karar ağaçlarının tahminini iyileştirmek için Rassal Ormanlar algoritmasından önce geliştirilmiş olup boosting (artırma) yöntemine dayanmaktadır. Sınıflandırma ve regresyon probleminde de kullanılabilmektedir (Freidman, 2001).

Karar ağaçlarında olduğu gibi her bir veri seti için bir ağaç oluşturulmaktadır. Bir sonraki ağacı oluşturmada önce, oluşturulan ağaçların hata oranları göz önünde bulundurulmaktadır. Diğer ağaçlar ile etkileşim halinde ilerlenerek model kurulumu tamamlanmaktadır. Modelin amacı, oluşan hata oranını en aza indirmeye çalışarak ağaç oluşturmaya devam etmesidir. Model artıkları üzerine tek bir tahminsel model formunda olan modeller serisi kullanılmaktadır (Freidman, 2001).

Gradyan Artırma Regresyon modeli, özetle beş aşamadan oluşmaktadır. Öncelikle Regresyon Ağacı oluşturulmaktadır. İkinci adımda, her bir ağacın yaprağı için hata oranı hesaplanmaktadır. Hesaplanan hata oranları yeni gözlem verileri olarak kullanılmaktadır. Kurulan ikinci ağaç ikinci adımdaki hata oranlarını en aza indirmeyi hedeflemektedir. İkinci ve dördüncü adımlar tekrar edilerek ilerleme sağlanmaktadır (Şahinarslan, 2019).

GBM, zayıf öğrencilerin bir araya gelerek güçlü bir öğrenci oluşturması fikrine dayanan AdaBoost (Freund and Schapire, 1999) algoritmasının genelleştirilmiş ve kolayca uyarlanabilen versiyonudur. Regresyon problemleri için Adaboost algoritması formülasyonu aşağıdaki gibi tanımlanmaktadır.

Girdi: $S = \{(x_1, y_1), \dots, (x_N, y_N)\}$ olarak etiketlenmiş N dizisine ait S örnekleme,

S örnekleme ait D dağılımı,

A , Zayıf öğrenme algoritması (Weak learn),

T , Tamsayı ve iterasyonların özel sayısı,

Δ , bir gerçel sayı.

Başlangıç: $i = 1, 2, \dots, N$ için ağırlık vektörü $w_i^1 = D(i)$;

olacak şekilde 7 adımlık bir döngü olarak formülasyon oluşturulmaktadır.

$t = 1, 2, \dots, T$ için;

$$1. \quad p^{(t)} = \frac{w^{(t)}}{\sum_{i=1}^N w_i^{(t)}} \text{ başlanır.} \quad (5.4)$$

2. S 'den $S^{(t)}$ örneklemini $p^{(t)}$ dağılımı ile rastgele seçerek, A öğrenme algoritması ve $h_t = A_{S^{(t)}}$ hipotezi elde edilir.

3. Hata

$$4. \quad \varepsilon_{(t)} = \sum_{i=1}^N p_i^{(t)} H S (||h_t(x_i) - y_i|| - \Delta) \quad (5.5)$$

Eğer $\varepsilon_{(t)} > \frac{1}{2}$ sonra $T = t - 1$ ve döngü yarıda bırakılır.

$$5. \quad \beta_{(t)} = \frac{\varepsilon_{(t)}}{1 - \varepsilon_{(t)}} \text{ hesaplanır.} \quad (5.6)$$

6. Yeni ağırlık vektörü,

$$7. \quad w_i^{(t+1)} = w_i^{(t)} \beta^{1 - H S (||h_t(x_i) - y_i|| - \Delta)} \text{ olarak ayarlanır} \quad (5.7)$$

Çıktı hipotezi: $\alpha_t = \log \frac{1}{\beta_t}$ olduğunda;

$$h_f = \operatorname{argmax}_{y \in [0,1]^m} \sum \alpha_t H S (\Delta - ||h_t(x_i) - y_i||) \text{ 'dir.} \quad (5.8)$$

(Bertoni vd., 1997)

5.2.3. Ekstrem Gradyan Artırma (Extreme Gradient Boost, XGBoost)

Ekstrem Gradyan Artırma (XGBoost) algoritması sonuçların performansı ve çalışma hızı açısından Gradyan Artırma algoritmasının geliştirilmiş halidir. Matematiksel olarak Gradyan Artırma algoritmasını kullanmaktadır ancak hız ve tahmin performansı bakımından optimize edilen daha gelişmiş versiyonu olarak tanımlanmaktadır.

Makine öğrenmesi alanında büyük veri setine sahip problemleri çözmek için kullanılan bir algoritma paketidir. Milyonlarca örnekleme az miktarda kaynak kullanarak çözebilen verimli bir algoritmadır. Tianqi Chen tarafından, 2014 yılında geliştirilip açık kaynak kodlu makine öğrenmesi algoritmalarının bulunduğu DMLC (Distributed Machine Learning Community) adlı kütüphaneye eklenmiştir (Chen ve Guestrin, 2016).

XGBoost algoritması, milyonlarca veri setini kısa zamanda ve diğer algoritmalara göre daha az işlemci gücü ve geçici belleğe (ram) ihtiyaç duyarak analiz edebilmektedir. Bu nedenle, son dönemlerde artan verinin yoğunluğu ve verileri analiz edebilmek için gerekli olan işlem gücünün maliyetini azaltmak için bu modele yönelim artmıştır.

Örneğin, Avrupa Nükleer Araştırma Merkezi (CERN), Büyük Hadron Çarpıştırıcısı'nın ürettiği yıllık 3 petabayt (PB, 1024 terabayt) veriyi analiz etmek için XGBoost algoritmasını kullanmıştır.

XGBoost, verideki her değeri incelemek yerine veriyi parçalara böler ve bu parçalara göre çalışmaktadır. Bu nedenle hız bakımından Gradyan Artırma algoritmasına göre iyi bir performans göstermektedir. Ayrıca XGBoost algoritması eksik verilerle çalışma imkanı sağlamaktadır.

5.2.4. Light Gradyan Artırma (Light Gradient Boosting, LightGBM)

LightGBM algoritması karar ağaçlarını kullanan farklı bir gradyan artırma modelidir. 2017 yılında Microsoft tarafından geliştirilmiştir.

XGBoost algoritmasının eğitim süresi performansını arttırmaya yönelik geliştirilen bir algoritmadır. Diğer karar ağaçları analiz sürecinde dikey bir şekilde dallanma yaparken, LightGBM yapraklar üzerinden yatay dallanma yoluna gitmektedir. Bu nedenle performans bakımından eğitim süresi daha kısa olmaktadır (Ke vd., 2017).

LightGBM diğer Gradyan Artırma modelleri ile karşılaştırıldığında hızlı işlem yapması, daha büyük verileri işleyebilmesi, daha az bilgisayar hafızasına (ram) duyması, eksik verilerle çalışma imkanı sağlaması gibi avantajları vardır.

LightGBM, veri seti üzerinde diğer modellere nispetle daha kolay ezberle öğrenme (overfitting) sağladığından büyük veri setlerinde daha iyi sonuçlar vermektedir. Farklı veri setleri üzerinde yapılan çalışmalarda LightGBM modelinin veriyi öğrenme sürecinin diğer modellere göre 20 kat daha hızlı olduğu sonucuna ulaşılmıştır (Ke vd., 2017).

5.2.5. Kategorik Artırma (Category Boosting, CatBoost)

Kategorik değişkenlerle çalışması için Yandex tarafından 2017 yılında geliştirilen hızlı ve başarılı bir Gradyan Artırma algoritması olarak son dönemlerde tercih edilen algoritmalarından biridir. Kategorik değişkenlerin otomatik olarak modele alınmasını sağlayan kod blokları algoritmaya gömülmüştür. Karar Ağaçları modellerinde olduğu gibi kategorik değişkenlerin modele alınması için sayısallaştırma işlemi yapılmasına gerek yoktur. Kategorik değişkenlerin kardinalitelerinin (sınıf sayılarının) fazla olduğu durumlarda bu özellikle hız kazandırmakta ve model performansını arttırmaktadır. Bu

nedenle, kategorik deęişkenlerin veri setinde daha çok olduęu durumlarda daha iyi performans göstermektedir.

Veri ön hazırlıęının kısa olması, boş deęerler ile çalışabilmesi, hızlı ve ölçeklenebilir GPU desteęi kılması, modeli performans bakımından iyi bir noktaya getirmektedir.

Catboost simetrik ağaęlar kurarak, derin ağaęlar kurmadan yüksek tahmin oranı yakalamaya çalışmaktadır. Simetrik ağaęlar ile rassallıęı sağladığından aşırı öğrenme sorununu da aşmaktadır.

6. UYGULAMA

Bu bölümde, Ağaç Temelli Makine Öğrenmesi yöntemlerinin, 3. Bölümde anlatılan B2C (Business to Customer) türünde olan Google Merchandise Store e-ticaret sitesi verilerine uygulanması gerçekleştirilmiştir. Uygulama kapsamında veri setine Karar Ağaçları, Rastal Ormanlar, GBM, LightGBM, XGBoost ve CatBoost modelleri uygulanmıştır.

Uygulama aşamalarında modellerin kurulumu, hızlı ve kullanışlı bir dil olan Python 3.7 programlama dili kullanılmıştır. Veri setini yerel bilgisayara indirmeye gerek duymadan, internet yoluyla çalışma yapılabilen Google Colab Pro platformu ile çalışma gerçekleştirilmiştir.

6.1.Problemin Tanımlanması

E-ticaret sektörünün büyük ölçüde artış göstermesiyle web sitesi üzerindeki müşterilerin davranışlarını da incelemek ve buna göre aksiyon almak çok önemli bir hal almaktadır. Buna göre web sitesi üzerinde gezinen veya alışveriş yapan müşterilerin web sitesi üzerinde bıraktığı bilgileri üzerinden; müşteri segmentasyonu, kampanya - reklam başarı oranları, trafik bilgileri ve ürün satın alma tahmin işlemlerine ulaşmak şirketler için hayli önem kazanmaktadır.

E-ticaret sektöründe web sitesi üzerindeki müşteri bilgileri ile satın alım yapan veya almayan müşterilerin tahmini ve bunları etkileyen faktörler yakın zamanda incelenmiştir (Özdemir ve Turanlı, 2021). Bu çalışma kapsamında ise yine web sitesi üzerindeki müşteri bilgileri kullanılarak tüketicinin harcamaları üzerine tamin modellerinin kurulması amaçlanmıştır. Ağaç temelli makine öğrenmesi yöntemleri kullanılarak en iyi hata metriğine ulaşan model ve harcama üzerinde etkili olan değişkenler tespit edilmiştir. Böylece çalışmanın sonucu olarak şirket ulaşacağı satış öngörülleri ile aynı zamanda harcamayı etkileyen faktörlere (değişkenlere) de erişmiş olacaktır. Bu öngörülere göre; kampanya, müşteriye özel kuponlar ve müşteri takibinin sağlanması gibi aksiyonlar almasına imkan olacaktır.

6.2.Verinin Toplanması

Çalışma kapsamında global bir marka olan Google Merchandise Store e-ticaret sitesi verileri kullanılmıştır. Veriler, kaggle.com üzerinde Google tarafından erişim izni verilerek yayınlanmıştır. Veriler, 2 Eylül 2016 ile 04 Ocak 2017 tarihleri arasında yer almakta, 903653 gözlem ve 55 değişken içermektedir.

Çalışma kapsamında alınan veriler, web sitesi içine gömülen bir kod ve şirketin onayı ile Google Analytics tarafından toplanmaktadır. Böylece müşteri davranışları verileri Google Analytics platformu üzerinde yer almakta, istenen zaman aralıklarında filtreler kullanılarak bu verilere erişim sağlanmaktadır. Anlık olarak davranışların takip edilmesini sağlayan bu platform birçok şirket tarafından tercih edilmektedir. Yine aynı şekilde platform üzerinde müşteri davranışlarının görselleştirilip filtrelenmesi ve raporlanması mümkün olmaktadır.

Çalışma kapsamında alınan veri seti, büyük hacim kapladığından bazı değişkenler (Totals, Device, Geo, Traffic Source) json formatında ayrıca kümelenerek tutulmaktadır. Bu nedenle ilk olarak yayınlandığı haliyle Çizelge 6.1’de olduğu gibi değişkenler ve anlamları gösterilmektedir. Veri Ön İşleme bölümünde detaylı olarak değişkenler ele alınmaktadır.

Çizelge 6.1. Kümelenmiş değişkenler

Değişken	Açıklama
Full Visitor ID	Müşteri numarası
Channel Grouping	Müşterinin siteye giriş yaptığı kanal
Date	Müşterinin web sitesine giriş yaptığı tarih
Device	Müşterinin web sitesine giriş yaptığı cihazların bilgilerini içerir.

SocialEngagementType	Katılımcı türü (Sosyal etkileşimli – sosyal etkileşimli değil)
Totals	Oturum boyunca toplanan numerik metrikleri içerir.(Tıklama, oturum sayısı, sayfa görüntüleme vb.)
Traffic Source	Oturumun başlatıldığı trafik kaynağı hakkında bilgiler içerir. (Müşterinin hangi reklamdan, kampanyadan veya sayfadan geldiği gibi bilgiler içerir.)
Visit Number	Müşterinin sayfayı oturum boyunca yaptığı ziyaret sayısı
Hits	Sayfa ziyaretinde oturum boyunca olan tıklama sayısı
Geo	Müşterilerin coğrafi bilgilerini içerir. (Ülke, bölge, şehir, enlem, boylam vs.)

6.3. Veri Ön İşleme

Bu bölümde, makine öğrenmesi model kurulumu öncesi gerekli olan veri ön işleme adımları uygulanmıştır. Bu ön işleme adımlarında, Python programlama dili altında yer alan Pandas Kütüphanesi kullanılmıştır. Pandas kütüphanesi yardımı ile veri setinde satır-sütun dönüşümleri, değişkenler üzerine analizler ve eksik veri tespitleri hızlı bir şekilde yapılabilmektedir.

Veri ön işleme aşamasında ilk olarak, bu aşamada önemli bir yer tutan değişken mühendisliği yapılmıştır. Veri setinin satır ve değişken sayısının fazla olması nedeniyle veri setinin daha az yer tutması için json dosyası formatında bazı değişkenler kümelenerek tutulmuştur. Kümelenen değişkenlerin json dosyası formatından çıkarılması ile 55 tane değişkene ulaşılmıştır. Ancak bu değişkenlerin hepsi model için uygun değildir. Bu nedenle, öncelikle hangi değişkenlerin anlamlı ve model için uygun olup olmadığı ele alınmıştır.

Veri setinde ID içeren değişkenler, satırları tamamıyla birbirinden farklı karakter içeren kategorik değişkenler olmak üzere 17 tane değişken veri setinden çıkarılmıştır. (Müşteri numarası, oturum numarası, yönlendirme sitesi isimleri vb.)

Veri seti incelendiğinde; 19 tane değişkenin varyansının 0 olduğu yani tüm değişken boyunca aynı veriyi içerdiği görülmüştür. Bu değişkenler; “web sitesinin rengi”, “enlem”, “boylam”, “sitede reklam olup olmadığı” gibi web sitesinin topladığı yüzeysel bilgilerden oluşmaktadır. Bu değişkenlerin modele konulması uygun değildir. Çünkü herhangi bir bilgi taşımadığından sadece modelin performans yükünü arttırıp model sonucu üzerinde anlamlı bir etkisi olmamaktadır.

Tarih formatında veri içeren Tarih değişkeni, gün ve ay olarak değişkenlere ayrılıp veri setine eklenmiştir. Tarih formatında veri içeren bu değişkenin modele konulması mümkün değildir. Bu nedenle tarih değişkeni içinde gün ve ay kırılımının önemli bir bilgi içereceği varsayıldığından ayrıca değişken olarak veri setine eklenmiştir. Gün ve Ay değişkenleri ile müşterinin web sitesine giriş yaptığı bu bilgileri ayrı değişken olarak ele almak modele ekstra olarak bilgi girişi sağlamıştır. Böylece müşterilerin siteye giriş yaptığı günler ve aylar üzerinden bir kıyas yaparak analiz fırsatı doğmuş, hem de bağımlı değişken üzerinde etkili olup olmadığı model aracılığıyla test edilmiştir.

Veri seti çok fazla sayıda kategorik değişken, ancak az sayıda sürekli değişken içermektedir. Bu sürekli değişkenler; bağımlı değişken olan Harcama Miktarı değişkeni ile Sayfa Görüntülenme Sayısı, Tıklama Sayısı ve Ziyaret Sayısı değişkenlerinden oluşmaktadır. Çizelge 6.2’de sürekli değişkenler İngilizce ve Türkçe olarak detaylı şekilde gösterilmektedir. Google Analytics platformunda değişken isimleri İngilizce olarak yer aldığı için İngilizce gösterimi de çalışmaya eklenmiştir ancak Türkçe karşılıkları ile çalışmaya devam edilmiştir.

Çizelge 6.2. Sürekli değişkenler

Değişken	Türkçesi	Açıklama
Totals transaction Revenue	Harcama Miktarı	Hedef (bağımlı) değişken. Müşterinin oturum içinde yaptığı harcama miktarı
Totals Pageviews	Görüntülenme Sayısı	Sayfa görüntüleme sayısı
Totals Hits	Tıklama Sayısı	Sayfada yapılan tıklama sayısı
Visit Number	Ziyaret Sayısı	Müşterinin sayfayı oturum boyunca yaptığı ziyaret sayısı

Çizelge 6.2’de gösterilen sürekli değişkenler içinde Görüntülenme Sayısı değişkeninde 100 tane eksik veri değerine rastlanmış ve bu satırlar çıkarılmıştır. Diğer değişkenler içinde eksik veriye rastlanmamıştır.

Veri setinde kullanılan kategorik değişkenler Çizelge 6.3’te yer almaktadır. Makine öğrenmesi modellerinde kategorik (karakter içeren) değişkenlerin sayısallaştırılması gerektiğinden Python yardımı ile modele konulmak üzere dummy değişken dönüşümü yapılmıştır. Tüm kategorik değişkenler nominal olduğu için, Python dili altında yer alan Encoder yardımı ile bu işlem yapılmıştır. Böylece, modellerin bu değişkenleri sayısal olarak görerek çalışmasına imkan sağlanmıştır.

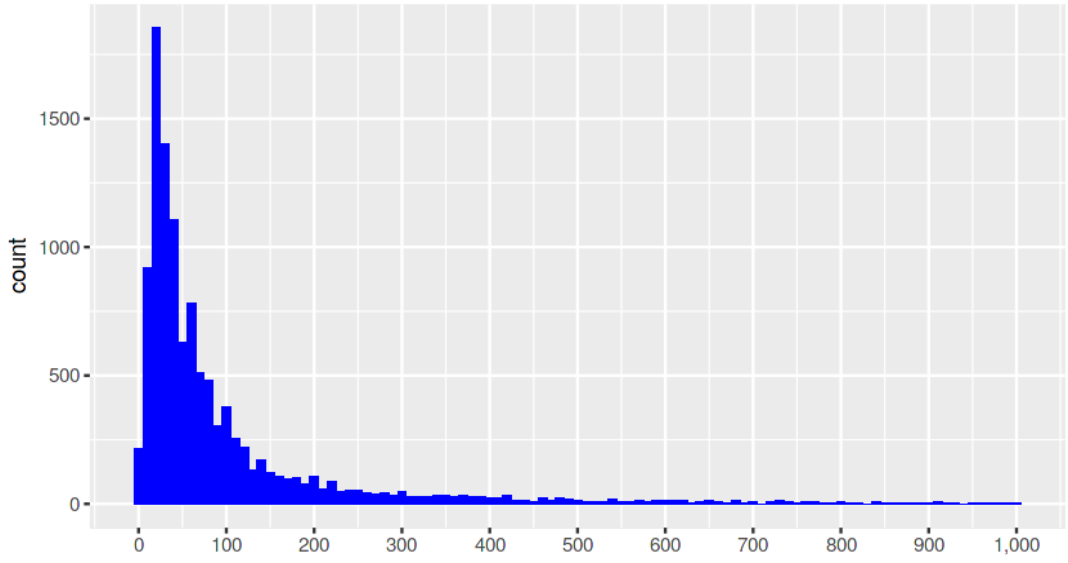
Çizelge 6.3. Kategorik değişkenler

Değişken	Türkçesi	Açıklama
Totals.New Visits	Eski/Yeni Müşteri	Müşterinin, daha önceden siteyi ziyaret edip veya etmediği bilgisini içerir. (Yeni müşteri:1, Eski müşteri:0)
Channel Grouping	Kanal	Müşterinin siteye giriş yaptığı kanal

totals.bounces	Çıkma Hatası	Müşterinin web sitesinde oturum içinde çıkma hatası alıp almadığı bilgisini içerir. (Hata aldı, Hata almadı)
Device Browser	Tarayıcı	Müşterinin web sitesine ulaştığı tarayıcı bilgilerini içerir. (Chrome, Safari vb.)
device.deviceCategory	Cihaz	Müşterinin web sitesine giriş yaptığı cihaz (Telefon, tablet, Masaüstü bilgisayar)
device.operatingSystem	İşletim Sistemi	Müşterinin giriş yaptığı cihazın işletim sistemi bilgisi (Windows, Macintosh vb.)
geoNetwork.city	Şehir	Müşterinin web sitesine giriş yaptığı şehir
geoNetwork.continent	Kıta	Müşterinin web sitesine giriş yaptığı kıta
geoNetwork.country	Ülke	Müşterinin web sitesine giriş yaptığı ülke
geoNetwork.subcontinent	Alt Kıta	Müşterinin web sitesine giriş yaptığı alt kıta (Kuzey Amerika, Batı Amerika vb.)
trafficSource.isTrueDirect	Kaynak Kullanımı	Müşterinin kaynaktan gelip gelmediği bilgisini içerir. (Doğru, Yanlış)
trafficSource.campaign	Kampanya Adı	Müşterinin trafik kaynağının kampanyası
trafficSource.source	Kaynak	Müşterinin hangi kaynaktan (Google, Youtube vb.) geldiği bilgisini içerir.
trafficSource.medium	Organik/Referans	Müşterinin organik veya referans üzerinden geldiği bilgisini içerir.

trafficSource.adContent	Reklam Adı	Müşterinin hangi reklam üzerinden geldiği bilgini içerir.
-------------------------	------------	---

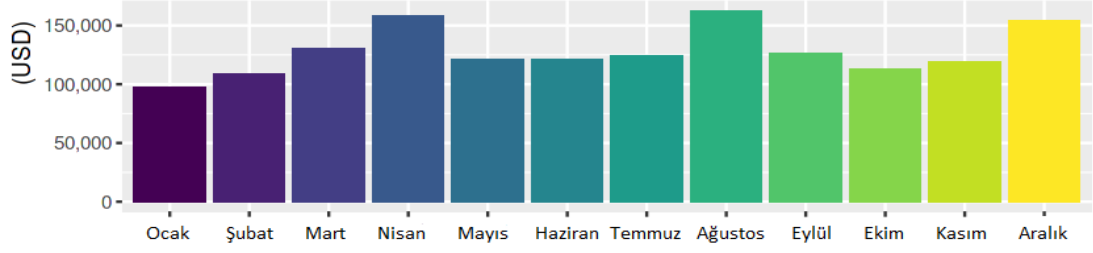
Değişken mühendisliği yapıldıktan sonra model kurulumu öncesi analiz için görselleştirme yapılmıştır. Görselleştirme işlemi için Python dili altında yer alan Matplotlib ve Seaborn Kütüphaneleri kullanılmıştır. İlk olarak, bağımlı değişkenin dağılımı incelenmiştir. Bağımlı (hedef) değişken olan Harcama Miktarı (Totals Transaction Revenue) dağılımı Şekil 6.1’de incelendiğinde sağdan çarpık olduğu görülmektedir. Bu çarpıklıktan dolayı, çalışma kapsamında seçilen ağaç temelli modellerin performansı önem teşkil etmektedir. Çünkü bu modeller, aykırı değerlerden fazla etkilenmeyerek çarpık dağılıma sahip bağımlı değişkenlerin olduğu problemlerde de başarılı tahmin performansları göstermektedir.



Şekil 6.1. Harcama Miktarı Histogramı

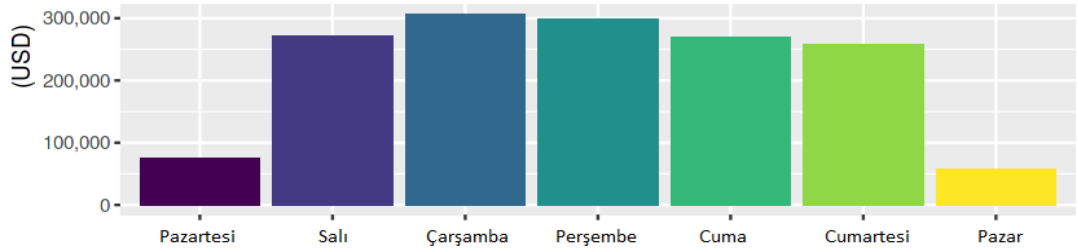
Bağımlı değişken üzerinde diğer bazı değişkenlerin kırılımlarını görmek de veri ön işleme adımları içinde önemlidir. Böylece model öncesi analiz yapma fırsatı olmakta ve problem daha iyi anlaşılmaktadır.

Şekil 6.2’de Harcama Miktarı’nın aylara göre toplam değerlerine bakıldığında ağustos ayında en çok toplam harcamanın olduğu görülmektedir. Ağustos ayından sonra sırasıyla nisan ve aralık ayları gelmektedir.



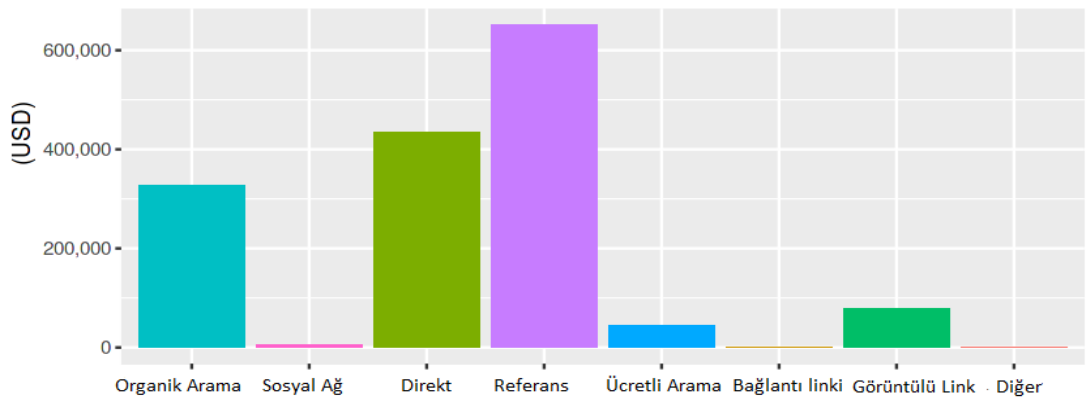
Şekil 6.2. Aylara göre toplam Harcama Miktarı

Şekil 6.3'te Harcama Miktarı'nın günlere göre toplam değerlerine bakıldığında, çarşamba gününde en çok toplam harcamanın olduğu görülmektedir. Ardından sırasıyla perşembe ve salı günleri gelmektedir.



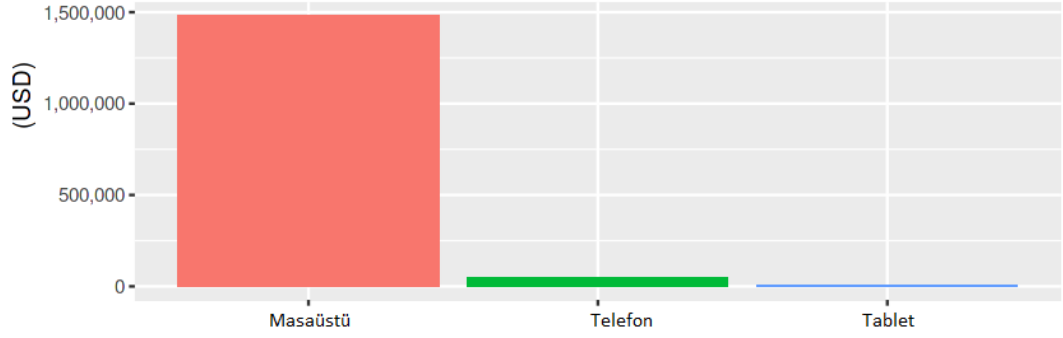
Şekil 6.3. Günlere göre toplam Harcama Miktarı

Şekil 6.4'de Harcama Miktarı ile Kanal kırılımı incelenmektedir. Grafikte görüldüğü gibi, Harcama Miktarı toplamı en çok Referans linki ile web sitesine gelen müşterilerden oluşmaktadır. Ardından sırasıyla, Direkt (Anasayfadan siteye gelen müşteri) ve Organik Arama (Arama motoru üzerinden gelen müşteri) kanalları gelmektedir.



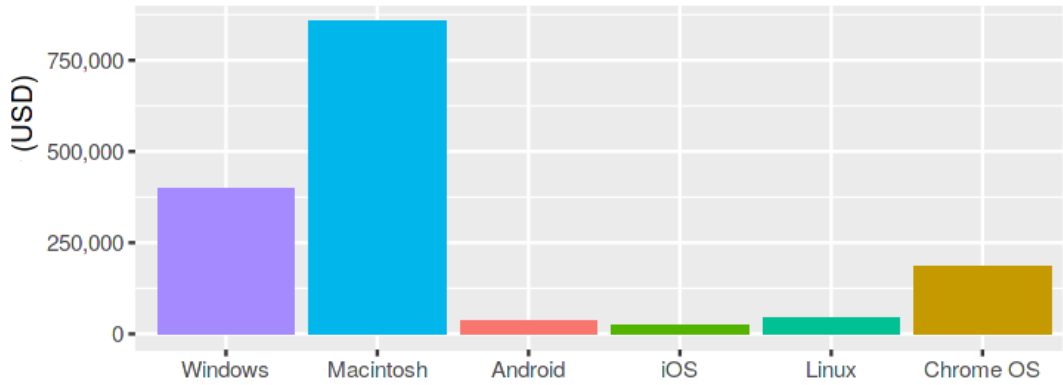
Şekil 6.4. Harcama Miktarı ile Kanal kırılımı

Şekil 6.5’de Harcama Miktarı ile Cihaz arasındaki kırılım incelenmektedir. Masaüstü kullanıcılarının, Telefon ve Tablet üzerinden gelen müşterilere göre oldukça fazla harcama toplamına ulaştığı görülmektedir.



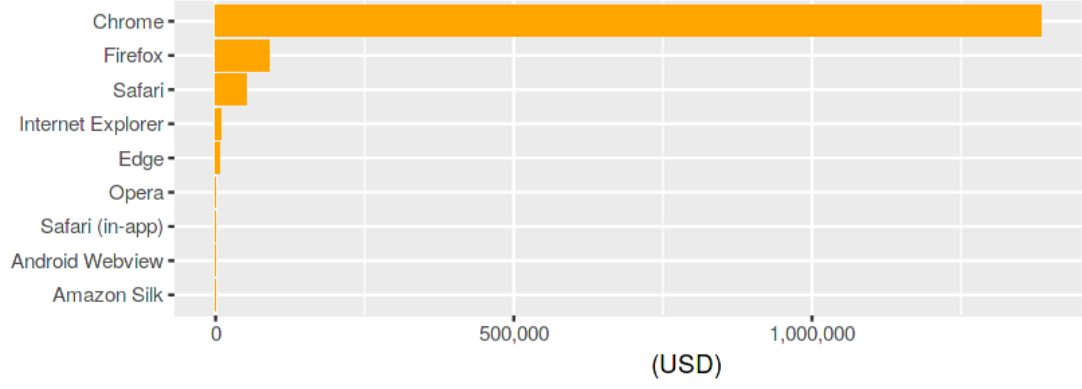
Şekil 6.5. Harcama Miktarı ile Cihaz kırılımı

Şekil 6.6’da Harcama Miktarı ile İşletim Sistemi kırılımı incelenmektedir. Macintosh işletim sistemini tercih eden kullanıcılarının en çok harcama miktarına ulaştığı görülmektedir. Ardından sırasıyla Windows ve Chrome OS işletim sistemleri yer almaktadır.



Şekil 6.6. Harcama Miktarı ile İşletim Sistemi kırılımı

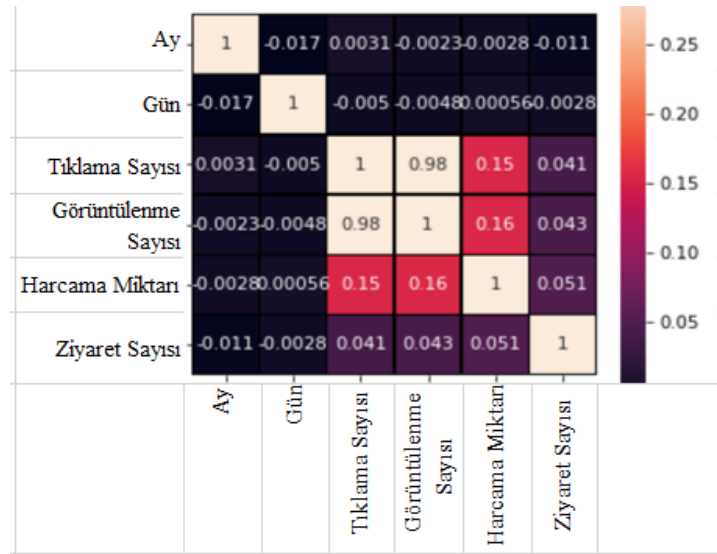
Harcama Miktarı ile Tarayıcı kırılımı Şekil 6.7’de incelenmektedir. Chrome tarayıcısını tercih eden kullanıcıların, oldukça büyük bir farkla, toplam harcama miktarında ilk sırada yer aldığı görülmektedir. Ardından sırasıyla Firefox ve Safari tarayıcıları gelmektedir.



Şekil 6.7. Harcama Miktarı ile Tarayıcı kırılımı

Veri setinde yer alan sürekli değişkenlerin aralarındaki ilişkinin belirlenmesi için Pearson Korelasyon analizi uygulanmıştır. Değişkenler arasındaki korelasyon katsayıları, Matplotlib kütüphanesi yardımıyla Şekil 6.8’de olduğu gibi tablo şeklinde gösterilmektedir. Bu tablo ile değişkenler arasındaki ilişkinin şiddeti ve yönü korelasyon katsayıları ile incelenebilmektedir.

Korelasyon katsayısı, değişkenler arasındaki ilişkinin yönü ve büyüklüğünü belirten katsayıdır. Bu katsayı, (-1) ile (+1) arasında bir değer almaktadır. Pozitif değerler; pozitif yönlü doğrusal ilişkiyi; negatif değerler ise negatif yönlü bir doğrusal ilişkiyi belirtmektedir. İlişkinin büyüklüğü genellikle; korelasyon katsayısı 0 ile 0.29 arasında ise düşük düzeyde ilişki, 0.30 ile 0.70 arasında ise orta düzeyde ilişki ve 0.71 ile 1 arasında ise yüksek düzeyde ilişki olduğu ifade edilmektedir. (James vd., 2017)



Şekil 6.8. Korelasyon matrisi

Şekil 6.8 incelediğinde; bağımlı değişken olan Harcama Miktarı ile sırasıyla yüksekten düşüğe olmak üzere Görüntülenme Sayısı, Tıklama Sayısı, Ziyaret Sayısı, Ay ve Gün değişkenleri arasında düşük düzeyde ilişki bulunmaktadır. Ayrıca Tıklama Sayısı ile Görüntülenme Sayısı arasında yüksek düzeyde ve pozitif yönlü korelasyon olduğu görülmektedir. Lineer modellerde bu tür durumlarda genel yaklaşım gereği, bağımlı değişken üzerinde daha az etkili olan değişkeni çıkarmak olasıdır. Ancak ağaç temelli modeller, korelasyon ve çoklu bağlantı problemlerinden fazla etkilenmemektedirler. Bu nedenle aralarında yüksek korelasyon olan iki değişken de çalışmada kullanılmıştır.

6.4.Modelin Oluşturulması ve Sonuçların Değerlendirilmesi

Çalışma kapsamında; sürekli, kategorik, Gün ve Ay değişkenlerinin oluşturduğu 21 değişkenli veri seti ile modeller kurulmuştur. 20 bağımsız değişken ve 1 bağımlı değişkenden oluşan veri seti Şekil 6.9’da detaylı olarak gösterilmektedir. Veri ön işleme yapılarak kullanılan 21 değişken, web sitesi üzerinden benzer şekilde alınan verilerle olan çalışmalara da kaynak olacaktır.

Değişkenler
Harcama Miktarı (Bağımlı değişken)
Görüntülenme Sayısı
Tıklama Sayısı
Ziyaret Sayısı
Gün
Ay
Eski / Yeni Müşteri
Kanal
Çıkma Hatası
Tarayıcı
Cihaz
İşletim Sistemi
Şehir
Kıta
Ülke
Alt Kıta
Kaynak Kullanım
Kampanya Adı
Kaynak
Organik / Referans
Reklam Adı

Şekil 6.9. Modelde kullanılan değişkenler

Makine öğrenmesi süreci gereği, model öncesi ezber öğrenmeyi engellemek ve performans olarak daha iyi sonuç almak adına Şekil 6.9’da yer alan değişkenler ile oluşturulan veri seti; eğitim ve test olarak ikiye ayrılmıştır. Çalışma için veri seti; %70 eğitim ve %30 test veri seti olmak üzere Holdout yöntemi ile ayrılmıştır.

Veri seti, eğitim ve test seti ayrıldıktan sonra modellerin çalıştırılması için Python dili ve kütüphaneleri kullanılmıştır. Bağımlı değişken Harcama Miktarı “Y” ve geriye kalan bağımsız değişkenler “X” diye ayrılarak Python dili yardımıyla Y ile X arasında denetimli olarak ağaç temelli makine öğrenmesi modellerinin çalışması sağlanmıştır. Ağaç temelli makine öğrenmesi modellerinde yer alan matematik formülasyonları gereği bu modelleri oluşturmak oldukça zordur. Bu nedenle Python dilinde gömülü olarak yer alan, makine öğrenmesi modelleri için kullanılan, Sklearn kütüphanesinden yararlanılmıştır. Sklearn kütüphanesi, Python diliyle geliştirilen açık kaynak kodlu istatistiksel modellerin yer aldığı oldukça sık kullanılan bir araçtır. Diğer kodlama dilleri olan R veya Spark dillerinde de benzer kütüphaneler mevcuttur ancak kullanım kolaylığı ve sürekli güncellenmesi nedeniyle çalışma kapsamında Python dili tercih edilmiştir. Modellerin her biri ayrıca alt yapı gerektirdiği için Sklearn kütüphanesi içinde, modellerin kendi adlarında alt kütüphaneler yer almaktadır. Çizelge 6.4’de, çalışma kapsamında kullanılan modeller ve kod blokları yer almaktadır. Modeller için kullanılan kütüphaneler, kod yardımıyla çağrılarak modellerin kurulumu yapılmaktadır.

Örneğin, Karar Ağaçları modelini kurmak için Sklearn.tree kütüphanesinden, DecisionTreeRegressor kütüphanesi gereklidir. DecisionTreeRegressor kütüphanesi içinde, Karar Ağaçlarını modelini kurmak için gerekli olan parametreler ve matematiksel altyapı mevcuttur.

Çizelge 6.4. Kullanılan modeller, kütüphaneler ve kod blokları (Scikit-Learn, 2021)

Modeller	Kütüphaneler	Kod Blokları
Karar Ağaçları	Sklearn.tree	from sklearn.tree import DecisionTreeRegressor
Rassal Ormanlar	Sklearn.ensemble	from sklearn.ensemble import RandomForestRegressor
GBM	Sklearn.ensemble	from sklearn.ensemble import GradientBoostingRegressor
LightGBM	Lightgbm	from lightgbm import LGBMRegressor
XGBoost	Xgboost	from xgboost import XGBRegressor
CatBoost	Catboost	from catboost import CatBoostRegressor

Çalışmada, Çizelge 6.4’de yer alan kütüphaneler yardımıyla Karar Ağaçları, Rassal Ormanlar, GBM, LightGBM, XGboost ve CatBoost modelleri eğitim seti üzerinde eğitilmiş ve test verisi üzerinde model başarı metrikleri ile performanslarına erişilmiştir. Ayrıca, K-Fold Cross Validation yöntemi eğitim seti üzerinde kullanılmıştır. Veri setinin satır sayısının büyüklüğü göz önüne alınarak K değeri 5 olarak seçilmiştir. K değerine daha büyük bir sayı verilmesi hem modellerin çalışmasını süresini uzatacak hem de aşırı öğrenmeye yaklaştıracaktır. Bu nedenle K değeri ile eğitim setinin 5 parçaya bölünerek modellerin kendilerini 5 kere test etmesi sağlanmıştır.

Oluşturulan modellerin hata metrikleri, bağımlı değişkenin sürekli olmasından dolayı regresyon metrikleri olarak yaygın olarak kullanılan MSE, RMSE ve MAE metrikleri seçilmiştir. Oluşturulan modellerin test verisi üzerindeki hata metrikleri aşağıdaki Çizelge 6.5’te gösterilmektedir.

Çizelge 6.5. Model hata metriklerinin karşılaştırılması

Modeller	Hata Metrikleri		
	Ortalama Hata Kare (MSE)	Kök Ortalama Hata Kare (RMSE)	Ortalama Mutlak Hata (MAE)
Karar Ağaçları	5,269	2,295	0,321
Rassal Ormanlar	2,820	1,680	0,316
GBM	2,696	1,641	0,324
LightGBM	2,685	1,608	0,320
XGBoost	2,644	1,627	0,312
CatBoost	2,599	1,612	0,324

Regresyon hata metrikleri gereği hata oranının düşmesi, modelin başarısını arttırmaktadır. Modellerin, Ortalama Mutlak Hata oranları birbirine yakın görünmektedir. Ancak modeller En Küçük Karelere göre çalıştırıldığından MSE ve RMSE metrikleri üzerinden kıyaslama yapmak daha doğru olacaktır. Ortalama Kare Hata'nın birbirine yakın olduğu veya çok büyük olduğu durumlarda RMSE'ye başvurmak daha tercih edilen bir yoldur. Bu nedenle RMSE'ye göre kıyaslama yapıldığında en başarılı olan modelin LightGBM olduğu görülmektedir. Ardından çok küçük bir farkla CatBoost yer almaktadır. Ardından sırasıyla XGBoost, GBM, Rassal Ormanlar ve Karar Ağaçları gelmektedir. Ağaç Temelli Topluluk Öğrenme modellerinin tümü, Karar Ağaçları'ndan daha başarılı sonuçlar elde etmektedir.

Çizelge 6.5'ten görüldüğü üzere, en iyi tahmin başarısını LightGBM göstermiştir. Eğer bu veri seti ile bir tahmin modeli oluşturulacaksa daha az hata değerine sahip olan LightGBM modelini kullanmak doğru olmaktadır. LightGBM modeli ile 20 tane bağımsız değişken girdisi kullanarak Harcama Miktarı değerinin tahmini, anlık ve canlı olarak gerçekleştirilebilir. Böylece müşterilerin harcamalarını, gerçeğe en yakın şekilde Google Analytics verilerini kullanarak tahmin etmek mümkün olmaktadır. Ancak modeli daha iyi hale getirmek için model optimize edilebilir. Modeller, Çizelge 6.4'te gösterildiği gibi kütüphaneler ile kurulmaktadır. Ancak bu modellerin kütüphaneleri içinde yer alan; ağacın derinliği, yaprak sayısı ve öğrenme hızı gibi modelin kullandığı parametreler sadece temel model kurulmasına yardımcı olmaktadır. Temel modeller kurulduktan sonra, en iyi sonucu veren model üzerine parametre iyileştirilmesi yapılmaktadır. Böylece modelin veri setine daha da uygun hale getirilmesi sağlanmaktadır. Çalışma kapsamında da LightGBM modeli, en iyi

tahmin algoritması olduğu için, bu model üzerine parametre optimizasyonu yapılmaktadır. Ağaç temelli modellerin genel olarak barındırdığı parametreler benzer olmaktadır. Ancak model kurulumu için birçok parametre olduğundan, modeli iyileştirmek için tüm parametreleri ele almak mümkün değildir. Ayrıca modeller, barındırdıkları tüm parametreleri optimize etme imkanı vermemektir. Bu nedenle, LightGBM modelinin iyileştirilmesi için genel olarak ele alınan önemli parametreler, Çizelge 6.6.'da gösterilmektedir.

Çizelge 6.6. Parametreler ve tanımları (LightGBM, 2021)

Parametreler	Tanımları
Objective	Problemin tipinin tanımlandığı parametredir. Problemin tipine göre sınıflandırma, regresyon veya çoklu regresyon şeklinde tanımlanabilir.
Metric	Modelin başarısını değerlendirmek için kullanılan değerlendirme ölçütünün olduğu parametredir. Objective parametresine göre belirlenir. (MSE, RMSE, Accuracy vb.)
Max_depth	Kurulacak ağacın derinliğini belirlemek için kullanılır. Eğer değeri çok arttırılırsa aşırı öğrenmeye, değeri olması gerekenden az olarak belirlendiğinde ise eksik öğrenmeye neden olur.
Num_leaves	Ağacın karmaşıklığını kontrol etmede kullanılan parametredir. Ezbere öğrenmeden (overfitting) kaçınmak için num_leaves değerinin $2^{(max_depth)}$ değerinden küçük olması gerekmektedir. Örneğin; max_depth 7 olarak belirlendiğinde, num_leaves değerinin 127 olarak ayarlamak aşırı öğrenmeye neden olabilir. Bu nedenle, bu parametre değerini 70 veya 80 olarak ayarlayarak daha iyi doğruluk elde edilebilir.
Min_data_in_leaf	Ağacın her bir yaprağında bulunan minimum gözlem sayısını ifade etmektedir. Aşırı öğrenmeyi engellemek için kullanılan önemli parametrelerden biridir. Parametrenin değeri, veri büyüklüğü ve num_leaves parametrelerine bağlıdır.
Num_iterations	Gerçekleştirilecek olan boosting turlarının sayısını kontrol eder. LightGBM, öğrenciler olarak karar ağaçlarını kullandığından, bu parametre aynı zamanda “ağaç sayısı” olarak da düşünülebilir. Num_iterations sayısı arttıkça eğitim süresi de artar.
Learning_rate	Kurulan ağaçları ölçeklendirmek için 0-1 arasında verilen bir değerdir. Learning rate değeri küçüldükçe tahmin performansı artar ancak bununla birlikte öğrenim süresi ve aşırı öğrenme ihtimali de artış gösterir.
Colsample_by_tree	Her bir ağacı eğitmek için kullanılan rastgele seçilen özelliklerin (değişkenlerin) oranıdır. Özellikle değişken sayısının büyük olduğu durumlarda kullanışlıdır. Örneğin, 50 değişken var ise bu parametre değeri 0.4 alındığında, rastgele seçilen 20 değişken ile model sınanmış olur.

Parametrelerin her birini tahmini olarak belirlemek mümkün değildir. Kullanılacak parametrelerin kendi içindeki etkileşimi ve çarpımları da modele etki etmektedir. Bu nedenle, Grid Search Algoritması kullanılmaktadır. Bu algoritma ile sözlükler şeklinde tanımlanan, modelde denenmesi istenen parametreler ve değerleri için bütün kombinasyonlar ile eğitim verisi üzerinde ayrı ayrı modeller kurulmaktadır. Belirlenmiş olan metriğe göre başarılı olan parametre değerleri algoritma yardımıyla bulunmaktadır (Burkov, 2019).

Çalışma kapsamında, Grid Search algoritması ile LightGBM modeli üzerinde, parametreler Çizelge 6.7’de gösterilmektedir. Listeler şeklinde verilen değer aralıkları ile eğitim seti üzerinde denetlenmektedir. Listeler içindeki değerlerin kombinasyonları, eğitim seti üzerinde denenerek modelin tahmin performansını optimize edecek şekilde en iyi parametre değerleri bulunmaya çalışılmaktadır.

Çizelge 6.7. Parametreler ve sınanan değerleri

Parametreler	Sınanan Değerleri
Colsample_bytree	[0.4, 0.5, 1]
Learning_rate	[0.01, 0.1, 0.15, 0.5]
Num_leaves	[5, 10, 50, 100]
Max_depth	[5, 10, 50]
Num_iterations	[40, 100, 200, 1000]

Grid Search algoritması sonucunda en iyi parametre değerleri Çizelge 6.8’de gösterilmektedir.

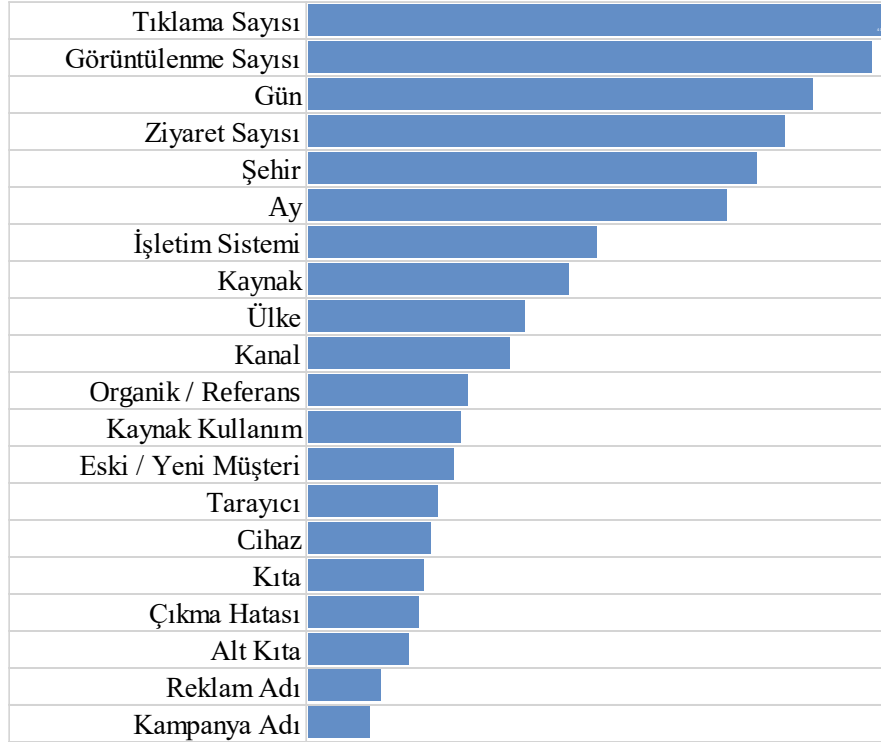
Çizelge 6.8. Parametreler ve bulunan değerleri

Parametreler	Bulunan Değerleri
Colsample_bytree	[1]
Learning_rate	[0.15]
Num_leaves	[64]
Max_depth	[10]
Num_iterations	[1000]

Çizelge 6.8’de bulunan parametreler ile birlikte, bağımlı değişkene göre seçilen parametrelerden olan Objective parametresi “regresyon” olarak, RMSE’ye göre de modeller kıyaslandığından Metric parametresi de “RMSE” olarak seçilmektedir. Tüm bu parametreler ile model tekrardan çalıştırıldığında, LightGBM modelinin RMSE

değeri 1,562 olarak bulunmaktadır. Çizelge 6.5’te gösterilen LightGBM modelinin RMSE değeri 1,608 iken parametre değerlerinin optimizasyonu sonucu RMSE değeri 1,562 olarak bulunmaktadır. Böylece model, hata değeri daha aşağıya çekilerek veri setine daha uygun hale getirilmiştir. Hata tahmin değerinde olan bu iyileşme, modelin uygulama performansını da arttırmıştır.

Başarı performansı yüksek modeller kurmanın yanında, anlamlı model kurmanın önemli noktalarından biri de bağımlı değişken üzerinde etkili olan değişkenlere ulaşmaktır. Ağaç temelli modeller çalışma mantığı gereği; rastgele olarak değişkenleri ve gözlemleri seçtiğinden, model kurarken hangi değişkenin, bağımlı değişken üzerinde etkili olup olmadığını da test etmiş olur. Model çalışırken eklenen bir değişken; hata değerini arttırmakta veya azaltmaktadır. Tüm değişkenlerin kombinasyonlarının, seçilen model üzerinde denenmesi ile bağımlı değişken üzerindeki etkili olan değişkenlere de ulaşılmaktadır. Çalışma kapsamında, en iyi tahmin performansı gösteren LightGBM modeli üzerinden, rastgele seçilen değişkenler arasından kurulan modeller neticesinde, bağımlı değişkeni en çok etkileyen faktörlere ulaşılmıştır. Şekil 6.10’da, Harcama Miktarı üzerinde en çok etkili olan değişkenler önem sırasına göre sıralanmaktadır.



Şekil 6.10. Değişken önem düzeyleri

Şekil 6.10 incelendiğinde harcama üzerinde en etkili olan değişken Tıklama Sayısı değişkeni olmuştur. Yani müşterinin, web sitesinde fazla zaman geçirerek sayfada gezinmesi, harcaması üzerinde etkilidir. Tıklama Sayısı değişkeni ile bağlantılı olan Görüntülenme Sayısı değişkeni önem sırasında ikinci olarak görünmektedir. Müşterinin, web sitesi üzerinde içerikleri incelenmesi, sayfaları görüntülenmesi, web sitesinde fazla zaman geçirmesi harcaması üzerinde etkili olmaktadır. Görüntülenme Sayısı değişkeninin ardından, Gün değişkeni 3. Sırada önem teşkil etmiştir. Müşterinin harcaması üzerinde, web sitesini hangi günler ziyaret ettiğinin önemli olduğu sonucuna ulaşılmıştır. Bu durumda; örneğin, günlere özel indirim ve fırsatların müşteri davranışlarını yönlendirmede etkili olduğu sonucuna ulaşılmaktadır. Ziyaret sayısı, Şehir ve Ay değişkenleri de önem düzeyinde ilk sıralarda yer almaktadır. Önem düzeyi yüksek değişkenler çerçevesinde; Harcama Miktarı üzerinde etkili olan değişkenlerin bilinmesiyle; müşterilerin davranışları üzerinde etki oluşturacak aksiyonlar alınabilmektedir. Örneğin; şirket; müşterilerine özel kampanyalar düzenleyebilir. Güne veya şehre özel kampanyalar sunarak satın almasını etkileyecek aksiyonlar geliştirebilir ve harcama oranlarını yükseltebilir.

7. SONUÇ VE ÖNERİLER

Günümüz dünyasında internet birçok insan tarafından bilgiye ulaşma, iletişim kurma, eğlence, sosyal etkileşim, online alışveriş ve özellikle sosyal medyanın takip edilmesi gibi pek çok amaç için kullanılmaktadır. Bunun yanında internet, sadece kişilerin değil aynı zamanda kurum/kuruluş ve sektörler tarafından da etkin biçimde kullanılmaktadır. Özellikle işletmeler, lojistik, reklamcılık, bankacılık ve pazarlama gibi alanlarda internetten elde ettikleri verilerden faydalanmakta ve dolayısıyla zamanla bu sektörler gelişme göstermektedir. Sektörlerin bu şekilde hızlı gelişmesi yeni iş alanlarının oluşmasına fırsat yaratmaktadır. Bunlardan birisi de ticaretin yeni bir biçim kazandığı, kısaca e-ticaret olarak tanımlanan elektronik ticarettir. E-ticaret, geleneksel ticaretin çok daha ötesinde hem tüketiciler hem de işletmeler açısından birçok olumlu yönü ile önemli bir ticaret türü haline gelmiştir. İnternetin temeli üzerine kurulan e-ticaret, aynı internetin sahip olduğu pek çok avantajı toplumlara sunmaktadır. Dünyada ve Türkiye’de, mevcut ticaret sistemlerine ek olarak, ticaret sistemlerini ve satış faaliyetlerini e-ticaret sistemi ile entegre etmiş işletmeler zamanla hem ticaret hacimlerini arttırmış hem de daha fazla ürün yelpazesini çeşitli kitlelere ulaştırmıştır.

E-ticaretin gelişmesi ile birlikte şirketler müşterileri hakkında birçok bilgi elde etme imkânı sağlamışlardır. Şirketler elde ettikleri bu verileri kullanarak farklı modeller geliştirmiş ve bu modelleri uygulayarak hem karlılıklarını hem de müşteri memnuniyetleri gibi birçok konuda oran ve sayılarını arttırmaya çalışmışlardır. Şüphesiz elde edilen bu verilerin işlenmesi ve anlamlı sonuçlar çıkarılmasının temelinde son zamanlarda giderek trend olan makine öğrenmesi yöntemlerinin kullanılması oldukça önemlidir. Şirketlerin makine öğrenmesi gibi yöntemler kullanması şirketlere zaman, otomatizasyon ve hız gibi oldukça hayati avantajlar sağlayacaktır. Çalışma kapsamında, Google Merchandise Store e-ticaret sitesi verilerinden yararlanılarak ağaç temelli makine öğrenmesi yöntemleri kullanılarak müşterilerin harcamalarının tahmin edilmesini, en az hata değeri ile sağlayan modelin belirlenmesi amaçlanmıştır. Ayrıca kurulan model yardımıyla, harcamaları üzerinde etkili olan değişkenlerin tespit edilmesi gerçekleştirilmiştir.

Çalışmada kullanılan veriler Google Analytics aracılığı ile toplanarak 2 Eylül 2016 ile 04 Ocak 2017 tarihleri arasında yer alan, 903653 satır ve 55 değişken içermektedir. Veriler üzerine makine öğrenmesi aşamalarından olan veri ön işleme yapılarak boş veya modele yük olan anlamsız değişkenler (Müşteri numarası, enlem, boylam gibi) çıkarıldıktan sonra 21 değişken ile çalışma tamamlanmıştır. Çalışmada Python 3.7 programlama dili ile bulut ortamında verilerin işlenmesini sağlayan Google Colab platformu kullanılmıştır.

Çalışmada veri seti %70 eğitim ve %30 test verisi şeklinde bölünerek eğitim verisi üzerinde ayrıca Cross Validation yöntemi ile modeller kurulmuştur. Karar Ağaçları, Rassal Ormanlar, GBM, LightGBM, Xgboost ve Catboost modelleri eğitim seti üzerinde eğitilmiş ve test verisi üzerinde uygulanarak test skorları alınmıştır. Oluşturulan modellerin hata metrikleri, bağımlı değişken sürekli olduğu için regresyon metrikleri olarak MSE, RMSE ve MAE olarak seçilmiştir. Genel olarak hata oranları birbirine yakın çıkmıştır. Ancak modellerin MAE değerlerinin birbirine oldukça yakın değerler çıkmasından ve modellerin en küçük karelere göre oluşturulmasından dolayı RMSE metriğine göre kıyaslamalar yapılmıştır. RMSE'ye göre kıyaslama yapıldığında en başarılı olan modelin 1,608 değeri ile LightGBM olduğu görülmektedir. Hemen ardından 1,612 RMSE değeri ile CatBoost yer almaktadır. Ardından sırasıyla GBM, XGBoost, Rassal Ormanlar ve Karar Ağaçları gelmektedir. En başarılı sonucu veren LightGBM üzerine parametre optimizasyonu gerçekleştirilerek modelin tahmin performansı arttırılmıştır. Ayrıca tüm sonuçlar göz önüne alındığında; bagging veya boosting yöntemlerini kullanan topluluk öğrenme modellerinin Karar Ağaçları'ndan daha başarılı sonuçlar verdiği ulaşılmıştır.

En iyi tahmin performansı gösteren LightGBM modeli ile harcama üzerinde etkili olan en önemli değişkenler tespit edilmiştir. Buna göre, harcama üzerinde en etkili olan değişken Tıklama sayısı)değişkeni olmuştur. Tıklama Sayısı değişkeni ile ilişkili olan Görüntülenme Sayısı değişkeni önem sırasında ikinci olarak tespit edilmiştir. Ardından, Gün değişkeni 3. Sırada önem teşkil etmiştir. Ziyaret Sıklığı, Şehir ve Ay değişkenleri de önem düzeyinde üst sıralarda yer almışlardır. Bu önemli değişkenler çerçevesinde, müşterilerin davranışları üzerinde etkili olacak aksiyonlar alınabilmektedir. Şirketler; müşterilerine özel kampanya, şehre özel fırsatlar ve indirimli günler kampanyası gibi satın almasını etkileyecek düzeyde aksiyonlar

geliştirebilir ve harcama oranlarını daha da yükseltebilir. Böylece şirket, veriye dayalı karar alma süreçlerini devam ettirerek e-ticaret alanında pazar payını arttırabilir. Oluşturulan makine öğrenmesi modelleri ile hızlı karar alarak kendi süreçleri içinde hızlı pozisyon alma fırsatını elde etmiş olur.

Çalışmanın, e-ticaret alanında çalışmak isteyen araştırmacılara birçok adımda yardımcı olabileceği öngörülmektedir. E-ticaret alanında problemlerin çözülmesi adına; makine öğrenmesi ile model kurmak, sonuçlarını test etmek ve takibini sağlamakta fikir oluşturabilir. Çalışmanın geliştirilebilmesi, farklı alanlarda uygulanması veya çalışma devamı olarak ağaç temelli modeller ile diğer makine öğrenmesi yöntemlerinin başarı oranlarının kıyaslanabileceği düşünülmektedir. Böylece literatürde önemli yer tutan modellerin çoğu, e-ticaret verileri üzerinden performansları karşılaştırılarak bu alanda makine öğrenmesi kullanımının gelişmesinde rol oynayabilir.

KAYNAKLAR

- Akkartal, E., Mendes, M., 2009. Regression Tree Analysis For Predicting Slaughter Weight In Broilers. *Italian Journal of Animal Science*, 615-624.
- Akpınar, H., 2000. Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği. *İ.Ü. İşletme Fakültesi Dergisi*, 1-22.
- Alpaydın, E., 2010. Introduction to Machine Learning. USA: Massachusetts Institute of Technology Press.
- Altaş, D., Gülpınar, V., 2012. Karar Ağaçları ve Yapay Sinir Ağlarının Sınıflandırma Performanslarının Karşılaştırılması: Avrupa Birliği Örneği. *Trakya Üniversitesi Sosyal Bilimler Dergisi*, 14(1).
- Baysal, A., 2011. CHAID ve Regresyon Ağacı Analiz Yöntemlerinin Hemşirelerde Stres Düzeyini Etkileyen Faktörler Üzerine Uygulaması. *Atatürk Üniversitesi Fen Bilimleri Enstitüsü Yüksek Lisans Tezi*, Erzurum.
- Behgounia, F., Zohuri, B., 2020. Machine Learning Driven An E-Commerce. *International Journal of Computer Science and Information Security*, 61-70.
- Bell, J., 2014. Machine Learning : Hands-On for Developers and Technical Professional. Somerset, John Wiley & Sons, Incorporated, USA.
- Bertoni, A., Campadelli, P., Parodi, M., 1997. A Boosting Algorithm for Regression, Artificial Neural Networks - ICANN '97, 7th International Conference, Lausanne, Switzerland.
- Bhasker, B., 2013. Electronic Commerce: Framework, Technologies, and Applications (4 b.). Tata McGraw Hill Education Private Limited.
- Breiman, L., 2001. Random Forests. *Machine Learning*, 45, 5-32.
- Breiman, L., Freidman, J., Olshen, R., Stone, C., 1984. Classification and Regression Trees. CRC Press.
- Breiman, L., Freidman, J., Olshen, R., Stone, C., 1993. Classification and Regression Trees. Chapman & Hall, New York
- Brownlee, J., 2016. Master Machine Learning Algorithms: Discover How They Work and Implement Them From Scratch. Jason Brownlee, 163.
- Bruce, P., Bruce, A., 2017. Practical Statistics for Data Scientists. USA: O'Reilly Media.
- Chaffey, D., 2002. E-Business and E-Commerce Management. Pearson Education Limited.

- Chang, L., Wang, H., 2006. Analysis of Traffic Injury: An Application of Non-Parametric Classification Tree Techniques. *Accident Analysis & Prevention*, 38(5), 1019-1027.
- Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System. *The 22nd ACM SIGKDD International Conference*.
- Çelik, C., Yelkikalan, N., 2021. Makine Öğrenme Yöntemlerinin Depo Yönetim Süreçlerinde Uygulanması: Azure ML Studio Örneği. *Journal of Yasar University*, 16(62), 659-682.
- Çetin, H., 2014. Online (Çevrimiçi) Alışverişte Akademisyen Davranışları ve Alışverişe Yönelten Etkenler. *Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, 19(4), 65-76.
- Demirdöğmez, M., Gültekin, N., Taş, H., 2018. Türkiye'de E-Ticaret Sektörünün Yıllara Göre Gelişimi. *Uluslararası Toplum Araştırmaları Dergisi*, 8(15), 2216-2236.
- Ekici, K., 2013. E-Ticaret. Savaş Yayınevi, Ankara
- Fidan, H., 2018. E-ticaret Müşteri Bağlılığı Gri İlişkisel Kümeleme Analizi. *AJIT-e: Bilişim Teknolojileri Online Dergisi*, 9(32), 163-182.
- Freidman, J., 2001. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189-1232.
- Freund, Y., Schapire, R., 1996. Experiments with a New Boosting Algorithm. In *Machine Learning, Proceed-ings of the Thirteenth International Conference*, 148-156. Bari, Italy.
- Geron, A., 2019. Hands-on Machine Learning with Scikit-Learn Keras & Tensorflow. USA: O'Reilly Media.
- Günaydın, İ., Benk , S., 2003. Globaleşmenin Vergi Sistemlerinde Meydana Getirdiği Değişiklikler. *Vergi Dünyası*, 267, 160-174.
- Gürcanok, T., 2020. E-ticaret Satış Verileri Üzerinde Bir Veri Bilimi Vaka Çalışması. *Maltepe Üniversitesi, Fen Bilimleri Enstitüsü, Yayınlanmamış Yüksek Lisans Tezi, İstanbul*.
- Hermogeno, D., 2019. E-Commerce: History and Impact on the Business and Consumers. *International Journal of Embedded and Software Computing*, 9(3).
- Hussung, T., 2016. From storefronts to search engines: A history of e-commerce. 2(3), 29-47.
- Fogg, A. 2021. A History of Machine Learning and Deep Learning. Erişim Tarihi: 05.07.2017. <https://www.import.io/post/history-of-deep-learning/>

- James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. Introduction to Statistical Learning. Springer.
- Karadeniz, T., Marař, H., Tokdemir, G., 2021. Ensemble Methods for Heart Disease Prediction. New Generation Computing, 569-581.
- Kavaklı, İ., 2018. E-Ticaret Rehberi. Kodlab, İstanbul.
- Kayahan, L., 2016. İktisadi Geliřme Ve Uluslararası İktisat Programı Türkiye’deki E-Ticaretin Ekonomideki Yeri Ve Geleceęi. Manisa Celal Bayar Üniversitesi, Yüksek Lisans Tezi, Manisa.
- Kayri, M., Boysan, M., 2008. Biliřsel Yatkınlık İle Depresyon Düzeyleri İliřkisinin Sınıflandırma ve Regresyon Ağacı İle İncelenmesi. Hacettepe Üniversitesi Eğitim Bilimleri Dergisi, 168-177.
- Kayri, M., Elkonca, F., řevgin, H., Ceyhan, G., 2014. Ortaokul Öğrencilerinin Fen ve Teknoloji Dersine Yönelik Tutumlarının CHAID Analizi ile İncelenmesi. Eğitim Bilimleri Arařtırması Dergisi, 4(1), 301-316.
- Kazan, S., Karakoca, H., 2019. Makine Öğrenmesi ile Ürün Kategorisi Sınıflandırma. Sakarya University Journal of Computer and Information Sciences, 2(1), 18-27.
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Liu, T., 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. NIPS.
- Küçükyılmazlar, A. (2006). Elektronik Ticaret Rehberi. İstanbul Ticaret Odası Yayınları, 3, 80-81, İstanbul.
- Lantz, B., 2013. Machine Learning with R. Olton: Packt Publishing Ltd.
- Lewis, D., 2017. Machine Learning Made Easy with R: An Intuitive Step by Step Blueprint for Beginners. CreateSpace Independent Publishing Platform.
- LightGBM, 2021. LighGBM Library. Eriřim Tarihi:01.05.2021. lightgbm.readthedocs.io/en/latest/pythonapi/lightgbm.LGBMRegressor.html
- Mingers, J., 1989. An Empirical Comparison of Pruning Methods for Decision Tree Induction. Machine Learning, 4, 227-243.
- Odabaşı, M., 2019. Makine Öğrenmesi Algoritmalarına DayalıYazılım Hata Tahmini: Çevresel Metriklerin Etkisi. İstanbul Şehir Üniversitesi Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, İstanbul
- OECD, 2011. OECD Guide to Measuring the Information Society. OECD Publishing.<https://doi.org/10.1787/9789264113541-en>
- Otero, F., Freitas, A., Johnson, C., 2012. Inducing Decision Trees with an Ant Colony Optimization Algorithm. Applied Soft Computing, 14(1), 3615-3626.

- Özdemir, R., Turanlı, M., 2021. Comparison Of Machine Learning Classification Algorithms For Purchasing Forecast. JOURNAL OF LIFE ECONOMICS, 8(1), 59-68.
- Perera, S., Ingirige, B., Ruikar, K., Obonyo, E., 2017. Advances in Construction ICT and e-Business. Routledge.
- Pitt, F., Berthon, P., Watson, R., Zinkhan, G., 2002. The Internet and the birth of real consumer power. Business Horizons, 45(4), 7-14.
- Qin, Z., 2009. Introduction to E-commerce. Berlin: Springer Science and Business Media.
- Quinlan, J., 1987. Simplifying Decision Trees. International Journal of Man-Machine Studies, 27, 221-234.
- Rokach, L., Maimon, O., 2005. Data Mining and Knowledge Discovery Handbook. Springer.
- Russell, S., Norvig, P., 1995. Artificial Intelligence A Modern Approach. Prentice Hall, Englewood Cliffs, New Jersey
- Saitoğlu, Y., 2015. Sınıflama ve Regresyon Ağaçları. Marmara Üniversitesi Sosyal Bilimler Enstitüsü, Doktora Tezi, İstanbul
- Sastry, P., Krishnan, R., Ram, B., 2010. Classification and Identification of Telugu and Written Characters Extracted from Palm Leaves using Decision Tree Approach. ARPN Journal of Engineering and Applied Sciences, 5(3), 22-32.
- Scikit-Learn. 2021. Machine Learning in Python. Erişim Tarihi:09.09.2021. <https://scikit-learn.org/stable/>
- Shelly, G., Rosenblatt, H., 2011. Systems Analysis and Design. IL, Course Technology, USA.
- Silahtaroglu, G., 2013. Veri Madenciliği (Kavram ve Algoritmaları, Papatya Yayıncılık Eğitim, İstanbul.
- Statista, 2020. Türkiye'de ve Dünyada E-Ticaret Verileri. Erişim Tarihi: 05.10.2020. <https://www.statista.com/outlook/dmo/ecommerce/turkey#market-arpu>
- Sutton, D., 2005. Classification and Regression Trees, Bagging and Boosting. Handbook of Statistics, 24, 303-329.
- Şahinarslan, F., 2019. Makine Öğrenmesi Algoritmaları ile Nüfus Tahmini: Türkiye Örneği, İstanbul Teknik Üniversitesi, Sosyal Bilimler Enstitüsü, Yüksek Lisans Tezi, İstanbul.
- Şimşek Gürsoy, T., 2010. Customer churn analysis in telecommunication sector. İstanbul Üniversitesi İşletme Fakültesi Dergisi, 39(1), 35-49.

- Tapkan, P., Özbakır, L., Baykasoğlu, A., 2011. Weka ile Veri Madenciliği Süreci ve Örnek Uygulama, Endüstri Mühendisliği Yazılımları ve Uygulamaları Kongresi, İzmir, 247-262.
- Temel, G., 2004. Sınıflama ve Regresyon Ağaçları. Mersin: Mersin Üniversitesi Sağlık Bilimleri Enstitüsü, Yüksek Lisans Tezi.
- Tübisad, Etid, Çalışması, Deloitte, 2018. Türkiyede E-Ticaret Pazar Büyüklüğü, https://www.tubisad.org.tr/tr/images/pdf/tubisad_2019_e-icaret_sunum_tr.pdf
- Yamamoto, G., 2013. E-Ticaret Kavramlar Gelişim ve Uygulamalar. Kriter Yayınları, İstanbul
- Yohannes, Y., Hoddinot, J., 1999. Classification and Regression Trees: An Introduction, International Food Policy Research Institute, Washington.
- Yücel, Y., 2017. Yaşam Memnuniyetini Etkileyen Faktörlerin Sınıflama Ve Regresyon Ağacı İle Belirlenmesi. İstanbul Ticaret Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, İstanbul.

ÖZGEÇMİŞ

Adı Soyadı : Mehmet YALÇIN

Eğitim Durumu

Lise : Ziya Gökalp (Y.D.A.) Süper Lise, 2007

Lisans : Diyarbakır Dicle Üniversitesi, Fen Fakültesi, Matematik Bölümü, 2017

Yüksek Lisans : İstanbul Ticaret Üniversitesi, Fen Bilimleri Enstitüsü, İstatistik Anabilim Dalı, 2021

Yayınları

Yalçın, M., Bağdatlı Kalkan, S., 2021. E-Ticaret Sitesi Müşterilerinin Harcamalarının Makine Öğrenmesiyle Tahmini, 7th International Management and Social Research Conference, 19-2 Mayıs, İstanbul, 116-117.