T.C.
İSTANBUL AYDIN ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

LIGHTGBM ALGORİTMASI İLE YENİ BİR SATIŞ TAHMİN MODELİNİN OLUŞTURULMASI VE PERAKENDE SEKTÖRÜNE UYGULANMASI

YÜKSEK LİSANS TEZİ

Yelda ARSLAN

Bilgisayar Mühendisliği Anabilim Dalı
Bilgisayar Mühendisliği Programı

EKİM, 2020

T.C.
İSTANBUL AYDIN ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

LIGHTGBM ALGORİTMASI İLE YENİ BİR SATIŞ TAHMİN MODELİNİN OLUŞTURULMASI VE PERAKENDE SEKTÖRÜNE UYGULANMASI

YÜKSEK LİSANS TEZİ

Yelda ARSLAN
(Y1813.010038)

Bilgisayar Mühendisliği Anabilim Dalı
Bilgisayar Mühendisliği Programı

Tez Danışmanı: Prof. Dr. Ali GÜNEŞ

EKİM, 2020

ONAY FORMU

ONUR SÖZÜ

Yüksek Lisans Tezi olarak sunduğum "LightGBM Algoritması ile Yeni Bir Satış Tahmin Modelinin Oluşturulması ve Perakende Sektörüne Uygulanması" adlı çalışmanın, tezin proje safhasından sonuçlanmasına kadarki bütün süreçlerde bilimsel ahlak ve geleneklere aykırı düşecek bir yardıma başvurulmaksızın yazıldığını ve yararlandığım eserlerin Bibliyografya'da gösterilenlerden oluştuğunu, bunlara atıf yapılarak yararlanılmış olduğunu belirtir ve onurumla beyan ederim. (28/10/2020)

Yelda ARSLAN

ÖNSÖZ

Yüksek lisans eğitim hayatım boyunca desteklerini hiçbir zaman esirgemeyen ve tez çalışmamda büyük katkıları bulunan danışman hocam Sayın Prof. Dr. Ali GÜNEŞ ve Dr. Öğr. Üyesi Peri GÜNEŞ'e teşekkürlerimi bir borç bilirim.

Tezimin yazım aşamasında sağladığı değerli katkılarından dolayı sevgili kardeşim Esra ARSLAN'a ve bu süreçte manevi desteklerini hiç esirgemeyen değerli aileme teşekkürlerimi sunarım.

Ekim, 2020 Yelda ARSLAN

LIGHTGBM ALGORİTMASI İLE YENİ BİR SATIŞ TAHMİN MODELİNİN OLUŞTURULMASI VE PERAKENDE SEKTÖRÜNE UYGULANMASI

ÖZET

Günümüz dünyasında gittikçe artan rekabet ortamında işletmeler varlıklarını sürdürebilmek için çeşitli stratejiler gerçekleştirmektedirler. Gün geçtikçe bu stratejilerden en çok tercih edileni müşterilerin taleplerine göre hareket etmek olmuştur. Böylece müşteri odaklı çalışma kavramı hayatımıza girmiştir. Şirketler için en önemli unsur olan müşteri memnuniyetini en üst seviyede tutmanın ancak müşteri odaklı çalışma anlayışı ile mümkün olacağı anlaşılmaktadır. Müşterilerin gelecek dönemlerde oluşacak isteklerini bilmek, işletmeler için büyük önem taşımaktadır. Böylece işletmeler öz kaynaklarının planlamalarını daha iyi yapabilmektedirler ve rakiplerine göre daha avantajlı duruma geçmektedir. İşletmeler, müşterilerin gelecekte oluşacak isteklerini ancak tahmin yöntemleri ile bilmektedirler.

Teknolojinin gelişmesi ve internetin yaygınlaşması ile birlikte işletmeler bu gelişmelerden daha çok faydalanmaya başlamışlardır. Günümüzde birçok işletme tahmin modellerinden yararlanmaktadır. Her işletmenin ihtiyacı ve sektördeki konumu farklılık gösterdiği için oluşturulacak tahmin modelleri işletmelerin yapısına uygun olmalıdır. Tahmin modelleri oluşturulurken işletmenin geçmişteki verileri referans alınmakta ve gelecekte oluşabilecek koşullar göz önünde bulundurulmaktadır. Böylece tahmin modellerinden daha iyi sonuçlar alınabileceği görülmektedir. İşletmeler yol haritalarını çizerken, uzun veya kısa dönemli planlamalarında firmalarının yapısına uygun olarak oluşturulan bu tahmin modellerinden yararlanmaktadırlar. Başarılı çalışan tahmin modelleri, işletmeleri fazla stok ve fazla mesai gibi maddi manevi birçok kayıptan kurtarmaktadır.

Bu çalışmada, perakende sektöründe faaliyet gösteren bir işletmenin E-Ticaret müşterilerine ait verileri kullanılmaktadır. Müşterilerin geçmişe ait verileri kullanılarak gelecek bir dönemdeki satış planlamalarına ışık tutacak, satış tahmin modeli oluşturulmaktadır. Çalışmanın uygulama kısmında, karar ağacı algoritmalarından biri olan LightGBM algoritması kullanılmaktadır. Daha verimli

sonuçlar alabilmek için algoritma üzerinde geliştirmeler yapılmaktadır. Oluşturulan satış tahmin modelinin uygulanması sonucunda elde edilen sonuçlar ve gerçek hayatta oluşan sonuçlar karşılaştırıldığında başarılı bir model olduğunun söylenmesi mümkündür. Bu kanıya, iki sonuç kümesi arasındaki standart sapmanın az olması ile varılmıştır. Çalışma sonucunda işletmenin stok yönetimi ve satış planlaması gibi önemli kararlarına ışık tutacak ve işletme kaynaklarının verimli kullanılmasına olanak tanıyacak bir çalışma olması hedeflenmektedir.

Anahtar Kelimeler: E-Ticaret, Karar Ağacı, LightGBM, Müşteri Davranışı, Regresyon Analizi, Satış Tahmini

DEMAND AND THE FORECASTING BASE A SPITING WAS MACHINE WORKING FOR CASTING BY APPLILAING LIGHTGBM ALGORITHM ON RETAIL DEMANDS

ABSTRACT

Today, companies are implementing various strategies to sustain their existence in a strong competition environment. The most preferred of these strategies is 'to act according to the demands of the customers'. Thus, the concept of customer-oriented marketing has occurred. It is understood that keeping customer satisfaction at the highest level, which is the most important unique for companies, can only be possible with a customer-oriented approach. It is very important to know your customers future demands. Thus, companies can plan their own resources better and become more advantageous than their competitors. These businesses can only know the future demands of their customers by using forecasting methods.

Thanks to the development of technology and the widespread use of the internet, many businesses are now making use of forecast models. The needs of each business and its position in the industry are different, so the forecast models to be used must be designed specifically for them. While designing the forecast models, the historical data of the business and the conditions that may occur in the future should be taken into consideration. While making their long or short-term planning, companies benefit from these forecast models specially designed for them. Successful forecasting models prevent negative situations such as over-stock and overtime in businesses.

In this study, e-commerce customers data of a company that exists in the retail sector is used. Using the customers historical data, a sales forecasting model will be created to shed light on future sales planning. In the application part of the study, LightGBM algorithm, one of the decision tree algorithms, is used. In order to get more efficient results, improvements are made on the algorithm. It is possible to say that it is a successful model when the results obtained from the application of the sales forecast model created and the real life results are compared. The standard deviation between the two sets of results is quite small. The outcome to be achieved through the study is

intended to shed light on important decisions of the business, such as inventory management and sales planning, and also to enable efficient use of business resources.
Key Words: Customer Behavior, Decision Tree, E-Commerce, LightGBM, Regression Analysis, Sales Forecast

İÇİNDEKİLER

ÖNSÖZ ... **iii**
ÖZET ... **iv**
ABSTRACT ... **vi**
İÇİNDEKİLER ... **viii**
KISALTMALAR LİSTESİ ... **x**
ŞEKİLLER LİSTESİ ... **xi**
ÇİZELGELER LİSTESİ ... **xiii**
I. GİRİŞ ... **1**
A. Problemin Tanımı ve Kapsam ... 3
II. LİTERATÜR ... **4**
III. GENEL BİLGİLER ... **7**
A. Tahmin ve Tahmin Yöntemleri ... 7
1. Yargısal Yöntemler ... 9
2. Özel Amaçlı Yöntemler ... 11
3. İstatistiksel Yöntemler ... 12
4. Birleşik Yöntemler ... 13
B. Perakende Sektörü ... 14
C. Veri Madenciliğinin İşletmeler Arası Rekabette Kullanılması ... 15
D. Makine Öğrenmesi ... 16
1. Makine Öğrenmesi ve Makine Öğrenmesi Yöntemleri ... 16
i. Eğiticili öğrenme ... 18
ii. Eğiticisiz öğrenme ... 18
iii. Yarı eğiticili öğrenme ... 18
iv. Destekleyicili öğrenme ... 19
2. Makine Öğrenmesinde Yaygın Olarak Kullanılan Algoritmalar ... 22
i. Doğrusal destek vektör makineleri ... 23
ii. Doğrusal olmayan destek vektör makineleri ... 24
i. LightGBM ... 28

IV. UYGULAMA 34
A. Veri Seti 34
B. Modelin Eğitime Hazırlık Süreci ve Özellikler Arasındaki İlişkilerin İncelenmesi 37
1. Kullanılan Yardımcı Kütüphaneler 37
2. Normalizasyon 38
3. Pivot Tabloların Oluşturulması 39
4. Ürün Özellikleri Arasındaki İlişki 39
5. Eğitim ve Test Verilerinin Belirlenmesi 45
C. Modelin Eğitimi 46
D. Sonuçların Alınması 50
V. DENEYSEL ÇALIŞMALAR 53
A. Regresyon Modelinin Seçimi 53
B. Algoritmanın Seçimi 55
C. Uygulamada Kullanılan Teknolojilerin Seçimi 56
D. Yeni Bir Model Oluşturma Gereksinimi 57
VI. BULGULAR VE SONUÇ 60
VII. KAYNAKÇA 62
EKLER 67
ÖZGEÇMİŞ 69

KISALTMALAR LİSTESİ

E-Ticaret : Elektronik Ticaret

E-Commerce : Electronic Commerce

GBM : Gradient Boosting Machine

LGBM : LightGBM

CNN : Convolutional Neural Network - Evrişimsel Sinir Ağları

DVM : Destek Vektör Makineleri

kNN : K-En Yakın Komşu

RBF : Radial Basis Function - Radyal Tabanlı Çekirdek Fonksiyonu

RO : Rastgele Orman

DFS : Depth First Search - Derin Öncelikli Arama

GOSS : Gradient Based One Side Sampling - Tek Taraflı Örnekleme

EFB : Exclusive Feature Bundling - Özel Değişken Paketi

BPNN : Backpropagation Neural Network - Geri Yayılımlı Ağlar

ŞEKİLLER LİSTESİ

Şekil 1 Tahmin Yöntemleri (Makridakis vd., 1979) 9
Şekil 2 Makine Öğrenmesi Yapay Zekâ ve Derin Öğrenme İlişkisi 17
Şekil 3 Nöron 20
Şekil 4 Konvolüsyonel Yapay Sinir Ağı İçin Girdi Örneği 21
Şekil 5 Evrişim Katmanı 21
Şekil 6 Maksimum Ortaklama ve Ortalama Ortaklama 22
Şekil 7 Tam Bağlı Katman 22
Şekil 8 Destek Vektör Makinesi Çalışma Prensibi 23
Şekil 9 Optimum Hiper Düzlem ve Destek Vektörleri 24
Şekil 10 Doğrusal Olmayan Destek Vektör Makineleri 25
Şekil 11 K-En Yakın Komşu Algoritması 25
Şekil 12 Karar Ağacı Yapısı 28
Şekil 13 Örnek Karar Ağacı 29
Şekil 14 Rastgele Ormanı Oluşturacak Karar Ağaçları 30
Şekil 15 Rastgele Orman Yapısı 30
Şekil 16 Lojistik Sigmoid Fonksiyonu 31
Şekil 17 Yardımcı Kütüphaneler 37
Şekil 18 Import Pandas 38
Şekil 19 Ortalama Değer Atama 39
Şekil 20 Pivot Tabloların Oluşturulması 39
Şekil 21 Pivot Tabloların İsimlendirilmesi 39
Şekil 22 Satış Adedi, Favori Sayısı ve Görüntülenme Sayısı Arasındaki İlişki 41
Şekil 23 Satış Adedi, Favori Sayısı, Görüntülenme Sayısı ve Ürün Özellikleri Arasındaki İlişki 43
Şekil 24 Satış Adedi, Favori Sayısı, Görüntülenme Sayısı ve Gelir Miktarı Arasındaki İlişki 45
Şekil 25 Eğitim ve Test Verilerinin Formatlanması 46
Şekil 26 Gradient Boosting 47

Şekil 27 Regresyon Analizi ve Ayarlamaları 47
Şekil 28 Gelir Hesaplama 47
Şekil 29 Gerekli Sütunlara Ağırlık verilmesi 47
Şekil 30 İterasyonun Belirlenmesi 47
Şekil 31 Ağacın Maksimum Uzaklığının Belirlenmesi 48
Şekil 32 Alt Düğüm Ayarları 48
Şekil 33 İşlemci Performansı 48
Şekil 34 Uyarıların Çıkarılması 48
Şekil 35 Kök Ortalama Kare Hatası 49
Şekil 36 Veri Setinin Yeni Model Üzerindeki Ağırlıkları 50
Şekil 37 Tahmin edilecek Ürün Numaralarının Bulunduğu Dosya Sisteme Tanımlanıyor 51
Şekil 38 Tahminler Satış Sütununa Ekleniyor 51
Şekil 39 Sonuçlar İçin Csv Dosyası Oluşturuluyor 51
Şekil 40 Veri Setinin LightGBM Algoritmasındaki Ağırlıkları 58

ÇİZELGELER LİSTESİ

Çizelge 1 Golf Oynama Sıklığı Veri Seti 32
Çizelge 2 Veri Seti Değerlerinin Bütüne Uyarlanması 32
Çizelge 3 Ürün Özelliklerini İçeren Veri Seti Örneği 34
Çizelge 4 Ürün Satış Bilgisini İçeren Veri Seti Örneği 35
Çizelge 5 Tahmin Edilmesi İstenilen Ürünlerin Olduğu ve Sonuçların Ekleneceği Veri Seti Örneği 36
Çizelge 6 Tahmin Edilmesi İstenilen Ürün Veri Seti 51
Çizelge 7 Yeni Oluşturulan Model Tahmininden Alınan Sonuçlar 52
Çizelge 8 LightGBM Algoritması Hata Metrikleri 52
Çizelge 9 Doğrusal Regresyon ve Regresyon Ağacı Hata Metrikleri 55
Çizelge 10 Algoritmaların Eğitim Bakımından Hız Faktörüne Göre Kıyaslanması .. 55
Çizelge 11 Test Verileri Üzerinde Hata Metriklerinin Karşılaştırılması 56
Çizelge 12 LightGM Algoritması Model Tahmininden Alınan Sonuçlar 59
Çizelge 13 Satış Adedi Sonuçları 60

I. GİRİŞ

Küreselleşmeyle birlikte oluşan rekabet ortamında işletmeler varlıklarını sürdürebilmek için pazar payını ve müşteri payını korumayı hedeflemektedirler. İşletmelerin rekabet içinde olması nedeniyle tüketici davranışları da zaman içerisinde büyük farklılıklar göstermiştir (Ekmekçi, 2006). İşletmeler bu değişime ayak uydurabilmek için müşteri odaklı çalışma anlayışını benimsemeye başlamışlardır. Benimsenen müşteri odaklı yaklaşım ile işletmeler, mevcut müşteriler ile ilişkilerini sürdürürken markaya olan bağlılıklarını arttırmayı ve yeni müşteriler kazanmayı amaçlamaktadırlar.

Yapılan birçok çalışmada işletmelerin mevcut müşterilerini kaybettikten sonra tekrar kazanma olasılığı ile yeni müşteri kazanma olasılıkları karşılaştırılmış ve araştırmalar sonucu mevcut müşterilerini kaybettikten sonra yeniden kazanmanın daha zor olduğu görülmüştür (Tosun, 2006). İşletmelerin mevcut müşterilerini kaybetmeleri ile satış kaybı arasında doğrudan bir bağlantı bulunmaktadır. Bu nedenle işletmeler müşteri taleplerine en hızlı şekilde cevap verebilir yapıda olmalılardır. İşletmecilerin en önemli isteklerinden biri müşterilerin hangi ürünleri daha çok talep ettiklerinin bilincinde olarak bu ürünleri yeterli sayıda ve istenilen kalitede üreterek müşterilerin istediği anda ürünlere ulaşabilmelerini sağlamaktır. Bu amaç doğrultusunda işletmeler, satılacak ürünleri pazara hızlıca çıkarma, müşterilere göre strateji geliştirme ve inovasyon yöntemlerini izlemektedirler. Ürünlerin doğru zamanda ve doğru sayıda pazara sunulması müşteri memnuniyetini sağlamak için önemli bir unsurdur. Bu nedenle tedarik zinciri yönetimini işletmelerin en başarılı şekilde yapmaları gerekmektedir. Beklenenden daha geç yapılan üretim ve tedarik süreçleri gibi arzdaki verimsizlikler şirketlerin büyütmesini yavaşlatmakta, müşteri memnuniyetsizliği oluşturmakta, stok seviyelerinin artmasına neden olmakta ve satış kayıpları yaşatmaktadır. Aynı zamanda şirketin sahip olduğu öz kaynakların verimsiz şekilde kullanılmasına neden olmaktadır. Bu durum beraberinde şirkete maddi ve zamansal bir kayıp yaşatmaktadır (Aydın, 2019).

Müşteri taleplerinin bilincinde olan işletmeler, müşterilerin istedikleri ürünlere daha kolay ulaşmaları için gerekli olan üretim ve tedarik gibi hizmetleri sunabilmektedirler. Böylece gün geçtikçe artan rekabet ortamında işletmelerin hayatta kalmaları daha kolaylaşmaktadır. Bu amaca hizmet edebilmek için işletmeler yapmış oldukları dönemsel veya yıllık planlarında, gelecekteki dönemlerin satış tahminlerini yapmaktadırlar ve buna göre üretim veya tedarik planlamalarını yapmaktadırlar (Demirtaş, 2011).

Satış tahmini; belirli bir zaman aralığı içinde satılması tahmin edilen ürün adedini ifade etmektedir. İşletmelerin aldığı stratejik kararlar için yapılan satış tahminleri büyük önem taşımaktadır. Satış tahmininin yapılacağı veriler yıllık, aylık veya günlük gibi zaman serisi verilerinden oluşmaktadır. Zaman serileri, belirli bir konu hakkındaki gözlemler ve konu hakkındaki geçmiş değerler ile bağlantılıdır (Washington vd., 2011).

Satış tahmin modelleri işletmeler tarafından yaygın olarak kullanılmaktadır. Çevresel birçok etkenin satış üzerinde doğrudan etkisi olduğu için, tahmin modellerine çevresel etkenleri dahil ederek oluşturulan satış tahmin modelleri karmaşık yapılarda olabilmektedir. Tahmini yapılan dönem içerisinde gerçekleşen koşullar göz önünde bulundurularak oluşturulan tahmin modelleri daha iyi sonuçlar vermektedirler. Tahmin edilen sonuçlar ile gerçekleşen sonuçlar arasında en az sapmanın olduğu tahmin modelleri başarılı kabul edilmektedir. Başarılı bir tahmin modeli sayesinde işletmeler gelecekte oluşacak satışa göre işletmelerine yön verip, karlılığı arttırırken oluşabilecek zararı en az seviyeye indirebilecek stratejileri izlemektedirler (Fantazzinia ve Toktamysovab, 2015).

Gelişen teknolojiler sayesinde bilişim sektörü, işletmelerin müşteri odaklı çalışmalarını destekler yapıdadır. Bilişim sektörü, müşterilere ait verilerin saklanabilmesine, işlenebilmesine ve bu verilerin işlenerek anlamlı bilgilere dönüştürülebilmesine olanak tanımaktadır. Böylece işletmelerin satış tahmini, müşteri kaybı analizi gibi hayati önem taşıyan geleceğe yönelik, işletmenin yol haritasını belirleyici bilgilere ulaşmaları kolaylaşmaktadır. İşletmeler ise bu verilere göre yıllık hatta dönemlik planlarını yapmaktadırlar ve işletmeleri için hayati önem taşıyan stratejik kararlar almaktadırlar (Shearer, 2000).

Günümüzde sektör ayrımı yapmaksızın faaliyet gösteren hemen her işletme veri madenciliği sayesinde müşterilerinin verilerini anlamlı hale getirmek için çalışmalar yapmaktadırlar. Veri madenciliği, işletmelerin müşteri davranışlarını analiz etmekte

en çok kullandıkları yöntemlerdendir. Böylece müşterilerin isteklerine göre işletmelerine yol haritası çizebilirler ve müşteri memnuniyetini mümkün olan en üst seviyede tutabilirler. Mevcuttaki müşterileri tek tek tanıyabilmek ve ihtiyaçlarını analiz edebilmek, işletmelerin müşteri, dolayısıyla satış kaybetme riskini en aza indirmektedir. Böylece işletmelerin mevcut durumlarını koruyabilmelerine yardımcı olmaktadır. Bu bağlamda satış tahmini faaliyetleri işletmeler için büyük rol oynamaktadır (Guo vd., 2013).

A. Problemin Tanımı ve Kapsam

Bu çalışmada perakende sektöründe faaliyet gösteren Dünya çapında lider firmalardan birinin E-Ticaret müşterilerine ait verileri kullanılmaktadır. Çalışmanın amacı, işletmenin hali hazırda çalışmakta olduğu müşterilerinin satın alma davranışlarını analiz ederek; üretim, tedarik, stok gibi planlama çalışmalarına yön verebilmeleri için gelecek dönemlerde satabilecekleri ürün adetlerini tahmin etmektir. Hatta işletmelerin müşteri profillerinin benzerlik gösterdiğinin göz önünde bulundurulmasıyla birlikte, doğru ürünleri doğru sayılarda sektöre arz ederek işletmenin pazardaki payını koruması amaçlanmaktadır.

Bu amaçlar doğrultusunda müşterilerin son altı ayda pazardaki satın alma faaliyetleri incelenerek, gelecekteki bir haftalık döneme ait satış tahmini yapılmaktadır. En iyi sonuçlara ulaşabilmek amacıyla; kullanılacak verilerin kapsadığı dönem ve tahmin edilen dönemin büyüklüğü her işletmeye özel olarak değişiklik gösterebilen yapıda olmalıdır. Bu nedenle çalışmada firmanın özellikleri ve müşteri kitlesi göz önünde bulundurularak, altı aylık satış dönemine bakılarak, bir haftalık satış periyodunun tahmin edilmesi uygun bulunmuştur.

Çalışma sonucunda hangi üründen kaç adet satılabileceği verisinin elde edilmesi amaçlanmaktadır. Böylece söz konusu işletmenin müşterilerini tanıyarak, müşterilerin taleplerine göre üretim ve tedarik planlarını yapmalarını, pazardaki yerlerini korumaları, hatta işletmenin büyümesine katkı sağlanması hedeflenmektedir.

II. LİTERATÜR

Alizadeh (2011), ürünlerin geçmiş satış verilerini kullanarak, gelecek dönemlerde oluşacak fiyatlarını tahmin etmeyi amaçlamaktadır. Çalışmada yapay sinir ağı ve oluşturulan yeni bir model performans açılarından karşılaştırılmaktadır. Uygulama Matlab ortamında yapılmıştır ve uygulama sonucu iki modelden elde edilen sonuçlar karşılaştırılmaktadır.

Asilkan (2009), yapay sinir ağları kullanılarak ikinci el otomobillerin gelecekteki fiyatları tahmin edilmeye çalışılmaktadır. Uygulamada kullanılan veriler Avrupa temelli birçok web sitesinden elde edilen ikinci el araba ilanlarından temin edilmiştir. Uygulama sonunda yapay sinir ağlarından elde edilen sonuçlar ve zaman serisi analizinden elde edilen sonuçlar karşılaştırılmıştır. Çalışma sonucunda yapay sinir ağlarından elde edilen sonuçların satış tahmininde başarılı sonuçlar verdiğini söylemek mümkündür.

Aydın (2019), çalışmasında piyasaya yeni çıkacak bir boyanın satış tahminini yapmayı hedeflemektedir. Tahmin yapılırken sadece satış verisinden yararlanılmamış, bağımlı ve bağımsız birçok nicel değişkenler arasında sebep sonuç ilişkilerine yer verilmiştir. Oluşturulan tahmin modelinde bulanık regresyon kullanılmaktadır. Bulanık regresyon kullanılmasının nedeni, henüz tanınmayan bir ürünün satışının tahmin edilmesidir.

Demirtaş (2011), bir işletmenin müşterilerinin pazardaki tercihlerini gösteren verileri anlamlı hale getirerek farklı birkaç yöntemi bir arada kullanarak satış tahmin uygulaması geliştirmektedir. Böylece müşteri odaklı çalışma anlayışını benimseyen işletmeler için hangi ürünlerin daha çok piyasaya sürülmesi gerektiği konusunda işletmelere yön vermeye çalışılmaktadır. Ek olarak Demirtaş çalışmasında, satış tahmin yöntemlerinin doğruluğu konusundaki araştırmalara da yer vermektedir.

Hamzaçelebi ve Kutay (2004), yapay sinir ağlarını kullanarak uzun vadeli elektrik enerjisi tüketimi üzerine bir tahmin çalışması yapmışlardır. Çalışmada kullandıkları farklı yapay sinir ağları yöntemleri ile elde edilen sonuçlar Box-Jenkins ve regresyon analizi yöntemleri ile karşılaştırılmıştır. Çalışmanın sonucunda yapay sinir ağlarının enerji tüketim tahmininde başarılı sonuçlar verdiği vurgulanmaktadır.

İşsever (2016), müşteri memnuniyetinin sağlanması ve işletmelerin rekabet ortamında hayatta kalabilmeleri için gereken bir satış tahmin modeli geliştirmektedir. Çalışmasında tekstil sektöründe faaliyet gösteren bir firmanın günlük satış verileri kullanılmaktadır. Modelden en verimli sonucu alabilmek için, özel günler ve hava durumu gibi etkenleri göz önünde bulundurmaktadır. Otoregresif hata ile regresyon metodunu kullanmaktadır.

Karataş (2011), yüksek lisans tezi çalışmasında yapay sinir ağlarının yazılım projeleri maliyet tahminlerinde nasıl kullanılabileceğini araştırmaktadır. Çalışmada yapay sinir ağlarının eğitiminde ve test edilmesinde COCOMO veri kümesi kullanılmaktadır. Uygulama kısmında ise XOR bilinmeyeninin çözüm sisteminden yararlanarak yeni bir yapay sinir ağı modeli oluşturulmuştur.

Özer (2011), şirketlerin varlıklarını sürdürebilmeleri ve büyümelerini sağlayabilmeleri için işletmelerin alması gereken stratejik kararlarda önemli rol oynayacak çeşitli çözümler sunmaktadır. Bunun için işletmenin mevcut müşterilerinin sergiledikleri davranışlara göre bulanık kümelemeden yararlanmaktadır.

Öztemiz (2017), Apriori algoritmasını kullanarak müşterilerin sepet ürün analizini yapmayı hedeflemektedir. Çalışmada ortam olarak Matlab kullanılmadır ve yapay sinir ağları ile satış tahmini yapılması hedeflenmektedir.

Sarıoğlu (2019), çalışmasında satış tahminlerinde kullanıcı-ürün etkileşiminin önemi üzerinde durmaktadır. Gittikçe artan sosyal medya ve internet kullanımının bir sonucu olarak; kullanıcıların bir ürünü tercih edeceği zaman öncelikle sosyal medya veya internet üzerinden daha önce kullanan kişilerin yorumlarına dikkat ettiğini, satın alma kararı verirken bu yorumların büyük bir etken olduğunu açıklamaktadır. Özellikle E-Ticaret sitelerinde yapılan ürün önerilerinin kullanıcı etkilerine bağlı olarak yapıldığını

savunmaktadır. Çalışmada birkaç yöntem aynı veri seti ile kullanılıp sonuçları karşılaştırılmaktadır.

Lee vd. (2012), gıda sektörü üzerine yapmış oldukları satış tahmini çalışmalarında lojistik regresyon ve Geri Yayılımlı Ağları (BPNN) kullanmışlardır. Bu iki yöntem performans bakımından çalışmada karşılaştırılmaktadır.
Yeğen (2020), gıda sektöründe hizmet veren bir işletmenin satış verilerinin analizini yaparak satış tahmini modeli geliştirmiştir. Çalışmada yapılan satış tahmin modeli IBM SPSS Modeler ile kurulmaktadır ve zaman serisi ve TCM modelleri ile analiz yapılmaktadır. Uygulama sonucunda zaman serisi modellerinin başarılı sonuçlar verdiği görülmektedir.

Yılmaz (2020), hazır yemek firmaları için, geleneksel zaman serisi modellerini kullanarak en performanslı ve başarılı modelin bulunmasını hedeflemektedir. Karşılaştırılan modellerde karşılaştırma kriterleri olarak Korelasyon ve Kısmi Korelasyon grafikleri, Akaike Bilgi Kriteri, Bayesian Bilgi Kriteri, Ortalama Hata Kareleri Kökü ve Standart Sapma ölçüm kriterleri kullanılmaktadır. Çalışma sonucunda en başarılı sonuçların Özbağlanım ve Vektör Özbağlanım modellerinden en iyi sonuçların alındığı sonucuna varılmıştır.

III. GENEL BİLGİLER

Genel bilgiler başlığı altında, tahminin tanımı, satış tahmininin önemi, tahmin yöntemleri, satış tahmini üzerinde yapılmış çalışmalar, veri madenciliğinin işletmeler arası rekabette kullanılması, perakende sektörü ve makine öğrenmesi konularına yer verilecektir.

A. Tahmin ve Tahmin Yöntemleri

Tahmin; bilinen değerlerin kullanılarak, bilinmeyen değerleri kestirme işlemidir (Demirtaş, 2011).

Tahmin; istatistiksel tesadüfi değişkenlerin büyüklüğünü gelecekteki bir zaman için öngörme işlemidir (İşsever, 2016).

Tahmin; işletmenin hedeflerini gerçekleştirebilmesi amacıyla, işletme planlarının yöneticiler tarafından düzenlenmesini sağlayan araçlardır (Kumar, 2009).

Tahmin; işletmenin gelecekte oluşacak satışlarının miktarını ve tutarını, çevre ve diğer rekabet faktörlerine bağlı olarak öngörmesi işlemidir (Chang vd., 2007).

Tahmin; geçmişte oluşan verilere, sezgilere ve akla dayalı olarak henüz gerçekleşmemiş bir durum hakkında sonucun kestirilmesi işlemidir. İşletmelerin gelecekteki belirsiz durumlar için de karar alabilme yeteneğine sahip olması gerekmektedir. Karar verme sürecinde işletme yöneticilerine yol gösterecek tahminlere ihtiyaç duyulmaktadır (Demirtaş, 2011).

Tahmin çalışmalarında geçmişteki veriler referans alınmaktadır. Kullanılan verilerin doğruluğu ve yeterliliği sonucunda gerçekleşene çok yakın öngörülerin yapıldığı ve başarılı sonuçlar elde edildiği görülmektedir. Bir tahminin başarılı olması, gerçekleşen durum ile beklenen durum arasındaki sapmanın en az olmasına göre değerlendirilmektedir. En başarılı tahmin çalışmalarında bile hata payı mutlaka vardır. İşletmelerin karar alırken bu hata payını göz önünde bulundurmaları gerekmektedir. Tahmin edilmek istenilen dönemin uzunluğu arttıkça, göz önünde bulundurulması gereken parametreler de artmaktadır. Bu parametreler, üzerinde tahmin çalışması

yapılacak sektöre bağlı olmakla birlikte genel olarak; finansal koşulları, rekabetçi firmaların faaliyetlerini, politik ve siyasal durumları kapsamaktadır (İşsever, 2016).

Günümüz dünyasında yaşanan teknolojik gelişmeler, işletmelerin ürünlerini pazara hızlıca sunabilmesine olanak tanırken tüketicilerin de pazardaki satın alma davranışları üzerinde hızlı değişiklikler yaşanmasına neden olmaktadır. Yaşanan bu hızlı değişimlere işletmelerin ayak uydurmaları, işletmeler için hayati önem taşımaktadır. Bu nedenle işletmeler ürünlerini daha hızlı ve etkin bir şekilde pazarlamak amacıyla tahmin yöntemlerini satış alanında sıkça kullanmaktadırlar. Elde edilen sonuçlar göz önünde bulundurularak, tüketicilerin daha az tercih ettikleri ürünler daha az, daha çok tercih ettikleri ürünler daha çok üretilmelidir. Böylece arz ve talep arasındaki ilişki dengede tutulacak ve kaynaklar en verimli şekilde kullanılacaktır. Bu durum işletme sermayesi için büyük önem taşımaktadır. Satış tahminleri işletmelerin geçmiş dönemlerinde yaşanan satışlara bakılarak elde edilmektedir. Geçmişteki veriler ele alınarak yapılan bir varsayım işlemidir (Chang vd., 2007).

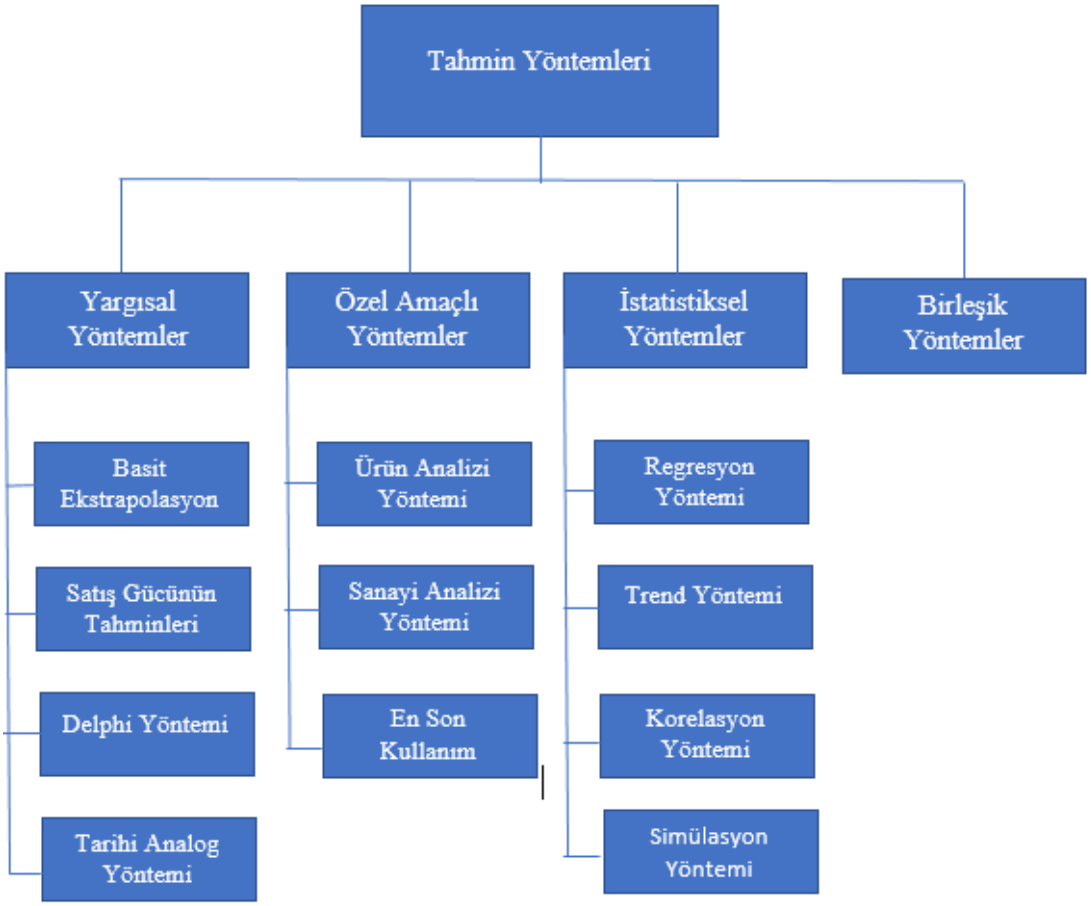
İşletmeler varlıklarını sürdürebilmek amacıyla, adımlar atarken gelecekte yaşanabilecek riskleri bilmek isterler. Çevresel faktörlerden minimum seviyede etkilenmek için bu riskleri önceden görerek önlemler almaları işletmeler için büyük önem arz eder. Bu nedenle geçmişte elde edilen veriler ışığında oluşan tahmin çalışmalarının satış planlamaları üzerinde etkisi büyüktür.

İşletmelerin gelecekteki durumunu tahmin etmek amacıyla yapılan çalışmalar, günümüz dünyasında işletmelerin sektördeki birçok davranışına yön vermektedir. Üretim, stok ve pazarlama faaliyetleri bunlara örnek olarak verilebilir. Bu nedenle tahmin çalışmaları, işletme yönetimi için büyük önem arz etmektedir. Tahmin çalışmaları yaygın olarak ticaret ve sanayi alanlarında kullanılmaktadır. Veri madenciliği kullanılarak yapılan bu çalışmalar sayesinde işletmeler, yatırımlarını, üretimlerini ve gelecekte oluşacak ihtiyaçlarını planlamaktadırlar (Wang vd., 2009).

Satış tahminleri, müşterilerin belirli zamanlarda ürünlere olan taleplerini öngörmeyi hedeflemektedir. Satış tahminleri yapılırken, işletmelerin üretim kapasiteleri göz önünde bulundurulmaksızın, pazar potansiyeli ele alınmaktadır. Üreticiler, toptancılar ve perakendeciler pazardaki potansiyele göre stok durumlarını yönetmektedir. Bu nedenle pazardaki talebi önceden tahmin etmeleri gerekmektedir. Talep edilen üründen daha azının pazara sunulması, işletmelere satış kaybı yaşatmaktadır. Hatta mevcut müşterilerinin rakip işletmelere yönelmesiyle müşterinin kaybedilmesine neden olabilmektedir. Talep edilenden çok daha fazla üretilen ürünler ise, işletmeye

hammadde, üretim, tedarik ve pazarlama alanlarında ek maliyetler getirmektedir. Önceden alınan önlemler sayesinde yaşanabilecek tüm olumsuz olayların önüne satış tahminleri sayesinde geçmemiz mümkündür. Tüm bu sebepler nedeniyle satış tahminleri işletmeler için büyük rol oynamaktadır (İşsever, 2016).

Talep tahmin çalışmaları için çeşitli yöntemler bulunmaktadır. Şekil 1'de görüleceği gibi bu yöntemler; yargısal talep tahmin yöntemleri, özel amaçlı yöntemler, istatistiksel talep tahmin yöntemleri ve birleşik yöntemler olarak dört ana gruba ayrılmaktadır.



Şekil 1: Tahmin Yöntemleri (Makridakis vd., 1979)

1. Yargısal Yöntemler

Yargısal tahmin yöntemlerinde geçmişe ait veriler bulunmamaktadır ya da geçmişten farklı koşulların oluşması gerekmektedir. Pazarda hiç bulunmayan bir ürünün ilk kez satışının yapılması ya da daha önce yaşanmamış olan büyük bir afetin sektörü etkilemesi buna örnek verilebilir. Bu gibi durumlarda geçmişteki veriler referans

alınamayacağı için tecrübeye dayalı tahmin çalışmaları yapılmaktadır. Yargısal tahmin yöntemleri nitel verilere göre yapıldığı için dezavantajları olduğu görülmektedir.

Bunlar;

- Değişimlere karşı verilen tepkilerin çok az ya da çok fazla olması,
- Geçmiş verilerdeki tutarsızlıklar,
- Kişisel görüşlerin tahminler üzerindeki olumsuz etkisi şeklinde ifade edilmektedir.

Yargısal yöntemler; basit ekstrapolasyon, satış gücünün tahminleri, delphi yöntemi ve tarihi analog yöntemleri olarak kendi içerisinde gruplara ayrılmaktadır (Serttaş, 2011).

a. Basit Ekstrapolasyon

Basit ekstrapolasyon yöntemi genellikle kısa zamanlı tahminlerde iyi sonuçlar vermektedir. Bu yöntemde geçmişte yapılan satışlar referans alınarak, yakın gelecekte oluşacak satışları tahmin etmektedir. Ekstrapolasyon yöntemi uygulanırken iki tarih aralığındaki veriler kullanılır. Kısa bir dönem referans alınarak kullanılan bu yöntemde verilerin az olması nedeniyle basit bir tahmin çalışması yapılabilmektedir (Olgun, 2009).

b. Satış Gücünün Tahminleri

Bu tahmin çalışması yöntemi genellikle, işletmelerin daha önce piyasaya sürmediği bir ürünün satışlarını tahmin etmekte kullanılmaktadır. Satış gücü tahmin yönteminde, ilgili ürünün geçmişine ait hiçbir satış verisi bulunmayacağı için işletmenin satış departmanı tahminde bulunmaktadır.

İşletmelerin bünyesindeki satış departmanı çalışanlarının, işletmenin ürün yelpazesine ve pazara olan hakimiyetleri sayesinde satışı konusunda başarılı tahminler yapabilecekleri görülmüştür. Satış gücünün tahminleri yöntemi, satış departmanının tecrübeleri, müşteriler ile aralarındaki ilişkileri ve pazara olan hakimiyetleri ile doğru orantılı başarı sağlanabilecek bir yöntemdir. Çalışanların bu konulardaki yetkinliklerinin yüksek olması işletme için avantajlı bir durum oluşturacaktır. Satış departmanın yapmış olduğu tahminler, yönetim departmanına gönderilerek ihtiyaç halinde burada revize edilip son halini almaktadır (Türk, 2019).

c. Delphi Yöntemi

Delphi satış tahmini yöntemi, alanlarında uzmanlaşmış yöneticiler ile birlikte yapılmaktadır. Bu yöntemde yöneticiler birbirlerinden habersiz bir şekilde tahminler yapmaktadırlar. Daha sonra tüm tahminler tüm yöneticilerle isimsiz şekilde paylaşılır ve gerekli ise her yöneticinin kendi satış tahminini revize etmeleri beklenmektedir. Bu yöntemde en az 2 tur olacak şekilde birkaç kez tahmin yapılmaktadır. Tahminler sonuçlandığında, ihtiyaç duyuluyorsa ilgili yöneticilerle toplantılar yapılabilir.

Delphi yönteminin en önemli özellikleri; katılımcıların kimliğinin açıklanmaması, sonuçların açıklanmasıyla yöneticilerin daha önce verdikleri tahminleri revize edebilmesi ve tüm yöneticilerden alınan tahmin sonuçlarına bakılarak ortak bir paydada buluşabilmektir (Chang vd., 2007).

d. Tarihi Analog Yöntemi

Tarihi analog yöntemi genellikle piyasaya sürülecek yeni ürünler için kullanılmaktadır. Ürün hakkında herhangi bir satış geçmişi olmaması nedeniyle, ürünün muadili sayılabilecek işletmenin kendi ürünleri veya rakip firmaların ürünlerinin satışları incelenmektedir. Yapılan bu pazar araştırmasına göre yeni ürünlerin ne kadar satabileceğinin tahmini yapılmaktadır (Karafakıoğlu, 2012).

2. Özel Amaçlı Yöntemler

Bu bölümde tahmin yöntemlerinden; ürün analizi yöntemi, sanayi analizi yöntemi ve en son kullanım yöntemi işlenmektedir.

a. Ürün Analizi Yöntemi

İşletmeler aylık veya yıllık gibi uzun soluklu planlarını yaparken gelecekte hangi ürünlerin ne kadar satılabileceğinin bilgisine erişebilmek isterler. Bu durumlarda işletmeler genellikle ürün analizi yöntemini kullanmaktadırlar. Ürün analizi yönteminde, işletme bünyesindeki tüm ürünler tek tek incelenerek gelecekteki satış tahminleri yapılmaktadır. Böylece işletmeler üretim planlamalarını yaparken ilgili ürünlerin stokunu arttırabilir ya da azaltabilir. Tüm ürünlerin tek tek satış tahminlerinin çıkarılmasıyla birlikte kümülatif olarak işletmenin satış tahmini de yapılabilmektedir. Ürün analiz yöntemi işletme için maddi kazançlar sağlamaktadır. Her bir ürünün tek tek analiz edilmesi sayesinde işletmelerin üretim kapasitelerini yönetmeleri kolaylaşmaktadır. Şirketin maddi kaynaklarını kullanmasında büyük rol oynayan üretim planlamasının en verimli şekilde yapılması sağlanmış olur.

b. Sanayi Analizi Yöntemi

Sanayi analizi yönteminde işletmeler, iş yaptıkları sektörün pazar hacmine göre hareket etmektedirler. Pazar hacmini bilen işletmeler, ilgili ürün için gerekli kaynak planlamasını yaparlar ve buna göre bir kaynak ataması yapılmaktadır. Bu aşamadan sonra, işletmelerin tahmin yapabilmek için ellerinde iki güçlü veri bulunmaktadır. Bunlar; pazar hacmi ve kaynak atamasıdır. Böylece işletmelerin ne kadar satış yapabileceğini tahmin etmeleri daha kolay hale getirilmektedir.

c. En Son Kullanım Yöntemi

En son kullanım yöntemi genellikle üretim yapan firmalarına parça üreten işletmeler için kullanılmaktadır. Bu işletmelerin müşterileri başka işletmeler olduğu için, satış yaptıkları işletmelerin kapasiteleri bu işletmelerin satış hacmini doğrudan etkilemektedir. Bu tür işletmeler satış tahmini yaparken müşterileri olan işletmelerin satış tahminlerinden yararlanırlar. İşletmeler bağımlı oldukları iş ortaklarının satış planlarına göre üretim ve satış planlamalarını yapmaktadırlar.

3. İstatistiksel Yöntemler

İstatistiksel tahmin yöntemlerinde, geçmişteki veriler matematiksel modellerde kullanılarak geleceğe yönelik tahminler yapılmaktadır. İstatistiksel tahmin yöntemlerinin de dezavantajları olduğu bilinmektedir. Bunlar;

- Bazı durumlarda oluşabilecek değişimlerin tahmin edilememesi,
- Geçmişteki verilerin ayıklanarak kullanılması, tüm bilgilerin göz önünde bulundurulmaması,
- Gelecekteki belirsizlik durumlarının dikkate alınmaması şeklinde özetlenebilir (Olgun, 2009).

a. Regresyon Yöntemi

Regresyon tahmin yöntemi; geçmişteki veriler referans alınır ve geçmişteki faktörlerin sürekliliğinin olacağı esasına dayanır. Bu tahmin yönteminde satışı etkileyen unsurların matematiksel olarak bir ağırlığı bulunur ve bu ağırlığın satışa olan etkisi hesaplanmaktadır.

Regresyon yöntemi, iki ya da daha fazla değişken arasındaki ilişkinin hesaplanmasında kullanılmaktadır. Regresyon tahmin yönteminde tek bir değişken kullanılıyorsa buna; tek değişkenli regresyon, birden fazla değişken kullanılıyorsa; çok değişkenli regresyon adı verilmektedir. Bu satış tahmin yönteminin kullanılmasındaki asıl amaç,

ilgili ürünün satış tahmini ile genel ekonomik göstergeler arasında bir ilişki olduğu gerçeğidir. Böylece satışı etkileyen ekonomik göstergeler belirlenir. Ortaya çıkan durumlara göre, satış tahmini yapılmaktadır (Karaca ve Karacan, 2016).

b. Trend Yöntemi

Trend tahmin yönetiminde işletmelerin geçmişe ait satış verileri kullanılmaktadır. Bu yöntemde işletmenin belirleyeceği herhangi bir dönem başlangıç dönemi olarak alınır ve içinde bulunulan döneme kadar yapılmış bütün satış verileri kullanılmaktadır. Trend yönteminde işletmenin bütün çevresel faktörlerinin aynı şekilde devam edeceği kabul edilmektedir (Aksoy, 2008).

c. Korelasyon Yöntemi

Korelasyon yönteminde satışı etkileyen değişkenler arasındaki ilişki hesaplanmaktadır. Bu yöntemde değişkenlerin herhangi birinde oluşacak değişimin, başka bir değişken üzerindeki etkisi izlenmektedir. Farklılıklar sonrası oluşacak ağırlıklar sayesinde, ekonomik değişkenler göz önünde bulundurularak satış tahmini yapılmaktadır.

Değişkenlerden faydalanılarak yapılan bu testler sonucunda; değişkenler arası ilişkinin görülmesi durumunda bir korelasyonun varlığı söz konusu olmaktadır. Tıpkı regresyon yöntemi gibi, korelasyon yönteminde de satışları etkileyen unsurlar analiz edilir ve değişkenler arası ilişkiler belirlenmektedir (Sun, 2010).

d. Simülasyon Yöntemi

Simülasyon yöntemi; üzerinde çalışmalar yapılan bir sisteminin, bir zaman aralığında karakteristiklerini tahmin etmeyi amaçlamaktadır. Bu amaç doğrultusunda simülasyon yöntemi; matematiksel ve mantıksal bir model geliştirilerek ilgili testler yapılması ve sistem planlama sürecini kapsamaktadır.

4. Birleşik Yöntemler

Birleşik tahmin yöntemlerinde; yargısal tahmin yöntemleri, özel amaçlı tahmin yöntemleri ve istatistiksel tahmin yöntemleri birlikte kullanılmaktadır. Bu yöntemlerden hangilerinin bir arada kullanılacağı, işletmenin iç dinamiklerine, ekonomik değişkenlere ve çevresel faktörlere bağlı olarak değişkenlik göstermektedir. Birleşik tahmin yönteminin benimsenmesinin asıl amacı; farklı tahmin yöntemlerinin, farklı durumlarda daha iyi sonuç verebiliyor olmasıdır.

Satış tahmin modelleri içinde bulunulan durumlara göre daha fazla maliyetli ya da daha karmaşık yapılardan oluşabilmektedir. Böyle bir durumda hangi tahmin modelinin seçileceği büyük bir soru işareti oluşturmaktadır. Birleşik yöntemler kullanılarak işletmenin yapısına ve içinde bulunulan duruma uygun tahmin modelleri bir arada kullanılarak daha doğru sonuçlar alınabileceği gözlemlenmektedir (Chang vd., 2007).

B. Perakende Sektörü

Ürünlerin tek tek ya da birkaç parça olarak satılmasına dayanan satış biçimine perakende olarak tanımlanmaktadır. Üretici firmalardan temin ettikleri ürünleri son kullanıcılarla buluşturan firmalara perakende firmaları denilmektedir. Perakende firmaları satışını yaptıkları ürünlerin satış sonrası desteklerini de sunmaktadırlar. Perakende firmalarının ürünün üretim süreci dışında tüm süreçleri yönettiklerini söylemek mümkündür. Müşteri ile ürünün buluşmasını perakende firmaları sağladığı için, müşteri ilişkileri büyük rol oynamaktadır. Müşterilerin ihtiyaç ve tercihlerini bilmek perakende firmalarının sektör içerisinde yer edinebilmeleri için büyük önem arz etmektedir. Perakende sektörü, rekabetin en çok görüldüğü sektörlerden biridir. Bu nedenle ürünün fiyatının, pazarlanmasının ve satış sonrası hizmetlerin en iyi şekilde sağlanması perakende firmaları için, avantaj sağlamaktadır (Uzunkaya, 2019).

Üretici firmalardan ürünleri satın alarak bu ürünleri son tüketicilerle buluşturan firmalara perakende firmaları denilmektedir. Perakende firmaları günümüzde hemen her sektörde yerlerini almaktadırlar. Perakende firmaları, üretici firmalar için pazarlama görevini üstlenirken, son tüketiciler için ise satın alma görevini üstlenmektedirler. Bu nedenle perakende firmaları dünya ticareti üzerinde önemli bir yere sahiptir.

Üretici firmaların sadece üretim ile ilgilenmesi ve ürünlerin pazarlama faaliyetlerinin perakende firmalarının üstlenmesiyle, ürünlere ve müşterilerin almış oldukları hizmetlere değer kazandırılmaktadır. Perakende firmalarının kendi aralarında rekabet ortamı oluşturması potansiyel müşteriler açısından büyük avantajlar sağlamaktadır. Özellikle büyük ölçekli perakende firmaları, yapmış oldukları işlere profesyonel bir bakış açısı kazandırmaktadırlar. Müşteri ile temas ettikleri her alan başta olmak üzere işletme genelinde tüm çalışanları kendi işlerinde uzman ve eğitimli kişilerden oluşmaktadır. Bu anlamda şirketin kendi bünyesine kazandırdığı her olumlu durum, müşterilere olan yaklaşım ve bakış açısıyla doğrudan ilişkilidir. İçinde bulunduğumuz

rekabet dünyasında işletmelerin ayakta kalabilmeleri için bu tür değerlere önem vermeleri ihtiyaç halini almaktadır. Benimsenen bu yaklaşım ise işletmeleri daima öne taşıyacak adımların parçası olmaktadır.

C. Veri Madenciliğinin İşletmeler Arası Rekabette Kullanılması

Günümüzde işletmelerin amacı, varlıklarını sürdürebildikçe daha fazla tüketiciye ulaşmak ve işletmelerini daha da ileriye taşımaktır. Bu amaç doğrultusunda işletmeler, teknolojinin faydalarından yararlanmaktadırlar. Gelişen teknoloji sayesinde, tüketicilerin pazardaki davranışları ve tercihleri analiz edilebilmektedir. İnternet üzerinde kullanıcıların yaptıkları işlemler sonucu elde edilen veriler işlenerek anlamlı bilgilere dönüştürülmektedir ve işletmeler gelişimlerinde bu bilgilerden yararlanmaktadırlar. Verilerin bu denli önemli hale getirilmesiyle büyük veri kavramı hayatımıza girmiştir. Bu verileri birleştirmek, anlamlandırmak ve işleyerek bilgiye dönüştürmek gün geçtikçe daha önemli hale gelmektedir. Verinin işlenerek anlamlı hale getirilebilmesinde veri madenciliği yöntemlerinden yararlanmaktayız.

Veri madenciliği yöntemleri günümüzde hemen her sektörde kullanılmaktadır. Bankacılık, E-Ticaret, Telekomünikasyon, sigortacılık, sağlık gibi insanın bulunduğu her sektörde veri madenciliği kullanılmaktadır. Veri madenciliği; Lineer Regresyon, Lojistik Regresyon, Karar Ağaçları ve K-Means gibi istatistiksel algoritmaları ve makine öğrenmesi yöntemleri olan; Destek Vektör Makineleri, Genetik Algoritmalar ve Yapay Sinir Ağları'nı kullanmaktadır (Haykin, 1999). Makine öğrenmesi yönteminde, kullanılan yazılım sayesinde öncelikle büyük veriler arasında işimize yarayan verilerin ayıklanması işlemi yapılmaktadır. Daha sonra bu verilerden anlamlı sonuçlar elde edilmektedir. Kazanılan bu deneyimlerle yeni veriler hakkında bir karar verebilmesi temeline dayanmaktadır.

Veri madenciliği sürecinde öncelikle problemin tanımlanması gerekmektedir. Veri madenciliği çalışmasının neden yapılması gerektiği ve nasıl yapılacağı konularının ilk aşamada belirlenmesi büyük önem taşımaktadır. Daha sonra verilerin toplanması ve bu verilerin hazırlanması gerekmektedir. Veri hazırlık aşaması, verinin anlamlı bilgilere dönüşmesine kadar olan süreci kapsamaktadır. Makine öğrenmesi aşamasında kullanılacak verilerin birbirleri arasında tutarlı olması gerekmektedir. Bu verilerde bulunan istisnai durumlar çıkarılmalıdır böylece maksimum düzeyde tutarlılık sağlanmalıdır. Verilerin hazırlık aşamasından sonra, veri modelleme aşamasına geçilmektedir. Veri modelleme, tahmin edilme sürecine denilmektedir. Veri

modelleme üç başlık altında toplanmaktadır. Bunlar; sınıflama, kümeleme ve birliktelik kurallarıdır (Koçtürk, 2010).

D. Makine Öğrenmesi

Bu bölümde makine öğrenmesi, makine öğrenmesi yöntemleri ve makine öğrenmesinde yaygın olarak kullanılan algoritmalar hakkında genel bilgiler verilmektedir.

1. Makine Öğrenmesi ve Makine Öğrenmesi Yöntemleri

Makine öğrenmesi; bilgisayar sistemlerinin, geliştirilen modelleri veya algoritmaları kullanarak bir işleyişi öğrenmesi ve çıkarım yapabilmesi olarak tanımlanabilir (Bishop, 2006).

Başka bir deyişle makine öğrenmesi; sistemlerin otomatik olarak öğrenme yeteneği kazanması olarak tanımlanmaktadır (Altan, 2019).

Bilgisayar sistemleri bir sonuca ulaşmaya çalışırken veya çıkarım yaparken bir görevi gerçekleştirmeye çalışmaktadırlar. Bilgisayar sistemleri makine öğrenmesini gerçekleştirirken öğrenme verisi olarak nitelendiren veri setlerinden yararlanmaktadırlar. Kullanılan bu veri setleri sayesinde sistemler bir model oluştururlar ve öğrenme süreci tamamlandığında bu sistemlerden tahminlerde bulunmaları beklenmektedir. Makine öğrenmesinin ilk amacı, bilgisayar sistemlerine insan müdahalesi olmadan öğrenmeyi sağlamak ve öğrendiklerinden çıkarımlar yapmasını beklenmektedir. Gerçek değerlere en yakın çıkarımlarda bulunan sistemlerin başarılı çalıştıklarını söylemek mümkündür. Şekil 2'de gösterildiği gibi makine öğrenmesi aynı zamanda yapay zekanın alt dallarındandır (Bishop, 2006).

Şekil 2: Makine Öğrenmesi Yapay Zekâ ve Derin Öğrenme İlişkisi

Makine öğrenmesi daha çok istatistiksel konularda başarılı sonuçlar elde etmemizi sağlasa da insani beceriler gerektiren konularda istenilen başarıyı sağlayamamaktadır. Bu nedenle makine öğrenmesinin alt dallarından biri olan derin öğrenme ortaya çıkarılmıştır. Derin öğrenme ile kazanılan, çok katmanlı yapı becerisi sayesinde çoklu soyutlama gerektiren verilerin gösterimi mümkün olmuştur. Derin öğrenme ile birlikte teknolojik olarak çok yol kat edilmiştir. Kazanılan bu özellikler, daha çok insana yönelik olan, ses algılama ve yüz tanıma gibi özelliklerdir.

Makine öğrenmesi için çeşitli algoritmalar kullanılmaktadır. Bu algoritmalar öğrenme sürecinde kullandıkları farklı yaklaşımlara göre üç ana gruba ayrılmaktadır. Bu gruplar; gelenekse yöntemler, derin öğrenme yöntemleri ve evrişimli yapay sinir ağları olarak tanımlanabilir. Geleneksel yöntemler de kendi içinde; öğreticili öğrenme, öğreticisiz öğrenme, yarı öğreticili öğrenme ve destekleyici öğrenme olarak dört ana başlık altında toplanmaktadır (Alpaydın, 2010).

a. Geleneksel Yöntemler

Geleneksel makine öğrenmesi yöntemleri, öğreticili öğrenme, öğreticisiz öğrenme, yarı öğreticili öğrenme ve destekleyici öğrenme olarak dört gruba ayrılmaktadır. Bu bölümde makine öğrenmesi yöntemlerinden geleneksel yöntemler konuları ele alınmaktadır (Hacıefendioğlu, 2012).

i. Eğiticili öğrenme

Eğiticili öğrenmenin temel mantığı; bir veri seti ile sisteme belirli davranışların öğretilmesi esasına dayanmaktadır. Bilgisayar sistemleri öğrenme aşamasında kullanılan bu veriler ile belirli kazanımlar sağlarlar ve yeni gelen veri setlerinde edindikleri bu kazanımları uygulamaktadırlar. Bu nedenle, eğiticili öğrenmede bir öğrenme süresi bulunmaktadır. Veri seti ile öğrenme yapan sistem, öğrenme süreci içerisinde bir model geliştirmeye başlamaktadır. Geliştirilen bu model daha sonra, henüz hiç kullanılmamış veriler üzerinde sonuç elde etmek için kullanılmaktadır.
Eğiticili öğrenme ürettiği çıktılara bakılarak kendi içerisinde iki gruba ayrılmaktadır. Eğer kullanılan eğiticili öğrenmede çıktılar bir değer seti ile sınırlandırılıyorsa; temel olarak sınıflandırma adını alırlar. Diğer bir eğiticili öğrenme çeşidi ise; eğri uydurmadır. Eğri uydurmada temel esas; elde edilen çıktıların sayısal bir aralığa sahip olmasına dayanmaktadır (Hacıefendioğlu).

ii. Eğiticisiz öğrenme

Eğiticisiz öğrenmede, eğiticili öğrenmeden farklı olarak kullanılan veri setleri üzerinde etiketleme yapılmamaktadır. Sistemin, kullanılan veri setlerinden kendi kendine çıkarımlar yapması beklenilmektedir. Eğiticisiz öğrenmede kullanılan veriler herhangi bir sınıfa dahil edilmemektedir. Bilgisayar sistemleri bu verileri işleme alırlar ve bu verilerdeki ortak noktaları bulmaya çalışmaktadırlar. Veri setleri arasında bulunan ortak noktalar ile veriler arasında bir sınıflandırma yapılmaktadır. Eğiticisiz öğrenme kullanan sistemler tüm verileri işledikten sonra ortaya çıkan ortak noktalara göre yaptıkları sınıflandırmaları çıktı olarak veren sistemlerdir (Hacıefendioğlu, 2012).

iii. Yarı eğiticili öğrenme

Yarı eğiticili öğrenme sistemi, adından da anlaşıldığı gibi eğiticili ve eğiticisiz öğrenme yapman sistemleri ifade etmektedir. Bu sistemler, eğiticili öğrenme yapan veri setlerinin yetersiz kaldığı durumlarda kullanılmaktadır. Yarı eğiticili öğrenme sistemleri, etiketlenmiş veri setlerinin öğrenme sürecinde yetersiz kaldığı durumda, etiketlenmemiş veri setlerinin kullanılmasıyla ortaya çıkmış sistemlerdir (Şeker, 2016).

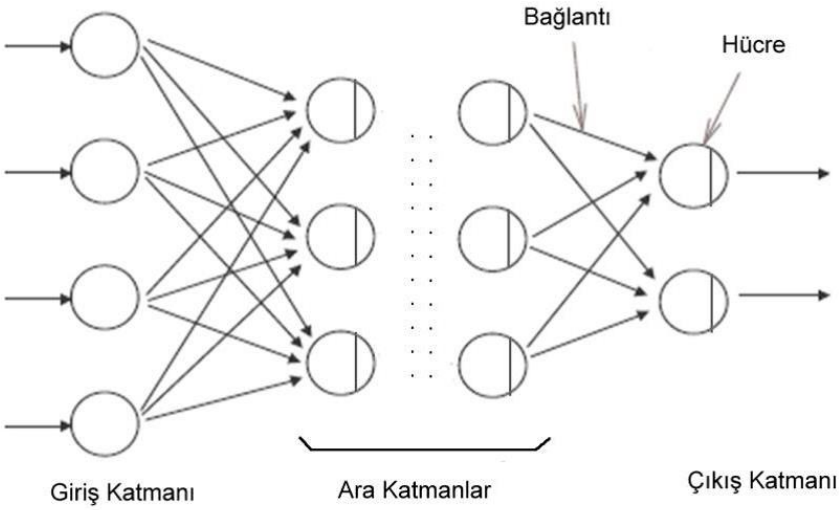
iv. Destekleyicili öğrenme

Destekleyici öğrenme diğer iki öğrenme yönteminden çok daha farklı çalışmaktadır. Bu öğrenme yönteminde bir yazılım ajanı konu olmaktadır ve temel davranış psikolojisine göre çalışmaktadır. Destekleyici öğrenme modelinde, yazılım ajanının çevre ile etkileşime geçip bir davranış kazanması temeline dayanmaktadır. Bu öğrenme yöntemini açıklayabilmek için bir bebeğin davranışları örnek gösterilebilir. Henüz sıcak-soğuk kavramını bilmeyen bir bebeğin cisimlere dokunarak bu davranışı kazanması destekleyici öğrenmeye örnek verilebilir (Altan, 2019).

b. Derin öğrenme

Derin öğrenme, günümüzde ses tanıma, görüntü tanıma, görsel algılama, nesne tanıma ve gen bilimi alanlarında sıklıkla kullanılan bir makine öğrenmesi yöntemidir. Derin öğrenme, girdi olarak bir veri seti kullanır ve bu veri seti ile bir yapay zeka modeli oluşturmamızı sağlamaktadır. Oluşturulan bu model ile veri seti kullanılarak çıktılar tahmin edilmeye çalışılmaktadır.

Derin öğrenme çok katmanlı bir mimariye sahiptir. Derin öğrenme, tıpkı bir insan beyni gibi çalışmaktadır. İnsan beyni içindeki nöronlar, yapay sinir ağları içinde de bulunmaktadır. Şekil 3'te örnek bir nöron gösterilmektedir. Nöronlar üç katmandan oluşmaktadır. Bu katmanlar giriş katmanı, ara katmanlar ve çıkış katmanı olarak adlandırılmaktadır. Giriş katmanı; verilerin girdi olarak alınıp ara katmanlara iletildiği katmandır. Ara katmanlar birden çok katmandan oluşabilmektedir. Bu durum problemin büyüklüğüne ve veri tipine bağlı olarak değişiklik göstermektedir. Veriler bu katmanlarda işlem görmektedir. Çıkış katmanı ise; ara katmanlarda işlenen verilerden oluşan çıktıları ulaştırmaktadır.



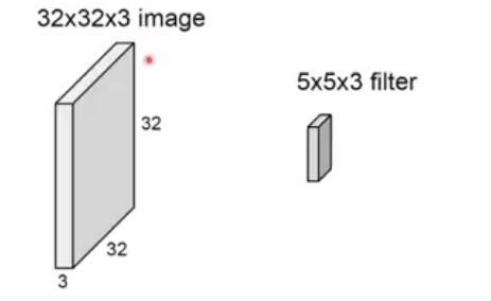
Şekil 3: Nöron

Derin öğrenmede, nöronlar arasındaki her bağlantının bir ağırlığı bulunmaktadır. Derin öğrenme yapılırken dikkat edilmesi gereken en önemli noktalardan biri bu ağırlıkları belirlemektir. Bu ağırlıklar bize ilgili girdinin önemini göstermektedir. Ağırlık değerleri en iyi modele ulaşılana kadar birkaç deneme ile belirlenmektedir. İlk değerler rastgele tahminlerle verilirken daha sonra bu değerler nedenlere dayandırılarak değiştirilir ve en optimal sonuca ulaşmak hedeflenmektedir.

Derin öğrenme yöntemlerinde ardışık veriler kullanıldığında çok başarılı sonuçlar elde edilmektedir. Geleneksel yöntemlere göre, derin öğrenme yöntemlerinden daha iyi sonuçlar alındığı görülmektedir. Ancak derin öğrenmedeki öğrenme sürecinin uzun olması bu yöntemin dezavantajı olarak kabul edilmektedir (Altan, 2019).

c. Konvolüsyonel (evrişimli) yapay sinir ağları

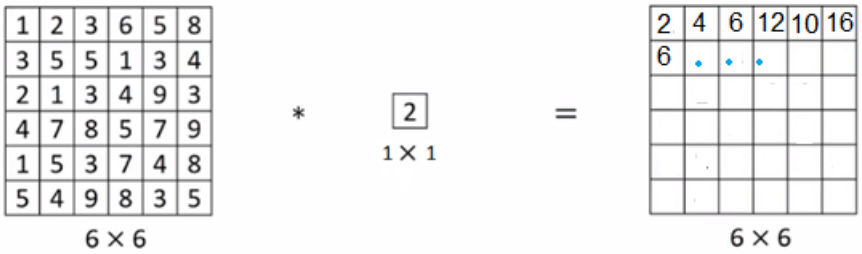
Konvolüsyonel ya da diğer adıyla evrişimli yapay sinir ağlarının çalışma mekanizmaları gereği girdileri resim ya da videolardır. Konvolüsyonel yapay sinir ağlarına verilen girdiler matrislere çevrilmektedir. Şekil 4'te görüldüğü üzere, girdi olarak bir resim ve filtre verilmektedir ve her ikisi de üç boyutlu dizi ile ifade edilmektedir. Resim için verilen 32x32x3 matrisindeki 32x32 resmin boyutunu ifade ederken, 3 renkli bir görsel olduğunu belirtmektedir.



Şekil 4: Konvolüsyonel Yapay Sinir Ağı İçin Girdi Örneği

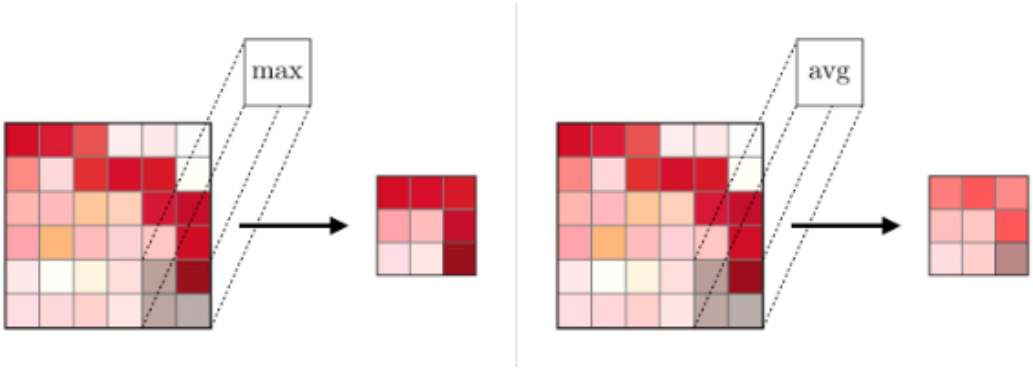
Evrişimsel yapay sinir ağları (CNN) birkaç farklı katmandan oluşmaktadır. Bunlardan en yaygın olarak kullanılanları; evrişim katmanı, ortaklama katmanı ve tam bağlı katmanlar olarak gösterilebilir.

Evrişim katmanında, işleme alınan resim matrisi üzerinde bir filtre matrisi kullanılır. Filtre matrisi, resim matrisinin üzerinde birer birer kaydırılarak işlenir ve resim tanımlanmaya çalışılır. Bu işleme evrişim adı verilmektedir. Şekil 5'te 6x6'lık bir resim matrisi üzerinde 1x1 matrislik bir filtre uygulanması örneklendirilmektedir.



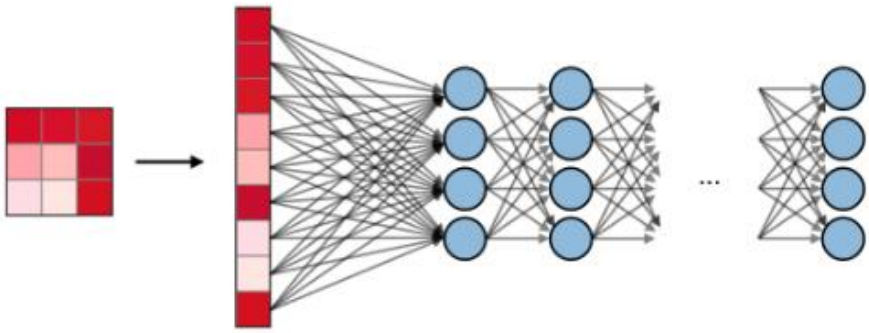
Şekil 5: Evrişim Katmanı

Aynı ya da birbirine benzer görüntülerin açıları değiştirildiğinde, evrişim katmanı bu görüntülerin ikisi için de benzerlik olduğunu anlayamaz ve farklı öznitelik haritaları çıkarmaktadır. Ancak bu görüntüler aynı oldukları için görüntü işlemede bu iki resim arasında bir benzerlik tespit edilmesine ihtiyaç duyulmaktadır. Ortaklama katmanı, bu iki görüntünün arasındaki benzerlikleri ortaya çıkarmakta kullanılmaktadır. Ortaklama katmanında maksimum ortaklama ve ortalama ortaklama olmak üzere iki farklı işlem uygulanmaktadır. Maksimum ortaklama yapılırken; işlem gören resim matrisinin en büyük değeri dikkate alınırken, ortalama ortaklamada, ilgili matris değerlerinin ortalaması esas alınmaktadır. Şekil 6'da maksimum ortaklama ve ortalama ortaklama gösterilmektedir.



Şekil 6: Maksimum Ortaklama ve Ortalama Ortaklama

Tam bağlı katmanda her bir giriş bir katmana bağlıdır. Bir evrişimli yapay sinir ağında tam bağlı katman bulunuyorsa bu katman evrişimli yapay sinir ağı modelinin sonlarında yer almaktadır. Bunun nedeni, tam bağlı katmanların genellikle sınıf skorlarını optimize etmekte kullanılmalarıdır. Şekil 7'de tam bağlı katman gösterilmektedir (Kurt, 2018).



Şekil 7: Tam Bağlı Katman

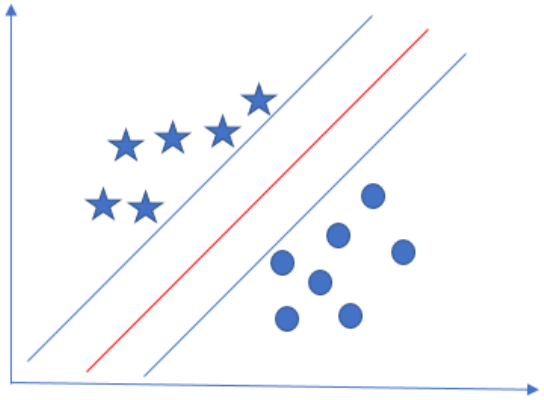
2. Makine Öğrenmesinde Yaygın Olarak Kullanılan Algoritmalar

Bu bölümde, makine öğrenmesinde en yaygın olarak kullanılan algoritmalardan olan; destek vektör makineleri, k-en yakın komşu, karar ağaçları, rastgele orman, lojistik regresyon ve naive bayes algoritmaları hakkında genel bilgilendirme yapılmaktadır.

a. Destek Vektör Makineleri (DVM)

Destek Vektör Makineleri (DVM), sınıflandırma ve regresyon problemleri için en yaygın kullanılan algoritmalardan biridir. Destek vektör makinelerinin temelinde istatistiksel öğrenme yer almaktadır ve destek vektör makineleri denetimli öğrenme yapmaktadır. Destek vektör makinelerinin çalışma mantığı; bir düzlem üzerinde iki grup oluşturulacak şekilde bir sınır belirlenmesine dayanmaktadır. Belirlenen bu sınır her iki gruptaki üyelere de en uzak sınır olarak belirlenmektedir. DVM, bu sınırın

belirlenmesinde rol oynamaktadır. Bu iki grubu sınıra maksimum uzaklıkta ayırabilecek tek bir nokta olduğu bilinmektedir. Bu noktaya optimum sınır adı verilmektedir. Şekil 8 bize optimum sınırı göstermektedir. Destek vektör makineleri verilerinin doğrusal olarak ayrılıp ayrılmama durumuna göre iki grupta incelenmektedir (Suykens, 1999).



Şekil 8: Destek Vektör Makinesi Çalışma Prensibi

i. Doğrusal destek vektör makineleri

Doğrusal olarak ayrılabilen verilen veriler üzerinde kullanılan hiper düzlem Denklem 1'de gösterilmektedir.

$$f(x) = \mathrm{w}^{\mathrm{T}}.x + b = \sum_{i=1}^{n} w_i.x_i + b$$

Denklem 1: İki Sınıfı Birbirinden Ayıran Hiper Düzlem Denklemi

Bu denklemi şu şekilde açıklamak mümkündür;

n : Veri kümesinin eleman sayısı

$X = \{xi,yi\}, i = 1,2,\ldots,n$: Veri kümesi

$yi \in \{-1,1\}$: Etiket değerleri

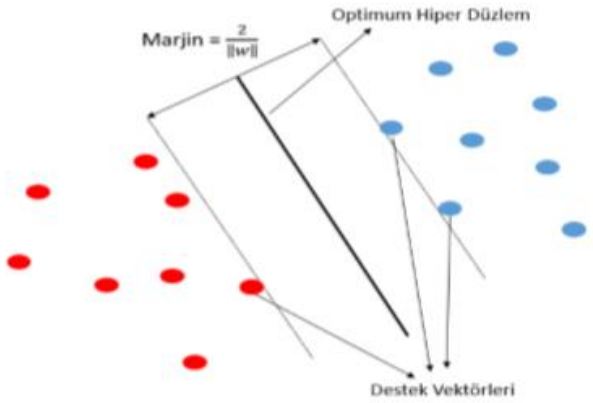
$xi \in \Re d$: Özellik vektörü

w : Ağırlık

x : Veriler

b : Eğitim terimi

Şekil 9'da doğrusal ayrılabilen veri kümeleri kullanımında hiper düzlem ve destek vektörleri gösterilmektedir.



Şekil 9: Optimum Hiper Düzlem ve Destek Vektörleri

ii. Doğrusal olmayan destek vektör makineleri

Bazı veri setleri doğrusal olarak ayrılabilirken, bazılarını doğrusal olarak ayırmak mümkün değildir. Doğrusal olarak ayrılmayan veri setleri için Denklem 1'in kullanımı doğru sonuç vermemektedir. Bu gibi durumda, veriler çekirdek fonksiyonundan geçirilir ve özellik uzayına taşınmaktadırlar daha sonra burada sınıflandırılmaktadırlar. Doğrusal olmayan destek vektör makinelerinde en yaygın olarak kullanılan fonksiyonlar radyal tabanlı çekirdek fonksiyonu, polinom çekirdek fonksiyonu, doğrusal çekirdek fonksiyonudur. Bu denklemler Denklem 2, Denklem 3 ve Denklem 4'te, doğrusal olarak ayrılmayan destek vektör makinesi Şekil 10'da gösterilmektedir.

$$K(x_i, x_j) = \left(\frac{-||x_i - x_j||^2}{2_{o^2}}\right)$$

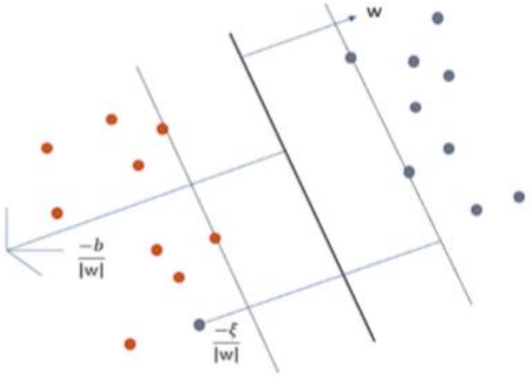
Denklem 2: Radyal Tabanlı Çekirdek Fonksiyonu (RBF)

$$K(x_i, x_j) = K(x_{i^T}, x_j)^d$$

Denklem 3: Polinom Çekirdek Fonksiyonu

$$K(x_i, x_j) = K(x_{i^T}, x_j)$$

Denklem 4: Doğrusal Çekirdek Fonksiyonu

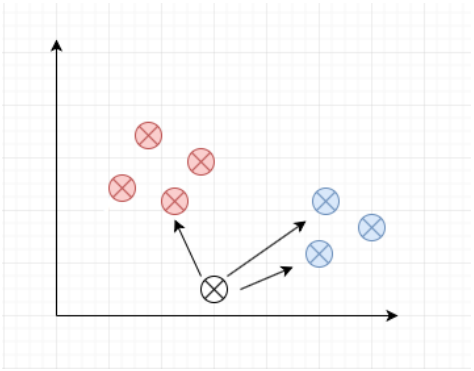


Şekil 10: Doğrusal Olmayan Destek Vektör Makineleri

b. K-En Yakın Komşu (kNN)

K-en yakın komşu algoritması denetimli öğrenme yapmaktadır. Yani bu durum algoritmada kullanılan modelin, eğitim verileri üzerinden bir öğrenme yapmasını açıklamaktadır. K-en yakın komşu algoritması, en temel makine öğrenmesi algoritmalardan biridir. Büyük verilerde kullanıma uygun olmayıp küçük ölçekli veriler üzerinde algoritmadan başarılı sonuçlar alınabilmektedir.

Algoritmanın çalışma mantığı; veri seti ile birbirine benzeyen verilerin gruplanarak, benzer özellikteki verilerden oluşan birkaç sınıfın oluşturulması temeline dayanmaktadır. Eğitim verileri sayesinde modelin eğitimi tamamlanmaktadır. Tahmin edilmesi istenilen veri, bilinen sınıflardan hangisine yakınsa o sınıfa dahil edilerek bir sonuç elde edilmektedir (Çataloluk, 2012). Şekil 11'de k-en yakın komşu algoritması gösterilmektedir.



Şekil 11: K-En Yakın Komşu Algoritması

K-en yakın komşu algoritmasında birkaç farklı uzaklık hesaplama algoritması kullanılmaktadır. Kullanılacak uzaklık hesaplama algoritması veri setine göre seçilmektedir. Veri setlerindeki farklılıklar farklı algoritmalar kullanılmasını zorunlu hale getirmektedir. En yaygın olarak kullanılan uzaklık hesaplama algoritmaları

Denklem 5, Denklem 6 ve Denklem 7, Denklem 8, Denklem 9, Denklem 10 ve Denklem 11'de gösterilmektedir.

Öklid Uzaklığı; iki nokta arasındaki uzaklığı hesaplamayı amaçlamaktadır. İki nokta arasındaki doğrunun uzaklık Öklid uzaklığı olarak adlandırılmaktadır.

$$S_i^* = \sqrt{\sum_{j=1}^{n} \left(v_{ij} - v_j^*\right)^2}$$

Denklem 5: Öklid Uzaklığı

Manhattan Uzaklığı; iki nokta arasındaki uzaklığın mutlak farkları toplamını ifade etmektedir.

$$S_i^* = \sum_{j=1}^{n} \left| v_{ij} - v_j^* \right|$$

Denklem 6: Manhattan Uzaklığı

Chebyshev Uzaklığı; İki nokta arasındaki en uzak mesafeyi ifade etmektedir.

$$S_i^* = \max_i \left| v_{ij} - v_j^* \right|$$

Denklem 7: Chebyshev Uzaklığı

Lorentzian Uzaklığı; İki nokta arasındaki uzaklığın negatif çıkmaması için; uzaklık iki nokta arasındaki farkın doğal logaritma değerine eşitlenmesiyle hesaplanmaktadır.

$$S_i^* = \sum_{j=1}^{n} \ln\left(1 + \left| v_{ij} - v_j^* \right|\right)$$

Denklem 8: Lorentzian Uzaklığı

Pearson Uzaklığı; iki nokta arasındaki farkın karesinin ideal çözüme oranlanması ile elde edilmektedir.

$$S_i^* = \sum_{j=1}^{n} \frac{\left(v_{ij} - v_j^*\right)^2}{v_j^*}$$

Denklem 9: Pearson Uzaklığı

Kosinüs Uzaklığı; iki nokta arasındaki açıyı ölçmektedir.

$$S_i^* = \frac{\sum_{j=1}^{n} v_{ij} v_j^*}{\sum_{j=1}^{n} v_{ij}^2 \sum_{j=1}^{n} v_j^{*2}}$$

Denklem 10: Kosinüs Uzaklığı

Jaccard Uzaklığı; kosinüs uzaklığı denkleminin farklı bir versiyonu şeklinde tanımlanmaktadır.

$$S_i^* = \frac{\sum_{j=1}^{n} \left(v_{ij} - v_j^*\right)^2}{\sum_{j=1}^{n} v_{ij}^2 + \sum_{j=1}^{n} v_j^{*2} - \sum_{j=1}^{n} v_{ij} v_j^*}$$

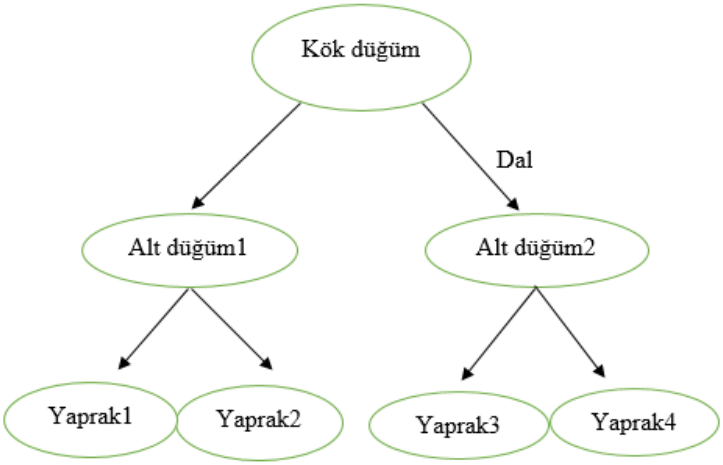
Denklem 11: Jaccard Uzaklığı

c. Karar Ağacı

Karar ağaçları denetimli öğrenme yapmaktadırlar ve sınıflandırma problemleri için kullanılmaktadırlar. Karar ağaçlarındaki temel amaç veri setindeki tüm elemanları sınıflandırmaya dahil etmektir. Bu amaç doğrultusunda, karar kuralları tüm verilere uygulanır ve ilgili veriler bir sınıfa dahil edilmektedir. Makine öğrenmesinde sınıflandırma problemlerinde yaygın olarak karar ağaçları kullanılmaktadır. Sayısal ve kategorik veri tipleri karar ağaçlarının kullanımı için oldukça uygundur.

Karar ağaçları düğüm, dal ve yapraklardan oluşan yapılardır. Bu yapı Şekil 12'de gösterilmektedir. Veri sınıflandırma yapabilmek amacıyla veri setindeki özniteliklerden yararlanılmaktadır. Her bir öznitelik bir düğümü oluşturmaktadır. Düğümlerde belirlenen kriterlere göre, veriler iki alt sınıfa bölünmektedir. Kök düğümün ve alt düğümlerin iki veya daha fazla alt düğümü bulunmaktadır.

Kök düğümlerin herhangi bir girdisi bulunmamaktadır. Çıktıları ise alt düğümleri oluşturmaktadır. Alt düğümlerin girdileri ve çıktıları bulunmaktadır. Alt düğümlerin çıktıları, yaprakların girdilerini oluşturmaktadır. Yaprakların ise çıktıları bulunmamaktadır. Yapraklar, ağacın dallarının ulaştığı son nokta olarak tanımlanmaktadır.

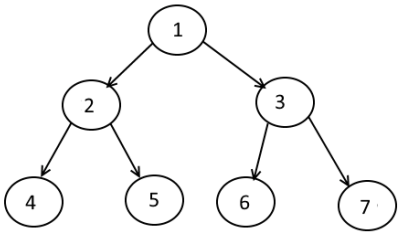


Şekil 12: Karar Ağacı Yapısı

Karar ağaçlarında kullanılan nitelikler arasından ayırt edici nitelikleri bulabilmek amacıyla bilgi kazancı ölçümü yapılmaktadır. Bilgi kazanımını ölçebilmek için entropi yönteminden yararlanılmaktadır. Karar ağacında kullanılan tüm özelliklerde bilgi kazancı hesaplanır ve en yüksek bilgi kazancını sağlayan özellik kök düğüm olarak belirlenmektedir. Dallanma bu kök düğümden başlayarak yapılmaktadır ve yapraklara ulaşana kadar dallanma devam etmektedir. Bir düğüme gelen verinin hangi dala gideceğine, düğümdeki özelliğin eşik değerinden verinin büyük ya da küçük olması durumuna göre karar verilmektedir. İlgili verilerin sınıfı, dallanma bittiğinde ulaşılan yaprağın sınıfına dahil olmaktadır. Şekil 12'de görülen Alt Düğüm1'e dahil olan bir veri, Yaprak1'e ulaşıyorsa, Yaprak1'in temsil ettiği sınıfa dahil olmaktadır (Savaş vd., 2012).

i. LightGBM

LGBM'nin adında geçen "Light" kelimesi çok hızlı çalışması nedeniyle ışık veya hafif olarak Türkçe 'ye çevrilmektedir. Algoritma makine öğrenmesi yöntemlerinde kullanılmaktadır ve karar ağacı temeline dayanan yüksek hızlı bir gradient boosting framework'dür. LGBM algoritması derin öncelikli arama (DFS) kullanmaktadır. DFS; algoritmasının kullanılması durumunda, Şekil 13'teki karar ağacında dolaşma sıralaması, 4-5-2-6-7-3-1 olacaktır.



Şekil 13: Örnek Karar Ağacı

GBM algoritmasının formülü Denklem 12'de, LightGBM algoritmasının formülü ise Denklem 13'te gösterilmektedir.

$$GBM = Desicion\ Tree + Boosting + Gradient\ Descent$$

Denklem 12: GBM Algoritması Formülü

$$\text{LightGBM} = GBM + GOSS + EFB$$

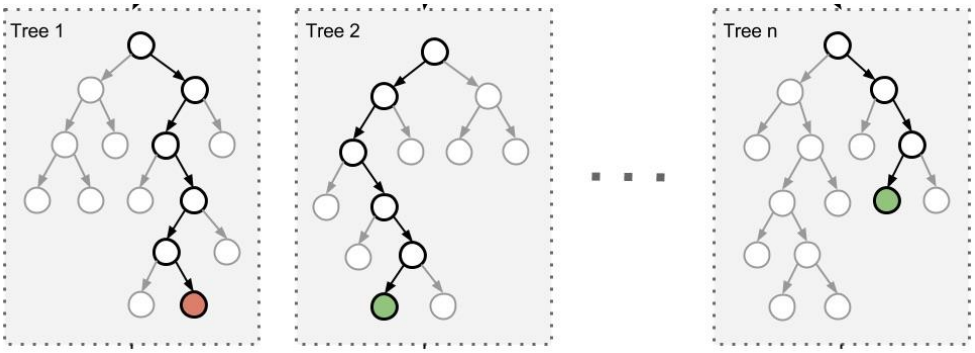
Denklem 13: LGBM Algoritması Formülü

LightGBM algoritması da tüm karar ağaçları gibi tümevarım yöntemiyle çalışmaktadır. Boosting özelliği sayesinde zayıf olan öğrenme kabiliyeti olan noktalar birleştirilerek daha güçlü yapılar elde edilmesini sağlamaktadır. Gradyan arttırma özelliği sayesinde ise yaşanacak kayıplar bulunup en aza indirilmesi sağlanmaktadır. Bu üç bileşen sayesinde Denklem 12'de gösterilen GBM algoritması elde edilmektedir. GBM algoritmasına GOSS ve EFB özelliklerinin kazandırılması ile birlikte LGBM algoritmasının yapısı elde edilmektedir. GOSS ile birlikte, ağacın büyümesinde daha iyi öğretilmiş örnekler kullanılmaktadır. EFB ile birlikte ise, özellikler bir araya getirilir ve öğrenme hızlandırılmış olur. Büyük veriler için kazandırılmış bir özelliktir.

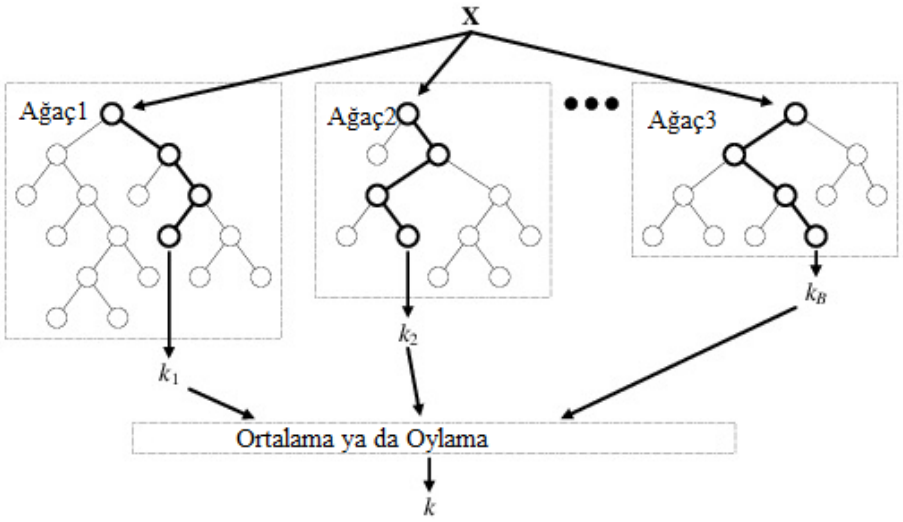
Yapısındaki bu özellikler nedeniyle LGBM algoritması, büyük verilerle çalışabilir yapıdadır ve minimum seviyede ram tüketmektedir. LGBM algoritmasının minimum 10.000 satır veri ile kullanılması tavsiye edilmektedir. Algoritmanın küçük veriler üzerinde takılmalar yaşattığı bilinmektedir. Ayrıca LGBM'in tercih edilmesinin en önemli unsurlarından biri doğruluğa odaklı bir algoritma olmasıdır. Algoritmada 100'den fazla parametre bulunmaktadır. Bu özelliği sayesinde algoritmaya istenen esnek yapının katılması sağlanmaktadır. LightGBM algoritması, performansının diğer karar ağaçlarına göre yüksek olması ve büyük verilerde çalışma anında (runtimeda) hızlı sonuçlar verebilmesi nedeniyle sıkça tercih edilmektedir (Ke vd., 2007).

d. Rastgele Orman

Rastgele orman (RO) algoritması, tıpkı karar ağaçları gibi denetimli öğrenme yapan, sınıflandırma ve regresyon için kullanılan makine öğrenmesi algoritmalarından biridir. Rastgele orman algoritması temelde, öğrenme aşamasına gelmiş birkaç karar ağacının bir araya gelmesiyle oluşturulmaktadır. Kullanılan karar ağaçları birbirinden ayrı olarak eğitilmiştir ve aynı durumda farklı kararlar verebilecekleri bilinmektedir. Rastgele orman algoritması, bu karar ağaçlarından elde edilen sonuçların ortalamasının alınması ve yeni bir tahmin yöntemi kullanılması esasına dayanmaktadır. İhtiyaç doğrultusunda istenildiği sayıda karar ağacı kullanılabilmesi sayesinde diğer sınıflandırma yöntemlerine göre daha avantajlıdır (Nisbet vd., 2009). Şekil 14'te rastgele ormanı oluşturacak karar ağaçları gösterilmektedir. Her bir karar ağacından elde edilen sonuçlara göre, her bir ağaç için en uygun yol belirlenmektedir. Şekil 15'te ise yöntemiyle yeni sonuçlar elde edilmektedir. Rastgele orman çalışma mantığı nedeniyle çok büyük bir ağaca benzetilmektedir.



Şekil 14: Rastgele Ormanı Oluşturacak Karar Ağaçları



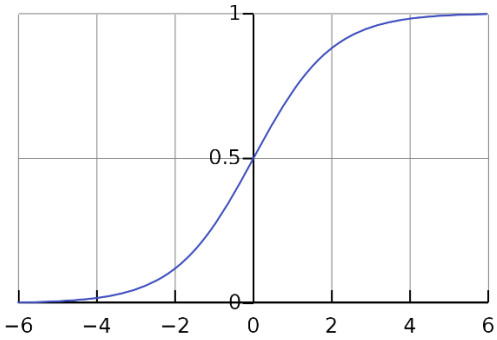
Şekil 15: Rastgele Orman Yapısı

e. Lojistik Regresyon

Lojistik regresyon algoritmasında kullanılan veri seti bir veya daha fazla bağımsız değişkenden oluşmaktadır. Lojistik regresyonun amacı bu veri kümesini analiz etmektir. Lojistik regresyon; sayısal bir değer elde etmekten çok, bir sınıfa dahil olma olasılığını hesaplayan istatistiksel bir yönteme dayanan makine öğrenmesi algoritmasıdır ve ikili çıktılar elde edilen binary değişkenlerin modellenmesi için kullanılmaktadır. Binary sonuçlar yaygın olarak 1ve 0 olarak tanımlanmaktadır. Denklem 14'te lojistik fonksiyon formülü, Şekil 16'da ise fonksiyon gösterilmektedir (Nisbet vd., 2009).

$$f(x) = \frac{L}{1 + \mathrm{e}^{-k(x-x_0)}}$$

Denklem 14: Lojistik Fonksiyon Formülü



Şekil 16: Lojistik Sigmoid Fonksiyonu

f. Naive Bayes

Naive Bayes algoritması olasılıkları sınıflandırmak için kullanılan ve eğitimli öğrenme gerçekleştiren bir makine öğrenmesi algoritmasıdır. Sınıflandırma yapılırken veri setindeki verilerin sıklığı ve kombinasyonları kullanılmaktadır. Naive Bayes algoritması büyük ölçekli verilerde ve küçük ölçekli verilerde performanslı ve başarılı sonuçlar vermektedir. Veri setinde kullanılan özellikler arasında ilişkiler algoritmaya tam olarak tanımlanmadığı durumlarda bile algoritma bu durumu tölere edebilir yapıdadır.

Naive Bayes algoritmasında eğitim süreci yoktur ve sınıflandırma yapmak için değişkenler arasındaki bağımlılıklardan yararlanılmaktadır. Naive Bayes algoritması çalışma mantığını, Denklem 15'te gösterilen Thomas Bayesin Koşullu Olasılıklar

Teoreminden (Bayes Teorimini) esas almaktadır. Algoritmanın adında bulunan Naive ifadesi, değişkenlerin bağımsızlığından gelmektedir (Dimitoglou, 2012).

$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$$

Denklem 15: Bayes Teorimi

Bayes Teoremini şu şekilde açıklayabiliriz;

P : Olasılığı,

C : veri setini,

P(c) : c sınıfının ilk olasılığını,

P(c|x) : x'in c sınıfından olma olasılığını,

P(x) : bir verinin x olma olasılığını ifade etmektedir.

Çizelge 1: Golf Oynama Sıklığı Veri Seti

Hava Durumu	Golf Oynama	
	Evet	Hayır
Güneşli	3	2
Bulutlu	4	0
Yağmurlu	2	3

Çizelge 1'de gösterilen örnek veri setinde hava durumuna göre golf oynama sıklıkları verilmektedir. Bu veri setini Çizelge 2'de gösterilen veri setine dönüştürüp değerlendirilebilir.

Çizelge 2: Veri Seti Değerlerinin Bütüne Uyarlanması

Hava Durumu	Golf Oynama		
	Evet	Hayır	
Güneşli	3/9	2/5	5/14
Bulutlu	4/9	0/5	4/14
Yağmurlu	2/9	3/5	5/14
	9/14	5/14	

Çizelge 2'de veri setleri bütüne bakılarak yeniden düzenlenmektedir. Örneğin, güneşli havada golf oynamaya evet diyenlerin olasılığı şu şekilde hesaplanmaktadır;

P(x | c) = P(Güneşli | Evet) = 3/9 = 0.33

P(x) = P(Güneşli) = 5/14 = 0.36

P (c) = (Evet) = 9/14 = 0.64

P(x | c) = P (Evet | Güneşli) = 0.33 x 0.64 + 0.36 = 0.60 olarak elde edilmektedir.

IV. UYGULAMA

Uygulama bölümünde veri seti, modelin eğitime hazırlık süreci, özellikler arasındaki ilişkiler, modelin eğitimi ve sonuçların alınması aşamaları incelenmektedir.

A. Veri Seti

Çalışmada işletmeye ait üç farklı veri seti girdi olarak kullanılmaktadır. Bu veri setlerinden ilki Çizelge 3'te gösterilen ürün özelliklerini içermektedir. İkinci veri seti, Çizelge 4'te gösterilen satış bilgilerini tutmaktadır ve üçüncü veri seti ise, Çizelge 5'te gösterilen satış tahminine konu olması beklenen ürün numaralarını bulundurmaktadır. Ürün özelliklerini ve satış bilgilerini bulunduran iki veri seti de (Çizelge 3 ve Çizelge 4) uzman görüşü alınarak karar verilen özelliklerden oluşmaktadır (EK A).

Çizelge 3: Ürün Özelliklerini İçeren Veri Seti Örneği

Ürün Numarası	Cinsiyet	Renk	Kategori No	Marka No	Alt Kategori No	Fiyat
1	1	476	938	603	83	95
2	2	476	407	1113	20	80
3	2	1886	407	1113	20	40
4	2	435	432	458	29	70
5	2	476	407	458	20	85

Çizelge 3'te gösterilen veri setinde ürüne ait özellikleri içermektedir ve 5 adet ürün bilgisi örnek olarak gösterilmektedir. Veri setinde kullanılan alanlar;

- Ürün Numarası: Ürün sürecini takip edebilmek adına verilen numara,
- Cinsiyet: Ürünün cinsiyetini ifade eder.
  - 0- Unisex veya cinsiyetsiz ürünler,
  - 1- Kadın,
  - 2- Erkek,
  - 3- Kız Çocuk,
  - 4- Erkek Çocuk,

5- Kız Bebek,
6- Erkek Bebek olarak ifade edilmektedir,

- Renk: Veri tabanı seviyesinde her rengin bir sayısal karşılığı bulunmaktadır,
- Kategori: Kadın-dış giyim bir kategoriye örnek verilebilir,
- Marka: Verileri kullanılan işletme, birkaç markanın ürününü online web sitesinde satışa açmaktadır. Her marka numarası bir markayı temsil etmektedir,
- Alt kategori: Kadın-dış giyim kategorisinin alt kategorisi olarak mont örnek verilebilir,
- Fiyat: Ürünün satış fiyatını temsil eder.

Çizelge 4: Ürün Satış Bilgisini İçeren Veri Seti Örneği

Ürün Numarası	Tarih	Satış Adedi	Stok	Görüntülenme Sayısı	Favori Sayısı
1	2018-11-08	0	72	27	0
2	2018-12-06	1	147	990	11
3	2018-12-21	0	69	9	1
4	2018-11-10	0	1	21	0
4	2018-11-19	0	1	35	0

Çizelge 4'te ise;

- Ürün numarası,
- Ürünün satış tarihi,
- Satış adedi,
- Stok adedi,
- Ürünü görüntüleyen kullanıcıların sayısı,
- Ürünü favoriye ekleyen kullanıcı sayısı bilgilerini bulundurmaktadır.

Ürün özelliklerini içeren veri setindeki bilgiler her ürün için farklılık gösterebilir. Ancak ürün özellikleri gibi verilerin tutulduğu tablolar tanım tablolarıdır ve bu tablolardaki veriler değiştirilmez, silinemezler.

Ürünün satış bilgisini içeren Çizelge 4'te tutulan veriler ise, çevresel ve ekonomik birçok faktöre göre, zamana göre değişiklik gösterebilecek verileri barındırmaktadır. Bu tablodaki verilerin yapılacak tahmin çalışması üzerindeki ağırlıklarının, Çizelge 3'e göre daha fazla olması beklenmektedir. Ürünün satış sayısı, görüntülenme ve

favoriye alınma sayılarının satış üzerinde doğru orantılı olarak bir etki yaratması geçmiş deneyimler göz önünde bulundurularak, büyük olasılık olarak değerlendirilmektedir.

Çizelge 3 ve Çizelge 4'teki veriler ayrı dosyalar içerisinde uygulamaya dahil edilmiştir. Uygulama içinde ise, ürün numarası alanı sayesinde birleştirilerek gereken durumlarda kullanılmaktadır.

Çizelge 5: Tahmin Edilmesi İstenilen Ürünlerin Olduğu ve Sonuçların Ekleneceği Veri Seti Örneği

Ürün Numarası	Tahmin Edilen Satış Adedi
1279	
8298	
8859	
15339	
21911	
36559	
100070	
103193	
108909	
118165	

Uygulamada kullanılan tüm ürünlerin satış tahmini yapılmaktadır. Çizelge 5'te ise tahmin edilmesi istenilen ürünlerin numara bilgileri bulunmaktadır. Modelin eğitiminin tamamlanması ve sonuçların oluşması işleminden sonra, yalnızca Çizelge 5'deki tahmin edilmesi istenilen ürünlerin kalacağı şekilde diğer ürünler için bir silme işlemi yapılmaktadır. Çalışma sonucunda bu veri setindeki ürünlerin kaç adet satılacağı tahmin edilmeye çalışılacaktır. Verilerin bir sınıfa dahil edilmesinden ziyade sayısal bir veri elde etmek amaçlanmaktadır.

Uygulamada geçmişe dönük 6 aylık satış verisi olmak üzere toplam 8611623 satış verisi ve 193732 ürün bilgisi kullanılmaktadır. Satış verileri eğitim ve test için ayrılırken haftalık olarak değerlendirilmiştir. Uygulamada her çalıştığında 6 aylık dönem içerisinde herhangi bir hafta tahmin edilmektedir. Verilerin yaklaşık %90'ı modelin eğitiminde kullanılırken %10'luk bir kısmı test verisi olarak ayrılmaktadır.

Modelin eğitime başlamadan veri setinden; günümüz dünyasında, dünyanın birçok ülkesinde her yıl gerçekleşen Black Friday, en uzun gece, yılbaşı ve geleneksel bayramlar gibi büyük ölçekli kampanyalar ve işletmelerin satış arttırmaya yönelik dönemsel olarak yaptığı küçük ölçekli kampanyalar çıkarılmıştır. Bunun nedeni; bu kampanyaların, elde edilecek sonuçları değiştirebileceğinin öngörülmesidir. Böylece gerçeğe en yakın, en başarılı tahminin yapılması hedeflenmektedir.

B. Modelin Eğitime Hazırlık Süreci ve Özellikler Arasındaki İlişkilerin İncelenmesi

Bu bölümde modelin eğitiminden önce yapılacak olan; verilerin okunması, normalizasyon, özelliklerin belirlenmesi ve özellikler arasındaki ilişkilerin incelenmesi ele alınacaktır. Özellikler arasındaki ilişkilerin bilinmesi, modelin eğitim sürecinde hangi özelliklerin ağırlıklarının daha fazla verilmesi konusunda yön verecektir.

1. Kullanılan Yardımcı Kütüphaneler

Uygulamada ihtiyacımız olan kütüphanelerin tamamı Python ile birlikte otomatik olarak gelmemektedir. Bu kütüphaneleri uygulama içerisine Şekil 17'de gösterildiği gibi dahil ederek kullanmamız gerekmektedir.

```
# Import libraries
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

Şekil 17: Yardımcı Kütüphaneler

Pandas kütüphanesi Python programlama dili için yüksek performans sunar ve veri analizi için kullanılmaktadır. Pandas ile dataframe'den veri eklenebilir ya da veri silinebilir. Pandas csv ve text dosyalarını açmak için kullanılmaktadır. Bu dosyaların içerisindeki verilerin okunmasını sağlar ve işlem yapılmasına olanak tanımaktadır. Pandas kütüphanesi çok hızlı çalıştığı için özellikle büyük veriler için büyük avantajlar sağlar. Uygulama içerisinde kullanılan veri setleri cvs dosyalarından oluşmaktadır. Pandas kütüphanesinin bu uygulamada kullanım amacı bu veri setlerini okuyarak işlem yapmaktır.

Numpy; temelde bilimsel işlemleri yapmamızı sağlayan bir kütüphanedir. Matematiksel işlemleri yapmamızı sağlar. Uygulama içerisinde numpy kütüphanesi,

modelin performansının hesaplanmasında ve tahmin hesaplamalarında kullanılmaktadır.

Seaborn kütüphaneleri verileri görselleştirmek için kullanılmaktadır. Pandas kütüphanesiyle uyumlu bir şekilde çalışmaktadır. Daha çok istatistiksel verilerden çıktı elde etmek için kullanılmaktadır. Uygulama içerisinde;

- Satış miktarı ve tarih arasındaki ilişki,
- Satış miktarı, ürün özellikleri ve tarih arasındaki ilişki,
- Oluşturulan modelden elde edilen sonucun gösterilmesinde seaborn kütüphanesi kullanılmaktadır.

Matplot kütüphanesi sayesinde iki boyutlu grafik çizimleri oluşturulabilir. İleri seviyede yüksek görselli çıktılar sunamaz fakat verilerden elde edilecek grafikler için başarılı çıktılar oluşturmaktadır. Matplot kütüphanesi uygulama içerisinde, elde edilen gelirin görselleştirilmesinde ve grafikler elde etmekte kullanılmıştır.

2. Normalizasyon

Veritabanı seviyesinde yapılan normalizasyon işlemi; büyük verilerden oluşmuş bir tabloda, tekrarlı veya boş kalan kayıtların ayıklanması işlemidir. Yazılacak uygulama, çok fazla satır ve sütundan oluşmuş bir veri tabanından beslenecekse eğer, uygulamadan en doğru çıktıları alabilmemiz için verilerin tutarlı olmasına dikkat etmeliyiz. Kullanılan veri tabanında boş kayıtlar olması ya da aynı kayıtların tekrar ediyor olması sonucu doğrudan etkileyen etkenlerdir. Bu nedenle büyük ölçekli verilerde normalizasyon büyük önem taşır.

Uygulama içerisinde yapılan ilk işlem, veri setlerini okuyarak boş kayıtların olup olmadığını kontrol etmektir. Şekil 18'deki komut, öncelikle train_file isimli dosyayı okumaktadır ve daha sonra varsa bu dosyadaki boş satır içeren kayıtlar bulunmaktadır.

```
import pandas as pd
df = pd.read_csv(train_file, parse_dates=['date'], infer_datetime_format=True, dayfirst=True)
df.isna().sum()
```

Şekil 18: Import Pandas

Boş olduğu tespit edilen kayıtlara, Şekil 19'daki komut sayesinde ortalama değerler atanmaktadır. Bu işlem tüm sütunları içerecek şekilde yapılmıştır ve yine Şekil 19'da

bu işlemin görüntülenme sayısı ve favoriye alınma sayıları için yapıldığı görüntülenmektedir.

```
df['clickcount'].fillna((df['clickcount'].mean()), inplace=True)
df['favoredcount'].fillna((df['favoredcount'].median()), inplace=True)
df.head()
```

Şekil 19: Ortalama Değer Atama

3. Pivot Tabloların Oluşturulması

Uygulamada girdi olarak ele alınan veriler günlük verilerden oluşmaktadır. Ancak uygulamadan tahmin edilmesi beklenen zaman dilimi bir haftalık bir periyottur. Verilerin daha kolay işlenebilmesi ve daha anlaşılır yapıya getirmek amacıyla veriler Şekil 20'de haftalık olarak pivot tablolara dönüştürülmektedir. Şekil 21'de ise her haftanın başladığı gün ile isimlendirme yapılacak şekilde veriler isimlendirilmektedir.

```
df['date'] = pd.to_datetime(df['date']) - pd.to_timedelta(7, unit='d')
df = df.groupby(['productid', pd.Grouper(key='date', freq='W-MON')])
['soldquantity','clickcount','favoredcount'].sum().reset_index().sort_values('date')
df = df.pivot_table(index=['productid'], columns='date'
,values=['soldquantity','clickcount','favoredcount'],fill_value=0).reset_index()
```

Şekil 20: Pivot Tabloların Oluşturulması

```
df.columns=[i+'_'+str(j)[:-9] if i!='productid' else i for i,j in df.columns]
df.head()
```

Şekil 21: Pivot Tabloların İsimlendirilmesi

4. Ürün Özellikleri Arasındaki İlişki

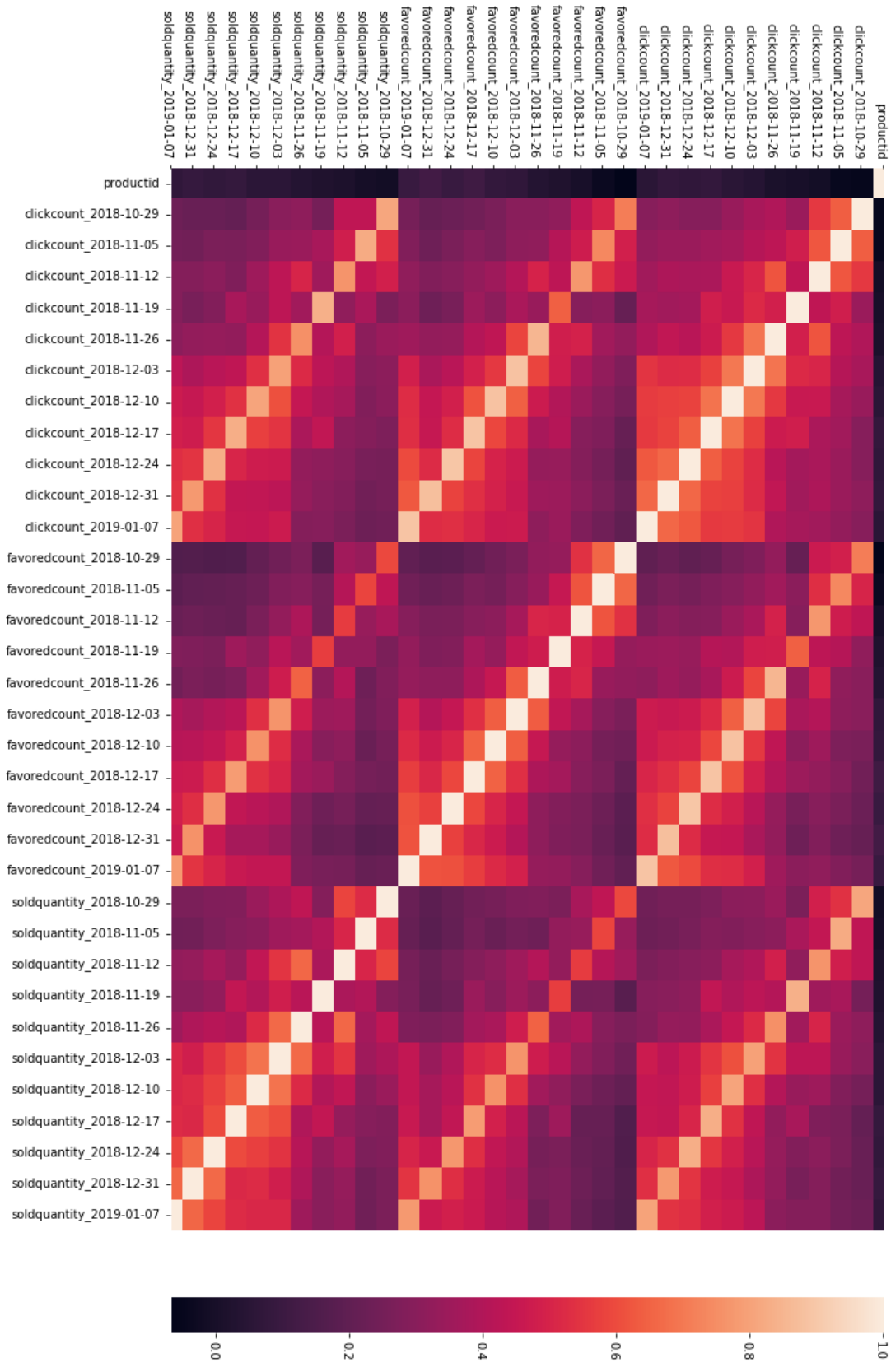
Uygulamada regresyon analizi yapılmaktadır. Regresyon analizi, ürün özellikleri arasındaki ilişkilerin çıkarılması ve bu ilişkilere göre tahmin yapılması esasına dayanmaktadır. Bu nedenle, bu bölümde uygulamada kullanılan özellikler arasındaki ilişkiler incelenmektedir. Elde edilen bu veriler modelin eğitiminde büyük rol oynamaktadır.

a. Satış Adedi, Favori Sayısı ve Görüntülenme Sayısı Arasındaki İlişki

Şekil 22'de ürünlerin satış miktarı, görüntülenme ve favoriye alınma sayılarının bulunduğu haftalar arasındaki ilişki görüntülenmektedir. Ürün numarası (productid) alanının daha koyu renkte olması, ürün numarasının bu tablo üzerinde hiçbir etkisi olmadığını göstermektedir. Bu sonuç elde edilirken Çizelge 3'teki ürün özelliklerine ait veriler dahil edilmemiş, yalnızca Çizelge 4'teki ilgili ürünlere ait; ürünün satış

adedi, ürüne görüntülenme sayısı, ürünün favoriye alınma sayısı gibi veriler kullanılmıştır.

Şekil 22'de bulunan grafiğin en altındaki değerler bize, kesişim noktalarındaki ağırlıkları ifade etmektedir. Grafikteki durum çubuğuna bakıldığında, koyu renklerden açık renklere gidildikçe kesişen noktaların ağırlıklarının daha fazla olduğu görülmektedir. Örneğin, ürün numarası (productid) sütunlarının kesiştiği yerlerin tamamen koyu renkli olduğu görülüyor. Bu değer bize ürün numarası (productid) sütunun hiçbir ağırlığı olmadığını göstermektedir. Başka bir örnek ile; clickcount_2018-10-29 isimli sütunun yine aynı isimdeki clickcount_2018-10-29 isimli sütunla olan kesişim noktasının ağırlığının 1 olduğunu görmekteyiz. Bu durum iki sütun arasındaki ilişkinin çok yüksek olduğunu bize ifade etmektedir. Bu grafiğin oluşturulması ile asıl elde edilmesi amaçlanan durum; görüntülenme sayısı, favoriye alınma sayısı ve satış miktarları arasında bir ilişki olup olmadığının gözler önüne serilmesidir.

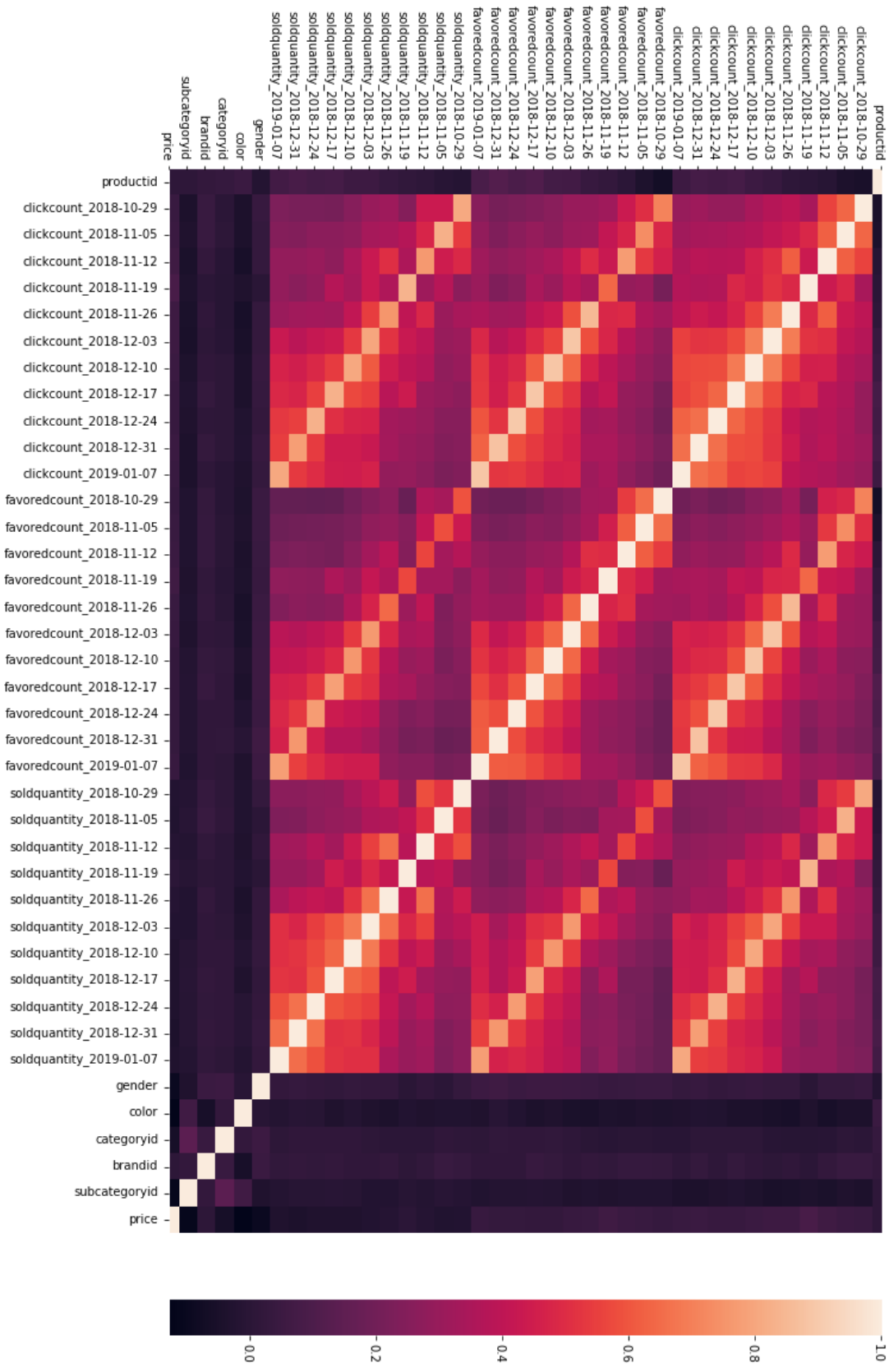


Şekil 22: Satış Adedi, Favori Sayısı ve Görüntülenme Sayısı Arasındaki İlişki

b. Satış Adedi, Favori Sayısı, Görüntülenme Sayısı ve Ürün Özellikleri Arasındaki İlişki

Şekil 22'deki sonuçlar, Çizelge 4'teki veriler kullanılarak elde edilmektedir. Şekil 23'teki sonuçlar ise, Çizelge 3 ve Çizelge 4'teki verilerin birleşiminden elde edilmektedir. Çizelge 4'teki, ürünün görüntülenme sayısı, favoriye alınma sayısı ve satış adedine ek olarak; Çizelge 3 ile birlikte; ürün numarası, renk, kategori, alt kategori, marka ve fiyat sütunları da sonuca dahil edilmektedir. Böylece tüm bu sütunların korelasyon üzerinde etkisi olmadığı görülmektedir.

Şekil 23'te; ürünün numarası, cinsiyeti, rengi, kategorisi, alt kategorisi ve markasının satışa olan etkisinin sıfıra yakın hatta sıfır olduğunu görülmektedir. Modelin eğitim sürecine başlamadan bu bilgilere sahip olmak bize avantajlı bir durum sağlamaktadır. Bir tahmin modeli oluştururken, en iyi sonuç verilere bakılarak elde edilmektedir. Bu nedenle veri modelindeki özellikler arasındaki ilişki çok açık bir şekilde bilinmesi başarılı bir model oluşturabilmek adına büyük önem taşımaktadır. Çalışmanın buraya kadar olan kısmında elde ettiğimiz her iki grafik de bize modelin eğitimine başladığımızda hangi özelliklerin ağırlıklarının daha fazla verilmesi konusunda yön verecektir.



Şekil 23: Satış Adedi, Favori Sayısı, Görüntülenme Sayısı ve Ürün Özellikleri Arasındaki İlişki

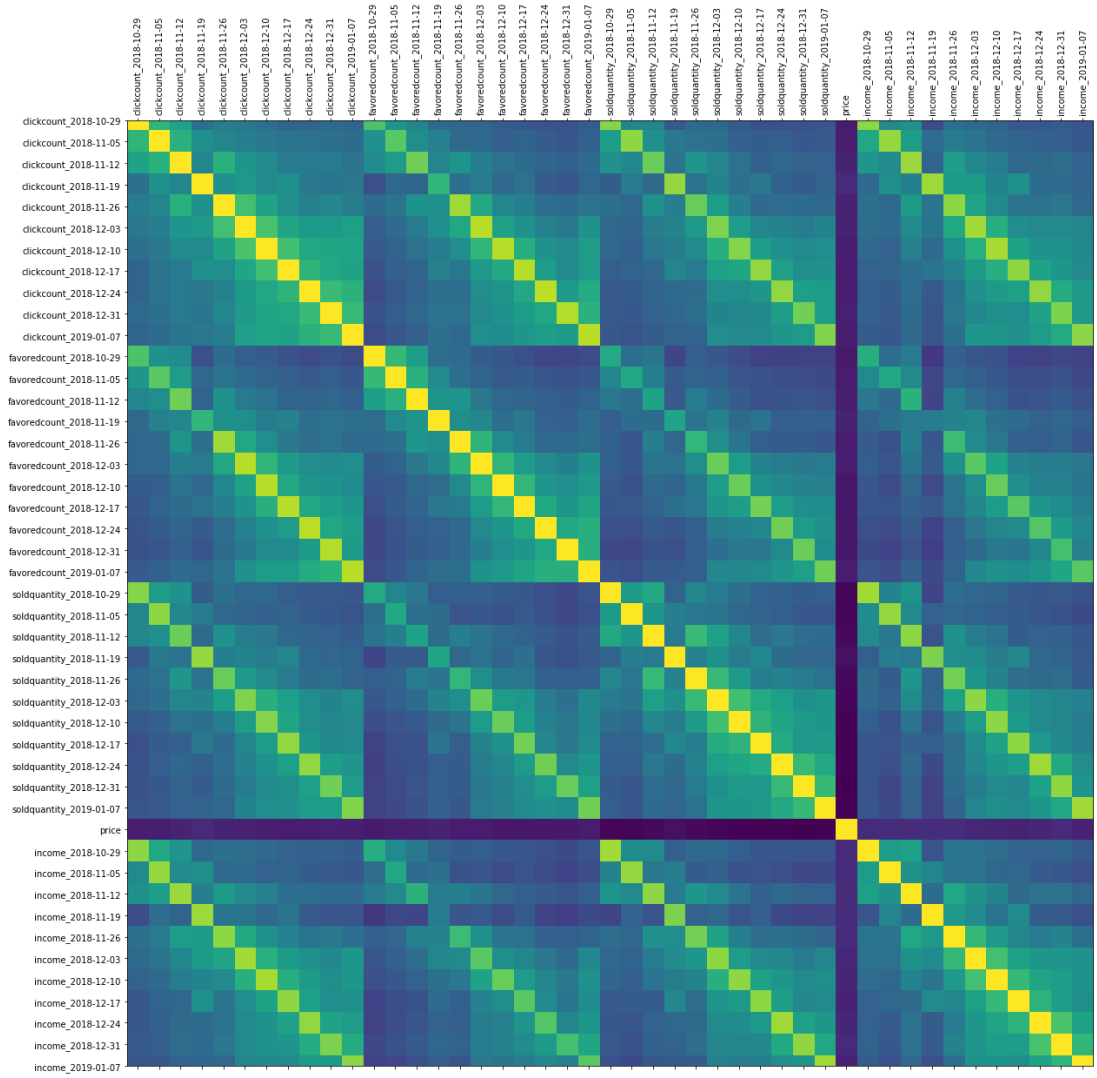
c. Satış Adedi, Favori Sayısı, Görüntülenme Sayısı ve Gelir Arasındaki İlişki

Şekil 24'ü elde etmek için öncelikle ürünlerden haftalık olarak elde edilen gelirler hesaplanmaktadır. Gelir hesaplanırken Denklem 16'dan yararlanılmıştır. Hesaplanan haftalık gelir ile, ürünlerin favoriye eklenme, görüntülenme sayıları ve satış adetleri arasındaki ilişki Şekil 24'te görülmektedir.

$$Gelir = Fiyat * Satış\ Adedi$$

Denklem 16: Gelir Hesaplama Denklemi

Hesaplanan gelirler yeni sütunlarda tutulmaktadır ve tarihlere göre yeni pivot tabloları uygulama içerisinde oluşturulmuştur. Oluşan korelasyon grafiğine bakıldığında satış adedi ve gelir arasında bir ilişki olduğu gözlemlenmektedir. En ideal sonuca ulaşabilmek adına elde edilen bu haftalık gelir sütunlar daha sonra modelin eğitimi aşamasında kullanılacaktır.



Şekil 24: Satış Adedi, Favori Sayısı, Görüntülenme Sayısı ve Gelir Miktarı Arasındaki İlişki

5. Eğitim ve Test Verilerinin Belirlenmesi

Uygulama her çalıştırıldığında veri setindeki herhangi bir hafta test verilerine ayrılmaktadır ve teste ayrılan veri seti tahmin edilmeye çalışılmaktadır. Bu yapıya uygun bir uygulama geliştirebilmek için yapılan gerekli çalışmalar Şekil 25'te görülmektedir. Bu kısmı kavrayabilmek için birkaç komuta hâkim olmamız gerekmektedir. Bu komutlar aşağıda kısaca açıklanmaktadır.

- **Sklearn.Preproccessing:** Bu kütüphane sayesinde, minimum ve maksimum olacak şekilde iki değer aralığı belirlemekteyiz. Bu aralık scaler adını verdiğimiz bir değişkene atanmaktadır. Burada belirlediğimiz aralık -1 ve 1'dir. Bu aralığı vermemizin nedeni, belirli bir hassasiyet yakalamaktır. Bu değerlerin yüksek olduğu bir durumda performanslı çözümler elde edemeyebiliriz.

- **As_Matrix:** veri serisini numpy dizisine çevirmek için kullanılmaktadır.
- **Scaler.fit_transform:** Bir ölçeklendirme yöntemi olarak kullanılmaktadır. Denklem 17'ye göre ölçeklendirme mekanizması çalışmaktadır.

$$\frac{X - Xmin}{Xmax - Xmin}$$

Denklem 17 : Scaler.fit_transform Denklemi

- **Reshape:** Numpy dizisini yeniden şekillendirmek için kullanılmaktadır. Burada satır ve sütun sayılarını değiştirmek için kullanılmıştır. Reshape(193731,40) komutu ile 193731 adet satır ve 40 adet sütun oluşacak şekilde bir dizi oluşturulmaktadır.

```
# # Model: **LightGBM**

# ### **Feature Scaling**
from sklearn.preprocessing import StandardScaler,MinMaxScaler
scaler = MinMaxScaler((-1,1))

X_train = X_train.as_matrix()
X_train = scaler.fit_transform(X_train)
X_train = X_train.reshape((193731, 40))

y_train = y_train.as_matrix()
y_train = y_train.reshape(193731, 1)


print(y_train.shape)
print(X_train.shape)

X_test = df.drop(labels=['soldquantity_2018-10-29','gender', 'color', 'categoryid', 'brandid','subcategoryid'],axis=1)

X_test = X_test.drop(columns=black_friday_cols)
X_test = X_test.as_matrix()
X_test = scaler.fit_transform(X_test)

X_test = X_test.reshape((193731, 40))
print(X_test.shape)
```

Şekil 25: Eğitim ve Test Verilerinin Formatlanması

C. Modelin Eğitimi

Satış verileri kullanılan firmanın geçmişe dayalı verileri analiz edildiğinde, geçmişteki haftaların satışa az da olsa etkisi olacağı bilinmektedir. Veri analiz ekibinin uzman görüşleri dikkate alınarak ve verilerin incelenme sürecinden sonra bu sonuca ulaşılmaktadır (EK A). Algoritmanın ihtiyaca göre modellenmesi durumunda, veri setindeki haftaların ağırlıklarının beklenen durumla örtüştüğü görülmektedir. En iyi sonucu elde edebilmek amacıyla kullanılan LightGBM algoritmasında aşağıdaki değişiklikler yapılarak yeni bir tahmin modeli oluşturulmaktadır.

Model eğitimi süresince bir iterasyonda yapılan hatanın bir sonraki iterasyonda öğrenilmiş bir değer olabilmesi için, Gradient Boosting yöntemi kullanılmaktadır.

Şekil 26'da görüleceği üzere bu yöntem sayesinde daha başarılı sonuçlar üreten bir model olması beklenmektedir.

```
import lightgbm as lgb
try:
    del model
except:
    pass
model=lgb.LGBMModel(boosting_type= 'gbdt',
```

Şekil 26: Gradient Boosting

Modelden beklenen sonuç bir haftalık satışı tahmin etmesi olduğu için, Şekil 27'de görüldüğü üzere modeli regresyon analizi yapabilecek şekilde tasarlamaktayız.

```
objective= 'regression'
reg_alpha= 0.1,
reg_lambda= 0.1,
```

Şekil 27: Regresyon Analizi ve Ayarlamaları

Şekil 24'de elde edilen haftalık gelir miktarının diğer özelliklere oranla daha fazla ağırlığının olması gerektiğini gözlemlemekteyiz. Şekil 28'de haftalık gelir miktarının hesaplanışını görüyoruz. Şekil 29'da ise haftalık gelir miktarı (income) adını verdiğimiz sütunun ağırlığı arttırılmıştır. Böylece gelir miktarı özelliği, modelimiz için daha büyük bir önem kazanmaktadır.

```
sold_qs=[col for col in df.columns if 'soldquantity' in col]
sold_qs
for col in sold_qs:
    date=col.split('_')[1]
    df['income_'+date]=df['price']*df[col]

df.head()
```

Şekil 28: Gelir Hesaplama

```
class_weight= None,
```

Şekil 29: Gerekli Sütunlara Ağırlık verilmesi

Şekil 30'da yeni yapılandırılan model, bir önceki model gibi 10000 iterasyon eğitilmeye devam ediliyor. Fakat 50 iterasyon boyunca herhangi bir iyileşme sağlanmıyorsa, en iyi sonucu veren iterasyonda modelin kalması sağlanmaktadır. Böylece model eğer daha performanslı sonuçlar üretemezse eğitimi durduracak ancak model daha hızlı çalışır yapıda olacaktır.

```
num_boost_round= 10000,
num_leaves= 50.
```

Şekil 30: İterasyonun Belirlenmesi

Şekil 31 ‘de karar ağacına kazandırılmış bazı özellikler gösterilmektedir. Ağacın ulaşabileceği en uzak noktalar, kullanılan veri setine uygun olarak belirlenmektedir. Bu değer aynı zamanda eğitimin ne kadar uzun süreceği ile doğru orantılıdır.

```
min_data_in_leaf= 9,
min_split_gain= 0.0,
max_depth= 20,
n_estimators= 100,
```

Şekil 31: Ağacın Maksimum Uzaklığının Belirlenmesi

Şekil 32’de alt düğümlerde bulunacak maksimum veri sayısı ve ağırlık gibi öğrenmeyle ilgili sürecin iyileştirilmesi gösterilmektedir. Öğrenme sürecinde düğümler arasında geçişte yaşanacak veri kayıpları en aza indirilmeye çalışılmıştır.

```
min_child_samples= 20,
min_child_weight= 0.001,
```

Şekil 32: Alt Düğüm Ayarları

Şekil 33’te işlemci ile ilgili yapılan ayar gösterilmektedir. Uygulamanın daha performanslı olması için bu ayar modele kazandırılmıştır.

```
n_jobs= -1,
```

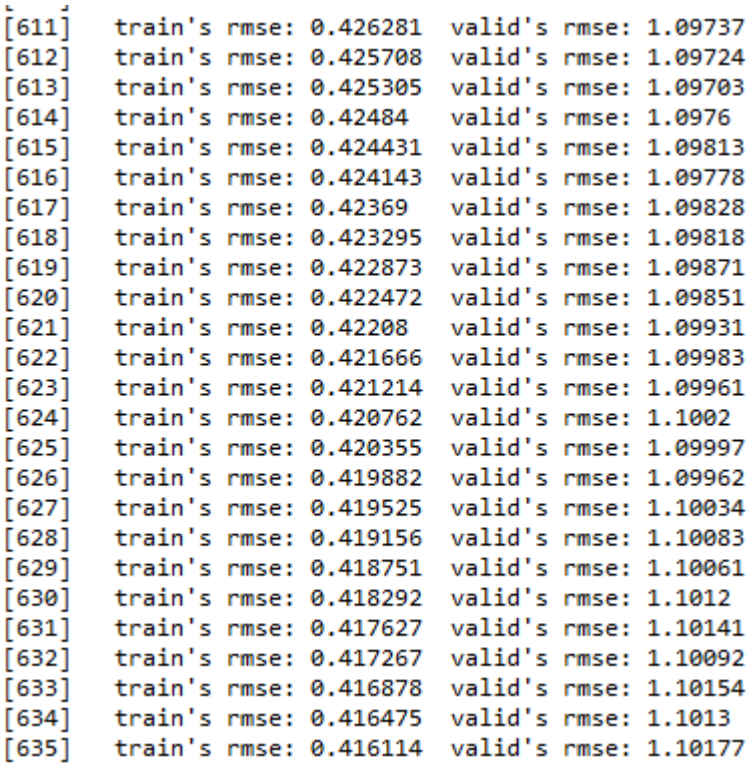
Şekil 33: İşlemci Performansı

Modelin herhangi bir durumda hata ile karşılaşması veya modelde takılmalar yaşanıp hep aynı sonuçları üretmesi modelin başarısını düşürmektedir. LightGBM böyle durumlarda ve hatta veri kayıplarında çalışmaya devam eden bir algoritmadır. Ancak bu durumun sonuçlara olacak olumsuz etkisi göz önünde bulundurulduğu için olası hatalarda sistemi durduran ve uyarı veren Şekil 34’ de gösterilen kontroller eklenmiştir.

```
silent= True
```

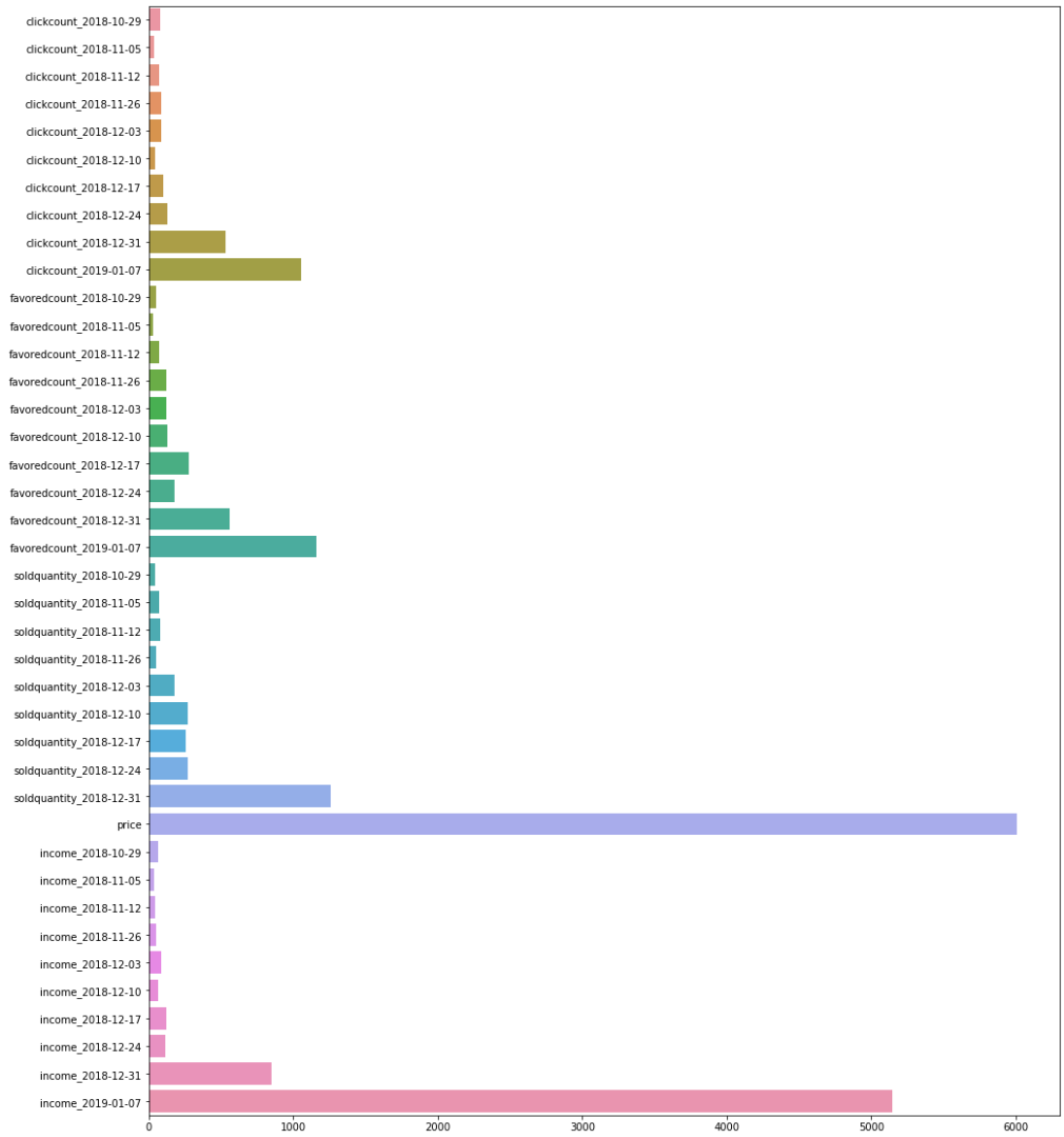
Şekil 34: Uyarıların Çıkarılması

Şekil 35’te modelden elde edilen kök ortalama kare hataları gösterilmektedir. Bu değer 0’a ne kadar yakınsa modelin o kadar başarılı olduğunu söylemek mümkündür. Eğitime devam edildikçe değerin 0’a yaklaştığı görülmektedir.



Şekil 35: Kök Ortalama Kare Hatası

Uygulamada kullanılacak modelin eğitimi tamamlandıktan sonra ürünün hangi özelliklerinin ağırlıklarının satış tahminine etkisinin fazla olduğunu Şekil 36'da elde etmekteyiz. Oluşan özellik önemi grafiği sonuçlarına göre tüketicilerin ürünleri tercih ederken dikkat ettikleri en önemli özelliğin ürün fiyatı olduğunu gözlemlemekteyiz. Aynı zamanda tüm haftaların satış tahminine etkisi olduğunu gözlemlemekteyiz. Geçmişe dönük satış verilerinin matematiksel olarak analizi yapıldığında bu beklenen bir durum olduğu için, modelin tahmine etkisi olan ağırlık değerleri bize daha güvenilir sonuçlar alabileceğimizi göstermektedir (EK A).



Şekil 36: Veri Setinin Yeni Model Üzerindeki Ağırlıkları

D. Sonuçların Alınması

Uygulamada kullanılan iki modelin de eğitimi yukarıdaki adımlar ile tamamen tamamlanmış olmaktadır. Test edilecek verilerin bulunduğu cvs dosyasının modele tanımlanması Şekil 37'de gösterilmektedir. Tanımlanma işleminden sonra dosya içerisindeki ürün numaraları okunmaya başlanmaktadır. Bu numaralar test aşamasında kullanılmak üzere numpy dizisi olarak kaydedilmektedir.

```
sub_file = 'Submission.csv'
submission = pd.read_csv(sub_file)
submission.head()
```

Şekil 37: Tahmin edilecek Ürün Numaralarının Bulunduğu Dosya Sisteme Tanımlanıyor

Yapılan tahminler Şekil 38'de gösterildiği gibi Submission.csv içerisindeki Sales sütununa eklenmeye başlıyor.

```
preds = pd.DataFrame(y_pred, columns=['sales'])
```

Şekil 38: Tahminler Satış Sütununa Ekleniyor

Son olarak Şekil 39'da görüleceği üzere sonuçların listeleneceği dosya oluşturuluyor.

```
last.to_csv('Result.csv',index_label='productid')
```

Şekil 39: Sonuçlar İçin Csv Dosyası Oluşturuluyor

Çizelge 6'da gösterilen veri seti, tahmin edilmesi beklenen ürünlerin numaralarını içermektedir. Veri setinin içerdiği bir diğer sütun olan tahmin edilen satış adedi ifade etmektedir. Eğitim süreci biten model, test verileri ile test sürecine sokulmaktadır. Çizelge 6'da gösterilen örnek tablo, csv formatında modele dahil edilmektedir. Test işleminin bitmesiyle modelin, bu tablodaki (csv dosyasındaki) tahmin edilen satış adedi sütununu doldurması beklenmektedir. Çizelge 7'de yeni oluşturulan tahmin modelinden alınan sonuçlar görülmektedir.

Çizelge 6: Tahmin Edilmesi İstenilen Ürün Veri Seti

Ürün No	Tahmin Edilen Satış Adedi
1279	
8298	
8859	
15339	
21911	
36559	
100070	
103193	
108909	
118165	

Çizelge 7: Yeni Oluşturulan Model Tahmininden Alınan Sonuçlar

Ürün No	Tahmin Edilen Satış Adedi
1279	8
8298	34
8859	10
15339	10
21911	13
36559	33
100070	3
103193	16
108909	41
118165	15

Çizelge 8'de oluşturulan modelin R-Square, Rmse ve Mape hata metriklerinde değerleri gösterilmektedir. Satış tahmin uygulamalarında geleceğe yönelik bir tahmin yapıldığı için ve gelecekteki dönem sürekli olarak değişkenlik gösterebileceği için hata metriklerinde tam bir başarı sağlanması mümkün olmamaktadır. Hata metriklerinin tek başına değerlendirmeye alınması doğru sonuca götürmeyecektir (Demirtaş, 2011). Yapılan çalışmanın satış tahmini olması nedeniyle elde edilen hata metriklerinin başarılı olduğunu söylemek mümkündür.

Çizelge 8: LightGBM Algoritması Hata Metrikleri

R-Square	RMSE	MAPE
0.7210	2.5601	49.4809

V. DENEYSEL ÇALIŞMALAR

Bu bölümde uygulamada kullanılacak regresyon modelinin, algoritmanın ve teknolojilerin seçilmesi, LightGBM temelli yeni bir model oluşturma ihtiyacı konuları ele alınacaktır.

A. Regresyon Modelinin Seçimi

Veri madenciliğinde kullanılan modeller kestirici ve tanımlayıcı olarak iki ana gruba ayrılmaktadır. Kestirici modeller sınıflandırma ve regresyon yöntemlerini içermektedir. Tanımlayıcı modellere ise kümeleme ve birliktelik yöntemlerinin kullanılmasında başvurulmaktadır. Kullanılacak yöntemin belirlenmesi, problemin tanımıyla doğrudan ilişkilidir. Yapılan çalışmada ihtiyaç sayısal bir verinin tahmin edilmesi olduğu için regresyon analizi kullanılmaktadır.

Regresyon yöntemlerinde sayısal veriler kullanılmaktadır ve sayısal girdi değişkenlerine bakılarak yanıt (response) olarak adlandırılan değişkenin değerini bulmak amaçlanmaktadır. Regresyon analizlerini diğer yöntemlerden ayıran en önemli özellik kategorik veri tiplerinde kullanılması ve sayısal bir değerin tahmin edilmesinin amaçlanmasıdır.

Çalışmada kullanılan veriler bir süreklilik ifade ettiği için en uyumlu regresyon yöntemlerinin doğrusal regresyon veya regresyon ağaçları olacağı kanısına daha önce yapılmış tahmin çalışmalarına dayandırılarak varılmıştır (Sutton, 2005). Bu bölümde en uygun yöntemin bulunabilmesi için doğrusal regresyon ve regresyon ağaçları karşılaştırılmaktadır.

Denklem 18'de doğrusal regresyonun formülü gösterilmektedir. Formülde gösterilen xi değerleri girdi değişkenlerini, yi değişkenleri ise yanıt (response) değişkenlerini belirtmektedir. Denklemde görüntülenen w x + b formülü ile hata kareler toplamını en küçük yaparak tahmin yapılmasını sağlamaktadır.

$$_{w,b}min \sum_{i=1}^{n} (y_i - (w^T x_i + \mathrm{b}))^2)$$

Denklem 18: Doğrusal Regresyon

Doğrusal regresyonda genellikle değişken sayısı, gözlem sayısından küçüktür. Eğer değişken sayısı gözlem sayısından büyük olursa denklem sonucu sıfıra eşitlenir ve aşırı uyum oluşur. Kullanılan yöntemde aşırı uyum olup olmadığı modelin eğitim ve test verileri üzerinde verilen hata oranlarına bakıldığında anlaşılmaktadır. Normal şartlar altında modelin test verileri üzerinde verdiği hata oranlarının daha yüksek olması beklenmektedir. Model test verisi üzerinde daha düşük hata oranı veriyorsa burada aşırı uyum olduğu söylenebilir.

Regresyon analizinin karar ağalarında kullanılması yöntemi ise 1984 yılında Bremain tarafından açıklanmıştır (Sutton, 2005). Ağaçlar gelişim dönemlerinde özyinelemeli olarak iki farklı gruba bölünürler. Bölünmeler son durdurma kriterlerine kadar devam eder. Oluşan modellerde genellikle öncelikli olarak büyük bir ağaç oluşturulur ve sonra ağacın zayıf noktaları budanır.

Geri budama işlemi ağacın aşırı uyumlanmasını engeller. Bu sayede regresyon ağaçlarından iyi sonuçlar alınabilmektedir. Regresyon ağaçları performanslı ve analiz sonucu çıktıların başarılı olduğunu söylemek mümkündür. Regresyon ağaçları karar verme aşamasında yanıt değişkeni ve açıklayıcı değişkenler arasındaki ilişkilere baktığı gibi doğrusal olmayan ilişkiler de göz önünde bulundurmaktadır. Aynı zamanda en büyük etkiye sahip olan değişkenlerin belirlenmesini sağlar ve o değişkenlerin ağırlığının arttırılmasına imkân verir.

Çizelge 9'da doğrusal regresyon ve regresyon ağacı kullanımında oluşan R-Square, Rmse ve Mape hata metrikleri karşılaştırılmaktadır. İki yöntem karşılaştırıldığında; R-Square ve Rmse hata metriklerinde doğrusal regresyonun daha başarılı sonuçlar verdiği görülse de Mape değerine bakıldığında regresyon ağacının çok daha başarılı sonuçlar verdiği görülmektedir. Satış tahminlerinin çok fazla değişkenlik göstermesi nedeniyle gelecek dönemdeki durumlarını öngörmek zor olmaktadır. Bu nedenle tek bütün parametrelerin birlikte değerlendirilmesi ve doğru yöntemin seçilmesi büyük önem taşımaktadır (Demirtaş, 2011). Tüm bu bilgiler doğrultusunda satış tahmininde yaygın olarak kullanılan ve en uygun görülen iki regresyon yöntemi arasından regresyon ağaçlarının uygulamada kullanılması uygun görülmüştür.

Çizelge 9: Doğrusal Regresyon ve Regresyon Ağacı Hata Metrikleri

	R-Square	**RMSE**	**MAPE**
Doğrusal Regresyon	0.8461	1.1903	93.5196
Regresyon Ağacı	0.7210	2.5601	49.4809

B. Algoritmanın Seçimi

Veriler karar ağacı algoritmalarından ve Boosting tabanlı olan; Xgboost, Catboost ve LightGBM tabanlı oluşturulan yeni algoritma olmak üzere toplam üç adet algoritma ile test edilmiştir. Algoritmalar eğitim ve test aşamalarını aynı bilgisayar sistemi üzerinde tamamlamış olup eşit koşullarda kıyaslama yapılmıştır. Algoritmalar 8611623 adet satış verisi ve 193732 adet ürün bilgisi kullanılarak eğitilmiştir.

Çizelge 10: Algoritmaların Eğitim Bakımından Hız Faktörüne Göre Kıyaslanması

Algoritma	Geçen Süre (Saat)
Xgboost	0:03:31.005703
Catboost	0:10:09.039207
LightGBM	0:00:22.535972

Çizelge 10'da görülen sonuçlara göre, Xgboost Algoritması 3 dakika 31 saniyede, Catboost Algoritması 10 dakika 9 saniyede ve LightGBM Algoritması 22 saniyede eğitim sürecini tamamlamıştır. LightGBM Algoritmasının büyük verilerde performanslı sonuçlar verdiği kanısına varılmıştır.

Bir diğer kıyaslama yöntemi olarak algoritmalar hata metriklerine göre kıyaslanmıştır. R-Square; modelin ne kadar iyi tahmin yapabildiğini ölçmektedir. Bu değer 0.12 altında ise modelin başarılı olmadığını söylemek mümkündür. En iyi değer 1 olarak kabul edilmektedir. Çizelge 11'de Xgboost ve Catboost algoritmaları için bu değerin 0.9, LightGBM için 0.7 aralığında olduğu görülmektedir. Bu nedenle üç modelin de başarılı olduğunu söylemek mümkündür.

RMSE; tahmin edilen değerler ile gerçekleşen değerler arasındaki farka bakarak algoritmanın doğruluğunu test eden yöntemdir. RMSE değeri 0 noktasına ne kadar yakınsa modelin o kadar başarılı olduğunu söylemek mümkündür. Eğitim verileri için

üç algoritma da göz önünde bulundurulduğunda; en iyi sonucu Xgboost Algoritmasının verdiği görülmektedir.

MAPE; tahmin değerlerinin gerçekleşen değerlere oranını veren hata metriğidir. Gerçekleşen tahminlerin düşük olduğu durumlarda bu değer büyük çıkmaktadır. Çizelge 11'deki MAPE değerlerine baktığımızda en iyi sonucu LightGBM Algoritmasının, daha sonra Catboost Algoritmasının ve Xgboost Algoritmalarının verdiği görülmektedir.

Çizelge 11: Test Verileri Üzerinde Hata Metriklerinin Karşılaştırılması

Algoritma	R-Square	RMSE	MAPE
Xgboost	0.9536	1.8725	80.1100
Catboost	0.9198	2.5736	80.6773
LightGBM	0.7210	2.5601	49.4809

Hata metrikleri kıyaslanırken üç değerin de birbiriyle ilişkili olduğu göz önünde bulundurulmuştur. Bir diğer kıyaslama kriteri olan algoritmanın performansına göre en başarılı sonucun LightGBM'den alındığı görülmektedir. Her iki algoritma üç dakikanın üzerinde eğitim sürecini sonlandırırken LightGBM Algoritmasının yalnızca 22 saniyede eğitimi tamamlaması ve hata metriklerinde başarılı sonuçlar vermesi doğrultusunda algoritmanın uygulamada kullanılması uygun bulunmuştur.

C. Uygulamada Kullanılan Teknolojilerin Seçimi

Bu çalışmada, karar ağacı türlerinden biri olan LightGBM kullanılmaktadır. Programlama dili olarak Pyhton kullanılırken, geliştirme ortamı olarak Anaconda kullanılmaktadır. Yapılan denemeler ve literatür taramaları sonucunda, çalışmada yöntem olarak bir karar ağacı türü olan LightGBM (LGBM)'in kullanılması kararlaştırılmıştır. Uygulamada karar ağacı kullanılmasının nedeni, karar ağaçlarının satış tahmini uygulamalarından elde edilen çıktıların doğruluk oranlarının yüksek olduğunun gözlemlenmesidir. Diğer yandan karar ağaçlarından LGBM'nin kullanılmasının en büyük nedeni performansının diğer karar ağaçlarına göre çok yüksek olması ve büyük verilerde çalışma anında hızlı sonuçlar elde edilebilmesidir.

Geliştirilen uygulamada programlama dili olarak Python kullanılmıştır. Python, sentaks olarak İngilizce 'ye oldukça yakındır. Konuşma diline çok yakın olması

nedeniyle Python üzerinde kolaylıkla hakimiyet sağlanabilmekte ve öğrenilme kolaylığı sunmaktadır. Aynı zamanda Python esnek ve dinamik yapıya sahip olan bir programlama dilidir. Python'ın kolay öğrenilebilen yapısı, gelişmiş kütüphaneleri ve hızlı çalışması nedeniyle bu çalışmada kullanılması uygun görülmüştür. Gelişen teknolojik dünyada eski programlama dilleriyle satırlarca yazılan kod karşılığında elde edilen çıktı, Python ile birkaç satır kod yazılarak elde edilebilir. Bu sayede hem kodun okunurluğu ve kalitesi artar hem de zamandan tasarruf edilmektedir. Python'un yazılım sektöründe tercih edilmesinin en önemli nedenlerinden bir diğeri ise, farklı işletim sistemleri üzerinde kolaylıkla çalışabilir uyumlu bir yapıda olmasıdır.

Uygulama ortam olarak Anaconda Spyder üzerinde çalıştırılmaktadır. Kullanıcı dostu arayüzü ve renkli grafik çıktıları sağlayabilmesi nedeniyle Anaconda Spyder tercih edilmiştir. Ayrıca Anaconda, Python'un istenilen sürümünün kolaylıkla çalıştırılmasına izin veren esnek bir yapı sağlamaktadır. Modüler yapışı sayesinde Anaconda, birçok yetki problemini aşarak sınırsız bir çalışma ortamı sağlayabilmektedir. Aynı zamanda yüksek işlem performansı sağlamaktadır ve paketler arası uyumun sağlanmasında başarılı sonuçlar elde edildiği gözlemlenmektedir.

D. Yeni Bir Model Oluşturma Gereksinimi

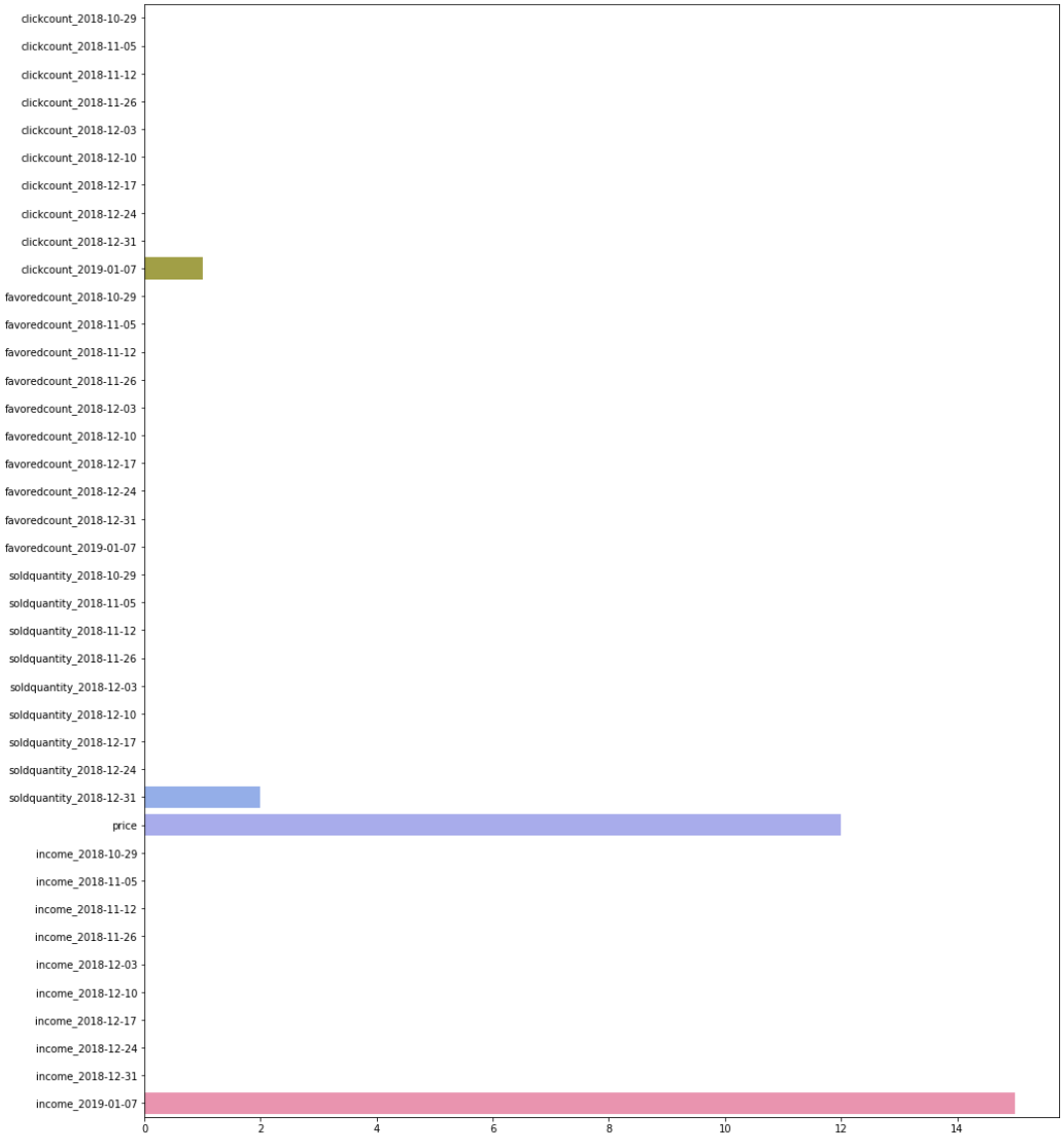
Uygulamada bir Karar Ağacı türü olan LightGBM algoritması kullanılmaktadır. Algoritma ilk denemelerde hiçbir değişikliğe uğratılmadan kullanılmıştır. Şekil 42'de LightGBM algoritması ile eğitim sürecinin tamamlanmasından sonra tahmine etki eden sonuçlar görülmektedir. Algoritmanın genel yapısı üzerinde hiçbir değişiklik yapılmadan kullanılması durumunda;

- Görüntülenme sayısına göre 2019-01-07,
- Satış miktarına göre 2018-12-31,
- Gelir miktarına göre 2019-01-07,
- Fiyatın sonuca olan etkilerinin çok büyük olduğu görülmektedir.

Modelin eğitiminde kullanılan diğer haftaların, sonuç üzerinde hiçbir etkisi olmadığı gözlemlenmektedir. Çizelge 12'de ise algoritmadan alınan satış tahminleri ve gerçekleşen satışlar gösterilmektedir. Modelin tüm haftalar üzerinde bir etkisinin olmaması nedeniyle, model gerçeğe yakın sonuçlar verse bile, modelin veri setine göre eğitilerek daha gerçekçi sonuçlar alınabileceğini söylemek mümkündür. Beklenen

durumda; tüm haftaların az da olsa model üzerinde etkisinin olması gerekmektedir. Bu sonuca uzman görüşü alınarak varılmıştır (EK A).

Makine öğrenmesi yöntemlerinde kullanılan algoritmalar çok esnek bir yapıya sahip olsa da kullanılacak veri setlerine göre bu algoritmalarda geliştirmeler yapmak modelden daha başarılı çıktılar almayı sağlayacaktır. Bir model tasarımı yapılırken, kullanılan verilere ve hangi özelliklerin ağırlığının fazla olacağına karar vermek, hangi özelliklerin kullanılacağını belirlemek model başarısını arttırabilmek adına büyük önem arz etmektedir. Bu nedenle yeni bir model geliştirme ihtiyacı doğmaktadır.



Şekil 40: Veri Setinin LightGBM Algoritmasındaki Ağırlıkları

Çizelge 12: LightGM Algoritması Model Tahmininden Alınan Sonuçlar

Ürün No	Tahmin Edilen Satış Adedi	Gerçekleşen Satış Adedi
1279	5	10
8298	20	31
8859	7	13
15339	6	14
21911	8	18
36559	20	46
100070	2	4
103193	9	18
108909	21	42
118165	9	15

VI. BULGULAR VE SONUÇ

Bu çalışmada Dünya çapında perakende sektöründe önde gelen firmalarından birinin E-Ticaret müşterilerine ait satış verileri kullanılarak satış tahmini yapabilmek adına bir model geliştirilmiştir. İşletmenin müşterilerine ait veriler, veri madenciliği yöntemleri kullanılarak anlamlı hale getirilmiştir.

Söz konusu işletme ile yapılan görüşmeler sonucu satış tahminlerinin geçmiş haftaları veya dönemleri referans alarak, uzman kişiler tarafından sezgisel ve deneyimsel olarak yapıldığı, satış tahmini için herhangi özel bir yöntem kullanılmadığı bilgisine ulaşılmıştır. Bu nedenle oluşturulan modelden elde edilen sonuçları karşılaştırmak için herhangi bir sistem bulunmamaktadır. Bu sorunu aşabilmek adına, modelin doğruluğu tahmin edilen sonuçlar ile gerçekleşen sonuçların karşılaştırılması ile elde edilmiştir. Karşılaştırma sonucunda modelin başarılı sonuçlar verdiği kanısına varılmıştır. Örneklem olarak seçilen 10 adet ürün verisi üzerinde yapılan karşılaştırmalardan elde edilen sonuçlar Çizelge 13'te gösterilmektedir. Gerçekleşen satışlar ve elde edilen tahmin sonuçları karşılaştırıldığında başarılı tahminler yapıldığını söylemek mümkündür.

Çizelge 13: Satış Adedi Sonuçları

Ürün No	Tahmin Edilen Satış Adedi	Gerçekleşen Satış Adedi
1279	8	10
8298	34	31
8859	10	13
15339	10	14
21911	13	18
36559	33	46
100070	3	4
103193	16	18
108909	41	42
118165	15	15

Ek olarak çalışma sonucunda elde edilen verilere göre; müşterilerin ilgili hafta veya takip eden haftalar içerisinde yapmış olduğu ürünleri favoriye alma ve ürünleri görüntüleme sayısı gibi etkileşimler, satın alınma oranları hakkında fikir vermektedir. Müşterilerin ürün ile olan bu tür etkileşimlerinin arasında bir ilişki bulunmaktadır. Bunun yanı sıra söz konusu firmanın E-Ticaret müşterilerinin ürünleri tercih etmelerindeki en önemli özelliğin ürünün fiyatı olduğu sonucuna ulaşılmaktadır.

VII. KAYNAKÇA

KİTAPLAR

HAYKİN, S. (1999), "**Neural Networks: A Comprehensive Foundation**", New Jersey: Prentice Hall Inc.

KARAFAKIOĞLU M. (2012), "**Örnek Olaylarla Satış Yönetimi**", Literatür Yayınları, İstanbul, ss. 9-11.

MAKALELER

ALPAYDIN, E. (2010), "Introduction to Machine Learning", **Second Edition**, The MIT Press, Cambridge

ASİLKAN Ö., Irmak S. (2009), "İkinci El Otomobillerin Gelecekteki Fiyatlarının Yapay Sinir Ağları ile Tahmin Edilmesi", **Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi** Y.2009, C.14, S.2 s.375-391.

BİSHOP, C. M. (2006), "Pattern Recognition and Machine Learning", New York, NY : **Springer**, 2006. - 738 p.

CHANG C., Wang Y., Liu C. (2007), "The development of a weighted evolving fuzzy neural network for PCB sales forecasting", **Elsevier**, Volume 32, Issue 1, January 2007, Pages 86-96

DİMİTOGLOU, G., Adams, J. A., & Jim, C. M. (2012). "Comparison of the C4. 5 and a Naïve Bayes classifier for the prediction of lung cancer survivability". arXiv preprint **arXiv**:1206.1121.

FANTAZZINIA D., Toktamysovab, Z. (2015), "Forecasting German car sales using Google data and multivariate models", **International Journal of Production Economics**, 170, 97- 135.

GUO, Z.X, Wong, W.K., Li, M. (2013), "A multivariate intelligent decision-making model for retail sales forecasting", **Decision Support Systems**, 55, 247–255.

KUMAR, S.A. ve Suresh, N.(2009), "Operations Management", **New Age International (P) Ltd.**, New Delphi.

LEE, W., Chen, C., & Chen, K. (2012), "A comparative study on the forecast of fresh food sales using logistic regression, moving average and BPNN methods", **Journal of Marine Science and Technology**. 20 (2), pp.142–152.

MAKRIDAKIS S., Hibon, M., & Moser, C. (1979). "Accuracy of forecasting: An empirical investigation**". Journal of the Royal Statistical Society**. Series A (General), 97-145

MERIGO J.M., Palacios, M.D., Ribeiro, N. B. (2015), "Aggregation systems for sales forecasting", **Journal of Business Research**, 68, 2299–2304.

NISBET R., Elder, J., Miner, G. (2009). Handbook Of Statistical Analysis And Data Mining Applications, **Elsevier**, Canada.

ÖZER M. (2001), "User Segmentation of Online Music Services Using Fuzzy Clustering", **Omega**, 29 (2), s: 193.

SAVAŞ, Topaloğlu S., Yılmaz N., Mithat (2012), "VeriMadenciliği ve Türkiyedeki Uygulama Örnekleri", **İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi**, Cilt 11, Sayı 21

SHEARER C. (2000), "The Crisp-DM model: The new blueprint for data mining", **Journal of Data Warehousing**, 5 (4): 13-23.

SHUKLA, M. and Jharkharia, S., 2013. "Applicability of ARIMA models in the wholesale vegetable market: an investigation". **Int. J. Inf. Syst. Supply Chain Manag**. 6 (3), pp. 105–119.

SUTTON, C.D. (2005). "Classification and regression trees, bagging and boosting. Handbook Of Statistics", **Data Mining and Data Visualization**, 24, 303-329.

SUN, X., Gauri, D. K., & Webster, S. (2011), "Forecasting for cruise line revenue management". **Journal of Revenue and Pricing Management**, 29 Ocak 2010; doi:10.1057/rpm.2009.55

SUYKENS, J. A. K., Vandewalle, J. (1999), "Least Squares Support Vector Machine Classifiers". **Neural Processing Letters**, 9(3), 293–300.

TÜRK E. (2019), "Yapay Sinir Ağları ile Talep Tahmini Yapma: Beyaz Eşya Üretim Planlaması için YSA Uygulaması", **İstanbul Sabahattin Zaim Üniversitesi Fen Bilimleri Enstitüsü Dergisi,** İstanbul, Cilt:1 Sayı:1

WANG X., Smith-Miles K., Hyndman R. (2009), "Rule induction for forecasting method selection: meta-learning the characteristics of univariate time series", **Neurocomputing** 72, 2581–2594.

WAHHINGTON S., Karlaftis, M., Mannering, F. (2011), "Taşımacılık Veri Analizi için İstatistiksel ve Ekonometrik Yöntemler", **Second Edition**, CRC Press.

ELEKTRONİK KAYNAKLAR

HYNDMAN,R.J.(2009),"ForecastingOverview", http://www.robjhyndman.com/papers/forecastingoverview.pdf (Erişim Tarihi: 11.11.2020)

TEZLER

AKSOY S. Z. (2008), "Kurumsal Kaynak Planlaması Yazılımlarında Talep Tahmin Yöntemleri ve Uygulamaları", İstanbul Üniversitesi Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, İstanbul

ALIZADEH M. (2011), "Yapay Sinir Ağları ile Fiyat Tahmin Analizi", İstanbul Üniversitesi Fen Bilimleri Enstitüsü Yüksek Lisans Tezi, İstanbul.

ALTAN E. (2019), "Genetik Algoritmalar ve Makine Öğrenmesi Yöntemleriyle Görüntü Sınıflandırma", Ege Üniversitesi Fen Bilimleri Enstitüsü Yüksek Lisans Tezi, İzmir

AYDIN B. (2019), "Satış Tahmini İçin Entegre Bulanık Bir Yaklaşım", Galatarasay Üniversitesi Fen Bilimleri Enstitüsü Yüksek Lisans Tezi, İstanbul

ÇATALOLUK H. (2012), "Gerçek Tıbbi Veriler Üzerinde Veri Madenciliği Yöntemlerini Kullanarak Hastalık Teşisi", Bilecik Üniversitesi Fen Bilimleri Enstitüsü Yüksek Lisans Tezi, Bilecik

DEMİRTAŞ D. (2011), "Satış Tahminlerinin Doğruluğu ve Bir Uygulama", Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü Yüksek Lisans Tezi, İstanbul

EKMEKÇİ A.S. (2006), "Endüstriyel Pazarda Satış Tahmin Yöntemlerinin Kullanılabilirliği ve Hazır Beton Sektöründe Bir Uygulama", Marmara Üniversitesi Sosyal Bilimler Enstitüsü Yüksek Lisans Tezi, İstanbul

HACIEFENDİOĞLU Ş. (2012), "Makine Öğrenmesi Yöntemleri ile Glokom Hastalığının Teşhisi", Selçuk Üniversitesi Fen Bilimleri Enstitüsü Yüksek Lisans Tezi, Konya

HAMZAÇELEBİ C., Kutay F. (2004), "Yapay Sinir Ağları ile Türkiye Elektrik Enerjisi Tüketiminin 2010 Yılına Kadar Tahmini", Gazi Üniversitesi Endüstri Mühendisliği Lisans Tezi, Ankara

İŞSEVER T. (2016), "Sales Forecasting in Textile Industry", Marmara Üniversitesi Fen Bilimleri Enstitüsü Yüksek Lisans Tezi, İstanbul

KARACA C., Karacan H. (2016), "Çoklu Regresyon Metoduyla Elektrik Tüketim Talebini Etkileyen Faktörlerin İncelenmesi", Gazi Üniversitesi, Mühendislik Fakültesi Fen Bilimleri Enstitüsü Yüksek Lisans Tezi, Kocaeli

KARATAŞ E.K. (2011), "Yapay Sinir Ağları ile Yazılım Projesi Maliyet Tahmini", İstanbul Üniversitesi Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, İstanbul.

KE G., Meng Q., Finley T., Wang T., Chen W., Ma W., Ye Q., Liu T. (2007), "LightGBM: A Highly Efficient Gradient Boosting Decision Tree" PhD thesis, The University of Waikato, 2007. 9

KOÇTÜRK Y. (2010), "Veri Madenciliğinde Bağlılık", İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, İstanbul.

KURT F. (2018), "Evrişimli Sinir Ağlarında Hiper Parametrelerin Etkisinin İncelenmesi", Hacettepe Üniversitesi Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, Ankara.

OLGUN, S. (2009), "Tedarik Zinciri Yönetiminde Talep Tahmini Yöntemleri ve Yapay Zeka Tabanlı Bir Talep Tahmini Modelinin Uygulanması", İstanbul Üniversitesi Fen Bilimleri Enstitüsü Yüksek Lisans Tezi, İstanbul

ÖZTEMİZ F. (2017), "Apriori algoritması ile müşteri bazlı market sepet analizi ve ürün satış tahmini", İnönü Üniversitesi Fen Bilimleri Enstitüsü Yüksek Lisans Tezi, Malatya

SARIPĞLU F. (2019), "Farklı kullanıcı-ürün etkileşim türlerini kullanarak özyineli sinir ağları ile ürün ve satış tahminlemesi", İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü Yüksek Lisans Tezi, İstanbul

SERTTAŞ, S. Z. (2011), "Türkiye'de Perakende Sektöründe Talebi Etkileyen Etmenler ve Yapay Sinir Ağlarıyla Talep Tahmini Uygulaması", Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü, İstanbul

TOSUN T. (2006), "Veri Madenciliği Teknikleriyle Kredi Kartlarında Müşteri Kaybetme Analizi", İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, İstanbul.

UZUNKAYA E. (2019), "Gıda Perakendesinin Organize Gıda Perakendesine Geçişindeki Aşamalarda Müşteri Profili", İstanbul Gedik Üniversitesi Sosyal Bilimler Enstitüsü Yüksek Lisans Tezi, İstanbul

YEĞEN N. (2020), "Perakende sektöründe veri madenciliği ile satış tahmini", Bahçeşehir Üniversitesi Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, İstanbul.

YILMAZ K. (2020), “Zaman Serisi Verileri Kullanılarak Hazır Yemek Sektöründe Satış Tahmini”, Bahçeşehir Üniversitesi Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, İstanbul.

DİĞER KAYNAKLAR

ŞEKER A., Amasyalı M. F. (2016), "Topluluk Algoritması Destekli Yarı-eğiticili Öğrenme”, Uluslararası Bilgisayar Bilimleri ve Mühendisliği Konferansı, vol.1, Tekirdağ, Türkiye, pp.1-4.

EKLER

EK A: Uzman Görüşü

EK A

Yelda Arslan'ın "LightGBM Algortiması ile Yeni Bir Satış Tahmin Modelinin Oluşturulması ve Perakende Sektörüne Uygulanması" adlı yüksek lisans tezi için yapmış olduğu satış tahmin uygulamasında kullanılan veriler uygulama süresince incelenmiştir. Veri setinde girdi olarak kullanılan; ürün numarası, cinsiyet, renk, kategori numarası, marka numarası, alt kategori numarası, fiyat, satış tarihi, satış adedi, stok, görüntülenme sayısı ve favori sayısının, gelecekte yapılacak satış adedinin tahmin edilmesi için yeterli olduğu bilgisine geçmişe dönük satış hareketleri analizlerinden ve tecrübelerimizden ulaşabilmekteyiz.

Gelecekteki bir haftanın ürün özelinde satış adedi tahmini söz konusu olduğunda, geçmişe dönük altı aylık satış verisinin incelenmesi yeterli olmaktadır. Uygulamaya konu olan geçmiş dönem satış verilerinin yapılan tahmin üzerinde etkileri olması beklenmektedir. Bu kanıya geçmişteki ürün satış verileri incelenerek varılmıştır.

E-Ticaret Veri Analiz Uzmanı

Emre Çakmak

ÖZGEÇMİŞ

Ad-Soyad **:** Yelda Arslan

E-posta **:** arslan.yelda@gmail.com

ÖĞRENİM DURUMU

Lisans **:** 2015, Trakya Üniversitesi, Bilgisayar Teknolojisi ve Bilişim Sistemleri

Yüksek Lisans **:** 2020, İstanbul Aydın Üniversitesi, Bilgisayar Mühendisliği

MESLEKİ DENEYİM

LC Waikiki

Yazılım Geliştirme Uzmanı — Ağustos 2017 – Halen

GNS Yazılım

Yazılım Geliştirme Uzmanı — Aralık 2016 – Ağustos 2017

Protel Bilgisayar A.Ş

Yazılım Eğitim ve Destek Uzman Yardımcısı — Eylül 2015 – Mayıs 2016