# *MAKİNE ÖĞRENMESİ ALGORİTMALARI İLE SATIŞ TAHMİNİ

Emine Nur NACAR[1*], Babek ERDEBİLLİ (B.D.ROUYENDEGH)[2]

[1]Ankara Yıldırım Beyazıt Üniversitesi, Mühendislik ve Doğa Bilimleri Fakültesi, Endüstri Mühendisliği Bölümü, Ankara
ORCID No: https://orcid.org/0000-0003-3785-1928
[2]Ankara Yıldırım Beyazıt Üniversitesi, Mühendislik ve Doğa Bilimleri Fakültesi, Endüstri Mühendisliği Bölümü, Ankara
ORCID No: https://orcid.org/0000-0001-8860-3903

**Anahtar Kelimeler**

*Makine öğrenmesi, satış tahmini, gözetimli öğrenme, regresyon*

**Öz**

*Günümüz dijital dünyasında satın alma gittikçe arttığından veriler çok büyük boyutlara ulaşmıştır. Endüstrinin getirdiği kavramlardan en belirgini ise boyutluluk laneti olmuştur. Bu sebeple işletmeler satın alma kararlarını alırken büyük zorluk yaşamaktadır. Uzun ya da kısa vadede satış tahmininin doğru yapılamaması müşteri memnuniyetsizliği, para kaybı, ham madde ihtiyacı gibi birçok soruna yol açacaktır. Tedarik zinciri elemanlarından üretici, perakendeci, tedarikçi ve müşteriye kadar birçok taraf yanlış ya da eksik satış tahmininden zarar görebilir. Yapay zekâ çağının getirdiği yeniliklerden olan makine öğrenmesi de birçok mühendislik uygulamasının getirdiği sorunlara olduğu gibi satış tahmini problemlerine de hızlı şekilde cevap verebilecek bir alandır. Bu çalışmada uçtan uca bir makine öğrenmesi proje süreci ele alınmıştır. Herhangi bir makine öğrenmesi projesinin adımları ve veriye yaklaşım boyutu tanıtılmıştır. Uygulama bölümünde makine öğrenmesi algoritmalarından doğrusal regresyon, Ridge, Lasso, Elastic Net, K-en yakın komşu ve Rastgele Orman algoritmaları kullanılarak gerçek veri seti için bir satış tahmin modeli geliştirilmiştir. Geliştirilen modelde en düşük hatayı veren algoritma Rastgele Orman algoritması olmuştur.*

## SALES FORECASTING VIA MACHINE LEARNING ALGORITHMS

**Keywords**

*Machine learning, sales forecasting, supervised learning, regression*

**Abstract**

*As purchases increase in today's digital world, data has reached enormous dimensions. The most prominent of the concepts brought by the industry has been the multidimensionality curse. For this reason, businesses have great difficulty in making purchasing decisions. Failure to make a correct sales forecast in the long or short term will cause many problems such as customer dissatisfaction, loss of money, and the need for raw materials. Many parties, from supply chain elements to manufacturers, retailers, suppliers and customers, may suffer from wrong or incomplete sales forecast. Machine learning, which is one of the innovations brought by the age of artificial intelligence, is an area that can quickly respond to sales prediction problems as well as the problems brought by many engineering applications. In this study, an end-to-end machine learning project process is discussed. The steps of any machine learning project and the approach to data dimension are introduced. In the application part, a sales forecast model has been developed for the real data set by using machine learning algorithms such as linear regression, Ridge, Lasso, ElasticNet, K-nearest neighbor and Random Forest algorithms. The algorithm with the lowest error in the developed model was the Random Forest algorithm.*

| | | | |
|---|---|---|---|
| Araştırma Makalesi | | Research Article | |
| Başvuru Tarihi | : 15.10.2020 | Submission Date | : 15.10.2020 |
| Kabul Tarihi | : 01.07.2021 | Accepted Date | : 01.07.2021 |

*Sorumlu yazar; e-posta : ennacar@ybu.edu.tr
*Bu çalışma, Emine Nur Nacar'ın yüksek lisans tezinden üretilmiştir.

## 1.Giriş

Üretimin dahi dijitalleştiği günümüz dünyasında, tüketim gittikçe artan bir ivme göstermektedir. İnsanların artık tek tıkla alışveriş yapabiliyor olmaları bu durumun başlıca sebebidir. Satın almanın giderek arttığı bu çağda, satışları tahmin edebilmek ve ona göre stok tutmak ya da üretim planlaması yapmak büyük önem kazanmıştır. Tahmin yapılmadan işletmelerin ne gelirleri ne de giderleri öngörülebilir. İşletmenin üretim planlama, satın alma, satış ve pazarlama gibi birçok departmanı bu tahminler üzerinden analiz yapmaktadır. Uzun ya da kısa vadede satış tahmininin doğru yapılamaması müşteri memnuniyetsizliği, para kaybı, ham madde ihtiyacı gibi birçok soruna yol açacaktır. Tahmin yapılmadan geleceğe dönük stratejik kararlar alınamayacağından işletmenin plan ve politikalar oluşturamamasına sebep olur. Bu durum sadece işletmenin kendisini etkilemekle kalmayıp tedarik zinciri elemanlarından üretici, perakendeci, tedarikçi ve müşteriye kadar birçok tarafın zarar görmesine sebep olur. Durum bireysel olmaktan çıkar ve birçok kişi, yapılamayan ya da hatalı yapılan tahminden muzdarip olur. İyi bir satış tahmini çalışması sonucunda ise sadece şirket kâr etmekle kalmaz, tedarik zincirinin tüm elemanlarında verimlilik artışına rastlanabilir. Böylelikle işletmeler sektörde devamlılığını sürdürmekle kalmayıp tedarik zincirinin tüm elemanları için kâr marjı sunabileceklerdir.

Hiçbir yöntem kullanılmadan büyük çaplı satışları tahmin edebilmek, işletmeler için imkânsız hale gelmiştir. Klasik yöntemler ise bu büyük veri yığınlarının oluşturduğu satışlara cevap verememekte ya da esneklikleri ve özel durumları hesaplayamamaktadır. Bu sebeple literatüre yeni kazandırılan yapay zekâ bileşenleri benzer sonuçlara cevap vermiştir. Ne kadar yapay zekâ yöntemlerinin robotik ya da telekomünikasyon gibi alanlarda kullanıldığı izlenimi olsa da zamanla bu önyargı değişecek ve verinin olduğu her yerde kullanılmaya başlanacaktır. Yapay zekâ çağının getirdiği yeniliklerden olan makine öğrenmesi de birçok mühendislik uygulamasının getirdiği sorunlara hızlı şekilde cevap verebilecek bir alandır. Bazı uygulamalara rastlansa da henüz çok yeni bir alan olup literatürde birçok varsayım bulunmaktadır. Çoğu işletme, veri biliminin benzeri sorunlara geleneksel yöntemlere kıyasla kısa sürede ve doğrulukta cevap verdiğini düşünmediğinden, bu çalışmanın işletmelere bir kılavuz niteliğinde olması beklenmektedir.

Bu çalışmada lineer regresyon, Ridge, Lasso, ElasticNet, K-En Yakın Komşu ve Rastgele Orman yöntemleri kullanılarak satış tahmini için model oluşturulmuş ve tahmin yöntemlerinde makine öğrenmesi algoritmalarının kullanılmasının gerekliliği üzerinde durulmuştur.

Çalışmanın ikinci ve üçüncü bölümünde sırasıyla satış tahmini ve makine öğrenmesi konuları ile ilgili literatür araştırması yapılmıştır. İzlenen metotun işlendiği üçüncü bölümde bir makine öğrenmesi modeli oluşturulurken izlenmesi ve dikkat edilmesi gereken adımlar keşifçi veri analizinden model optimizasyonuna kadar verilmiştir. Beşinci bölümde gerçek veri seti ile yapılan örnek bir satış tahmini çalışması yapılmış, sonuçlar altıncı bölümde işlenmiştir.

## 2. Makine Öğrenmesi

Yapay zekâ (AI), bilgisayar bilimlerinin bir alt dalı olan ve bilgisayarların akıllıca davranmasını sağlayarak akademi ve endüstride kendine yer bulan bir araştırma alanıdır (Nilsson, 2014). Bu macera ve yapay zekaya yönelik eğilim 2016 yılında, Google DeepMind'a ait bir yazılım şirketinin, DeepMind tarafından üretilen bir yazılım olan AlphaGo ile 9-dan rakibini avantajsız bir şekilde yenmesiyle başlamıştır (Siau ve Wang, 2018). Yapay zekâ, makine öğrenmesi ve derin öğrenme birbiriyle çok karıştırılan kavramlardır. Yapay zekâ ile makine öğrenmesinin ilişkisi Şekil 1'de verilmiştir. Yapay zekâ, programların insanlar gibi öğrenebilmesi ve davranabilmesi iken makine öğrenmesi ise aynı amaç için yazılan algoritmalardır.


Şekil 1. Yapay zekâ bileşenleri

Biz bilgisayarlara sahip olmadan çok önce, insanlar verilerde kalıplar bulmaya çalışmışlardır. Bu tür kalıpları ortaya çıkarmak için astronomik verilerin analizi, doğrusal denklemleri çözme yöntemleri (Newton-Gauss), gradyan iniş yoluyla optimum öğrenme (Newton), polinom enterpolasyonu (Lagrange) ve en küçük kareler uydurma (Laplace) gibi matematiksel tekniklerin ortaya çıkmasına neden olmuştur. On dokuzuncu ve yirminci yüzyılın başlarında, içerdiği kalıpları ortaya çıkarmak için verileri analiz etmek için geniş bir matematiksel yöntem yelpazesi ortaya çıktı. Yirminci yüzyılın ortalarında dijital bilgisayarların yapımı, veri analizi tekniklerinin otomasyonuna izin vermiştir. Son yarım yüzyılda, teknolojinin hızlı ilerlemesi ile sektörlerin ihtiyacı giderek arttığından farklı veri analizi teknikleri ve karmaşık yöntemler geliştirilmiştir. Aynı zaman dilimi içinde, dijital bilgisayarların gelişimi ve hızlı ilerlemesi, yeni makine öğrenimi yöntemlerini ortaya çıkarmıştır (Biamonte, Wittek, Pancotti, Rebentrost, Wiebe ve Lloyd, 2017).

Makine öğrenmesi, bilgisayarların insanlara benzer şekilde öğrenmesini sağlamak için çeşitli algoritma ve tekniklerin geliştirilmesi için çalışan bilimsel bir çalışma alanıdır. Bilgisayarların deneyim yoluyla otomatik olarak gelişmesini sağlayarak en hızlı büyüyen alanlardan biri olmuştur (Jordan ve Mitchell, 2015). Makine öğrenmesi ile verilerden istatistiksel çıkarımlar veya tahminler yapılabilir. Bu şekilde makineler daha doğru tahminler yapabilirler. Makinelerin öğrenme süreci, geçmiş bilgilere dayanarak tahminlerde bulunmaktır.

Makine öğrenmesi ilk kez 1959'da Arthur Samuel tarafından bir dama oyununu analiz etmek için kullanılmıştır. Arthur Samuel, makine öğrenmesini, bilgisayarlara açıkça programlanmadan öğrenme yeteneği veren araştırma alanı olarak tanımlamıştır (Samuel, 1959). 1998 yılında Tom Mitchell tarafından yapılan tanıma göre, "Bir bilgisayar programının, T ile ölçülen performansı P ile ölçüldüğünde iyileşirse, bazı T görevlerine göre E deneyiminden ve bazı performans ölçülerinden P öğrendiği söylenir." (Mitchell, 1997)

Makine öğrenmesi; olasılık, istatistik ve optimizasyon gibi matematiğin alt alanları üzerine inşa edilmiştir. Model oluşturulur ve makinenin geçmiş verileri öğrenerek geleceği tahmin etmesi beklenir. Bu bakımdan makine öğrenmesi, diğer yazılım mühendisliği yöntemlerinden farklıdır. Çünkü bu geleneksel yöntemlerde yazılım geliştiricinin veri setine bakarak bazı kalıplar oluşturması beklenir. Bu prosedürler oldukça önyargılıdır ve gerçekçi değildir (Grigorev, 2020).

Makine öğrenmesi gittikçe yayılan bir alan olmuştur ve olmaya da devam edecektir. Bunun birçok nedeni vardır. Birincisi, sembolik makine öğrenmesi, hesaplamalı öğrenme teorisi, sinir ağları, istatistikler ve örüntü tanıma alanındaki ayrı araştırma toplulukları birbirlerini keşfetmişler ve birlikte çalışmaya başlamışlardır. İkinci olarak, makine öğrenmesi teknikleri, veri tabanlarında bilgi keşfi, dil işleme, robot kontrolü ve kombine optimizasyon gibi yeni problem türlerine ve ayrıca konuşma tanıma, yüz tanıma, el yazısı tanıma, tıbbi gibi daha geleneksel problemlere uygulanabilmektedir (Dietterich, 1997).

Makine öğrenmesi türleri genel olarak gözetimli ve gözetimsiz olmak üzere ikiye ayrılır. Gözetimli öğrenme, önceden atanmış doğru sınıflandırma ile veri kaynağından bir veri örneğini eğitmeye dayanır. Gözetimsiz öğrenmede ise kendi kendini organize eden sinir ağları, etiketlenmemiş giriş verilerindeki gizli kalıpları belirlemek için gözetimsiz öğrenme algoritmasını kullanmayı öğrenir. Gözetimsizlik, çözümü değerlendirmek için bir hata sinyali sağlamadan bilgiyi öğrenme ve organize etme yeteneğini ifade eder (Sathya ve Abraham, 2013). Gözetimli öğrenme türü ise sınıflandırma ve regresyon olarak ikiye ayrılır. Veri setinin içeriğine göre çeşitli algoritmaları kullanmak mümkündür.

## 3. Satış Tahmininde Makine Öğrenmesi Uygulamaları

Satış tahmini konusu günümüze dek çoğunlukla geleneksel zaman serisi yöntemleri ile çözülen ve uzun zamandır literatürde var olan bir konudur. Temel olarak satış tahminleri geçmiş verilerden faydalanarak gelecekteki satışları öngörmeyi hedefler. Satış tahmin modelleri iki türe ayrılabilir: (i) zaman serisi yöntemleri ve (ii) derin öğrenme ve makine öğrenmesi algoritmalarını içeren yapay zeka yöntemleri (Şener, 2019). Yapay zeka yaklaşımları geliştirilmeden önce, AR, MA, ARIMA, VARMA ve Holt-Winters gibi yöntemler kullanılmıştır. Bu yöntemler Python gibi programlama dillerinde uygulanabilir. Ancak belirli ihtiyaçları karşılayamazlar. Armstrong çalışmasında tahmin modellerin birlikte kullanıldığında daha iyi sonuçlar verdiğini gözlemlemiştir (Armstrong, 1989). Papacharalampous ve arkadaşları tarafından yapılan başka bir çalışmada, zaman serisi

yöntemlerinde gecikmeli değişken seçimi, hiperparametre işleme ve makine öğrenmesi ile klasik algoritmalar arasında yapılan karşılaştırmaya ilişkin ampirik kanıtlar sunulmuştur. Yunanistan'daki sıcaklık verileri üzerine yapılan bu çalışmada geleneksel zaman serisi yöntemleri ile makine öğrenmesi algoritmaları karşılaştırılmıştır. Çalışmada, dört geleneksel yöntem ve makine öğrenmesi yöntemleri Destek Vektör Regresyonu ve Sinir Ağları kullanılarak bir karşılaştırma yapılmıştır. Sonuçlar, klasik algoritmaların ve makine öğrenmesi algoritmalarının belirli sınırlar altında eşdeğer performans gösterdiğine işaret etmiştir. Sonuçlar veri setinden veri setine değişebilir, ancak klasik algoritmaların herhangi bir üstünlüğü yoktur (Papacharalampous, Tyralis ve Koutsoyiannis, 2018).

Pavlyshenko'ya göre, klasik zaman serisi yöntemlerinin birçok sınırlaması vardır. Bu nedenle satış tahminleri, zaman serisi yöntemlerinden çok denetimli öğrenme yöntemlerinden biri olan regresyon problemlerinin konusudur. Regresyon yöntemleri, zaman serisi yöntemlerinden daha iyi sonuçlar vermektedir. Çünkü regresyon modelleri geçmiş verilere dayalı genel bir fonksiyon oluşturmayı hedefler. Yazara göre, zaman serisi modellerinin birçok kısıtlaması bulunmaktadır. Eksik ya da aykırı değer problemleri, dış faktörler, yeni ürünler. Satış verilerinde çok fazla eksik veya aykırı değer olabilir. Bu nedenle keşifçi veri analizi gereklidir. Böylelikle daha doğru sonuçlara ulaşmak mümkün olacaktır. Makine öğrenmesi algoritmaları bu gibi durumlara klasik yöntemlere kıyasla yanıt verir. Satışları etkileyen birçok dış faktör vardır. Bunlar, veri analistinin tecrübesiyle gerçekleştirilebilir. Aksi takdirde klasik yöntemler sadece satış rakamlarına göre yapılacaktır. İnsan hayatını etkileyen pek çok dış faktörün ele alınmadığı klasik yöntemlerde dış etkenler sadece yorumlanabilir ancak makine öğrenmesi algoritmaları ile bunu verilere yansıtmak mümkündür. Pazara yeni bir ürün çıkması durumunda geçmiş verilere sahip olunması mümkün değildir. Ancak makine öğrenmesi algoritmaları ile pazar içindeki başka bir ürünün verileri kullanılarak bir satış tahmini yapmak mümkündür (Pavlyshenko, 2019).

XGBoost yöntemi, Avrupa eczane perakende şirketinin olası satışlarını tahmin etmek için Rastgele Orman ve Doğrusal Regresyon yöntemleriyle karşılaştırılmıştır. Sonuca göre XGBoost yönteminin diğerlerine göre çok daha başarılı bir şekilde çalıştığı görülmüştür. İleride yapılacak çalışmalar için, satış tahmininin yanı sıra şu 6 alanda da aynı eylemlerin gerçekleştirilebileceği önerilmiştir: reklam, öneriler, talep tahmini, müşteri bazlı fiyatlandırma, tatil / uzatılmış satış planlaması ve ürün sınıflandırması (Jain, Menon ve Chandra, 2015).

2018 yılında hava koşullarının etkileri dikkate alınarak Rincon-Patino ve diğ. tarafından yapılan bir çalışmada tarım sektörüne ilişkin veriler makine öğrenimi teknikleriyle analiz edilmiştir. Çalışmada doğrusal regresyon, çok katmanlı algılayıcı, destek vektör makineleri ve çok değişkenli regresyon tahmin modelini içeren dört makine öğrenmesi tekniği incelenmiştir. Avokado satışlarının incelendiği bu çalışmada, en iyi sonuçları destek vektör makineleri ve çok değişkenli regresyon modellerinin verdiği görülmüştür (Rincon-Patino, Lasso ve Corrales, 2018).

Microsoft Azure Machine Learning Studio platformunda Kaggle'dan Walmart satış verilerini kullanan satış tahmini çalışmasında, klasik zaman serisi yöntemleri makine öğrenmesi algoritmalarıyla karşılaştırılmıştır. Doğrusal, Bayes, Sinir Ağları, Karar Ağacı ve Artırılmış Karar Ağacı regresyonları kullanılmıştır. Doğrulama fonksiyonu olarak ortalama kare hata ve kök ortalama kare hata tercih edilmiştir. Karar Ağacı Regresyon modeli başlangıçta en düşük hatayı verirken, model tuning uygulandıktan sonra Artırılmış Karar Ağacı regresyonunun en düşük hatayı verdiği görülmüştür. Bu sonuç ise regresyon modellerinin klasik zaman serisi modellerinden daha iyi çalıştığının bir başka kanıtıdır (Catal, Ece, Arslan ve Akbulut, 2019).

R programı ve Jupyter Notebook kullanılarak yapılan Pavlyshenko'nun çalışmasında Rossmann mağazalarının satış tahminleri incelenmiştir. Çalışmada Lasso, Sinir Ağları, Rastgele Orman ve ExtraTree modelleri incelenmiştir. Satış tahminlerinde regresyon modellerinin zaman serisi yöntemlerinden çok kullanılması gerektiği vurgulanmıştır (Pavlyshenko, 2019).

Weng, Liu ve Xiao tarafından tedarik zinciri için satış tahmini çalışması için LightGBM ve LSTM yöntemlerinden oluşan birleşik bir model önerilmiştir. Vaka çalışması için Kaggle'da Corporación Favorita Market Sale ve Rossman mağaza satışları ve Jollychic E-ticaret Platformunda tedarik zinciri veri setleri tercih edilmiştir. Model için değişken seçiminin çok önemli olduğu vurgulanmıştır. Veri setini incelemek için veri görselleştirme tekniklerinden biri olan dağılım

grafiği tercih edilmiştir. Değişken mühendisliği bölümünde, verilerin satış verisi olup olmamasına göre eksik veriler 0 veya -1 ile doldurulmuştur. Alakasız değişkenler veri setinden çıkarılmış ve verilerde anormal bir eğilim olup olmadığı analiz edilmiştir. Değerlendirme için Normalleştirilmiş Ağırlıklı Kök Ortalama Kare Logaritmik Hata (NWRMSLE) kullanılmıştır. Ridge, Destek Vektör Makineleri, Rastgele Orman, XGBoost, LightGBM ve LSTM yöntemleri kullanılmış ve birleşik modelin doğruluğunu değerlendirmek için karşılaştırılmıştır. Sonuçlar, kombine modelin üç örnek veri seti için verimli bir şekilde çalıştığını göstermiştir (Weng ve diğ., 2019).

Helmini ve arkadaşlarına göre ARIMA gibi istatistiksel tahmin modelleri yaygın olarak kullanılmasına rağmen doğrusal olmayan problemleri analiz etmek için uygun değildir. Doğruluğu diğerlerinden daha yüksek olduğu için bu modelleri analiz etmek için makine öğrenmesi yöntemleri kullanılmalıdır. Özgün bir LSTM modeli yazarlar tarafından geliştirilmiştir. Modele hiperparametre optimizasyonu yapılmıştır. Geliştirilen bu model XGBoost ve Rastgele Orman ile karşılaştırılmıştır. Modeli uygulamak için Kaggle'daki Rossmann mağaza satış veri seti kullanılmıştır. RMSE ve MAE, hata fonksiyonları olarak kullanılır. İlk model, geliştirilen model ve diğer makine öğrenimi yöntemleri birbirleriyle karşılaştırılmıştır. Geliştirilmiş LSTM modelinin doğruluğunun diğerlerinden çok daha iyi olduğu sonucuna varılmıştır (Helmini, Jihan, Jayasinghe ve Perera, 2019).

Verstraete ve diğ. Lasso yöntemi kullanarak bir taktik satış tahmini çalışması gerçekleştirmiştir. Tedarik zincirinin tüm unsurları için taktiksel satış tahmininin gerekli olduğu belirtilmiştir. Yazarlara göre, geleneksel satış tahmin yöntemleri bazı makroekonomik faktörlere cevap veremez. Değişiklikler öngörülemediğinden, uzman görüşleri veya çalışanlar tarafından yapılan birtakım varsayımlara başvurulur. Bu gibi senaryolar oldukça pahalı ve analitik olarak kanıtlanamayan yaklaşımlardır. Bu nedenle, Lasso yöntemi makroekonomik değişiklikleri tahmin etmek için kullanılmıştır. Kullanılan veri seti, otomotiv sektöründeki bir firmadan alınan aylık veri setidir ve 62 aylık verileri içermektedir. Ortalama mutlak yüzde hata kullanılarak modelin hatası hesaplanmıştır. Bu uzun dönemli tahmin çalışması sonucunda, klasik tahmin yöntemlerine göre hatanın %54,5 oranında azaldığı görülmüştür (Verstraete, Aghezzaf ve Desmet, 2020).

GBM, Rastgele Orman ve ElasticNet yöntemlerini karşılaştıran Antipov ve Pokryshevskaya, tüm satış tahminlerinde kullanılması öngörülen genel bir çalışma yapılmıştır. Yorumlanabilir makine öğrenmesi yöntemleri ile eyleme geçirilebilir tahminler elde edilmiştir. Çalışmada kullanılan veri seti 156 haftalık olup bir perakende analitik şirketi tarafından sağlanmıştır (Antipov ve Pokryshevskaya, 2020).

## 4. Yöntem

Yöntem bölümünde keşifçi veri analizinden başlanarak model optimizasyona kadar uçtan uca makine öğrenmesi model kurma ve veriyi analiz etme süreci ifade edilmiştir.

### 4.1 Keşifçi Veri Analizi

İlk olarak 1977'de Tukey tarafından tanıtılan keşif veri analizi (KVA), veri setini anlamak için gerekli ilk adımdır (Tukey, 1977). KVA sayesinde insanların verilere bakış açısı değişti. Birçok anlamsız sayıdan analiz yapmak daha kolay hale geldi (Velleman ve Hoaglin, 1981).Veri setindeki anormallikler, aykırı değerler ve dağılım böylece analiz edilir (Komorowski, Marshall, Salciccioli ve Crutain, 2016).

Model KVA olmadan yapılırsa hata artar. Çünkü veri setinde bazı bozulmalar olabilir. Bunu önceden analiz etmek gerekmektedir. Verileri anlamak için öncelikle veri okuryazarlığı gereklidir. Bu aynı zamanda günlük hayatta ihtiyaç duyulan bir özelliktir. Veri setindeki değişkenleri tanımak için değişken türleri, veri setinin dağılımı ve eğilimi incelenir. Bunları grafik olarak veya bazı teknik yöntemlerle yapmak mümkündür. Veri seti ile ilgili genel yorumlar kutu grafiği, dağılım grafiği ve ısı haritası gibi grafikler kullanılarak yapılır.

Gerçek hayattan alınan bir veri seti çoğu zaman istenen özelliklere sahip değildir. Veriler düzgün tutulmadığı veya model odaklı olmadığı için KVA çok önemlidir. Modele başlamadan önce veri ön işleme bölümünde önemli veya anlamsız sorunlar olup olmadığı kontrol edilmelidir. Bu işlemler veri temizleme, veri standardizasyonu, veri indirgeme ve değişken dönüşümü ile yapılır.

### 4.1.1 Veri Ön İşleme

Makine öğrenmesi modelinin amacı genelleştirilebilir yapıları ortaya çıkarmaktır. Bu şekilde belirli olaylar gözlemlendiğinde belirli tahmin sonuçları elde edilir. Verilerin kalitesi çok önemlidir. Elimizdeki veriler kötüyse, hangi makine öğrenmesi aracını kullanırsak kullanalım sonucumuz işe yaramayacaktır. Veriler ne kadar iyi olursa, model o kadar iyi olur. Tablo 1'de KVA süreçleri verilmiştir.

Tablo 1
KVA Süreçleri

| Veri Temizleme | Veri Standardizasyonu | Veri İndirgeme | Değişken Dönüşümü |
|---|---|---|---|
| Aykırı veri analizi | Normalizasyon | Gözlem sayısını indirgeme | Sayısal değişken dönüşümü |
| Eksik veri analizi | Standardizasyon | Değişken sayısını indirgeme | Kategorik değişken dönüşümü |
| Gürültülü veri | Logaritmik Dönüşüm | | |

### 4.1.2 Aykırı Veri Analizi

Verilerdeki genel eğilimden sapan veya diğer gözlemlerden oldukça farklı olan gözlemlere aykırı değer denir. Bir veri kümesindeki gruplar genellikle veri kümesinin ortasında toplanır. Kuyruklar genellikle bir fark yaratan noktadır ve veri setinde düzeltilmesi gerekir. Aykırı değerler burada yakalanır (Velleman ve Hoaglin, 1981).

Aykırı değer, veri setinde karşılaşılan en ciddi sorunlardan biridir. Genelleştirilebilirlik kaygılarıyla oluşturulan kural kümelerini veya işlevleri yanlış yönlendirir ve yanlılığa neden olur. Bu nedenle aykırı değerler çeşitli tekniklerle düzeltilmeye çalışılır.

Aykırı değerin belirlenmesi, veri seti hakkında genel bilgi gerektirir. Tarafsız modeller oluşturmak, incelenen veriler hakkında sektör bilgisi gerektirir. Kurulan model genelleştirilebilir olmalıdır. Aksi takdirde, model sadece verilen veri seti için iyi sonuçlar verecektir. Bu nedenle, seyrek görülen senaryolar ve genel olarak uymayan yapılar çalışmanın dışında tutulmalıdır.

### 4.1.3 Eksik Veri Analizi

Veri analizinde sık karşılaşılan durumlardan biri olan eksik veriler, incelenen veri setinde gözlemlerin olmamasıdır. Genellikle "NA" olarak ifade edilir. Bu eksik veriler silinebilir veya çeşitli tekniklerle doldurulabilir. Her iki durum da bazı sorunlara neden olabilir. Bu nedenle veri setinin iyi analiz edilmesi gerekmektedir.

Eksik gözlemleri doğrudan çalışmadan çıkarmak, analizin güvenilirliğini azaltabilir. Bu nedenle, eksik değerler rastgele ortaya çıkarsa, eksik değerleri olan gözlemler silinir. Aksi takdirde silme yanlışlıklara neden olabilir. Veri setindeki eksikliğin yapısal bir eksiklik olup olmadığı bilinmelidir. "NA" her zaman eksik değer anlamına gelmez. İlgili değişkenin ölçülmediği veya 0 olduğu da bulunabilir.

Birçok değişken içeren bir veri kümesindeki birkaç "NA" değerinin tüm satırlarını silmek mantıksız olacaktır. Örneğin 50 satır ve 100 sütunlu bir veri kümesi analiz edilir. 1 "NA" değeri için 49 değişkenin bilgisinin silinmesi büyük kayıplara neden olacaktır. Bu nedenle, eksik veriler silinmeden önce bu durumlar dikkate alınmalıdır.

Eksik verilerde ilk adım, eksik veri türünün ne olduğunu belirlemektir. Çünkü kullanılacak yöntem eksik verilerin türüne göre belirlenir (Rubin, 1976):

Tamamen rastgele eksik (MCAR): Tamamen rastgele olan, diğer değişkenlerden veya yapısal bir sorundan kaynaklanmayan gözlemler.

Rastgele kayıp (MAR): Diğer değişkenlere bağlı olarak ortaya çıkan eksiklik türüdür.

Eksik rastgele değil (MNAR): Yapısal sorunlar nedeniyle oluşan eksiklik türüdür. Diğer değişkenlerden veya mantıksal nedenlerden dolayı ortaya çıkmış olabilir.

### 4.2. Makine Öğrenmesi Algoritmaları

Bu bölümde makine öğrenmesi algoritmalarından doğrusal regresyon, Ridge, Lasso, ElasticNet, Rastgele Orman ve K-En Yakın Komşu algoritmaları ele alınmıştır.

#### 4.2.1 Doğrusal Regresyon

Basit doğrusal regresyonun temel amacı, bağımlı ve bağımsız değişken arasındaki ilişkiyi ifade eden doğrusal fonksiyonu bulmaktır. Bu, hata kareler toplamını en aza indirerek yapılır. Basit doğrusal regresyon, makine öğrenmesi modellerinin temelini oluşturduğu için çok önemlidir. Çok değişkenli doğrusal regresyonun temel amacı ise basit doğrusal regresyonda olduğu gibi -bağımlı ve bağımsız değişkenler arasındaki ilişkiyi ifade eden fonksiyonu bulmaktır. Bağımlı değişken, bağımlı değişkeni etkilediği belirlenen bağımsız değişkenler yardımıyla bulunur. Bağımlı değişkeni etkileyen bağımsız değişkenlerin ne ölçüde veya ne şekilde etkilediği belirlenir ve aralarındaki ilişki tanımlanır. Örneğin, bağımlı değişken olumsuz ya da olumlu etkilenebilir. Bu denklemler, Monte Carlo, en küçük kareler gibi çeşitli yöntemlerle çözülebilir (Seber ve Lee, 2012).

#### 4.2.2 Ridge Regresyon

Hata kareler toplamını minimize eden katsayılar, bu katsayılara uygulanan ceza ile bulunur (Hoerl ve Kennard, 1970). $L_2$ düzenlenme durumunu ifade ederken, λ ayar parametresidir. Bazı kaynaklarda büzülme yöntemi olarak anılır (Gokpinar, Ebegil ve Gokpinar, 2017). Cezayı kontrol etme yeteneği sağlar ve λ bazı şekillerde optimize edilebilir. Eşitlik (1)'de Ridge regresyon ifade edilmiştir.

$$SSE_{L_2} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{P} \beta_j^2 \qquad (1)$$

Ridge regresyonunun özellikleri;

Aşırı öğrenmeye karşı dayanıklıdır.

Önyargılıdır, ancak varyansı küçüktür. Bazen önyargılı modeller daha çok tercih edilir (Gokpinar ve diğ., 2017).

Çok fazla parametre olduğunda En Küçük Karelerden daha iyidir.

Boyutluluk lanetine(curse of dimensionality) karşı çözüm sunar.

Birden çok doğrusal bağlantı sorunu olduğunda etkilidir.

Tüm değişkenlerle bir model oluşturur. İlişkisiz değişkenleri modelden çıkarmaz, katsayılarını sıfıra yaklaştırır.

λ kritik bir rol oynar. İki terimin göreceli etkilerini kontrol etmeye izin verir.

λ için optimum bir değer bulmak önemlidir. Bunun için çapraz doğrulama yöntemi kullanılır.

#### 4.2.3 Lasso Regresyon

Katsayılar, Ridge regresyonunda olduğu gibi hata kareleri toplamını minimize eden katsayılara ceza uygulanarak bulunur. Ridge regresyonundan farkı cezaların, katsayıların sıfır olacağı şekilde uygulanmasıdır. Böylece değişken seçimi yapılır (Tibshirani, 1996). $L_1$ düzenleme yöntemi olarak da anılır. Eşitlik (2)'de Lasso regresyon ifade edilmiştir.

$$SSE_{L_1} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{P} |\beta_j| \qquad (2)$$

Modelde tüm ilgili-ilgisiz değişkenleri bırakarak Ridge regresyonunun dezavantajının üstesinden gelinmesi önerilmiştir.

Katsayıları sıfıra yaklaştırır.

λ yeterince büyük olduğunda katsayıları sıfır yapar. Böylelikle değişken seçimi yapılır.

Doğru λ seçimi için çapraz doğrulama kullanılır.

Ridge ile Lasso arasında bir üstünlük yoktur. Ridge regresyon $L_2$ düzenleme parametresini kullanırken Lasso regresyon $L_1$ düzenleme parametresini kullanır.

#### 4.2.4 Elasticnet Regresyon

ElasticNet ayrıca Ridge ve Lasso'da olduğu gibi cezalar uygulayarak SSE'yi en aza indiren bir değere sahiptir. ElasticNet, $L_1$ ve $L_2$ yaklaşımlarını birleştirir (Zou ve Hastie, 2005). Böylelikle daha etkili düzeltme elde edilir. Ridge tipi ceza, Lasso tipi değişken seçimi ile birlikte yapılır. Regresyon modellerinin geliştirilmiş bir versiyonudur. Ancak en gelişmiş haliyle olması her zaman iyi sonuçlar vereceği anlamına gelmez. Bu durum veri setinden

veri setine farklılık gösterecektir. Eşitlik (3)'te ElasticNet regresyon ifade edilmiştir.

$$SSE_{Enet} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda_1 \sum_{j=1}^{P} \beta_j^2 + \lambda_2 \sum_{j=1}^{P} |\beta_j| \quad (3)$$

### 4.2.5 Rastgele Orman

Bir torbalama tekniği olan Rastgele Orman, sınıflandırma ve regresyon için toplu öğrenme yöntemini kullanan bir gözetimli öğrenme algoritmasıdır. Rastgele ormanlardaki ağaçlar paralel olarak uzanır. Ağaçları inşa ederken bu ağaçlar arasında etkileşim yoktur. Eğitim zamanında çok sayıda karar ağacı inşa ederek ve sınıfların modu (sınıflandırma) veya ayrı ağaçların ortalama tahmini (regresyon) olan sınıfı çıkararak çalışır. Rastgele bir orman, birçok karar ağacını bazı yararlı değişikliklerle bir araya getiren bir meta tahminleyicidir. Her düğümde bölünebilen özelliklerin sayısı, toplamın bir yüzdesi ile sınırlıdır. Bu, topluluk modelinin herhangi bir özelliğe çok fazla güvenmemesini sağlar ve tüm potansiyel olarak öngörücü özelliklerin adil kullanımını sağlar. Her ağaç, bölünmelerini oluştururken orijinal veri kümesinden rastgele bir örnek alır ve aşırı uyumu önleyen başka bir rasgelelik öğesi ekler (Liaw ve Wiener, 2002).

### 4.2.6 K-En Yakın Komşu

Gözlem benzerliğine göre tahminler yapılır. Sınıflandırma problemleri için ortaya çıkmış ve daha sonra regresyon problemlerine uygulanmıştır. Parametrik olmayan bir öğrenme türüdür (Peterson, 2009). Adımları:

1. Komşuların sayısı (k) belirlenir.
2. Bilinmeyen nokta ile diğer tüm noktalar arasındaki mesafe hesaplanır.
3. Uzaklıklar artan sırada listelenir ve en yakın k gözlem seçilir.
4. Sınıflandırma en sık sınıf olarak, regresyon ise ortalama değer tahmin değeri olarak verilir.

### 4.3 Model Performans Değerlendirme

Model oluşturulduktan sonra performansı değerlendirilmelidir. Aksi takdirde oluşturulan modelin verimliliği analiz edilemez. Bu model hakkında yorum yapılmasını ve doğru sonuca ulaşılmasını engelleyen bir durumdur. Bunun için modelin önce Hold-out, K-katlı çapraz doğrulama, Leave One Out ve Bootstrap gibi yöntemlerle doğrulanması gerekir. Model doğrulandıktan sonra hata fonksiyonları incelenmelidir. Böylelikle eğitim seti ve veri setinden parçalanmış test seti için modelin hataları bulunacaktır. Daha sonra eğitim seti ile test seti arasındaki ilişkinin hatalar açısından incelenmesine olanak sağlar. Yanlılık ve varyans durumları incelenir ve modelde aşırı uyum veya yetersiz uyum olup olmadığı kontrol edilir.

### 4.3.1 Model Doğrulama

Modellerin ürettiği sonuçların doğru değerlendirilmesi çalışmasıdır. Böylelikle model, başarı sonuçlarını daha doğru değerlendirebilecektir. Eğitim seti üzerine kurulan modelin ürettiği sonuçların doğruluğu çeşitli yöntemlerle değerlendirilmelidir.

Hold-out Yöntemi: Veri seti, eğitim seti ve test setine bölünür. Eğitim seti ile kurulan modelin performansı, test seti ile ölçülür. Genellikle orijinal veri setinin 2/3'ü eğitim seti ve 1/3'ü test seti olarak ayrılır (Omary ve Mtenzi, 2010). Veri setindeki gözlem sayısının az olduğu durumlarda, eğitim ile test setlerinin birbirinden ayrılmasında güçlükler yaşanacağı için başka yöntemler kullanılmaktadır.

K-Katlı Çapraz Doğrulama: Orijinal veri seti, eğitim ve test setleri olarak ikiye ayrılır. Hold-Out yönteminden farklı olarak, veri seti k alt bölüme ayrılmıştır. Belirtilen alt kümelerden biri dışarıda bırakılır. Kalan k-1 parçası ile oluşturulan model 1 adet dışarıda bırakılarak test edilir. Bu durum, veri setinin tüm alt kümeleri için aynı şekilde tekrarlanır. Son durumda, k eğitim hatası oluşturulur. Eğitim kümesinin hatası, bu hataların aritmetik ortalaması alınarak bulunur. Eğitim seti hatası, test hatasının kötü bir öngörücüsüdür (Hastie, Tibshirani ve Friedman, 2001). Eğitim seti hatasının iyi bir test hatası öngörücüsü yapmanın yolu, onu doğrulamaktır. Orijinal veri setindeki bir eğitim setinin hatasını test seti hatası üzerinde karşılaştırmak yerine, eğitim setini kendi başına parçalara bölerek hata doğrulanır. Eğitim seti hatasının kötü bir tahminleyici olması azaltılmaya çalışılır.

Leave One Out Yöntemi: K-Katlı Çapraz Doğrulama yönteminin özel bir durumudur. Set sayısı, gözlem sayısına eşittir (n gözlem, n set). Her yinelemede bir gözlem atlanır ve diğerleriyle test hatası hesaplanır. Model n-1 numune ile eğitilir ve dışarıda kalan 1 numune ile test edilir. Bu, her numunenin test için bir kez kullanılması için yapılır (Wong, 2015).

Bootstrap Yöntemi: Örnek, orijinal veri kümesinden değiştirilerek seçilir. Modelin hatası, elde edilen her örnek üzerinde hata hesaplanarak belirlenir. Eğitim ve test setlerine ayrılma burada da söz konusudur. Test seti ayırma için 2 yöntem vardır. İlkinde numunelerden biri test seti olarak seçilebilir. İkinci olarak, test seti önce ayrılabilir. Doğrulama bu şekilde yapılır (Efron ve Tibshirani, 1985).

### 4.3.2 Yanlılık-Varyans İkilemi

Modellerin tahmin başarısının değerlendirilmesidir. Her şeyden önce esneklik, verilerin fonksiyonel yapısının algoritma tarafından uygun şekilde değerlendirilmesidir. Algoritmanın bunu yapma esnekliği, yanlılık ve varyans fenomenini ortaya çıkarır.

Yanlılık, gerçek değerler ile tahmin edilen değerler arasındaki mesafeyi ifade eder. Modelin yüksek yanlılığı, modelin daha az öğrenmesi anlamına gelir. Fonksiyon, tüm noktaları temsil edemez ve gerçek değerleri tahmin etmekten uzaklaşır. Bazı gözlemlere göre yüksek bir yanlılık oluşturur. Bu da az öğrenmeye neden olur.

Varyans, değişkenlik anlamına gelir. Bir modelin hassasiyetinin bir ölçüsüdür. Varyans ne kadar yüksekse, esneklik o kadar yüksek olur. Tahmin fonksiyonunun yapıyı veri seti içinde temsil etme yeteneği arttıkça varyansı artar ve yanlılığı azalır. Bu, modeli bir noktaya kadar başarılı kılar. Ancak bu durum arttıkça model daha çok öğrenir. Varyans arttıkça model özelleşir. Amaç, bir model oluştururken genellenebilir bir yapı sunmak olduğu için model, veri setini ezberlemektedir. Örnek veri setinin yapısını öğrenmekten uzaklaşır ve veri setini aynen kopyalar. Bu durum aşırı öğrenmeye neden olur. Hiç görmediği veriler veri setine yüklendiğinde hatalar üretecektir.

Doğru modelde esneklik doğru seçilmelidir. Yüksek varyansa ve yüksek önyargıya sahip olmamalıdır. Yüksek tahmin başarısı her zaman veri setinde iyi sonuçlar vermez. Önemli olan veri setinin algoritmasının doğru oluşturulması ve genelleştirilebilir modelin yakalanmasıdır.

### 4.3.3 Model Optimizasyonu

Model tahmininin performansını iyileştirmeyi amaçlayan optimizasyon için kullanılan bazı kavramlar vardır. Teorik olarak verilerde bulunmayan kavramsal model parametresi olarak adlandırılır ve bunlara erişim için bazı hatalar dikkate alınır, katsayılar ve ağırlıklar hesaplanır. Model hiperparametresi, kullanılan makine öğrenmesi algoritmalarının harici parametresidir. Model parametresi verilerden çıkarılabilecek katsayı iken, model hiperparametresi verilerden elde edilemeyen değerleridir. Model hiperparametreleri kullanıcı tarafından belirlenir. Modeli optimize etmek için kullanılan parametrelere model hiperparametreleri denir. Modelin kendisi tarafından üretilen katsayılara parametre adı verilir. Aralarındaki en önemli fark, bir tanesinin kullanıcı tarafından verilen veri setinden üretilmesidir. Model hiperparametrelerinde deneyler yaparak en küçük hatayı veren katsayıyı kullanmak mümkündür (Seyedzadeh, Rahimian, Rastogi ve Glesk, 2019).

Modelin parametrelerinin ayarlanması, model parametre ayarı olarak adlandırılır. Bu durum parametreye ve hiperparametreye bağlı olarak dahili veya harici olabilir. Model ayarlama konsepti, hepsini kapsayan genel bir kavramdır. Modelin parametrelerini hiperparametreler aracılığıyla bulmak, hiperparametre aracı olmadan değişken mühendisliği yapmak, değişkenleri seçmek, model seçimi yapmak, hiperparametrelerle modelleri geliştirmek, teste ulaşmak ve model doğrulama yöntemleri ile doğrulanmış eğitim hatalarının hepsi model ayarlamasıdır. Bu çalışmada araştırma ve yayın etiğine uyulmuştur.

## 5. Ofis Malzemeleri Satışı Yapan Firma İçin Satış Tahmini

Örnek çalışma, perakende ofis malzemeleri satışı yapan bir firmada gerçekleştirilmiştir. Alınan veri seti 01.01.2019 ile 31.08.2020 tarihlerini kapsamakta olup yirmi aylık veri içermekle birlikte 1470 satır ve 26 sütundan oluşmaktadır. Python programlama dili kullanılarak JupyterLab platformunda yazılan veri setine ilişkin bilgiler Şekil 2'de verilmiştir. İlgili veri setindeki 26 sütundaki eksik veri sayıları verilmiştir. Bu sütunlar 4 adet veri tipi içermektedir. Kategorik ve numerik değişkenler içeren bu veri setinde, eksik veri ve aykırı veriler veri tipine göre analiz edilmiştir.

Çünkü kategorik değişkenler için istatistiksel analizleri kullanmak mümkün değildir.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1470 entries, 0 to 1469
Data columns (total 26 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   KODU                            1470 non-null   object
 1   ADI                             1470 non-null   object
 2   TARİH                           1470 non-null   datetime64[ns]
 3   SERİ                            1470 non-null   object
 4   SIRA NO                         1470 non-null   int64
 5   BELGE NO                        0 non-null      float64
 6   BELGE TARİHİ                    1470 non-null   datetime64[ns]
 7   EVRAK TİPİ                      1470 non-null   object
 8   HAREKET CİNSİ                   1470 non-null   object
 9   TİPİ                            1470 non-null   object
 10  N/İ                             1470 non-null   object
 11  DEPO                            1470 non-null   object
 12  GİREN MİKTAR                    1470 non-null   int64
 13  ÇIKAN MİKTAR                    1470 non-null   float64
 14  BİRİM ADI                       1470 non-null   object
 15  ANA DÖVİZ BRÜT BİRİM FİYATI     1470 non-null   float64
 16  ANA DÖVİZ NET BİRİM FİYATI      1470 non-null   float64
 17  ANA DÖVİZ BRÜT TUTAR            1470 non-null   float64
 18  ANA DÖVİZ NET TUTAR             1470 non-null   float64
 19  VERGİLİ TUTAR (YEREL DÖVİZ)     1470 non-null   float64
 20  VERGİLİ TUTAR (ALTERNATİF DÖVİZ) 1470 non-null  float64
 21  AÇIKLAMA                        552 non-null    object
 22  AÇIKLAMA2                       229 non-null    object
 23  FATURA SERİ                     1454 non-null   object
 24  FATURA SIRA NO                  1454 non-null   float64
 25  MÜŞTERİ NO                      1470 non-null   int64
dtypes: datetime64[ns](2), float64(9), int64(3), object(12)
memory usage: 298.7+ KB
```

Şekil 2. Ham veri setine ait bilgiler

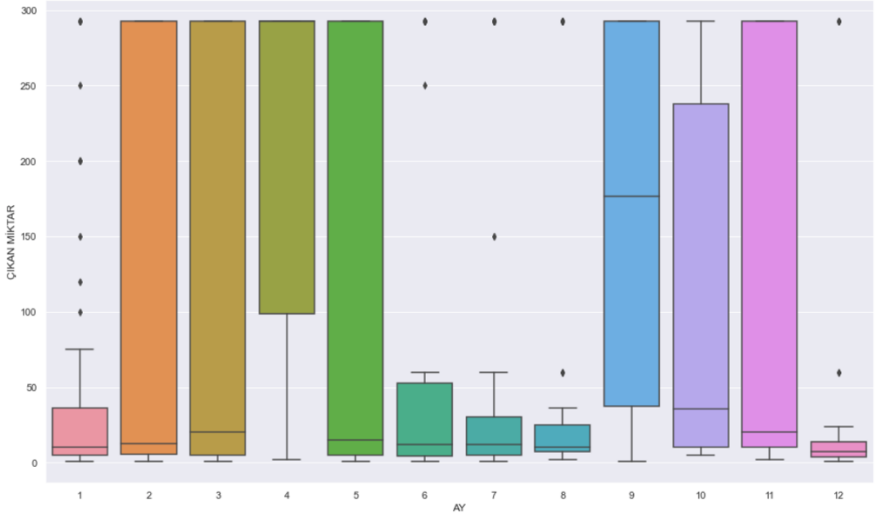
Keşifçi veri analizi: Verinin yapısını anlayabilmek adına keşifçi veri analizi yapılmıştır ve modelde kullanılmayacak olan 18 sütun düşürülmüş, kalan 8 sütun üzerinden model çalışılmıştır. Modelde kullanılan sütunlar KODU, ADI, TARİH, ÇIKAN MİKTAR, BİRİM ADI, VERGİLİ TUTAR (YEREL DÖVİZ), FATURA SIRA NO ve MÜŞTERİ NO sütunlarıdır. Zaman bazında analiz yapabilmek adına veriye ilgili tarihi baz alacak şekilde GÜN, HAFTANINGÜNÜ, HAFTA, AY ve YIL sütunları eklenmiştir.

Eksik veri analizi: Eksik veriler analiz edildiğinde Şekil 3'te verildiği gibi FATURA SIRA NO'ya dair 16 eksik veriye rastlanmıştır. Eksik veriler sıra no içerdiğinden ortalama ya da herhangi bir yöntem ile verileri doldurmak mantıksız olacaktır. Bu sebeple ilgili veriler veri setinden çıkarılmıştır.

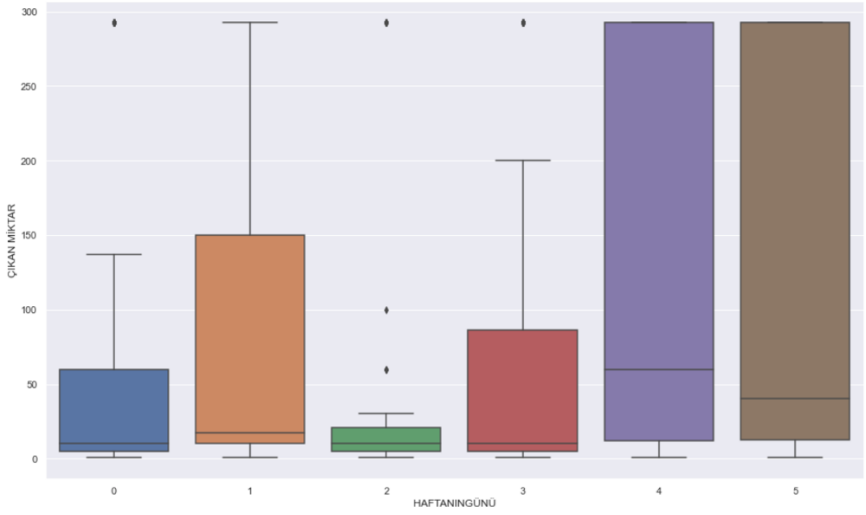
```
KODU                          0
ADI                           0
TARİH                         0
ÇIKAN MİKTAR                  0
BİRİM ADI                     0
VERGİLİ TUTAR (YEREL DÖVİZ)   0
FATURA SIRA NO               16
MÜŞTERİ NO                    0
dtype: int64
```

Şekil 3. Eksik veri analizi

Şekil 4'teki kutu grafiği aylara göre satış miktarını gösterirken, Şekil 5 haftanın günlerine göre satış miktarını göstermektedir. Pazar günü satış yapılmadığından altı gün gösterilmiştir. Örneğin grafiğe göre cuma ve cumartesi günleri en çok satış yapılırken çarşamba günü yapılan satış en azdır. Bu da şirkete, alacağı aksiyon kararlarında fikir verecektir. Benzer durum aylara göre yapılan satışlarda da geçerlidir.




Şekil 4. Aylara göre satış kutu grafiği




Şekil 5. Haftanın günlerine göre satış kutu grafiği

Aykırı değerler iyi bir satış tahmini kararı alınmasında karşılaşılan en büyük engellerdendir. Çünkü veri setinde çok fazla rastlanmayan bu aykırı değerler, tahmin kararının sapmasına yol açacaktır. Bu nedenle satış miktarındaki aykırı değerler için çeyrekler açıklığı (IQR) kullanılmıştır. Birinci çeyrek değerinin 5, üçüncü çeyrek değerinin ise 120 olduğu gözlemlenmiştir. Bu nedenle birinci ve üçüncü çeyrekten 1,5 açıklık verilerek alt sınır ve üst sınır elde edilmiştir. Aykırı değerler verinin üst sınıra uygun alanlarında gözlemlendiğinden üst sınır olan 292,5 değeri aykırı değerler için kullanılmıştır. Üst sınırdan büyük aykırı değerler analiz edildiğinde 72 gözlem birimine rastlanılmıştır. İlgili gözlem birimleri sayıca görece fazla görülmüş ve bu satırlar düşürülmek yerine baskılama yöntemi kullanılmıştır. Baskılama yöntemi ile üst sınır değerlerinden yüksek olan gözlem birimleri 292,5 değerine indirilmiş ve yeni veri çerçevesi (dataframe) oluşturulmuştur.

Modelin kurulması aşamasında ise öncelikle LinearRegression, Ridge, Lasso, ElasticNet, KNeighborsRegressor ve RandomForestRegressor kütüphaneleri indirilmiştir. ÇIKAN MİKTAR ve TARİH üzerinden sklearn.model_selection() kütüphanesinden train_test_split() fonksiyonu çağırılarak model validasyonu gerçekleştirilmiş, modele en uygun eğitim ve test verileri oluşturulmuştur. Eğitim ve test verilerinin fonksiyon kullanılarak seçilmesinin sebebi, ilgili fonksiyonun veri setini okuyarak uygun şekilde bölmesinden ileri gelmektedir. x_cv ve y_cv olmak üzere oluşturulan eğitim verileri, x_train ve y_train olarak oluşturulan test setleri üzerinde test edilmiştir. Kendi içinde iki ayrı veri seti olarak gözükmesine rağmen ilgili veri setleri bağımlı ve bağımsız değişkenleri ifade etmesine göre ayrılmıştır. Burada y_cv ve y_train veri setleri sadece hedef değişken olan ANA DÖVİZ NET TUTAR sütununu içermektedir. Eğitim ve test verisi lin.fit(x_train, y_train) komutu ile fit edilmiş ve lin_predict() komutu ile x_cv veri seti tahmin edilmek için çağırılmıştır. İlgili algoritmaları içeren, yukarıda belirtilen kütüphaneler ve $r^2$ için iki dizi oluşturularak çağırılmış ve veri setleri bu haliyle girdi olarak algoritma kütüphanelerinde işlenmek üzere okutulmuştur. Modelden models.sort_values(by="r_2", ascending=False) komutu ile $r^2$ skorunu azalan şekilde vermesi istenmiş ve Şekil 6'daki $r^2$ skorları elde edilmiştir. Elde edilen skorlara göre en iyi tahmincinin Rastgele Orman yöntemi olduğu görülmüştür.

Tablo 2
Algoritmaların $r^2$ skorları

| | Yontem | Sonuc |
|---|---|---|
| 2 | RandomForestRegressor | 0.839980 |
| 0 | LinearRegression | 0.670193 |
| 4 | ElasticNet | 0.663131 |
| 5 | Ridge | 0.662641 |
| 3 | Lasso | 0.661614 |
| 1 | KNeighborsRegressor | 0.440423 |

Rastgele Orman yöntemine göre model hipertune edilmiştir. K-En Yakın Komşu algoritması için best_params_ fonksiyonu kullanılarak en iyi parametreler 2 olarak bulunmuştur. effective_metric fonksiyonu ile veri setine uygun ölçümün Öklid olduğu belirlenmiştir. Bu işlemlere göre knn_tuned adında bir dizi tanımlanmış ve bulunan veri setine ait özellikler içine yerleştirilmiş ve KNeighborsRegressor kütüphanesi tekrar çağırılarak x_train ve y_train veri setleri fit edilmiştir. Yapılan işlemlerden sonra $r^2$ skorunun 0,5302 olduğu görülmüştür. Rastgele Orman algoritması için ise max_depth, max_features ve n_estimators parametreleri kullanılarak rf_tuned adında bir dizi oluşturulmuştur. Veri seti ilgili parametreler için işleme koyulduktan sonra en iyi parametrelerin max_depth için 9, max_features için 5 ve n_estimators için 200 olduğu bulunmuştur. Bu verilere göre RandomForestRegressor kütüphanesi ile algoritma tekrar çağırılmış ve $r^2$ skorunun 0,9726 olduğu hesaplanmıştır. Lineer regresyon ile 0,7213 ile elde edilen skor, Ridge, Lasso ve ElasticNet algoritmaları ile sırasıyla 0,7134, 0,7128, 0,7189 skorları elde edilmiştir. Yapılan analizin sonucunda 8,3 dakika ile en uzun süre Rastgele Orman algoritmasına ait olmasına rağmen en iyi sonucun elde edildiği gözlemlenmiştir.

Model hipertune edildikten sonra elde edilen $r^2$ skorunun 0.9726 olduğu görülmüştür. Bu da Rastgele Orman yöntemiyle yapılan satış tahminin diğer yöntemlere göre iyi bir şekilde çalıştığının göstergesi niteliğindedir. Model oluşturulduktan sonra veriler istenilen tarih aralığı için tahmin yapmaya uygun haldedir.

## 6. Sonuçlar

Günümüz dijital dünyasında satın alma gittikçe arttığından veriler çok büyük boyutlara ulaşmıştır. Endüstrinin getirdiği kavramlardan en belirgini ise boyutluluk laneti olmuştur. Böylelikle işletmeler satın alma kararlarını alırken büyük zorluk yaşamaktadır. Geleneksel satış tahmini yöntemleri işletmelerin ihtiyaçlarına istenilen ölçüde cevap veremediğinden akademi ve endüstri farklı arayışlara girmiştir. İçinde bulunduğumuz yapay zekâ çağında, makine öğrenmesi algoritmaları iyi bir tahminci olduklarından satış tahmini problemlerinde de uygulanması mümkündür. Bu çalışmada uçtan uca bir makine öğrenmesi proje süreci ele alınmıştır. Herhangi bir makine öğrenmesi projesinin adımları ve veriye yaklaşım boyutu tanıtılmıştır. Uygulama bölümünde makine öğrenmesi algoritmalarından doğrusal regresyon, Ridge, Lasso, ElasticNet, K-en yakın komşu ve Rastgele Orman algoritmaları kullanılarak JupyterLab üzerinden Python programlama dilinde bir satış tahmini çalışması gerçekleştirilmiştir. Algoritmalar arasından en düşük hatayı veren algoritma Rastgele Orman algoritması olmuştur. Bu sebeple ilgili veri setinde belirtilen algoritmalar arasından Rastgele Orman algoritması kullanılarak tahminin yapılması uygun görülmüştür. Aylık veya haftalık olarak istenen verilerden tahmin yapılması mümkündür. Aynı zamanda ilgili işletmenin pazara yeni bir ürün koyması durumunda işletmenin verileri analiz edilmiş olduğundan yeni bir ürün için de tahmin yapılabilmesi mümkündür. Bu çalışmada geleneksel tahmin yöntemlerinde eksik veri ve aykırı verilerin analiz edilmemesinden doğan problemler aşılmaya çalışılmıştır. Satış tahmini ile ilgili literatürdeki yöntemlere kıyasla bu çalışmanın iki ana avantajı vardır. Öncelikle geçmişte diğer yöntemlere dayalı satış tahmini çalışmalarında karşılaşılan hız sorunu çözülmüştür. Bir firma yeni bir ürünü ilgili sektörün pazarına sürmek istediğinde benzer davranış gösteren ürünler dikkate alınarak satış tahmini kolaylaşacaktır. İkincisi, aykırı değerlere ve eksik değerlerden doğan tahmin eksiklikleri Keşifçi Veri Analizi ile çözülmüştür. Yapılan satış tahmini çalışması, farklı sektörlere de uygulanabilir olup diğer makine öğrenmesi algoritmaları da kullanılarak çalışmanın içeriğini genişletmek mümkündür. Bu çalışmanın sadece belirli alanlarda uygulanan makine öğrenmesi çalışmalarının, gerçek hayata adapte olması sebebiyle ilgili araştırmacılara rehber olması beklenmektedir. Gelecek çalışmalar birçok farklı sektörde gerçekleştirilebilir ve endüstrinin gerektirdiği diğer tahmin konularına uygulanabilir. Yöneticilerin tahminleri analiz etmesine izin verecek bir ara yüz geliştirmek de mümkün olacaktır.

## Araştırmacıların Katkısı

Bu araştırmada; Emine Nur NACAR, problemin belirlenmesi, bilimsel yayın araştırması, veri toplanması, bu verilerin bilgisayar ortamına aktarılarak yöntemin belirlenmesi ve uygulanması, problemin çözümü ve makalenin hazırlanması; Babek ERDEBİLLİ, bilimsel yayın araştırması, çalışmanın gözden geçirilmesi ve sonuçların yorumlanması konularında katkı sağlamışlardır.

**Çıkar Çatışması**

Yazarlar tarafından herhangi bir çıkar çatışması beyan edilmemiştir.

**Kaynaklar**


Antipov, E. A., & Pokryshevskaya, E. B. (2020). Interpretable machine learning for demand modeling with high-dimensional data using Gradient Boosting Machines and Shapley values. *Journal of Revenue and Pricing Management*, *19*(5), 355–364. doi: https://doi.org/10.1057/s41272-020-00236-4

Armstrong, J. S. (1989). Combining Forecasts: The End of the Beginning or the Beginning of the End? *International Journal of Forecasting*, *5*(4), 585–588. doi: https://doi.org/10.1016/0169-2070(89)90013-7

Biamonte, J., Wittek, P., Pancotti, N., Rebentrost, P., Wiebe, N., & Lloyd, S. (2017). Quantum machine learning. *Nature*, *549*(7671), 195–202. doi: https://doi.org/10.1038/nature23474

Catal, C., Ece, K., Arslan, B., & Akbulut, A. (2019). Benchmarking of Regression Algorithms and Time Series Analysis Techniques for Sales Forecasting. *Balkan Journal of Electrical and Computer Engineering*, *7*(1), 20–26. doi: https://doi.org/10.17694/bajece.494920

Dietterich, T. G. (1997). Machine-Learning Research. *AI Magazine*, *18*(4), 97–136. doi: https://doi.org/10.1145/1056743.1056744

Efron, B., & Tibshirani, R. (1985). The bootstrap method for assessing statistical accuracy. *Behaviormetrika*, *12*(17), 1-35.

Gokpinar, E., Ebegil, M., & Gokpinar, F. (2017). A Review on Shrinkage Parameters in Ridge Regression. *GU Journal of Science*, *30*(4), 565–582. Erişim adresi: https://dergipark.org.tr/tr/download/article-file/380312

Grigorev, A. (2020). *Machine Learning Bookcamp MEAP V06* (A. Books (ed.)). Copyright 2020 Manning Publications. Erişim adresi: https://www.manning.com/books/machine-learning-bookcamp

Hastie, T., Tibshirani, R., & Friedman, J. (2001). The Elements of Statistical Learning. In *Vol.1 No.10*. New York: Springer series in statistics. doi: https://doi.org/10.1007/978-1-4419-9863-7_941

Helmini, S., Jihan, N., Jayasinghe, M., & Perera, S. (2019). Sales forecasting using multivariate long short term memory network models. *PeerJ PrePrints*, *7*, e27712v1. doi: https://doi.org/10.7287/peerj.preprints.27712

Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, *12*(1), 55–67. doi: https://doi.org/10.1080/00401706.1970.10488634

Jain, A., Menon, M. N., & Chandra, S. (2015). *Sales Forecasting for Retail Chains*. 1–6. Erişim adresi: https://pdfs.semanticscholar.org/76a2/44f4da1d29170a9f91d381a5e12dc7ad2c0f.pdf

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260. Erişim adresi: https://science.sciencemag.org/content/349/6245/255

Komorowski, M., Marshall, D. C., Salciccioli, J. D., & Crutain, Y. (2016). Exploratory Data Analysis. In *Secondary Analysis of Electronic Health Records* (pp. 185–203). doi: https://doi.org/10.1007/978-3-319-43742-2

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, *2*(3), 18–22.

Mitchell, T. M. (1997). Machine Learning. In *McGraw-Hill Science/Engineering/Math*. doi: https://doi.org/10.1007/978-3-642-21004-4_10

Nilsson, N. J. (2014). Principles of Artificial Intelligence. In *Morgan Kauffmann*. Erişim adresi: https://stacks.stanford.edu/ file/ druid:zd294jv9941/zd294jv9941.pdf

Omary, Z., & Mtenzi, F. (2010). Machine learning approach to identifying the dataset threshold for the performance estimators in supervised learning. *International Journal for Infonomics (IJI)*, *3*(3), 314-325.

Papacharalampous, G., Tyralis, H., & Koutsoyiannis, D. (2018). Univariate Time Series Forecasting of Temperature and Precipitation with a Focus on Machine Learning Algorithms: a Multiple-Case Study from Greece. *Water Resources Management*, *32*(15), 5207–5239. doi:

https://doi.org/10.1007/s11269-018-2155-6

Pavlyshenko, B. M. (2019). Machine-Learning Models for Sales Time Series Forecasting. *Data*, *4*(1), 1–11. doi: https://doi.org/10.3390/data4010015

Peterson, L. E. (2009). K-nearest neighbor. *Scholarpedia*, *4*(2), 1883.

Rincon-Patino, J., Lasso, E., & Corrales, J. C. (2018). Estimating Avocado Sales Using Machine Learning Algorithms and Weather Data. *Sustainability*, *10*(10), 12. doi: https://doi.org/10.3390/su10103498

Rubin, D. (1976). Inference and Missing Data. *Biometrika*, *63*(3), 581–592. doi: https://doi.org/10.1093/biomet/63.3.581

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, *3*(3), 210–229. doi: https://doi.org/10.1147/rd.441.0206

Sathya, R., & Abraham, A. (2013). Comparison of supervised and unsupervised learning algorithms for pattern classification. *International Journal of Advanced Research in Artificial Intelligence*, *2*(2), 34–38. Erişim adresi: https://www.researchgate.net/publication/273246843_Comparison_of_Supervised_and_Unsupervised_Learning_Algorithms_for_Pattern_Classification

Seber, G. A., & Lee, A. J. (2012). *Linear regression analysis* (Vol. 329). John Wiley & Sons.

Seyedzadeh, S., Rahimian, F. P., Rastogi, P., & Glesk, I. (2019). Tuning machine learning models for prediction of building energy loads. *Sustainable Cities and Society*, *47*, 101484.

Siau, K., & Wang, W. (2018). Building Trust in Artificial Intelligence, Machine learning, and Robotics. *Cutter Business Technology Journal*, *31*(2), 47–53. Erişim adresi: https://www.researchgate.net/publication/324006061_Building_Trust_in_Artificial_Intelligence_Machine_Learning_and_Robotics

Şener, S. (2019). Makine Öğrenmesi Yardımıyla Zincir Restoran Gıda Satışlarının Tahmin Edilmesi Ve Hava Durumunun Etkisinin İncelenmesi, İstanbul Teknik Üniversitesi, İstanbul.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. doi: https://doi.org/10.1093/ej/31.121.133

Tukey, J. W. (1977). *Exploratory Data Analysis* (Vol. 2). Pearson. doi: https://doi.org/10.1007/978-3-662-45006-2_9

Velleman, P. F., & Hoaglin, D. C. (1981). Applications, Basics, and Computing of Exploratory Data Analysis. In *Duxbury Press*. Erişim adresi: https://ecommons.cornell.edu/handle/1813/78

Verstraete, G., Aghezzaf, E. H., & Desmet, B. (2020). A leading macroeconomic indicators' based framework to automatically generate tactical sales forecasts. *Computers and Industrial Engineering*, *139*(August 2019), 106169. doi: https://doi.org/10.1016/j.cie.2019.106169

Weng, T., Liu, W., & Xiao, J. (2019). Supply chain sales forecasting based on lightGBM and LSTM combination model. *Industrial Management and Data Systems*, *120*(2), 265–279. doi: https://doi.org/10.1108/IMDS-03-2019-0170

Wong, T. T. (2015). Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognition*, *48*(9), 2839-2846.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *67*(2), 301–320. doi: https://doi.org/10.1111/j.1467-9868.2005.00503.x