İSTANBUL TEKNİK ÜNİVERSİTESİ ★ LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

SATIŞ ADEDİNİ ETKİLEYEN DEĞİŞKENLERİN KEŞFİ VE DUYARLILIK ANALİZİ UYGULAMASI: E-TİCARET ÖRNEĞİ

YÜKSEK LİSANS TEZİ

Rabia AYDIN

Endüstri Mühendisliği Anabilim Dalı

Endüstri Mühendisliği Programı

HAZİRAN 2023

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

SATIŞ ADEDİNİ ETKİLEYEN DEĞİŞKENLERİN KEŞFİ VE DUYARLILIK ANALİZİ UYGULAMASI: E-TİCARET ÖRNEĞİ

YÜKSEK LİSANS TEZİ

Rabia AYDIN
(507191123)

Endüstri Mühendisliği Anabilim Dalı

Endüstri Mühendisliği Programı

Tez Danışmanı: Prof. Dr. Fethi ÇALIŞIR

HAZİRAN 2023

ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL

DISCOVERY OF VARIABLES AFFECTING THE NUMBER OF SALES AND APPLICATION OF SENSITIVITY ANALYSIS: E-COMMERCE EXAMPLE

M.Sc. THESIS

Rabia AYDIN
(507191123)

Department of Industrial Engineering

Industrial Engineering Programme

Thesis Advisor: Prof. Dr. Fethi ÇALIŞIR

JUNE 2023

İTÜ, Lisansüstü Eğitim Enstitüsü'nün 507191123 numaralı Yüksek Lisans Öğrencisi Rabia AYDIN, ilgili yönetmeliklerin belirlediği gerekli tüm şartları yerine getirdikten sonra hazırladığı 'Satış Adedini Etkileyen Değişkenlerin Keşfi ve Duyarlılık Analizi Uygulaması: E-Ticaret Örneği' başlıklı tezini aşağıda imzaları olan jüri önünde başarı ile sunmuştur.

Tez Danışmanı :	**Prof. Dr. Fethi ÇALIŞIR** İstanbul Teknik Üniversitesi	.....
Jüri Üyeleri :	**Dr.Öğr.Üyesi Ömer Faruk BEYCA** İstanbul Teknik Üniversitesi	.....
	Prof. Dr. Selim ZAİM İstanbul Sabahattin Zaim Üniversitesi	.....

Teslim Tarihi : 22 Mayıs 2023
Savunma Tarihi : 21 Haziran 2023

Değerli Aileme,

ÖNSÖZ

Başta eğitim hayatım olmak üzere her alanda beni destekleyen sevgili aileme emekleri ve güvenleri için teşekkür ederim.

Bu çalışmanın gerçekleşmesinde bana yol gösteren saygıdeğer hocam Prof. Dr. Fethi ÇALIŞIR'a sonsuz teşekkürlerimi sunarım.

Haziran 2023 Rabia AYDIN

İÇİNDEKİLER

Sayfa

KISALTMALAR

LGBM **:** Light Gradient Boosting

RF **:** Random Forest

CART **:** Decision Tree

XGBoost **:** Extreme Gradient Boosting

CatBoost **:** Category Boosting

GBM **:** Gradient Boosting

RMSE **:** Root Mean Squared Error

MAE **:** Mean Absolute Error

MSE **:** Mean Squared Error

SEMBOLLER

Z	**:** Z-score
x	**:** Gözlem sayısı
μ	**:** Ortalama değer
σ	**:** Standart sapma
Q1	**:** %25'inci çeyrek değer
Q3	**:** %75'inci çeyrek değer
p	**:** Bağımsız değişken sayısı
n	**:** Örneklem büyüklüğü
y	**:** Gerçek değer
$\hat{y}$	**:** Tahmin edilen değer

ÇİZELGE LİSTESİ

Sayfa

ŞEKİL LİSTESİ

Sayfa

SATIŞ ADEDİNİ ETKİLEYEN DEĞİŞKENLERİN KEŞFİ VE DUYARLILIK ANALİZİ UYGULAMASI: E-TİCARET ÖRNEĞİ

ÖZET

Şirketler geçmiş verilerini analiz ederek gelecekteki durumları hakkında fikir sahibi olmayı ve daha etkili bir yol haritası oluşturmayı mümkün kılabilir. Bir şirketin gelirlerini belirleyen temel faktör olan satış adetleri, gelecekteki durumlarını değerlendirebilmek adına öngörüler oluşturmak için kullanılan geçmiş verilere dayalı bir metriktir. Bu çalışma, bir şirketin gelecekteki durumunu tahmin etmek için geçmiş verilere dayalı olarak satış adedi tahminleri gerçekleştirme amacını taşımaktadır.

Çalışmanın ilk aşaması, veri setinde yer alan bilgilerin keşfedilmesiyle başlamıştır. Veri keşfi, veri setindeki değişkenlerin özelliklerini, dağılımlarını ve ilişkilerini anlamak için gerçekleştirilmiştir.

Veri seti, çalışmanın gerekliliklerine uygun hale getirilmek üzere işlenmiş ve düzenlenmiştir. Daha sonra satış adedi tahminlerini gerçekleştirmek için makine öğrenmesi algoritmaları kullanılmıştır. Bu çalışmada, Karar Ağaçları (Decision Tree, CART), Rastgele Orman (Random Forest, RF), Kategori Artırma (Category Boosting, CatBoost), Gradyan Artırma (Gradient Boosting, GBM) ve Ekstrem Gradyan Artırma (Extreme Gradient Boosting, XGBoost) olmak üzere beş farklı makine öğrenmesi algoritması kullanılmıştır.

Modellerin performansları karşılaştırılarak değerlendirilmiştir. Performans karşılaştırmasından sonra CatBoost, Gradient Boosting ve Random Forest algoritmalarının en iyi performansı gösterdiği belirlenmiştir. Bu algoritmalara dayalı olarak seçilen değişkenler incelenmiş ve bu algoritmalarda en önemli değişkenlerin benzer olduğu tespit edilmiştir.

Seçilen en önemli değişkenler, gelecekteki durumların daha iyi anlaşılabilmesi amacıyla kullanılmak üzere belirlenmiş ve bu değişkenler üzerinden duyarlılık analizi çalışması gerçekleştirilmiştir. Bu çalışma için Random Forest algoritması kullanılmış ve her bir değişkenin hedef değişken üzerindeki etkisi incelenmiştir.

Değişkenlerin bir birim artması durumunda tahmin edilen hedef değişkeniyle gerçekleştirilen karşılaştırmalar, değişkenlerin hedef değişken üzerindeki etkilerini ortaya çıkarmıştır.

Bu çalışmada, bir e-ticaret uygulaması üzerinden taze sebze ve meyve satışı yapan bir şirkete ait olan 2022 Ocak-Kasım dönemine ait veri seti kullanılmıştır. Veri seti, sipariş bilgilerini ve sipariş detaylarını içermektedir. Çalışmanın kapsamını genişletmek adına, veri setinde bulunmayan ancak anlamlı olabileceği düşünülen değişkenler veri setine eklenmiştir.

Bu çalışma, şirketlere geçmiş verileri kullanarak gelecekteki durumlarına yönelik fikirler sağlama ve daha etkili bir yol haritası oluşturma imkanı sunmaktadır. Satış

adedi tahminleri ve duyarlılık analizi sayesinde, şirketler satışlarını etkileyen faktörleri anlayabilir ve gelecekteki durumları hakkında daha sağlam bir temel üzerinde planlama yapabilirler.

DISCOVERY OF VARIABLES AFFECTING THE NUMBER OF SALES AND APPLICATION OF SENSITIVITY ANALYSIS: E-COMMERCE EXAMPLE

SUMMARY

Analyzing historical data empowers companies to gain insights into their future circumstances and develop a more effective strategic roadmap. Sales volume, a pivotal determinant of a company's revenue, serves as a metric to glean knowledge about their future prospects by leveraging past data. This study endeavors to forecast sales volume through the utilization of machine learning algorithms, utilizing historical data sourced from an e-commerce platform specializing in the distribution of fresh fruits and vegetables.

The dataset encompassed the period from January to November 2022, comprising comprehensive information pertaining to customer orders and associated details. Furthermore, additional relevant variables were integrated into the dataset to encompass potential influential factors that were initially absent.

The research methodology employed encompassed a systematic approach involving data exploration, data preparation, model development, and performance evaluation. Initially, a comprehensive exploration of the dataset was conducted to acquire a profound understanding of its structure, characteristics, and patterns. This step facilitated the identification of potential variables that could impact sales volumes, including delivery-related factors, customer membership duration, pricing, discounts, temporal factors (e.g., month, weekends), and specific product IDs.

Subsequently, meticulous preparation of the dataset was conducted to ensure compatibility with the employed machine learning algorithms. This entailed addressing missing data, transforming variables when necessary, encoding categorical variables, and splitting the dataset into training and testing sets.

Five machine learning algorithms were selected for the task of sales volume prediction, namely Decision Trees (CART), Random Forest (RF), Category Boosting, (CatBoost), Gradient Boosting (GBM), and Extreme Gradient Boosting (XGBoost). These algorithms were chosen due to their effectiveness in handling intricate non-linear relationships and their capacity to effectively process large datasets.

To assess model performance and determine the most suitable algorithm, a comprehensive comparison utilizing appropriate evaluation metrics such as mean squared error, mean absolute error, and R-squared was conducted. Model performance was evaluated using both the training and testing datasets to ensure generalizability and mitigate overfitting.

Based on the performance evaluation, the CatBoost, Gradient Boosting, and Random Forest algorithms exhibited superior results compared to others. Consequently, these

algorithms were selected for further analysis owing to their ability to capture complex interactions, effectively handle categorical variables, and provide insights into feature importance.

To gain deeper comprehension regarding the impact of individual variables on sales volume, a sensitivity analysis was conducted employing the selected Random Forest model. This analysis involved examining changes in predicted sales volume as each variable was incremented by one unit while keeping other variables constant. By comparing predicted values before and after the increment, the influence of each variable on sales volume was quantified. This analysis provided valuable insights into variables that exerted the most significant effects on sales volume and highlighted how manipulating these variables could potentially drive business growth.

The findings of this study contribute to companies' ability to make informed decisions, develop effective marketing strategies, optimize pricing and discounts, enhance supply chain management, and improve overall operational efficiency. By comprehending the factors influencing sales volume and their respective magnitudes, businesses can align their resources, investments, and marketing efforts accordingly, positioning themselves for sustainable growth and gaining a competitive edge in the market.

Overall, this research serves as a practical and valuable framework for businesses operating in the e-commerce sector, specifically those involved in the sale of fresh fruits and vegetables. The employed methodology, encompassing data exploration, machine learning modeling, and sensitivity analysis, establishes a robust foundation for sales volume prediction and facilitates the acquisition of actionable insights into the underlying drivers of business performance.

Upon conducting a thorough analysis and interpreting the results, several key findings emerged from this research. The predictive models utilizing machine learning algorithms, particularly CatBoost, Gradient Boosting, and Random Forest, demonstrated robust capabilities in accurately forecasting sales volume. The insights provided by the sensitivity analysis shed light on the variables that have the most substantial impact on sales volume. For instance, it was observed that discounts offered on specific products during weekends had a significant positive effect on sales volume, indicating the potential benefits of strategic promotions.

Furthermore, customer membership duration was found to play a crucial role in driving sales. Longer-term memberships exhibited a positive correlation with higher sales volume, indicating the importance of customer loyalty and retention programs. By understanding the factors affecting sales, businesses can tailor their marketing initiatives and customer engagement strategies to boost revenue and customer satisfaction.

The findings of this research offer practical implications for companies operating in the fresh produce e-commerce sector. By leveraging historical data and predictive models, businesses can make data-driven decisions to optimize inventory management, streamline supply chain operations, and reduce wastage. Accurate sales volume forecasting enables companies to plan their procurement and distribution processes efficiently, minimizing stockouts and maximizing revenue.

Additionally, the insights gained from the sensitivity analysis can guide companies in fine-tuning their pricing and discount strategies to attract more customers while maintaining profitability. Aligning these strategies with customer preferences and behaviors can strengthen brand loyalty and foster repeat business.

In conclusion, this study demonstrates the value of analyzing historical data and employing machine learning algorithms to forecast sales volume in the fresh fruits and vegetables e-commerce domain. The research methodology, encompassing data exploration, modeling, and sensitivity analysis, provides a robust framework for sales prediction and actionable insights into the factors influencing business performance.

By embracing data-driven approaches, companies can adapt to the dynamic nature of the market, make informed decisions, and stay ahead of the competition. The knowledge gained from this study empowers businesses to optimize their operations, develop targeted marketing initiatives, and cultivate long-term growth in the fast-evolving landscape of e-commerce.

1. GİRİŞ

Günümüzde internet üzerinden alışveriş yapma oranı geçmişe kıyasla önemli ölçüde artmıştır. Bu durumda Covid-19 sürecinin etkisi büyük bir rol oynamıştır. Covid-19 salgınıyla birlikte insanların alışveriş alışkanlıkları önemli ölçüde değişmiştir. İnsanlar, ihtiyaçlarını internet üzerinden evlerinin konforunda karşılayabilme imkanıyla daha fazla internete yönelmişlerdir. Bu da şirketleri e-ticarete yönelmeye teşvik etmiştir.

E-ticaret sadece alışveriş yapan kişiler için değil aynı zamanda satış yapan şirketler için de bir konfor alanı sağlamaktadır. E-ticaret sayesinde şirketler daha geniş bir müşteri tabanına ulaşabilir ve potansiyel satışlarını artırabilir. Ayrıca farklı sektörlerdeki şirketler e-ticaret üzerinden satış yaparak yeni satış kanalları oluşturma tercihinde bulunabilirler. Örneğin giyim, elektronik, kozmetik ve gıda sektörleri internet üzerinden satış yapan sektörlere örnek olarak gösterilebilir.

Günümüz iş dünyasında şirketler, rekabetçi piyasada ayakta kalabilmek için hizmetlerini verimlilik, güvenilirlik ve kullanılabilirlik açısından geliştirmek zorundadır. Satış tahmini ve etkili talep planlaması, şirketlerin performansını olumlu yönde etkileyebilmektedir (Mitra vd., 2022).

E-ticaret ortamında potansiyel müşterilerin verdiği bilgiler, müşteri ihtiyaçlarının ve ürün satışlarının tahmin edilmesinde kullanılmaktadır. Ürün satış tahmini, günümüz modern iş ortamında önemli bir gereksinim haline gelmiştir. Ürün satış tahminleri sayesinde şirketler kayıplarını azaltabilir ve ekonomik seviyelerini yükseltebilirler (Dharshini & Vijila, 2021).

Yöneticiler genellikle satışları tahmin etmek için kendi algılarına, sezgilerine ve deneyimlerine güvenirler. Ancak bu, kişiden kişiye değişebilir, nitelikli ve deneyimli yöneticilerden sürekli olarak güvenilir girdiler almak zor olabilir. Sonuç olarak, bilgisayar ağları gelecekteki satışları tahmin etmede karar verme sürecine yardımcı olabilir. Makine öğrenimi (ML), büyük miktarda veri ve ilgili bilgileri kullanarak etkili satış tahmin modelleri oluşturmak için kullanılabilir (Mitra vd., 2022).

Satış, günümüz rekabetçi iş dünyasında şirketler için son derece önemlidir ve doğru satış tahmini her başarılı perakende işinde kilit bir rol oynamaktadır. Doğru satış tahmini, fazla üretimi önlemeye ve fazla stok miktarını azaltmaya yardımcı olarak envanter yönetiminde etkili olabilmektedir. Ayrıca maliyet tahmini ile bir şirketin karlılığını artırmak için de harika bir araç olabilir (Ensafi vd., 2022).

Gıda sektöründe, market ürünlerinin yanı sıra sebze ve meyve satışı da internet üzerinden tercih edilen alışveriş alanlarından biridir. Müşterilerin taze ürünlere erişebilmesi, sebze ve meyve alışverişi yapan kişiler için önemli bir faktördür. Bu nedenle, birçok e-ticaret platformunda taze sebze ve meyve satışı yapılmaktadır.

Veri kavramı, e-ticaretin önemli bir parçasıdır. Veri madenciliği yöntemleri, değerli bilgilerin bulunduğu verilerden anlamlı bilgiler çıkarmak için kullanılır. Şirketler, müşteri davranışlarını inceleyerek gelecekteki davranışları tahmin etme gibi analizler yapmak için veri madenciliği yöntemlerini kullanmaktadır. Şirketler, müşteri odaklı olmanın yanı sıra şirket içi organizasyonları planlamak için de veri madenciliği yöntemlerinden yararlanmaktadır. Örneğin, satış tahminlemesi ve gelir tahmini gibi çalışmalar bu kapsamda değerlendirilebilir.

Bu çalışma, taze meyve ve sebze satışı yapan bir e-ticaret platformuna ait veriler kullanılarak müşterilerin geçmiş alışveriş davranışları üzerinden geliri etkileyen faktörleri keşfetmeyi ve bu faktörlerin duyarlılık analizini gerçekleştirmeyi amaçlamaktadır. 2022 yılı Ocak-Kasım dönemine ait veriler kullanılarak veri madenciliği yöntemlerinden makine öğrenmesi algoritmalarıyla özellik seçimi yapılmış ve daha sonra önemli olarak belirlenen özelliklerin duyarlılık analizi gerçekleştirilmiştir.

1.1 Literatür Taraması

Literatürde makine öğrenmesi algoritmaları ile özellik seçilimi yapılan birçok farklı çalışma bulunmaktadır:

Bir çalışmada stratejik bir konumdan uzaklık, çevrimiçi kullanıcı derecelendirmeleri, ağızdan ağıza derecelendirme, otel tarifesi ve müşteri yorumları gibi faktörler dikkate alınarak satış sıralaması tahmini için özelleştirilmiş bir otel öneri modeli geliştirilmesi amaçlanmıştır. Trivago.com sitesinden elde edilen veriler üzerinden çalışma gerçekleştirilmiştir. Yapay sinir ağları modeli kullanılan diğer Random Forest ve Gradient Boosting modellerine göre daha iyi sonuç vermiştir. Otellerin satış sıralamasını tahmin etmede kullanılan değişkenlerin önemine ve değişkenlerin aralarındaki etkileşime bakılmıştır (Srivastava vd., 2022).

Geçmiş satış verilerini kullanarak satış tahmini ve duyarlılık analizi çalışması yapılmıştır. Bu çalışmada ürünün satış tahminini tahmin etmek için makroekonomik göstergelere odaklanılmıştır. Satış tahmini yaparken lineer regresyon modeli, duygu analizi, bas model ve ekonometrik model kurularak hangi modelin daha iyi tahmin ettiği incelenmiştir (Dharshini & Vijila, 2021).

Aynı ürünü birden fazla mağazada satan büyük firmaların talep tahminine yeni bir yaklaşım getiren çalışmada, denetimli makine öğrenmesi algoritmaları ve yapay sinir ağları modeli kullanılarak kıyaslama gerçekleştirilmiştir. Özellik seçimi, veri dönüşümü ve verinin keşfi çalışmada önemli rol oynamıştır. Özellik seçimi sayesinde talep tahminine etki eden özelliklerin keşfi sağlanmıştır (Thivakaran & Ramesh, 2022).

Bu çalışmada, kimya endüstrisi alanındaki veriler üzerinde çalışma gerçekleştirilmiştir. Çalışmada duyarlılık analizi ve aktif öğrenme kullanarak makine öğrenimi ile modellenen nonlineer işlemlerin hesaplama verimliliğini artırmak için bir model indirgeme yöntemi geliştirilmiştir. Öncelikle duyarlılık analizi, model çıktıları ve girdileri arasındaki önemli bağlantıları belirlemek için kullanılmıştır. Duyarlılık analizi ile elde edilen önemli girdi özelliklerini kullanan azaltılmış sıralı sinir ağları (RNN) algoritması, nonlineer sistemi yaklaşık olarak tahmin etmek için geliştirilmiştir. Ayrıca model öngörülü kontrol (MPC) içinde kullanılarak nonlineer sistemi denge durumunda sabitlemek için kullanılmıştır. Bu çalışmanın esas amacı yüksek boyutlu girdi ve çıktılara sahip verinin dezavantajlarını

azaltmak için duyarlılık analizi çalışmasını gerçekleştirerek ilk etapta işlem çıktıları üzerinde büyük etkiye sahip girdilerin belirlenmesini sağlamak olmuştur (Zhao vd., 2022).

Makine öğrenmesi yöntemlerini uygularken performansı artırmak için duyarlılık analizi tabanlı özellik seçilimi yöntemi kullanılarak çalışma gerçekleştirilmiştir. Özellik seçimi yöntemini uygularken toplam duyarlılık indeksine dayalı yaklaşım kullanılarak bu yöntemin avantajları gözlemlenmiş, önerilen yöntemin performansını ölçmek için farklı veri setleri kullanılmış ve diğer yöntemlerle karşılaştırılması gerçekleştirilmiştir. Yapılan çalışmada toplam duyarlılık indeksinin veri setindeki önemli özellikleri etkili şekilde tanımladığı sonucu çıkmıştır. Yapılan çalışmalar sonucunda toplam duyarlılık indeksinin diğer modern özellik seçme modelleriyle rekabet edebileceği sonucu çıkmıştır (Kamalov, 2018).

Tavuklardaki genetik belirteçler ve bağışıklık tepkisi arasındaki ilişkileri incelemek için makine öğrenmesi algoritmasına dayalı duyarlılık analizi çalışması gerçekleştirilmiştir. Bu çalışmada Random Forest algoritması kullanılarak elde edilen modeller aynı değişkenler kullanılarak oluşturulan doğrusal modellerden daha iyi bir sonuç elde etmiştir. Özellik seçimi adımında Boruta algoritması kullanılmıştır. Bu çalışmada kullanılan bu algoritmanın hesaplama açısından zor olduğu sonucuna varılmıştır. Duyarlılık analizinin değişkenler arasında karmaşık doğrusal olmayan ve katkısal olmayan etkileşimlerinin mevcut olabileceği sistemdeki ilgili değişkenlerin tanımlanması için duyarlı olabileceği sonucu bu çalışmada gösterilmiştir (Polewko-Klim vd., 2020).

Gerçekleştirilen çalışmada otomatik makine öğrenimi alanındaki bileşik veri odaklı boruların performansını artırmak için duyarlılık analizi ele alınmıştır. Ancak bileşik boruların performansını etkileyen birçok faktör olduğundan, bileşik boruların parametrelerinin doğru yapılandırılmasının öneminden bahsedilmiştir. Bu nedenle duyarlılık analizleri, bileşik boruların parametrelerinin optimize edilmesine yardımcı olabilir ve otomatik makine öğrenimi çözümlerinin performansını artırabilir. Performansı etkileyen faktörlerden birinin özellik seçimi yöntemi olduğundan bahsedilmiş ancak bu çalışma duyarlılık analizi uygulanarak bileşik boruların olası parametrelerinin performans üzerindeki etkisini değerlendirmesi amaçlanmıştır. Bileşik boruların parametrelerinin doğru yapılandırılmasının modellerin performansını önemli ölçüde etkilediğini belirtmektedir. Bu nedenle duyarlılık

analizleri yoluyla bileşik boruların parametrelerinin optimize edilmesi, modellerin performansını artırabilir. Çalışma sonucunda duyarlılık analizinin, bileşik boruların performansını artırmak için önemli bir araç olabileceği sonucu elde edilmiştir. Makalede, farklı veri kümelerinde yapılan deneyler sonucunda bileşik boruların hassasiyetinin parametrelerin doğru yapılandırılmasına bağlı olduğu gösterilmiştir. Bu nedenle duyarlılık analizleri, bileşik boruların parametrelerinin optimize edilmesine yardımcı olabilir ve AutoML çözümlerinin performansını artırabilir (Barabanova vd., 2021).

Moda perakendesinde önemli konu olan yeni bir ürünün satışlarının tahmin edilmesi hakkında gerçekleştirilen çalışmanın amacı şirketin verimliliğini ve müşteri memnuniyetini artırmaya yardımcı olmaktır. İki katmanlı model kullanılan çalışmanın ilk katmanında talep tahmini lineer regresyon yöntemi kullanılarak gerçekleştirilmiştir. İkinci katmanda satışlar aynı zamanda envanter olarak modele dahil edilmiştir. Çalışmada özellik seçimi uygulanırken Gradient Boosting algoritması kullanılmıştır. Ürün kümelerini belirlemek için K-means algoritması kullanılmıştır. Her kümenin parametrelerini belirlemek için Genetik algoritması kullanılmıştır. Kurulan modelin veri seti Singapurlu bir şirket verisini içermektedir. Kurulan iki katmanlı modelin diğer modeller ile kıyaslaması gerçekleştirilmiş ve diğer modellere göre daha başarılı bir sonuç elde edildiği sonucuna varılmıştır. Ayrıca çalışma kapsamında ürünlerin rekabet gücünü ölçmek ve optimum envanter seviyesini araştırmak için iki indikatör çalışmaya dahil edilmiştir (Chen vd., 2022).

Literatür araştırmasına ait özet tablo Ek A.’ da bulunmaktadır.

2. VERİ MADENCİLİĞİ

Günlük yaşamımızda gerçekleştirdiğimiz çeşitli davranışlar ve hareketler, günümüzde veri olarak kaydedilebilmektedir. Bilgisayar teknolojisinin gelişimiyle birlikte, bu büyük boyuttaki verilerin depolanması daha da kolay hale gelmiştir. Sosyal medya etkileşimleri, internet alışverişleri, sağlık ve yaşam verileri gibi birçok kaynaktan elde edilen veriler, veritabanlarında saklanmaktadır. Bu veriler, veri madenciliği yöntemleri kullanılarak anlamlı bilgilere dönüştürülebilmektedir.

Veri madenciliği, büyük hacimli veri kümelerini analiz ederek içerdikleri örüntüleri ve eğilimleri keşfetme sürecini ifade eden bir disiplindir. Bu süreç, istatistik, algoritmalar, veri yapıları ve bilgisayar mimarisi gibi bilgisayar bilimlerinin çeşitli alanlarının birleşimini içermektedir. Veri madenciliği yöntemleri, verileri işleyerek içerdikleri gizli bilgileri ve ilişkileri ortaya çıkarma amacını taşır.

Veri madenciliği, çeşitli disiplinlerin birleşimiyle oluşan bir alan olarak kabul edilmektedir. Şekil 2.1'de veri madenciliğini oluşturan disiplinlerin ilişkisini gösterilmektedir.

Şekil 2.1 : Veri Madenciliği Disiplinleri.

2.1 Makine Öğrenmesi

Makine öğrenimi, bilgisayar biliminin bir dalı olarak ortaya çıkmış ve çeşitli endüstrilerde kullanılabilen çıkarıma dayalı problemleri algoritmalar aracılığıyla çözümleyen bir bilim dalıdır. Özellikle karmaşık ve doğrusal olmayan veriler için istatistiksel modellere kıyasla daha doğru tahmin etmeyi sağlar. Makine öğrenimi yöntemleri, uygun modellerin karmaşık ve doğrusal olmayan ilişkilere sahip veriler üzerine uygulanmasıyla başarılı sonuçlar elde edebilir.

Makine öğrenimi algoritmaları, büyük ölçekte işlenmesi zor olan verileri istatistiksel modeller kullanarak programlayarak, örneklemden anlamlı çıkarımlar yapmaya çalışır. Makine öğrenimi aynı zamanda, örnek verileri kullanarak performans kriterini optimize eden bir bilgisayar programı olarak da tanımlanabilir. Hem gelecek için tutarlı tahminler yapabilme yeteneğine sahiptir hem de geçmişe yönelik anlamlı sonuçlar üretebilir (Şahinarslan, 2019).

Farklı türde makine öğrenimi teknikleri, çeşitli uygulama alanlarında etkili modeller oluşturmak için kullanılabilir. Modellerin doğası ve hedeflenen sonuç, daha önce incelenen verilerin özelliklerine ve amaçlarına bağlı olarak öğrenme yeteneklerine önemli bir etki eder.

Makine öğrenimi, şirketlere birçok açıdan yardımcı olur. Büyümeyi destekleyecek çıktılar üretebilir, gelir akışlarını ortaya çıkarabilir ve zorlu sorunları çözmeye yardımcı olabilir. Şirketler, büyük miktardaki verilerini hızlı bir şekilde analiz ederek istenen sonuca hızla ulaşabilirler.Makine öğrenimş algoritmaları problemin girdi ve çıktı değerlerine göre farklı gruplara ayrılmaktadır:

1. Denetimli makine öğrenimi
2. Denetimsiz makine öğrenimi
3. Pekiştirmeli makine öğrenimi

2.1.1 Denetimli Makine Öğrenimi

Denetimli öğrenme, verilerin bir kısmını kullanarak veriler arasındaki ilşkiyi öğrenen ve bu öğrenilen bilgiyi kullanarak kullanılmayan verileri tahmin etme veya sınıflandırma amacıyla kullanılan bir makine öğrenme yöntemidir. Temel amacı,

bilinen veri setinden elde edilen sonuçlarla bilinmeyen veri seti için etkili tahminler yapmaktır.

Denetimli öğrenmede, tahmin edilmek istenen değişkenin türüne bağlı olarak problem grupları farklılık gösterir. Eğer tahmin edilecek değişken sayısal ve sürekli bir değer ise regresyon problemi olarak ele alınırken, eğer tahmin edilecek değişken kategorik bir değer ise sınıflandırma problemi olarak ele alınır.

Sınıflandırma problemlerinde, beklenen hedef çıktı kategorik bir değişken şeklindedir ve amacımız bu kategorik değişkeni tahmin etmektir. Örneğin, belirli semptomlar gösteren bir hastanın hasta olup olmadığını tahmin etmek için bu semptomları kullanabiliriz. Benzer şekilde, bir kişinin operatör değiştirme veya değiştirmeme durumunu tahmin etmek için de çözümler üretilebilir.

Regresyon problemlerinde ise beklenen hedef çıktı sayısal ve sürekli bir değişkendir ve amacımız bu değişkeni tahmin etmektir. Örneğin, bir evin fiyatını tahmin etmek için evin konumu, oda sayısı, banyo sayısı, bahçe olup olmaması gibi değişkenleri kullanabiliriz. Bu değişkenlerle evin fiyatı arasındaki ilişkiyi kullanarak yeni bir evin fiyatını tahmin etme işlemini gerçekleştirilebilir.

Sonuç olarak, denetimli öğrenme, bağımsız değişkenleri kullanarak bağımlı değişkenleri tahmin etme işlemidir. Problemin türü, bağımlı değişkenin niteliğine göre belirlenir.

2.1.2 Denetimsiz Makine Öğrenimi

Denetimsiz öğrenme, verileri analiz ederek bu verilerle ilgili bilgileri ortaya çıkarma yaklaşımına dayanır. Bu tür çalışmalarda veriler, bağımlı veya bağımsız değişken olarak sınıflandırılmaz. Veriler, birbirleriyle olan ilişkilerine göre kümelere ayrılır. Veri kümesindeki verilerin ortak özellikleri üzerinden kümeleme işlemi yapılır ve anlamlı veriler elde edilir. Veriler, yakınlık, uzaklık, benzerlik gibi ölçütlere göre analiz edilerek kümelenir. K-ortalama algoritması bu tür problemlerde kullanılan bir yöntemdir.

Denetimsiz makine öğrenmesi algoritması sayesinde veriler arasındaki ilişkileri bulma imkanı sağlanır. Birliktelik kuralları bu tür problemlerin çözümünde kullanılır. Örneğin, bir ürün satın alan bir kişiye, başka hangi ürünleri satın alabileceği tahmin

edilebilir veya geçmiş veriler incelenerek bir kişiye bir ürün aldığında farklı bir ürünün tavsiye edilebilmesi mümkündür.

Veri seti kümelere ayrıldığında, iyi bir kümeleme elde edilip edilmediğini anlamak için kümeler dikkatlice incelenmelidir. Aynı kümeye ait veriler arasında yüksek benzerlikler olması arzulanırken, farklı kümelerdeki veriler arasında düşük benzerlikler olması beklenir.

2.1.3 Pekiştirmeli Makine Öğrenimi

Pekiştirmeli öğrenme, mevcut durumu analiz ederek çevreden gelen tepkilerle eğitilen bir makine öğrenmesi türüdür. Sistem, belirli bir hedefe veya amaca ulaşmak için kendisini yeni durumlarla eğiterek doğru sonuçlar elde etmeyi amaçlar. Doğru sonuçlar elde edildiğinde sistem ödüllendirilirken, yanlış kararlar verdiğinde cezalandırılır. Bu şekilde, en doğru kararı elde etme amacına yönelir. Bu amaca ulaşabilmek için yeterli sayıda yeni durum ile karşılaşması gereklidir.

Pekiştirmeli öğrenme yaklaşımının dört temel bileşeni bulunmaktadır.

1. Davranış prensipleri: Sistem tarafından gerçekleştirilebilecek aksiyonları belirler.
2. Ödül: Gerçekleştirilen aksiyona karşılık gelen ödül puandır.
3. Değer fonksiyonu: En yüksek ödülü almak için uzun vadeli stratejilerin temelini oluşturur. Ödülü düşük bir aksiyonun değeri yüksek olabilir. Amaç toplamda maksimum ödülü elde etmektir.
4. Model: Simülasyon yapabilme amacıyla oluşturulan bir çevre modelidir.

Pekiştirmeli öğrenme modeli canlıların öğrenme yapısına benzetilebilir. Örneğin bebek yürümeyi öğrenirken düşüp kalkarak deneyim kazanır. Her düşüşüyle birlikte biraz daha deneyim kazanarak sonunda yürümeyi başarır.

2.2 Makine Öğrenmesi Aşamaları

Makine öğrenmesi yöntemini seçmeden önce, veri bilimi projelerinde aşağıdaki adımların yapılması gerekmektedir. Bu adımlar, veriyi modele hazırlamak amacıyla gerçekleştirilen işlemleri içermektedir:

1. Verilerin toplanması ve analizi
2. Verilerin hazırlanması
3. Modelin oluşturulması
4. Sonuçların değerlendirilmesi

2.2.1 Verilerin toplanması ve analizi

Problemin ve istenilen çıktıların belirlenmesi sonucunda çalışmada kullanılacak verinin toplanması ilk aşamayı oluşturmaktadır. Veri toplanması aşamasında problemin verisi veri toplama yöntemleri kullanarak toplanabilir. Anket çalışması, mülakat, gözlem, deney, şirket raporlarının incelenmesi gibi yöntemler veri toplamak için kullanabilir (Şahinarslan, 2019). Ayrıca şirketlerden elde edilen veriler veya internet üzerinden elde edilen verilerde kullanılabilir.

Toplanan verinin ilk olarak analizi gerçekleştirilir. Verinin analizi yapılırken istatistiksel yöntemler kullanarak inceleme gerçekleştirilebilir. Veriyi anlamak makine öğrenmesi çalışmalarında önemli bir adımdır. Uygun makine öğrenmesi modeli seçmek için veriyi iyice anlamak ve analiz etmek gerekmektedir.

2.2.2 Verilerin hazırlanması

Veri analizinde genel hatlarıyla veriyi inceledikten sonra verinin çalışma için hazırlanması aşamasına geçilir. Bu aşama özellik mühendisliği olarak da adlandırılabilir. Bu aşamada ilk olarak çalışma kapsamında uygun olarak görünen değişkenlerin seçimi yapılır, veri türleri kontrol edilir, yeni değişkenler oluşturulur, aykırı değerler tespit edilir, eksik değerler doldurulur, ölçeklendirme gerektiren değişkenler uygun yöntemler ile ölçeklendirilir.

2.2.2.1 Aykırı değer analizi

Veri setinde yer alan değişkenlerin her birinin kendi içerisinde diğer değerlerle karşılaştırıldığında veri setine uygun olmadığı tespit edilen aşırı değerlere aykırı değer (outlier) denir. Bu aykırı değerler olma ihtimali çok düşük olasılıkta olan değerler içermektedir ve yapılan analizlerde analiz sonucunun sapmasına neden olabilir. Aykırı değerleri tespit etmek için çeşitli yöntemler bulunmaktadır. Aykırı değerleri tespit ederken istatistiksel yöntemlerden ve grafiksel yöntemlerden

faydalanılır. Z - skoru yöntemi istatistiksel yöntemlere örnel olarak verilebilir. Kutu grafiği yöntemi ise grafiksel yöntemlere örnek olarak verilebilir.

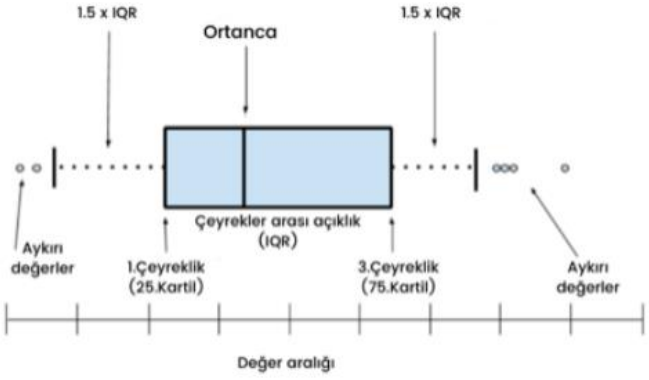
Z-skoru yöntemi normal dağılıma sahip bir veri setinde, bir gözlemin ortalamadan ne kadar standart sapma yaptığını gösterir (Dayanıklı, 2021). Veri setinde yer alan gözlemlerin, normal veya normale yakın dağılıma sahip olduğu durumlarda Z- skoru yöntemi kullanılabilir. Z-skoru gözlemden ortalama değer çıkarıldıktan sonra standart sapmaya bölünmesiyle hesaplanır.

$$Z = \frac{(x - \mu)}{\sigma} \tag{2.1}$$

Z-skoru -3 ile 3 arasında olan değerlerin normal olduğu bu sınırların dışında kalan değerlerin ise aykırı değerler olduğu kabul edilir. Aykırı değerler grafik üzerinde gözlemlenebilirler. Şekil 2.2'de gösterildiği gibi bu gözlem için kutu grafik yöntemi kullanılmaktadır. Bu yöntemde aykırı değerleri tespit etmek için aykırı değer sınırları belirlenir. Bu sınırların dışında kalan değerler aykırı değer olarak adlandırılır.

$$Alt\ Aykırı\ Değer\ Sınırı\ =\ Q1\ -\ 1.5x(Q3 - Q1) \tag{2.2}$$

$$Üst\ Aykırı\ Değer\ Sınırı\ =\ Q3\ +\ 1.5x(Q3 - Q1) \tag{2.3}$$



Şekil 2.2 : Kutu Grafik Yöntemi.

Aykırı değerler tespit edildikten sonra bu değerler üzerinde işlemler gerçekleştirilir. Aykırı değerleri ortadan kaldırmak için uygulanan en basit yöntem bu değerleri silme işlemidir. Ancak veri silme işlemleri, verinin değişkenliğini etkileyebileceğinden ilk tercih olmamalıdır. Diğer bir yöntem ise aykırı değerlerin yerine değer atama

yöntemidir. Atanacak değerler ortalama, medyan, mod gibi istatistiksel değerler kullanılarak seçilebilir. Veri setinde aykırı değer incelemesi yapıldığında sayısal değişkenler üzerinde kutu grafiği yöntemi uygulanarak grafik üzerinden aykırı değerler gözlemlenebilir.

2.2.2.2 Eksik değer analizi

Veri seti üzerinde gerçekleşen çalışmalarda, eksiksiz bir veri setine sahip olmak önemlidir. Eksik verilere sahip veri seti, analiz sonuçlarının güvenilirliğini ve tutarlılığını etkileyerek yanlış sonuçlar elde etmemize neden olabilir. Kayıp veriler, genellikle üç farklı şekilde kategorize edilir (Dayanıklı 2021).

Tamamen Rastgele Kayıp (MCAR): Eksik verilerin tamamen rastlantısal şekilde ve kontrol dışı şekilde eksik olduğu durumdur. Teknik veya insan kaynaklı hatalardan kaynaklanır.

Rastgele Kayıp (MAR): Eksik veriler ve ölçülen değerler arasında sistematik bir ilişki vardır. Eksik veriler rastgele olmuşmuş olsa da diğer ölçülen değişkenler sayesinde eksik veri tahmin edilebilir.

Rastgele Olmayan Kayıp (MNAR): Eksik veriler, değişkenin kendisiyle veya veri setinde ölçülmemiş farklı bir değişkenle ilişkilidir.

Eksik verinin neden olduğu sorunları gidermek için bazı teknikler kullanılmaktadır. Silme yöntemi, eksik verilere sahip gözlemlerin veri setinden tamamen çıkarılmasıdır. Değer atama yöntemi ise eksik verilerin basit veya model tabanlı gelişmiş teknikler kullanılarak tahmin edilen değerler ile eksikliğin giderilmesidir. Eksik değerlere ortalama, medyan, mod veya sabit bir değer atanabilir. K- en yakın komşu yöntemi (KNN) kullanılarak da eksik verilere değer ataması gerçekleştirilebilir. Eksik veriler üzerinde yapılan işlemlerin başarısını değerlendirmek önemlidir. Değişkenlerin R-kare değeri karşılaştırılarak yapılan işlemlerin başarılı bir etki oluşturup oluşturulmadığı kontrol edilmelidir.

2.2.2.3 Ölçeklendirme

Değişkenler farklı ölçeklerde olabilir. Bir değişken 1-10 arasında değer alabilirken farklı bir değişken 1-1000 arasında değer alabilir. Bu durumda yüksek değere sahip değişken diğer değişkene baskınlık sağlayabilir. Bu değerleri normal hale getirmek

ve baskınlığı azaltmak adına bazı yöntemler kullanılmaktadır. Bu yöntemler normalizasyon, standardizasyon gibi yöntemlerdir.

Normalleştirme yöntemi değişkenlerin 0 ve 1 arasındaki değerlere atanmasıdır. Burada yeni değerlerin dağılımı ile önceki halinin dağılımı benzer şekildedir.

$$X' = (X - X_{min})/(X_{max} - X_{\min)} \quad (2.4)$$

Değişkenin ortalama değeri 0 standart sapma değerinin ise 1 olacak şekilde dağılımın normale yaklaşmasını sağlayan yöntem standardizasyon yöntemidir.

2.2.3 Modelin oluşturulması

Veri hazırlama aşamasından sonra çalışma kapsamında uygulanacak makine öğrenmesi algoritmasının seçilmesi aşaması gelmektedir. Veri setine en uygun olabilecek algoritma seçilerek makine öğrenmesi modeli kurulur. Tek bir algoritma tüm çalışmalar için uygun çözüm bulamaz. Her bir algoritmanın her bir veri setinde birbirine üstünlük sağladığı alanları bulunmaktadır (Wolpert & Macready, 1997)

Veri setini en iyi açıklayan algoritmayı bulmak adına birçok algoritma bu aşamada denenebilir. Sonrasında en uygun algoritma seçilerek çalışmaya devam edilebilir. Çalışma için en uygun modeli seçerken modelin performansına bakılması gerekmektedir. Modelin performansına bakmak için veri setini ayrıştırarak test edilmesi gerekiyor. İki farklı yöntem ile veri seti ayırmak mümkündür. Hold-out yöntemi ve cross-validation yöntemi bu yöntemlerdir.

Hold-out yöntemi veri setini 2 veya 3 parçaya ayırmayı sağlayan yöntemdir. Veri setini 3 ayrı parçaya ayırdığımızda modelin çalıştırıldığı kısım train, iyileştirildiği kısım validation ve modelin test edildiği kısım test kısmıdır. Veri setinin parçalarına ayrılması Şekil 2.3'te gösterilmiştir.

Şekil 2.3 : Hold-out Yöntemi.

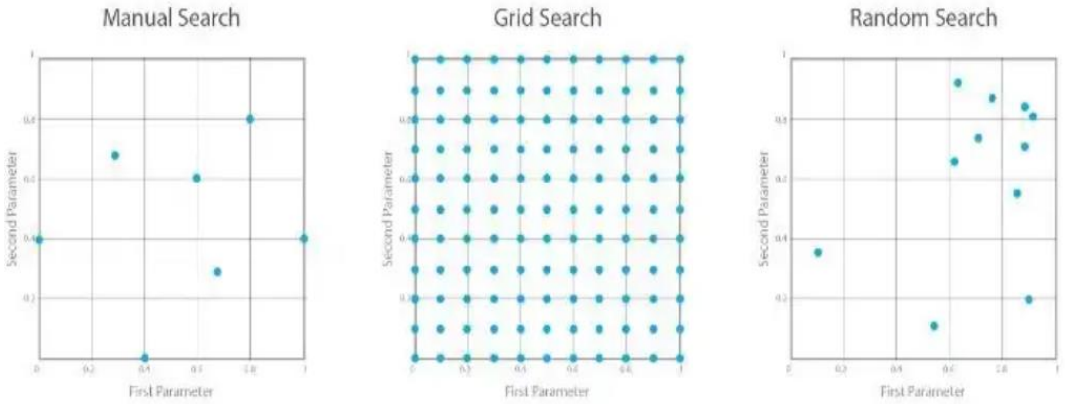
Train kısmında makine öğrenmesi algoritmaları uygulanarak uygun olan algoritma seçilir. Bu veri setinde modelin veriyi anlayabilmesi için veri setinin büyük kısmı burada yer alır. Veri setinin en az %60 ı burada yer almalıdır.

Validation kısmı train kısmı içerisinden seçilir. Buradaki amaç uygulanan makine öğrenmesi modelinin iyileştirilmesini sağlamaktır. Modelin iyileştirilmesi için hiper parametre optimizasyonu uygulanarak en uygun parametreler seçilerek modelin performansı artırılabilir. Validation kısmı train kısmındaki veriden seçildiği için az miktarda seçilebilir.

Her makine öğrenmesi modelinin kendine ait parametreleri bulunmaktadır. Probleme veri setine bağlı olarak parametrelerin alması gereken değerler değişkenlik göstermektedir. Uygun parametre değerlerini seçmek için hiper parametre optimizasyonu yapılmaktadır. Parametreler manuel seçilebileceği gibi seçmek için kullanılan yöntemlerde bulunmaktadır. Grid Search ve Random Search uygulanan yöntemlere örnek olarak verilebilir. Hiper parametre optimizasyon yöntemlerinin çalışma manıtğı Şekil 2.4'te gösterilmiştir.

Grid search ile uygun parametre değerlerini bulurken parametreler için uygun aralıklar belirleyerek o belirlenen aralıklardaki en iyi kombinasyon oluşturularak değerleri seçmemizi sağlar.

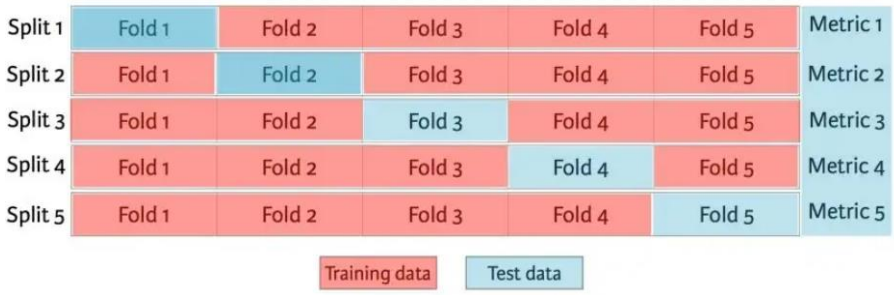
Random search ile uygun parametre değerlerini bulmak için grid search yönteminde yapılması beklenen gibi parametreler için uygun değer aralıkları belirlenir. Bu yöntemde değerlerin her birinin denenmesi yerine rastgele seçilen değerler denenerek uygun olan kombinasyon oluşturulur.



Şekil 2.4 : Hiper Parametre Optimizasyonu Yöntemleri.

Test kısmında makine öğrenmesi algoritması modeli train kısmıyla öğrendiğinde modelin başarısını, ne kadar doğru çalıştığını ölçmek için modeli kurarken modelin görmediği bir veri seti üzerinden modelin başarısının ölçülmesi gerekmektedir. Test kısmı ile veri setindeki değişkenler üzerinden hedef değişkenini tahmin etmesi sağlanır ve gerçek hedef değişkeni ile karşılaştırılması gerçekleştirilir. Böylece kurulan modelin ne kadar performanslı çalışıp çalışmadığı gözlemlenmektedir.

Cross validation yöntemi veri kümesini rastgele k tane gruba ayırma yöntemidir. Gruplardan bir tanesi test veri seti olarak kullanılırken geriye kalan gruplar train veri seti olarak kullanılır. Her bir grup için model tekrarlanarak eğitilir ve test grubu ile test edilir. Cross validation yönteminin görselleştirilmesi Şekil 2.5'te gösterilmiştir.



Şekil 2.5 : Cross Validation Yöntemi.

Örneğin, veri seti 5 gruba ayrıldığında model 5 kere eğitilip test edilmiş olacaktır. Her modelde test veri seti değişkenlik gösterdiğinden modelin daha iyi sonuç vermesi beklenmektedir.

Modelin performansını değerlendirirken bakılması gereken değerlendirme kriterleri bulunmaktadır.

Makine öğrenmesi algoritmalarından denetimli öğrenme yöntemlerinden regresyon ve sınıflandırma yöntemlerinin değerlendirme kriterleri birbirinden farklıdır.

Regresyon modelini değerlendirirken kullanılan değerlendirme kriterlerine R^2, *Adjusted* R^2, *MAE, MSE* örnek olarak verilebilir.

R^2, veri setinde bulunan değişkenlerin modele olan uyumunun göstermektedir. R^2 değerinin 1'e yakın bir değer olması modelin performansını yüksek olduğunu gösterirken aynı zamanda yüksek R^2 değeri modelin overfitting durumu olduğunu gösterebilmektedir.

$$R^2 = 1 - \frac{Explained\ Variation}{Total\ Variation} \tag{2.5}$$

Modele eklenen her değişken modeli karmaşıklaştırabilir. Bu durumda overfittinge yol açmaktadır. Bu durumun önüne geçebilmek adına Adjusted R^2 değerine bakılmalıdır. İki metriğin birbirinden farkı ise gereksiz eklenen değişkenlerin Adjusted R^2 hesaplanırken cezalandırılıyor olmasıdır.

$$Adjusted\ R^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1} \tag{2.5}$$

Mutlak hata değerini hesaplarken gerçek değerlerden tahmin edilen değerlerin çıkarılması ile oluşan farklardır. Mutlak hatanın düşük olması modelin iyi bir performansa sahip olduğunu göstermektedir.

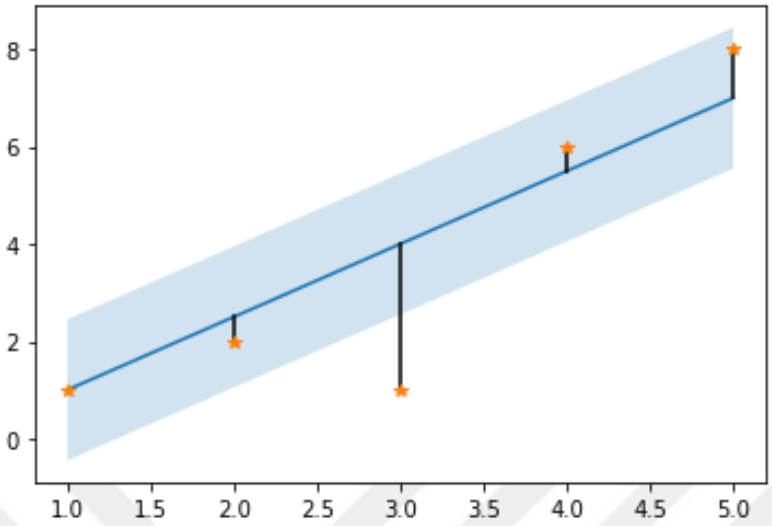
$$MAE = \frac{1}{N}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2.6}$$

Ortalama kare hata değerini hesaplarken tüm veri setinde örnek başına ortalama kare kaybını kullanır. Örnekler için hesaplanan tüm kare kayıplarının toplamının örnek sayısını bölünmesiyle hesaplanır.

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2.7}$$

Hatanın ortalama karekökü değeri artık değer olarak değerlendirilen tahmin hatalarının standart sapmasıdır. Artık değerler veri noktalarının regresyon çizgisine

olan uzaklığın ölçüsünü göstermektedir. Artık değerlerin gösterimi Şekil 2.6'da gösterilmiştir.



Şekil 2.6: Artık Değer Gösterimi.

$$RMSE = \sqrt{MSE} \tag{2.8}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2.9}$$

Ortalama mutlak yüzde sapma değeri sistemin doğruluğunu göstermektedir. Doğruluk değerini yüzde olarak ölçer ve gerçek değerden tahmin edilen değerin çıkarılması ve gerçek degree bölünmesi ile hesaplanır. Regresyon analizinde ve model değerlendirmesinde kullanılır.

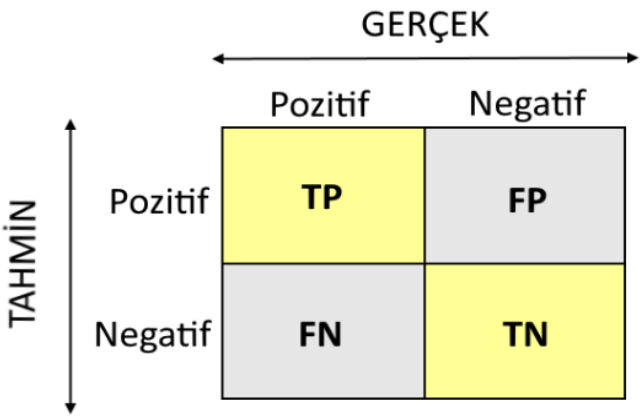
$$MAPE = \frac{1}{N}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{2.10}$$

Sınıflandırma modelini değerlendirirken kullanılan değerlendirme kriterlerine Karışıklık matrisi (Confusion Matrix), F1-score, Auc-Roc Curve örnek olarak verilebilir.

Karışıklık matrisi performans ölçüm metriğidir. Tahmin edilen ve gerçek değerlerin 4 farklı gruba ayrılmış şekilde matris halinde göstermeyi sağlar. Matrisi

anlayabilmek için bilinmesi gereken bazı terimler bulunmaktadır. Bu terimlerin matris hali Şekil 2.7'de gösterilmiştir.

- TP (True Positive- Doğru Pozitif): 1 olarak sınıflandırılan ve gerçekte de 1 olan değerlerin sayısıdır.
- FP (False Positive- Yanlış Pozitif): 1 olarak sınıflandırılan ve gerçekte de 0 olan değerlerin sayısıdır.
- TN (True Negative - Doğru Negatif): 0 olarak sınıflandırılan ve gerçekte de 0 olan değerlerin sayısıdır.
- FN (False Negative - Yanlış Negatif): 0 olarak sınıflandırılan ve gerçekte de 1 olan değerlerin sayısıdır.



Şekil 2.7 : Karışıklık Matrisi.

Güvenilir sonuçlar elde etmek için karışıklık matrisi içindeki değerler kullanılarak bazı performans ölçüm metrikleri hesaplanır.

- Kesinlik (Precision): Doğru sınıflandırılan verilerin oranını verir.
- Duyarlılık (Recall): Sadece pozitif değerlerden doğru sınıflandırılanların oranını verir.
- Doğruluk (Accuracy): Doğruluk değeridir.
- F1-score: Precision ve Recall değerlerinin harmonik ortalamasını verir. 0-1 aralığında değer alır.

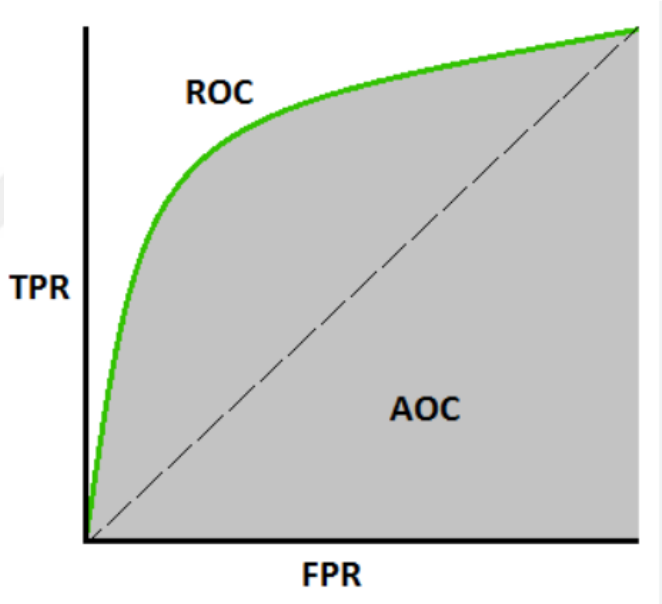
$$Precision = \frac{TP}{TP + FP} \tag{2.11}$$

$$Recall = \frac{TP}{TP + FN} \tag{2.12}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{2.13}$$

$$F1_{Score} = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{2.14}$$

Auc-Roc eğrisi performans ölçme yönteminde Roc olasılık eğrisidir. Auc ayrılabilirliğin ölçüsünü temsil etmektedir. Auc yüksek olması modelin iyi bir tahminleme yaptığının göstergesidir. Auc-Roc eğrisinin grafiksel gösterimi Şekil 2.8'de gösterilmiştir.



Şekil 2.8 : Auc-Roc Eğrisi.

2.2.4 Sonuçların değerlendirilmesi

Modellerin çalıştırılması sonucu bazen veri üzerinde işlemler yapılması gerekebilir. Bu aşamada geriye dönük adımlar tekrarlanarak model tekrardan kurulabilir. Modellerin performans ölçütleri ile değerlendirildiğinde en iyi sonuç elde edilen model çalışma için kullanılır. Kurulan model, başlangıçta tanımlanan problemin çözümü için kullanılır.

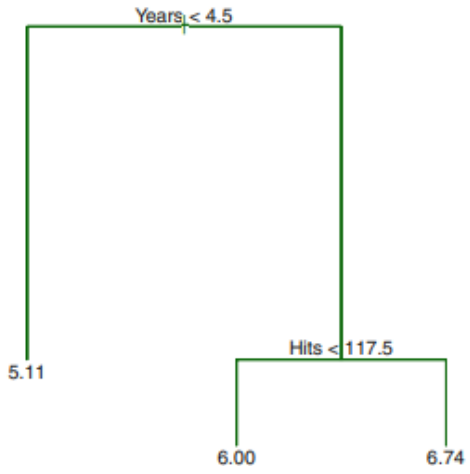
3. MAKİNE ÖĞRENMESİ ALGORİTMALARI

3.1 Karar Ağacı Algoritması

Karar ağaçı sınıflandırma problemlerinde ve regresyon problemlerinde kullanılan denetimli öğrenme algoritmasıdır. Bu algoritma meydana gelen olayların olasılığına dayalı olarak değerlendirilecek çözümleri olasılık hesabı ve ağaç benzeri grafik ile karşılaştırarak göreli optimum çözümü tahmin eder (Zhou, Sun, Fu, Jiang & Xue, 2019).

Karar ağaçları kök değişken ile başlar. Veriler kök değişkeninden sonra düğüm noktalarında karar verme işlemi gerçekleştirdikten sonra dallara ayrılır. En son aşamada yapraklara ulaşır. Yapraklar çıktı değerlerini ifade etmektedir.

Karar kuralları girdi ve çıktı özellikleri arasındaki ilişkiye bakılarak belirlenir. Karar ağacı algoritması veri kümesini karar kurallarına göre temel olarak birkaç dallara Şekil 3.1'de gösterildiği gibi ayırır (James, 2013).



Şekil 3.1 : Karar Ağacı.

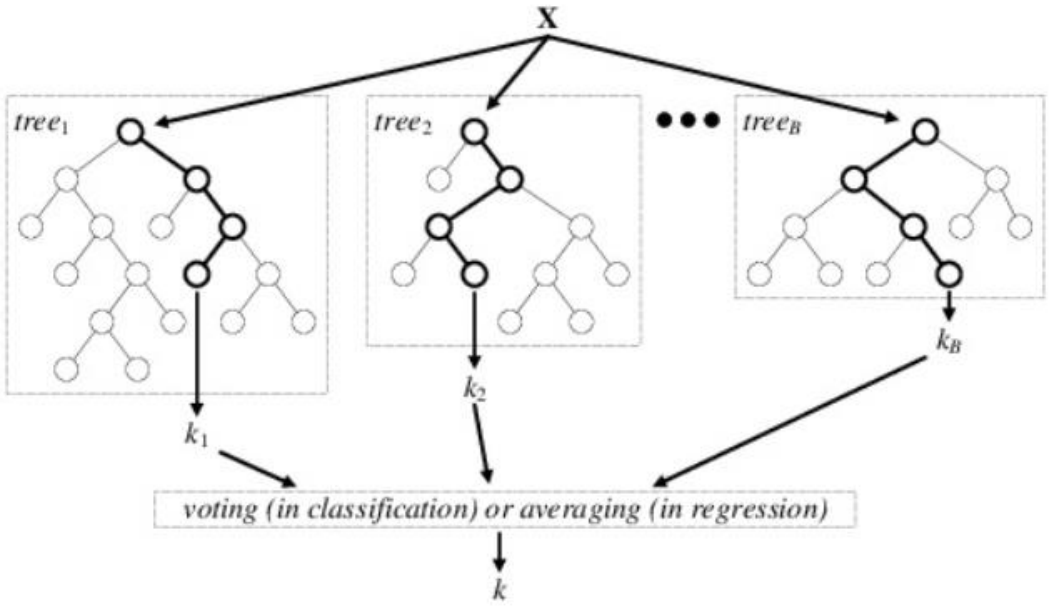
Şekilde yer alan karar ağacını incelediğimizde; beyzbol oyuncusunun büyük liglerde oynadığı yıl sayısına ve bir önceki yılda yaptığı vuruş sayısına bağlı olarak günlük maaşını tahmin etmeye yarayan regresyon modeli yer almaktadır. Her yapraktaki değer oraya düşen gözlemlere verilen yanıtın ortalamasıdır.

3.2 Rastgele Orman Algoritması

Rastgele orman yöntemi büyük verilerin sınıflandırılmasında kullanılan algoritmadır. Rastgele Orman (RF), önyükleme toplama (torbalama) yöntemi ve rastgele öznitelik seçimi uygulanarak Sınıflandırma ve Regresyon Ağacı (CART) yönteminin geliştirilmiş halidir (Saputra, Suharjito, 2019).

Rastgele Orman algoritmasının uygulanması için iki parametrenin başlangıçta belirlenmesi gerekmektedir. Bu parametrelerden ilki orman içinde yer alması istenilen ağaç sayısı bilgisidir. Diğer parametre ise her bir düğümde olması istenilen değişken değeridir. Veri seti belirlenen ağaç sayısı kadar alt kümeye ayrılır. Ardından her bir alt küme için kök düğüm ve dallanma oluşturulur. Oluşan ağaçların çözümler oylama yoluyla veya çözümlerin ortalaması alınarak sonuç belirlenir.

Rastgele orman yönteminin Şekil 3.2'de gösterildiği gibi birden fazla karar ağaç üretir. Birden fazla karar ağaçları bir araya gelerek karar ormanını oluşturur. Her bir ağaçtan elde edilen sonuçlar bir araya getirilerek en son tahmin yapılır.



Şekil 3.2 : Rastgele Orman Algoritması.

3.3 Gradyan Arttırma Algoritması

Gradient Boosting algoritması özellikle büyük ve karmaşık veri kümelerinde tahmin hızı ve doğruluğu ile öne çıkan bir yöntemdir. Algoritma önyargı hatasını en aza indirmemize yardımcı olur.

Bu algoritma, makine öğrenmesinde tahminleri sırayla yapan bir çeşit topluluk algoritmasıdır. Algoritmanın amacı zayıf olarak gördüğü tahminleri kademeli olarak güçlü tahminlere dönüşmesini sağlamaktır.

Gradient Boosting algoritmasında ilk iterasyonda tahminleri üreten bir fonksiyon oluşturulur. Tahmin edilen değerler ile hedef değerler arasındaki fark hesaplanır. Hesaplanan fark değerlerini içeren "Loss" fonksiyonunu oluşturulur. İkinci iterasyonda ilk iterasyonda oluşturulan fonksiyon ve "Loss" fonksiyonu birleştirilir. Tahmin edilen değerler ile hedef değerler arasındaki farklar tekrardan hesaplanır. Algoritma bu şekilde çalışmaya devam eder. Bu şekilde fonksiyonun başarısının artması ve tahmin edilen değerler ile hedef değişkeni arasındaki farkları minimize etmeye çalışılır.

3.4 Ekstrem Gradyan Arttırma Algoritması

Ekstrem Gradyan Artırma (Extreme Gradient Boosting, XGBoost), Gradient Boosting algoritmasının çeşitli düzenlemeler ile optimize edilmiş daha performanslı çalışma sağlayan halidir. Yüksek tahmin gücü elde edebilmesi, aşırı öğrenmenin önüne geçebilmesi, boş verileri yönetebilmesi ve bunları hızlı yapabilmesi algoritmanın en önemli özellikleridir.

Ekstrem Gradyan Artırma algoritmasında ilk adımda ilk tahmin değerini (base score) belirlemek gerekmektedir. Belirlenen tahmin değeri sonraki adımlarda gerçekleşecek olan işlemler ile yakınsanacak bir değer olmalıdır.

Yapılan tahmin işleminin ne kadar iyi sonuç verdiği modelin hata değerine bakarak yorum yapılabilir. Hata değeri gözlemlenen değerlerden tahmin edilen değerlerin çıkarılması ile elde edilmektedir.

3.5 Kategori Artırma Algoritması

Kategori Artırma (Category Boosting, CatBoost) ağaç tabanlı entegrasyon algoritmasıdır. Değişkenleri filtrelemeye ve hangi değişkenlerin tahmin üzerinde önemli bir etkiye sahip olduğunu görmemize yardımcı olur. Modelde yer alan değişkenlerin önemini ortaya çıkarmaktadır.

CatBoost algoritması, temeli simetrik karar ağaçlarına (oblivious trees) dayanan yeni bir gradyan artırma algoritmasıdır. Algoritma sayesinde çok derin ağaçlar kurmaya gerek kalmadan yüksek tahmin oranı elde edilmesini sağlar. Bu sayede aşırı öğrenme sorununu ortadan kaldırır. CatBoost algoritması kategorik özelliklerle başa çıkmak için tercih edilen bir algoritmadır. Aynı zamanda gradyan sapması, tahmin sapması sorunlarının önüne geçmeyi sağlar.

3.6 Light Gradyan Artırma Algoritması

Light Gradyan Artırma algoritması karar ağaçları kullanan farklı bir gradyan artırma yöntemidir. Karar ağaçları yöntemlerinde dallanma dikey yönde gerçekleşirken, LightGBM algoritmasında dallanma yatay yönde gerçekleşir. Yaprakların delta değerleri karşılaştırılarak en büyük delta değerine sahip yaprak üzerinden yatay şekilde dallanmaya devam edilir. LightGBM algoritması kategorik değişkenlerin sayısal şekilde ifade edilmesine gerek kalmadan uygulanabilen bir algoritma olması ile tercih edilen bir algoritmadır.

4. ÖZELLİK SEÇİMİ

Özellik seçimi, makine öğrenmesi algoritmaları kullanılarak oluşturulan modellerde veriyi en iyi şekilde açıklayan değişkenlerin seçilmesi ile performasın optimize edilmesi sağlayan önemli bir adımdır. Bu nedenle, özellik seçimi makine öğrenmesinde önemli bir adım olarak görülmektedir. Özellik seçimi modelin performansını olumsuz yönde etkileyen veya modele etkisi olmayan değişkenlerin filtrelenmesini sağlar.

Özellik seçimi orjinal özelliklerin kolayca yorumlanabilen bir alt kümesini seçerek model boyutunu azaltır ve özel süreç bilgisine dayalı olarak verilerin gelişmiş bir şekilde anlaşılmasını sağlar. Geleneksel özellik seçim yaklaşımları üç kategoriye ayrılabilir: Sarmalayıcı, gömülü ve filtre tabanlı yöntemler (Zhao vd., 2022).

Özellik seçimi işlemi, makine öğrenmesi modellerinin performansını optimize etmek amacıyla değişkenlerin önem sıralamasına göre gerçekleştirilir. Bu süreç hedef değişkeni tahmin etmek ve hangi değişkenlerin önemli olduğunu belirlemek için kullanılır. Önemli olan değişkenler, modelin hedef değişkenini en çok etkileyen değişkenler olarak kabul edilir.

Veri setinde yer alan her bir değişkenin önemini belirlemek için özellik önemi kullanılır. Değişkenlerin önem sıralaması modelin performansına olan katkılarına göre oluşturulur. Bu sıralama daha sonra elenmesi gereken değişkenlerin belirlenmesi ve daha performanslı modellerin oluşturulması için kullanılır. Özellik seçimi, elde edilen önem sıralamasına dayanarak gerçekleştirilir ve bu sayede modelin performansı artırılır.

Özellik önemi çeşitli uygulama alanlarında kullanılabilir. Örneğin, tıp alanında, bir hastalığın tanısında önemli olabilecek değişkenlerin belirlenmesinde kullanılabilir. Bu sayede hangi değişkenlerin hastalığın teşhisinde en etkili olduğu belirlenebilir. Finansal analizlerde yatırım stratejilerini belirlerken önemli olabilecek değişkenlerin seçimi için kullanılabilir. Bu, yatırımcıların hangi faktörlerin en büyük etkiye sahip olduğunu anlamalarına yardımcı olur. Pazarlama alanında, müşteri segmentasyonu

ve hedefleme çalışmalarında, belirli değişkenlerin belirleyici olduğu müşteri gruplarının belirlenmesinde önemli olabilir.

Özellik önemi analizi, makine öğrenmesi modellerinin performansını artırmak ve veri setindeki önemli değişkenleri tanımlamak için kullanılır. Bu analiz, modelin genel anlamda daha iyi performans göstermesini sağlarken, aynı zamanda modele etkisi olmayan veya zayıf etkiye sahip değişkenlerin yükünü azaltır. Bu sayede daha etkili ve anlamlı sonuçlar elde etmek mümkün olur.

Özellik seçimi, açıklanabilirlik kolaylığı sunması, öğrenme verimliliğini artırması, tahmine dayalı iyileşme sağlaması ve etkili veri toplama imkanı sunması gibi avantajlarla öne çıkar. Ayrıca modelin hedef değişkeni üzerinde önemsiz olduğu düşünülen özelliklerin elemesi, makine öğrenmesinde önemli bir bileşen olarak kabul edilir. Özellik seçimi, modelin açıklanabilirliğini artırırken aynı zamanda gereksiz özelliklerin çıkarılmasıyla öğrenme verimliliğini artırır. Bu süreç tahminlerin doğruluğunu artırırken daha az veri kullanarak daha etkili sonuçlar elde etmeyi sağlar (Zhao vd., 2022).

Özellik seçimi farklı makine öğrenmesi algoritmaları ve teknikleri kullanılarak hesaplanabilir. Örneğin, Karar Ağacı, Random Forest, XGBoost ve Gradient Boosting algoritmaları gibi algoritmalar özellik önemini hesaplamak için kullanılabilir.

Özellik seçimi duyarlılık analizi çalışmalarında da önemli rol oynamaktadır. Duyarlılık analizi bir değişkenin hedef değişkeni üzerindeki etkisini ölçerek değişkenler arasındaki ilişkiyi anlamamıza yardımcı olur. Bu analiz, bir değişkenin hedef değişkenin üzerindeki etkisini nicel olarak görmemizi sağlar. Önemli ve etkili değişkenleri seçmemizde yol gösterir. Özellik seçimi kurulan modelin performansını artırırken aynı zamanda modele etkisi olmayan değişkenlerin yükünü azaltır.

Sonuç olarak özellik seçimi analizi, veri kümesindeki değişkenlerin hedef değişkene olan etkisini ölçerek makine öğrenmesi modelinin performansını iyileştirmek ve veri kümesindeki önemli değişkenleri tanımlamak amacıyla kullanılmaktadır. Bu işlemi gerçekleştirmek için çeşitli yöntemler ve algoritmalar uygulamaktadır.

5. DUYARLILIK ANALİZİ

Duyarlılık analizi, bir sistem veya sürecin belirli bir girdi veya koşul değişikliğine nasıl tepki vereceğini belirlemek için kullanılan bir yöntemdir. Bu analiz sistemin hassasiyetini ve direncini değerlendirerek potansiyel riskleri ve fırsatları ortaya çıkarmayı amaçlar.

Duyarlılık analizi genellikle bir model veya simülasyon yoluyla gerçekleştirilir. Bu modeller, sistemin davranışını matematiksel denklemler veya kural tabanlı yaklaşımlar kullanarak tanımlar. Bu modeller, sistemin davranışını değiştiren faktörleri ve bunların etkilerini hesaba katarak sistemin duyarlılığını belirler.

Duyarlılık analizi klasik modelleme ve simülasyon alanında yaygın olarak kullanılmaktadır. Model çıktılarındaki belirsizliklerin kaynaklarını ortaya çıkarır. Aynı zamanda parametrelerin önem sıralamasını belirlemeye, önemsiz olanları elemeye ve yüksek varyanslı alt uzayları belirlemeye yardımcı olabilir (Barabanova vd., 2021).

Duyarlılık analizi çeşitli disiplinlerde kullanılır. Örneğin ekonomik bir modelin duyarlılık analizi, belirli bir vergi artışının veya faiz oranı değişikliğinin ekonomiyi nasıl etkileyeceğini belirlemek için kullanılabilir. Benzer şekilde bir mühendislik projesinin duyarlılık analizi, belirli bir malzemenin özelliklerindeki değişikliklerin ürün performansı üzerindeki etkilerini belirleyebilir.

Duyarlılık analizi karar verme süreçlerinde de kullanılabilir. Özellikle belirsizliklerin olduğu durumlarda bir sistemin duyarlılığı, risklerin ve fırsatların belirlenmesine ve bunlara uygun stratejilerin oluşturulmasına yardımcı olabilir.

Sonuç olarak duyarlılık analizi, bir sistemin veya sürecin belirli bir değişikliğe nasıl tepki vereceğini belirlemek için kullanılan bir yöntemdir. Bu analiz bir sistemin duyarlılığına ilişkin önemli bilgiler sağlayarak, riskleri ve fırsatları belirlemeye ve karar verme süreçlerine katkıda bulunmaya yardımcı olabilir.

Makine öğrenmesi, son yıllarda birçok alanda büyük ilgi gören ve uygulamaları yaygınlaşan bir teknolojidir. Bu teknolojinin en önemli uygulama alanlarından biri de

duyarlılık analizidir. Makine öğrenmesi teknikleri büyük veri kümelerindeki desenleri belirleyerek, sistemlerin belirli değişikliklere nasıl tepki vereceğini tahmin etmeye yardımcı olabilir.

Makine öğrenimi alanında duyarlılık analizi uygulamaları da bulunmaktadır. Derin sinir ağları gibi modellerde, çeşitli mimari unsurların tahmin kalitesi üzerindeki etkisini keşfetmek için duyarlılık analizi kullanılabilir. Ayrıca makine öğrenme modelinin hiperparametrelerinin analizi ve özellik seçimi görevleri de duyarlılık analiziyle bağlantılıdır (Barabanova vd., 2021).

Duyarlılık analizi çalışması geliştirilmesi için öncelikle bir makine öğrenmesi modeli seçilmeli ve eğitilmelidir. Verilerin toplanması, işlenmesi ve modelin eğitimi için kullanılacak özelliklerin belirlenmesi modelin doğruluğu ve güvenilirliği açısından önemlidir. Daha sonra kurulan model, belirli parametrelerin değiştirilmesi durumunda ürün performansının nasıl etkileneceğini tahmin edebilmek için kullanılabilir. Bu tahminler ürün veya sistemin duyarlılığı hakkında önemli bilgiler sağlayarak, riskleri ve fırsatları belirlemeye yardımcı olabilir.

Bahsedilen çalışma birçok farklı endüstride kullanılabilir. Örneğin bir otomobil üreticisi, bu uygulamayı kullanarak araç performansını etkileyen farklı faktörleri belirleyebilir ve bu faktörlerdeki değişkenliklerin araba performansı üzerindeki etkilerini tahmin edebilir. Bu tahminler araba üreticisinin yeni araçlar tasarlarken veya mevcut araçların performansını artırmaya çalışırken doğru kararlar vermesine yardımcı olabilir.

Makine öğrenmesi teknikleri kullanılarak geliştirilen bir duyarlılık analizi uygulaması, bir ürün veya sistemdeki parametrelerin değiştirilmesi durumunda performansın nasıl etkileneceğini tahmin edebilir. Duyarlılık analizi bir sistemin veya ürünün belirli parametrelerinde yapılan değişikliklerin, performansı veya sonuçları üzerindeki etkisini tahmin etmeye yardımcı olan bir analiz yöntemidir. Bu analiz sonuçları, yapılan değişikliklerin sistemin veya ürünün performansına olan etkisini ölçmek ve değerlendirmek için kullanılabilir.

Duyarlılık analizi sonuçları, belirli bir değişkenin değerindeki değişikliğin sistemin çıktısı üzerinde ne kadar etkisi olduğunu belirleyebilir. Örneğin, bir araç üreticisi arabanın hızını artırmak için motorun gücünü artırmak istediğinde duyarlılık analizi yaparak motor gücündeki artışın aracın hızı üzerindeki etkisini tahmin edebilir. Bu

sonuçlar, araç üreticisinin motor gücünü artırıp artırmama kararını vermesinde yardımcı olabilir.

Duyarlılık analizi sonuçları, sistemin veya ürünün belirli parametrelerindeki değişikliklerin performans veya sonuçlar üzerindeki etkisini ölçmek ve karşılaştırmak için kullanılabilir. Bu sonuçlar, ürün veya sistemin performansının iyileştirilmesine veya optimize edilmesine yönelik kararların verilmesinde yardımcı olabilir.

Sonuç olarak duyarlılık analizi sonuçları, belirli bir değişkenin değerindeki değişikliğin, sistemin veya ürünün performansı üzerinde ne kadar etkisi olduğunu belirleyebilir. Bu sonuçlar, ürün veya sistemin performansını iyileştirmeye yönelik kararlar vermek veya riskleri ve fırsatları belirlemek için kullanılabilir.

Duyarlılık analizi yönteminin adımları ayrıntılı olarak açıklanmaktadır:

1. Veri toplama: İlk adım, analiz için kullanılacak verilerin toplanmasıdır. Bu veriler, değişkenlerin ve çıktıların belirlendiği bir dizi örnektir. Duyarlılık analizi için değişkenler belirli aralıklarla değiştirilerek çıktı üzerindeki etkileri ölçülmelidir. Veri toplama sürecinde değişkenlerin ve çıktıların doğru bir şekilde ölçülmesi ve kaydedilmesi önemlidir.
2. Veri ön işleme: Veriler toplandıktan sonra önceden belirlenmiş değişkenler ve çıktılar arasındaki ilişkiyi belirlemek için veri ön işleme adımı gerçekleştirilir. Bu adım verilerin temizlenmesi, normalleştirilmesi ve özellik mühendisliği gibi tekniklerle veri setinin hazırlanmasını içerir. Veri setindeki aykırı değerlerin ele alınması ve eksik verilerin dikkate alınması da bu adımda gerçekleştirilmelidir.
3. Model seçimi: Duyarlılık analizi için uygun bir makine öğrenmesi modeli seçilir. Bu model, değişkenlerin çıktı üzerindeki etkisini ölçmek için kullanılacak bir fonksiyon olmalıdır. Bu modeller arasındaki doğrusal regresyon, karar ağaçları, K-NN (en yakın komşu) ve destek vektör makineleri (SVM) gibi popüler algoritmalar bu çalışmada kullanılabilir.
4. Model eğitimi: Seçilen model önceden toplanan verilerle eğitilir. Model eğitimi değişkenlerin çıktı üzerindeki etkilerini tahmin etmek için kullanılan bir algoritmadır. Eğitim süreci, modelin doğruluğunu artırmak için hiperparametrelerin ayarlanması ile değişkenler ve çıktılar arasındaki ilişkiyi

belirlemek için modelin uygun şekilde öğrenmesini sağlamak için gerçekleştirilir. Eğitim süreci boyunca veri seti eğitim ve doğrulama kümelerine ayrılır ve modelin genelleme yeteneği değerlendirilir.

5. Duyarlılık analizi: Eğitilmiş model, değişkenlerin ve çıktıların belirlendiği veri setindeki değişikliklerin çıktı üzerindeki etkilerini tahmin etmek için kullanılır. Bu adım değişkenlerin değerlerindeki değişikliklerin performans veya sonuçlar üzerindeki etkisini ölçmek ve karşılaştırmak için kullanılır. Duyarlılık analizi sonuçları, değişkenlerin önem sırasını belirleyebilir ve hangi değişkenlerin çıktı üzerinde daha fazla etkiye sahip olduğunu gösterebilir. Ayrıca duyarlılık analizi sonuçları, farklı değişken değerleri için çıktıların tahmin edilmesinde kullanılabilir.

Duyarlılık analizi sonuçları, değişkenlerin çıktı üzerindeki etkisini belirleyebilir ve hangi değişkenlerin performans üzerinde daha büyük bir etkiye sahip olduğunu ortaya koyabilir. Bu çalışmanın sonuçları, ürün veya sistem üzerinde yapılabilecek iyileştirmeleri ve optimizasyonları belirlemek için kullanılabilir. Duyarlılık analizi sonuçları ayrıca riskleri ve fırsatları tanımlayabilir ve karar verme sürecinde önemli bir yol gösterici olarak kullanılabilir.

6. UYGULAMA

6.1 Aşamalar ve Metodoloji

Bu çalışmanın amacı, taze sebze ve meyve ürünleri satışı yapan bir e-ticaret uygulamasının verilerini kullanarak, satış adedini etkileyen değişkenleri keşfetmek ve bu değişkenlerin duyarlılık analizini gerçekleştirmektir. Çalışma kapsamında ilk olarak çeşitli makine öğrenmesi algoritmaları uygulanarak en iyi sonucu veren algoritma seçilmiş, ardından en iyi sonucu veren algoritma kullanılarak özellik seçimi uygulanmıştır. Son aşamada ise özellik seçimi sonucunda elde edilen değişkenlerle duyarlılık analizi yapılmıştır.

Çalışmada kullanılan veri seti, 2022 Ocak - 2022 Kasım ayları arasındaki 11 aylık verileri içermektedir. Veri seti, uygulama üzerinden sipariş oluşturan kullanıcıların sipariş bilgilerini içermektedir.. Hedef değişkeni olarak siparişi edilen ürünlerin adet bilgisi kullanılmıştır. Veri setinde model kurulumu öncesinde özellik mühendisliği (feature engineering) kapsamında veri ön hazırlığı adımı gerçekleştirilmiştir. Daha sonra çeşitli makine öğrenmesi algoritmaları uygulanarak en iyi sonucu veren algoritma seçilmiştir. Seçilen algoritma kullanılarak özellik seçimi (feature selection) gerçekleştirilmiştir. Bu sayede, modele etkisi bulunan değişkenler keşfedilerek, bu değişkenlerin duyarlılık analizi çalışması için kullanılması sağlanmıştır.

Bu çalışma, taze sebze ve meyve ürünleri satışı yapan bir e-ticaret uygulamasının verilerini kullanarak, satış adedini etkileyen değişkenleri belirlemeyi ve bu değişkenlerin duyarlılık analizini gerçekleştirmeyi hedeflemektedir.

6.2 Problem Tanımı

Günümüzde birçok kuruluş, şirket ve devlet kurumu dijitalleşme çağına ayak uydurabilmek için çalışmalar gerçekleştirmektedir. Kurumların hizmet kalitelerini artırması, insanlara erişmesi dijitalleşmenin sayesinde daha kolay ve hızlı bir şekilde gerçekleşmeye başlamıştır. Hızlı ve kolayca erişimin olduğu bu dönemde çoğu işlem artık online platformlar üzerinden gerçekleştirilmektedir. E-ticaret online alışveriş

veya ürünlerin online ticareti alanında kullanılmakta ve tüm dünyada dijitalleşmenin etkisiyle kullanımı büyük bir hızla artmaktadır.

Alışveriş alışkanlıklarının değişmesi ile birlikte e-ticaret uygulamaları veya siteleri üzerinden gerçekleştirilen satış veya satın alma işlemleri geçmişe göre artış göstermektedir. Şirketler bunu yeni bir satış kanalı olarak değerlendirmekte ve bu alana yönelmektedir. İnternet aracılığıyla herhangi bir yerden alışveriş yapma kolaylığına sahip olan kullanıcılar, bu şekilde alışveriş yapmaya teşvik edilmektedir.

Kullanıcılar e-ticaret platformlarında alışveriş yaparken kendilerine ait çeşitli bilgileri paylaşmaktadır. Bunlar arasında kart bilgileri, kişisel bilgiler, finansal işlem kayıtları, davranış alışkanlıkları gibi kullanıcının kayıt altına alınan önemli bilgiler yer almaktadır. Şirketler, bu verileri kullanarak çeşitli analizler yapabilmektedir. Bu analizler, kullanıcı bazlı analizlerin yanı sıra şirketin gelecek planlamaları için de değerli bilgiler sunmaktadır. Örneğin, kullanıcının sepetine eklediği ürünlerle birlikte neler alabileceğini öneren bir tavsiye sistemi oluşturulabilir veya geçmiş satışların analizi sonucunda gelecek için stratejik planlar oluşturulabilir.

E-ticaret üzerinden satış yapan birçok sektör bulunmaktadır. Bu çalışma kapsamında taze sebze ve meyve satışı gerçekleştiren e-ticaret uygulaması verileri üzerinde yapılan araştırma ele alınmaktadır. Şirketin geçmiş verileri üzerinde makine öğrenmesi algoritmaları kullanılarak, satış adedini etkileyen en önemli değişkenlerin keşfi gerçekleştirilmekte ve bu değişkenlerin duyarlılık analizi sonuçları ile ilgili fikir edinilmektedir.

6.3 Verilerin Hazırlanması ve Analiz Edilmesi

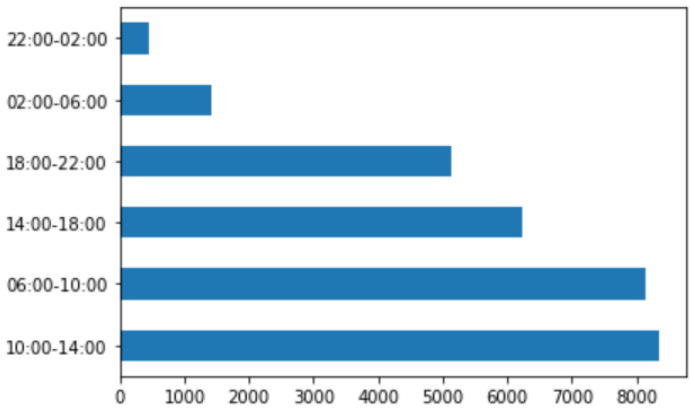
Bu çalışmada Python dili kullanılmıştır. Python dili sahip olduğu açık kütüphaneler sayesinde makine öğrenmesi algoritmalarını uygulamakta kolaylık sağlamaktadır. Bu çalışmada arayüz olarak Jupyter kullanılmıştır. Açık kaynaklı yazılım olduğundan ücretsiz olarak kurulumu gerçekleştirilmiştir.

Taze sebze ve meyve satışı yapan e-ticaret uygulaması üzerinden elde edilen veri setinde birçok değişken bulunmaktadır. İlk olarak iptal edilen siparişler veri setinden çıkarılmıştır. İstanbul satışları üzerinden bir çalışma gerçekleşeceğinden diğer illere ait sipariş bilgileri veri setinden çıkarılmıştır. Kategorik değişken olarak adlandırılan değişkenlerin aynı anlama gelen ama farklı iki grup gibi davranmasına sebep olan

yazım yanlışlıkları ortadan kaldırılmıştır. Veri seti incelendiğinde eksik değer bulunmadığından eksik veriler üzerinde gerçekleştirilmesi gereken işlemler bu çalışmada kullanılmamıştır. Hazırlanan veri setinde kullanmak üzere yeni değişkenler oluşturulmuştur. Sipariş tarihi ve teslimat tarihi arasındaki geçen gün sayısını ifade eden 'ORDER_DELIVERY_DIFF' değişkeni oluşturulmuştur. Siparişi veren kullanıcının üye olduktan sonra geçen gün sayısını ifade eden 'Customer_MembershipDay' değişkeni oluşturulmuştur.

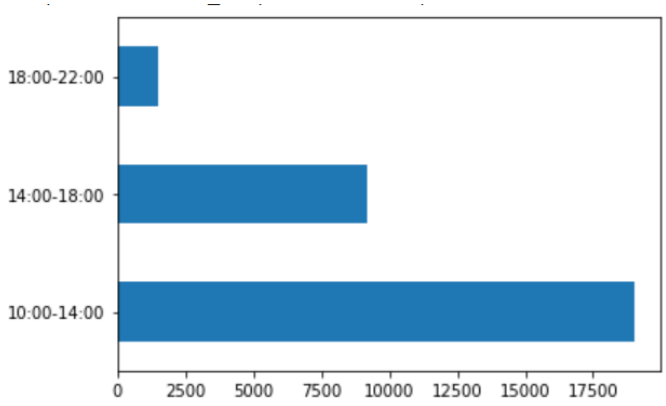
Veriye ilk baktığımızda değişkenlerin neler olduğunu nasıl bilgiler barındırdığını anlamamız gerekmektedir. Bunun için değişkenlerin davranışlarına bakılmış, ilk etapta kategorik değişkenlerin analizi gerçekleştirlmiştir.

Kullanıcıların siparişlerini verdikleri saat aralılkarına ait bir değişken bulunmaktadır. Bu değişkenin dağılımına baktığımızda sabah saatlerinde verilen siparişlerin daha fazla olduğu gözlemlenmiştir. Şekil 4.1'de görsel hali bulunmaktadır.



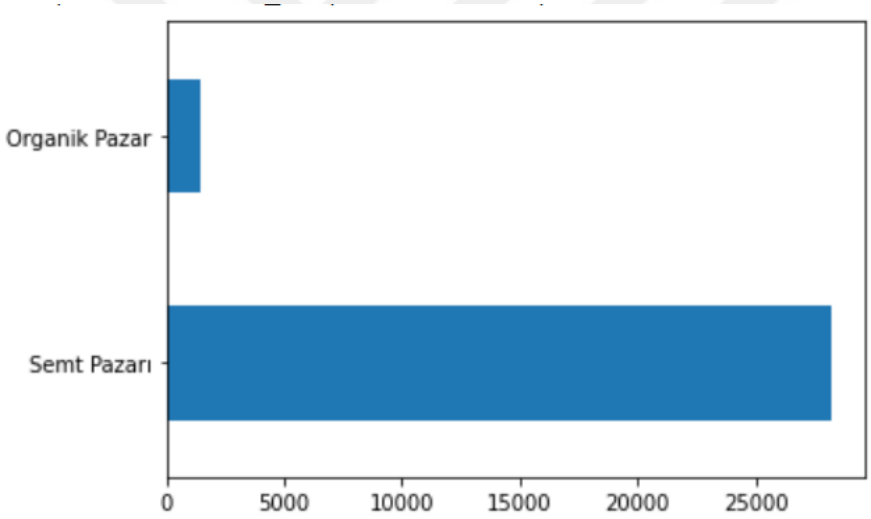
Şekil 4.1 : Siparişlerin Oluşturulduğu Saat Aralıkları.

Siparişlerin teslimat zaman aralıklarını incelediğimizde; siparişlerin büyük oranda 10:00 – 14:00 saat aralığında teslim edildiği gözlemlenmiştir. Şekil 4.2'de dağılımı gösterilmiştir.



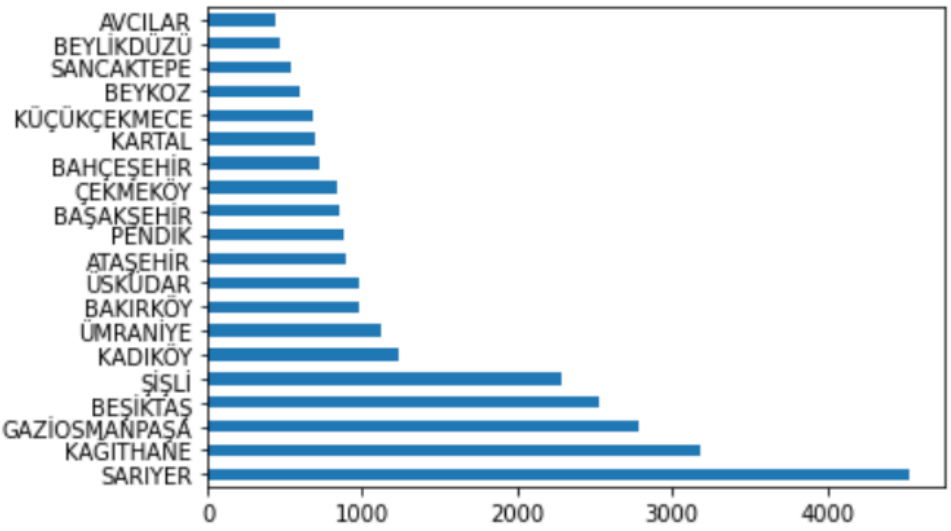
Şekil 4.2 : Siparişlerin Teslim Edildiği Saat Aralıkları.

Pazar türlerinin dağılımına bakıldığında semt pazarlarından satın alınan ürünlerin daha çoğunlukta olduğu gözlenmiştir. Organik pazarların sayısının semt pazarlarının sayısından çok daha az olmasından kaynaklı böyle bir sonucun çıktığı da söylenebilir. Şekil 4.3'de dağılımı gösterilmiştir.



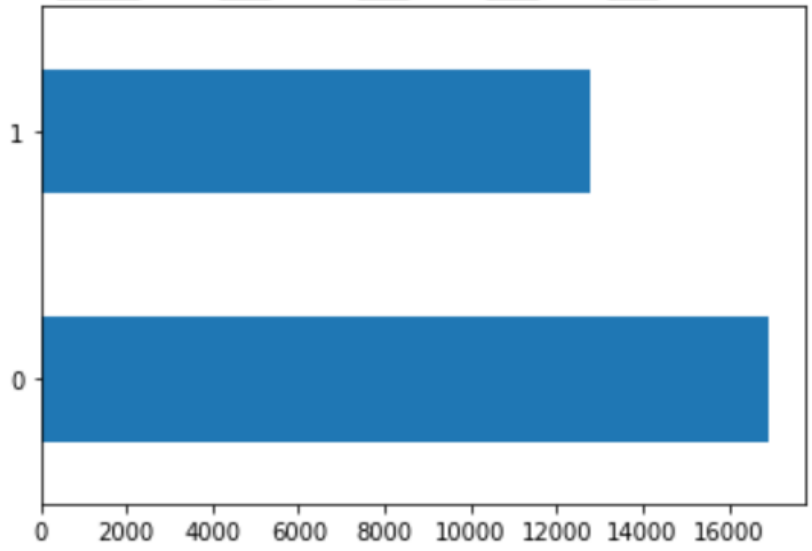
Şekil 4.3 : Pazar Türlerinin Dağılımı.

Kullanıcıların sipariş oluşturdukları bölgeleri incelediğimizde ilk 20 ilçenin dağılımı Şekil 4.4'te gösterilmiştir. Sonuçlar incelendiğinde siparişlerin büyük ölçüde Sarıyer ilçesinden verildiği gözlemlenmiştir.



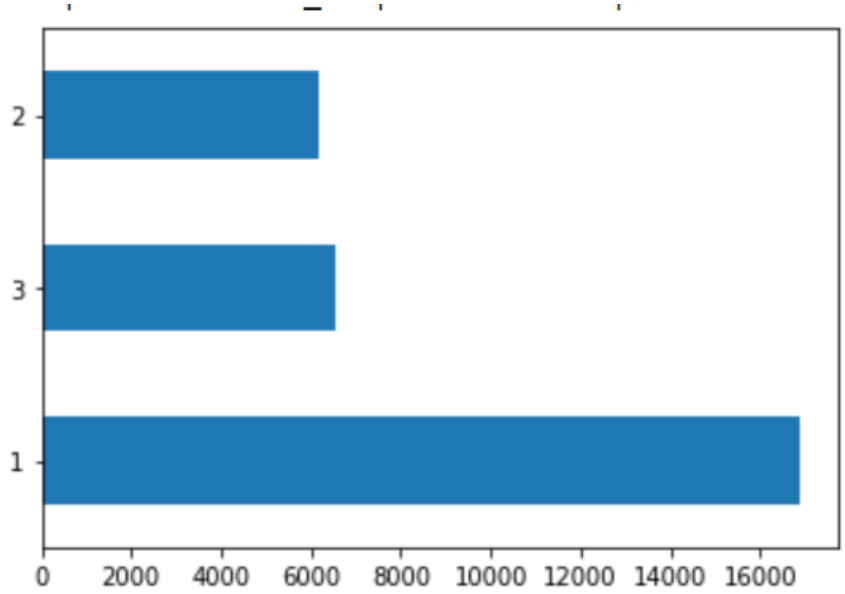
Şekil 4.4 : Siparişlerin Bölge Dağılımları.

Sipariş oluştururken kullanıcıların kendilerine tanımlanan kampanya, indirim tutarlarını kullanıp kullanmadığına bakıldığında siparişlerin %55' inde indirim kuponu kullanıldığı gözlenmiştir. Şekil 4.5'de dağılımı gösterilmiştir.



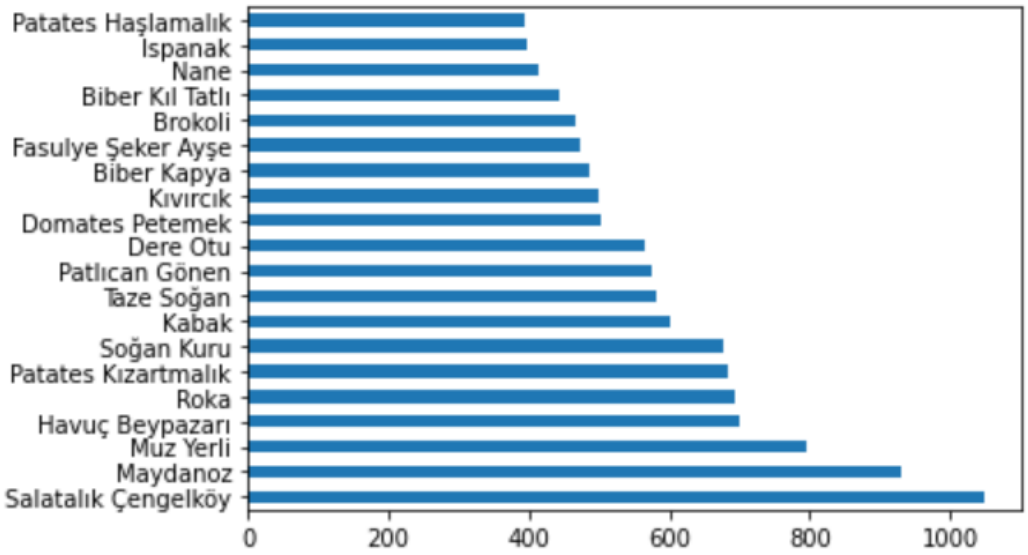
Şekil 4.5 : Siparişlerin Kampanya Dağılımı.

Satın alınan ürünlerin kategorilerinin dağılımı Şekil 4.6'da gösterilmiştir. Ürünlerin büyük bir kısmının birinci kategoride bulunan ürünlerden tercih edildiği gözlemlenmiştir.



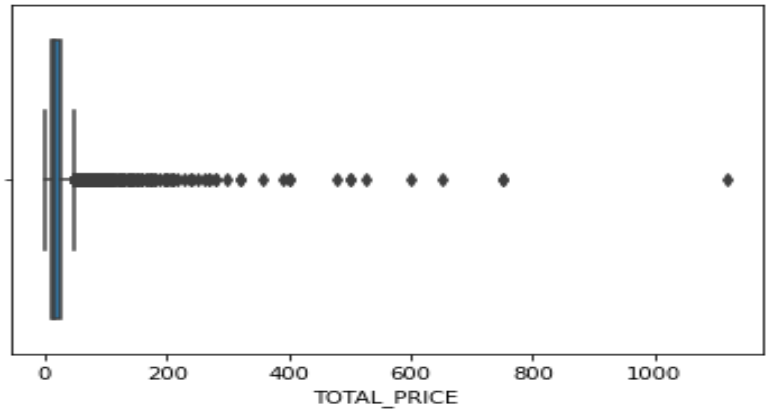
Şekil 4.6 : Ürün Kategori Dağılımı.

Satın alınan ürünlerden en çok satın alınan 20 ürün Şekil 4.7'de gösterilmiştir. En çok satın alınan ürünün salatalık olduğu gözlemlenmiştir.



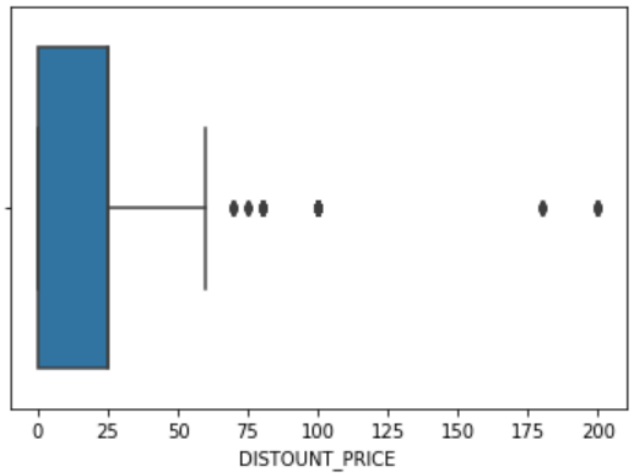
Şekil 4.7 : Sipariş Ürün Dağılımı.

Sayısal değişkenler incelendiğinde; değişkenin dağılımına bakmak ve aykırı değere sahip olup olmadığını gözlemlemek için box-plot grafiği kullanılmıştır. Satış gelirini incelemek için grafiğe baktığımızda aykırı değerlerin olduğu görülmüştür. Bu aykırı değere sahip sipariş bilgisini çıkarılarak çalışmaya devam edilmiştir. Şekil 4.8'de dağılımı gösterilmiştir.



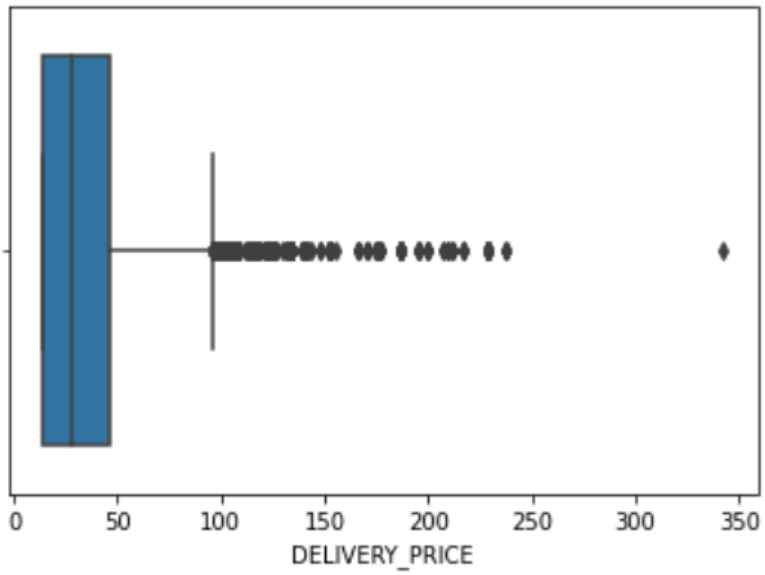
Şekil 4.8 : Satış Geliri Dağılımı.

Bir diğer sayısal değişken olan bir siparişte kupon kullanıldığında kullanılan kuponun fiyat bilgisini içeren değişkenin box-plot grafiğini incelediğimizde yüksek tutarda indirim yapılan siparişlerin olduğu görülmektedir. Bu siparişler incelendiğinde 200 TL indirim yapılan bir sipariş ve 180 TL indirim yapılan bir sipariş bulunmaktadır. Bu siparişler genel dağılımın dışında aykırı değerler olduğundan çıkarılmıştır. Şekil 4.9'da dağılımı gösterilmiştir.



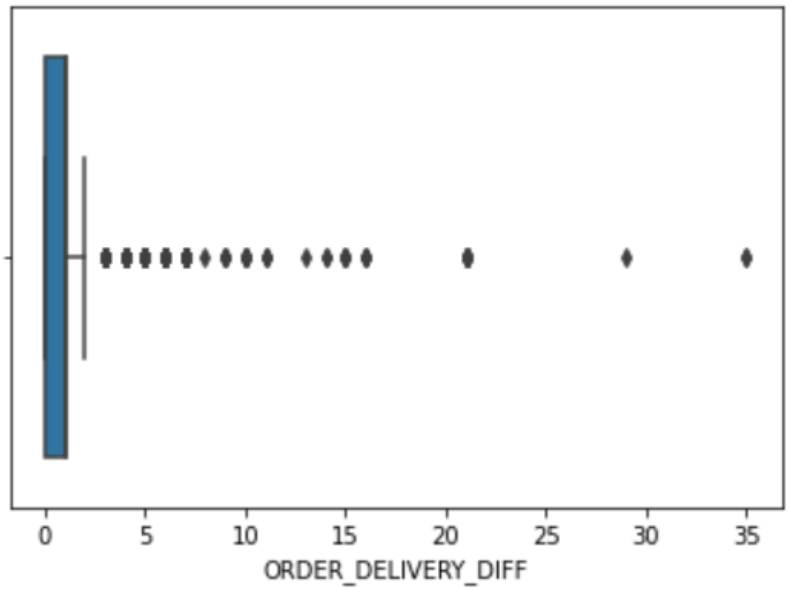
Şekil 4.9 : İndirim Tutar Dağılımı.

Teslimat için kargo firmalarına ödenen ücret incelendiğinde ortaya çıkan box-plot grafiği Şekil 4.10'da gösterilmiştir. Aykırı değer baskılama yöntemi kullanılarak düzenlenmiştir.



Şekil 4.10 : Teslimat Ücreti Dağılımı.

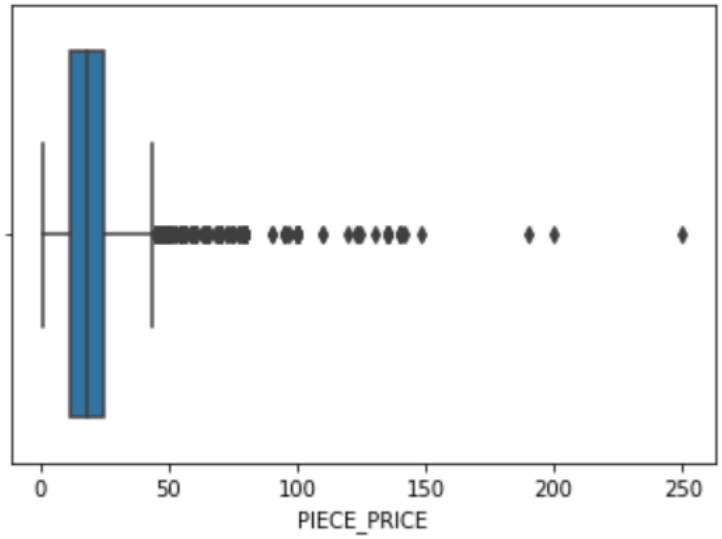
Sipariş tarihi ile teslimat tarihi arasında geçen süre hesaplanarak oluşturulan yeni değişken incelendiğinde kullanıcılar ileri tarihli teslimat seçerek sipariş oluşturabilmektedir. Bu şekilde oluşan siparişler aykırı değer olarak değerlendirilmemiş ve bu değerler için bir işlem gerçekleştirilmemiştir. Şekil 4.11'de dağılımı gösterilmiştir.



Şekil 4.11 : Sipariş ve Teslimat Arasındaki Geçen Süre.

Siparişlerde alınan ürünlerin birim fiyatları incelendiğinde yüksek fiyatlı ürünlerinde satışının gerçekleştiği görülmektedir. Yüksek fiyatlı ürün incelendiğinde bu ürünün

taze zerdeçal olduğu keşfedilmiştir. Bu değerler de aykırı değer olarak değerlendirilmemiştir. Şekil 4.12'de dağılımı gösterilmiştir.



Şekil 4.12 : Siparişlerdeki Birim Fiyat Dağılımı.

Tarihsel değişkenler de oluşturularak veri setine eklenmiştir. Bu değişkenler aşağıda belirtilmiştir.

- month: Siparişin ait olduğu ay bilgisi
- week_of_year: Siparişin yılın kaçıncı haftasına ait olduğu bilgisi
- is_wknd: Siparişin hafta sonu olup olmadığı bilgisi
- is_month_start: Siparişin ayın başlangıç gününe ait olup olmadığı bilgisi
- is_month_end: Siparişin ayın bitiş gününe ait olup olmadığı bilgisi

Veri setine ilk etapta baktığımızda; bir siparişin detayını gösterecek şekilde oluşturulduğunu görmekteyiz. Çalışma kapsamında veri setini sipariş tarihi ve ürün bazlı gruplayarak yeni bir veri seti oluşturularak model kurulması planlanmıştır. Yeni oluşturulan veri seti Şekil 4.13'te gösterilmiştir

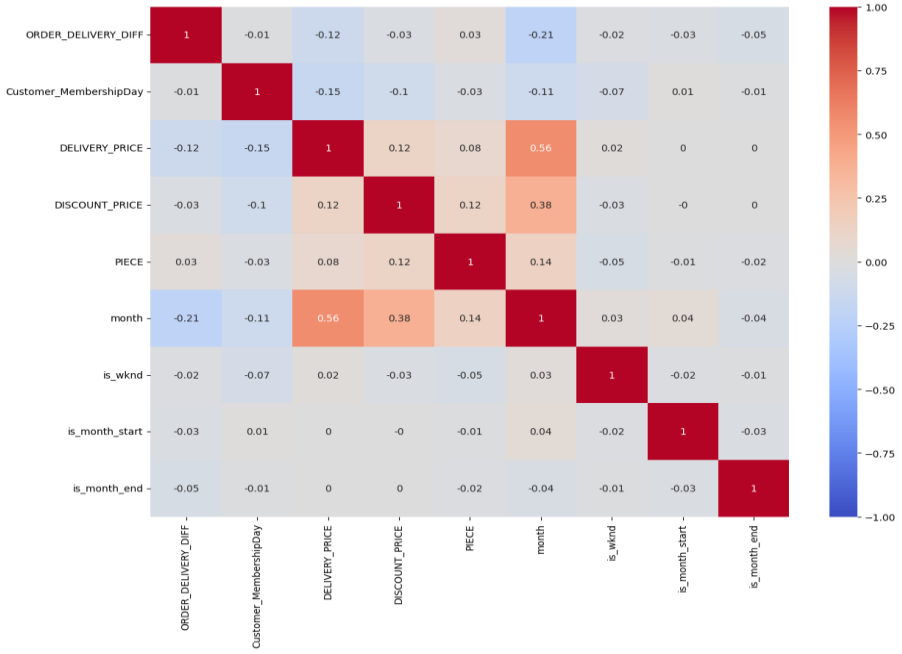
	ORDER_DATE	PRODUCT_ID	ORDER_DELIVERY_DIFF	Customer_MembershipDay	DELIVERY_PRICE	DISCOUNT_PRICE	PIECE	month	is_wknd	is_month_start	is_month_end
0	2022-01-02	43	1.0	0.0	13.900	0.0	1.0	1	1	0	0
1	2022-01-02	45	1.0	0.0	13.900	0.0	2.0	1	1	0	0
2	2022-01-02	77	1.0	0.0	13.900	0.0	1.5	1	1	0	0
3	2022-01-02	96	1.0	0.0	13.900	0.0	2.0	1	1	0	0
4	2022-01-02	109	1.0	0.0	13.900	0.0	2.0	1	1	0	0
...	...	...	...	...	...	...	...	...	...	...	...
11435	2022-11-20	148	0.0	0.0	122.890	25.0	1.0	11	1	0	0
11436	2022-11-20	149	0.0	91.0	55.950	0.0	1.0	11	1	0	0
11437	2022-11-20	150	0.0	45.5	89.420	12.5	2.0	11	1	0	0
11438	2022-11-20	168	0.0	91.0	55.950	0.0	1.0	11	1	0	0
11439	2022-11-20	170	0.0	45.5	43.905	12.5	3.0	11	1	0	0

Şekil 4.13 : Yeni Veri Seti.

Yeni veri setini oluştururken gün - ürün bazlı bir gruplama işlemi gerçekleştirildiğinden sayısal olan değişkenler üzerinden hesaplama yapılması gerekmektedir.

'ORDER_DELIVERY_DIFF', 'Customer_MembershipDay', 'DELIVERY_PRICE' ve ''DISCOUNT_PRICE' değişkenlerin ortalaması alınarak veri setine eklenmiştir. Hedef değişkeni olan 'PRICE' değişkeninin toplamı alınarak veri setine eklenmiştir.

Veri setinde bulunan değişkenlerin birbirleri ile olan istatistiksel ilişkisine korelasyon denir. Değişkenlerin birbirleri ile olan ilişkilerine bakıldığında ortaya çıkan sonuç Şekil 4.14'te yer almaktadır. Değişkenler arasındaki korelasyon değerlerine bakıldığında 0.7'den yüksek olan değer bulunmadığından; değişkenler arasında yüksek korelasyon ilişkisi yoktur.



Şekil 4.14 : Korelasyon Matrisi.

Veri setinde nümerik değişkenlerin yanı sıra kategorik değişkenler de bulunmaktadır. Bu değişkenleri sayısal biçimde ifade edebilmek için One Hot Encoding yöntemi uygulanmıştır. Bu işlem sonrasında başlangıçta 11 değişkene sahip veri seti 349 değişkene ulaşmıştır. Her ürün ayrı bir değişken olarak veri setine eklenmiştir. Veri üzerinde incelemeler ve analizler bittikten sonra, veri model kurmak için hazır hale gelmiştir.

6.4 Modelin Kurulması

Veri seti üzerinden model kurmadan önce, veri setini modelin öğrenmesi ve test etmesi için ayırma işlemi yapılması gerekmektedir. Veri setinin 2 parçaya ayrılması gerekmektedir.

Train kısmı modelin veri setini anlayacağı kısmı oluşturuyor. Test kısmı ise train ile veriyi anlayan modelin validation kısmı ile kendini iyileştirdikten sonra modelin ne kadar iyi çalıştığını ölçmek için tahminler üreteceğimiz kısmı oluşturuyor.

Bu çalışma kapsamında veri seti sipariş tarihlerine göre parçalara ayrılmıştır. 2022 Ocak-Kasım arasındaki verileri içeren veri setinde ilk 10 ay train verisi olarak kullanılmıştır. Veri setindeki Kasım ayına ait 20 günlük veri, test verisi olarak ayrılmıştır. İlk model kurulumunda birçok algoritma ile model kurma gerçekleştirilmiştir. Böylece veri setini en iyi analiz eden ve en iyi performansa sahip tahminleme çalışmasını yapan algoritmanın keşfi gerçekleştirilmiştir

İlk kurulan modeller DecisionTree, RandomForest, CatBoost, GradientBoosting, XGB'dir. 5 farklı algoritma ile modelin eğitilmesi gerçekleştirilmiştir. Model eğitimi gerçekleşirken cross validation yöntemi kullanılmıştır.

Performans skorları Çizelge 4.1'de gösterilmiştir. Skorlar karşılaştırıldığında en iyi sonucu veren algoritmanın CatBoost algoritması olduğu sonucuna ulaşılmaktadır.

Çizelge 4.1: Model Performans Skorları.

Model	RMSE
CART	2.0063
RF	1.8048
CB	1.7213
GBM	1.7597
XGBoost	1.8094

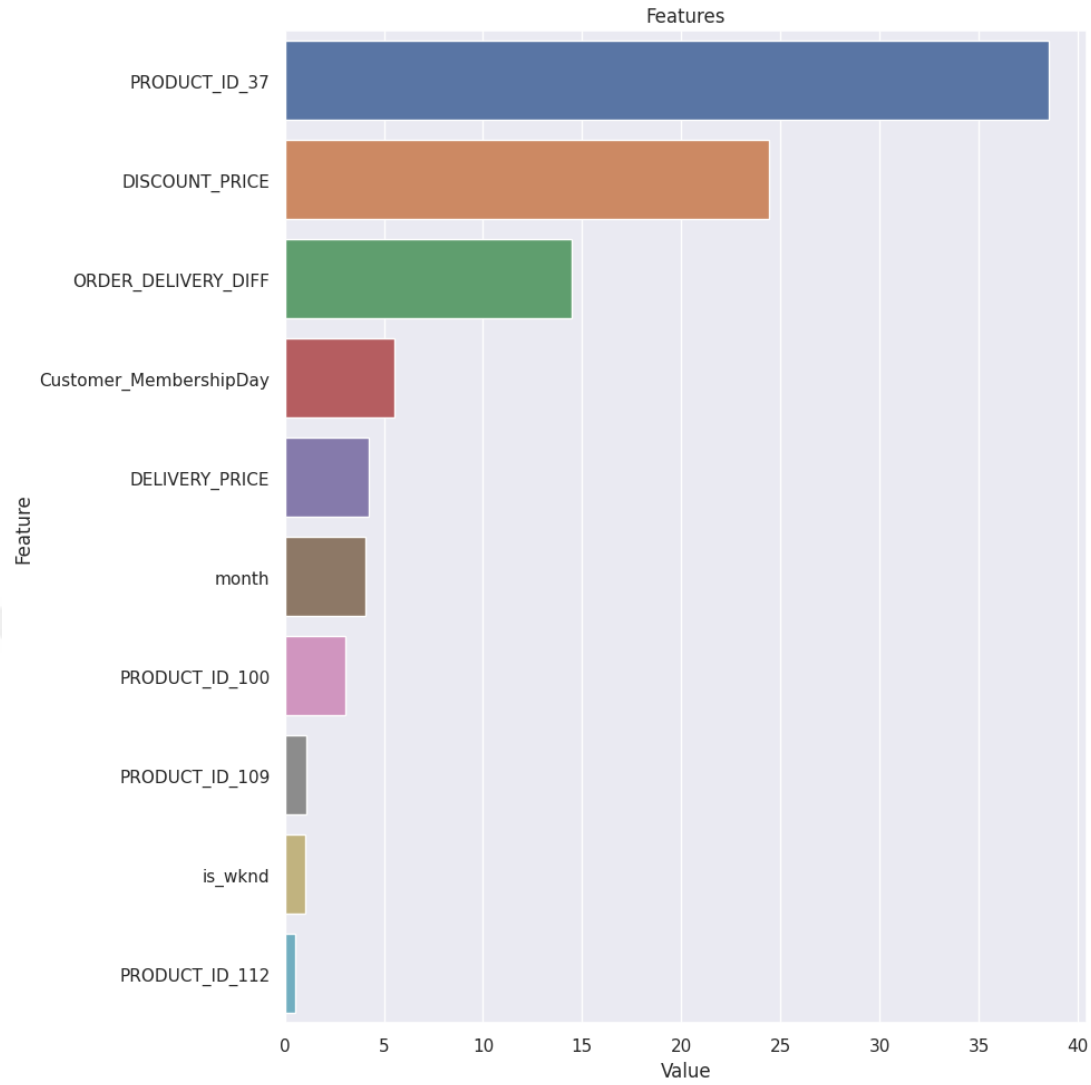
6.4.1 Özellik seçimi

En iyi performansa sahip algoritma; veri setinde yer alan değişkenlerin modeldeki önemine bakılması aşaması olan özellik seçimi aşamasında kullanılmaktadır.

İncelenen skorlar sonucunda en iyi performansa sahip Catboost algoritması kullanılarak değişkenlerin önem analizine çalışma kapsamında bakıldı. Değişkenlerin önem analizinin sonucunda oluşan ilk 10 değişken Şekil 4.15'de gösterilmiştir. Tahminleme yaparken sonucu etkileyen en önemli değişkenlerden bir tanesinin 'Product_Id_37 değişkeni olduğu gözlemlenmiştir. Bu değişken sipariş oluşturulan 37 id'li ürünü ifade etmektedir. 'Discount_Price' ve 'Order_Delivery_Diff' değişkenlerinin de önemli değişkenler olduğu sonucuna varılmıştır.

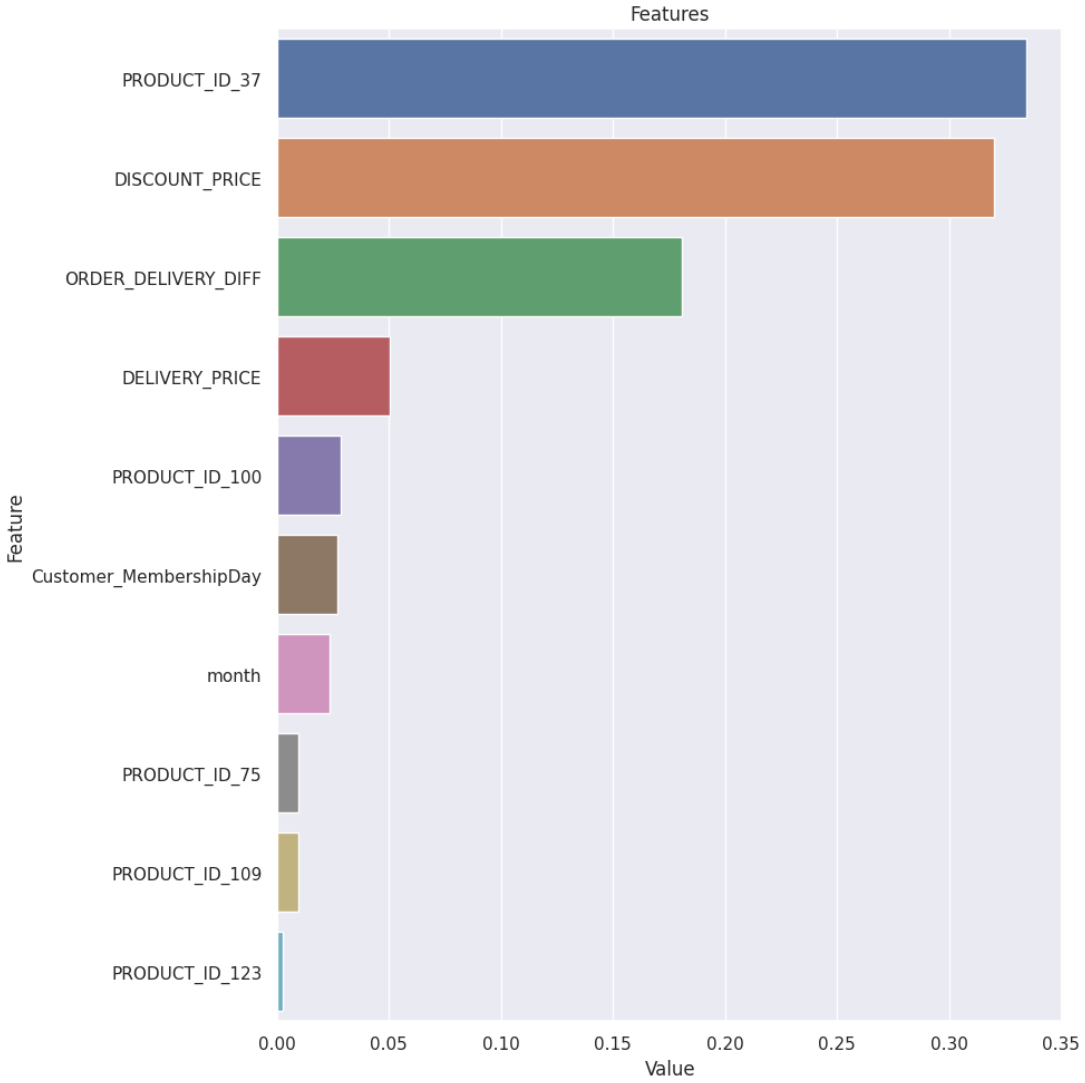
En iyi skoru veren ikinci algoritma olan Gradient Boosting algoritması kullanılarak değişkenlerin önem analizine bakıldığında ortaya çıkan önemli ilk 10 değişken Şekil 4.16'da yer almaktadır. Product_ID_37 değişkeni de bu algoritma kullanıldığında önemli olarak karşımıza çıkmıştır. Gradient Boosting algoritması uygulanan özellik seçimi ile CatBoost algoritması uygulanan özellik seçiminin sonuçları karşılaştırıldığında ortak değişkenler bulunmaktadır ancak önem sıraları algoritmalara göre değişiklik göstermiştir.

En iyi skoru veren üçüncü algoritma olan Random Forest algoritması kullanılarak değişkenlerin önem analizine bakıldığında ortaya çıkan önemli ilk 10 değişken Şekil 4.17'de yer almaktadır Bu algoritma uygulandığında en önemli değişken 'Discount_Price' değişkeni olduğu sonucu çıkmıştır. Üç algoritmanın özellik seçiminin sonuçları karşılaştırıldığında ortak değişkenler bulunmaktadır ancak önem sıraları algoritmalara göre değişiklik göstermiştir.

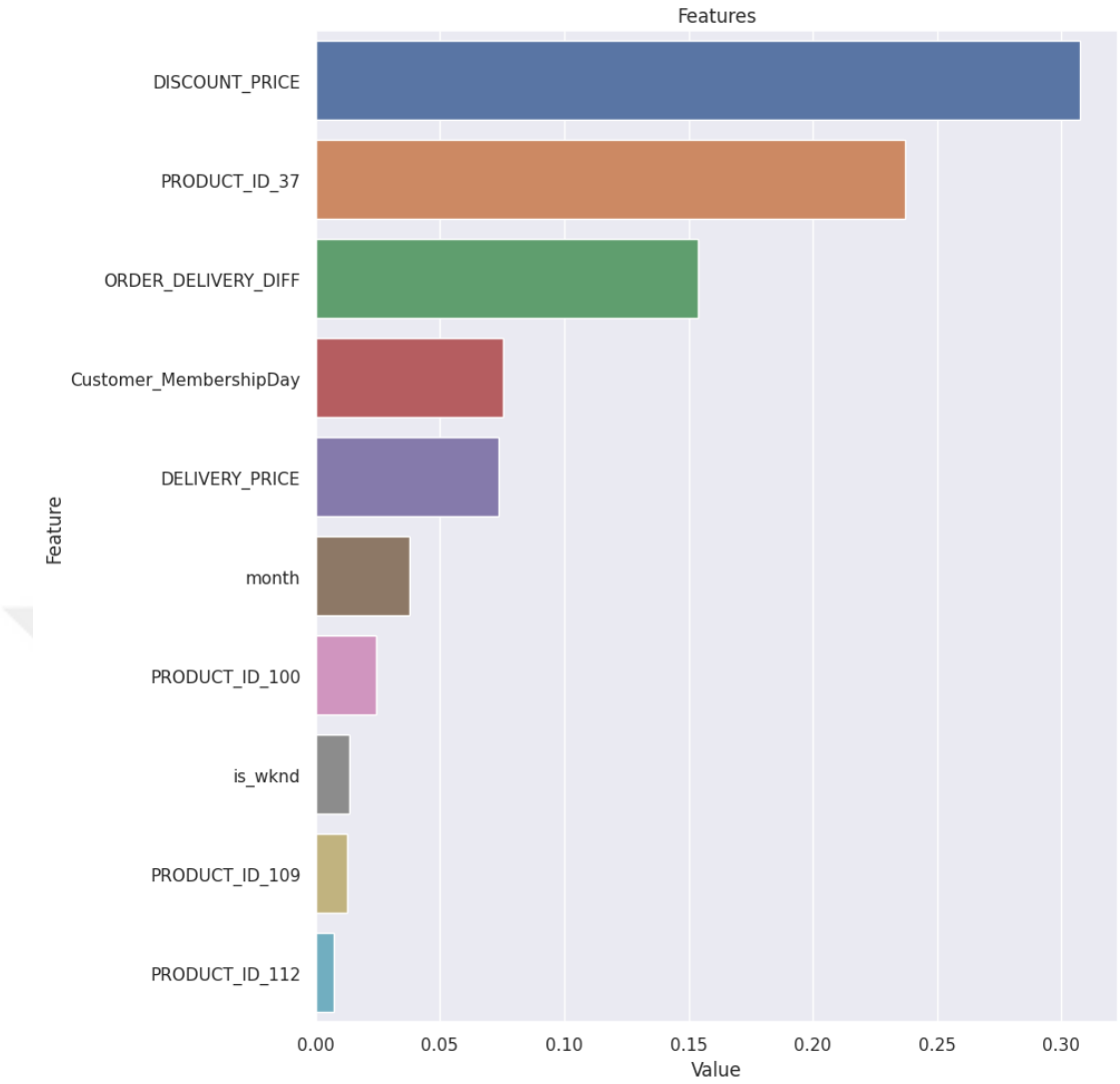


Şekil 4.15 : CatBoost Algoritması Özellik Seçimi Grafiği.

.



Şekil 4.16 : Gradient Boosting Algoritması Özellik Seçimi Grafiği.



Şekil 4.17 : Random Forest Algoritması Özellik Seçimi Grafiği.

CatBoost algoritması uygulandığında önemli olarak değerlendirilen değişkenler ve önem değerleri Çizelge 4.2'de yer almaktadır.

Çizelge 4.2 : CatBoost Algoritması Özellik Seçimi Sonuçları.

Değişken	**Önem Değeri**
PRODUCT_ID_37	38.513996
DISCOUNT_PRICE	24.426390
ORDER_DELIVERY_DIFF	14.487409
Customer_MembershipDay	5.546306
DELIVERY_PRICE	4.220278
Month	4.045705
PRODUCT_ID_100	3.029915
PRODUCT_ID_109	1.102881
is_wknd	1.031713
PRODUCT_ID_112	0.543367
PRODUCT_ID_123	0.407113
PRODUCT_ID_75	0.306247
PRODUCT_ID_12	0.274668
PRODUCT_ID_162	0.250851
PRODUCT_ID_27	0.247930
PRODUCT_ID_34	0.210094
PRODUCT_ID_135	0.153826
PRODUCT_ID_18	0.143688
PRODUCT_ID_108	0.112230
PRODUCT_ID_53	0.091639

Gradient Boosting algoritması uygulandığında önemli olarak değerlendirilen değişkenler ve önem değerleri Çizelge 4.3'de yer almaktadır.

Çizelge 4.3 : Gradient Boosting Algoritması Özellik Seçimi Sonuçları.

Değişken	**Önem Değeri**
PRODUCT_ID_37	0.334225
DISCOUNT_PRICE	0.319844
ORDER_DELIVERY_DIFF	0.179663
DELIVERY_PRICE	0.050486
PRODUCT_ID_100	0.028519
Customer_MembershipDay	0.028140
month	0.023676
PRODUCT_ID_75	0.009581
PRODUCT_ID_109	0.009525
PRODUCT_ID_123	0.002763
PRODUCT_ID_112	0.002164
PRODUCT_ID_12	0.002058
PRODUCT_ID_34	0.001817
PRODUCT_ID_162	0.001716
PRODUCT_ID_135	0.001141
PRODUCT_ID_50	0.001099
PRODUCT_ID_108	0.000698
PRODUCT_ID_122	0.000666
is_wknd	0.000490
PRODUCT_ID_102	0.000368

Random Forest algoritması uygulandığında önemli olarak değerlendirilen değişkenler ve önem değerleri Çizelge 4.4'de yer almaktadır.

Çizelge 4.4 : Random Forest Algoritması Özellik Seçimi Sonuçları.

Değişken	**Önem Değeri**
DISCOUNT_PRICE	0.307287
PRODUCT_ID_37	0.237045
ORDER_DELIVERY_DIFF	0.153902
Customer_MembershipDay	0.075617
DELIVERY_PRICE	0.073799
month	0.037734
PRODUCT_ID_100	0.024267
is_wknd	0.013804
PRODUCT_ID_109	0.012893
PRODUCT_ID_112	0.007334
PRODUCT_ID_123	0.004712
PRODUCT_ID_75	0.004468
PRODUCT_ID_12	0.003764
PRODUCT_ID_53	0.002852
PRODUCT_ID_18	0.002607
is_month_start	0.002532
PRODUCT_ID_34	0.002514
PRODUCT_ID_162	0.002511
PRODUCT_ID_135	0.002020
PRODUCT_ID_108	0.001945

Özellik seçimi aşamasında üç farklı algoritma uygulandığında gerçeleştirildiğinde; birçok değişkenin ortak olarak seçildiği görülmektedir. Aşağıda, seçilen değişkenlerin bir listesi yer almaktadır:

- ORDER_DELIVERY_DIFF
- Customer_MembershipDay
- DELIVERY_PRICE
- DISCOUNT_PRICE
- Month
- is_wknd
- PRODUCT_ID_12
- PRODUCT_ID_37
- PRODUCT_ID_75
- PRODUCT_ID_100
- PRODUCT_ID_109
- PRODUCT_ID_112
- PRODUCT_ID_123

Seçilen değişkenlerin hedef değişkeni üzerindeki etkisini ölçmek adına duyarlılık analizi çalışması gerçekleştirilmiştir.

6.4.2 Duyarlılık analizi

Özellik seçimi bölümünde, seçilen 13 değişken ile yeni bir model kurulumu gerçekleştirilmiştir. Duyarlılık analizi için tercih edilen algoritma, Random Forest algoritmasıdır ve bu algoritma kullanılarak yeni bir model oluşturulmuştur.

Duyarlılık analizi, değişkenlerin tahmin edilen hedef değişkeni olan 'Piece' üzerindeki etkisini ölçmeyi amaçlamaktadır. Bu amaç doğrultusunda her bir değişkenin etkisini kontrol etmek için bir örnek oluşturulur ve ilgili değişken bir birim artırılarak yeni bir örnek oluşturulur. Ardından, bu yeni örnekteki hedef değişken değeri ve orijinal örnekteki hedef değişken değeri tahmin edilir. Tahmin edilen değerler karşılaştırılır. İlgili değişkenin artırılmasının hedef değişkenini nasıl etkilediği gözlemlenir.

Bu yöntem sayesinde, seçilen değişkenlerin hedef değişkenini nasıl etkilediği ve hangi değişkenlerin daha fazla önem taşıdığı analiz edilmektedir.

Seçilen değişkenler ile Random Forest talgoritması uygulanarak tahmin çalışması sonucunda çıkan sonuçlar Çizelge 4.5'te yer almaktadır.

Çizelge 4.5 : Duyarlılık Analizi Sonuçları.

Değişken	Hedef Değişkene Etkisi
ORDER_DELIVERY_DIFF	Artırır
Customer_MembershipDay	Azaltır
DELIVERY_PRICE	Azaltır
DISCOUNT_PRICE	Artırır
Month	Değiştirmez
is_wknd	Artırır
PRODUCT_ID_12 (Barbunya)	Artırır
PRODUCT_ID_37 (Domates Petemek)	Artırır
PRODUCT_ID_100 (Mısır)	Artırır
PRODUCT_ID_109 (Patates Kızartmalık)	Artırır
PRODUCT_ID_112 (Patlıcan Gönen)	Artırır
PRODUCT_ID_123 (Salatalık Çengelköy)	Artırır
PRODUCT_ID_75 (Kavun Kırkağaç)	Azaltır

7. SONUÇ VE ÖNERİLER

Bu çalışmada farklı makine öğrenme algoritmaları ve satış verileri kullanılarak adet tahmini yapılmış ve önemli olan değişkenlerin keşfi sağlanmıştır. Sonrasında bu değişkenlerin hedef değişkeni üzerindeki etkisine bakılmıştır.

İlk olarak, farklı makine öğrenmesi algoritmaları kullanılarak modellerin performansı ölçülmüştür. Catboost algoritmasının en iyi tahmin başarısına sahip olduğu belirlenmiştir. Kurulan modellerde ilk etapta çok sayıda değişken modele dahil edilmiştir. Her bir satılan ürün kategorik değişkenden sayısal değişkene dönüşümü sağlanarak modele dahil edilmiştir.

Satış tahmini yaparken bazı değişkenlerin diğer değişkenlere göre daha çok etki ettiği gözlemlenmiştir. Bu değişkenlerin keşfi özellik seçimi aşamasında gerçekleştirilmiştir. Bu aşamada en iyi performansa sahip üç algoritma için özellik seçimi aşaması uygulanmıştır. 3 algoritma sonucu karşılaştırıldığında modele etkisi bakımından en etkili değişkenlerin aynı değişkenler olduğu gözlemlenmiştir.

Product_ID_37 değişkeni en önemli değişken olarak gözlemlenmiştir. Bu ürüne baktığımızda Domates Petemek ürünüdür. Bu ürün satış adedi tahmin ederken en önemli değişken olduğu sonucu çıkmıştır. Uygulanan indirim, satış adedi tahmininde önemli bir değişken olarak belirlenmiştir. Sipariş tarihi ile teslim tarihi arasında geçen süreyi ifade eden değişken de satış adedi tahmininde önemli rol oynamaktadır. Bir kullanıcının üye olduğu tarihten günümüze kadar geçen süre satış adedini tahmin ederken önemli bir değişken olarak karşımıza çıkmaktadır. Teslimat ücreti satışı tahmin etmede önemli değişkenler arasındadır.

Önemli değişken olarak seçilen 13 değişken ile yeni model kurulumu gerçekleştirilmiştir. Yeni model kurulumunda duyarlılık analizi çalışmalarında tercih edilen Random Forest algoritması kullanılmış ve yeni kuralan model ile duyarlılık analizi çalışması gerçekleştirilmiştir. Amacı, önemli değişkenlerin hedef değişkeni üzerindeki etkisini ölçmek olan bu analizler sonucunda bazı önemli sonuçlara ulaşılmıştır.

Elde edilen sonuçların analizine göre; ‘ORDER_DELIVERY_DIFF’ değişkeninin bir birim artmasının hedef değişkeninin artmasını sağladığı ve bu durumda sipariş tarihi ile teslimat tarihi arasında geçen sürenin artması yani ileri tarihli sipariş oluşturulması kullanıcıların daha fazla ürün satın almalarını teşvik ettiği tespit edilmiştir. Kullanıcılar sipariş verdikten sonra ürünlerini en kısa sürede teslim almayı beklerler ancak bu çalışmada modele dahil edilen ‘ORDER_DELIVERY_DIFF’ değişkeninin artması, siparişin teslimatının hemen gerçekleşmeme senaryosunda satış adedine nasıl etki edildiğini göstermektedir. Analizlerimiz, bazı kullanıcıların bekleme süresi boyunca daha fazla ürün satın alma eğiliminde olduklarını göstermektedir. Kullanıcılar ileri tarihli teslimatı tercih ettiklerinde; ilgilerini çeken ek ürünleri de satın alma işlemi gerçekleştirdikleri düşünülebilir.

Kullanıcının üye olduğu tarih üzerinden geçen gün sayısını ifade eden ‘Customer_MembershipDay’ değişkeninin hedef değişkeni üzerindeki etkisine baktığımızda; kullanıcının üyelik süresinin artmasıyla satış adedinin düştüğünü göstermektedir. Bu sonucun ortaya çıkmasının sebeplerinden biri müşteri sadakati olabilir. Uzun süreli üyelere, başlangıçta ilk alışverişe özel tanımlanan kuponlar daha yüksek miktarda satış yapılmasını sağlamış olabilir. Zamanla bu müşterilerin daha az satın alma sıklığının azalması beklenir. Müşterinin sadakati ve alışkanlıkları, üyelik süresiyle birlikte değişebilir. Bir başka sebep olarak kullanıcıların değişen tercihleri olabilir. Kullanıcıların üyelik süresi arttıkça, tercihleri ve ihtiyaçları da satış adedini etkiliyebilir. Kullanıcıların ihtiyaçları zamanla değişebilir ve üyelik süresi arttıkça farklı ürün veya hizmetlere olan talepleri de değişebilir. Kullanıcılar belli bir süre sonra ihtiyaçlarını karşılamış olabilir veya satın alma işleminin doygunluğuna ulaşmış olabilirler.

‘DELIVERY_PRICE’ değişkeninin hedef değişkeni üzerindeki etkisi incelendiğinde; kullanıcıların ürün ve teslimat fiyatlarına duyarlı olduğu ve bu fiyatların satın alma kararlarını etkilediği gözlemlenmiştir. Kullanıcılar daha yüksek teslimat ücreti nedeniyle satın alma miktarını veya sıklığını azaltabilirler. Bu durumda işletmenin fiyat politikasını gözden geçirmesi ve rekabetçi teslimat fiyatları sunması önemlidir. Kullanıcıların rekabetçi fiyatlarla ürün satın almalarını teşvik etmek satış adedini artırmayı sağlayabilir. Teslimat ücretinin artması satış adetlerini olumsuz yönde etkilediğinden alternatif teslimat seçenekleri sunulabilir. Dönemsel veya bölgesel

satış adedini artırmaya yönelik ücretsiz teslimat veya indirimli teslimat seçenekleri ile avantajlar sunmak kullanıcıların taleplerini artırabilir. Teslimat ücretinin yüksek olması kullanıcıların satın alma motivasyonlarını da etkileyebilir. Kullanıcılara değerli olduğunu hissettirmek ve iyi bir alışveriş deneyimi sunmak için teslimat maliyetlerinin optimize edilmesi sağlanabilir.

'DISCOUNT_PRICE' değişkeninin artması ile satış adedinin artması durumu pazarlama stratejileri ve müşteri davranışı üzerinde düşünülmesi gereken noktaları içermektedir. Kullanıcılar kendilerine indirimli fiyatlar sunulduğunda satın alma işlemine daha fazla ilgi göstermektedir. Yüksek indirimler kullanıcının daha fazla ürün satın almasını teşvik etmektedir. Uygulanan indirimler kullanıcıların fiyat hassasiyetini azaltabilir ve satış adetlerini artırabilir. İndirim uygulanması müşteri çekme ve sadakat oluşturmak için etkili bir yöntemdir. Yeni üyelik oluşturan kullanıcıların yanı sıra uzun vadeli kullanıcıların da satın alma sıklıklarını artırmak sağlanabilir.

'Month' değişkeninin artmasının satış adedi üzerinde anlamlı bir etki oluşturmaması elde edilen önemli sonuçlardan biridir. Ay bilgisinin değişiminden satış adedinin etkilenmemesi, kullanıcıların davranışlarının mevsimselliğe bağlı olarak değişmediğini göstermektedir. Bir başka deyişle, ayın değişmesi kullanıcıların alışkanlıklarını veya satın alma davranışlarını etkilememektedir. Kullanıcıların satın alma kararlarını kişisel tercihler, ürün özellikleri veya ihtiyaçları gibi faktörler belirleyebilir. Ay bilgisinin satış adedine anlamlı bir etkisinin olmaması işletmenin gelecek stratejilerini belirlerken diğer değişkenlere odaklanmasını sağlayabilir. Diğer değişkenlerin satış adedine olan etkisinin daha detaylı bir şekilde analizi yapılabilir.

'Is_Wknd' değişkeni ile hedef değişkeni arasında pozitif bir ilişki olduğu sonucu elde edilmiştir. Hafta sonu olması kullanıcıların daha fazla ürün satın alma eğiliminde olduğunu göstermektedir. Hafta sonu kullanıcıların daha fazla zamanlarının olduğu ve ihtiyaçlarına odaklandıkları dönemdir. İşletme hafta içine özel kampanya veya indirim uygulayarak kullanıcıların satın alma eğilimlerini artırabilir. Kullanıcıların yoğun olduğu saat aralığında bu kampanya veya indirimin tanımlanması kullanıcıların etkileşimini ve satın alma potansiyelini artırabilir.

Ürünlerin hedef değişkeni üzerindeki etkisine bakıldığında ürün bazında değişen etkilerin olduğu gözlemlenmiştir. Barbunya ürününü temsil eden 'Product_Id_12'

değişkeninin artması veya mısır ürününü temsil eden ‘Product_Id_100’ değişkeninin artması veya kızartmalık patates ürününü temsil eden ‘Product_Id_109’ değişkeninin artması satış adedini artırır. Taze ve kaliteli ürünleri satışa sunan tedarikçi kullanıcıların ürüne olan ilgisini artırabilir. Ürünlerin mevsimsel olarak tercih edildiği dönemde satış adedinin artması beklenir. Satış adedini daha fazla artırmak adına satış adedini artıran ürünler üzerinden özel indirimler veya kuponlar sunulabilir. Kavun kırkağaç ürününü temsil eden ‘Product_Id_75’ değişkeninin artması satış adedini azaltmaktadır. Rekabetin yüksek olduğu düşünülürse, farklı kavun çeşitlerinin olması kavun Kırkağaç ürününün satış adedinin azalmasına neden olabilir. Kullanıcıların satın aldıkları kavundan memnun olmamaları satın alma sıklığının azalmasına neden olabilir. Kullanıcılarda oluşan memnuniyetsizliği önlemek adına çalışmalar gerçekleştirilebilir. Satın almış oldukları ürünü tedarik eden pazarcılar ile görüşülebilir. Ürün kalitesinin düşmemesi adına tedarikçiler ile anlaşma şartları tekrardan değerlendirilebilir. Patlıcan gönen ürününü temsil eden ‘Product_Id_112’ değişkeninin artması satış adedini artırmaktadır. Patlıcan yemekleri ve meze çeşitlerinin popüler olduğu bölgelere özel çalışmalar gerçekleştirilerek satış adedinin artması sağlanabilir. Salatalık Çengelköy ürününü temsil eden ‘Product_Id_123’ değişkeninin artması satış adedini artırmaktadır. Taze ve kaliteli ürünlerin uygun fiyatlı sunulması satış adedini artırır. Satış adedini artırmak adına uygulama üzerinden sağlıklı tarifler veya sağlıklı yaşam önerileri gibi ek bilgiler sunulabilir.

Her bir ürünün satış adedi üzerindeki etkisi farklı olabilir. Ürün bazında stratejiler belirlenebilir. Talebi artan bir ürün için stok düzenlemeleri gerçekleştirmek gerekirken, talebi azalan etkisi fazla ürünler için pazarlama stratejileri geliştirilebilir.

Bu çalışmadan elde edilen sonuçlar seçilen veri setine bağlı oluşmuştur. Kurulan modeller ve tahmin çalışmaları farklı veri setinde farklı sonuçlar oluşturabilir. Yeni değişkenlerin eklenmesi ile satış adedini etkileyen önemli değişkenlerin değişmesi mümkündür. Gelecekteki çalışmalarda farklı makine öğrenmesi algoritmaları kullanılarak veya farklı değişkenler eklenerek gerçekleştirilebilir.

KAYNAKLAR

Barabanova, I. V. & Vychuzhain, P. & Nikitin, N.O. (2021). Sensitivity Analysis of the Composite Data-Driven Pipelines in the Automated Machine Learning, *Procedia Computer Science, 193, 484-493.*

Chen, D. & Liang, E. & Zhou, K. & Liu, F. (2022). Sales Forecasting for Fashion Products Considering Lost Sales, *Applied Sciences,12,7081.*

Dharshini, M. P. & Vijila, S. Antelin. (2021). Comparative Study of Product Sales Forecasting Methods, *International Research Journal on Advanced Science Hub (IRJASH),03.*

Ensafi, Y. & Amin, S. H. & Zhang, G. & Shah, B. (2022). Time-series forecasting of seasonal items sales using machine learning – A comparative analysis, *International Journal of Information Management Data Insights,2,100058.*

James, G. & Witten, D. & Hastie, T. & Tibshirani, R. (2013). Statistical learning. An Introduction to Statistical Learning, *Springer: 15-57.*

Kamalov, F. (2018). Sensitivity Analysis for Feature Selection, 2018 *17th IEEE International Conference on Machine Learning and Applications (ICMLA), 1466-1470.*

Mitra, A. & Jain, A. & Kishore, A. & Kumar, P. (2022). A Comparative Study of Demand Forecasting Models for a Multi-Channel Retail Company: A Novel Hybrid Machine Learning Approach, *Operations Research Forum,3,*58.

Polewko-Klim, A. & Lesiński, W. & Kitlas Golińska, A. & Mnich, K. & Siwek, M. & Rudnicki, W. R. (2020). Sensitivity analysis based on the random forest machine learning algorithm identifies candidate genes for regulation of innate and adaptive immune response of chicken, *Poultry Science, 99(12), 6341-6354.*

Saputra. A. & Suharjito. (2019). Fraud Detection using Machine Learning in e-Commerce. (IJACSA) *International Journal of Advanced Computer Science and Applications, 10(9),332-339.*

Srivastava, P. R. &Eachempati, P. & Charles, V. & Rana, N. P. (2022). A hybrid machine learning approach to hotel sales rank prediction, *Journal of the Operational Research Society.*

Şahinarslan, F. V. (2019). MAKİNE ÖĞRENMESİ ALGORİTMALARI İLE NÜFUS TAHMİNİ: TÜRKİYE ÖRNEĞİ. (Yüksek Lisans tezi). İstanbul Teknik Üniversitesi, Lisansüstü Eğitim Enstitüsü, İstanbul.

Thivakaran, T. K. & Ramesh, M. (2022). Exploratory Data analysis and sales forecasting of bigmart dataset using supervised and ANN algorithms, *Measurement: Sensors,23,100388.*

Wolpert, D. H. ve W. G. Macready (1997). "No free lunch theorems for

Optimization., *IEEE transactions on evolutionary computation 1(1): 67-82.*

Zhou, H. & Sun, G. & Fu, S. & Jiang, W. & Juan Xue, J. (2019). A Scalable Approach for Fraud Detection in Online E-Commerce Transactions with Big Data Analytics. *CMC-COMPUTERS MATERIALS & CONTINUA, 60(1), 179-192.*

Zhao, t. & Zheng, Y. & Wu, Z. (2022). Improving computational efficiency of machine learning modeling of nonlinear processes using sensitivity analysis and active learning. *Digital Chemical Engineering 3, 100027.*

Dayanıklı, A. S. (2021). Aykırı Değer (Outlier) Analizi Nedir? Uç Değerler Nasıl Tespit Edilir?, Erişim: 12 Ocak 2022, https://ravenfo.com/2021/02/11/aykiri-deger-analizi/.

Dayanıklı, A. S. (2021). Kayıp Veri Nedir? Kayıp Veri Analizi Nasıl Yapılır?, Erişim: 14 Ocak 2022, https://ravenfo.com/2021/06/06/kayip-veri-analizi/.

Dayanıklı, A. S. (2021). Normal Dağılım ve Python ile Normallik Testi, Erişim: 18 Ocak 2022, https://ravenfo.com/2021/07/11/normal-dagilim-python-normallik-testi/.

EKLER

EK A: Litaratür Taraması Özet Çizelge

EK A: Litaratür Taraması Özet Çizelge

Sıra	Yıl	Yazar	Yayın Adı	Method
1	2022	(Zhao vd., 2022)	Improving computational efficiency of machine learning modeling of nonlinear processes using sensitivity analysis and active learning	Recurrent Neural Networks, Feature Selection, Sensitivity Analysis, Active Learning
2	2022	(Srivastava vd., 2022)	A hybrid machine learning approach to hotel sales rank prediction	Random Forest, Gradient Boosting
3	2021	(Dharshini & Vijila, 2021)	Comparative Study of Product Sales Forecasting Methods	Regression, Bass model
4	2022	(Thivakaran & Ramesh, 2022)	Exploratory Data analysis and sales forecasting of bigmart dataset using supervised and ANN algorithms	Random Forest, XGboost, ANN, Regression
5	2022	(Chen vd., 2022)	Sales Forecasting for Fashion Products Considering Lost Sales	LR, GBDT, SVR, ANN
6	2018	(Kamalov, 2018)	Sensitivity Analysis for Feature Selection	Feature Selection, Sensitivity Analysis, Total Sensitivity Index
7	2020	(Plewko-Klim vd., 2020)	Sensitivity analysis based on the random forest machine learning algorithm identifies candidate genes for regulation of innate and adaptive immune response of chicken	Random Forest, Sensitivity Analysis, Feature Selection
8	2021	(Barabanova vd., 2021)	Sensitivity Analysis of the Composite Data-Driven Pipelines in the Automated Machine Learning	Sensitivity Analysis, AutoML

Çizelge A. 1 Literatür Araştırması Özet Tablo

ÖZGEÇMİŞ

Ad-Soyad **:Rabia AYDIN**

ÖĞRENİM DURUMU:

- **Lisans** **:** 2019, İstanbul Teknik Üniversitesi, Fen ve Edebiyat Fakültesi, Matematik Mühendisliği