

T.C.
İSTANBUL ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
İŞLETME ANABİLİM DALI

YÜKSEK LİSANS

BÜYÜK VERİ VE FİNANSAL PİYASALARDA
İSTATİSTİKİ MAKİNE ÖĞRENMESİ
METODLARININ YATIRIM KARARLARINDA
KULLANILMASI

Fehim KURUCAN

2501160206

TEZ DANIŞMANI

Prof. Dr. Ali HEPŞEN

İSTANBUL – 2019

TEZ ONAY SAYFASI



T.C.
İSTANBUL ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ



YÜKSEK LİSANS
TEZ ONAYI

ÖĞRENCİNİN;

Adı ve Soyadı : FEHİM KURUCAN Numarası : 2501160206
Anabilim Dalı /
Anasanat Dalı / Programı : FİNANS Danışmanı : PROF.DR.ALİ HEPŞEN
Tez Savunma Tarihi : 25.09.2019 Saati : 11.00
Tez Başlığı : BÜYÜK VERİ VE FİNANSAL PIYASALARDA İSTATİSTİKİ MAKİNE ÖĞRENMESİ
: METODLARININ YATIRIM KARARLARINDA KULLANILMASI.

TEZ SAVUNMA SINAVI, İÜ Lisansüstü Eğitim-Öğretim Yönetmeliği'nin 36. Maddesi uyarınca yapılmış,
sorulan sorulara alınan cevaplar sonunda adayın tezinin **KABULÜNE** OY BİRLİĞİ / OYÇOKLUĞUYLA karar verilmiştir.

JÜRİ ÜYESİ	İMZA	KANAATİ (KABUL / RED / DÜZELTME)
PROF.DR.MURAT KIYILAR		Kabul
PROF.DR.ALİ HEPŞEN		KABUL
DOÇ.DR.GÜÇLÜ OKAY		KABUL

YEDEK JÜRİ ÜYESİ	İMZA	KANAATİ (KABUL / RED / DÜZELTME)
DOÇ.DR.SİBEL YILMAZ TÜRKMEN		
DR.ÖĞR.ÜYESİ EMİR OTLUOĞLU		

ÖZ

BÜYÜK VERİ VE FİNANSAL PİYASALARDA İSTATİSTİKİ MAKİNE ÖĞRENMESİ METOTLARININ YATIRIM KARARLARINDA KULLANILMASI

Fehim KURUCAN

Bu tez çalışmasının amacı BİST100 endeksinin gün içindeki maksimum fiyatının tahmin edilmesidir. Bu kapsamda, makine öğrenmesi yöntemleri kullanılarak oluşturulan modellerden yatırım stratejileri geliştirilmesi hedeflenmiştir. Zaman serisi analizi olarak ele alınabilecek olan BİST100 endeksinin tahmin edilmesi, finansal kararların verilmesinde ve bu kararlarla oluşturulan yatırım stratejileriyle getiri elde edilmesinde önemli bir yere sahiptir. Makine öğrenmesi yöntemleriyle sermaye piyasalarında varlık fiyatlarının tahminine yönelik birçok çalışma literatürde mevcuttur. Günlük trendin tahmini ve kapanış fiyatının tahmini genel olarak uygulama alanlarıdır. Bu çalışmada, hem kullanılacak makine öğrenmesi yönteminin geliştirilmesi, hem de gün içi maksimum fiyatı kullanılarak, veri setinin genişletilmesiyle yatırım stratejisi modeli farklılaştırılmıştır.

BİST100 endeksinin gün içi en yüksek fiyatının tahmin edilmesinde dört ve beş yıllık uzunluğa sahip veri setleri kullanılmıştır. Bu sayede modelin başarısı iki ayrı veri seti ile test edilmiştir. Endeksin tahmininde; tahminci olarak endeksin geçmiş verileri gecikmeli olarak kullanılmıştır. Bununla birlikte, tahminci veri setinde; döviz kuru, faiz oranları, emtia ve CDS gibi ekonomik ve dışsal faktörlerde ele alınmıştır. Çalışmada iki ayrı makine öğrenmesi tekniği kullanılarak kapsam genişletilmiştir. Bunlar, literatürde öne çıkan Yapay Sinir Ağları ve XGBoost modelleridir. Modellerin karşılaştırılmasında literatürde yaygın kullanılan performans ölçütleri kullanılmıştır. Bunlardan bazıları; Ortalama Mutlak Hata (MAE), Sinyal Doğruluğu (SA), Ortalama Kare Hata (MSE) ve Kök Ortalama Kare Hata (RMSE)'dir.

Ampirik çalışma sonucunda, XGBoost yönteminin bu durumda Yapay Sinir Ağlarına göre daha iyi performans gösterdiğine yönelik bulgular elde edilmiştir. Çalışma kapsamında oluşturulan yatırım stratejileri elde edilen model performansları ışığında oluşturulmuştur. XGBoost ile oluşturulan yatırım stratejileri, aynı dönem aralığında yapılacak pasif yatırıma göre, daha iyi sonuç verdiği tespit edilmiştir. Modellerin tahmin gücünün artırılması için yapılan denemelerde, aynı günün açılış fiyatının tahminci veri setine eklenmesi, modellerin performansını artırmıştır.

Anahtar Kelimeler: Makine Öğrenmesi, XGBoost, Yapay Sinir Ağları, Tahmin, Fintech, Yatırım Stratejisi, BİST100, Hisse Senedi

ABSTRACT

**USING BIG DATA AND STATISTICAL MACHINE LEARNING METHODS
IN INVESTMENT DECISIONS IN FINANCIAL MARKETS**

Fehim KURUCAN

This paper aims to forecast the daily maximum prices of the BİST100 index. In this context, the target is to develop investment strategies from models created by using machine learning methods. Forecasting of the BİST100 index, which can be considered as time series analysis, has an important role in making financial decisions and generating returns with investment strategies formed by these decisions. There are many studies in the literature on the prediction of asset prices in capital markets using machine learning methods. The forecast of the daily trend and closing price are generally areas of application. In this study, the investment strategy model is differentiated both by developing the machine learning method to be used and by expanding the data set by using intraday maximum price.

In order to forecast the highest intraday price of the BİST100 index, four and five years of data sets were used. Thus, the power of the model was tested with two separate data sets. In predicting the index, historical data of the index were used with delay as a predictor data. In addition, the predictor data set includes economic and external factors such as exchange rate, interest rates, commodity, and CDS. In this study, the content was expanded by using two different machine learning techniques. These are Artificial Neural Networks and XGBoost models that are prominent in the literature. In the comparison of models, many performance criteria were used which are commonly used in literature. Some of those; Mean Absolute Error (MAE), Signal Accuracy (SA), Mean Square Error (MSE) and Root Mean Square Error (RMSE).

As a result of the empirical study, it was found that the XGBoost method showed better performance than the Artificial Neural Networks in this situation. The investment strategies created within the scope of the study were formed through with the obtained model performances. The investment strategies created with XGBoost

have been found that give better results than passive investment for the same period of time. In the experiments to increase the predictive power of the models, it was observed that adding the opening price of the same day to the predictor data set increased the performance of the models.

Keywords: Machine Learning, XGBoost, Artificial Neural Network, Forecasting, Fintech, Investment Strategy, BIST100, Stock



ÖNSÖZ

Bu tez çalışmasının amacı BİST100 endeksini tahmin ederek yatırım stratejileri oluşturmaktır. Tahmin modellerinin oluşturulmasında makine öğrenmesi metotları kullanılmıştır. Zaman serisi analizi olarak ele alınabilecek olan BİST100 endeksinin tahmin edilmesi, finansal kararların verilmesinde ve bu kararlarla oluşturulan yatırım stratejileriyle getiri elde edilmesinde önemli bir yere sahiptir. Bu çalışmada çeşitli analizler ışığında yatırım stratejileri oluşturularak, BİST100 endeksinden getiri elde edilmesi hedeflenmiştir.

Öncelikle desteğini her daim hissettiğim kıymetli abim Ahmet Salih Kurucan'a, çalışmamda desteklerini esirgemeyen tez danışmanım Prof. Dr. Ali HEPŞEN'e ve maddi manevi desteklerini esirgemeyen her zaman yanımda olan sevgili aileme teşekkürü bir borç bilirim.

Fehim KURUCAN

İSTANBUL, 2019

İÇİNDEKİLER

ÖZ.....	iii
ABSTRACT	v
İÇİNDEKİLER	viii
TABLolar LİSTESİ.....	x
ŞEKİLLER LİSTESİ.....	xi
KISALTMALAR LİSTESİ.....	xii
GİRİŞ	1

BİRİNCİ BÖLÜM

BÜYÜK VERİ VE MAKİNE ÖĞRENMESİ MODELLERİ

2.1	Büyük Veri	3
2.2	Alternatif Veri	4
2.3	Makine Öğrenmesi	4
2.4	Çalışmada Kullanılan Makine Öğrenmesi Yöntemleri	6
2.4.1	Yapay Sinir Ağları	6
2.4.2	XGBoost.....	11

İKİNCİ BÖLÜM

LİTERATÜR İNCELEMESİ

2.1	Borsa İstanbul ve Makine Öğrenmesi Üzerine Yapılan Çalışmalar.....	20
2.2	Makine Öğrenmesi ile Modellenmiş Zaman Serisi Çalışmalarının Sınıflandırılması	23

ÜÇÜNCÜ BÖLÜM

KULLANILAN VERİLER VE PERFORMANS ÖLÇÜTLERİ

3.1	Kullanılan Veriler	25
3.1.1	Borsa Endeksleri	25
3.1.2	Ekonomik Parametreler.....	29
3.1.3	Döviz Kurları ve Emtialar	31
3.2	Performans Ölçütleri	34

DÖRDÜNCÜ BÖLÜM

BİST100 ENDEKSİ ÜZERİNE UYGULAMA

4.1	Metedoloji	36
4.2	Veri Düzenlemeleri	36
4.3	Makine Öğrenmesi Modelleriyle Tahmin Yapma ve Performans Karşılaştırmaları	38
4.4	Yatırım Stratejisi Uygulamaları	43
SONUÇ.....		50
KAYNAKÇA		52
EKLER.....		55

TABLÖLAR LİSTESİ

Tablo 2.1 : Makine Öğrenmesi Modellerine Göre Çalışmaların Sınıflandırılması (Krollner a.e., 2010).....	23
Tablo 2.2 : Zaman Periyotlarına Göre Çalışmaların Sayısı (Krollner a.e., 2010).....	24
Tablo 2.3 : Kullanılan Verilere Göre Çalışmaların Sayısı (Krollner a.e., 2010)	24
Tablo 4.1 : Ertesi Gün En Yüksek Fiyat Tahmini Birinci Veri Seti Performans Sonuçları	39
Tablo 4.2 : Aynı Gün En Yüksek Fiyat Tahmini Birinci Veri Seti Performans Sonuçları	40
Tablo 4.3 : Ertesi Gün En Yüksek Fiyat Tahmini İkinci Veri Seti Performans Sonuçları	41
Tablo 4.4 : Aynı Gün En Yüksek Fiyat Tahmini İkinci Veri Seti Performans Sonuçları	42
Tablo 4.5 : Getiri Karşılaştırma Tablosu (İkinci Veri Seti)	47
Tablo 4.6 : Getiri Karşılaştırma Tablosu (İkinci Veri Seti)	49

ŞEKİLLER LİSTESİ

Şekil 1.1 : Yapay Sinir Ağları Yapısının Görselleştirilmesi (Qui, 2016).....	8
Şekil 1.2 : İleri Beslemeli Yapay Sinir Ağı Örneği (Aşkın, 2011).....	9
Şekil 1.3 : Geri Beslemeleri Yapay Sinir Ağı Örneği (Aşkın, 2011)	10
Şekil 1.4 : XGBoost Modelinin Gelişimi (Morde, 2019)	12
Şekil 1.5 : Karar Ağacı Örneği (Chen, 2016)	12
Şekil 1.6 : Karar Ağacı Topluluk Öğrenme Yöntemi (Chen a.e., 2016)	13
Şekil 1.7 : Gradient Boosting.....	16
Şekil 1.8 : Karar Ağacının Dallara Ayrılması (Chen a.e., 2016).....	18
Şekil 1.9 : Karar Ağacı Dal Ayrımı (Chen a.e., 2016)	19
Şekil 3.1 : BİST100 Günlük En yüksek Fiyat Değişim Grafiği	25
Şekil 3.2 : BİST Bankacılık Endeksi Günlük Değişim Grafiği	27
Şekil 3.3 : BİST30 Endeksi Günlük Değişim Grafiği	28
Şekil 3.4 : S&P500 Endeksi Günlük Değişim Grafiği	28
Şekil 3.5 : 1 Yıllık Tahvil Faizi Günlük Değişim Grafiği	29
Şekil 3.6 : 10 Yıllık Tahvil Faizi Günlük Değişim Grafiği	30
Şekil 3.7 : CDS Günlük Değişim Grafiği	31
Şekil 3.8 : USD/TRY Günlük Değişim Grafiği.....	32
Şekil 3.9 : EUR/USD Günlük Değişim Grafiği.....	32
Şekil 3.10 : Brent Petrol Günlük Değişim Grafiği	33
Şekil 3.11 : XAU/USD Günlük Değişim Grafiği	33
Şekil 4.1 : Metodoloji	36
Şekil 4.2 : Öğrenme (Train) ve Test Ayrımlı Birinci Veri Seti.....	37
Şekil 4.3 : Öğrenme (Train) ve Test Ayrımlı İkinci Veri Seti.....	38
Şekil 4.4 : BİST100 Birinci Test Verisi Grafiği	43
Şekil 4.5 : Borsa İstanbul İşlem Saatleri (Borsa İstanbul, 2019).....	44
Şekil 4.6 : Birinci Test Verisi ile Oluşturulan Birinci Yatırım Stratejisi Getirileri...	45
Şekil 4.7 : Birinci Test Verisi ile Oluşturulan İkinci Yatırım Stratejisi Getirileri.....	46
Şekil 4.8 : İkinci Test Verisi ile Oluşturulan İkinci Yatırım Stratejisi Getirileri	48
Şekil 4.9 : İkinci Test Verisi ile Oluşturulan İkinci Yatırım Stratejisi Getirileri	49

KISALTMALAR LİSTESİ

ABD	: Amerika Birleşik Devletleri
ANN	: Artificial Neural Network
BİST	: Borsa İstanbul
CDS	: Credit Default Swap (Kredi Temerrüt Takası)
<i>e</i>	: Hata
k	: Gün Veri Sayısı
MAD	: Mean Absolute Deviation
MAE	: Mean Absolute Error
MAPE	: Mean Absolute Percentage Error
ME	: Ortalama Hata (Mean Error)
MSE	: Mean Squared Error
NN	: Neural Network
RMSE	: Root Mean Square Error
SA	: Sinyal Doğruluğu (Sign Accuracy)
SD	: Standard Sapma (Standard Deviation)
SVM	: Support Vector Machine
XGBoost	: Extreme Gradient Boosting

GİRİŞ

Sermaye piyasalarında varlık fiyatlarının tahminine yönelik günümüze kadar birçok çalışma yapılmıştır. Bu çalışmaların büyük bir kısmı da borsa endekslerinin veya hisselerin çeşitli periyotlarda yönlerinin ve fiyatlarının tahmin edilmesi ile ilgilidir. Ülkemizde Borsa İstanbul bütün bireylere açık işlem görmektedir. Borsa fiyat ve hacim verileri 15 dakika gecikmeli olarak ücretsiz erişime sunulmakta ve belirli ücretlerle de anlık veriye ulaşılabilir. Borsaya açık hisselerin finansalları ve diğer duyuruları da Kamuyu Aydınlatma Platformundan paylaşılmaktadır.

Borsaya açık hisselerin fiyatlaması ülkenin ekonomik yapısından, küresel değişimlerden, şirketin finansal tablolarından ve şirketle ilgili çıkan veya duyurulan haberlerden etkilenmektedir. Borsanın yönünün belirlenmesi üzerine yapılan birçok çalışma bu göstergeleri temel almıştır. Bir kısım çalışma ise borsada sıkça görülen teknik analiz yaklaşımını temel almıştır. Teknik analiz olarak RSI, CCI, ağırlıklı hareketli ortalama gibi birçok teknik gösterge veri setlerine eklenerek çalışmalar yapılmıştır.

Günümüzde yazılım teknolojisinin gelişmesiyle birlikte çalışmalarda çok çeşitli modeller geliştirilmekte ve kullanılmaktadır. Bunlardan makine öğrenmesi modellerinin birçok çalışmada kullanıldığı görülmektedir. Makine öğrenmesi modellerinin de birçok çeşidi bulunmaktadır. Ayrıca bazı çalışmalarda hibrit modeller oluşturulmakta ve bu modeller çeşitlendirilmektedir.

Bu çalışmanın temel amacı makine öğrenmesi teknikleriyle Borsa İstanbul 100 endeksinin tahmin edilmesi ve bu üretilen modellerle yatırım stratejileri oluşturulmasıdır. Bu çalışmada makine öğrenmesi tekniklerden yapay sinir ağları ve XGBoost modelleri ile tahminler yapılmıştır. Bu yapılan tahminlerle de yatırım stratejileri oluşturulmuştur.

Bu tez çalışmasında makine öğrenmesi modelleri oluşturulurken çeşitli veriler kullanılmıştır. Kullanılan veriler günlük bazda zaman serisi olarak oluşturulmuştur. Borsa İstanbul 100 endeksinin günlük olarak 5 verisi oluşmaktadır. Uygulamada ülkemizdeki en temel borsa endeksi olan Borsa İstanbul 100 endeksinin günlük en

yüksek fiyatının tahmin edilmesi hedeflenmiştir. Bu kapsamda birçok model oluşturulmuş ve bu modellerin performansları ölçülerek en iyi modeller seçilmiştir. Seçilen modellerle günlük yatırım stratejileri oluşturularak getirileri karşılaştırılmıştır.

Çalışmanın ilk kısmında, kullanılan makine öğrenmesi modellerinin temelleri ayrı ayrı anlatılmıştır. İkinci kısmında literatür araştırması yapılmıştır. Literatür araştırması hem yapılan çalışmaların modelleri bazında hem de kullanılmış veriler bazında yapılmıştır. Daha sonraki kısımda kullanılacak veriler ve performans ölçütleri incelenmiştir.

En son kısımda BİST100 üzerine uygulama yapılmıştır. Belirlenen verilerle yapay sinir ağları ve XGBoost modelleri kurularak BİST100 günlük en yüksek fiyatı tahminleri yapılmıştır. Modeller çeşitli parametrelerde denenmiş ve birden çok performans ölçütü kullanılarak karşılaştırılmıştır. Performans ölçütlerine göre çıkan en iyi modellerle yatırım stratejisi simülasyonları oluşturulmuştur. Bu simülasyonlar birbirleriyle karşılaştırılmış ve analiz edilmiştir.

BİRİNCİ BÖLÜM

BÜYÜK VERİ VE MAKİNE ÖĞRENMESİ MODELLERİ

1.1 Büyük Veri

Büyük veri deyince akla ilk gelen verinin boyutu olsa da aslında büyük veri sadece çok fazla veri demek değildir. Büyük veri ihtiyaca uygun şekilde hızlı akışa sahip, farklı formatlarda depolanması gereken verilerdir. Temel olarak büyük verinin 3 boyutu vardır. Bunlar hacim, hız ve çeşitliliklerdir.

Büyük verinin boyutunu ifade eden ifade hacimdir. Günümüzde veri hacmi petabayt boyutlarına ulaşmıştır. Özellikle sosyal medya gibi birçok insanın veri ürettiği web siteleri günlük olarak terabayt boyutunda veri üretmektedir. Verinin bu denli büyük olması işlenmesini zorlaştırmaktadır. Bu veriyi işleyebilmek ve buradan anlamlı sonuçlar elde edebilmek için kullanılan sistemlerin buna uygun optimize edilmesi gerekmektedir.

İnternetin de yaygınlaşmasıyla birçok kaynaktan veri oluşmaktadır. Veriler aynı zamanda çok çeşitli olabilmekte ve çok hızlı üretilmektedir. Verinin hızlı üretilmesi işlem sayısını artırmaktadır. Sürekli hareket halinde olan verilere geleneksel yöntemlerle analiz etmek mümkün olmamaktadır.

Günümüzde teknolojinin de gelişimiyle veri çeşitliliği artmıştır. Geleneksel verilere ek olarak web sitelerinde oluşan veriler, sosyal medya paylaşımları, e-postalar, paylaşılan çeşitli dokümanlar, elektronik cihazlardaki sensörler gibi çeşitli kaynaktan veriler oluşmaktadır. Bu verilerin hepsi farklı yapıdadır. Verilerin birleştirilebilmesi ve hepsinden bir sonuç çıkarılabilmesi için farklı yöntemlerle işlenmesi gerekmektedir.

Büyük veri bilginin büyüklüğüne değil, onunla ne yapılacağı üzerine yoğunlaşır. Büyük verinin yaygınlaşmasıyla birlikte birçok şirket çok veri tutmak yerine verilerin verimli bir şekilde analizine yoğunlaşmıştır.

Verilerin boyutları büyüdükçe, analiz edilmesi de zorlaşmaktadır. Bu tür verilerin analizinin günlük kullanılan normal bilgisayarlarla analiz edilmesi ve

buradan sonuçlar çıkartılması da imkansız hale gelmektedir. Makine öğrenmesi modelleri bu çeşidi çok ve büyük boyuttaki verileri çok daha hızlı ve kullanışlı şekilde analiz etmemize imkan sağlar.

1.2 Alternatif Veri

Teknolojinin gelişimiyle birlikte geleneksel verilere ek olarak birçok alternatif veri kaynağı oluşmuştur. Bunları sınıflandırdığımızda 3 ana grupta inceleyebiliriz. Bunlar insanların oluşturduğu, şirket faaliyetlerinden oluşan ve sensörlerle toplanan verilerdir.

İnsanların oluşturduğu verilere sosyal medya paylaşımları, haberler, analizler, yorumlar ve kişisel veriler örnek verilebilir. Sosyal medyada birçok çeşit paylaşım yapılmakta ve finansal piyasalarla, hisse senetleri ile ilgili bireysel düşünceler de çokça paylaşılmaktadır. Ekonomik değişimler, merkez bankası açıklamaları, yatırım kuruluşlarının analizleri, hükümetlerin yaptığı açıklamalar, şirkete özel haberler ve yorumlar sıkça internet üzerinden paylaşılmaktadır. Bu tarz veriler doğal dil işleme yöntemleriyle analiz edilebilmekte ve finansal piyasalara etkisi ölçülebilmektedir. Ayrıca bireylerin oluşturduğu arama motoru verileri de alternatif veriye örnek verilebilir.

Şirket faaliyetlerinin oluşturduğu veriler, şirketlerin kendi içlerinde toplayabileceği veya resmi kuruluşların açıkladığı verilerdir. Bu tür verilere kredi kartı işlem kayıtları, şirket tedarik zinciri kayıtları ve Türkiye İstatistik Kurumunun paylaştığı veriler örnek olarak verilebilir.

Sensörlerle oluşan verilere uydu verileri ve coğrafi konum verileri örnek verilebilir. Uydu görüntüleri, hava durumu sensörlerinden oluşan, telefonlardan oluşan ve market kameralarından oluşan veriler bu tipte verilerdir. Bu verilerin çoğu yapısal olmayan verilerdir ve diğer veri türlerine göre boyutları çok daha yüksektir.

1.3 Makine Öğrenmesi

Makine öğrenmesi istatistiksel yöntemler kullanarak elimizdeki veriler arasında bağlantı oluşturan ve bu bağlantılarla çıkarımlar yapan algoritmalarıdır.

Makine öğrenmesi algoritmaları regresyon, kümeleme gibi yöntemlerle çalışabilmektedir. Bu çalışmada zaman serileriyle çalışılmıştır ve regresyon tabanlı algoritmalar kullanılmıştır.

Makine öğrenmesi temel olarak sınıflandırıldığında üçe ayrılır. Bunlar gözetimli öğrenme, gözetimsiz öğrenme ve takviyeli öğrenmedir.

- **Gözetimli Öğrenme (Supervised Learning):** Bu algoritmalarda veriye ait kategoriler, sınıflandırmalar veya sonuçlar daha önceden bilinir. Algoritma bilinen bu sonuçlara göre bilgiler arasında bağlantı kurarak öğrenir. Tahmin yapmak istenilen bilgilerde bu bağlantılar kullanılır. Bunlar aşağıdaki şekilde ikiye ayrılır.

- Regresyon
- Sınıflandırma

- **Gözetimsiz Öğrenme (Unsupervised Learning):** Bu kısımdaki algoritmalar gözetimli öğrenmede olan sonuç bilgisine sahip değildirler. Bu yöntem önceden etiketlenmemiş verilerin üzerinde çalışarak bağlantılar kurar. Bu bağlantılanan verileri birbirleri arasında kümeler. Bunlar aşağıdaki şekilde üçe ayrılır.

- Kümeleme (Clustering)
- Birlikte Kuralı (Association Rule Mining)
- Boyut Azaltma (Dimensionality Reduction)

- **Takviyeli Öğrenme (Reinforcement Learning):** Bu yöntem öğrenme işlemini geri bildirim ile ilerletir. Algoritmada devamlı denemeler yapılır ve hedefe ulaşamadığı sürece yapılan hareketleri iyi veya kötü olarak sınıflandırır. Öğrendiği deneyimlerden yanlış yapılan hareketleri bir daha tekrarlamayarak hep ileriye doğru çalışma mantığıyla çalışan algoritmalar.

1.4 Çalışmada Kullanılan Makine Öğrenmesi Yöntemleri

Bu çalışmada iki farklı makine öğrenmesi modeli kullanılmıştır. Bunlar yapay sinir ağları ve karar ağacı temelli olan XGBoost modelidir.

1.4.1 Yapay Sinir Ağları

Yapay zeka uygulamalarından biri olan yapay sinir ağları, insan beyninin çalışma yapısını taklit ederek mevcut verileri analiz edip, bu verilerden farklı öğrenme algoritmaları ile yeni bilgiler oluşturan bilgi işlem teknolojisidir.

Yapay sinir ağları insan beynin çalışmasını temel almış ve bu yapıyı taklit etmeyi hedeflemiştir. Sinir hücreleri biyolojik olarak synapsler yoluyla iletişim kurarlar ve birbirlerine bağlıdırlar. Yapay sinir ağlarında da buna benzer olarak birbirlerine bağlıdır ve gelen bilgileri toplar. Bu bilgileri aktivasyon fonksiyonundan geçirerek çıktıları üretir. Yapay sinir ağları veriden veri üretme yöntemiyle çalışmaktadırlar. Haykin (1994:24) yapay sinir ağlarını “Yapay sinir ağı, deneyime dayalı bilgiyi depolamaya ve bu bilgiyi kullanıma sunmaya yönelik doğal bir eğilim içinde olan yoğun paralel dağıtılmış bir işlemcidir” şeklinde tanımlamıştır.

Yapay sinir ağlarının herhangi bir olay hakkındaki girdi ve çıktılar arasındaki ilişkiyi, doğrusal olsun veya olmasın, elde bulunan mevcut örneklerden öğrenerek daha önce hiç görülmemiş olayları, önceki örneklerden çağrışım yaparak ilgili olaya çözümler üretebilmesi, algoritmik çözümü bulunamamış problemlerin çözümlenmesinde kullanımını artırmıştır (Yakut, 2014:141).

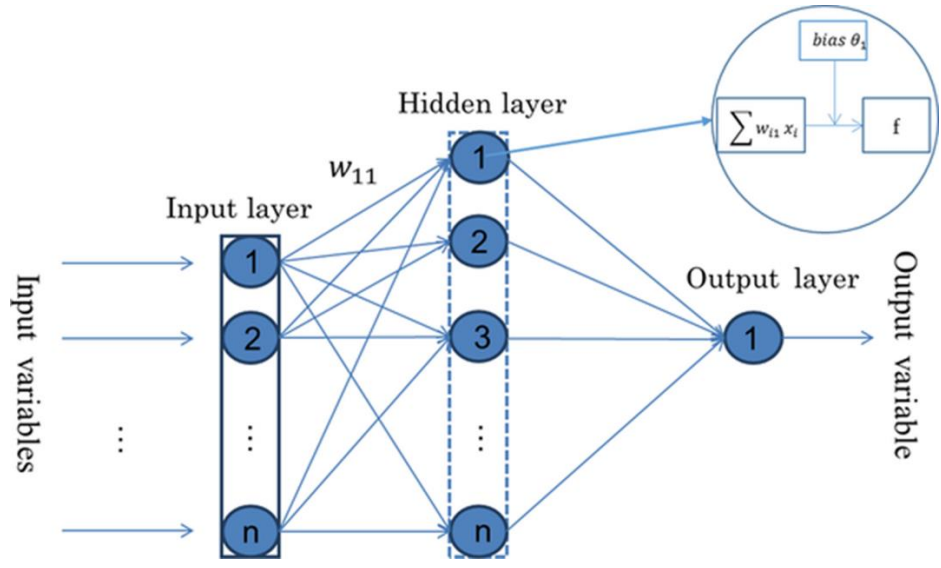
Yapay sinir ağları nöronlar yani yapay sinir hücreleri ile çalışmaktadır. Bu nöronlar birbirlerine bağlıdır ve aralarındaki bağlar ağırlık derecelerini ifade eder. Her bir nöron çeşitli bağlantılardan birbirine fonksiyonlar yardımıyla bağlanmasından oluşmaktadır.

Yapay sinir ağları temel olarak 3 ana gruptan oluşmaktadır. Bunlar;

- Girdi katmanı
- Ara katmanlar

- Çıktı katmanı

İlk kısım giriş katmanı, girdi verisinin bulunduğu kısımdır. İkinci kısım ise gizli katmanın bulunduğu ara kısımdır. Bu gizli katman birden çok olabilir. Son olarak, gizli katmandan gelen bilgilerin birleştiği çıktılarının bulunduğu katmandır. Temel olarak verilen giriş bilgileri, ağa girdi katmanından iletilir. Ara katmanlarda gelen veriler işlenir ve çıktı katmanına gönderilirler. Ara katman hem girdi katmanı ile hem de çıktı katmanı ile ağlarla bağlıdır. Bu ağların hepsi istatistiksel olarak belirli ağırlık derecelerini ifade etmektedir. Modelin verilen girdilerle doğru sonuçların elde edilebilmesi için ağırlık derecelerinin doğru değerleri alması gerekmektedir. Bu ağırlık derecelerinin doğru değerlerinin bulunması için farklı ağırlık dereceleriyle tekrar tekrar denemeler yapılmaktadır. Bu denemelere modelin eğitilmesi adı verilir. Değerler en başında rastgele değerler ile başlar. Modelin doğruluğu, ağırlık derecelerinin öğrenme kural yapısına göre her deneme ile geliştirilir. Bu denemeler modelden en doğru sonuçlar alınana kadar devam ettirilir. Doğru sonuçlara yaklaşma modelden beklenen performansa göre çalışmadan çalışmaya değişiklik göstermektedir. Modelin eğitim aşaması bittikten sonra modelimizin yapısı oluşmuş olur. Bu belirlenen yapı ile test verisi ile kurulan örneklerin çıktıların performansları ölçülür. Oluşan model beklenen performans değerlerine ulaşmadıysa, modelin giriş verisi değiştirilerek yeni modeller oluşturulabilir. Buna ek olarak kullanılan yapay sinir ağı yapısı veya ara katmandaki nöron sayısı gibi kullanılan parametrelerin değiştirilmesi gibi yöntemlerle yeni modeller oluşturulabilir.



Şekil 1.1 : Yapay Sinir Ağları Yapısının Görselleştirilmesi (Qui, 2016)

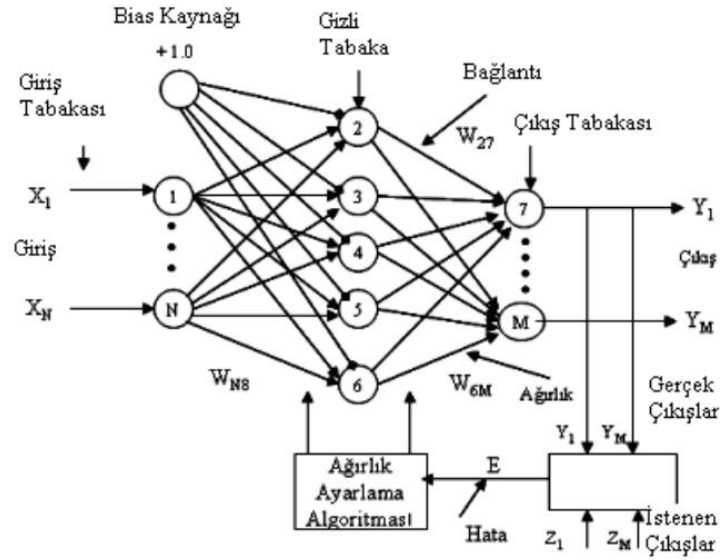
Yukarıdaki anlatılan nöronları ve bağlantılarını şekil 2.1. deki gibi ifade edebiliriz. Burada yer alan nöronlar saklı katmanları (Hidden Layer) ifade eder. Bu nöronlar birçok şekilde girdi ve çıktı verilerine bağlıdır. Aradaki bağlantılar (w) ağırlıkları gösterir. Çeşitli ağırlıklarında denenmesi ve bu denemelerden çıkan hataların minimize edilmesi ile de yapay sinir ağları ile makine öğrenmesi modelimiz oluşur. Ne kadar hatalar minimize edilirse o kadar kaliteli sonuçlar elde edilebilir.

Bu makine öğrenmesi yöntemi, gelişen bilgisayar teknolojisiyle birlikte insanların hesaplamasının zor olduğu büyük miktardaki verinin bilgisayar yardımıyla işlenmesinde kullanılabilir. Bu yöntemlerin geliştirilmesiyle birlikte çeşitli bilimsel alanlarda birçok çalışma yapılmakta ve birçok problemin çözümü sağlanmaktadır.

Yapay sinir ağları lineer regresyon gibi klasik yöntemlere göre farklı bir hesaplama yöntemi önermektedir. Oluşturulan verilere uyum sağlayabilen, adaptif, belirsizliklerde karar alıp devam edebilen, eksik bilgi verisi olsa dahi çalışabilen ve hatalar karşısında toleranslı olabilen bir yapıya sahiptir. Yapay sinir ağları modellerinin çok çeşitli kullanımları vardır. Günümüzde birçok farklı sektörde kullanılmaktadır.

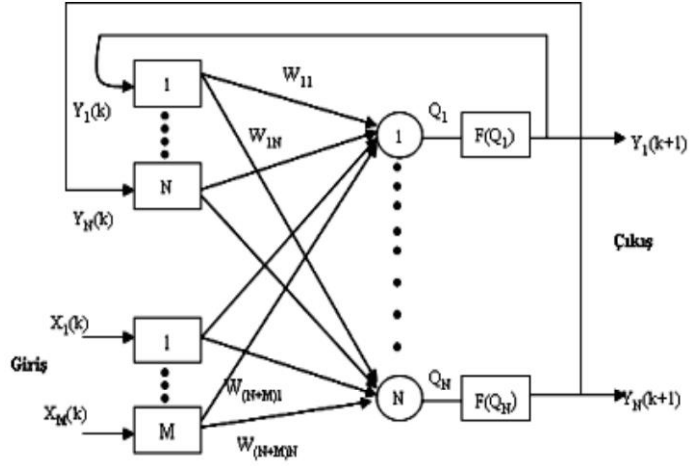
Yapay sinir ağırları çalışma biçimine göre iki farklı şekilde olabilmektedir. Bunlar; ileri beslemeli ve geri beslemeli ağırlar olarak adlandırılmaktadır.

- **İleri Beslemeli Ağlar:** İleri beslemeli ağırlar daima ileri yöne doğru veri akışı oluştururlar. Bu tür yapay sinir ağı modellerinde gizli tabaka sayısı, içindeki nöron sayıları ve girişteki nod sayıları tercihe göre belirlenir. Bu ara işlemcilerin artması modelimizin çalışma hızını düşürmektedir. Birden çok gizli tabakası olan modeller derin öğrenme olarak ifade edilir.



Şekil 1.2 : İleri Beslemeli Yapay Sinir Ağı Örneği (Aşkın, 2011)

- **Geri Beslemeli Ağlar:** Geri beslemeli ağırlar, ara katmanlardan giriş birimine veya çıkış katmanından ara katmanlara geri besleme yapılan ağlardır. Geri beslemeli sinir ağlarının dinamik hafızları vardır. Bu tip yapay sinir ağları zaman serisi analizlerinde sıkça kullanılmaktadır.



Şekil 1.3 : Geri Beslemeleri Yapay Sinir Ağı Örneği (Aşkın, 2011)

Yapay sinir ağı modelinin gelişimi ve oluşumu aşaması modelin eğitim kısmıdır. Modelin eğitimi giriş katmanı, ara katman ve çıkış katmanı arasındaki ağların optimize edilmesi ile gerçekleşir. Bu ağlar hepsi istatistiksel olarak bir ağırlığı ifade eder. Her aşamada ağırlıkların değiştirilmesi ile çıkan hataların minimize edilmesi hedeflenir. Modelin performansını ölçmede kullanılan hataların karelerinin ortalaması (mean square error) ile ölçüldüğünde aşağıdaki matematiksel ifade ortaya çıkar.

$$E(w) = \frac{1}{n} \sum_{\xi_i} (e_i^\xi)^2 = \frac{1}{n} \sum_{\xi_i} (d_i^\xi - y_i^\xi)^2 \quad (1.1)$$

Bu fonksiyonda d_i^ξ hedeflenen çıkış değeri ve y_i^ξ modelin çıkışında elde edilen tahmin değeridir.

$$w^{(k+1)} = w^{(k)} + \Delta w^{(k)} \quad (1.2)$$

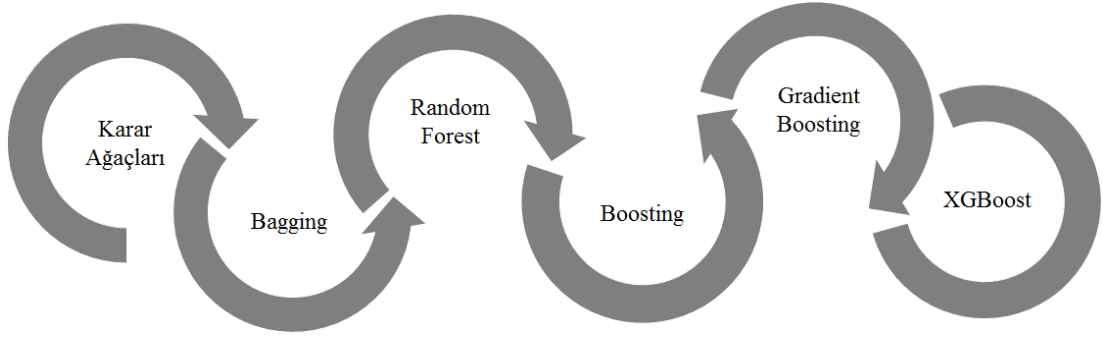
Yapay sinir ağlarının arasında bulunan ağırlıklar yukarıdaki formüldeki gibi ağırlığın değişimi $\Delta w^{(k)}$ birbiri üzerine eklenerek modeller eğitilir.

Yapay sinir ağı modelinin istenen sonuca doğru ilerlemesinde $E(w)$ 'yi mümkün olduğunca minimuma indirmek hedeflenir. Bunun için her seferinde $\Delta w^{(k)}$ ile kullanılan ağırlıklar değiştirilir. $\Delta w^{(k)}$ 'nın hesaplanması için birçok algoritma vardır. Bunlardan en yaygın kullanılanı geri yayımlı (back propagation) algoritmalarıdır. Bu algoritmalar Levenberg-Marquardt, Bayesian Regulation Fletcher-Powell Conjugate Gradient, ve Gradient Descent örnek olarak verilebilir.

- **Levenberg-Marquardt Method:** Levenberg-Marquardt temel olarak maksimum komşuluk üzerine kurulmuş en az kareler hesaplama metodudur. Bu algoritma Gauss-Newton ve gradient-descent algoritmalarının en iyi özelliklerinden oluşur ve bu iki metodun kısıtlamalarını kaldırır. Genel olarak bu metod yavaş yakınsama probleminden etkilenmez. Gauss-Newton algoritmasının kombinasyonu olan Levenberg-Marquardt algoritması optimizasyon problemlerinde eğimli düşüş algoritmasına göre (gradient descent) daha etkindir. Ayrıca daha hızlıdır (Aşkın, 2011).
- **Bayesian Regulation Method:** Bayesian regülasyonu Levenberg-Marquardt optimizasyonuna göre ağırlı ve bias değerlerini günceller. Karesel hata ve ağırlıkların kombinasyonu minimize eder ve ağı üretmek için doğru kombinasyonu belirler (Aşkın, 2011).

1.4.2 XGBoost

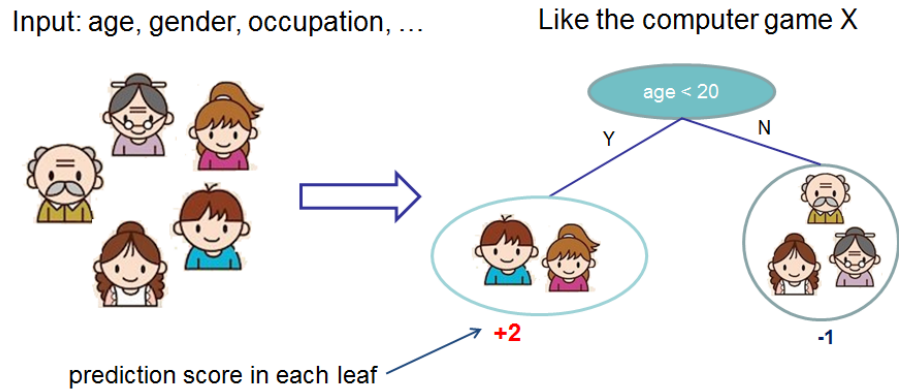
XGBoost modelinin temeli karar ağaçları ile başlamaktadır. Random forest ve gradient boosting modelleri de karar ağacı tabanlı modellerdir. XGBoost modeli gradient boosting yönteminin geliştirilmiş halidir. Bu bölümde sırasıyla karar ağacı, random forest, gradient boosting ve buna ek olarak bagging, boosting algoritmalarının temel yapıları incelenmiştir.



Şekil 1.4 : XGBoost Modelinin Gelişimi (Morde, 2019)

1.4.2.1 Karar Ağaçları

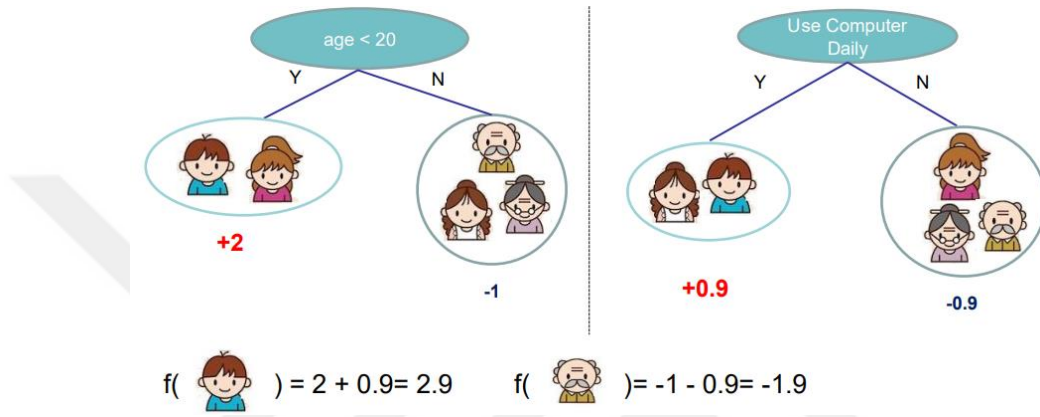
Karar ağaçları, temel olarak veri madenciliğinde sınıflandırma yöntemlerinden biridir. Bu yöntemle veriler dallara ayrılarak belirlenen kural setine göre sınıflandırılır ve analiz edilir. Daha sonrasında ise karar ağaçları regresyona da uyarlanarak zaman serilerinde de kullanılabilir hale gelmiştir. Random forest ve gradient boosting yöntemleri karar ağaçları oluşturularak regresyon yapılabilen modellerdir.



Şekil 1.5 : Karar Ağacı Örneği (Chen, 2016)

Karar ağaçları yukarıdaki örnekte olduğu gibi değişkenlere göre ayırım yapılarak oluşturulurlar. Ayırımların her birine dal ismi verilir. Buradaki örnekte bilgisayar oyunu sevenler, yaşa göre ayrılmıştır. Belirli bir topluluk örneği, 20 yaşından küçükler ve büyükler olarak ikiye ayrılmıştır. Karar ağaçları sınıflandırma için kullanılabileceği gibi, regresyon problemleri için de kullanılabilirler.

Karar ağaçlarında regresyon hesaplamaları da aşağıdaki örnekte benzer şekilde oluşmaktadır. Bu örnekte bilgisayar oyunu oynamayı seven kişileri tahmin etmek istediğimizde aşağıdaki gibi iki tane karar ağacı oluşturabiliriz. Burada elimizde 2 sınıf veri olduğunu varsayıyoruz. Bunların biri yaş birisi günlük bilgisayar kullanımı.



Şekil 1.6 : Karar Ağacı Topluluk Öğrenme Yöntemi (Chen a.e., 2016)

Topluluk öğrenme yöntemi birçok karar ağacı ile tek bir sonuca ulaşmak için kullanılan yöntemlerdir. Bu örnekte her bir karar ağacı topluluğu ilgili değişkene göre olasılıkları örneklerimizden çıkartır. Daha sonrasında iki karar ağacındaki olasılıklar toplanarak son sonuçlar elde edilir. Bu tip modellerde değişken kadar birbirinden ayrı yapılar kurulup en sonunda ulaşılmak istenen sonuca göre tekrar birleştirme yapılmaktadır. Topluluk öğrenme algoritmaları birçok değişkeni ayrı ayrı topladığı için temelde kullanılan tekil karar ağaçlarına göre daha başarılı sonuçlar verirler. Topluluk öğrenme yöntemlerden en çok kullanılanları bagging ve boosting algoritmalarıdır.

$$\hat{y}_i = \sum_{k=1}^K f_k(\hat{x}_i), f_k \in \mathcal{F} \quad (1.3)$$

Matematiksel olarak topluluk öğrenme yöntemi yukarıdaki formül ile ifade edilir. Formül 1.3’de K toplam karar ağacı sayısını, f olasılık fonksiyonunu ve $\mathcal{F}$ mümkün olan bütün karar ağacı kümesini göstermektedir.

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \quad (1.4)$$

Temel olarak karar ağaçları ile regresyon bu matematiksel model ile ifade edilir. Formül 1.4’deki l fonksiyonu, tahminimizle hedeflediğimiz doğru değer arasındaki farkı ifade eder. Ω fonksiyonu da modelin karmaşık (complexity) kalan kısmının düzeltme fonksiyonunu ifade eder.

Karar ağacı makine öğrenmesi modellerinin iki ana çalışma yöntemi vardır. Bunlar bagging ve boosting olarak adlandırılmaktadır. Bagging yapısı birbirinden bağımsız yapılar kurarak paralel olarak çalışır. Rastgele orman (random forest) modelleri bagging mantığı ile çalışan modellerdir. Boosting yapısı ise bağımsız paralel yapılar yerine, sırayla birbirinin üzerine eklenen yapıya sahiptir. Bu tahmin yapma yöntemleri, bir önceki tahminlerin hatalarından öğrenilen mantığı geliştirerek sonuca daha çok yaklaşmayı hedefleyen modellerdir. Bu modellerde her seferinde bir öncekinin hatalarını göz önünde bulundurarak ilerlediği için paralel çalışan modellere göre daha hızlı çalışırlar. Boosting yöntemi, bagging yönteminden farklı olarak paralel karar ağaçlarının birleştirilmesi yerine her üretilen karar ağacı, sırayla hatayı minimize edecek şekilde çalışır. Temelde, bu tip modellerde her yeni karar ağacı belirli ağırlık oranlarında bir öncekinin üzerine eklenir ve model her adımda iyileştirilerek istenen hedefe doğru ilerler.

1.4.2.2 Random Forest

Random Forests ile sınıflandırma, regresyon veya kümeleme yapılabilir. Random Forests, topluluk öğrenme yöntemi ile çalışan modellerdir. Tekil olarak oluşturulan birçok karar ağacı bir araya getirilerek karar ormanı

oluşturulur. Oluşturulan her karar ağacı veri setinden bootstrap yöntemi ile birbirinden ayrı örneklerle seçilmektedir. Karar ağaçlarındaki tahminler çeşitli yöntemlerle (ortalama alma vb.) bir araya getirilerek son tahmin çıktısı hesaplanır.

Random Forests yönteminde sonradan gelen veriye ait tahmin yapılmasının yanında, değişkenlerin önem derecesi de hesaplanmaktadır. Veri setinde çok sayıda değişken varsa, değişken önem derecesinin hesaplanması model indirgemesi açısından oldukça kullanışlıdır. Örneğin binlerce değişkenin bulunduğu veri setinde, Random Forests yöntemiyle elde edilen önem derecesine göre, kurulacak yeni modelde önem derecesi yüksek değişkenler kullanılarak daha etkin tahminlerin yapılması sağlanabilir (Akman, 2011).

Random Forest modellerinin çalıştırılması için 2 tane parametre seçilmesi gerekmektedir. Bu parametreler, karar ağacında en iyi dal bölünmesini sağlamak için düğümde kullanılan değişkenlerin sayısı ve oluşturulacak karar ağacı sayısıdır. Bu yöntemin bazı dikkat çekici üstün yönleri ve kısıtları aşağıda sıralanmıştır.

Üstün Yönleri;

- Kayıp verilere ve aşırı öğrenme problemlerine karşı güçlü modeller oluşturur.
- Çok fazla değişkene sahip veri setlerinde de kullanılabilir.
- Sadece 2 parametre belirlenir.

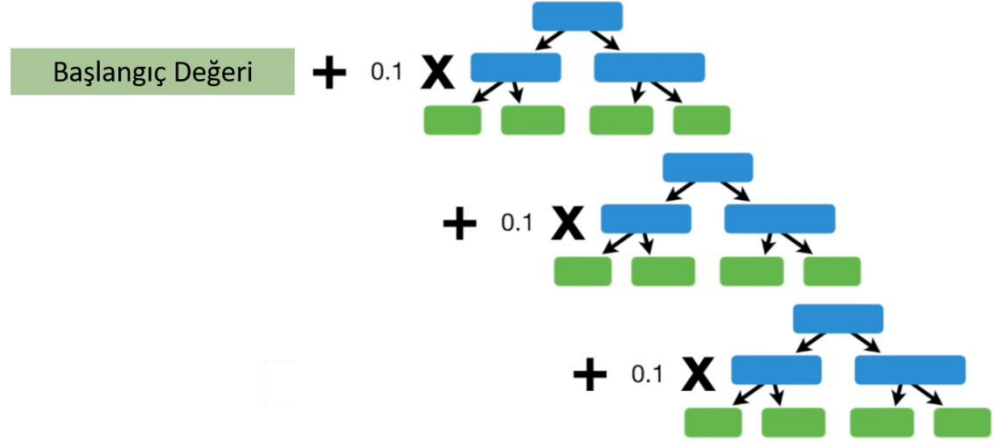
Kısıtları;

- Karar ağaçlarında budama yapılmaması performansı etkilemektedir.

1.4.2.3 Gradient Boosting

Gradyan artırma (Gradient Boosting), hem regresyon hem de sınıflandırma problemleri için bir makine öğrenmesi tekniğidir. Gradyan artırma modelleri karar ağaçlarının makine öğrenmesi ile birleşmesiyle oluşmuştur. Temelde bu algoritmalar karar ağacı oluştururlar ve bütün makine öğrenme algoritmalarında olduğu gibi çıkan

hataları en düşük seviyeye indirmeyi amaçlarlar. XGBoost algoritması gradyan artırma modelidir ve boosting algoritmasıyla çalışmaktadır.



Şekil 1.7 : Gradient Boosting

Gradyan artımlı modeller görselleştirildiğinde yukarıdaki şekildeki örnek verilebilir. Gradient Boosting yönteminde, Random Forests modelinden farklı olarak şekilde olduğu gibi karar ağaçlarının birbiri üzerine eklenmesiyle ilerlemektedir. Model, verilerimize uygun herhangi bir başlangıç değeriyle başlatılır. Daha sonrasında her yeni oluşturulan karar ağacı belirli oranlarda üst üste eklenerek, çıkarılan hatalar minimize edilene kadar model geliştirilmeye devam edilir.

$$\mathcal{L}^{(t)} = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \quad (1.5)$$

Gradyan artımlı modeller temel olarak formül 1.5'deki matematiksel modelle ifade edilir. Bu formülde bir önceki formüle göre f_t eklenmiştir. Bu makine öğrenmesindeki gelişimdeki hatayı minimize edene kadar eklenen fonksiyonları ifade eder. Her bir adımda yeni bir fonksiyon ilave edilerek istenen amaca doğru ilerlenir.

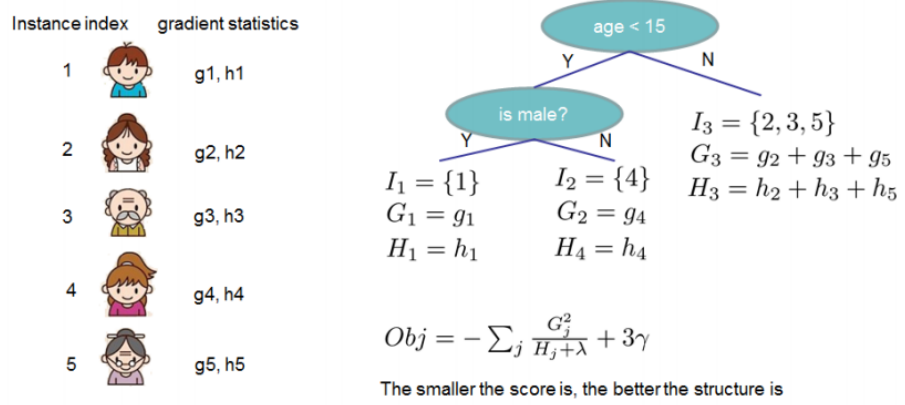
$$\begin{aligned}
\hat{y}_i^{(0)} &= 0 \\
\hat{y}_i^{(1)} &= f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i) \\
\hat{y}_i^{(2)} &= f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i) \\
&\dots \\
\hat{y}_i^{(t)} &= \sum_{k=1}^t f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i)
\end{aligned} \tag{1.6}$$

Daha önce yukarıda belirtilen f fonksiyonunun hesaplaması bu şekildedir. Burada her bir öğrenilen modeli, yeni eklenen karar ağacı ile düzeltilmesi şeklinde ilerlemektedir.

Modelimizin son kısmı $\Omega(f)$ ile ifade edilen karmaşık (complexity) olarak ifade edilen, düzeltme terimidir. Her makine öğrenmesi modelinde olduğu gibi burada da modelin düzeltme kısmı bulunmaktadır. XGBoost modelinde bu formül 1.7 ile hesaplanmaktadır.

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2 \tag{1.7}$$

Karar ağaçlarında önemli faktörlerden bir tanesi de ağacın dallarının nasıl ayrılacağıdır. Bunun için birçok yöntem kullanılmaktadır. XGBoost modelinde yöntem aşağıdaki şekildedir.



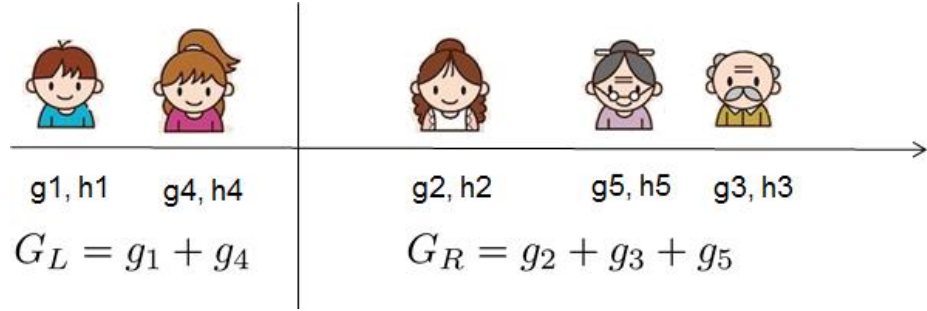
Şekil 1.8 : Karar Ağacının Dallara Ayrılması (Chen a.e., 2016)

$$g_i = \partial_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}) \quad (1.8)$$

$$h_i = \partial_{\hat{y}_i^{(t-1)}}^2 l(y_i, \hat{y}_i^{(t-1)}) \quad (1.9)$$

Bu fonksiyonda g_i kayıp fonksiyonunun birinci dereceden diferansiyelini, h_i ise ikinci dereceden diferansiyelini ifade eder. XGBoost modelinde dalların optimum ayrılması için aşağıdaki formül kullanılmaktadır. Burada hesapladığımız değerler ne kadar küçükse karar ağacının yapısı o kadar kalitelidir.

$$Kazanç = \frac{1}{2} \left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right] - \gamma \quad (1.10)$$



Şekil 1.9 : Karar Ağacı Dal Ayrımı (Chen a.e., 2016)

Bu formül 1) yeni oluşturulan dal ayrımında sol taraftaki yaprağın puanı 2) yeni oluşturulan dal ayrımında sağ taraftaki yaprağın puanı 3) orijinal yapraktaki puan 4) ek yapraktaki düzeltme terimi olarak ifade edilebilir. Eğer kazanç γ 'den küçükse, yeni dalı eklememenin daha iyi olduğu anlaşılır. Bu yapılan işleme karar ağacı modellerinde budama (pruning) adı verilir. XGBoost modelinde, bu yöntem kullanılmaktadır. Bu yöntemle karar ağaçlarının dallarının aşırı ve gereksiz büyümesi engellenmiş olur.

XGBoost modelleri gradyan artırmadan performansının çok daha yüksek seviyede olmasının başlıca nedenlerden bir tanesi de karar ağaçlarını kullanan bu algoritmalarda aşağıya doğru uzanan ağaç dallarını sınırlandırmasıdır. Bu sınırlandırmayı budama ile yapmaktadır. Karar ağacındaki dallar aşağıya doğru baştan belirlenen sınırlandırmayı geçerse algoritma bu dalları geriye doğru kısaltır. Bu yöntem sayesinde karar ağacı çok büyüyüp karmaşıklaşmadığı için performans klasik yöntemlere göre çok daha yüksek olmaktadır. İkinci olarak da XGBoost kütüphanesi kullanılan bilgisayardaki donanımı en yüksek seviyede kullanılabilecek şekilde tasarlanmıştır.

Bu modellerin performansının çok yüksek olması, günlük kullandığımız bilgisayarımızda çeşitli parametrelerde birçok deneme yapabilmemizi de sağlamaktadır. Bu çalışmada Python programında çalışma yapılmış ve (XGBoost Kütüphanesi, 2019) 'sinden faydalanılmıştır.

İKİNCİ BÖLÜM

LİTERATÜR İNCELEMESİ

2.1 Borsa İstanbul ve Makine Öğrenmesi Üzerine Yapılan Çalışmalar

Literatür incelemesinde öncelikle tez çalışmasının da amacı olan Borsa İstanbul üzerine yapılan ilgili çalışmalar ile makine öğrenmesi modelleriyle hisse senetleri üzerine yapılan çalışmalar incelenmiştir. Daha sonrasında ise benzer çalışmalarda kullanılan ve analiz edilen veriler, zaman dilimleri ve modellere göre sınıflandırmalar incelenmiştir.

Gradyan artıma yöntemi ilk olarak Friedman (2001) tarafından geliştirilmiştir. Daha sonrasında Chen (2016:785-794) gradyan artırma yönteminde bazı geliştirmelerle 2016 yılında XGBoost yöntemini makalesiyle paylaşmıştır. XGBoost, en az seviyede kaynakla gerçek problemleri çözme yeteneğine sahiptir Chen (2016:785). Bu çalışmayla birlikte XGBoost yöntemiyle zaman serilerinde tahmin etme çalışmaları yapılmıştır. Bunlardan bir kısmı aşağıda paylaşılmıştır. Zaman serilerinde yapay sinir ağları modelleri kullanılarak daha fazla çalışma yapıldığı görülmüştür.

Avcı (2009:443-461) çalışmasında BİST 30’da bulunan 5 tane hisse ile çalışma yapmış ve yapay sinir ağları ile al-tut stratejine göre daha iyi sonuçlar elde etmiştir.

Tsai (2009) çalışmasında karar ağaçları ve yapay sinir ağlarının birleşimiyle oluşan hibrit modelden yüzde 77’lik bir performansla sadece yapay sinir ağlarına göre daha iyi bir sonuç alındığı tespit etmiştir.

Uğur (2009) çalışmasında finansal oranları, makroekonomik değişkenler ve hisse senedi geçmiş fiyat verileri ile sanayi, mali ve hizmet sektöründeki hisselerin fiyatları tahmin etmeye çalışmıştır. Çalışmada tek ve iki yönlü sabit etkiler panel veri ve Arellano-Bond ile Arellano-Bover dinamik panel veri regresyon modelleri kullanılmıştır. Çalışmanın sonucunda bu fiyatları belirleyen önemli değişkenlerden bazılarının piyasa değeri / defter değeri oranı, enflasyon oranı, faiz oranı, hisse başına

kazanç olduğu ortaya çıkmıştır. Sektörler arasında bu parametreler değişiklik gösterebilmektedir

Yıldırım (2011:4076-4091) çalışmada Borsa İstanbul'da yer alan kağıt sektörü üzerine aylık bazda analizler yapmıştır. Lineer regresyona göre yapay sinir ağları ile daha iyi sonuç alınmıştır. Veri setine altın fiyatları, borsa hacmi ve faiz oranları da eklenmiştir.

Kara (2011:5311-5319) çalışmasında BİST 100 endeksinin günlük fiyatlarının yönünü tahmin etmeyi hedeflemiştir. Bu çalışmada iki farklı model olarak yapay sinir ağları ve destek vektör makineleri (SVM) kullanılmış ve birbirleriyle karşılaştırılmıştır. Yapay sinir ağları modeliyle yüzde 75,74'lük ve destek vektör makineleri ile yüzde 71,52'lik doğruluk payı yakalamıştır. Yapay sinir ağları modeli ile daha iyi bir sonuç elde etmiştir. Veri setinde belirli teknik indikatörler kullanmıştır.

Güresen (2011:10389-10397) çalışmasında 146 günlük öğrenme verisi kullanarak 36 günlük NASDAQ endeksini tahmin etmeyi amaçlamıştır. Performans ölçütü olarak MSE ve MAD kullanmıştır. Çalışmanın sonucu olarak zaman serisi tahmin modellerinde, klasik yapay sinir ağı modelinin kurmuş olduğu hibrit modellere göre daha iyi sonuç verdiğini ifade etmektedir.

Shen (2012) çalışmasında destek vektör makinesi (SVM) ile borsa endeksi trendi tahmin etmeye çalışmıştır. Sonuç olarak NASDAQ endeksi için yüzde 74, S&P500 endeksi için yüzde 76 ve DJIA endeksi için yüzde 77'lik bir doğruluk oranında tahmin yapılmıştır.

Kılıç (2014:759-765) çalışmasında BİST 100 endeksinin belli bir kalıba göre hareket ettiği ve piyasanın etkin olmadığı sonucuna varılmıştır.

Patel (2015:259-268) çalışmasında Hindistan borsa endekslerinden CNX Nifty ve S&P Bombay Stock Exchange (BSE) Sensex endeksleri üzerine yapılmıştır. 10 senelik günlük en yüksek, en düşük, açılış ve kapanış fiyatları veri setinde kullanılmış ve teknik indikatörler de buraya dahil edilmiştir. Yapay sinir ağları, destek vektör makinesi (SVM), random forest (rasgele orman) ve naive-bayes modelleri kurulmuş ve karşılaştırılmıştır.

Laboissiere (2015:66-74) çalışmasında Brezilya'nın 3 tane enerji dağıtım şirketinin hisse senetlerinin günlük en yüksek ve en düşük fiyatlarını tahmin etmeyi amaçlamıştır. Bu çalışmada yapay sinir ağları modelleri kullanılmıştır. Performans ölçütleri olarak da MAE, MAPE ve RMSE ile ölçüm yapıp karşılaştırmalar sunulmuştur. Çalışmanın tahmin etme veri setine günlük fiyatlara ek olarak dolar kuru da eklenmiştir. Çalışmanın sonunda MAPE performans kriterine göre en yüksek fiyat tahmininde yüzde 0,9 değerinden daha düşük olduğu sonucuna ulaşmıştır.

Şahin (2016) çalışmasında BİST100 endeksini tahmin etmeyi hedeflemiştir. Teknik indikatörlerin ve dolar, petrol fiyatı gibi verilerden iyi ayrı veri seti oluşturmuş ve birbirleriyle kıyaslamıştır. Ayrıca bu veri setleriyle yapay sinir ağları ve bulanık mantık modelleri oluşturmuş ve karşılaştırmıştır. Yapay sinir ağlarının daha iyi tahmin etme kabiliyetine sahip olduğu sonucuna varmıştır.

Dey (2016) çalışmasında Apple ve Yahoo şirketlerinin hisse senedi fiyatlarının trendlerini makine öğrenmesi modeli ile tahmin etmeyi hedeflemiştir. Makine öğrenmesi modellerinden eXtreme Gradient Boosting (XGBoost) modelini kullanmıştır. Kurduğu modellerde 60 ve 90 günlüğün periyotlar için yüzde 87'lik bir doğruluk oranına ulaşmıştır. Çıkardığı sonuçlardan diğer modellere göre önemli derecede daha başarılı sonuçlar elde ettiğini ifade etmektedir.

Jiao (2017) çalışmasında S&P 500 endeksindeki hisseler ve içinde bulunduğu hisselerin yönünü tahmin etmeyi hedeflemiştir. Veri setinde faiz oranı gibi çeşitli makroekonomik veriler kullanmıştır. Çalışmada yapay sinir ağları ve gradyan artırma yöntemleriyle birlikte çeşitli makine öğrenmesi modelleri kullanılmıştır. S&P 500 endeksinin tahmin edilmesinde diğerlerine göre daha iyi sonuçlar almıştır.

Gündüz (2017:4829-4840) çalışmasında Borsa İstanbul hisselerinden Garanti, Türk Hava Yolları ve İş Bankası şirketlerinin günlük kapanış fiyatları tahmin edilmeye çalışılmıştır. Veri setinde çeşitli teknik indikatörler kullanılmıştır. Gradyan artırma makinesi ile diğer kullandığı yöntemlere göre daha iyi sonuçlar elde etmiştir.

Shukla (2017) çalışmasında Hindistan borsa endekslerinden NSE Nifty 50 tahmin edilmesi amaçlanmıştır. Performans ölçütü olarak RMSE kullanılmıştır. Yapay sinir ağlarıyla, ARİMA modeline göre daha iyi sonuç alınmıştır.

Erkartal (2018) çalışmasında Garanti Bankası hisse senedinin 5 gün sonrası için alım satım sinyalleri oluşturmuştur. Bu sinyaller için karar ağacı, destek vektör makineleri ve geri yayımlı sinir ağları modelleri ile tahminler üretmiştir. Yüzde 80'lik bir performansla karar ağacı modeli daha iyi sonuç vermiştir.

2.2 Makine Öğrenmesi ile Modellenmiş Zaman Serisi Çalışmalarının Sınıflandırılması

Krollner (2010) çalışmasında zaman serisi tahminlerinde makine öğrenmesi yöntemlerinin kullanımını araştırmıştır. Bu çalışmada 2000 yılından 2009 yılına kadarki birçok çalışmayı kategorize etmiştir. Bu kısımda, kullanılan makine öğrenmesi çeşitleri, kullanılan zaman periyotları ve kullanılan verilere göre çalışmalar sınıflandırılmıştır.

Tablo 2.1 : Makine Öğrenmesi Modellerine Göre Çalışmaların Sınıflandırılması
(Krollner a.e., 2010)

Teknoloji	Toplam Çalışma Sayısı
Yapay Sinir Ağları Tabanlı	21
Geliştirme Yapılmış ve Optimize edilmiş Teknikler	4
Hibrit Modeller	15
Diğer	6

Makine öğrenmesi yöntemlerinden en çok yapay sinir ağları ve hibrit modeller üzerine çalışma yapılmıştır.

Tablo 2.2 : Zaman Periyotlarına Göre Çalışmaların Sayısı (Krollner a.e., 2010)

Zaman	Çalışma Sayısı
Günlük	31
Haftalık	3
Aylık	3
Çoklu/Diğer	9

Kurulan zaman serisi çalışmalarında diğer kullanılan periyotlara göre en çok günlük bazda çalışma yapıldığı görülmektedir.

Tablo 2.3 : Kullanılan Verilere Göre Çalışmaların Sayısı (Krollner a.e., 2010)

Veri	Çalışma Sayısı
Gecikmeli endeks verisi	35
Hacim	4
Teknik İndikatörler	13
Petrol Fiyatı	4
S&P 500 / NASDAQ / Dow Jones (ABD harici çalışmalar)	4
İşsizlik Oranı	1
Para Arzı	3
Döviz Kuru	3
Altın Fiyatı	3
Kısa ve uzun vadeli faiz oranı	6
Diğer	10

Çalışmaların çoğu gecikmeli endeks verisi kullanmıştır. Bu çalışmaların çoğunda ise en yüksek fiyat, en düşük fiyat, açılış fiyatı ve kapanış fiyatı şeklinde kullanılan zaman diliminin içinde oluşan veriler kullanılmıştır.

ÜÇÜNCÜ BÖLÜM

KULLANILAN VERİLER VE PERFORMANS ÖLÇÜTLERİ

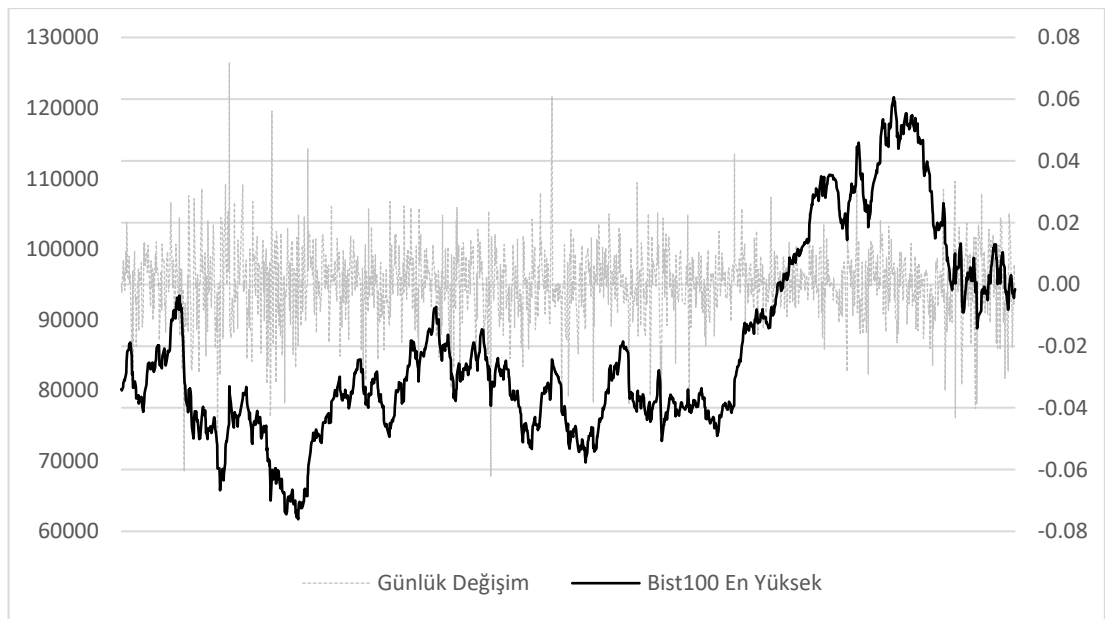
3.1 Kullanılan Veriler

Veri setinde kullanılan bütün veriler tahmin yapılması hedeflenen BİST100 fiyatının oluştuğu günlere göre uyarlanmıştır. BİST100 endeksinin kapalı olduğu günler için veri setlerinde ayrıştırma yapılmıştır. Çalışmada kullanılan veri setleri Borsa endeksleri, ekonomik parametreler, döviz kurları ve emtialar olarak 3 ana grupta incelenmiştir.

3.1.1 Borsa Endeksleri

BİST100 Endeksi

Borsa İstanbul Pay Piyasası için temel endeks olarak kullanılmaktadır. Yıldız Pazar ve Ana Pazar'da işlem gören şirketlerle, Kolektif Yatırım Ürünleri ve Yapılandırılmış Ürünler Pazarı'nda işlem gören gayrimenkul yatırım ortaklıkları ve girişim sermayesi yatırım ortaklıkları arasından seçilen 100 paydan oluşmakta olup, BİST 30 ve BİST 50 endekslerine dahil payları da kapsar (Borsa İstanbul, 2019).



Şekil 3.1 : BİST100 Günlük En yüksek Fiyat Değişim Grafiği

Borsa İstanbul'un günlük olarak 5 temel verisi oluşur ve yayınlanır. Bunlar Borsa'nın günlük açılış fiyatı, kapanış fiyatı, en yüksek fiyat, en düşük fiyat ve toplam hacimdir. Bu çalışmada bu verilerin hepsi geçmişe yönelik modele eklenmiştir.

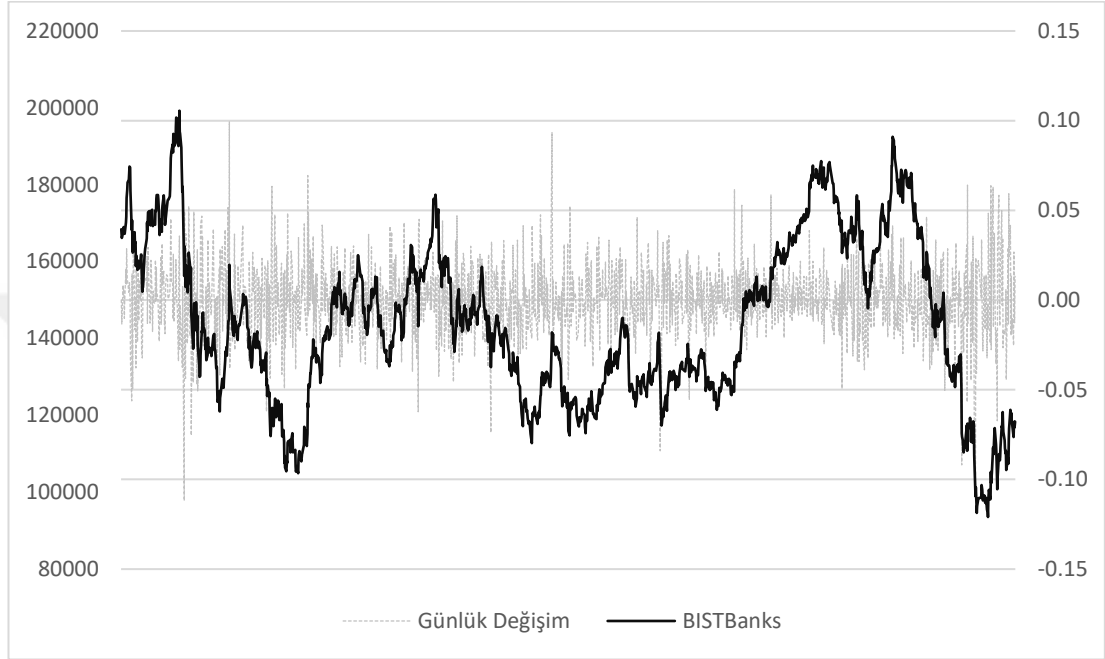
BİST 30, BİST 50 ve BİST 100 endeksi paylarının seçimi ve endekslere alınma kriterleri:

- Payların endekslere alınabilmesi için değerlendirme dönemi sonu itibarıyla A veya B listesinde yer almaları şarttır. C ve D listesinde yer alan paylar endekslere dahil edilmemektedir (Borsa İstanbul, 2019).
- BİST 30, BİST 50 ve BİST 100 endekslerinin kompozisyonu Yıldız Pazar ve Ana Pazar'da işlem gören şirketlerle, Kolektif Yatırım Ürünleri ve Yapılandırılmış Ürünler Pazarı'nda işlem gören gayrimenkul yatırım ortaklıkları ve girişim sermayesi yatırım ortaklıklarının paylarından oluşmaktadır (Borsa İstanbul, 2019).
- Payların endekslere alınabilmesi için, değerlendirme dönemi sonu itibarıyla Borsa İstanbul'da en az 60 gün süreyle işlem görme şartı aranmaktadır. Halka arz edilen kısmın piyasa değeri 1 milyar TL'den veya Yıldız Pazar ve Ana Pazar'daki fiili dolaşımdaki payların toplam piyasa değerinin %1'inden daha büyük olması nedeniyle değerlendirme dönemi içinde endekslere dahil edilen paylar için ise 60 gün süreyle işlem görme şartı aranmaz (Borsa İstanbul, 2019).
- Birden fazla grup payı ayrı sıralarda işlem gören şirketlerin, sadece bir grup payı endekslere dahil edilmektedir (Borsa İstanbul, 2019).

Endekslerde yer alacak payların seçimi piyasa değeri ve günlük ortalama işlem hacmine göre belirlenir. Paylar, değerlendirme döneminin sonunda fiili dolaşımdaki payların piyasa değerlerine ve değerlendirme dönemindeki günlük ortalama işlem hacimlerine göre büyükten küçüğe doğru sıralanarak seçilir.

BİST Bankacılık Endeksi

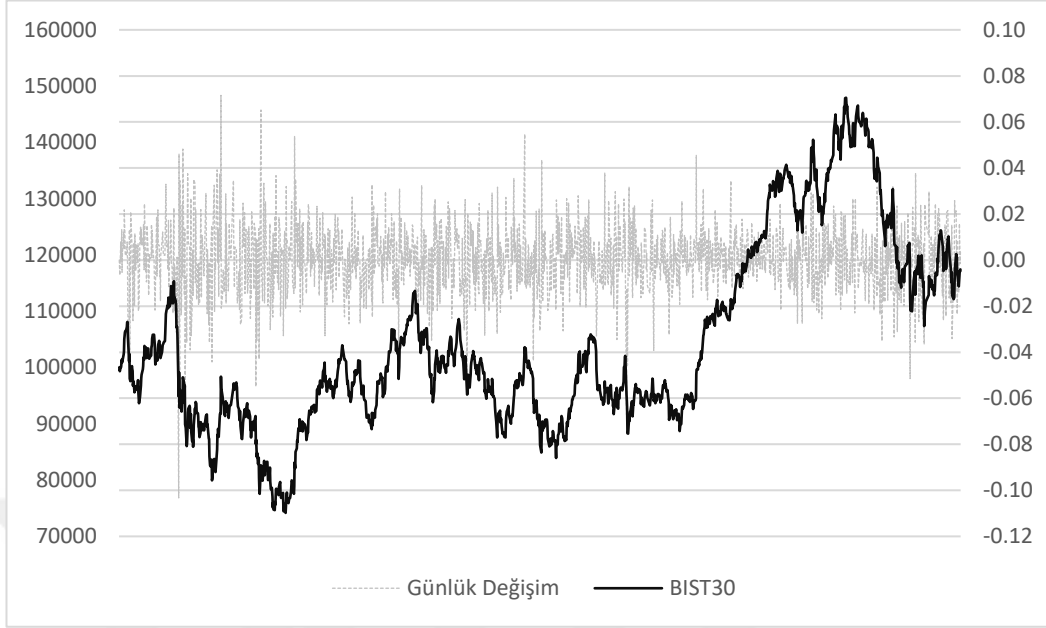
Borsa İstanbul Pay Piyasası bankacılık sektörü için temel endeks olarak kullanılmaktadır.



Şekil 3.2 : BİST Bankacılık Endeksi Günlük Değişim Grafiği

BİST30

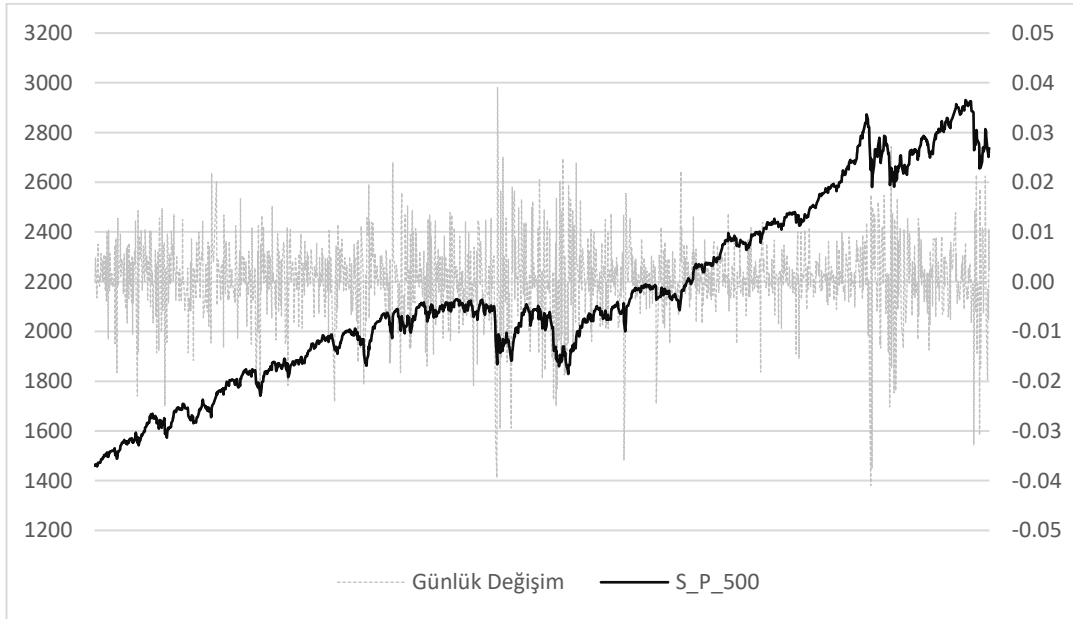
Yıldız Pazar ve Ana Pazar’da işlem gören şirketlerle, Kolektif Yatırım Ürünleri ve Yapılandırılmış Ürünler Pazarı’nda işlem gören gayrimenkul yatırım ortaklıkları ve girişim sermayesi yatırım ortaklıkları arasından seçilen 30 paydan oluşur (Borsa İstanbul, 2019).



Şekil 3.3 : BİST30 Endeksi Günlük Değişim Grafiği

S&P500 Endeksi

Amerika Birleşik Devletleri hisselerinin en büyük 500 hissesini kapsayan endekstir. Amerika Birleşik Devletleri hisselerinin yaklaşık %70 - %80'ini kapsar. Bu borsa endeksi, Standard & Poor's tarafından yapılmaktadır. Standard & Poor's, Amerika Birleşik Devletleri merkezli uluslararası kredi derecelendirme kuruluşudur.

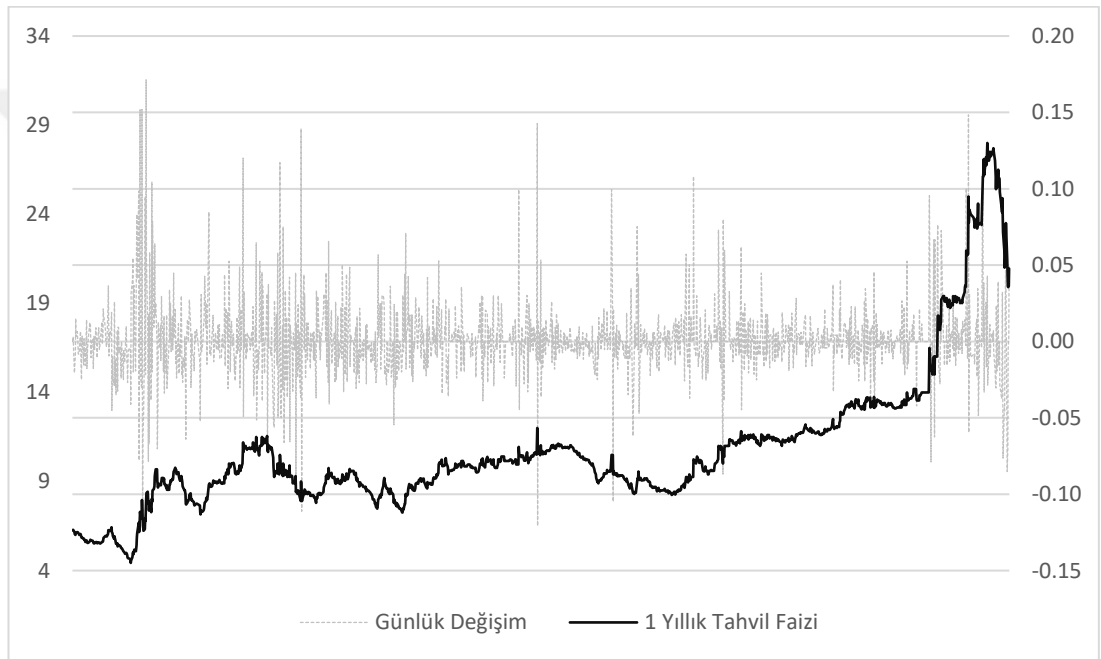


Şekil 3.4 : S&P500 Endeksi Günlük Değişim Grafiği

3.1.2 Ekonomik Parametreler

1 Yıllık Tahvil Faizi

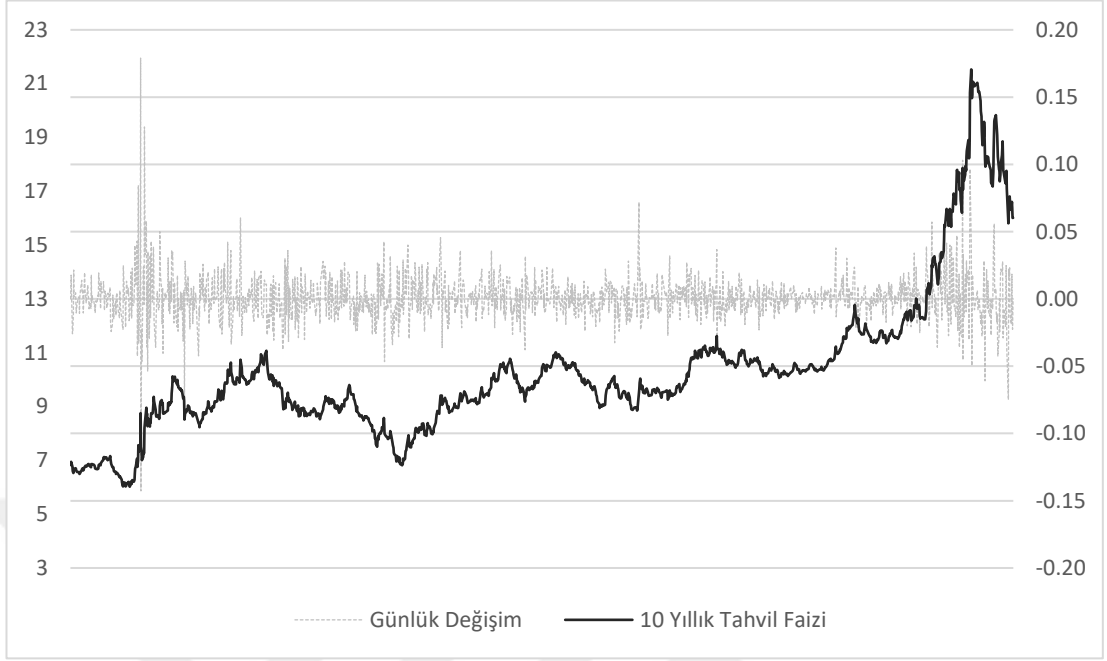
Kupon ödemesine az kalmış olan devlet tahvilleri arasında en çok işlem gören (veya son üç işlem günü içinde en çok işlem gören) devlet tahvilinin faizidir. 1 yıllık faizler için gösterge niteliği taşır.



Şekil 3.5 : 1 Yıllık Tahvil Faizi Günlük Değişim Grafiği

10 Yıllık Tahvil Faizi

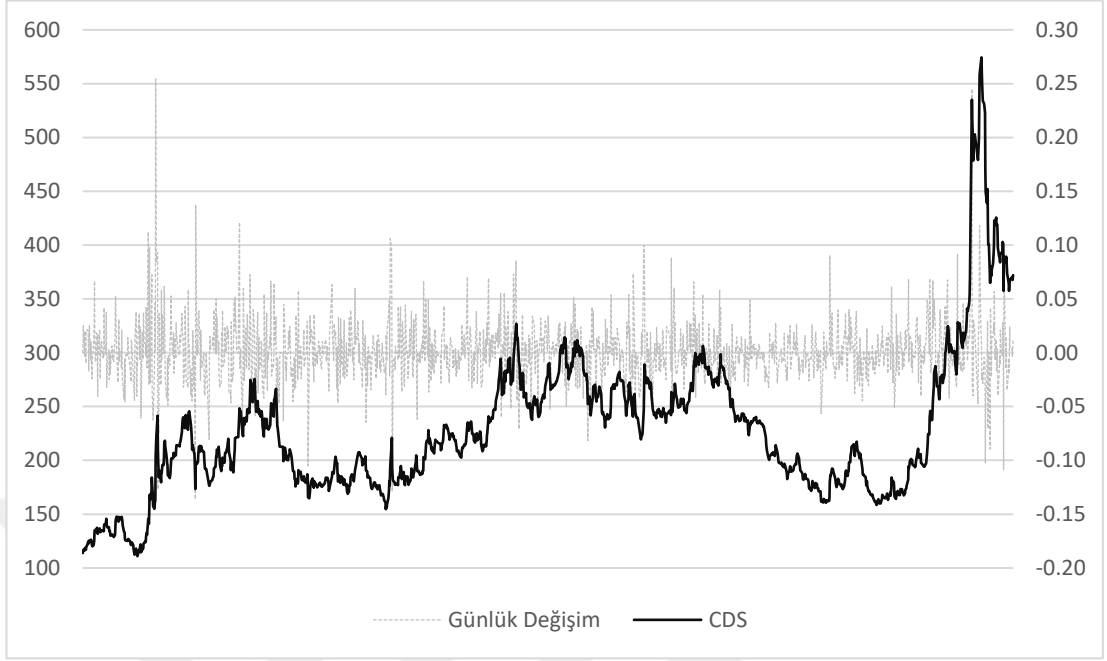
10 yıllık devlet tahvilleri arasında en çok işlem gören devlet tahvilinin faizidir. 10 yıllık faizler için gösterge niteliği taşır. Uzun vadeli olduğu için anlık merkez bankası faiz değişimlerinde 1 yıllıklara göre daha geç tepki verir.



Şekil 3.6 : 10 Yıllık Tahvil Faizi Günlük Değişim Grafiği

CDS

CDS (Credit Default Swap) dilimize çevrilmiş hali kredi temerrüt takasıdır. Kredi temerrüt takasları sigortacılık faaliyetlerinde kullanılan ve kredinin temerrüde uğraması haline karşılık bir sigorta çeşididir. Bir ülkenin kredi temerrüt takası ise o ülkeye kullanılacak bir kredinin sigortalanması için alınacak oran, aynı zamanda ülkenin yabancı yatırımcı gözündeki risk primini göstermektedir. Türkiye'nin 5 yıllık CDS verisi de gösterge olarak kullanılmakta ve JP Morgan tarafından yayınlanmaktadır. Bu veri baz puan olarak yayınlanmaktadır.

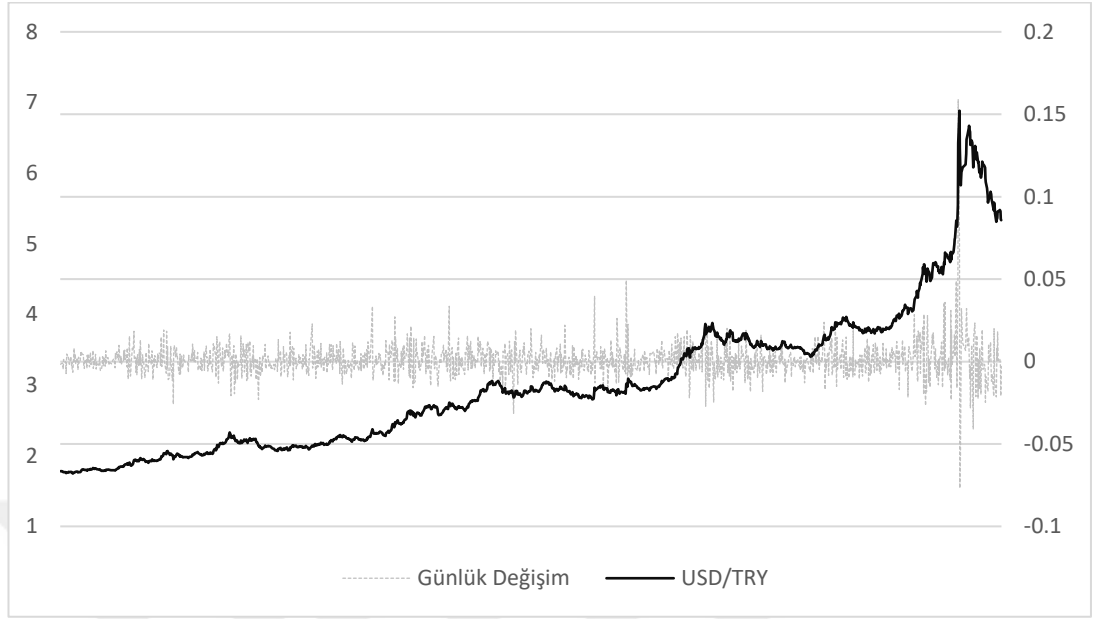


Şekil 3.7 : CDS Günlük Değişim Grafiği

3.1.3 Döviz Kurları ve Emtialar

USD/TRY Kuru

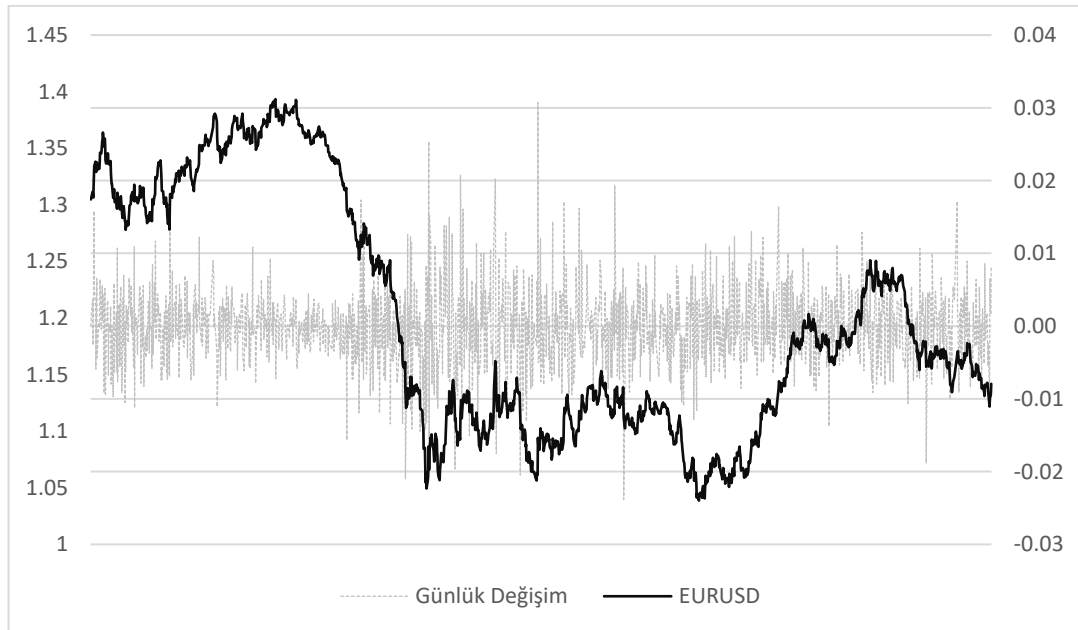
Dolar TL paritesini ifade eder. Özellikle gelişmekte olan ülkeler için yapılan çalışmalarda veri olarak önemli bir yere sahiptir. Amerikan doları, Merkez Bankası tarafından rezerv para olarak kabul edilmektedir. Dünyada uluslararası ödeme aracı olarak kullanılması, kullanıma hazır ve diğer para birimlerine çevrilebilme özelliğinin yüksek olması rezerv niteliğini artırmaktadır.



Şekil 3.8 : USD/TRY Günlük Değişim Grafiği

EUR/USD Kuru

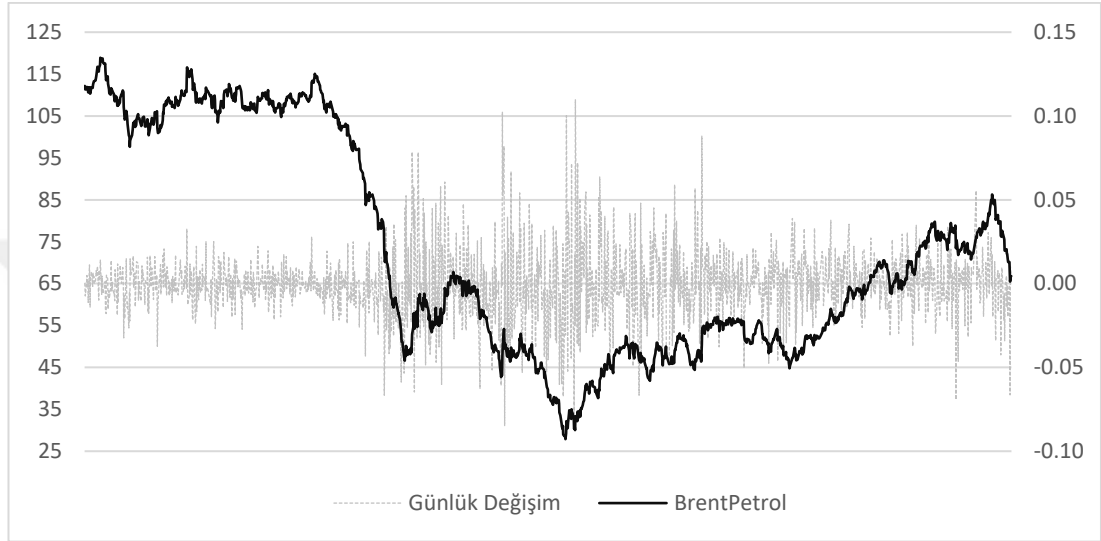
Euro, dünyada dolardan sonra en çok rezervi bulunan para birimidir. Euro, Avrupa Birliği ülkelerinin ortak kullandığı para birimidir. Bu para birimi 1 Ocak 1999’da bankacılık işlemlerinde kullanılmaya başlamıştır. Çalışmada Euro para biriminin dolar cinsinden paritesi veri setine eklenmiştir.



Şekil 3.9 : EUR/USD Günlük Değişim Grafiği

Brent Petrol

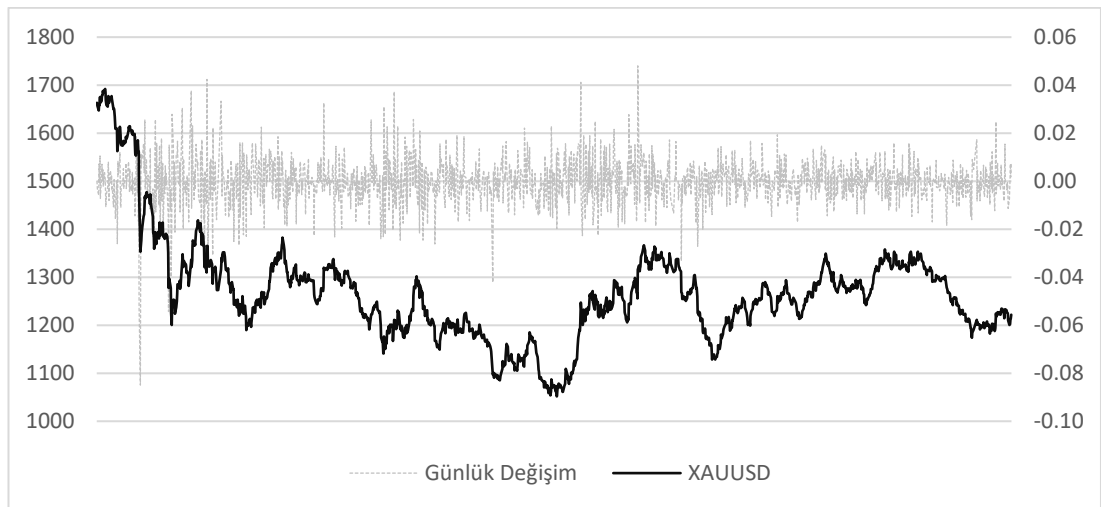
Petrol piyasasında geçerli olan ve en çok kullanılan petrol türü fiyatını ifade eder. Dünyada petrol piyasalarında kullanılan diğer petrol türleri WTI (Crude Oil) ve Umman petrolüdür. Dolar para birimi cinsinden veri setimize dahil edilmiştir.



Şekil 3.10 : Brent Petrol Günlük Değişim Grafiği

Altın

Altın dünyada güvenli liman olarak bilinen bir kıymetli madendir. Altının en büyük özelliği neredeyse dünyadaki bütün merkez bankalarında rezerv olarak tutulmasıdır. Dolar cinsinden ons fiyatı veri setimize dahil edilmiştir.



Şekil 3.11 : XAU/USD Günlük Değişim Grafiği

3.2 Performans Ölçütleri

Modellerimizi karşılaştırırken bir tane performans ölçütü kullanmak yerine birçok ölçüt kullanılması tercih edilmiştir. Seçilen ölçütler literatürde incelendiği şekilde çalışmalarda en yaygın kullanılan ölçütlerdir. Kullanılan ölçütler; ortalama mutlak hata, mutlak hataların standart sapması, ortalama kare hata, sinyal(yön) doğruluğu, ortalama kare hataların kara kökü, ortalama hata ve standart sapmadır.

- **MAE:** Ortalama mutlak hata, tahmin edilen ve gerçekleşmiş değer arasındaki farktır. Bu farklar toplanırken mutlak değer alınır. Bütün farklar toplandıktan sonra toplam miktarına bölünüp ortalama değer hesaplanır.

$$MAE = \frac{1}{n} \sum_{j=1}^n |e_j| \quad (3.1)$$

- **Mutlak Hataların Standart Sapması (SD Absolute Error):** Hataların standart sapma değerini gösterir.

$$SY.x = \sqrt{\frac{\sum_{j=1}^n e_j^2}{n-K \text{ (Parametre Sayısı)}}} \quad (3.2)$$

- **MSE:** Ortalama kare hata, hataların karelerinin toplamının ortalamasını gösterir.

$$MSE = \frac{1}{n} \sum_{j=1}^n e_j^2 \quad (3.3)$$

- **Sinyal Doğruluğu (Sign Accuracy):** Verilen sinyalin doğruluk oranını gösterir. Her tahmin değerinde tam değer bulunamayıp arada bir fark oluştursa da bu değer yüzdelik olarak yapılan tahminlerin yönünün yüzde kaçının doğru olduğunu gösterir.
- **RMSE:** Ortalama kare hataların karekökü (Sqrt of MSE), hataların karelerinin toplamının ortalamasının alınması sonrasındaki karekök değeridir.

$$RMSE = \sqrt{\frac{\sum_{j=1}^n e_j^2}{n}} \quad (3.4)$$

$$RMSE = \sqrt{MSE} \quad (3.5)$$

- **Ortalama Hata (ME):** Ortalama hata, tahmin değerleri ile gerçek değerlerin arasındaki ortalama hatadır. Hata, bu değerlerin arasındaki farktır.

$$ME = \frac{1}{n} \sum_{j=1}^n e_j \quad (3.6)$$

- **Standart Sapma:** Varyans, ortalamadan uzaklıkların karelerinin ortalamasıdır. Varyansın karekökü de standart sapmayı verir. Standart sapma finasta risk ölçütü olarak kullanılmaktadır. Bu çalışmada yatırım stratejilerinin riskini ölçmede kullanılmıştır.

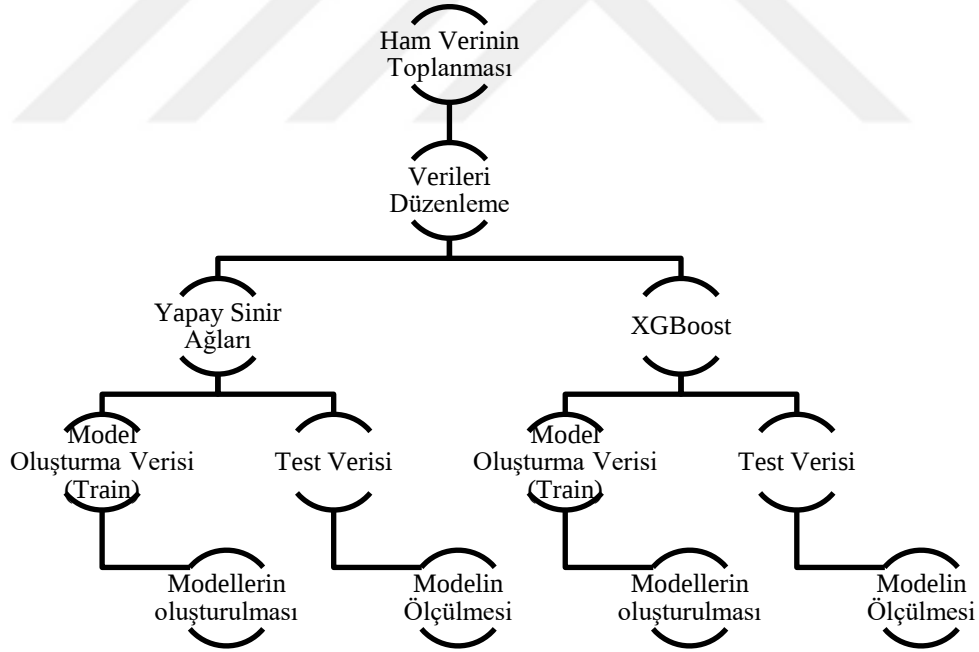
$$\sigma = \frac{1}{n} \sum_{j=1}^n (x_j - \bar{x})^2 \quad (3.7)$$

DÖRDÜNCÜ BÖLÜM

BİST100 ENDEKSİ ÜZERİNE UYGULAMA

4.1 Metodoloji

Makine öğrenmesi birçok programlama dilinde programlanabilmekte ayrıca sistematik olarak geliştirilmiş bazı programlar ile de herhangi bir kod yazmadan makine öğrenmesi modelleri oluşturulabilmektedir. Bu çalışmada, dünyada bu tarz problemlerin çözümünde sıkça kullanılan Python dili kullanılmıştır. Veri setlerinin hazırlanması, çıktıların tablo ve grafik haline getirilmesinde ise Microsoft Excel programından faydalanılmıştır. Genel olarak metodoloji, aşağıdaki şekilde özetlenmiştir. Buna ek olarak çıkan en iyi modellerden yatırım stratejileri oluşturulmuştur.



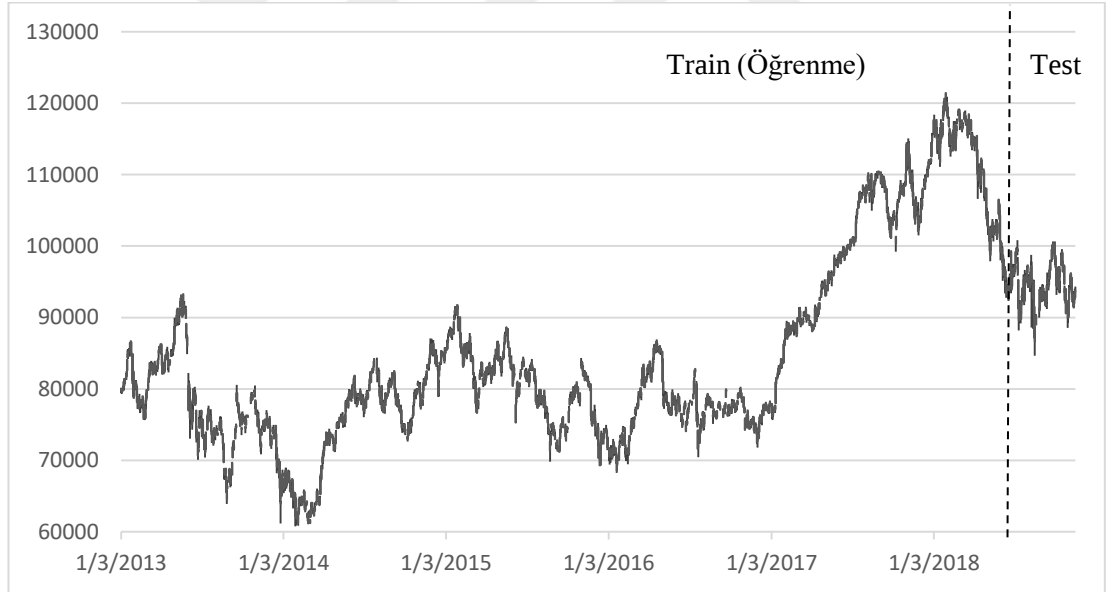
Şekil 4.1 : Metodoloji

4.2 Veri Düzenlemeleri

Bu çalışmada yaklaşık beş yıllık döneme ilişkin günlük veri seti kullanılmıştır. Veriler 3 Ocak 2013 tarihinden başlayıp 16 Kasım 2018 tarihinde son bulmaktadır.

Toplamda 1477 günlük veri kullanılmıştır. Bu verilerin yaklaşık yüzde 90'ı model oluşturma (öğrenme) ve yüzde 10'u da test olarak kullanılmıştır.

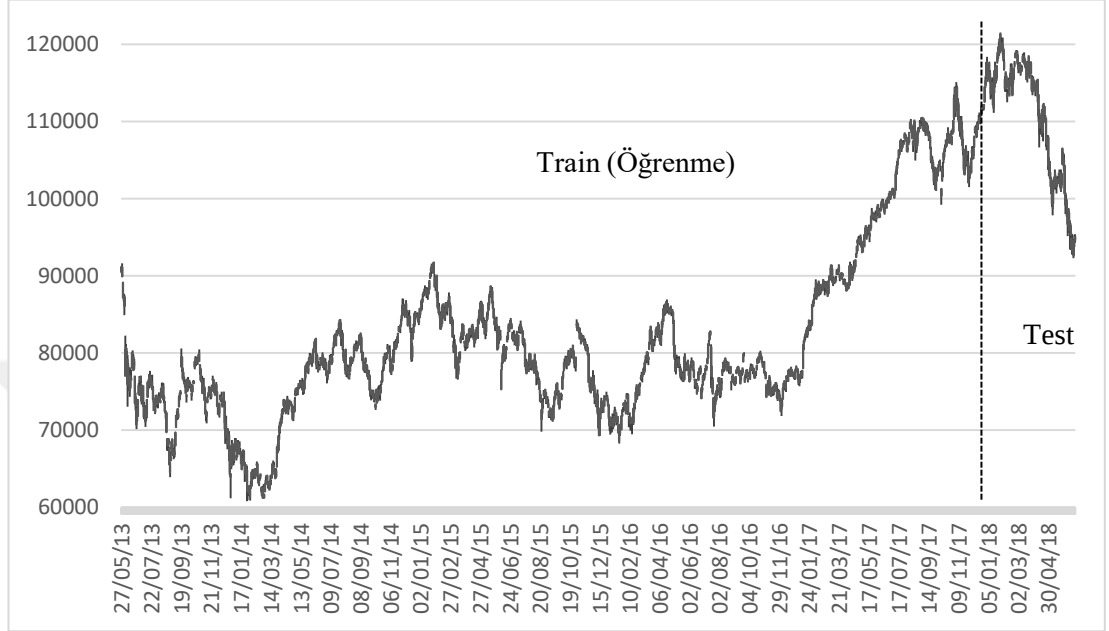
Verilerin düzenlenmesi adımı en önemli adımlardan bir tanesidir. Programlamada çıkan ve çalışmayı yavaşlatan hataların birçoğu verilerin hatalı olmasında kaynaklanmaktadır. Verilerin hepsi zaman serisi şeklinde olduğu için bir önceki günden yüzdelik değişim hesaplanarak modellere girdi olarak verilmiştir. Ayrıca veriler düzenlenirken modellerde hata vermemesi ve hatalı modeller oluşmaması adına bazı düzenlemeler yapılmıştır. Bu düzenlemelerden bir tanesi CDS verisinin bazı ulusal tatil günlerinde yayınlanmamasından kaynaklanmaktadır. Bu veriler ondan önceki en son açıklanan verilerden günlük değişim hesaplanarak düzenlenmiştir. Diğer veri setlerinde de aynı düzenleme yapılmıştır. Ayrıca Borsa İstanbul'un tatil olduğu günlere göre diğer veri setleri de düzenlenmiştir.



Şekil 4.2 : Öğrenme (Train) ve Test Ayrımlı Birinci Veri Seti

Borsa İstanbul endeksi ekonominin çeşitli dönemlerinde farklı trendlerde olabilmektedir. Buraya kadar olan çalışmaya ek olarak, modellerimizin doğruluğundan daha çok emin olabilmek adına, veri setimizin en baş ve en son tarafından 100 günlük veri eksiltilerek yeni bir zaman diliminde tahmin modelleri oluşturulmuştur. Yeni veri setinde 1277 veri bulunmaktadır. Bu verilerin 121 günü test

verisi olarak kullanılmıştır. Veri seti 27.05.2013 tarihinden başlayıp 21.06.2018 tarihinde sona ermektedir.



Şekil 4.3 : Öğrenme (Train) ve Test Ayrımlı İkinci Veri Seti

4.3 Makine Öğrenmesi Modelleriyle Tahmin Yapma ve Performans Karşılaştırmaları

Tez çalışmasının amacı Borsa İstanbul'daki günlük en yüksek fiyatı tahmin etmek ve bu üretilen tahminlerle yatırım stratejisi oluşturmaktır. Çalışmanın ilk aşamasında fiyat tahmini yapılmış ve en iyi modeller seçilmiştir. İkinci aşamasında ise bu seçilen modellerden yatırım stratejileri oluşturulmuştur.

Çalışmada 2 farklı stratejiyle tahmin yapma yöntemi geliştirilmiştir. Birincisi t ve t 'ye kadar olan veriler yani bugünün ve bugünden öncesinin verisi kullanılarak, $t+1$ yani yarınki BİST fiyatında görülecek en yüksek fiyat tahmin edilmeye çalışılmıştır. İkinci olarak ise bu verilere ek olarak o günün açılış fiyatı da dahil edilip yine o günün en yüksek fiyatı geçmiş veriye dayanarak tahmin edilmeye çalışılmıştır. İki veri seti için de aynı çalışma yapılmıştır.

Daha sonraki bölümde bu tahmin edilen en yüksek fiyatlarla 2 farklı yatırım stratejisi oluşturulmuş ve pasif yatırımla elde edilebilecek getiri ile karşılaştırılmıştır.

İlk çalışma olarak birinci veri seti ile ertesi günün en yüksek fiyatı tahmin edilmiştir. Kullanılan veri setleri aynı kalmak suretiyle program çeşitli parametrelerde tekrar tekrar çalıştırılmıştır. Bu çalışmada 1000'den fazla model üretilmiş ve bu modellerden en iyilerinden 8 tanesi karşılaştırma tablosu şeklinde aşağıda sunulmuştur.

Tablo 4.1 : Ertesi Gün En Yüksek Fiyat Tahmini Birinci Veri Seti Performans Sonuçları

No	Yöntem	k	MAE	SD Absolute Error	SA	MSE	Sqrt of MSE	Mean Error
1	XGBoost	18	0.00694	0.00680	0.70776	0.00009	0.00923	-0.00095
2	XGBoost	12	0.00679	0.00656	0.69318	0.00009	0.00944	-0.00051
3	XGBoost	1	0.00694	0.00660	0.66591	0.00009	0.00958	-0.00020
4	XGBoost	25	0.00723	0.00654	0.67890	0.00010	0.00975	0.00026
5	XGBoost	20	0.00738	0.00670	0.71167	0.00010	0.00996	-0.00075
6	Yapay Sinir Ağları	8	0.00774	0.00675	0.52835	0.00011	0.01027	0.00009
7	XGBoost	16	0.00761	0.00696	0.67426	0.00011	0.01031	0.00049
8	Yapay Sinir Ağları	2	0.00783	0.00703	0.57788	0.00011	0.01052	-0.00065

Ölçü birimlerine göre en iyi sonuçlar koyu renkli şekilde ifade edilmiştir. Bu modellerden birinci sıradaki model, birçok ölçüte göre en iyi, birkaçına göre ise de en iyiye yakın sonuçları vermiştir. Dolayısı ile 1 numaralı model test modeli olarak seçilmiştir. Bu model geçmişe dönük 18 günlük veri kullanmaktadır. Daha sonraki bölümde bu model kullanılarak yatırım stratejisi oluşturulacaktır.

Aynı zaman dilimiyle oluşturulan ikinci stratejide kullanılacak veride ise yukarıda da bahsedildiği üzere borsanın açılış fiyatı geldikten sonra o günün en yüksek fiyatının tahmin edilmesi hedeflenmiştir. Bu veri seti kullanılarak da yukarıda olduğu gibi binden daha fazla sayıda model oluşturulmuş ve bunların en iyileri aşağıda ayrıntılı performans ölçüleriyle sıralanmıştır. Buradaki çalışmada da birinci veri seti kullanılmıştır.

Tablo 4.2 : Aynı Gün En Yüksek Fiyat Tahmini Birinci Veri Seti Performans Sonuçları

No	Yöntem	k	MAE	SD Absolute Error	SA	MSE	Sqrt of MSE	Mean Error
1	XGBoost	48	0.005619	0.005407	0.799534	0.000061	0.007798	0.00001
2	XGBoost	29	0.005726	0.005523	0.790805	0.000063	0.007956	-0.00049
3	XGBoost	4	0.00595	0.005459	0.769231	0.000065	0.008074	0.00028
4	XGBoost	8	0.00605	0.005714	0.791383	0.000069	0.008322	-0.00031
5	XGBoost	8	0.006043	0.006072	0.764172	0.000073	0.008566	-0.00085
6	XGBoost	4	0.006908	0.006663	0.748869	0.000092	0.009598	0.00040
7	Yapay Sinir Ağları	13	0.008132	0.007858	0.53303	0.000128	0.011309	0.00064
8	Yapay Sinir Ağları	3	0.008229	0.007899	0.425339	0.00013	0.011406	-0.00113

Yukarıdaki tablodan da görüleceği üzere en iyi model 1 numaralı modeldir. Bu modelde geriye dönük 48 günlük veri, tahmin yapmak için kullanılmaktadır. Bütün hata birimi ölçütlerinde en iyi sonucu vermiştir. Daha sonraki bölümde bu model ikinci yatırım stratejisi için kullanılacaktır.

Test sonuçları birinci veri setiyle karşılaştırıldığında ikinci veri setinin yani tahmin edilmek istenen günün açılış fiyatının da veriye katıldığı model daha iyi sonuç vermiştir.

Çalışmanın sonuçlarının daha iyi analiz edilebilmesi adına oluşturulan ikinci veri setine göre de aynı makine öğrenmesi modelleri çalıştırılmıştır. Burada da çeşitli parametrelerde 1000'in üzerinde model oluşturulmuştur. Bu modellerden bazıları aşağıda sıralanmıştır. Bu modeller ertesi günün günlük en yüksek fiyatını tahmin etmek için oluşturulmuştur. Veri setinde tahmin edilecek günün verisi kullanılmamıştır.

Tablo 4.3 : Ertesi Gün En Yüksek Fiyat Tahmini İkinci Veri Seti Performans Sonuçları

No	Yöntem	k	MAE	SD Absolut e Error	SA	MSE	Sqrt of MSE	Mean Error
1	XGBoost	32	0.00634	0.00586	0.68182	0.00008	0.00863	0.00035
2	XGBoost	57	0.00615	0.00620	0.71858	0.00008	0.00873	0.00000
3	XGBoost	31	0.00675	0.00588	0.66310	0.00008	0.00895	0.00048
4	XGBoost	58	0.00655	0.00658	0.68033	0.00009	0.00929	0.00133
5	XGBoost	47	0.00681	0.00634	0.70732	0.00009	0.00931	-0.00046
6	XGBoost	5	0.00676	0.00663	0.67278	0.00009	0.00947	0.00029
7	Yapay Sinir Ağları	10	0.00783	0.00708	0.57368	0.00011	0.01056	-0.00014
8	Yapay Sinir Ağları	20	0.00787	0.00723	0.51194	0.00011	0.01068	0.00008

Bu kurulan modellerde de XGBoost modellerinin daha iyi sonuç verdiği görülmüştür. Bu modellerden gerek ortalama mutlak hata performans ölçütünün daha yüksek olması, gerekse genel olarak diğer performans ölçütlerindeki karşılaştırmalarda da iyi olması nedeniyle ikinci model seçilmiştir. Bu modelde geçmişe dönük 57 gün veri kullanılmıştır.

Diğerine göre zaman dilimi daha kısıtlı olan ikinci veri setinden bir çalışma daha yapılmıştır. Buradaki çalışmada da Borsa İstanbul 100 endeksinin en yüksek fiyatının tahmini yapılmaya çalışılmıştır. Aynı günün açılış fiyatı da geçmiş veri setine dahil edilerek modeller kurulmuştur. Borsanın açılış fiyatının gelmesiyle günün en yüksek fiyatının tahmin edilmesi üzerine kurgulanmıştır. Bu veri setimizde de 121 adet test verisi oluşmuştur. Performans ölçütlerine göre ayrılmış 8 model paylaşılmıştır.

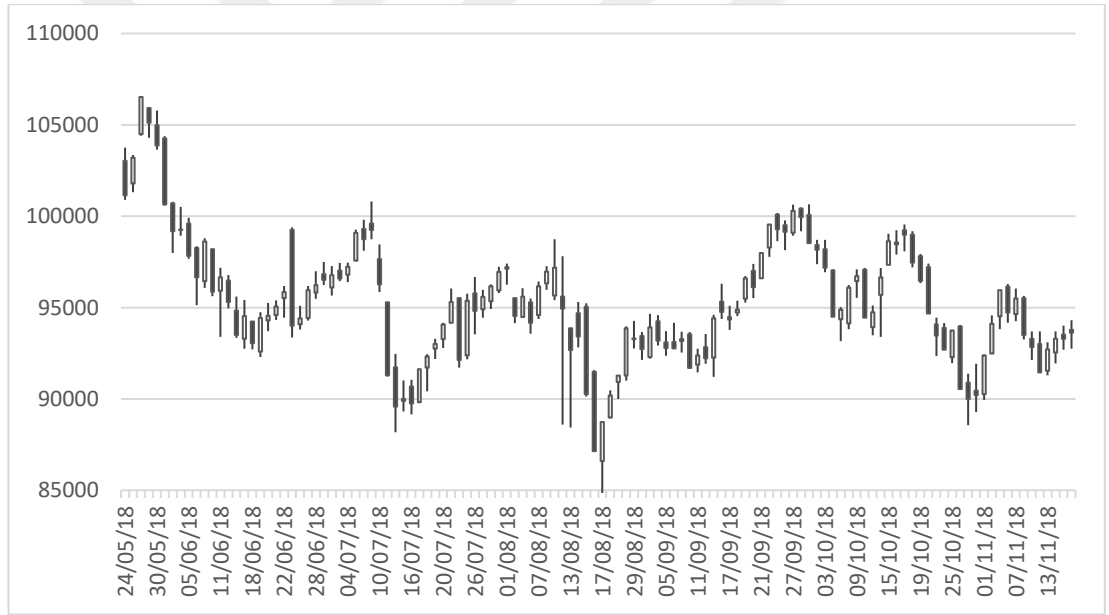
Tablo 4.4 : Aynı Gün En Yüksek Fiyat Tahmini İkinci Veri Seti Performans Sonuçları

No	Yöntem	k	MAE	SD Absolute Error	SA	MSE	Sqrt of MSE	Mean Error
1	XGBoost	58	0.005107	0.004433	0.762295	0.000046	0.006763	0.000466
2	XGBoost	20	0.005587	0.004827	0.774536	0.000055	0.007383	0.001091
3	XGBoost	41	0.005675	0.004725	0.757412	0.000055	0.007385	0.001348
4	XGBoost	5	0.005536	0.005465	0.769634	0.000061	0.007779	0.000103
5	XGBoost	34	0.005843	0.006059	0.769437	0.000071	0.008418	0.000036
6	XGBoost	50	0.006697	0.006386	0.752717	0.000086	0.009253	-0.000374
7	Yapay Sinir Ağları	15	0.007805	0.006799	0.498681	0.000107	0.010351	0.000489
8	Yapay Sinir Ağları	9	0.008218	0.008075	0.527559	0.000133	0.011522	-0.000031

Birçok performans ölçütünde birinci model diğerlerine göre daha başarılı olduğu görülmüştür. Bu modelde geçmişe dönük 58 gün veri bulunmaktadır. Bir sonraki bölümde yatırım stratejisi oluşturulurken bu model kullanılacaktır.

4.4 Yatırım Stratejisi Uygulamaları

Birinci veri setimize göre test tarihleri arasında yapılacak pasif bir yatırımın getirisi yaklaşık eksi %9 olacaktır. Yani test verisinin başında BİST100'e yapılacak bir yatırım verinin son tarihinde kapanış fiyatından satıldığı varsayılırsa, bu yatırım yüzde 9 zarar edecektir.



Şekil 4.4 : BİST100 Birinci Test Verisi Grafiği

Bu kısımda, elde ettiğimiz tahmin verilerinden 2 farklı yatırım stratejisi kapsamında 4 farklı yatırım stratejisi oluşturulmuş ve getiri grafikleri çizilmiştir. Borsa İstanbul'da anlık veri çeşitli kaynaklardan ücretli olarak paylaşılmaktadır. 15 dakikalık gecikmeli veri ücretsiz olarak verilmektedir. Ayrıca borsada aracı kurumlar işlem hacmi üzerinden komisyon oranı bazlı ücret tarifesi uygulanmaktadır. Bu yatırım stratejilerinin getirileri hesaplanırken Borsa İstanbul'daki verilerin ücretsiz ve anlık

ulaşıma açık olduğu varsayılmıştır. Ayrıca herhangi bir işlem masrafının da olmadığı varsayılmıştır.

PAY PİYASASI SÜREKLİ İŞLEM YÖNTEMİ (TAM İŞ GÜNLERİNDE)											
 BORSA İSTANBUL	AÇILIŞ (TEK FİYAT)		SÜREKLİ İŞLEM YÖN.	GÜN ORTASI (TEK FİYAT)		SÜREKLİ İŞLEM YÖN.	KAPANIŞ (TEK FİYAT)			KAPANIŞ FİYATINDAN / SON FİYATTAN İŞLEMLER	
	09:40– 09:55	09:55 (+)*	10:00– 13:00	13:00– 13:55	13:55 (+)*	14:00– 18:00	18:00– 18:01	18:01– 18:05	18:05 (+)*	18:07– 18:08	18:08– 18:10
	Emir Toplama	Fiyat Belirleme ve İşlem	İşlem	Emir Toplama	Fiyat Belirleme ve İşlem	İşlem	Kapanış Marj Yayını	Emir Toplama	Fiyat Belirleme ve İşlem	Marj Yayını	İşlem
Yıldız Pazar											
Ana Pazar											
GİP	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
KYÜYÜP (BYF, Varant ve Sertifikalar hariç)											
C Grubunda PY/LS olan paylar											
BYF, Varant ve Sertifikalar	---	---	✓	---	---	✓	---	---	---	---	---
Yeni Pay Alma Hakları											
Resmi Müzayede İşlemleri (*)	---	---	✓	---	---	✓	---	---	---	---	---
Birincil Piyasa İşlemleri (**)	---	---	10:30 ✓	---	---	---	---	---	---	---	---
Kısaltmalar BYF : Borsa Yatırım Fonu GİP : Gelişen İşletmeler Pazarı KYÜYÜP : Kolektif Yatırım Ürünleri ve Yapılandırılmış Ürünler Pazarı LS : Likidite Sağlayıcı PY : Piyasa Yapıcı						Açıklamalar ✓ : Pazarın / Grubun / Yatırım aracının, seansın ilgili alt sürecine dâhil olduğunu gösterir. --- : Pazarın / Grubun / Yatırım aracının, seansın ilgili alt sürecinin dışında olduğunu gösterir. * (+) : İşlemlerin tamamlanmasına bağlı olarak değişebilecek bir saati ifade eder. ** : ME, MF, MV, MC koduyla yapılan resmi müzayede işlemleri için geçerli olan seans saatlerini ifade eder. *** : Birincil Piyasa İşlemleri, saat 10:30 -13:00 arasındadır. Not : Emir toplama süreci, belirtilen saatlerden sonra 30 saniye içerisinde rastlantısal olarak kesilmektedir.					

Güncellenme Tarihi: 01/11/2016

Sayfa 1 / 4

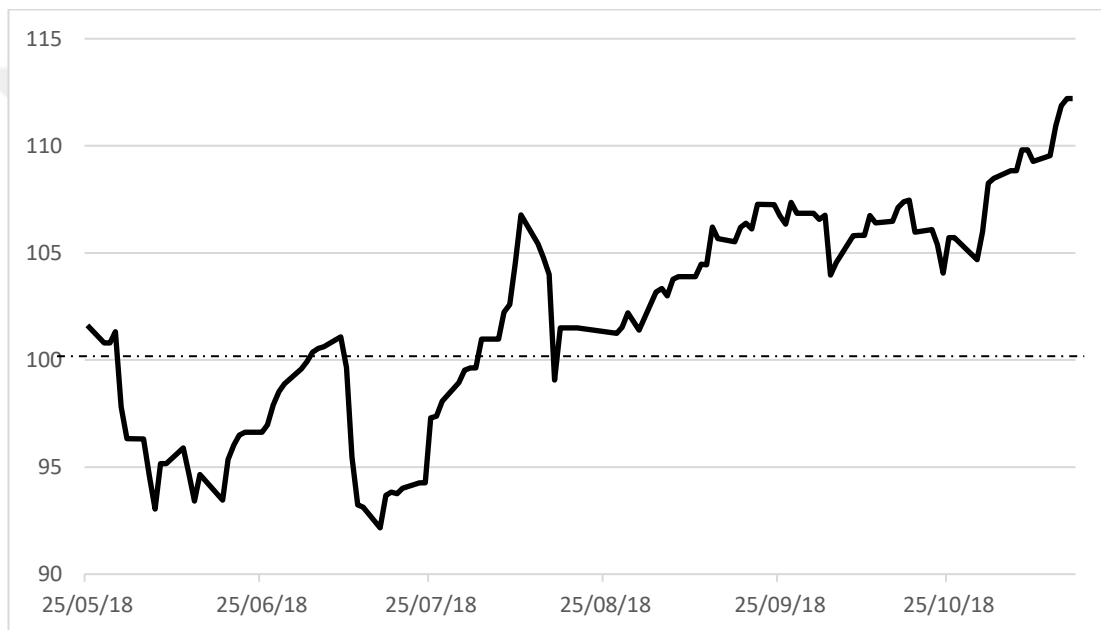
Şekil 4.5 : Borsa İstanbul İşlem Saatleri (Borsa İstanbul, 2019)

Borsa İstanbul her tam iş günlerinde yukarıdaki çizelgeye göre hareket etmektedir. Bu çizelgeye göre her sabah gelen emirlerle birlikte bir açılış fiyatı oluşmaktadır. Aynı zamanda her akşam kapanış fiyatında/son fiyattan işlem imkanı sunulmaktadır.

Yatırım stratejilerinden birincisi ertesi günün en yüksek fiyatını tahmin eden modelden üretilmiştir. Birinci veri setiyle oluşturulmuş modelin test verisi kullanılmıştır. Bu test verisi 120 günden oluşmaktadır. Bu 120 günlük tahminler kullanılarak borsanın sabah açılışıyla birlikte açılış fiyatından alım pozisyonuna girildiği varsayılmıştır. Hedef fiyat olarak da tahmin ettiğimiz fiyat kullanılmıştır. Eğer borsa gün içinde tahmin ettiğimiz fiyata ulaşırsa kazançlı işlem olarak getiri elde edildiği düşünülmüştür. Tam tersi olarak da bu fiyata ulaşmadığı takdirde tahmin ettiğimiz değer başarısız sayılmış ve günün sonundaki kapanış fiyatından pozisyonun

kapatıldığı varsayılmıştır. Her ne kadar kapanış fiyatından pozisyonlarımızı kapatsak da modellerimizde borsanın yönünü de tahmin ettiğimizden kapanan bu pozisyonlar da artı getiri oluşturabilmektedir.

Bu yatırım stratejisinde birinci veri seti ertesi gün en yüksek fiyat tahmini için oluşturulan model kullanılmıştır. Bu modelle oluşturulan yatırım stratejisinin günlük olarak getirisi hesaplanmıştır. Hesaplanan değerler birim değer 100'den başlamak kaydıyla grafiğe yansıtılmıştır:

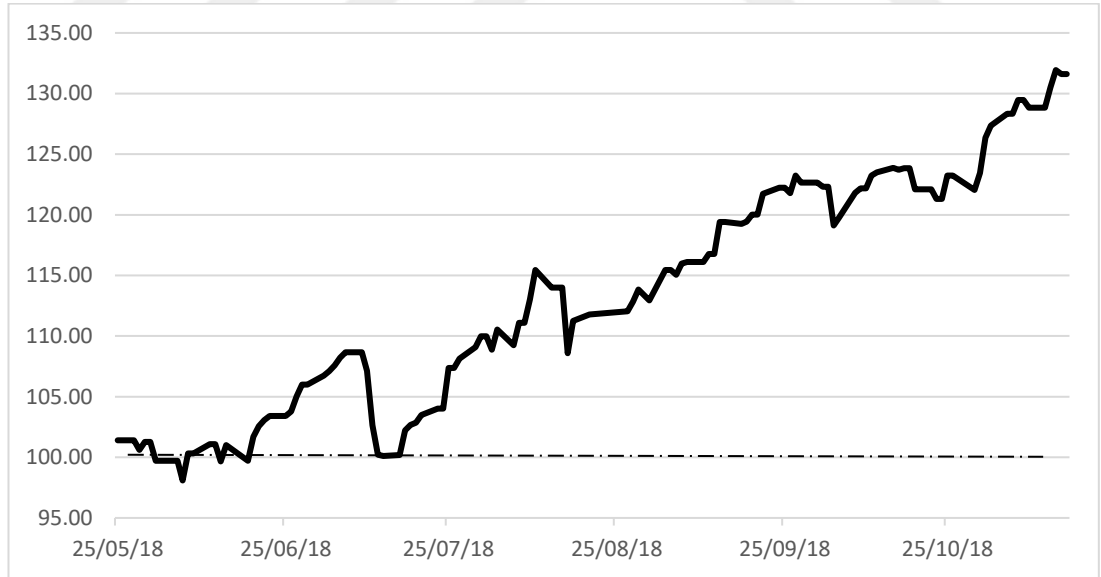


Şekil 4.6 : Birinci Test Verisi ile Oluşturulan Birinci Yatırım Stratejisi Getirileri
(Birim=100)

Verimizin ilk kısmında borsa düşüş trendinde olmasına rağmen modelimiz ortalama seviyeyi takip etmiştir. Daha sonrasında stratejimizin sonunda ise yaklaşık yüzde 12’lik bir getiri elde edilmiştir. Eğer Borsa İstanbul endeksine test verisinin en başında pasif yatırım yapılmış olsaydı yaklaşık yüzde 9’luk bir kayıp oluşacaktı. Bu yatırım stratejisinin pasif yatırıma göre daha yüksek getiri elde ettiği gözlemlenmiştir. Bu getirilerin standart sapma değeri ise 0,0117 olarak hesaplanmıştır. Pasif yatırımın günlük getirilerinden hesaplanan standart sapmasına göre (0,0161) de daha düşük çıkmaktadır.

Yatırım stratejilerimizden ikincisi ise aynı günün açılış fiyatının da öğrenme verisinde bulunduğu ve gün içi en yüksek fiyatı tahmin etmeyi hedeflediğimiz ikinci modelimiz kullanılmıştır. Bu stratejide de birinci veri seti kullanılmaktadır. Bu strateji günlük olarak borsanın sürekli işlem seansının açılmasıyla birlikte açılış fiyatından alış pozisyonu yapıldığı ve gelen açılış fiyatıyla birlikte gün içi en yüksek fiyatın tahmin edilmesiyle birlikte pozisyona devam etme veya kapatma kararı alınmıştır. Ey yüksek fiyatın düşük tahmin edilmesi durumunda açılış fiyatında pozisyonun tekrar aynı fiyattan kapatıldığı varsayılmıştır. Pozisyonumuzun devam etmesi yönü seçilmişse satış fiyatı olarak tahmin ettiğimiz en yüksek fiyat girilmiştir. Gün içinde en yüksek fiyata ulaşamaması durumunda günlük kapanış fiyatından işlemin kapatıldığı varsayılmıştır.

Bu varsayımlar göz önünde bulundurularak birim değer 100 ile başladığı getiri grafiği çizilmiştir.



Şekil 4.7 : Birinci Test Verisi ile Oluşturulan İkinci Yatırım Stratejisi Getirileri
(Birim=100)

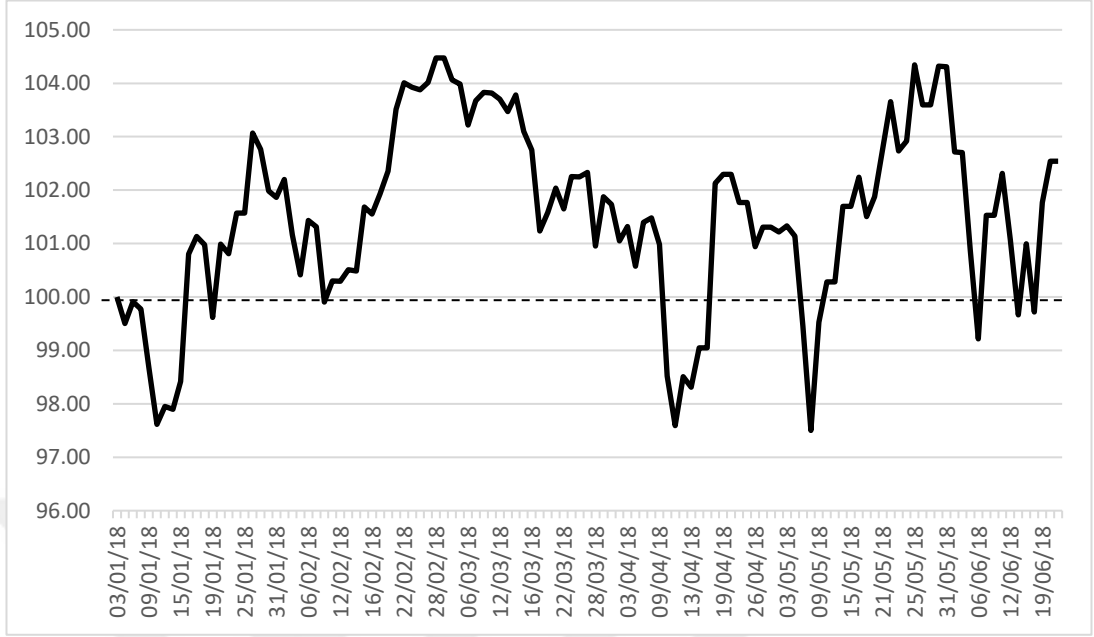
Bu stratejimizde yüzde 31,6 getiri elde edilmiştir. Bu modelimizde de pasif yatırıma göre daha yüksek getiri elde edilmiştir. Getirilerin standart sapma değeri de 0,0114 olarak hesaplanmıştır.

Tablo 4.5 : Getiri Karşılaştırma Tablosu (İkinci Veri Seti)

	Pasif Yatırım	Birinci Yatırım Stratejisi	İkinci Yatırım Stratejisi
Dönemsel Getiri	-9.1%	12.0%	31.6%
Standart Sapma	0.0161	0.0117	0.0114

Aynı günün açılış fiyatının dahil edildiği model diğerlerine göre daha iyi sonuçlar vermiştir. Bu strateji ile pasif yatırım ve önceki ertesi günün tahmin edildiği modele göre daha fazla getiri elde ettiği gözlemlenmiştir.

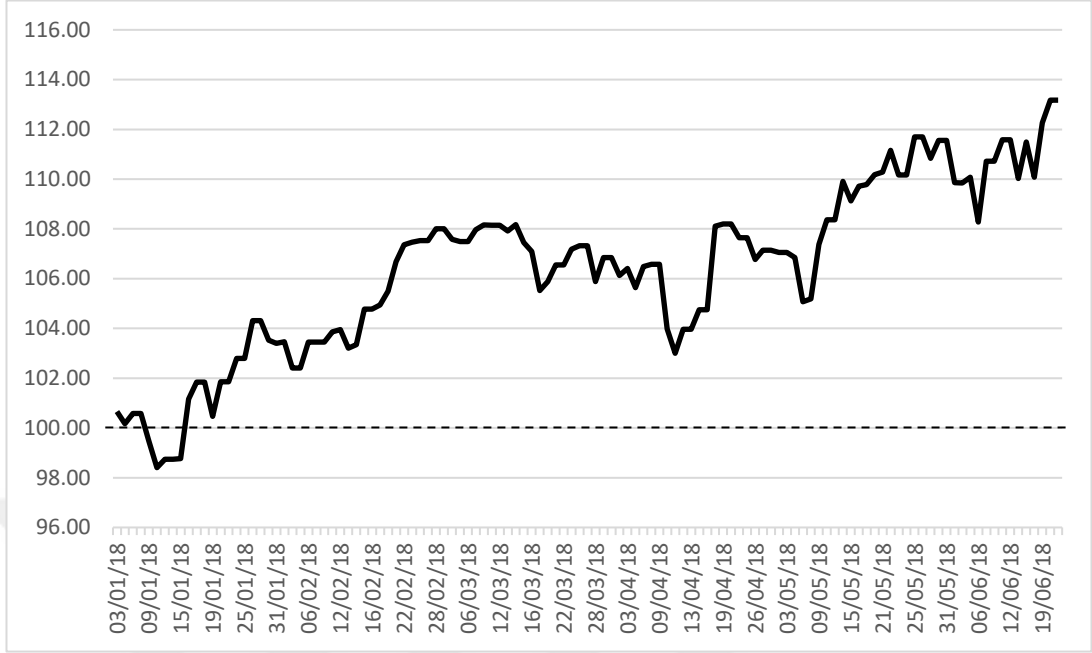
Yatırım stratejilerinin bu kısmında 1277 gün veriye sahip ikinci veri setiyle stratejiler oluşturulmuştur. Burada da önceki varsayımlar yapılmış ve günlük işlem yapıldığı düşünülmüştür. İkinci veri seti test verisine denk gelen kısım yüksek derecede düşüş trendine sahip olduğu için modellerimizin daha iyi test edilebileceği düşünülmüştür. Burada yatırım stratejilerinde kullanılacak modeller de bir önceki kısımda seçilmiştir. Burada kullanılan model ertesi günün en yüksek fiyatını ve yönünü tahmin etmektedir. Getiriler günlük olarak hesaplanmıştır.



Şekil 4.8 : İkinci Test Verisi ile Oluşturulan İkinci Yatırım Stratejisi Getirileri
(Birim=100)

Bu yatırım stratejisinde çok yüksek getiri elde edilememiştir. Ancak buradaki test verisinde borsanın güçlü bir düşüş trendinde olduğu görülmüştür. Bu zaman dilimden yapılacak pasif bir yatırım yaklaşık yüzde 20 zarar edecektir. Dolayısıyla düşük gözüken getirimiz aslında pasif yatırıma göre çok iyi bir performans göstermiştir. Bu yatırım stratejisinin standart sapması da 0,0091 ile pasif yatırıma göre daha iyi çıkmıştır.

Son olarak, ikinci veri seti ile aynı günün en yüksek fiyatını tahmin ettiğimiz en iyi modelle oluşturulmuştur. Bu yatırım stratejisinde borsada açılış fiyatının gelmesiyle tahmin modelinin çalıştırıldığı ve borsa endeksinde pozisyon alındığı düşünülmüştür. İşlemler günlük bazda gerçekleştirildiği için getiriler de günlük bazda hesaplanmıştır. Birim değer 100 ile başlatılarak görselleştirilmiştir.



Şekil 4.9 : İkinci Test Verisi ile Oluşturulan İkinci Yatırım Stratejisi Getirileri
(Birim=100)

Bu yatırım stratejisinden dönem sonunda yüzde 13,2'lik getiri elde edilmiştir. Bu getiri pasif getiriyle karşılaştırıldığında çok iyi sonuç vermiştir. Bir önceki stratejimizden de daha iyi sonuç vermiştir.

Tablo 4.6 : Getiri Karşılaştırma Tablosu (İkinci Veri Seti)

	Pasif Yatırım	Birinci Yatırım Stratejisi	İkinci Yatırım Stratejisi
Dönemsel Getiri	-17.9%	2.5%	13.2%
Standart Sapma	0.0126	0.0091	0.0085

Bütün yatırım stratejilerinin sonuçlarında bakıldığında ertesi gün tahminiyle üretilen modellerin pasif yatırıma göre daha iyi sonuç verdiği görülmüştür. Aynı günün en yüksek fiyatını tahmin ettiğimiz modeller her ikisinde de daha iyi sonuç vermiştir. Aynı yatırım stratejileri iki farklı veri seti ile denenmiştir. Her iki veri setinde de seçilen modellerin iyi sonuç aldığı görülmüştür.

SONUÇ

Gelişen bilgisayar ve veri teknolojileriyle birlikte makine öğrenmesi teknikleri dolaylı veya doğrudan birçok finansal istatistiki çalışmada yer almaktadır. Literatür çalışmasında incelediğimiz şekilde yapılan çalışmalarda en çok yapay sinir ağları ve türevi modellerin kullanıldığı görülmektedir. Finansal kararların verilmesinde ve yatırım stratejilerinin oluşturulmasında, bilgisayar teknolojisinin yardımıyla algoritmalar oluşturulmakta ve verilen finansal kararlar otomatik hale gelmektedir. Günümüzde birçok yatırım kuruluşu, algoritmik işlemler için bilgi işlem teknolojilerine yatırım yapmaktadır. Yapılan algoritmik alım-satım işlemleri, veri toplama ve işleme yeteneğinin geliştirilmesiyle çok hızlı hale gelmiştir. Yapılan işlemler milisaniye seviyelerine ulaşmıştır. Literatürde bu işlemler yüksek hızlı işlemler (High Frequency Trading) olarak adlandırılmaktadır. Yüksek hızlı işlem teknolojileriyle, yatırımcılar ve şirketler piyasadaki değişimlere anlık olarak tepki vererek, çeşitli stratejilerle alım-satım işlemleri gerçekleştirmektedir.

Bu çalışma Borsa İstanbul 100 endeksi üzerine yapılmıştır. Endeksin günlük en yüksek fiyatı, makine öğrenmesi modelleri ile tahmin edilmeye çalışılmıştır. Oluşan bu veriler ile de günlük yatırım stratejileri oluşturularak pasif yatırıma göre daha yüksek getiri elde edilmesi hedeflenmiştir. Çalışmada makine öğrenmesi yöntemlerinden yapay sinir ağları ve karar ağacı temelli gradient boosting modeli olan XGBoost modeli kullanılmıştır. Kullanılan modeller çeşitli performans ölçütlerine göre karşılaştırılmıştır. İki ayrı zaman diliminde iki farklı kurduğumuz modellerde XGBoost modellerinin yapay sinir ağları modeline göre daha iyi sonuç verdiği yönünde bulgular elde edilmiştir. Seçilen bu modellerle de iki farklı yatırım stratejisi oluşturulmuştur. Bu stratejiler oluşturulurken Borsa İstanbul fiyat oluşum seans yapısı göz önünde bulundurulmuştur.

Oluşturulan birçok modele göre borsanın yönünün tahmin edilmesinde, XGBoost modeli ile yapay sinir ağları modelinden daha yüksek performans elde edilebildiği gözlemlenmiştir. Borsa İstanbul 100 endeksinin en yüksek fiyatının günlük yönünün tahmin edilmesinde yüzde 72'lik oranlara ulaşılmıştır. XGBoost

modellerinin karar ağacı tabanlı model olması bunun nedenlerinden bir tanesi olduğu düşünülmektedir.

Tahmin modellerinde aynı günün açılış verisinin de geçmiş veri olarak kullanıldığı modeller, sadece önceki günlerin verisinin kullanıldığı modellere göre daha üstün performans sağlamıştır. Bu da Borsa İstanbul 100 endeksinde, en yüksek fiyatın tahmin edilmesinde günlük açılış fiyatının önemli bir veri olduğunu göstermektedir.

Kurgulanan stratejilerde işlem ve bilgi maliyetlerinin olmadığı varsayılmıştır. Borsada algoritmalarla yapılan alım-satım işlemlerinde çeşitli stratejiler uygulanabilmektedir. Bu çalışmada günlük olarak alım yönlü stratejiler oluşturulmuştur. Oluşturulan stratejilerde yapılan tahmin modelleri doğrultusunda pasif yatırıma göre daha yüksek getiri elde edilebileceği görülmüştür.

Yapılan çalışmada, 5 yıllık veri seti kullanarak oluşturulan modelde 6 aylık dönem için test yapılmıştır. Bu test döneminde al-tut stratejisi ile Borsa İstanbul 100 endeksine yapılacak yatırım yüzde 9 zarar edecektir. Makine öğrenmesi modeliyle oluşturulan tahminler vasıtasıyla günlük al-sat stratejileri oluşturulmuştur. İlk yatırım stratejisi ile aynı dönem için yüzde 12 dönemsel getiri elde edilmiştir. İkinci yatırım stratejisi ile de yüzde 31,6 getiri ile hepsinden yüksek dönemsel getiri elde edilebileceği görülmüştür.

Test edilen yatırım stratejilerinin hepsinde pasif yatırıma yani al-tut stratejisine göre daha yüksek performans gözlemlenmiştir. Bu yatırım stratejileriyle Borsa İstanbul trendinin aşağı olduğu dönemlerde bile getiri elde edilebileceği görülmüştür. Ayrıca aynı günün açılış fiyatının tahminci veri setine eklenmesi ile oluşturulan ikinci yatırım stratejisinde, modelin performansının arttığı gözlemlenmiştir. Gözlemlenen getirilerin tesadüfi olmadığı ikinci veri setiyle de kontrol edilmiştir.

KAYNAKÇA

- Akman, Muhammet, Genç, Yasemin, Ankaralı, Handan: "Random Forests Yöntemi ve Sağlık Alanında Bir Uygulama", **Türkiye Klinikleri J Biostat**, 3(1), 2011, s.36 - 48.
- Aşkın, Dildade, İskender, İres, Mamızadeh, Ali: "Farklı Yapay Sinir Ağları Yöntemlerini Kullanarak Kuru Tip Transformatör Sargısının Termal Analizi", **Gazi Üniv. Müh. Mim. Fak. Der.**, Cilt 26, No 4, 2011, s.905-913.
- Avcı, Emin: "Stock Return Forecasts With Artificial Neural Network Models", **Marmara Üniversitesi İ.İ.B.F. Dergisi**, 2009, C.XXVI, 1, s. 443 - 461.
- Borsa İstanbul Endeksler: "BİST Pay Endeksleri", (Çevrimiçi) <https://www.borsaistanbul.com/endeksler/BIST-pay-endeksleri>, 2019.
- Borsa İstanbul İşlem Saatleri: "Borsa İstanbul Pay Piyasası Yeni Seans Saatleri", (Çevrimiçi) <https://www.borsaistanbul.com/urunler-ve-piyasalar/piyasalar/pay-piyasasi/islem-saatleri>
- Chen, Tianqi, Guestrin, Carlos: "XGBoost: A Scalable Tree Boosting System", In Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, s.785–794
- Dey, Shubharthi , Kumar, Yash, Saha, Snehanthu, Basak, Suryoday: "Forecasting to Classification: Predicting the direction of stock market price using Xtreme Gradient Boosting", <https://www.researchgate.net>, 2016.
- Erkartal, Reşat B: "Forecasting İstanbul Stock Exchange", Yeditepe Üniversitesi, 2018.
- Friedman, Jerome H: "Greedy function approximation: a gradient boosting machine", **The Annals of Statistics**, V.29, No.5, 2001, 1189 - 1232.
- Guresen, Erkam, Kayakutlu, Gulgun, Daim, Tugrul U.: "Using artificial neural network models in stock market index prediction", **Expert Systems with Applications**, 38, 2011, s.10389 - 10397.

- Gündüz, Hakan,
Çataltepe, Zehra,
Yaslan, Yusuf: "Stock daily return prediction using expanded features and feature selection", **Turkish Journal of Electrical Engineering & Computer Sciences**, 25, 2017, s.4829 - 4840.
- Haykin, Simon: "Neural Networks: A Comprehensive Foundation", **Pearson Education**, 1999.
- Jiao, Yang,
Jakubowicz, Jérémie: "Predicting Stock Movement Direction with Machine Learning: an Extensive Study on S&P500 Stocks", **IEEE International Conference on Big Data**, 2017.
- Kara, Yakup,
Boyacioglu, Melek A.,
Baykan, Ömer K.: "Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange", **Expert Systems with Applications**, 38, 2011, s.5311 - 5319.
- Kılıç, Süleyman B.,
Paksoy, Semin, Genç,
Tolga: "Forecasting the Direction of BİST 100 Returns with Artificial Neural Network Models ", **International Journal of Latest Trends in Fin. and Eco. Sc.**, 4, 3, 2014, s.759 - 765.
- Krollner, Bjoern,
Vanstone, Bruce,
Finnie, Gavin: "Financial time series forecasting with machine learning techniques: a survey.", **European Symposium on Artificial Neural Networks - Computational Intelligence and Machine Learning**, https://works.bepress.com/bjoern_krollner/1/, 2010.
- Laboissiere, Leonel
A., Fernandes,
Ricardo A.S., Lage,
Guilherme G.: "Maximum and minimum stock price forecasting of Brazilian power distribution companies based on artificial neural networks", **Applied Soft Computing**, 35, 2015, s. 66 - 74.
- Morde, Vishal: "XGBoost Algorithm", (Çevrimiçi)
<https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d>, 2019.
- Patel, Jigar, Shah,
Sahil, Thakkar,
Priyank, Kotecha, K: "Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques", **Expert Systems with Applications**, 42, 1, 2015, s.259 - 268.

- Qiu M, Song Y "Predicting the Direction of Stock Market Index Movement Using an Optimized Artificial Neural Network Model", (Çevrimiçi) <https://doi.org/10.1371/journal.pone.0155133>, 2019.
- Shen, Shunrong, Jiang, Haomiao, Zhang, Tongda: "Stock Market Forecasting Using Machine Learning Algorithms ", pdfs.semanticscholar.org, 2012.
- Shukla, Suvarna, Yadav, Krishna: "Forecasting Indian Stock Market Index Using Singular Spectrum Analysis: A Comparison with ARIMA and Artificial Neural Network", (Çevrimiçi) <https://ssrn.com/abstract=3036432>, 2017.
- Şahin, Nazmi B.: "Predicting Stock Price Index of Borsa İstanbul (BİST) Using Different Machine Learning Techniques", İstanbul Üniversitesi Tez, 2016.
- Tsai, C.-F, Wang, S.-P: "Stock Price Forecasting by Hybrid Machine Learning Techniques", **Proceedings of the International MultiConference of Engineers and Computer Scientists, 1**, 2009.
- Uğur, Ahmet: "Hisse senedi getirilerinin panel veri analizi ile tahmini: İstanbul Menkul Kıymetler Borsasında bir inceleme", İnönü Üniversitesi, 2009
- XGBoost Kütüphanesi: "XGBoost Library", (Çevrimiçi) <https://xgboost.readthedocs.io>, 2019.
- Yakut, Emre, Elmas, Bekir, Yavuz, Selahattin: "Yapay Sinir Ağları ve Destek Vektör Makineleri Yöntemleriyle Borsa Endeksi Tahmini", **Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi**, C.19-1, 2014, s.139 - 157.

EKLER

Yapılan çalışmada Python programlama dili kullanılmıştır. Modellerin kodları sadeleştirilerek bu kısımda başlıkları ile paylaşılmıştır. İlk olarak Ek 1’de performans ölçütlerinin tanımları ve hesaplama şekilleri bulunmaktadır. Ek 2’de XGBoost kütüphanesinin kullanımı ve parametre giriş değerleri belirtilmiştir. Ek 3’de ise yapay sinir ağları modelinin kullanım şekli ve parametre değerleri gösterilmiştir. Yapay sinir ağları modeli için Tensorflow kütüphanesinden faydalanılmıştır.

EK 1 : Performans Ölçütleri Hesapları

```
import numpy as np
import pandas as pd

def stats_str(self, stats):
    return 'Mean absolute error: %.6f\n' % stats['mae'] + \ # Ortalama mutlak hata, tahmin edilen ve gerçekleşmiş değerler arasındaki farktır.
    'SD absolute error: %.6f\n' % stats['stdae'] + \ # Hataların standart sapma değerini gösterir.
    'Sign accuracy: %.6f\n' % stats['sign_accuracy'] + \ # Verilen sinyalin doğruluk oranını gösterir.
    'Mean squared error: %.6f\n' % stats['mse'] + \ # Ortalama kare hata, hataların karelerinin toplamının ortalamasını gösterir.
    'Sqrt of MSE: %.6f\n' % stats['sqrt_mse'] + \ # Ortalama kare hataların karekökü (Sqrt of MSE), hataların karelerinin toplamının ortalamasının alınması sonrasındaki karekök değeridir.
    'Mean error: %.6f\n' % stats['me'] + \ # Ortalama hata, tahmin değerleri ile gerçek değerlerin arasındaki ortalama hatadır

def _compute_stats(self, prediction, truth):
    residuals = np.abs(prediction - truth)
    return {
        'mae': np.mean(residuals),
        'stdae': np.std(residuals),
        'sign_accuracy': np.mean(np.equal(np.sign(prediction), np.sign(truth))),
        'mse': np.mean(np.power(prediction - truth, 2.0)),
        'sqrt_mse': np.mean(np.power(prediction - truth, 2.0)) ** 0.5,
        'me': np.mean(prediction - truth),
    }
```

EK 2 : XGBoost Modeli ve Kullanılan Parametreler

```
def iterate_xgb(job_info, job_runner, iterations=100, k_lim=60):
    if XgbModel is None:
        return

    job_runner.iterate(iterations, params_fun=lambda: {
        'target': job_info.target,
        'k': np.random.randint(1, k_lim),
        'model_class': XgbModel,
        'model_params': {
            'max_depth': np.random.randint(3, 8),
            'n_estimators': np.random.randint(100, 300),
            'learning_rate': 10 ** np.random.uniform(-2, -0.5),
            'gamma': np.random.uniform(0, 0.1),
            'subsample': np.random.uniform(0.5, 1),
        }
    })

import os
from xgboost import XGBRegressor
from sklearn.externals import joblib

from .model import Model
from util import vlog

class XgbModel(Model):
    def __init__(self, **params):
        Model.__init__(self, **params)
        del params['features']
        self._model = XGBRegressor(**params)

    def fit(self, train):
        self._model.fit(train.x, train.y)

    def predict(self, x):
        return self._model.predict(x)
```

EK 3 : Tensorflow Yapay Sinir Ağları Modeli ve Kullanılan Parametreler

```
def iterate_neural(job_info, job_runner, iterations=10, k_lim=21)
:
    if NeuralNetworkModel is None:
        return

    job_runner.iterate(iterations, params_fun=lambda: {
        'target': job_info.target,
        'k': np.random.randint(1, k_lim),
        'model_class': NeuralNetworkModel,
        'model_params': {
            'batch_size': np.random.choice([500, 1000, 2000, 40
00]),
            'epochs': 100,
            'learning_rate': 10 ** np.random.uniform(-10, -2),
            'init_sigma': 10 ** np.random.uniform(-10, -3),
            'layers': _random_layers(np.random.randint(1, 4)),
            'cost_func': np.random.choice(['l1', 'l2']),
            'lambda': 10 ** np.random.uniform(-8, -2),
        }
    })

    from __future__ import absolute_import

    import tensorflow as tf

    from .tensorflow_model import TensorflowModel
    from .nn_ops import ACTIVATIONS, COST_FUNCTIONS, dropout, batch
_normalization

    class NeuralNetworkModel(TensorflowModel):
        def __init__(self, **params):
            TensorflowModel.__init__(self, **params)

            self._layers = params.get('layers', [])
            self._init_sigma = params.get('init_sigma', 0.001)
            self._lambda = params.get('lambda', 0.005)
            self._cost_func = COST_FUNCTIONS[params.get('cost_func'
, 'l2')]

            self._compile()

        def _compile(self):
            TensorflowModel._compile(self)

            with tf.Graph().as_default() as self._graph:
                x = tf.placeholder(tf.float32, shape=[None, self._f
eatures], name='x')
                y = tf.placeholder(tf.float32, shape=[None], name='
y')

                mode = tf.placeholder(tf.string, name='mode')
```

```

        rand_init = lambda shape: tf.random_normal(shape=shape, stddev=self._init_sigma)

        layer = x
        dimension = self._features
        reg = 0
        for idx, layer_params in enumerate(self._layers):
            with tf.variable_scope('l_%d' % idx):
                size = layer_params.get('size', 50)
                W = tf.Variable(rand_init([dimension, size]), name='W%d' % idx)
                b = tf.Variable(rand_init([size]), name='b%d' % idx)
                layer = tf.matmul(layer, W) + b

                batchnorm = layer_params.get('batchnorm', False)
                if batchnorm:
                    batch_normalization(layer, tf.equal(mode, 'train'))

                activation_func = ACTIVATIONS[layer_params.get('activation_func', 'relu')]
                layer = activation_func(layer)

                dropout_prob = layer_params.get('dropout', 0.5)
                layer = dropout(layer, tf.equal(mode, 'train'), keep_prob=dropout_prob)

                reg += self._lambda * tf.nn.l2_loss(W)
                dimension = size

            with tf.variable_scope('l_out'):
                W_out = tf.Variable(rand_init([dimension, 1]), name='W')
                b_out = tf.Variable(rand_init([1]), name='b')
                output_layer = tf.matmul(layer, W_out) + b_out
                reg += self._lambda * tf.nn.l2_loss(W_out)

        cost = self._cost_func(output_layer, y) + reg
        optimizer = tf.train.AdamOptimizer(self._learning_rate).minimize(cost)

        init = tf.global_variables_initializer()

        self._x = x
        self._y = y
        self._mode = mode
        self._output = output_layer
        self._cost = cost
        self._optimizer = optimizer
        self._init = init

```