Ankara Hacı Bayram Veli Üniversitesi
İktisadi ve İdari Bilimler Fakültesi Dergisi 24/3 (2022) 1205-1230
E-ISSN 2667-405X

Makine Öğrenmesi ile Finansal Zaman Serisi Tahminleme

Seyyide DOĞAN[*] Yasin BÜYÜKKÖR[**]

Geliş Tarihi (Received): 18.09.2022– Kabul Tarihi (Accepted): 29.11.2022

Öz

Finans uygulamalarının önemli bir çalışma alanını oluşturan finansal zaman serisi tahminlemesi son yıllarda makine öğrenmesi (Machine Learning, ML) yöntemlerinin gelişimi ile finans ve akademi çevrelerinin daha fazla önem atfettiği bir konu olmuştur. Bu çalışmanın amacı, finansal zaman serisi gelecek değerinin tahmininde ML yöntemlerinin karşılaştırmalı olarak bir incelemesini sunmaktır. Çalışmada gelişmiş ve gelişmekte olan iki borsa endeksi ve İstanbul borsasının yüksek hacimli iki hisse senedinin son 5 yıllık kapanış verileri kullanılmıştır. Endeks tahmininde sıklıkla kullanılmış ve başarılı bulunan Destek Vektör Regresyonu (Suport Vector Regression, SVR) ve literatürde zaman serisi tahmininde izine az rastladığımız topluluk (ensemble) makine öğrenmesi yöntemleri olan Rassal Orman (Random Forest, RF) ve Extrem Gradyan Arttırma (eXtreme Gradient Boosting, XGBoost) yöntemleri tercih edilmiştir. Çalışmanın bulgularına göre, MAE, MAPE ve RMSE kriterleri göz önünde bulundurulduğunda en iyi tahmin yöntemi SVR olarak tespit edilmiştir.

Anahtar Kelimeler: finansal zaman serisi, makine öğrenmesi, destek vektör makinesi, rassal orman, xgboost

Financial Time Series Prediction Using Machine Learning

Abstract

Making up an important working area of finance applications, financial time series forecasting, with the advancements in Machine Learning (ML) methods in recent years, has become a topic that finance and academic circles attach more importance to. The aim of this study is to present a comparative review of ML methods in financial time series future value. In the study, the last 5-year closing data of two developed and emerging stock market indices and two high-volume stocks of the Istanbul stock market were used. Support Vector Regression (SVR), which is often used in index forecasting and found successful, and Random Forest (RF) and eXtreme Gradient Boosting (XGBoost) methods which are rarely used ensemble machine learning methods in time series forecasting in literature, are preferred. As a result of the study, when MAE, MAPE and RMSE criterions are taken into consideration, SVR was confirmed to be the best forecasting method.

Key Words: financial time series, machine learning, support vector machine, random forest, xgboost

[*] Dr. Karamanoğlu Mehmetbey Üniversitesi, Uluslararası Ticaret ve İşletmecilik Bölümü, Karaman, Turkiye, dogans@kmu.edu.tr, ORCID:0000-0001-7835-7905.

[**] Dr. Karamanoğlu Mehmetbey Üniversitesi, İşletme Bölümü, Karaman, Turkiye, yasinbuyukkor@kmu.edu.tr, ORCID:0000-0002-1006-0539.

Giriş

Finansal faaliyetlerin kurallarına hakim olabilmek ve gelecek eğilimlerini tahmin edebilmek her zaman akademi ve finans araştırmacılarının dikkatini çeken önemli bir konu olmuştur (Poon, 2003; Nonejad, 2017; Cao vd., 2019). Finans sektöründe yaşanan hızlı gelişmeler ve küresel ekonomik krizlerin finansal piyasalar üzerinde yarattığı olumsuz etkileri sebebiyle finansal tahminlemelerin yapıldığı çalışmalar son yıllarda daha da önemli hale gelmiştir.

Finansal tahmin uygulamalarında en sık çalışılan konuların başında, bir hisse senedinin fiyatının tahmin edilmesi gelmektedir (Sezer vd., 2020). Öte yandan, borsanın değişken doğası sebebi ile fiyat dalgalanmalarının güvenilir ve doğru bir şekilde tahmin edilmesi oldukça zor bir süreçtir (Nava vd, 2016; Cao ve Wang, 2018; Nava vd., 2018; Cao vd., 2019). Bu zorluk, finansal zaman serilerinin durağan olmaması ve istatistiksel özelliklerinin zaman, siyasi olaylar, genel ekonomik koşullar, yatırımcı beklenti ve etkileşimlerinden dolayı doğrusal olmamasından kaynaklanmaktadır (Huan vd., 2005; Di Matteo 2007; Nava vd. 2016)

Finansal endekslerin açılış, kapanış, hacim vb. gibi zaman serisi verilerine dayalı tahmini finansal düzenleyiciler, politika yapıcılar, risk ve portföy yöneticileri, finansal kuruluşlar ve yatırımcılar için karlı bir iş olduğundan, üzerinden pek çok farklı tahmin modeli geliştirilmiştir (Deviren vd.,2014).

Finansal zaman serisi tahmininde önceleri sıklıkla kullanılan, doğrusal regresyon, otoregresif hareketli ortalama (ARIMA), ARCH (otoregresif koşullu varyans), GARCH (genelleştirilmiş otoregresif koşullu varyans), hata düzeltme modelleri (ECM) gibi klasik istatistiksel yöntemler bugün hala kullanılan iyi birer yardımcı araçtır. Ancak, bu tür modeller verilerin istatistiksel dağılımı ve durağanlığı hakkında birtakım varsayımlar gerektirmesinden dolayı doğrusal ve durağan olmayan finansal zaman serilerinin uzun vadeli tahminlerinde zayıf yeteneklere sahip oldukları kabul edilmektedir (Kazem vd.,2013; Yu, vd., 2008; He vd., 2012; Niu vd., 2020). Son yıllarda, ML yöntemlerinin büyük ölçekli finansal zaman serisinin doğrusal olmayan modelleme kabiliyetleri bu açmazın üstesinden gelmekte ve klasik yöntemlere göre daha iyi performans göstermektedir (Niu vd., 2020; Karasu vd., 2020). ML yöntemlerinin içinde Yapay Sinir Ağları (Neural Netwok, NN)'nın önemli bir yeri vardır ve en sık çalışılan yöntemlerin başında gelmektedir (Crone ve Nokolopous,2007; Aras ve Deveci Kocakoç, 2016). Derin öğrenme alt başlığı altında NN'nin diğer türevleri olan Çok Katmanlı Algılayıcı (Multi Layer Percepton, MLP), Tekrarlı Sinir Ağları (Recurrent Neural Network, RNN), Uzun-Kısa Süreli Bellek (Long Short Term Memory, LSTM) yine yaygın bir çalışma alanı bulmuştur.

Ancak, NN uygulamalarında gizli katman sayısının, bağlantı ağırlık değerlerinin ve öğrenme oranının ne olacağı gibi birtakım zorluklar hala geçerliliğini korumaktadır (Kuremoto vd., 2014). ML yöntemleri içinde NN kadar sık kullanılan bir diğer yöntem ise Destek Vektör Makinesi (Support Vector Machine, SVM)'dir (Patel 2015(b)). SVM, deneysel risk minimizasyonu prensibine dayanan NN'nin aksine, yapısal risk minimizasyonuna dayalı tahmin modelleri geliştirdiğinden yüksek bir genelleme kabiliyetine sahiptir (Cao ve Tay, 2003; Yu vd., 2009). Örüntü tanıma problemlerindeki üstünlüklerine ilave olarak, Rassal Orman (RF), AdaBoost ve XGBoost gibi topluluk (ensemble) öğrenme yöntemleri ile finansal zaman serisi tahmin çalışmaları çoğunlukla hisse senedinin gelecek eğilime odaklanmıştır (Kumar ve Thenmozhi, 2006; Patel, vd., 2015(a); Balling vd., 2015). Bu çalışmalarda, SVR' ye karşılık topluluk makine öğrenmesi yöntemlerinden hangisinin daha iyi performans gösterdiğine dair net bir kanıt bulmak mümkün olmamıştır. Bununla birlikte, elde edilen bilgiler dahilinde, hisse senedi ve endekslerin gerçek fiyatının tahmininde topluluk makine öğrenmesi yöntemlerinin tercih edildiği ve SVR ile sonuçlarının kıyaslandığı bir çalışma Patel ve diğerlerine (2015(b)) aittir. Literatürdeki bu boşluk çalışmamızın temel motivasyonudur.

Bu çalışmanın ana katkısı, hisse senedinin gerçek değerinin tahmininde, topluluk yöntemlerinin (RF, XGBoost) ve tek sınıflandırıcı yöntemlerin (SVR) performanslarının kıyaslamalı olarak ortaya konulmasıdır. Çalışmanın bir diğer katkısı da pek çok çalışmada göz ardı edilmiş olan, ancak yöntemlerin tahmin performansı üzerinde önemli bir etkiye sahip olduğu bilinen hiper-parametrelerin optimize edilmesidir.

Bu çalışmada yer alan orijinal veriler gelişmiş ve gelişmekte olan New York ve İstanbul Borsasına ait iki adet birleşik endeks ve İstanbul borsasında işlem gören en büyük hacimli şirketlerden Ereğli ve Aselsan şirketlerinin hisse senetlerinin de dahil olduğu 4 finansal zaman serisi üzerinde gerçekleştirilmiştir.

Bu çalışma sonucunda ortaya konulan bulguların, araştırmacıların hangi yöntemin hisse senedi fiyat tahmininde daha iyi performans gösterdiğine dair daha açık bir fikre sahip olmasını sağlayacağı düşünülmektedir. Çalışmanın takip eden kısımlarında finansal zaman serisi tahmini için ML uygulamalarına odaklanan diğer çalışmalar geniş kapsamlı olarak sunulmuştur. Bölüm 3'de kullanılan yöntemin teorik çerçevesi ve Bölüm 4'te yöntemin pratik bir uygulaması tüm detayları ile açıklanmış ve model bulgularına ilişkin sonuçlar tartışılmıştır. Çalışmanın bulgularına ilişkin genel sonuçlarının yer aldığı son bölümü ise zaman serisi tahminlemesinde ML yöntemlerinin araştırmaya açık yanlarını ve öne çıkan taraflarını içermektedir.

1. Literatür Taraması

ML yöntemleri literatürde geniş çaplı uygulama alanı bulmuştur. Bu alanlarda yapılan kapsamlı liretatür incelemesi çalışmaları: finansal zaman serileri analizinde (Cavalcante vd., 2016; Sezer vd., 2020), eğitim alanında (Tosunoğlu vd. 2021), sağlık alanında (Shailaja vd., 2018), tarım uygulamaları alanında (Liakos vd., 2018), malzeme ve metalürji mühendisliği alanında (Wei vd., 2019), uygulamalı fizik ve astronomi çalışmalarında (Brunton vd., 2020), internet ve ağ uygulamalarında (Zander vd., 2005), tedarik zinciri çalışmalarında (Carbonneau vd., 2008) ve üretim süreci ve planlanmasında (Wuest vd., 2016) olmak üzere çeşitli alanlarda kullanılmaktadır.

Finansal zaman serisi tahminlemelerinde özellikle 1995 yılından önce yapılan çalışmaların çoğunda geleneksel istatistiksel yöntemler ile tahminleme yapılmaktadır. 2000'li yılların başında bilgisayarların işlem kapasitelerindeki gelişmeler ve yapay zeka yöntemlerindeki hızlı değişimlerle beraber araştırmacılar, geleneksel yöntemler yerine hem uygulaması daha kolay hem de daha anlaşılır olan ML yöntemlerine yönelmişlerdir. İlk ML çalışmaları iyi bilinen ve sıkça kullanılan NN ailesi ve SVM etrafında toplanmakta ve çalışmalar genellikle bu iki yöntemin kıyaslanması özelinde yürütülmüştür. Tay ve Cao (2001) finansal zaman serisi öngörüsünde SVM ile Çok Katmanlı Geri Beslemeli Sinir Ağları (BPNN) yöntemlerini karşılaştırırken Chicago Ticaret Borsasında işlem gören beş adet vadeli işlem sözleşmesine ait verileri kullanmışlardır. Analiz sonuçlarına göre SVM'nin BPNN'den çok daha iyi sonuçlar gösterdiği raporlanmıştır. Cao ve Tay (2001), SVM ile Geri Beslemeli Algoritma kullanılarak eğitim yapılan MLP yöntemini karşılaştırmışlardır. Çalışmanın uygulama bölümünde veri seti olarak S&P 500 günlük endeks değerleri kullanılmıştır. Araştırma sonuçlarına göre SVM'nin daha üstün performans gösterdiği ortaya konmuştur. Huang vd. (2005), hisse değerlerini tahmin etmek amacıyla SVM tabanlı model kullanmışlar ve kullanılan modelin Doğrusal Diskriminant Analizi (DA), Kuadratik Diskriminant Analizi ve Elman Geri Beslemeli Sinir Ağları yöntemlerine göre daha iyi sonuçlar elde ettiğini göstermişlerdir.

Her ne kadar yukarıda bahsi geçen çalışmalarda SVM'nin daha üstün bir yöntem olduğu düşünülse de, NN' nin bir türü olan derin öğrenme metotları ile de bu alana katkı sağlanmış ve iyi sonuçlar elde edilebilceği gösterilmiştir. Demirel vd. (2021) Borsa İstanbul (BIST)'da işlem gören 42 firmanın açılış ve kapanış değerlerini tahminlemek amacıyla MLP, SVM ve LSTM

makine öğrenmesi yöntemlerini kullanmışlardır. Analiz sonuçlarına göre MLP ve LSTM, SVM'den daha iyi sonuçlar göstermektedir. Akita vd. (2016), Paragraph Vector ve LSTM yöntemlerini kullanarak finansal zaman serisi öngörümlemesi yapmışlardır. Yu vd. (2021), beş farklı hisse senedinin Açılış, Kapanış, En Yüksek, En Düşük, Hacim, Düzeltilmiş Kapanış ve Dönüştürülmüş Zaman değerlerini XGBoost yöntemi ile eğitmişler ve bu eğitim setini LSTM modeli için girdi değişkenleri olarak kullanmışlardır. Oluşturulan LSTM-XGBoost yöntemi LSTM ve RNN'den daha iyi sonuçlar verdiği raporlanmıştır.

RF ve XGBoost gibi topluluk makine öğrenmesi yöntemler ile araştırmalarını yürüten diğer bazı çalışmalar ise şöyledir: Vijh vd. (2020) farklı sektörlere ait 5 şirketin bir gün sonraki kapanış fiyatlarını NN ve RF kullanarak tahminlemişlerdir. Hisse senetlerinin Açılış, En Yüksek, En Düşük ve Kapanış değerleri kullanılarak yeni değişkenler oluşturulmuş ve oluşturulan bu değişkenler girdi olarak kullanılmıştır. Yeni değişkenlerin eklenmesiyle beraber makine öğrenmesi yöntemlerinin daha iyi sonuçlar elde ettikleri görülmüştür. Kumar vd. (2021), Hindistan'ın Bombay Borsasında (BSE30) işlem gören Infosys şirketinin hisse değerlerini LSTM yöntemi ile tahminlemişlerdir. Kullanılan yöntemi ARIMA ve XGBoost modelleri ile karşılaştırmışlardır. LSTM'nin RMSE ve MAPE açısından diğer yöntemlere göre daha iyi sonuçlar gösterdiğini raporlamışlardır. Arslankaya ve Toprak (2021) makine öğrenmesi yöntemlerinden RF ve Polinom Regresyon ile derin öğrenme yöntemlerinden RNN ve LSTM yöntemleri kullanılarak Ereğli Demir ve Çelik Fabrikaları kapanış hisse değerleri tahminlemişlerdir. Araştırma bulgularına göre RF Regresyon en iyi performansı göstermiştir.

Karar ağaçlarının performansının araştırıldığı bazı çalışmalar ise şöyledir: Zhang vd. (2017), Ensemle Empirical Mode Decomposition (EEMD) ve Çok Boyutlu k- En Yakın Komşuluk (MKNN) yöntemlerini kullanarak iki aşamalı olarak hisse senedi fiyatlarının kapanış ve en yüksek değerlerini öngörümlemişlerdir. Önerilen ensemle yöntem Empirical Mode Decomposition (EMD), EMD-KNN, KNN ve ARIMA yöntemlerine göre daha üstün performans göstermiştir. Asfaq vd. (2021) karar ağaçlarının da içinde bulunduğu oldukça geniş sayıda çalışma ile alana önemli bir katkı sağlamıştır. Bu çalışmada NASDAQ borsasında işlem gören 10 şirketin kapanış fiyatlarını önceki kapanış fiyatlarını ele alarak tahminlemeye çalışmışlardır. Bu nedenle 9 farklı makine öğrenmesini karşılaştırmışlar ve en düşük Ortalama Karesel Hataya (RMSE) sahip olan makine öğrenmesi yöntemlerini belirlemişlerdir. Uygulama sonuçlarına göre SVR, Lasso, Elastic Net ve Ridge Regresyon, Karar Ağaçları, Extra Tree ve

Ransac yöntemlerine göre daha yüksek doğruluk oranlarına sahip bir yöntem olarak bulunmuştur.

ML yöntemlerinin bireysel olarak kullanılması kolay ve anlaşılır olsa da araştırmacılar genellikle teorik alt yapısı birbirine benzeyen yöntemleri bir arada kullanmayı denemişlerdir. Böylece bireysel olarak karşılaştırdıkları yöntemleri hibrit hale dönüştürmüş ve performanslarını incelemişlerdir. Hibrit çalışmalarda, öncelikle bir ML yöntemi ile tahminleme yapılır ve elde edilen sonuç başka bir ML yöntemi için girdi olarak kullanılır. Bu yöntemin avantajı veri seti üzerinde birden fazla eğitim yapılarak ML yönteminin tahminleme performansını artırmasıdır. Abraham vd. (2001), zaman serisi öngörümlemesinde ilk defa hibrit yöntemlerden birini önermişlerdir. Bu çalışmada önerilen hibrit yöntem Sinir Ağları- Bulanık Mantık ve NN yöntemlerini içermektedir ve hibrit olmayan yöntemlere göre daha karmaşık olmasına rağmen daha iyi sonuçlar vermektedir. Enke ve Thawornwong (2005), geleneksel veri madenciliği yöntemleriyle NN'yi birleştirerek hisse değerlerini tahmin etmişlerdir. Araştırmacılar üç ileri beslemeli katmana sahip Olasılıksal Sinir Ağları (PNN) yöntemini kullanmışlardır. Bulanık mantık temelli bir diğer çalışma Fu vd. (2007)'ne aittir. Bu çalışmada Fuzzy Cerebellar Model Articulation Controller- Bayesyan Ying Yang NN yöntemlerini önermişlerdir. Tsai ve Wang (2009), hisse senedi öngörümlemesinde Karar Ağaçları (Decision Tree, DT) ve NN yöntemlerini birleştirmişlerdir. Birleştirilen model DT-NN'nın diğer yöntemlerden daha doğru sonuçlar göstermiştir. Meta sezgisel yöntemlerin avantajlarından yararlanarak ana tahmin modeli güçlendiren diğer çalışmalarda umut vadeci sonuçlar elde etmişlerdir. Kim ve Han (2000), NN bağlantı ağırlıklarının belirlenmesinde Genetik Algoritmalar (Genetic Algorithm, GA) yaklaşımını hisse senedi fiyat tahminlemesinde kullanmışlardır. Deneysel sonuçlar önerilen yöntemin üstün performansa sahip olduğunu göstermektedir. Chaudhary ve Garg (2008), borsa tahmininde GA-SVM yöntemini önermişlerdir. Önerilen bu yöntem GA ile SVM'nin hibrit versiyonudur. Yöntem Hindistan'ın üç en büyük şirketi TCS, Infosys ve RIL ile otuz farklı şirket üzerinde uygulanmıştır. GA-SVM yöntemi diğer yöntemlere göre daha iyi sonuçlar göstermiştir.

Geleneksel yöntemler ile ML yöntemlerini kıyaslayan çalışmalar olduğu gibi iki yöntemin avantajlarından faydalanmak adına hibrit yöntemlerin önerildiği çalışmalarda bu alana katkı sağlamıştır. Egeli vd. (2003), MLP ve Genelleştirilmiş İleri Beslemeli Ağ Tabanlı mimari yöntemlerini önermişlerdir. Araştırma bulgularına göre önerilen yöntemler ARIMA gibi geleneksel yöntemlere göre daha iyi sonuçlar vermiştir. Pai ve Lin (2005), geleneksel

ekonometrik yöntem ARIMA ile SVM yöntemini birleştirerek finansal zaman serisine uygulamışlardır. Seçilen 10 farklı hisse senedi fiyatı analiz edilmiş ve önerilen hibrit modelin SVM ve ARIMA yöntemlerinden daha iyi olduğu raporlanmıştır.

Finansal zaman serisi tahminlemelerinde en yaygın problem verilerin yüksek gürültü içermesidir. Bazı araştırmacılar bu soruna eğilerek zaman serisi tahminlemesinde bulunmuştur. Lu vd. (2009) yüksek gürültü problemini aşmak amacıyla Bağımsız Bileşenler Analizi (ICA) ve SVR yöntemlerini sırasıyla kullanmayı önermişlerdir. Çalışmanın uygulama bölümünde Nikkei 225 açılış endeksi ile TAIEX kapanış endeksini veri seti olarak ele almışlardır. Önerilen ICA-SVR yönteminin SVR' ye göre daha üstün performans gösterdiğini raporlamışlardır.

Finansal zaman serisi tahminlemede en sık çalışılan uygulama alanı borsa endeks tahminleri olmuştur. Kim (2003) Kore birleşik fiyat endeksini (KOSPI); Lu vd.,2009 Tayvan Borsası (TAIEX) ve Hang Seng Endeksini (HSI) ; Chen vd. (2014) Tayvan Borsa endeksini (TAIEX) ve Hang Seng Endeksini (HSI); Yakut vd. (2014) Borsa İstanbul (BIST) endeksini; Yetis vd. (2014) ve Moghaddam vd. (2016) NASDAQ borsasının endeks değerlerini; Rasel vd. (2015) Dhaka Stock Exchange (DSE) endeksini, S&P 500 endeksini ve IBM endeksini; Yan vd. (2016) Shangai Birleşik Endeksini; Akita vd. (2016) Tokyo Borsası Endeks değerlerini tahmin etmeye çalışmışlardır. Bunula beraber borsa endeksleri içerisinde yer alan belli başlı şirketlerin borsa değerlerine ilişkin tahmin yapan çalışmalar da bulunmaktadır. Chen vd. (2017), S&P 500'de bulunan 100 hisse senedini; Henrique vd. (2018), Çin, Brezilya ve Amerika borsalarında faaliyet gösteren toplam 18 şirkete ait verileri; Demirel vd. (2021) Borsa İstanbul (BIST)'da işlem gören 42 firmanın açılış ve kapanış değerlerini; Asfaq vd. (2021) NASDAQ borsasında işlem gören 10 şirketin kapanış fiyatlarını incelemişlerdir.

Bazı araştırmacılar makine öğrenmesi yöntemleriyle ekonomistlerin bilgiye dayalı yöntemleri arasında var olan zıtlıkları tartışmışlardır ve günlük borsa verilerine ilave olarak teknik göstergeleri girdi değişken olarak kullanmayı tercih etmişlerdir (Kim 2003; Yan vd. 2016). Hsu vd. (2016), otuz dört finansal endeksi altı yıldan uzun bir süre için incelenmiş ve sonuç olarak makine öğrenmesi yöntemlerinin ekonomistlerin kullandıkları yöntemlerinden daha üstün olduklarını ileri sürmüşlerdir. Kumar vd. (2016), veri setinden hesapladıkları 55 adet teknik göstergeyi kullanarak borsa endeksini tahminlemeye çalışmışlardır. Bu teknik göstergeler içerisinden en uygun göstergeleri hangi yöntemin ortaya koyabileceğini tartışmışlardır.

2. Yöntem

Bu bölümde literatürde sıklıkla kullanılan ve araştırmacılar tarafından güçlü sonuçlar elde ettikleri raporlanan makine öğrenmesi yöntemlerinden SVR, RF ve XGBoost'un teorik altyapısı sunulmuştur.

2.1. Destek Vektör Regresyonu (Support Vector Regression, SVR)

İlk defa Vapnik ve Cortes (1995) tarafından önerilen SVM, sıklıkla regresyon ve sınıflama problemlerinde tercih edilen bir makine öğrenmesi yöntemidir. Düşük boyuta sahip veri setlerinin doğrusal olarak ayrıştırılamadığı durumlarda SVM, bu veri setini yüksek boyuta sahip uzaya çıkararak doğrusal ayrıştırılabilir hale getirmektedir (Tay ve Cao 2001, Gunn 1998). SVM, regresyon problemlerinde kullanıldığında Destek Vektör Regresyonu (SVR) ismini almaktadır.

Bir eğitim setinde x_i ve y_i sırasıyla girdi ve çıktı değişkenleri olmak üzere $T = \{(x_1, y_1), \dots, (x_n, y_n)\}$ olduğu varsayılsın. SVR, eğitim verisindeki gerçek değer y_i'den en fazla ε kadar uzakta doğrusal bir fonksiyon bulabilmeyi amaçlar (Smola ve Scholkopf, 2004). Bu fonksiyonu elde etmek amacıyla Vapnik ve Cortes (1995), Eşitlik 1'deki optimizasyon fonksiyonunu önermişlerdir;

$$\begin{gathered} min\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}(\xi_i - \zeta_i) \\ kısıt(w^T x_i + b) - y_i \leq \varepsilon + \xi_i \\ y_i - (w^T x_i + b) \leq \varepsilon + \zeta_i \\ \xi_i, \zeta_i \geq 0, i = 1,2, \dots, N \end{gathered} \tag{1}$$

Eşitlik 1'deki amaç fonksiyonunda, $\|w\|^2$genelleme yeteneğini yansıtan güven aralığını, $\sum_{i=1}^{N}(\xi_i + \zeta_i)$ fonksiyonun öğrenme kapasitesini belirleyen deneysel riski, ξ_i ve ζ_i hatanın kabul edilebilir üst ve alt sınırlarını, $\varepsilon > 0$ duyarsız kayıp parametre ve $C > 0$ ise ceza parametresini temsil etmektedir. Eşitlik 1'deki dual problem çözülürken genellikle Lagrange çarpanları yöntemi kullanılır. Lagrange çarpanları yöntemi kullanılarak elde edilen regresyon fonksiyonu aşağıdaki gibi yazılabilir:

$$f(x) = w^T x + b = \sum_{i=1}^{N}(\hat{\alpha}_i - \alpha_i)K(\boldsymbol{x_i}, \boldsymbol{x}) + b \tag{2}$$

Eşitlik 2'de $\hat{\alpha}_i$ ve α_i Lagrange çarpanları, $\hat{\alpha}_i \alpha_i = 0$ ve $K(x_i, x)$ kernel fonksiyonudur. Kernel fonksiyonunun değerleri özellik uzayı $\boldsymbol{\phi(x_i)}$ ve $\boldsymbol{\phi(x_j)}$' de bulunan $\boldsymbol{x_i}$ ve $\boldsymbol{x_j}$ vektörlerinin iç çarpımlarına eşittir ve $\boldsymbol{K(x_i, x_j) = \phi(x_i)\phi(x_j)}$ olarak yazılabilir. (Cherkassky ve Ma, 2004; Vapnik,2000; Vapnik,1999;). Literatürde sıklıkla kullanılan kernel fonksiyonları Eşitlik 3-5'te verilmiştir. Bu fonksiyonlar sırasıyla doğrusal kernel, radial tabanlı kernel ve polinomial kernel fonksiyonlarıdır.

$$K(x_i, x_j) = x_i^T x_j \tag{3}$$

$$K(x_i, x_j) = e^{-\gamma \|x_i - x_j\|^2}, \gamma > 0 \tag{4}$$

$$K(x_i, x_j) = \left(x_i^T x_j + 1\right)^d \tag{5}$$

Burada γ ve d parametreleri genellikle araştırmacılar tarafından deneme yoluyla belirlenmektedir.

2.2. Rassal Orman Regresyonu (Random Forest Regression, RF)

Rassal Orman (RF), Breiman (2001) tarafından geliştirilen karar ağaçlarına dayanan sınıflama ve regresyon analizinde sıklıkla kullanılan bir topluluk makine öğrenmesi yöntemidir. RF, modelin sapmasını ve varyansını azaltmak amacıyla çok sayıda karar ağacını birbirlerine paralel olacak şekilde büyütür. Modelin eğitilmesi amacıyla veri setinden bootstrap yöntemiyle ile örneklemler çekilir ve her örnek sınıflama veya regresyon ağaçlarını büyütmek amacıyla kullanılır. Burada önemli olan adım ise tüm tahmin ediciler yerine, bölünmüş örneklemler içerisinden daha az sayıda ve uygun olan tahmin ediciler seçilir. Bu adım karar ağacı yeteri kadar büyüyene kadar devam ettirilir ve yeni veri seti bu karar ağaçlarından elde edilen tahminlerin toplanmasıyla oluşturulur. RF regresyonuna ait fonksiyon Eşitlik 6'daki gibi gösterilebilir:

$$f_{RF}^{C}(\boldsymbol{x}) = \frac{1}{C}\sum_{i=1}^{C} T_i(\boldsymbol{x}) \tag{6}$$

Eşitlik 6'da $\boldsymbol{x}$ girdi vektörü, C ağaç sayısı ve $T_i(\boldsymbol{x})$ ise bootstrap örnekleri kullanılarak elde edilen regresyon ağacıdır (Ahmad,2018). RF regresyonu kurulurken izlenecek adımlar aşağıdaki gibidir:

1. n adet bootstrap örneği yerine koyma yöntemiyle orijinal veri setinden çekilerek bireysel olarak regresyon ağaçlarının büyütülmesi amacıyla kullanılır.
2. Eğitim veri seti içerisinden elde edilen örneklere "başarılı örnekler" olarak adlandırılırken geriye kalan örnekler "başarısız örnekler" olarak adlandırılır.
3. Her bir karar ağacını inşa etmek amacıyla rastgele sayıda özellik seçilerek yapraklar ve düğümler oluşturulur. Böylece modelin doğrulunu artırmak amaçlanır.
4. RF regresyonun probleminde, bir özellik "kök düğüm" olarak atanır ve veri seti sırasıyla bölünür ve dallara ayrılır. Böylece yukarıdan başlanılarak aşağıya doğru bir regresyon ağacı kurulur (Wang vd., 2016).

2.3. Ekstrem Gradyan Arttırma (XGBoost)

Temeli Karar Ağaçlarına dayanan topluluk makine öğrenmesi yöntemi olan Extreme Gradient Boosting (XGBoost), genellikle regresyon ve sınıflandırma problemlerinde kullanılmaktadır. XGBoost'un en önemli özellikleri arasında; ölçeklenebilir olması ve yüksek hesaplama yeteneğine rağmen düşük hafızaya gereksinim duyması gösterilebilir (Chen ve Guestrin, 2016). Ayrıca XGBoost, makine öğrenmesi yöntemlerinde sıklıkla karşılaşılan bir problem olan aşırı öğrenme ve aşırı uyum sorunlarını en aza indiren bir "düzeltme terimine" sahiptir. Düşük hafıza gerekliliği ve yüksek hesaplama yeteneği nedeniyle veri boyutunun sürekli ve hızlı arttığı finansal zaman serisi alanında çalışan araştırmacı ve uygulamacılar için kullanışlı bir yöntem haline gelmiştir.

XGBoost topluluk makine öğrenmesi yönteminin amaç fonksiyonu;

$$L(\phi) = \sum_i \ell(y_i, \hat{y}_i) + \sum_k \Omega(f_k) \tag{7}$$

şeklinde tanımlanabilir. Eşitlik 7'de, ℓ, konveks ve türevlenebilir bir kayıp fonksiyonudur. Bu fonksiyon, tahminlenen değer ($\hat{y}_i$) ile gerçek değer (y_i) arasındaki farkı ölçmektedir. Her bir f_k ise birbirlerinden bağımsız olan bir ağaç yapısına ve bağımsız ağaçların yaprak ağırlıklarına karşılık gelmektedir. Düzeltme terimi (regularization term) Ω, Eşitlik 7'in ikinci bileşeni ise;

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \tag{8}$$

ile gösterilebilir. Eşitlik 8'de, karmaşıklık parametresini γ, karar ağacındaki yaprak sayısını T, yaprak düğümlerinin ağırlıklarını w ve aşırı uyum parametresini λ göstermektedir. XGBoost yöntemine ait önemli özellikler aşağıda sıralanmıştır;

- Veri setinde içerisinde kayıp veya eksik gözlem varsa XGBoost bu gözlemleri otomatik olarak doldurur (sparsity-aware). Bu durumda araştırmacı tamamlanmış veri setini kullanabilir.

- Makine öğrenmesi yöntemlerinde genellikle kullanılan eşit ağırlıklı veri setlerine ek olarak XGBoost ağırlığın eşit olmadığı veri setleri üzerinde de rahatlıkla kullanılabilir (distributed weighted quantile sketch).

3. Uygulama

Model tahmini için seçilen tüm makine öğrenmesi algoritmaları Python'da scikit-learn[1] kütüphanesi kullanılmıştır.

3.1. Veri Seti

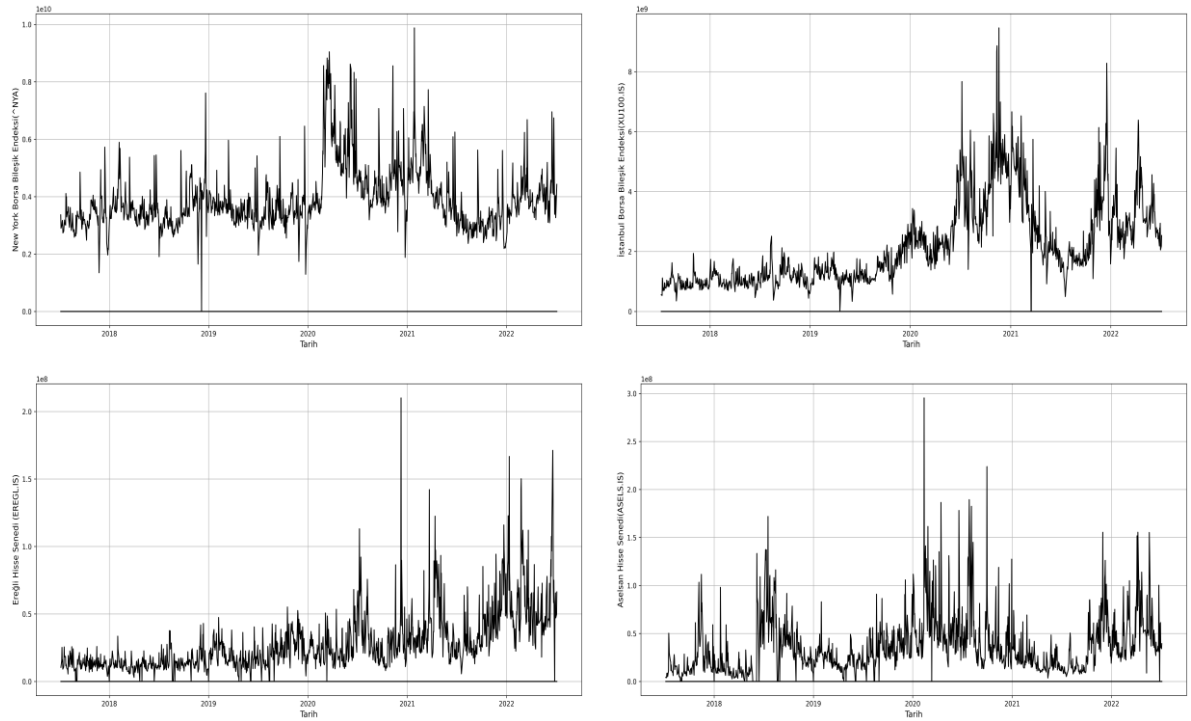
Önerilen tahmin modellerini eğitmek ve performanslarını test etmek için New York Borsası (NYSE) ve İstanbul Borsasına (BİST 100) ait birleşik endeksler tercih edilmiştir. Bu iki endeksten ilki Amerika ve ikincisi ise Asya piyasasını temsil etmektedir. Bu endeksler gelişmiş ve gelişmekte olan borsalara ait olmaları sebebi ile seçilmiştir. Endekslerin deneysel veri olarak seçilmesinin nedeni, bireysel hisse senetlerinin şiddetli dalgalanmalarının tesadüfiliğini dışlayabilen, finansal piyasaların oynaklığının yoğunlaştırılmış ifadesi olmasından kaynaklanmaktadır (Cao ve Wang, 2019). Bu nedenle çalışmamızda iki borsa endeksinin yanısıra Borsa İstanbul'da işlem gören iki bireysel hisse senedi de seçilmiştir. Hisse senedi olarak Borsa İstanbul'da işlem gören en yüksek hacimli 10 hisse senedinin içinden Ereğli Demir ve Çelik ile Aselsan Şirketlerine ait hisse senetleri seçilmiştir. Endeks ve hisse senetlerine ait günlük kapanış verileri Yahoo Finans'den (http://finance.yahoo.com) temin edilmiştir. Şekil 1'de 07-07-2017 ve 06-07-2022 aralığında 5 yıllık bir dönemi sunulan finansal zaman serisi verilerinin oldukça değişkenlik gösterdiği ve durağan olmadıkları gözlenmektedir. Zaman serilerine ilişkin temel istatistikler Tablo 1'de sunulmuştur. Her bir zaman serisi %80-%20 kuralına göre eğitim ve test seti olarak parçalanmıştır. Farklı koşul ve nedenlerden dolayı borsada işlem yapılmayan günler olduğundan bu veriler hariç tutulmuştur.

[1]scikit-learn: https://scikit-learn.org/stable/supervised_learning.html#supervised-learning

Tablo 1. Endeks ve Hisse Senedi Kapanış Fiyatlarına İlişkin Tanımlayıcı İstatistikler

	Gözlem	*Ortalama*	*S.Sapma*	*Max*	*Min*	*%25*	*%50*	*%75*
NYSE	1258	13708.6	1805.86	17353.8	8777.38	12495.4	12975.8	15320.1
BİST	1247	1268.96	405.563	2648.2	836.753	1000.49	1107.85	1413.65
EREGLİ	1276	9.86112	8.20462	36.5	3.47162	4.71002	5.7637	12.9687
ASELSAN	1276	14.713	4.2825	28.52	8.08813	11.6132	14.5798	16.7195

Hisse senedi piyasalarında, pratik deneyim, bir endeksin kapanış fiyatının bir önceki işlem gününe ait kapanış fiyatlarından, yani zaman serisinin t-1 dönemindeki değerlerinden etkilendiğini göstermiştir. Bu nedenle çalışmada t-7 dönemine kadar gecikmeli değerler dikkate alınarak tahmin modeli kurulmuştur. Şekil 1'de sırasıyla NYSE, BİST100, Ereğli Demir ve Çelik ile Aselsan endeks ve hisse değerlerinin zaman serisi grafikleri verilmiştir.



Şekil 1. Endeks ve Hisse Senedi Değerlerinin Grafikleri

3.2. Model Değerlendirme Kriterleri

Tahmin sonuçlarının performansları, önceki çalışmaların benimsediği yaklaşımlar dikkate alınarak Ortalama Mutlak Hata (MAE), Ortalama Karesel Hatanın Karekökü (RMSE) ve Ortalama Mutlak Yüzde Hata (MAPE) ölçütleri ile değerlendirilmiştir. Bu ölçütler Eşitlik 9-11'te verildiği gibi hesaplanmıştır.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \hat{Y}_i\right| \tag{9}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2} \tag{10}$$

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_i - \hat{Y}_i}{Y_i}\right| \times 100\% \tag{11}$$

Burada, n verilerin toplam sayısı; Y_i ve $\hat{Y}_i$, sırasıyla i zamanındaki gerçek ve tahmin değerleridir. MAE, RMSE ve MAPE değerlendirme kriterlerinin değerleri daha küçük olduğunda tahmin performansı daha iyi olarak değerlendirilmektedir.

3.3. Parametre Seçimi ve Yapılandırma

Makine öğrenmesi çalışmalarında veri seti için uygun olan parametreleri belirlemek tahmin performansını artırdığından bu çalışmada kullanılan yöntemlere ait en optimal parametreler (hyper-parameter tuning) Grid Arama (Izgara) yöntemi kullanılarak belirlenmeye çalışılmıştır. Zaman serisi verileri doğaları gereği geçmiş değerlerle ilişkili olduğundan Çapraz Doğrulama (Cross Validation) yöntemi yerine İleri Adımlı Doğrulama (Kayan Pencereler, Walk-Forward Validation) yöntemi (Hu vd., 1999) kullanılmıştır. Bu yöntemde eğitim veri seti bir sonraki değeri tahminlemek için sürekli olarak kaydırılır ve tahminlenen gözlem eğitim veri setine dahil edilir. Böylece zaman serisi verisinin önceki verilerle olan ilişkisi korunmuş olur. Tablo 2'de her bir makine öğrenmesi yönteminde denenen parametreler ve seçilen optimal parametrelere yer verilmiştir. NYSE ve BİST100 veri setinde kullanılan gamma parametresi için aralık genişletilmiş ve alt sınırı 2^{-25}'e kadar çekilmiştir.

Tablo 2: Hiper-parametreler ve Kullanılan Optimal Parametreler

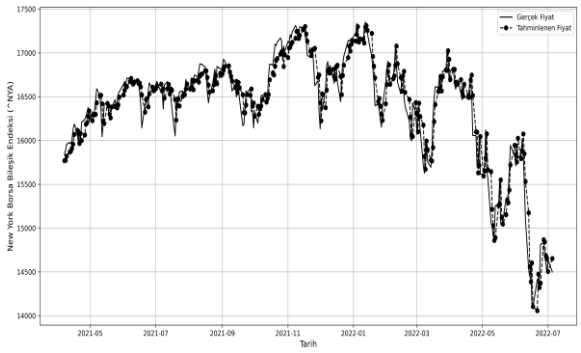
Yöntem	*Hiperparametreler*	*Parametre Aralığı*	*NYSE*	*BİST100*	*EREĞLİ*	*ASELSAN*
SVR	***C***	$2^{-5,-3,\ldots 15}$	2^{13}	2^{13}	2^{11}	2^{11}
	gamma	$2^{-15,-13,\ldots 5}$	2^{-25}	2^{-25}	2^{-15}	2^{-15}
RF	***n_estimators***	200, 300, 400, 500, 750, 1000, 2000	500	300	1000	500
	max_depth	5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100	5	20	50	10
	min_sample leaf	1, 2, 4, 8	2	1	2	2
	min_samples_split	2, 5, 10	2	2	1	2
XGBoost	***max_depth***	2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 50	6	6	7	5

learning_rate	0.0001, 0.001, 0.01, 0.1, 0.2, 0.3	0.1	0.1	0.1	0.1
n_estimators	5, 10, 50, 60, 70, 80, 90, 100, 200, 300, 500,1000	300	300	90	90
colsample_bytree	0.1, 0.5, 0.9, 1	0.9	0.9	0.9	1
min_child_weight	1, 2, 3, 4, 5, 10, 50, 100	2	3	2	4
subsample	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9	0.9	0.9	0.6	0.9

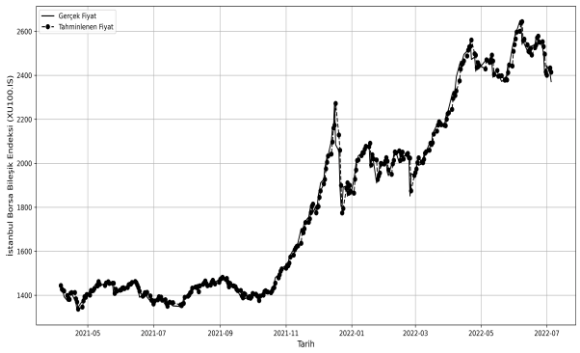
RMSE kriterine göre en düşük değeri veren parametre optimal değer olarak belirlenmiştir. Burada RF ve XGBoost yöntemlerinde denenen her bir parametre ile elde edilen RMSE değerlerinin sapmasının daha az, SVR'nin sapmasının daha fazla oluşu dikkati çeken bir nokta olmuştur.

3.4. Uygulama Sonuçları

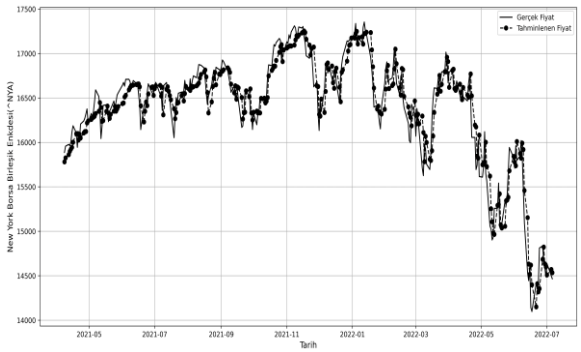
İki endeksin tahmin sonuçları Şekil 2'de ve iki hisse senedinin tahmin sonuçları Şekil 3'de gösterilmektedir. Şekillere göre, finansal zaman serisinin tahmin değerleri ile gerçek değerleri çok yakın hatta çoğu zaman dilimlerinde birebir örtüşmektedir. Önerilen modellerin performansını daha doğru bir şekilde değerlendirmek için MAE, RMSE ve MAPE ölçütleri tercih edilmiş ve tahmin sonuçları Tablo 3'de sunulmuştur.



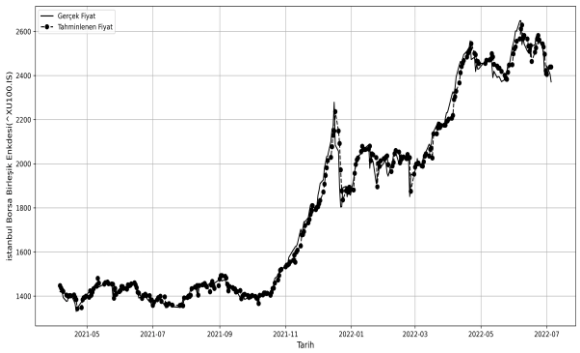
SVR -NYSE



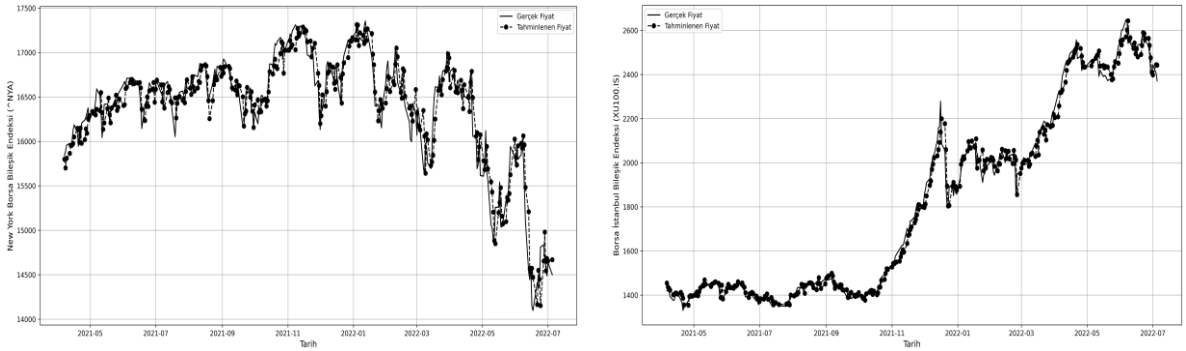
SVR BİST 100



RF -NYSE



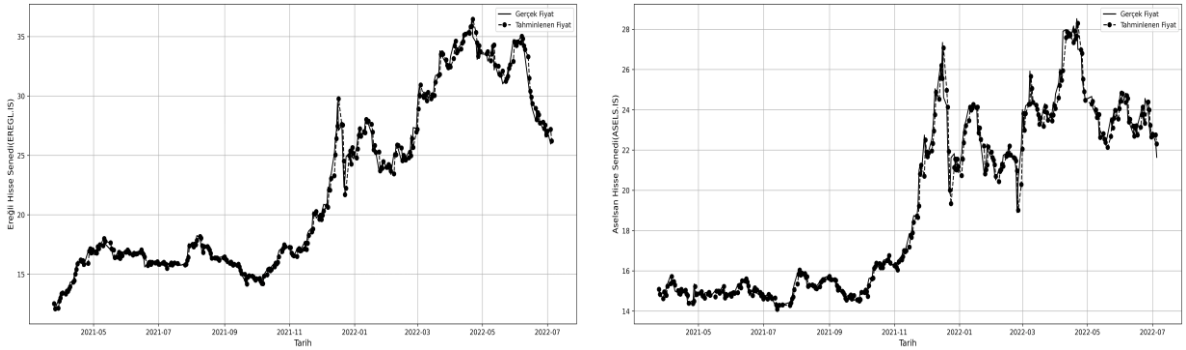
RF BİST 100



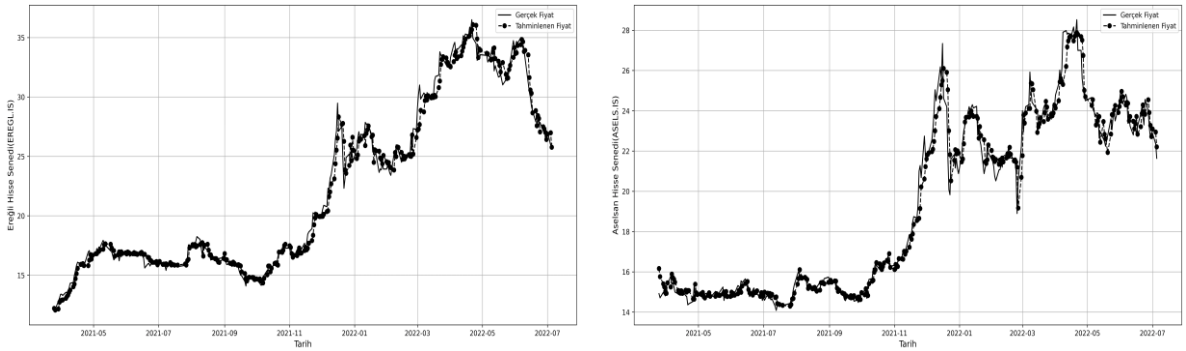
XGBoost -NYSE XGBoost BİST 100

Şekil 2: NYSE ve BİST 100 Veri Setleri için Gerçek ve Tahminlenen Endeks Değerleri

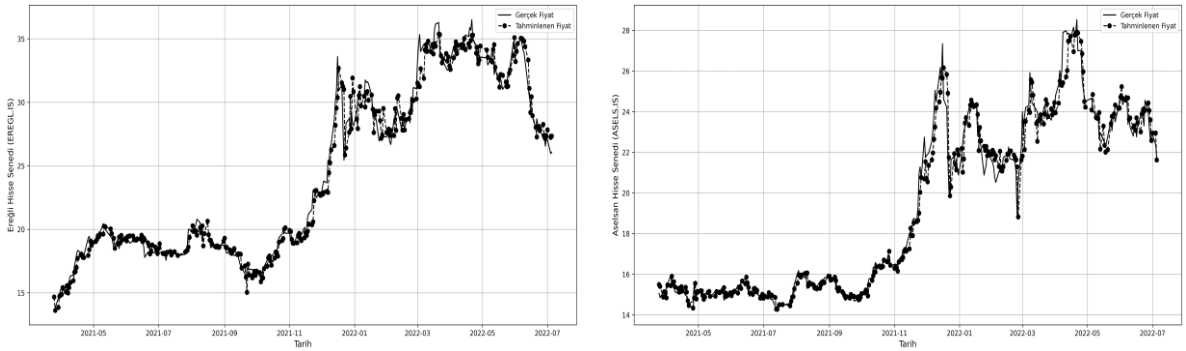
Borsaların nispeten istikrarlı olduğu dönemlerde tahmin sonucu gerçek değerine daha yakın olmaktadır. Ciddi bir artış gözlemlenen 2021 ekim ayından itibaren tırmanışa geçen İstanbul borsasının kapanış değerlerinin tepe ve dip yaptığı noktalarda tahmininin nispeten sapma gösterdiği dikkat çekmektedir. Özellikle RF ve XGBoost modelleri gerçek değerleri yakalamakta zorlanırken SVR'nin bu kritik değerleri daha doğru tahmin ettiği söylenebilir.



SVR - EREĞLİ SVR - ASELSAN



RF - EREĞLİ RF - ASELSAN



XGBoost - EREĞLİ XGBoost - ASELSAN

Şekil 3: Ereğli ve Aselsan Hisse Senetlerine ait Gerçek ve Tahminlenen Fiyatlar

Ortalama olarak son 300 güne karşılık gelen test veri seti üzerinde elde edilen tahmin sonuçlarının sapmaları Tablo 3'de yer almaktadır. MAE, MAPE ve RMSE performans ölçütlerinden gözlemlenebildiği üzere, tüm makine öğrenmesi yöntemlerinin kapanış fiyat tahminlerinin genel olarak iyi olduğu söylenebilir.

Tablo 3. Modellerin Tahmin Performansları

Endeks/Hisse senedi	*Kriter*	*RF*	*XGBoost*	*SVR*
NYSE	***MAPE***	0.0085	0.0091	**0.0083**
	MAE	139.6850	148.4244	**135.8598**
	RMSE	180.0295	191.5049	**175.3148**
BIST	***MAPE***	0.0142	0.01440	**0.0115**
	MAE	27.3783	27.4188	**22.0341**
	RMSE	40.2260	39.4288	**33.9484**
EREĞLİ	***MAPE***	0.0250	0.0268	**0.0199**
	MAE	0.5739	0.6735	**0.4569**
	RMSE	0.8258	0.9807	**0.6749**
ASELSAN	***MAPE***	0.0209	0.0210	**0.0183**
	MAE	0.4344	0.4384	**0.3769**
	RMSE	0.6792	0.6637	**0.5744**

SVR'nin tahmin sonuçlarına göre, NYSE endeksinde ölçülen MAE, RMSE ve MAPE değerleri sırasıyla 0.0083, 135.8598 ve 175.3148 ve BİST 100 endeksinde sırasıyla 0.0115, 22.0341 ve 33.9484 olup diğer modellerden çok daha küçüktür.

Benzer şekilde SVR Ereğli ve Aselsan hisse senetlerini tahmin etmede de diğer modellere göre daha iyi performans göstermiştir. SVR doğrusal olmayan regresyon tahmin problemlerini yüksek boyutlu bir özellik uzayında tanımlanan bir dizi doğrusal fonksiyon kullanarak risk minimizasyonu prensibi ile çözmektedir. SVR'nin bu doğası pek çok sınıflama probleminde olduğu gibi regresyon tahmininde de onu başarılı bir yöntem yapmıştır. RF ise nispeten daha küçük farklarla XGBoost'tan daha iyi sonuçlar verdiği gözlenmektedir.

4. Sonuç

Finansal zaman serilerinin tahminlenmesi/ öngörümlemesi hem yatırımcılar hem de araştırmacılar için her zaman merak uyandıran bir konu olmuştur. Önceleri geleneksel ekonometrik yöntemlerle yapılan bu tahminlemeler özellikle son 20 yılda gelişen bilgisayar teknolojileri ve yapay zeka ile yeni bir boyut kazanmıştır. Geleneksel yöntemlerin durağanlık, doğrusallık ve normallik gibi varsayımlarını gerektirmeyen ML yöntemleri hem işlem kapasitesi hem de uygulama kolaylığı açısından araştırmacılar için bir kurtarıcı olarak görülmektedir. Bu çalışmada, dünyanın en büyük borsası ve toplam değeri 20 trilyon dolardan daha fazla olan NYSE borsa endeksi ile BİST 100 endeksi kullanılarak SVR, RF ve XGBoost makine öğrenmesi yöntemleri MAPE, MAE ve RMSE kriterleri açısından karşılaştırılmıştır. Ayrıca çalışmada ek olarak BİST 100'de işlem gören en değerli 10 şirket arasında yer alan Aselsan ve Ereğli Demir ve Çelik hisse değerleri kullanılmıştır. Endeks değerleri ve hisse değerlerinin kullanılma amacı ise endeks değerlerinin oynaklığının daha az olmasına karşın hisse senedi değerlerinin oynaklığının daha fazla olması nedeniyle kullanılan yöntemlerin farklı yapıdaki veri setlerine karşı nasıl davranacağını belirlemektir. Çalışma genel olarak dikkate alındığında şu sonuçlara ulaşılabilir:

- Hisse senedi fiyat tahminlemesi, karmaşık ve öngörülemeyen hataları içerdiğinden kullanılan ML yöntemleri arasında SVR, temeli topluluk karar ağaçlarına dayanan RF ve XGBoost yöntemlerine göre karmaşık problemleri çözerken kullandığı yüksek boyutlu verilerde doğrusal olarak ayrıştırma yeteneği sayesinde daha iyi sonuçlar vermektedir.
- ML yöntemlerinin performansı genellikle kullanılan parametreler ile yakından ilişkilidir. Bu çalışmada ele alınan ML yöntemleri için hiper parametre ayarlaması Grid Tarama yöntemiyle yapılmış ve optimal parametreler kullanılarak analizler gerçekleştirilmiştir.
- Analiz sonuçları literatürde yapılan çalışmalarla paralellik göstermektedir. SVR yönteminin önceki çalışmalarda karar ağacı temelli algoritmalara göre daha iyi performansa sahip olduğu birçok araştırmacı tarafından raporlanmıştır.
- Karar Ağaçlarına dayanan RF ve XGboost algoritmaları SVR'den daha kötü ancak kendi aralarında birbirlerine yakın sonuçlar vermektedir.

Finansal zaman serisi tanminlemesi/ öngörümlemesinde ML yöntemlerinin hem yatırımcılar hem de araştırmacılar için kullanılabilecek faydalı bir araç olduğu görülmektedir. Sonraki çalışmalar için daha büyük veri setleri ve kullanılan yöntemlerin güçlü yanları dikkate alınarak oluşturulabilecek hibrit yöntemlerin performansları araştırılabilir.

Kaynakça

Abraham, A., Nath, B., & Mahanti, P. K. (2001, May). Hybrid İntelligent Systems for Stock Market Analysis. In *International Conference on Computational Science* (pp. 337-345), Springer, Berlin, Heidelberg.

Ahmad, M.W., Reynolds, J., Rezgui, Y. (2018). Predicti& Modelling for Solar Thermal Energy Systems: A Comparison of Support Vector Regression, Random Forest, Extra Trees And Regression Trees, *Journal of Cleaner Production*, 203, 810-821.

Akita, R., Yoshihara, A., Matsubara, T., & Uehara, K. (2016, June). Deep Learning for Stock Prediction Using Numerical and Textual İnformation. In *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)* (pp. 1-6). IEEE.

Aras, S., & Kocakoç, İ. D. (2016). A new model selection strategy in time series forecasting with artificial neural networks: IHTS. *Neurocomputing*, *174*, 974-987.

Arslankaya, S., & Toprak, Ş. (2021). Makine Öğrenmesi ve Derin Öğrenme Algoritmalarını Kullanarak Hisse Senedi Fiyat Tahmini. *International Journal of Engineering Research and Development*, *13*(1), 178-192.

Ashfaq, N., Nawaz, Z., & Ilyas, M. (2021). A Comparative Study of Different Machine Learning Regressors for Stock Market Prediction. *Arxiv Preprint Arxiv:2104.07469*.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5-32.

Brunton, S. L., Noack, B. R., & Koumoutsakos, P. (2020). Machine learning for fluid mechanics. *Annual review of fluid mechanics*, *52*, 477-508.

Cao, J., & Wang, J. (2019). Stock Price Forecasting Model Based on Modified Convolution Neural Network and Financial Time Series Analysis. *International Journal of Communication Systems*, *32*(12), e3987.

Cao, J., Li, Z., & Li, J. (2019). Financial Time Series Forecasting Model Based On CEEMDAN And LSTM. *Physica A: Statistical Mechanics and Its Applications*, *519*, 127-139.

Cao, L. J., & Tay, F. E. H. (2003). Support Vector Machine with Adaptive Parameters in Financial Time Series Forecasting. IEEE Transactions on Neural Networks, 14, 1506–1518. Doi:10.1109/TNN.2003.820556.

Cao, L., & Tay, F. E. (2001). Financial Forecasting Using Support Vector Machines. *Neural Computing and Applications*, *10*(2), 184-192.

Carbonneau, R., Laframboise, K., & Vahidov, R. (2008). Application of machine learning techniques for supply chain demand forecasting. *European Journal of Operational Research*, *184*(3), 1140-1154.

Cavalcante, R. C., Brasileiro, R. C., Souza, V. L., Nobrega, J. P., & Oliveira, A. L. (2016). Computational intelligence and financial markets: A survey and future directions. *Expert Systems with Applications*, *55*, 194-211.

Chen SM (1996) Forecasting Enrollments Based On Fuzzy Time-Series. Fuzzy Sets Syst 81:311–319

Chen, H., Xiao, K., Sun, J., & Wu, S. (2017). A Double-Layer Neural Network Framework for High-Frequency Forecasting. *ACM Transactions on Management Information Systems (TMIS)*, *7*(4), 1-17.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A Scalable Tree Boosting System. In *Proceedings of The 22nd Acm Sigkdd İnternational Conference on Knowledge Discovery and Data Mining* (pp. 785-794).

Chen, Y. S., Cheng, C. H., & Tsai, W. L. (2014). Modeling Fitting-Function-Based Fuzzy Time Series Patterns for Evolving Stock İndex Forecasting. *Applied İntelligence*, *41*(2), 327-347.

Cherkassky, V., Ma, Y. (2004). Practical Selection of SVM Parameters and Noise Estimation for SVM Regression, Neural Networks 17, 113–126.

Choudhry, R., & Garg, K. (2008). A Hybrid Machine Learning System for Stock Market Forecasting. *International Journal of Computer and Information Engineering*, *2*(3), 689-692.

Crone, S., Nikolopoulos, K.: Results of The NN3 Neural Network Forecasting Competition. The 27th International Symposium on Forecasting, Program, pp. 129 (2007).

Demirel, U., Çam H., & Ünlü R., (2021). Predicting Stock Prices Using Machine Learning Methods and Deep Learning Algorithms: The Sample of The Istanbul Stock Exchange. *Gazi University Journal of Science*, *34*(1), 63-82.

Deviren, B., Kocakaplan, Y., Keskin, M., Balcılar, M., Özdemir, Z. A., & Ersoy, E. (2014). Analysis of Bubbles and Crashes In The TRY/USD, TRY/EUR, TRY/JPY and

TRY/CHF Exchange Rate Within The Scope of Econophysics. *Physica A: Statistical Mechanics and Its Applications, 410*, 414-420.

Di Matteo, Tiziana. 2007. Multi-Scaling In Finance. Quantitative Finance 7: 21–36.

Egeli, B., Ozturan, M., & Badur, B. (2003). Stock Market Prediction Using Artificial Neural Networks. *Decision Support Systems, 22*, 171-185.

Enke, D., & Thawornwong, S. (2005). The Use of Data Mining and Neural Networks for Forecasting Stock Market Returns. *Expert Systems with Applications*, *29*(4), 927-940.

Fischer, T., & Krauss, C. (2018). Deep Learning with Long Short-Term Memory Networks for Financial Market Predictions. *European Journal of Operational Research, 270*(2), 654-669.

Fu, J., Lum, K. S., Nguyen, M. N., & Shi, J. (2007, June). Stock Prediction Using Fcmac-Byy. In *International Symposium on Neural Networks* (pp. 346-351). Springer, Berlin, Heidelberg.

Gerlein, E. A., Mcginnity, M., Belatreche, A., & Coleman, S. (2016). Evaluating Machine Learning Classification for Financial Trading: An Empirical Approach. *Expert Systems with Applications*, *54*, 193-207.

Gunn, S.R. (1998). Support Vector Machines for Classification and Regression. ISIS Technical Report (Available At: Http://Users.Ecs.Soton.Ac.Uk/Srg/Publications/Pdf/SVM.Pdf).

Hamzaçebi, C., Akay, D., & Kutay, F. (2009). Comparison of Direct and Iterative Artificial Neural Network Forecast Approaches In Multi-Periodic Time Series Forecasting. *Expert Systems with Applications, 36*(2), 3839-3844.

Hansen, J. V., Mcdonald, J. B., & Nelson, R. D. (1999). Time Series Prediction with Genetic-Algorithm Designed Neural Networks: An Empirical Comparison with Modern Statistical Models. *Computational Intelligence*, *15*(3), 171-184.

He K, Yu L, Lai KK. Crude Oil Price Analysis and Forecasting Using Wavelet Decomposed Ensemble Model. Energy 2012;46(1):564e74.

Henrique, B. M., Sobreiro, V. A., & Kimura, H. (2018). Stock Price Prediction Using Support Vector Regression on Daily And up to The Minute Prices. *The Journal of Finance and Data Science, 4*(3), 183-201.

Hsu, M. W., Lessmann, S., Sung, M. C., Ma, T., & Johnson, J. E. (2016). Bridging the Divide In Financial Market Forecasting: Machine Learners & Financial Economists. *Expert Systems With Applications, 61*, 215-234.

Hu, M. Y., Zhang, G., Jiang, C. X., & Patuwo, B. E. (1999). A Cross- Validation Analysis of Neural Network out- of- Sample Performance In Exchange Rate Forecasting. *Decision Sciences*, *30*(1), 197-216.

Huang, W., Nakamori, Y., & Wang, S. Y. (2005). Forecasting Stock Market Movement Direction With Support Vector Machine. *Computers & Operations Research*, *32*(10), 2513-2522.

Karasu, S., Altan, A., Bekiros, S., & Ahmad, W. (2020). A New Forecasting Model With Wrapper-Based Feature Selection Approach Using Multi-Objective Optimization Technique For Chaotic Crude Oil Time Series. *Energy*, *212*, 118750.

Kazem, A., Sharifi, E., Hussain, F. K., Saberi, M., & Hussain, O. K. (2013). Support Vector Regression with Chaos-Based Firefly Algorithm for Stock Market Price Forecasting. *Applied Soft Computing*, *13*(2), 947-958.

Kim, K. J. (2003). Financial Time Series Forecasting Using Support Vector Machines. *Neurocomputing*, *55*(1-2), 307-319.

Kim, K. J., & Han, I. (2000). Genetic Algorithms Approach to Feature Discretization In Artificial Neural Networks for The Prediction of Stock Price Index. *Expert Systems with Applications*, *19*(2), 125-132.

Kumar, D., Meghwani, S. S., & Thakur, M. (2016). Proximal Support Vector Machine Based Hybrid Prediction Models for Trend Forecasting In Financial Markets. *Journal of Computational Science*, *17*, 1-13.

Kumar, M., & Thenmozhi, M. (2006). Forecasting Stock index movement: A comparison of support vector machines and random forest. SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, January 24, 2006.

Kumar, R., Kumar, P., & Kumar, Y. (2021, January). Analysis of Financial Time Series Forecasting Using Deep Learning Model. In *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* (pp. 877-881), IEEE.

Kuremoto, T., Kimura, S., Kobayashi, K., & Obayashi, M. (2014). Time Series Forecasting Using A Deep Belief Network with Restricted Boltzmann Machines. *Neurocomputing*, *137*, 47-56.

Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. *Sensors*, *18*(8), 2674.

Lu, C. J., Lee, T. S., & Chiu, C. C. (2009). Financial Time Series Forecasting Using İndependent Component Analysis and Support Vector Regression. *Decision Support Systems*, *47*(2), 115-125.

Moghaddam, A. H., Moghaddam, M. H., & Esfandyari, M. (2016). Stock Market Index Prediction Using Artificial Neural Network. *Journal of Economics, Finance and Administrative Science*, *21*(41), 89-93.

Nava, N., Di Matteo, T., & Aste, T. (2018). Financial Time Series Forecasting Using Empirical Mode Decomposition and Support Vector Regression. *Risks*, *6*(1), 7.

Nava, Noemi, Tiziana Di Matteo, And Tomaso Aste. 2016(a). Time-Dependent Scaling Patterns in High Frequency Financial Data. The European Physical Journal Special Topics 225: 1997–2016.

Niu, T., Wang, J., Lu, H., Yang, W., & Du, P. (2020). Developing A Deep Learning Framework with Two-Stage Feature Selection for Multivariate Financial Time Series Forecasting. *Expert Systems with Applications*, *148*, 113237.

Nonejad N. Prediction Aggregate Stock Market Volatility Using Financial and Macroeconomic Predictors: Which Models Forecast Best, When and Why?. J Empir Financ. 2017;42:131- 154.

Pai, P. F., & Lin, C. S. (2005). A Hybrid ARIMA and Support Vector Machines Model In Stock Price Forecasting. *Omega*, *33*(6), 497-505.

Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015(a)). Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques. Expert Systems with Applications, 42, 259–268.

Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015(b)). Predicting Stock Market Index Using Fusion of Machine Learning Techniques. *Expert Systems with Applications*, *42*(4), 2162-2172.

Rasel, R. I., Sultana, N., & Meesad, P. (2015). An Efficient Modelling Approach for Forecasting Financial Time Series Data Using Support Vector Regression and Windowing Operators. *International Journal of Computational Intelligence Studies*, *4*(2), 134-150.

Ser-Huang Poon, Forecasting Volatility In Financial Markets: A Review, J. Econ. Lit. 41 (2) (2003) 478–539.

Sezer, O. B., Gudelek, M. U., & Ozbayoglu, A. M. (2020). Financial Time Series Forecasting with Deep Learning: A Systematic Literature Review: 2005–2019. *Applied Soft Computing*, *90*, 106181.

Shailaja, K., Seetharamulu, B., & Jabbar, M. A. (2018, March). Machine learning in healthcare: A review. In *2018 Second international conference on electronics, communication and aerospace technology (ICECA)* (pp. 910-914). IEEE.

Smola, A.J., Scholkopf, B., 2004. A Tutorial on Support Vector Regression. Stat. Comput. 14, 199–222.

Tay, F. E., & Cao, L. (2001). Application of Support Vector Machines In Financial Time Series Forecasting. *Omega*, *29*(4), 309-317.

Tosunoğlu, E., Yılmaz, R., Özeren, E., & Sağlam, Z. (2021). Eğitimde makine öğrenmesi: Araştırmalardaki güncel eğilimler üzerine inceleme. *Ahmet Keleşoğlu Eğitim Fakültesi Dergisi*, *3*(2), 178-199.

Tsai, C. F., & Wang, S. P. (2009, March). Stock Price Forecasting by Hybrid Machine Learning Techniques. In *Proceedings of The İnternational Multiconference of Engineers and Computer Scientists* (Vol. 1, No. 755, P. 60).

V.N. Vapnik, (2000). The Nature Of Statistical Learning Theory, Springer, New York.

Vapnik, V., Cortes, C. (1995). Support Vector Networks. Machine Learning. 20 (3), 273–297.

Vapnik, V.N. (1999). An Overview of Statistical Learning Theory, IEEE Transactions on Neural Networks 10 988–999.

Vijh, M., Chandola, D., Tikkiwal, V. A., & Kumar, A. (2020). Stock Closing Price Prediction Using Machine Learning Techniques. *Procedia Computer Science*, *167*, 599-606.

Wang,L., Zhou, X., Zhu, X., Dong, Z., Guo, W. (2016). Estimation of Biomass In Wheat Using Random Forest Regression Algorithm and Remote Sensing Data, *The Crop Journal*, 4(3),212-219.

Wei, J., Chu, X., Sun, X. Y., Xu, K., Deng, H. X., Chen, J., ... & Lei, M. (2019). Machine learning in materials science. *InfoMat*, *1*(3), 338-358.

Wuest, T., Weimer, D., Irgens, C., & Thoben, K. D. (2016). Machine learning in manufacturing: advantages, challenges, and applications. *Production & Manufacturing Research*, *4*(1), 23-45.

Yakut, Y., Yakut, E., & Yavuz, S. (2014). Yapay Sinir Ağları ve Destek Vektör Makineleri Yöntemleriyle Borsa Endeksi Tahmini. *Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, *19*(1), 139-157.

Yan, D., Zhou, Q., Wang, J., & Zhang, N. (2017). Bayesian Regularisation Neural Network Based on Artificial İntelligence Optimisation. International Journal of Production Research, 55(8), 2266-2287.

Yetis, Y., Kaplan, H., & Jamshidi, M. (2014, August). Stock Market Prediction by Using Artificial Neural Network. In *2014 World Automation Congress (WAC)* (pp. 718-722). IEEE.

Yu HK (2005) Weighted Fuzzy Time-Series Models for TAIEX Forecasting. Physica A 34, 609–624.

Yu L, Wang S, Lai KK. Forecasting Crude Oil Price with an EMD-Based Neural Network Ensemble Learning Paradigm. Energy Econ 2008;30(5):2623e35.

Yu, L., Chen, H., Wang, S., & Lai, K. K. (2009). Evolving Least Squares Support Vector Machines For Stock Market Trend Mining. IEEE Transactions On Evolutionary Computation, 38, 802–815. Doi:10.1109/TEVC.2008.928176.

Yu, S., Tian, L., Liu, Y., & Guo, Y. (2021, July). LSTM-XGBoost Application of The Model To The Prediction of Stock Price. In *International Conference on Artificial Intelligence and Security* (pp. 86-98). Springer, Cham.

Zander, S., Nguyen, T., & Armitage, G. (2005, November). Automated traffic classification and application identification using machine learning. In *The IEEE Conference on Local Computer Networks 30th Anniversary (LCN'05) l* (pp. 250-257). IEEE.

Zhang, N., Lin, A., & Shang, P. (2017). Multidimensional K-Nearest Neighbor Model Based on EEMD for Financial Time Series Forecasting. *Physica A: Statistical Mechanics and its Applications*, *477*, 161-173.

Extended Abstract

Financial time series analysis has always been a topic of interest to researchers and investors. While investors performed analyses using various financial and technical indicators in order to maximize their profits, researchers studied time series to develop the best forecasting/prediction methods.

Among the most frequently studied topics in financial forecasting are prediction of index values and prediction of stock prices. However, stock markets are volatile and unstable due to issues such as political conditions, economic outlook and investor expectations/desires. Therefore, index forecasting is a difficult subject (Nava et al, 2016; Cao and Wang, 2018; Nava et al, 2018; Cao et al, 2019).

In the early financial time series forecasting studies, researchers generally used linear models. The leading of these models are AR, MA and ARIMA models, which are linear regression analysis and time series models. However, using linear models is not an appropriate approach to the volatile nature of the data. Therefore, econometric methods that take into account volatility such as ARCH and GARCH have become more popular. These models, which have been used for a long time, require some assumptions about the distribution and stationary of the data. Since these assumptions are difficult to provide because of large and complex data, the researchers have applied to Machine Learning (ML) methods, which have better nonlinear modelling capabilities and have superior performance than traditional statistical / econometric methods (Niu et al, 2020; Karasu et al, 2020).

While the use of Machine Learning methods in financial time series has started a new era, the most used methods are gathered around GA, Artificial Neural Network (ANN) and SVM, whose theoretical and applied infrastructure is established and well-known (Kim and Han,

(2000); Cao and Tay (2001); Tay and Cao (2001); Huang et al. (2005). In the first ML applications, these methods were used separately and were compared with traditional statistical methods. However, with the development of ML methods and the increase in their capabilities, new methods have begun to be used. These methods are Multi-Layer Perceptron (MLP) and Generalized Feed Forward Network-Based Architecture Egeli et al. (2003), Back Propagation Neural Network (BPNN) and Case Based Reasoning (CBR) Kim (2003), Random Walk (RW) Henrique et al (2018), Random Forest (RF) Vijh et al. (2020), Lasso, Elastic Net and Ridge Regression, Decision Trees (DT), Extra Tree and Ransac Asfaq et al (2021), MLP and Long-Short Term Memory(LSTM) Demirel et al. (2021), Polynomial Regression and Recurrent Neural Network (RNN) Arslankaya and Toprak (2021), eXtreme Gradient Boosting (XGBoost) Kumar et al (2021) from deep learning methods. Although ML methods are easy to use separately, researchers have turned to hybrid methods that combine the strengths of the methods. In hybrid methods, the prediction, which is first made using an ML method, is used as the input of another ML method. Usually, ML methods are used together in hybrid studies. (Abraham et al (2001), Enke and Thawornwong (2005), Fu et al (2007), Chaudhary and Garg (2008), Tsai and Wang (2009), Yakut et al (2014), Akita et al (2016), Moghaddam et al (2016), Chen et al (2017), Yan et al (2016), Zhang et al (2017), Yu et al (2021). Other researchers preferred to use ML methods and traditional statistical / econometric methods (Pai and Lin (2005), Lu et al (2009), Chen et al (2014), Hsu et al (2016), Kumar et al (2016).

Although there are studies on financial time series forecasting in the literature, it has been determined that there are very few studies on comparing the methods in predicting the real values of different indices and stocks. In this study, SVM method, which is frequently used in the literature and RF and XGBoost, which are among the rare ensemble machine learning methods although they are powerful in financial time series forecasting, are preferred as forecasting models. It has been shown by Patel et al (2015 (b)) that RF outperforms SVM. However, there is not enough evidence on this subject and there is no detailed research. The performance of ensemble learning algorithms has been most extensively studied by Balling et al (2015). On the other hand, aforementioned study focused on the future trends of the stocks. From this point of view, in this study, which focuses on the forecasting of stock market index values and real values of stocks, the forecast performance of SVR, RF and XGBoost methods is presented comparatively on 4 different data sets in order to contribute to the applications of ML methods in financial time series forecasting.

In order to compare the method performances, the composite index values and stocks of developed and developing countries were used. Selected research data belong to New York Stock Exchange (NYSE) which is the world's largest stock market and total value is greater than 26 trillion dollars, and Borsa Istanbul Stock Exchange (BIST), which belongs to the developing Turkish stock market. In addition to the stock market index values in the study, the stock prices of Aselsan (ASELS) and Ereğli Demir ve Çelik (EREGL), which are among the 10 most valuable companies traded in the BIST 100, are predicted. The purpose of using both index and stock prices together is to examine the behaviour of ML methods in less volatile (index) and more volatile (stock) data sets.

In the application section of the study, the selected index and stock prices were analysed using 5-year time series data from the Yahoo Finance (http://finance.yahoo.com) website between the dates 07-07-2017 and 06-07-2022. The first 80% of the obtained datasets were allocated as training and the remaining 20% as test datasets. Also, weekends and holidays on which no trades are made on the stock market are excluded. Considering that time series data are affected by previous values, lagged values up to 7 days ago were included in the analysis as input variables in the study. In ML methods, it is important to determine the appropriate parameters (hyper-parameter tuning) in order to increase the forecasting performance and obtain better forecasting models. In the study, suitable parameters for ML models were

determined by Grid Search method. Since time series data is related to past values, Forward Validation (Sliding Windows, Walk Forward Validation) method was used since there was no Cross Validation on the data set (Hu et al, 1999). By using this method, the next value is predicted and predictions are made for following values by joining the training data set, respectively. Thus, the relationship of the time series data to the previous values is preserved. The range required for optimal parameters was obtained from studies frequently used in the literature. By testing each parameter, the parameters with the lowest RMSE value were determined as the optimal parameters. While comparing the models for each data set, MAE, RMSE and MAPE values were used as performance measures.

According to the application results, considering the periods when the stock markets are more stable, it has been observed that all methods predict the real values more accurately, and they show deviations in the rapid ups and downs experienced in the stock markets.

According to the experimental results, RMSE values of SVR in NYSE, BIST, ASELS and EREGL datasets are 175.3148, 33.9484, 0.6749 and 0.5744, respectively. The RMSE values of RF on the same datasets are 180.0295, 40.2260, 0.8258 and 0.6792, respectively. The RMSE values of XGBoost are 191,5049, 39.4288, 0.9807 and 0.6637, respectively. As a result, it appears that the ML method showing the best results in all data and all performance measures is SVR. The performance of the following RF and XGBoost methods are very close to each other. However, RF gave slightly better results than XGBoost.

As a result, despite some assumptions of traditional methods, the fact that ML methods do not require any assumptions makes them favourable both in terms of application and interpretation. In financial time series estimation, ML methods have been shown to be useful and highly performing methods for both investors and researchers. In this study, the performances of RF and XGBoost methods, which are rarely seen in time series forecasting, are compared with the SVR method, which is frequently preferred and proven to be successful against many methods. In few numbers of studies where this comparison is made, there is no clarity about which method is better. In this respect, the study contributed to the literature. According to the findings of the study, SVR was found to be the best method among ML methods. While RF and XGBoost, which are ensemble ML methods, gave worse results than SVR, they presented a close prediction performance among themselves.