T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

FUTBOLCULARIN PİYASA DEĞERİNDE PERFORMANSIN ÖNEMİ: BİR MAKİNE ÖĞRENİMİ YAKLAŞIMI

Hüseyin Ekrem BAYAR

YÜKSEK LİSANS TEZİ

İstatistik Anabilim Dalı

İstatistik Programı

Danışman

Dr. Öğr. Üyesi Erhan ÇENE

Haziran, 2022

T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

FUTBOLCULARIN PİYASA DEĞERİNDE PERFORMANSIN ÖNEMİ: BİR MAKİNE ÖĞRENİMİ YAKLAŞIMI

Hüseyin Ekrem BAYAR tarafından hazırlanan tez çalışması 14.06.2022 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalı, İstatistik Programı **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Dr. Öğr. Üyesi Erhan ÇENE
Yıldız Teknik Üniversitesi
Danışman

Jüri Üyeleri

Dr. Öğr. Üyesi Erhan ÇENE, Danışman
Yıldız Teknik Üniversitesi _____

Dr. Öğr. Üyesi Doğan YILDIZ, Üye
Yıldız Teknik Üniversitesi _____

Dr. Öğr. Üyesi Hasan Halit TALİ, Üye
Haliç Üniversitesi _____

Danışmanım Dr. Öğr. Üyesi Erhan ÇENE sorumluluğunda tarafımca hazırlanan Futbolcuların Piyasa Değerinde Performansın Önemi: Bir Makine Öğrenimi Yaklaşımı başlıklı çalışmada veri toplama ve veri kullanımında gerekli yasal izinleri aldığımı, diğer kaynaklardan aldığım bilgileri ana metin ve referanslarda eksiksiz gösterdiğimi, araştırma verilerine ve sonuçlarına ilişkin çarpıtma ve/veya sahtecilik yapmadığımı, çalışmam süresince bilimsel araştırma ve etik ilkelerine uygun davrandığımı beyan ederim. Beyanımın aksinin ispatı halinde her türlü yasal sonucu kabul ederim.

Hüseyin Ekrem BAYAR

İmza

Sevgili Aileme

TEŞEKKÜR

Akademik ve sektörel hayatımda başarılı olmama yardımcı olan, başta danışman hocam Dr. Öğr. Üyesi Erhan ÇENE'ye ve bölümdeki tüm hocalarıma, sevgili aileme, Dr. Fahrettin Fırat GÖNEN'e ve Vahit KESKİN'e teşekkürlerimi sunuyorum.

Hüseyin Ekrem BAYAR

İÇİNDEKİLER

SİMGE LİSTESİ vii

KISALTMA LİSTESİ viii

ŞEKİL LİSTESİ ix

TABLO LİSTESİ x

ÖZET xi

ABSTRACT xiiiii

1 GİRİŞ 1

1.1 Literatür Özeti 1

1.2 Tezin Amacı 4

1.3 Hipotez 4

2 FUTBOL'DA PİYASA DEĞERİ 6

3 MAKİNE ÖĞRENMESİ 12

3.1 Temel Bileşen Analizi 14

3.2 Regresyon ve Araştırmada Kullanılan Regresyon Modelleri 15

3.2.1 LGBM Algoritması 16

3.2.2 XGBoost Algoritması 17

3.2.3 Bagging Algoritması 17

3.2.4 Random Forest Algoritması 18

3.3 Başarı Ölçüm Metrikleri 18

3.4 Model Doğrulama Yöntemi 19

3.4.1 Holdout (Sınama Seti Yöntemi) 19

3.5 SHAP Analizi 20

3.6 Python Programlama Dili 20

4 VERİ SETİ 22

4.1 Veri Kaynakları: Transfermarkt, Football Reference, Statsbomb 23

4.2 Veri Toplama Metodu: Web Kazıma 23

4.3 Toplanan Veri 24

5 UYGULAMA 27

5.1 Veri Ön İşleme 27

5.2 Temel Bileşen Analizi 29

5.3 Modelleme 30

5.4 Değişken Önem Grafiği ve SHAP Analizi 31
5.5 Bonservis Bedelleri ile Tahmin Edilen Piyasa Değerlerinin Kıyaslanması 33
6 SONUÇ VE ÖNERİLER **37**
KAYNAKÇA **40**
EK **43**
TEZDEN ÜRETİLMİŞ YAYINLAR **50**

SİMGE LİSTESİ

$\forall$	Her
λ	Lambda
$\tilde{X}$	Ortalaması Çıkartılmış Veri Matrisi
β_0	Regresyon Sabiti
$\beta_1, \beta_2, \ldots, \beta_n$	Regresyon Katsayıları
$\sum$	Toplam

KISALTMA LİSTESİ

AVO	Açıklanabilir Varyans Oranı
FBRef	Football Reference
LGBM	Light Gradient Boosting
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
RF	Random Forest
RMSE	Root Mean Squared Error
SHAP	Shapley Additive Explanation
TBA	Temel Bileşen Analizi
TM	Transfermarkt
XGBoost	XG Boosting Machine

ŞEKİL LİSTESİ

Şekil 2.1 Beş büyük lig (Almanya, Fransa, İngiltere, İspanya, İtalya) takımlarının toplam transfer yatırımları....7
Şekil 2.2 Piyasa değerini etkileyen faktörler....9
Şekil 2.3 Burak Yılmaz Transfermarkt transfer geçmişi....10
Şekil 3.1 Bilgi katlama eğrisi....12
Şekil 3.2 Makine öğrenmesi, yapay zeka, veri bilimi ve derin öğrenme kelimelerinin Google Trends ile zaman içinde gösterilen ilgi grafiği....13
Şekil 3.3 LGBM algoritmasının yaprak odaklı büyüme stratejisi....16
Şekil 3.4 XGBoost algoritmasının seviye odaklı büyüme stratejisi....17
Şekil 3.5 Örnek karar ağacı....17
Şekil 3.6 Holdout yöntemi....19
Şekil 3.7 SHAP ile örnek model tahmini açıklanabilirliği....20
Şekil 4.1 Mesut Özil TM piyasa değeri akışı....24
Şekil 4.2 Mesut Özil FBRef performans istatistikleri örneği....24
Şekil 5.1 Performans istatistikleri veri görselleştirme....28
Şekil 5.2 Kümülatif açıklanabilir varyans oranı ve bileşen sayısı grafiği....29
Şekil 5.3 LGBM değişken önem grafikleri....31
Şekil 5.4 SHAP değişken önem grafiği....32
Şekil 5.5 Cristiano Ronaldo'nun SHAP ile değerlendirilmesi....33

TABLO LİSTESİ

Tablo 3.1 TIOBE indeksine göre en popüler 20 programlama dili.....21

Tablo 4.1 Web kazıma ile toplanan verilerin özet listesi 25

Tablo 5.1 Modelleme kullanılan bileşen sayısı ve kümülatif açıklanabilir varyans oranı.....30

Tablo 5.2 Referans model karşılaştırmaları.....30

Tablo 5.3 3.LGBM modelinin performansı 30

Tablo 5.4 Model farklılıklarının nedenleri 31

Tablo 5.5 Pozisyon sınıflarına göre ortalama mutlak hata 31

Tablo 5.6 TM piyasa değeri, model tahmini, bonservis bedeli özet istatistikleri .33

Tablo 5.7 TM piyasa değeri ve model tahmininin bonservis bedeli ile MAPE üzerinden kıyası 34

Tablo 5.8 Shapiro normallik testi.....34

Tablo 5.9 Mann Whitney U testi 35

Tablo 5.10 Spearman korelasyon testi 35

Tablo 5.11 2020-2021 sezonunda en yüksek bonservis değeriyle transfer olan ilk 10 futbolcu ve piyasa değerleri 36

Tablo A.1 Transfermark ve Football Reference web sitelerinden toplanan çeşitli veriler 43

Tablo A.2 Football Reference performans istatistikleri.....44

ÖZET

Futbolcuların Piyasa Değerinde Performansın Önemi: Bir Makine Öğrenimi Yaklaşımı

Hüseyin Ekrem BAYAR

İstatistik Anabilim Dalı

Yüksek Lisans Tezi

Danışman: Dr Öğr. Üyesi Erhan ÇENE

Modern spor dallarının icadından beri spor severler, sporcular ve yöneticiler sürekli olarak veri ile yaşamaktadırlar. Bilgisayar teknolojilerinin gelişimi ile birlikte verinin üretilmesi, toplanması, işlenmesi ve bilgiye dönüşümünün kolaylaşması, pek çok sektörde büyük etki yaratmıştır. Günümüzde veri bilimi uygulamaları her alanda olduğu gibi spor dallarında da hızla yaygınlaşmaktadır. Futbol, basketbol, beyzbol, hokey gibi takım sporlarında rekabet üstünlüğü sağlamak amacıyla; oyuncu keşfi ve gelişimi, taktik geliştirme, transfer süreçlerinin yönetilmesi vb. konularda kulüpler veri bilimi uygulamalarına sıklıkla başvurmaktadırlar. Bu çalışmada amaçlanan, makine öğrenmesi algoritmaları kullanılarak futbolcuların piyasa değerlerini tahmin etmek ve bir öneri sistemi olarak kullanılıp kullanılmayacağını araştırmaktır. Araştırmada kullanılan veri seti 2017-2021 sezonları içerisinde beş büyük Avrupa liginde (Almanya, Fransa, İngiltere, İspanya, İtalya) oynamış 3948 futbolcudan oluşmaktadır. Araştırma sonucunda piyasa değerinin bonservis bedelini açıklamada ve transfer süreçlerini yönetmede etkili bir faktör olabileceği düşünülmektedir. Tüm çalışmalar Python programla dili yapılmıştır. Temel Bileşen Analizi ile veri seti içerisinde çok fazla sayıda bulunan performans istatistikleri daha az bileşenle açıklanmıştır. Modelleme aşamasında LGBM, XGBoost, RF ve Bagging algoritmaları kullanılmış ve model başarıları değerlendirdikten

sonra referans model olarak LGBM modelinin araştırma probleminde en iyi performansı gösterdiği tespit edilmiştir. Referans modelin performansı, öznitelik mühendisliği ve parametre ayarlamaları ile iyileştirilmiştir. Model sonuçları, SHAP analizi ve hipotez testleri kullanılarak incelenmiştir. Bu yöntemler, model başarısı ve piyasa değeri ile bonservis bedeli arasındaki ilişki hakkında detaylı bilgi sunmaktadır.

Anahtar Kelimeler: Spor analitiği, veri bilimi, makine öğrenmesi, futbol, piyasa değeri.

ABSTRACT

The Importance of Performance of Football Players on Market Value: A Machine Learning Approach

Hüseyin Ekrem BAYAR

Department of Statistics

Master of Science Thesis

Supervisor: Assoc. Prof. Dr. Erhan ÇENE

Since the invention of modern sports, sports lovers, athletes and managers have constantly been living with data. With the development of computer technologies, the production, collection, processing and conversion of data into information has become easier and has had a great impact on many sectors. Today, data science applications are spreading rapidly in sports branches as in every field. In order to provide competitive advantage in team sports such as football, basketball, baseball and hockey; player discovery and development, tactics development, management of transfer processes, etc. clubs frequently apply to data science applications. The aim of this study is to predict the market values of football players using machine learning algorithms and to investigate whether they can be used as a recommendation system. The data set used in the research consists of 3948 football players who played in five major European leagues (Germany, France, England, Spain, Italy) during the 2017-2021 seasons. As a result of the research, it is thought that the market value can be an effective factor in explaining the transfer fee and managing the transfer processes. All work has been done in Python programming language. Performance statistics, which are too many in the data set, are explained with fewer components by using Principal Component Analysis. LGBM, XGBoost, RF and Bagging algorithms were used in the modeling phase and after evaluating

the model performance, it was determined that the LGBM model showed the best performance in the research problem as a reference model. The performance of the reference model has been improved by feature engineering and parameter tuning. Model results were analyzed using SHAP analysis and hypothesis testing. These methods provide detailed information about the model performance and the relationship between market value and transfer fees.

Keywords: Sports analytics, data science, machine learning, football, market value.

1
GİRİŞ

1.1 Literatür Özeti

Günümüzün popüler alanlarından olan veri bilimi, gerçek olayları verilerle anlamak ve analiz etmek için istatistikleri, veri analizini, makine öğrenimini ve ilgili yöntemlerini birleştirmek için kullanılan bir kavramdır [1]. Veri biliminin istatistik bilimi tarihiyle çok yakın bir ilişkisi bulunmaktadır [2]. Ünlü istatistikçi John Tukey, istatistik biliminde işin matematiğine fazla odaklanıldığını ancak verilerin analizine yeteri kadar önem gösterilmediğini vurgulamıştır [3]. Bilgisayar teknolojilerinin hızla gelişmesiyle birlikte veri bilimi kavramı günümüzde pek çok alanda yaygın bir şekilde kullanılmaktadır [4].

Veri biliminin alt başlıklarından olan makine öğrenmesi terimi ilk olarak 1959 yılında yapay zeka ve bilgisayar oyunları alanlarında önde gelen isimlerden biri olan Arthur Samuel tarafından kullanılmıştır. Makine öğrenmesi veri türlerine dayalı öğrenmeyi sağlayan algoritmaların geliştirme ve tasarım süreçlerini içeren bir alandır [5]. 21. yüzyılın ilk on yılının başlangıcı, makine öğrenimi tarihinde bir dönüm noktasıdır. Bu dönüm noktasıyla birlikte büyük veri, paralel hesaplama ve bellek maliyeti ve yeni algoritmaların geliştirilmesi olarak üç büyük eğilim ortaya çıkmıştır [6]. Bu eğilimler sayesinde günümüzde makine öğrenimi uygulamaları her alanda, çok büyük verilerle ve düşük maliyetlerle yapılabilmektedir.

Spor analitiği kullanımında en yaygın bilinen örnek beyzbol spor dalındadır. Oyuncu alımında veri analitiğini ciddi bir şekilde kullanan ilk spor dalı beyzbol olmuştur [7]. 1990'ların sonunda, Oakland Athletics Genel Müdürü Billy Beane, oyuncu keşifleri ve takım kadrosuyla ilgili kararlar için istatistiksel verileri kullanmaya başladı [8]. Bu olay spor dünyasında büyük merak uyandırmakla birlikte sadece spor dünyasıyla sınırlı kalmadı ve çekilen Moneyball filmi ile birlikte veri analitiğinin spor dallarında kullanılmasında büyük farkındalık oluşturdu [9]. Oyuncu istatistiklerinden elde edilen öngörüler ile düşük değerli ancak yetenekli

oyuncuları ve zirve noktasını aşmış olan aşırı değerli oyuncuları belirlemesine yardımcı oldu. Takip eden yirmi yılda, Athletics'in oyuncu alımına yönelik yenilikçi yaklaşımı, çoğu daha sonra Beane'in fikirlerini benimseyen tüm beyzbol takımlarının en düşük bütçelerinden birine sahip olmalarına rağmen, takımın kabaca her iki sezonda bir playofflara ulaşmasına yardımcı oldu [10]. Moneyball örneği ile birlikte veri analitiği özellikle takım sporlarında etkin bir şekilde kullanılmaya başlandı. Literatürde beyzbol dışında basketbol, futbol, Amerikan futbolu, hokey, voleybol gibi takım sporlarında spor analitiği uygulamaları mevcuttur.

Futbol literatüründe transfer süreçlerinde kullanılabilecek piyasa değeri tahmini ve bonservis bedelleri ile kıyaslanmasını araştıran çok sayıda çalışma bulunmaktadır. Temelde çalışmalar, kullanılan veri kaynağı ve algoritmalar açısından farklılık göstermektedir. Ücretsiz, zengin içerikli ve büyük örnekler içeren spor verilerine erişim çok mümkün olmadığından araştırmacıların farklı veri kaynaklarından veri toplamak durumundadır. Ayrıca ekonomik piyasalardaki rekabetle birlikte futbol transfer piyasasındaki rekabet ve futbol ücretlerini belirleyen faktörler hakkında da kesin görüşler bulunmamaktadır [11]. Bu da araştırmacıların, veri toplama süreçlerinde farklı kaynaklara yönelmesine ve farklı özelliklerin piyasa değeri üzerindeki katkısını görme şansına olanak sağlamıştır. Bu tezde ise piyasa değerlerini açıklamak için ağırlıklı olarak performans istatistikleri toplanmıştır.

Müller, Simons ve Weinmann (2017) beş büyük Avrupa liginden 4217 oyuncu ve 6 sezonluk bir veri seti ile çok değişkenli lineer regresyon modeli üzerine çalışmışlardır. Çalışmanın piyasa değeri tahminlemede başarılı sonuçlar ürettiği ve futbol yöneticileri, gözlemcileri ve menajerler için önemli etkilerinin olduğu düşünülmektedir [12].

Aydemir, Temizel ve Temizel (2022), literatürdeki mevcut yaklaşımların daha çok açıklayıcı modellere odaklandığını ve modellerin gelecek verileri tahmin etmede kullanılmadığını eleştirmiştir. Futbolcuların saha içi performans metriklerini, internetten alınan oyuncu popülerliği metriklerini ve gerçek transfer değerlerini birleştiren bir yöntem önermişlerdir. Önerilen yaklaşımın, literatürde yaygın olarak kullanılan modellerden daha iyi performans gösterdiğini vurgulamışlardır [13].

Ezzeddine (2020) çalışmasında, oyuncunun finansal değerini etkileyen en önemli faktörleri tahmin etmek için bir hedonik fiyatlandırma fonksiyonu ve üç makine öğrenme algoritması kullanılmıştır. Transferlerin, sürpriz maç sonuçlarının ve COVID-19'un piyasa değerine etkisini belirlemek için modelleme yapılmıştır [14].

Behravan ve Razavi (2021) çalışmasında FIFA 20 oyunu verileri kullanılarak hem denetimli hem de denetimsiz makine öğrenme algoritmaları kullanılarak piyasa değerini tahmin etmek için yeni bir yöntem önerilmiştir. Önerilen yöntem oyuncuların değerlerini %74 doğrulukla tahmin edebilmektedir [15].

Richau, Follert, Frenger ve Emrich (2010) çalışmasında, özellikle performans göstergelerine odaklanarak futbolcuların TM piyasa değerlerine odaklanarak futbolcuların farklı özelliklerinin göreli etkisini değerlendirmektedir. İngiltere Premier Ligi 2012-2017 sezonları içerisindeki 1897 oyuncunun piyasa değeri tahminlemesi için artırılmış regresyon ağaçları kullanmışlardır. Bireysel performans metriklerinin, bir oyuncunun piyasa değeri üzerindeki etkileri ve farklılıkları incelenmiştir. Farklılıklar hem değişken bazlı hem de pozisyon bazlı ele alınmış ve performans metriklerine geniş bir perspektif sunmuştur. Oyuncular hakkında ekonomik bir değerlendirme ve liglerin kalitesi üzerine inceleme yapmıştır [16].

Yiğit, Samak ve Kaya (2019) yaptıkları çalışmada futbol kulüplerinin transfer kararlarını desteklemek için makine öğrenmesi tekniklerini kullanarak futbolcu değer belirleme modeli oluşturmuştur. Önerilen modeller, esas olarak Football Manager video oyununda sağlanan bireysel oyuncuların içsel özelliklerine dayandırılmıştır. Bu kapsamda Avrupa ve Güney Amerika'dan 11 farklı büyük ligde aktif olan 5316 oyuncunun bireysel istatistiklerine dayalı olarak, Ridge ve Lasso regresyon modelleri, Random Forest ve XGBoost algoritmaları kullanılarak bir topluluk modeli önerilmiştir [17].

Geurkink, Boone, Verstockt ve Bourgois (2021) çalışmada, Belçika liginde oynanan 576 maç ile bir model eğiterek maç kazanma ve kaybetmeyi belirleyen en önemli değişkenleri tespit etmeyi amaçlamıştır. Bu nedenle değişkenler arasında bulunan çoklu doğrusal bağlantıyı azaltma ve boyut indirgeme çalışmaları yapılmıştır. SHAP analizi ile değişkenlerin modele katkıları incelenmiştir. Model, 576

maçtan 516'sını doğru bir şekilde sınıflandırarak %89,6 ± %3,1 isabet oranına sahiptir [18].

1.2 Tezin Amacı

Veri analitiğinin her geçen gün öneminin artması ve uygulandığı her alanda fark yaratmasıyla birlikte spor dallarında da aktif bir şekilde kullanılmaya başlanmıştır. Spor analitiği bir takımın veya bir sporcunun rakiplerine karşı rekabet üstünlüğü sağlamasına yardımcı olur. Transfer süreçleri de kulüplerin sportif başarısı ve finansal durumları açısından iyi bir şekilde yönetilmelidir. Takıma transfer edilmek istenen sporcu, kulübün rekabet üstünlüğünü arttıracak, gelecek vaad eden ve başka bir takıma satılmak istendiğinde de kulübe minimum zararı verecek bir sporcu olmalıdır. Futbolda da bu süreç farklı değildir. Bu yüzden transfer süreçlerinde ve öncesinde futbolcuların piyasa değerleri bir karar alma mekanizmasına dönüşür. Piyasa değerini etkileyen sayısız parametre bulunduğundan, bir futbolcunun piyasa değerinin belirlenmesi, karmaşıklığı yüksek ve belirsiz bir problemdir.

Bu çalışma, veri bilimi ve makine öğrenmesi yöntemlerinin futbolcuların piyasa değerindeki bu belirsizliği ortadan kaldırmada yardımcı olacağı öngörüsüyle tasarlanmıştır. Çok sayıda ve çeşitli parametreler ile futbolcuların piyasa değerleri, makine öğrenmesi algoritmaları kullanılarak tahmin edilmiş ve kıyaslanmıştır. Elde edilen tahmin sonuçları futbolcu bazında incelenmiş ve piyasa değeri tahminlerinin bonservis bedelleriyle olan ilişkisi araştırılmıştır.

1.3 Hipotez

Spor kulüplerinde transfer süreçlerinin doğru bir şekilde yönetilebilmesi için piyasa değeri tahmin modellerinin bir öneri sistemi olarak kullanılabileceği düşünülmektedir. Makine öğrenmesi algoritmaları ile kulüpler rakiplerine karşı rekabet üstünlüklerini arttıracak ve transfer ettikleri futbolcuların gelecek satışlardan maksimum kar veya minimum zarar elde etmelerine yardımcı olacaktır. İlk olarak futbolcuların çeşitli istatistikleri kullanılarak piyasa değerinin tahmin edilmesi ve modelleme sonucunda modele katkı sağlayan parametrelerin tespitinin edilmesiyle performans istatistiklerinin piyasa değeri tahmininde etkili olup olmadığı

araştırmanın öncelikli hipotezidir. Ek olarak piyasa değerinin bonservis bedeli arasında bir ilişkinin olup olmadığının incelenmesi, potansiyeli yüksek gelecek vaad eden futbolcuların keşfedilmesi ve piyasa değerinin normalden fazla olan ve balon olarak nitelendirilen futbolcuların tespit edilmesinde rahatlıkla kullanılabilecek bir makine öğrenmesi algoritması tasarlamak bu araştırmanın diğer hipotezleridir.

2
FUTBOL'DA PİYASA DEĞERİ

Futbol, dünyadaki en popüler sporlardan bir tanesidir. Futbol biri kaleci olmak üzere on birer oyuncudan oluşan iki takımla ve bir topla oynanan bir takım sporudur. Futbolun oynandığı alan, dikdörtgen şeklindeki bir sahadır. Dikdörtgenin kısa kenarlarında karşılıklı iki kale bulunmaktadır ve amaç, rakip takımın kalesine topu sokarak gol atmaktır. Atılan golün geçerli olması için oyuncular, elleri ve kolları hariç temelde ayaklarını ve vücutlarının diğer uzuvlarını kullanmaları gerekmektedir.

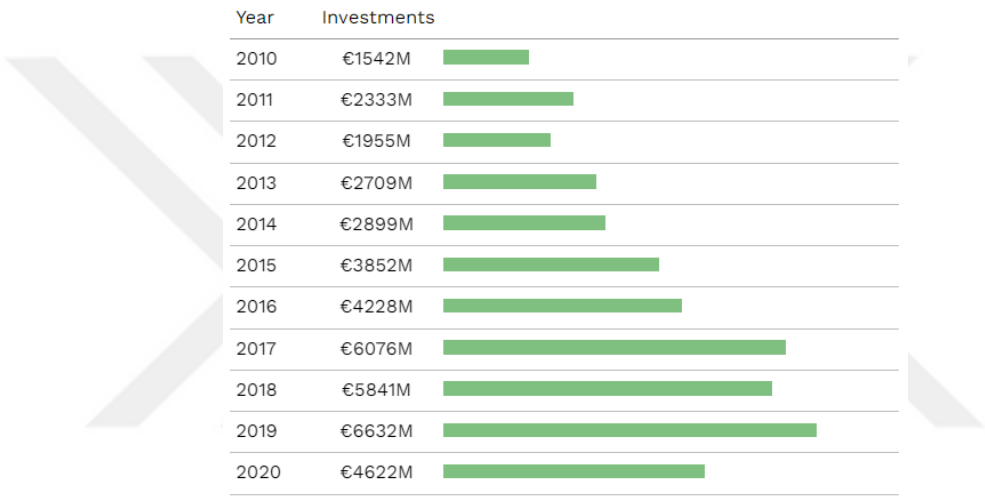
Tarih boyunca dünyanın çeşitli bölgelerinde futbola benzer oyunlar oynanmıştır. FIFA, futbola benzeyen ve bilimsel kanıtlara sahip olan ilk oyunu, MÖ 300-200 yıllarında Çin'de askerî eğitim amacıyla oynanan cuju olarak göstermektedir [19, 20]. Modern futbolun ortaya çıkışı ve yayılışı 19. yüzyıl ortalarında, İngiltere'deki özel okullarda farklı kurallarla oynanan futbol biçimlerine dayanmaktadır [21].

Günümüz futbolunda bir futbol takımı kurabilmek için izlenebilecek iki yol, oyuncuları alt yapıdan yetiştirmek ya da oyuncu bulmak yani transfer yapmaktır. Futbolda kış ve yaz dönemi olmak üzere iki transfer dönemi bulunmaktadır. Takımlar, transfer dönemi içerisinde başka takımlardan oyuncu transfer edebilirler ve aynı şekilde oyuncuları diğer takımlara satabilirler. Spor dallarında ana amaç, rakip takım veya sporculara karşı rekabet üstünlüğü sağlamaktır. Bu yüzden transfer süreçlerinde takıma dahil edilmek istenen futbolcu, kulübün maddi durumunu zedelemeden, maliyetini sözleşme sürecince çıkartabilecek, saha içi performansıyla takıma katkı sağlayabilecek ve sözleşmesi bitmeden maliyetinden daha yüksek bir fiyatla veya en kötü şartlar altında asgari zararla veya zarar elde etmeden elden çıkarılabilecek ve takımda rekabeti en üst seviyeye çıkarabilecek ve kulübün marka değerini yükseltecek bir futbolcu olmalıdır [22].

Transfer süreçleri, spor kulüpleri için çok önemli bir iştir. Kulüplerin en büyük gelir ve gider kaynağı olmakla birlikte en büyük endişe kaynağıdır. Transfer mali-

yetleri, yanlış oyuncu seçimi, pazarlık süreçlerinin doğru yönetilememesi, sözleşmeler gibi konular kulüpleri hem finansal açıdan hem de sportif başarı açısından zarara uğratmaktadır. Doğru kadro mühendisliği ile kulübün maddi, manevi ve sportif çıkarları kısa, orta ve uzun vadede korunabilmelidir. İyi bir kadro mühendisliği yapmak için de kulüpler, gözlemleme (scouting) yani futbolcu izleme ve futbolcu keşfi yapmaktadırlar.

COVID-19 virüsünün yayılması ve dünyadaki tüm işleyişi değiştirmesiyle birlikte virüs, profesyonel futbol ekonomisi, futbolcuların transfer piyasası, kadro mühendisliği ve gözlemleme üzerinde oldukça etkili olmuştur.



Şekil 2.1 Beş büyük lig (Almanya, Fransa, İngiltere, İspanya, İtalya) takımlarının toplam transfer yatırımları [23]

Şekil 2.1'de görüldüğü gibi COVID pandemisi transfer pazarındaki yatırım seviyesinde önemli bir düşüş getirdi. Ocak 2010'dan bu yana 5 büyük Avrupa liginin transfer yatırımları incelendiğinde sürekli bir artış ve futbol ekonomisinde büyüme gözlenirken COVID ile sürekli artış durmuş ve futbol ekonomisi 2016'daki haline geri dönmüştür. 2020'de ölçülen değer, rekor yılı olan 2019'da gözlemlenen değerden %30 daha düşüktür. Bu dönemde futbol kulüplerinin transfer stratejileri de değişerek daha çok kiralık ve serbest oyunculara yönelmişlerdir.

Futbol ekonomisindeki ve transfer piyasasındaki değişimler dünyadaki tüm kulüpleri ilgilendirmektedir ve tüm kulüpler bu değişimlere göre karar alma mekanizmaları oluştururlar. Ayrıca, takımlarda gözlemleme ve kadro mühendisliği doğru bir şekilde yönetilemediğinde transfer maliyetleri kısa ve uzun vadede kulüpleri

olumsuz etkileyecektir. Bu yüzden, futbolcuların piyasa değeri, performansı sürekli olarak takip edilerek doğru kararlar verilmedir. Transfer edilen oyuncu takımda maliyetini çıkarabilmeli ve ideal senaryoda takımlar da oyuncuyu satarken maliyetinden daha yükseğe satabilmelidir. Bu sebeple, kulüpler ayrıca gelecekte mevcut piyasa değerini arttıracağını düşündükleri oyuncuları transfer süreçlerinde tercih ederler.

Piyasa değeri, sporcuların transfer ücretlerini tahmin etmede, pazar içerisinde sporcunun konumunu belirlemede ve transfer görüşmeleri için referans noktası olarak kullanılmaktadır. Özetle piyasa değerleri, futbolcular için ödenmesi en muhtemel transfer ücretlerinin tahminleridir. Kavramsal farklılıklar olmasına rağmen, piyasa değerleri ve transfer ücretleri birbirlerinin yerine kullanılabilir [24].

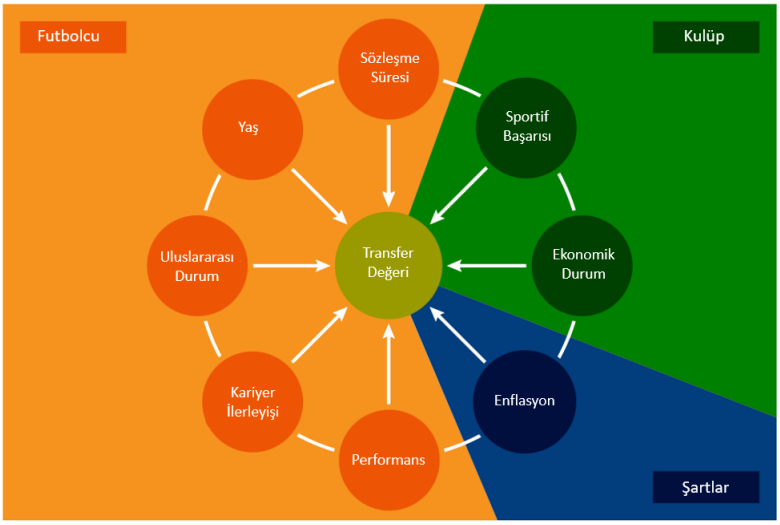
Sporcuların piyasa değerleri geleneksel olarak kulüplerin kendileri, gazeteciler, spor uzmanları gibi sporculara değer biçebilecek kişiler tarafından tahmin edilmektedir. Geleneksel yönteme ek olarak, taraftarların da futbol piyasası ve piyasa değerleri ile ilgilenmeye başlaması, piyasa değeri tahminleri üreten web sitelerinin de spor dünyasına girmesini sağladı. Böylelikle, kitle kaynak kullanımı yardımıyla futbolcuların piyasa değerleri tahmin edilmeye başlandı.

Transfermarkt, futbol transfer pazarında en bilinen web sitesidir [25]. Site içerisinde temel düzeyde istatistikler, haberler, maç sonuçları, transfer söylentileri, piyasa değeri tahminleri, bireysel ve takım düzeyinde bilgiler gibi çeşitli pek çok veri yer almaktadır. Kullanıcılar Transfermark'a üye olduktan sonra, futbolcular hakkında piyasa değerlerini, tartışmaları ve gündemi takip edebilmektedirler. Ek olarak, futbolcuların performansına, mevcut piyasa değerine, kişisel reklam gücüne, takımdaki konumuna vs. bağlı olarak, kullanıcılar futbolcuların değerleriyle ilgili kendi önerilerini paylaşabilir hatta diğer kullanıcılar ile de etkileşime geçebilirler. Bir futbolcu hakkında yeterli veri toplandıktan sonra Transfermarkt, bireysel tahminleri toplayarak futbolcunun mevcut piyasa değerini hesaplar.

James Surowiecki 2004 yılında Kalabalıkların Bilgeliği (The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations) kitabını yayınladıktan sonra kitle kaynak kullanımı popüler hale gelmiştir [26]. Kalabalıkların bilgeliği uygulamaları

tahmin piyasaları, Delphi yöntemleri ve geleneksel kamuoyu yoklamaları olarak üç genel kategoride özetlenebilir [27]. Kalabalıkların bilgeliği kavramı, bir grubun bilgi ve tecrübesi bir bireyinkinden daha iyi veya daha doğru olduğu durumları tanımlamaktadır. Spor dünyasında Transfermarkt gibi kitle kaynak kullanımı ile çalışan web sitelerinin arkasında "Kalabalıkların Bilgeliği" kavramı bulunmaktadır.

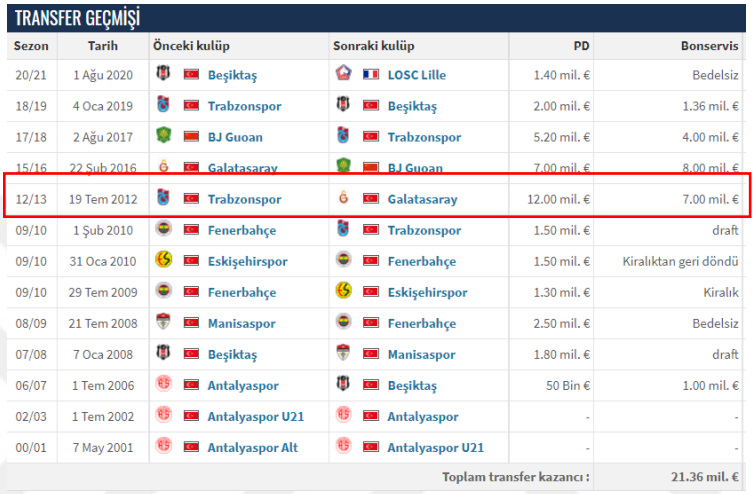
Transfermarkt, kitle kaynak kullanımı ile futbolcuların piyasa değerlerini hesaplarken; geleneksel hesaplama yöntemleri ise bir futbolcunun piyasa değerini çeşitli parametreler kullanarak hesaplar. Transfermarkt, kitle kaynak kullanımı ile futbolcuların piyasa değerlerini hesaplarken; geleneksel hesaplama yöntemleri ise bir futbolcunun piyasa değerini çeşitli parametreler kullanarak hesaplar. Literatürde piyasa değerini etkileyen faktörler araştırıldığında temelde 3 faktör ön plana çıkmaktadır. Bu 3 faktörün altında da farklı parametreler bulunabilir. Karşılaşılan ilk faktör, futbolcu hakkındaki bilgilerdir. Futbolcunun karakteristik ve demografik özellikleri, performans istatistikleri, sözleşme detayları, uluslararası başarıları, popülerliği, saha içi ve saha dışı davranışları, kariyer geçmişi ile birlikte bunlar dışında da pek çok özellik bir futbolcunun piyasa değeri hakkında fikir vermektedir.



Şekil 2.2 Piyasa değerini etkileyen faktörler [28]

Bir futbolcunun piyasa değerini belirlemede ikinci faktör, futbolcunun sözleşmeli olduğu spor kulübüdür. Futbol kulüplerinin sportif başarısı arttıkça kulübün futbolcuları daha değerli hale gelir. Sportif başarı ile birlikte aynı zamanda kulübün

finansal ve ekonomik durumu transfer süreçlerinde çok kritik rol almaktadır. Futbolcuların piyasa değerinin altında başka kulüplere transfer olması sık karşılaşılan bir durumdur. Bunun nedenlerinden biri kulübün finansal durumunun kötü olmasından dolayı daha düşük teklifleri kabul etmelerinden kaynaklanmaktadır. Örneğin, Burak Yılmaz, Trabzonspor'dan Galatasaray'a piyasa değerinden daha düşük bonservisle transfer olurken Trabzonspor'un finansal açıdan kötü bir durumda olmasaydı bu ücreti daha üst seviyelere taşıyabilirdi [29].



TRANSFER GEÇMİŞİ

Sezon	Tarih	Önceki kulüp	Sonraki kulüp	PD	Bonservis
20/21	1 Ağu 2020	Beşiktaş	LOSC Lille	1.40 mil. €	Bedelsiz
18/19	4 Oca 2019	Trabzonspor	Beşiktaş	2.00 mil. €	1.36 mil. €
17/18	2 Ağu 2017	BJ Guoan	Trabzonspor	5.20 mil. €	4.00 mil. €
15/16	22 Şub 2016	Galatasaray	BJ Guoan	7.00 mil. €	8.00 mil. €
12/13	19 Tem 2012	Trabzonspor	Galatasaray	12.00 mil. €	7.00 mil. €
09/10	1 Şub 2010	Fenerbahçe	Trabzonspor	1.50 mil. €	draft
09/10	31 Oca 2010	Eskişehirspor	Fenerbahçe	1.50 mil. €	Kiralıktan geri döndü
09/10	29 Tem 2009	Fenerbahçe	Eskişehirspor	1.30 mil. €	Kiralık
08/09	21 Tem 2008	Manisaspor	Fenerbahçe	2.50 mil. €	Bedelsiz
07/08	7 Oca 2008	Beşiktaş	Manisaspor	1.80 mil. €	draft
06/07	1 Tem 2006	Antalyaspor	Beşiktaş	50 Bin €	1.00 mil. €
02/03	1 Tem 2002	Antalyaspor U21	Antalyaspor	-	-
00/01	7 May 2001	Antalyaspor Alt	Antalyaspor U21	-	-
			Toplam transfer kazancı :		21.36 mil. €

Şekil 2.3 Burak Yılmaz Transfermarkt transfer geçmişi [30]

Piyasa değerini etkileyen üçüncü faktör ise futbol enflasyonu olarak tanımlanmaktadır. Ekonominin dünya genelinde ve ülke bazındaki değişimi ve etkisiyle birlikte futbol ekonomisinde de hareketliliğe neden olmaktadır. Bu yüzden piyasa değeri hesaplanırken ekonomik şartlar da göz önünde bulundurulmalıdır.

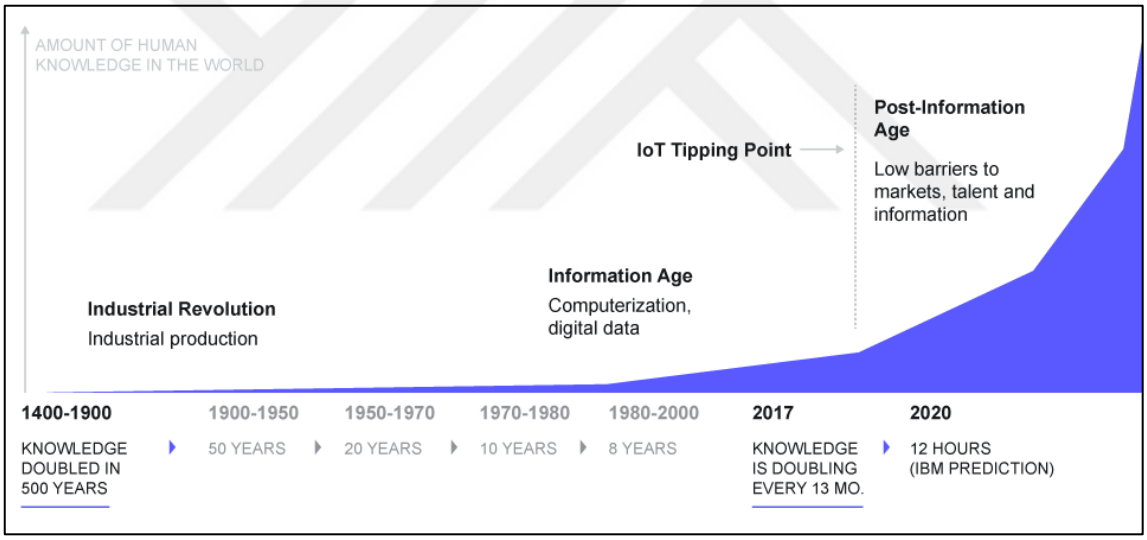
Transfermarkt'ın futbolcu piyasa değerlerini hesaplarken baktığı kriterler:

- Futbolcunun yaşı
- Futbolcunun kulüp ve milli takımda gösterdiği performans
- Futbolcunun oynadığı ligin sportif ve finansal açıdan seviyesi
- Futbolcunun gelişme potansiyeli
- Futbolcu ile ilgilenen kulüp sayısı ve seviyesi
- Futbolcunun tecrübesi
- Futbolcunun sakatlık yaşama duyarlılığı
- Transfer piyasasındaki genel eğilimler, talepler ve ödenen ücretler [31]

Piyasa değerini etkileyen çok fazla sayıda parametre olmasından dolayı veri bilimi ve makine öğrenmesi yöntemlerinin transfer süreçlerinde kullanılması kulüpler için bir fark yaratabilir. Tüm bu parametrelerin bir araya getirilmesi ve işlenmesi ile birlikte piyasa değeri tahmini bir analitik yöntem üzerine inşa edilebilir.

3
MAKİNE ÖĞRENMESİ

Veri çağımızın en önemli araçlarından bir tanesidir. Geçmişten bugüne insanlık sürekli olarak bilgi üretmektedir. Ancak, ürettiği bilgiyi hızlıca işleyip kullanabilmesi insanlık tarihinin son 100 yılında mümkün hale gelmiştir. Bunun nedeni, teknoloji ve bilgisayarlardaki ilerlemenin hızlanması ve bilgisayarların bilgi işleme, depolama kapasitelerinin artmasıdır. Ayrıca, bulut sistemleri ve veri tabanlarındaki gelişimeler de verilerin depolanmasını kolaylaştırdığı gibi veri maliyetini de azaltmıştır. Bu da sektörel anlamda veri biliminin uygulanmasını kolaylaştırmış ve yaygınlaştırmıştır.

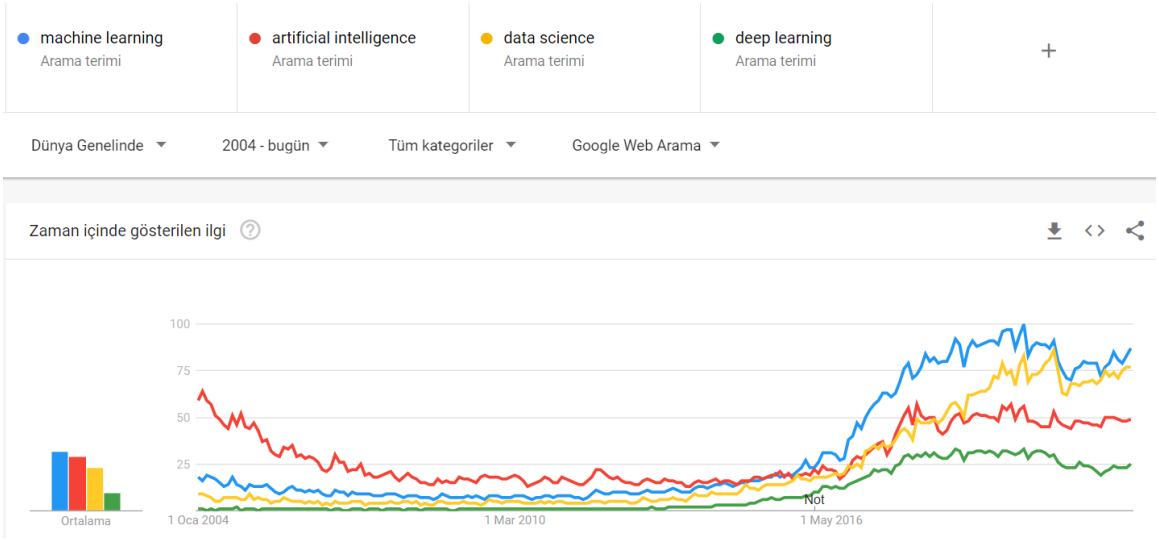


Şekil 3.1 Bilgi katlama eğrisi [32].

Makine öğrenimi, bilgisayarların, açıkça programlanmadan görevleri nasıl yerine getirebileceklerini keşfetmelerini içerir. Bilgisayarlara atanan basit görevler için, makineye eldeki sorunu çözmek için gereken tüm adımları nasıl uygulayacağını bildiren algoritmalar programlamak mümkündür; bu tür görevlerin bilgisayar tarafından öğrenilmesine gerek yoktur. Daha gelişmiş görevler için ise insanların gerekli algoritmaları elle oluşturması zor olabilir. Uygulamada, insan programcıların gerekli her adımı belirlemesinden ziyade, makinenin kendi algoritmasını geliştirmesine yardımcı olmak daha etkili olabilir [33].

Makine öğrenmesi, verileri anlamak için geniş bir araç setini ifade eder. Bu araçlar denetimli (gözetimli) veya denetimsiz (gözetimsiz) olarak sınıflandırılabilir. Genel olarak, denetimli istatistiksel öğrenme, bir veya daha fazla girdiye dayalı bir çıktıyı tahmin etmek için istatistiksel bir model oluşturmayı içerir. Denetimsiz öğrenmede girdiler vardır ancak denetleyici çıktılar yoktur; yine de bu tür verilerden ilişkileri ve yapıyı öğrenebilmek mümkündür [34].

Makine öğrenmesi günümüzde yaygın olarak kullanılmaktadır. Şekil 3.2'de görüldüğü üzere makine öğrenmesi ve ilişkili konular hakkında son yıllarda büyük ilgi bulunmaktadır.



Şekil 3.2 Makine öğrenmesi, yapay zeka, veri bilimi ve derin öğrenme kelimelerinin Google Trends ile zaman içinde gösterilen ilgi grafiği [35]

Makine öğrenmesinin denetimli ve denetimsiz öğrenme yöntemlerinde bulunan alt başlıkları şu şekildedir.

Denetimli öğrenme yöntemleri:

- Sınıflandırma
- Regresyon

Denetimsiz öğrenme yöntemleri:

- Kümeleme
- Boyut indirgeme

- İlişkilendirme kuralları

Araştırma kapsamında hem denetimli hem de denetimsiz öğrenme yöntemleri kullanılmıştır. Futbolcuların piyasa değerlerini tahmin etmek için regresyon yöntemlerinin uygulanması gerekmektedir. Bununla birlikte, piyasa değerini etkilemesi için aday değişken sayısı çok fazla olduğundan boyut indirgeme yöntemi olan Temel Bileşen Analizi uygulanarak veri seti boyutu küçültülmüştür.

3.1 Temel Bileşen Analizi

Temel Bileşen Analizi (TBA) çok değişkenli istatistiksel yöntemlerden bir tanesidir. Makine öğrenmesinde de denetimsiz öğrenme başlığı adı altındadır. TBA, yüksek boyutlu veri setlerinde veri setinin değişken bazında boyutunu azaltmayı amaçlar. Böylelikle veri seti daha az bileşenle en iyi şekilde orijinal veri setini temsil eder. Boyut indirgeme gürültü etkisini ortadan kaldırmaktadır fakat bir miktar bilgi kaybı da söz konusu olmaktadır. Oluşabilecek bilgi kaybının görece hata ve gürültü ile kıyaslanabilir düzeyde küçük olması beklenir [36].

TBA yönteminde eşitlik 3.1 ve 3.2'de gösterildiği gibi X matrisinin M tane N×1'lik gözlem vektöründen oluştuğu kabul edilmektedir.

$$X = [x^1 \mid x^2 \mid \dots \mid x^M]\ (NxM) \tag{3.1}$$

$$X = \begin{bmatrix} x_1^1 & x_1^2 & . & x_1^M \\ x_2^1 & x_2^2 & . & x_2^M \\ . & . & & \\ x_N^1 & x_N^2 & . & x_N^M \end{bmatrix} \tag{3.2}$$

Her değişkenin içerdiği bilgi ve ölçü birimleri bakımından farklılık göstermektedir. Bu yüzden TBA'da kullanılacak tüm değişkenler aynı birime dönüştürülmelidir. Bu işleme de standartlaştırma denir. Standartlaştırma işleminde değişkenlerin ortalaması sıfıra çekilir. Eşitlik 3.3'te vektörlerin ortalaması bulunur ve ardından ortalama vektörü her bir gözlem vektöründen çıkartılarak değişkenlerin ortalaması sıfıra eşitlenir.

$$m = \frac{1}{M} \sum_{i=1}^{M} x^i = \begin{bmatrix} m_1 \\ m_2 \\ . \\ m_N \end{bmatrix} \tag{3.3}$$

$$\tilde{x}^i = x^i - m\ , \forall\ i \tag{3.4}$$

$$\tilde{X} = [\tilde{x}^1 \mid \tilde{x}^2 \mid \dots \mid \tilde{x}^M] \ (NxM) \tag{3.5}$$

$$\tilde{X} = \begin{bmatrix} x_1^1 - m_1 & x_1^2 - m_1 & . & x_1^M - m_1 \\ x_2^1 - m_2 & x_2^2 - m_2 & . & x_2^M - m_2 \\ . & . & . & . \\ x_N^1 - m_N & x_N^2 - m_N & . & x_N^M - m_N \end{bmatrix} \tag{3.6}$$

Standartlaştırılan veri ile kovaryans matrisi eşitlik 3.7 ve 3.8'deki gibi bulunur.

$$C = \tilde{X}\tilde{X}^T \tag{3.7}$$

$$C = \begin{bmatrix} (x_1^1 - m_1)^2 & (x_1^2 - m_1)(x_2^1 - m_2) & . & (x_1^M - m_1)(x_N^1 - m_N) \\ (x_2^1 - m_2)(x_1^2 - m_1) & (x_2^2 - m_2)^2 & . & (x_2^M - m_2)(x_N^2 - m_N) \\ \left(x_N^1 - m_N \right)(x_1^M - m_1) & (x_N^2 - m_N)(x_2^M - m_2) & . & (x_N^M - m_N)^2 \end{bmatrix} \tag{3.8}$$

Özdeğer-özvektör ayrışması için eşitlik 3.9 kullanılır.

$$Cv = \lambda v \tag{3.9}$$

Özdeğerler büyükten küçüğe doğru sıralanır. Sıralanmış özdeğerlere göre özvektörlerden ilk p tanesi bir matrisin sütunlarını oluşturulacak şekilde yerleştirildiğinde W izdüşüm matrisi elde edilir.

$$W = \left[w_1 \mid w_2 \mid \dots \mid w_p\right] \tag{3.10}$$

$$y^i = W^T x^i, i = 1,2, \dots, M \tag{3.11}$$

3.2 Regresyon ve Araştırmada Kullanılan Regresyon Modelleri

Regresyon yöntemi ilk olarak Adrien Marie Legendre tarafından 1805 yılında ortaya atılmıştır ve ilk şekli ise en küçük kareler prensibidir [37]. Regresyon terimi ise İngiliz istatistikçi Francis Galton tarafından bir biyolojik inceleme için ortaya atılmıştır [38]. Regresyon analizi, nicel iki veya ikiden fazla değişken arasındaki ilişkiyi nedensellik bağlamında ölçmek için kullanılan bir yöntemdir. Regresyonda tek bir değişken kullanılarak analiz yapılıyorsa tek değişkenli regresyon, eğer birden çok değişken kullanılıyorsa çok değişkenli regresyon denir. Regresyonda kullanılan değişkenlerden birine bağımı değişken diğerlerine ise bağımsız değişken denmektedir. Bağımsız değişkenlerin bağımlı değişken ile olan ilişkisinde doğrusallık varsa doğrusal olan regresyon modelleri tercih edilir, doğrusallık yoksa doğrusal olmayan regresyon modelleri tercih edilir. Çoklu doğrusal regresyon modeline ilişkin denklemler aşağıdaki gibidir.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_n X_{in} + \varepsilon_i \quad (3.12)$$

$$\sum_{i=1}^{n} e_i^{\,2} = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \quad (3.13)$$

$$b = [X'X]^{-1} X'Y \quad (3.14)$$

Araştırma kapsamında bağımlı değişken ile bağımsız değişkenler arasında doğrusal bir ilişki bulunmadığından dolayı doğrusal olmayan regresyon yöntemleri tercih edilmiştir. Literatürdeki popüler makine öğrenmesi algoritmalarından LGBM, XGBoost, Bagging (Torbalama) ve Random Forest (Rasgele Orman) algoritmaları incelenmiştir.

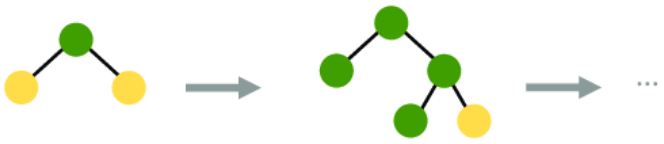
3.2.1 LGBM Algoritması

LGBM algoritması olarak bilinen Light Gradient Boosting Machine, Microsoft tarafından geliştirilen ücretsiz ve açık kaynak kodlu bir makine öğrenmesi algoritmasıdır [39]. Temeli karar ağaçlarına ve gradyan arttırımına dayanır.

LGBM algoritmasının sağladığı avantajlar:

- Daha hızlı eğitim hızı ve daha yüksek verimlilik.
- Daha düşük bellek kullanımı.
- Daha iyi doğruluk.
- Paralel, dağıtılmış ve GPU öğrenimi desteği.
- Büyük ölçekli verileri işleme yeteneğine sahiptir [40].

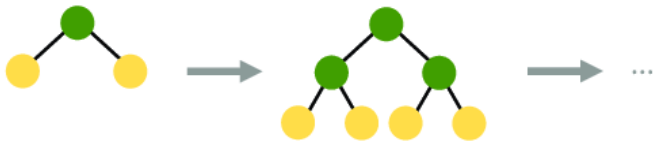
Yukarıdaki avantajlar ile makine öğrenimi yarışmalarının kazanan birçok çözümünde LGBM algoritması yaygın olarak kullanılmaktadır [41]. Ayrıca XGBoost algoritmasının daha performanslı halidir.



Şekil 3.3 LGBM algoritmasının yaprak odaklı büyüme stratejisi

3.2.2 XGBoost Algoritması

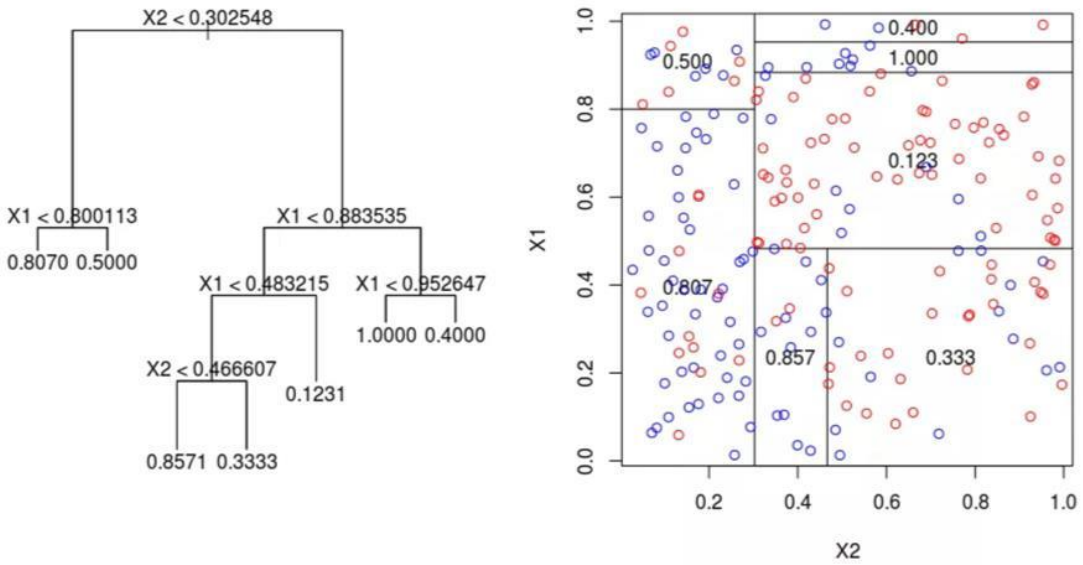
XGBoost (eXtreme Gradient Boosting) algoritması Gradient Boosting algoritmasının optimize edilmiş yüksek performanslı versiyonudur ve açık kaynak kodludur [42]. XGBoost algoritması makine öğrenmesi yarışmalarında kazandıran popüler algoritmalardan bir tanesidir [43].



Şekil 3.4 XGBoost algoritmasının seviye odaklı büyüme stratejisi

3.2.3 Bagging (Torbalama Yöntemi) Algoritması

Karar ağaçları makine öğrenmesinde hem sınıflandırma hem de regresyon problemlerinde kullanılan denetimli öğrenme yöntemlerindendir. Karar ağaçları heterojen veri setlerini belirlenmiş bir hedef değişkene göre homojen alt gruplara ayırır. Alt gruplara ayırma işlemi bir takım karar verme kuralları ile gerçekleştirilir [44].



Şekil 3.5 Örnek karar ağacı

Topluluk öğrenme yöntemlerinin temeli birden çok karar ağacının ürettiği tahminlerin bir araya getirilmesiyle belirlenir. Bagging ve Random Forest algoritmaları

topluluk öğrenme yöntemlerinde en yaygın bilinen ve kullanılan başlıca algoritmalardır.

Bagging yöntemi ile veri seti kullanılarak çok sayıda karar ağacı eğitilir. Eğitilen her bir ağaç bir tahmin sonucu üretir. Regresyon problemlerinde ağaçların ürettiği tahminlerin ortalaması alınır. Sınıflandırma problemlerinde ise tahminlerin oylaması yapılarak tahmin sonucuna karar verilir. Oluşturulan her bir ağaç için rasgele örneklem yöntemi olan bootstrap kullanılır [45].

3.2.4 Random Forest (Rasgele Orman) Algoritması

Random Forest, Bagging yöntemine benzer olarak çok sayıda karar ağacı eğitilerek tahminlerin bir araya getirilmesine dayanır ve bootsrap yöntemi ile veri setinden rasgele örnekler seçilerek her bir ağaç oluşturulur. Bagging yönteminde örnek seçimimi yapılırken Random Forest Bagging yöntemine ek olarak değişken seçiminde de rasgelelik kullanarak her bir ağacı oluşturur [46].

3.3 Başarı Ölçüm Metrikleri

Regresyon algoritmalarında model başarısının ölçülmesinde pek çok metrik kullanılmaktadır. Literatürde yaygın olarak kullanılan ve araştırmada kapsamında da kullanılan metrikler aşağıdaki gibidir:

- RMSE (Root Mean Square Error): Hata Kareler Ortalamasının Karekökü
- MAE (Mean Absolute Error): Ortalama Mutlak Hata
- MAPE (Mean Absolute Percentage Error): Ortalama Mutlak Yüzde Hatası
- R^2: Açıklayıcılık Katsayısı

RMSE; tahmin edilen değer ile gerçek değerler arasındaki farkın karelerinin ortalamasının kareköküdür.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \tag{3.15}$$

MAE yani ortalama mutlak hata; tahmin edilen değer ile gerçek değerler arasındaki farkın mutlak değerlerinin ortalamasıdır.

$$MAE = \frac{\sum_{i=1}^{n}|\hat{y}_i - y_i|}{n} \tag{3.16}$$

MAPE; bir tahmin yönteminin tahmin doğruluğunun bir ölçüsüdür.

$$MAPE = \frac{100\%}{n} \sum_i^n \left|\frac{A_i - F_i}{A_i}\right| \quad (3.17)$$

R^2; regresyon modeli tarafından açıklanabilen değişimin toplam değişim içindeki payı olarak tanımlanır ve bağımlı değişkendeki değişimin yüzde kaçının bağımsız değişken tarafından açıklanabildiği konusunda bilgi verir [36]. R^2 değeri 0 ile 1 arasındadır ve değer 1'e yaklaştıkça bağımlı değişkendeki değişimin büyük bir bölümünün bağımsız değişkenler ile açılandığını söylemektedir.

Varyans Açıklama Oranı veya Açıklanabilir Varyans Oranı; TBA'da başarı metriği olarak kullanılmaktadır. TBA ile oluşturulan bileşenlerin açıklanabilir varyans oranlarının toplamı %70'in üzerinde ise orijinal veri yapısını iyi açıklayan bileşenler olarak değerlendirilir [36].

3.4 Model Doğrulama Yöntemi

Makine öğrenmesi algoritmalarının başarısını ölçebilmek ve modeli daha iyi değerlendirebilmek için model doğrulama yöntemlerine ihtiyaç vardır. En sık başvurulan model doğrulama yöntemleri ise K-Katlı Çapraz Doğrulama, Boostrap, Leave-one-out ve Holdout olarak bilinmektedir. Bu doğrulama yöntemleri hem sınıflandırma hem de regresyon problemlerinde kullanılmaktadır.

3.4.1 Holdout (Sınama Seti Yöntemi)

Holdout yönteminde veri seti eğitim seti ve test seti olmak üzere iki parçaya ayrılmaktadır. Model eğitim verisi ile eğitilir ve test verisi ile de modelin başarısı yani performansı ölçülür.



Şekil 3.6 Holdout yöntemi

Yöntemin yaygın kullanımında veri setinin 2/3'ü eğitim seti olarak ayrılır, geriye kalan 1/3'lük set ise test seti olarak model performansını ölçmede kullanılır.

Veri setinde yeterli gözlem sayısı olduğu düşünüldüğünde veri seti eğitim test setlerinin yanı sıra eğitim, validasyon ve test seti olarak da bölünmektedir. Buradaki validasyon seti verinin iki parçaya bölündüğündeki test seti ile denklik göstermektedir. Model iyileştirmesi ve performans değerlendirmesi validasyon setine göre yapılır. Test seti de modelin genelleştirme başarısını göstermektedir. Eğer veri setinde gözlem sayısı az ise Holdout yöntemi ile veri seti iki parçaya ayrılamaz. Bunun nedeni, modeli eğitmek ve test etmek için kullanılanacak veri miktarının belirlenmesinde kafa karışıklığı yaşanması ve ayrımın doğru yapılamama ihtimalinden dolayı Holdout dışında farkı yöntemlerin geliştirilmesini zorunlu hale getirmiştir [47].

3.5 SHAP Analizi

SHAP (Shapley Additive Explanation), herhangi bir makine öğrenimi modelinin çıktısını açıklamaya yönelik bir oyun teorisi yaklaşımıdır [48]. Makine öğrenmesi, derin öğrenme ve yapay zeka modellerinin geleneksel yöntemlerle açıklanabilirliği zor olduğundan SHAP, modelleri daha açıklanabilir hale getiren alternatif bir yöntemdir. SHAP yöntemi, her bir özniteliğin model başarısına etkisinin gözlenmesine olanak sağlamaktadır. Tespit edilen en etkin özniteliklerle de aynı veya daha yüksek model başarısı oluşturulmasına fırsat sunabilmektedir [49].

Şekil 3.7'de bir tahminde bulunurken tahmin sonucuna etki eden özniteliklerin etkisi gösterilmektedir.



Şekil 3.7 SHAP ile örnek model tahmini açıklanabilirliği [50]

3.6 Python Programlama Dili

Guido van Rossum tarafından geliştirilen Python programlama dili, nesne yönelimli, yorumlamalı, birimsel (modüler) ve etkileşimli yüksek seviyeli bir dildir

[51]. Birçok farklı problem sınıfına uygulanabilen üst düzey bir genel amaçlı programlama dilidir. Sektörde ve akademide yaygın olarak kullanılmaktadır. Python; veri bilimi, makine öğrenmesi, derin öğrenme ve yapay zeka alanında kullanılan en popüler programlama dillerinin başında yer almaktadır. Araştırma kapsamında veri toplama, veri ön işleme, analiz, veri görselleştirme vb. uygulamalar Python kullanılarak yapılmıştır. Şekil 3.8'de görüldüğü üzere Python TIOBE indeksine göre Mart 2022 itibariyle en popüler programlama dilidir.

Tablo 3.1 TIOBE indeksine göre en popüler 20 programlama dili [52]

Mar 2022	Mar 2021	Programlama Dili	Puan	Değişim
1	3	Python	14.26%	+3.95%
2	1	C	13.06%	-2.27%
3	2	Java	11.19%	+0.74%
4	4	C++	8.66%	+2.14%
5	5	C#	5.92%	+0.95%
6	6	Visual Basic	5.77%	+0.91%
7	7	JavaScript	2.09%	-0.03%
8	8	PHP	1.92%	-0.15%
9	9	Assembly Language	1.90%	-0.07%
10	10	SQL	1.85%	-0.02%
11	13	R	1.37%	+0.12%
12	14	Delphi/Object Pascal	1.12%	-0.07%
13	11	Go	0.98%	-0.33%
14	19	Swift	0.90%	-0.05%
15	18	Matlab	0.80%	-0.23%
16	16	Ruby	0.66%	-0.23%
17	12	Classic Visual Basic	0.60%	-0.66%
18	20	Objective-C	0.59%	-0.31%
19	17	Perl	0.57%	-0.58%
20	38	Lua	0.56%	+0.23%

4
VERİ SETİ

Bir veri bilimi projesinde kullanılacak veri kaynağının erişilebilirliği, kalitesi, büyüklüğü ve çeşitliliği çok önemlidir. Veri setleri bu özelliklere sahip olduğu taktirde analizlerin güvenirlikleri ve başarısı artar. Ayrıca, sonuçların büyük kitlelere genelleştirilebilmesi de kolaylaşır.

Spor dallarında veri toplama süreçlerinin zahmetli olmasından dolayı, ücretsiz ve yukarıda bahsedilen özellikleri taşıyan kapsamlı veri setleri bulunmamaktadır. Spor verilerini toplayan şirketlerin ticari nedenlerden dolayı verilerini ücretsiz bir şekilde kamuoyu ile paylaşmamaktadır. Bu yüzden araştırmacılar, ilgilendikleri konu hakkında farklı kaynaklardan veri toplamaya ihtiyacı olabilir.

Kapsamlı bir literatür ve çevrimiçi incelemesinden sonra, spor istatistikleri içeren internet sitelerinden veri kazıma yöntemi ile veri toplanabileceği görülmüştür. Araştırma için kullanılabilecek pek çok internet sitesi, veri tabanları, haber siteleri mevcuttur.

Farklı kaynaklardan veri toplamak, daha sonra araştırmacılar için problemler yaratabilir. Verilerin bir araya getirilmesi, birleştirilmesi, verilerdeki hata ve anomalilerin tespit edilmesi, benzer veya işe yaramayan özelliklerin tespiti ve çıkartılması gibi çeşitli problemler araştırma süresini uzatmakla beraber kapsamını da daraltabilir. Ayrıca, çoğu internet sitesinde benzer istatistikler bulunabileceğinden veri toplama süreci veri çeşitliliğini, kalitesini arttırmaya yönelik ve ihtiyacı karşılayacak şekilde olmalıdır. Bu yüzden, araştırma sorunu yanıtlayabilecek veri kaynaklarından olan Transfermarkt ve Football Reference siteleri tercih edilmiştir. Ek olarak, Statsbomb ile olan partnerliği sebebiyle bazı ileri düzey performans istatistikleri FBRef sitesinde bulunmaktadır.

4.1 Veri Kaynakları: Transfermarkt, Football Reference, Statsbomb

TM (www.transfermarkt.com) ve FBRef (www.fbref.com) maç istatistikleri, transfer haberleri, piyasa değerleri gibi futbol bilgileri içeren internet sitelerinden bazılarıdır.

Bir futbolcunun piyasa değeri uzun yıllardır futbol uzmanları, kulüp yöneticileri tarafından belirlenirken, TM kalabalığın bilgeliği teorisini kullanarak futbolcuların piyasa değerlerini kitlelere tahmin ettiren bir platform haline gelmiştir. TM'tan futbolcuların piyasa değerleri, sakatlık bilgileri, karakteristik özellikleri, transfer geçmişi gibi özellikler toplanmıştır.

Futbolcuların performans istatistiklerini elde etmek için FBRef sitesinden yararlanılmıştır. FBRef sitesini diğer futbol sitelerinden ayıran en önemli özellik, site içerisinde çok fazla sayıda lig ve oyuncunun olmasıyla birlikte kapsamlı ve çeşitli performans istatistiklerini içeriyor olmasıdır. Ek olarak, performans istatistikleri içerisinde Statsbomb (www.statsbomb.com) şirketinin sağladığı ileri düzey istatistikler de bulunmaktadır.

4.2 Veri Toplama Metodu: Web Kazıma

Web çıkarma veya toplama olarak da bilinen web kazıma, World Wide Web'den (WWW) veri çıkarmak ve daha sonra almak veya analiz etmek için bir dosya sistemine veya veritabanına kaydetmek için bir tekniktir [53]. Veriler, bir kullanıcı tarafından manuel veya bilgisayar programları ve yazılımlar aracılığı ile otomatik olarak toplanabilir. İnternet ortamında sürekli olarak bilgi üretilmesinden dolayı verilerin toplanmasında hızlı ve güçlü bir yöntemdir.

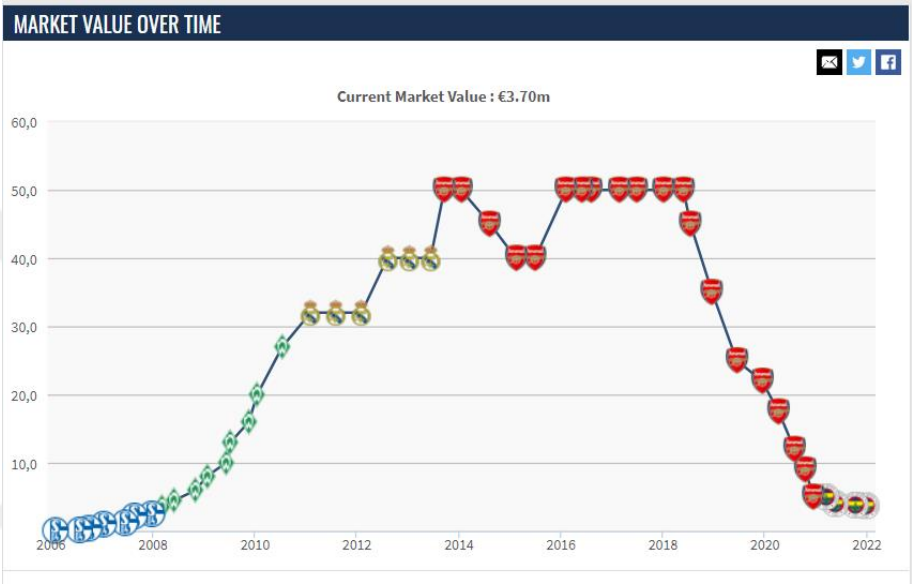
TM ve FBRef sitelerinden veri elde etmek için Python programlama dili ve veri kazıma kütüphaneleri olan Beautiful Soup, js2xml ve pyjsparser kütüphaneleri kullanılmıştır.

Veri toplama aşamaları:

- Hangi sezonların ve liglerin verilerinin toplanacağının kararlaştırılması.
- Her bir sezon ve lig içerisinde bulunan tüm futbolcuların site içerisindeki url uzantılarının bulunması ve futbolcu listesini oluşturma.

- Toplanmak istenen verilere özel kodların hazırlanması.
- Kodlar yardımıyla futbolcu listesi içerisinde bulunan tüm futbolcular için teker teker verilerin elde edilmesi ve depolanması.

Şekil 4.1 ve 4.2'de Mesut Özil adlı futbolcu için toplanmak istenen bazı verilerin kaynakları gösterilmektedir. Web kazıma ile grafikler ve tablolardan kolaylıkla her bir futbolcu için teker teker veri toplanabilir ve öznitelik çeşitliliği olan bir veri seti elde edilebilir.



Şekil 4.1 Mesut Özil TM Piyasa Değeri Akışı (14 Şubat 2022) [54]

Standard Stats Share & Export ▾ Glossary Toggle Per90 Stats

Club Expanded | Club Collapsed | Domestic Leagues | Domestic Cups | International Cups | National Team

					Playing Time				Performance							Per 90 Minutes					Expected				Per 90 Minutes				
Season	Age	Squad	Country	Comp	MP	Starts	Min	90s	Gls	Ast	G-PK	PK	PKatt	CrdY	CrdR	Gls	Ast	G+A	G-PK	G+A-PK	xG	npxG	xA	npxG+xA	xG	xA	xG+xA	npxG	npxG+xA
2021-2022	32	Fenerbahce		2. Europa Lg	2	2	141	1.6	1	0	1	0	0	0	0	0.64	0.00	0.64	0.64	0.64	0.5	0.5	0.4	0.9	0.32	0.23	0.55	0.32	0.55
2021-2022	32	Fenerbahce	TUR	1. Süper Lig	19	12	1,119	12.4	7	2	4	3	3	0	0	0.56	0.16	0.72	0.32	0.48									
2020-2021	31	Fenerbahce	TUR	1. Süper Lig	10	8	645	7.2	0	1	0	0	0	0	0	0.00	0.14	0.14	0.00	0.14									
2019-2020	30	Arsenal	ENG	EFL Cup	2	2	134	1.5	0	1	0	0	0	0	0	0.00	0.67	0.67	0.00	0.67									
2019-2020	30	Arsenal		2. Europa Lg	2	1	151	1.7	0	0	0	0	0	0	0	0.00	0.00	0.00	0.00	0.00	0.0	0.0	0.2	0.2	0.02	0.12	0.14	0.02	0.14
2019-2020	30	Arsenal	ENG	FA Cup	1	1	76	0.8	0	0	0	0	0	0	0	0.00	0.00	0.00	0.00	0.00									
2019-2020	30	Arsenal	ENG	1. Premier League	18	18	1,438	16.0	1	2	1	0	0	1	0	0.06	0.13	0.19	0.06	0.19	1.3	1.3	2.2	3.5	0.08	0.14	0.22	0.08	0.22
2018-2019	29	Arsenal		2. Europa Lg	10	8	646	7.2	1	1	1	0	0	1	0	0.14	0.14	0.28	0.14	0.28	1.1	1.1	1.4	2.5	0.16	0.20	0.35	0.16	0.35

Şekil 4.2 Mesut Özil FBRef Performans İstatistikleri Örneği (14 Şubat 2022) [55]

4.3 Toplanan Veri

Veri toplamaya başlamadan önce araştırmanın yapılacağı hedef grup belirlenmiştir. Her ne kadar dünya genelindeki tüm ligler ve futbolcular hakkında veri bulunsa da ücretsiz paylaşılan verilerde popülerliği ve gelişmişliği yüksek olan liglerde daha çok bilgi toplanabileceği görülmüştür. Bu liglerde, temel istatistiklerin dışında daha ileri düzey istatistikler paylaşılmaktadır.

Avrupa'nın beş büyük ligi olan Almanya, Fransa, İngiltere, İspanya ve İtalya liglerinde oynayan futbolcular hedef grup olarak seçilmiştir. Veri seti 2017-2021 sezonları arasında bu liglerde oynamış 3948 futbolcudan oluşmaktadır. Veri setinde 11013 gözlem ve 418 değişken vardır. Değişkenler içerisinde futbolcuların piyasa değerleri, sakatlık, popülerlik, demografik ve karakteristik özellikler gibi parametrelerle birlikte lig, kupa, Avrupa kupaları ve milli takım maçlarındaki performans istatistikleri bulunmaktadır.

Araştırma kapsamında toplanan tüm TM ve FBRef değişkenleri ek bölümünde detaylı bir şekilde gösterilmektedir. Toplanan verinin özet hali ise Tablo 4.3.1'de verilmiştir.

Tablo 4.1.'de TM ve FBRef'ten toplanan futbolcu kişisel bilgileri kategorisi altında, futbolcuların yaşı, boy, kilo, milliyetleri, kullandığı ayak, oynadıkları mevki gibi kişisel bilgileri içermektedir. Futbolcuların sakatlık geçmişinden ve cezalı olma durumundan dolayı kaçırdıkları maç sayısı, zamana bağlı piyasa değeri değişimleri, transfer süreçlerindeki bonservis bedelleri ve takım bilgileri TM'den alınmıştır. Performans istatistikleri ve haber olma sayıları ise FBRef'den alınmıştır.

Tablo 4.1 Web kazıma ile toplanan verilerin özet listesi

Veri Kaynağı	Değişken Kategorisi	Alt Kategori
Transfermarkt	Futbolcu Kişisel Bilgileri	Yaş, boy, kilo, milliyeti, mevki, kullandığı ayak vb.
	Piyasa Değeri	Piyasa değeri geçmişi
	Sakatlık ve Kaçırılan Maç Bilgisi	Sakatlık türü, kaçırdığı maç sayısı, sakat olduğu gün sayısı
	Transfer Geçmişi	Bonservis bedeli, transfer olunan takım bilgisi
	Cezalı Olma Durumu	Ceza nedeni ve kaçırdığı maç sayısı
Football Reference	Futbolcu Kişisel Bilgileri	Yaş, boy, kilo, milliyeti, mevki, kullandığı ayak vb.
	Popülerlik	Haber olma sayısı
	Performans İstatistikleri (Lig, Kupa, Milli Maçlar, Uluslararası Turnuvalar)	Çeşitli genel istatistikler
		Oynama süresi
		Performans
		Beklenen istatistikleri
		Pas
		Pas tipi
		Şut
		Defansif aksiyonlar
		Pozisyonlanma

Veri setinin 2017-2020 arasını kapsayan üç sezonluk kısmı eğitim verisi (n=8196) olarak kullanılmışken, veri setindeki son sezon olan 2020-2021 sezonu test verisi (n=2817) olarak kullanılmıştır.

5
UYGULAMA

Uygulama bölümünde araştırmada yapılan tüm çalışmalar sırasıyla açıklanmaktadır. Veri toplama sürecinden sonra modelleme ve modelleme öncesinde yapılan çalışmalar, özellik seçimi, model karşılaştırmaları, model performans incelemeleri anlatılacaktır.

5.1 Veri Ön İşleme

Makine öğrenimi modellerinin açıklanabilirliğini sağlamak, doğal olarak algoritmik şeffaflık ve adalet kavramlarıyla ilişkili olan aktif bir araştırma konusudur [56]. Model performansını, modelin davranışını ve tahmin sonuçlarını açıklamada veri ön işlemenin etkisi çok büyüktür. Bu yüzden araştırma kapsamında kullanılacak olan tüm verilerde hataların tespit edilmesi, düzeltilmesinde ve işe yarar bilgilerin ortaya çıkartılması konularına odaklanılmıştır.

TM ve FBRef sitelerinden web kazıma yöntemiyle elde edilen tüm verilere aşağıdaki işlem uygulanmıştır.

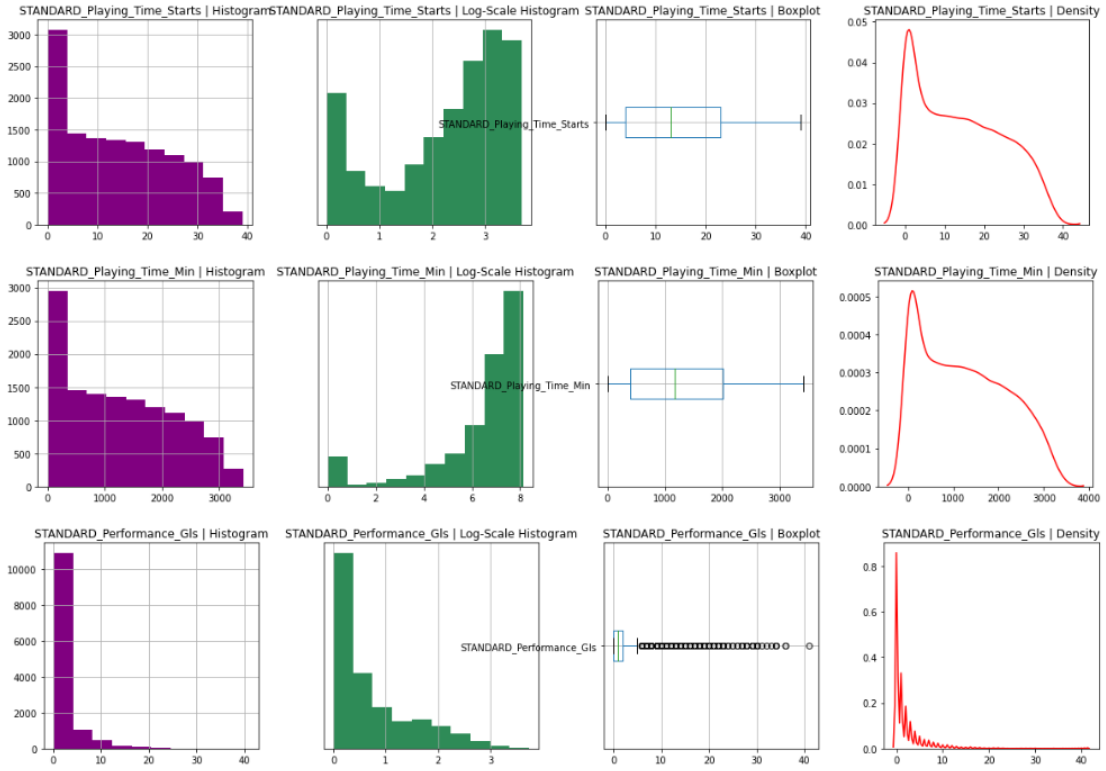
- TM ve FBRef veri setlerindeki veri tiplerinin ve değişken isimlendirmelerinin düzeltilmesi
- TM ve FBRef kulüp ve futbolcu isimlerinin, futbolcu kimlik bilgilerinin eşleştirilmesi
- Veri setlerinin birleştirmeye uygun hale getirilmesi
- Keşifçi veri analizi ve özet istatistiklerin incelenmesi
- Eksik gözlem, aykırı gözlem, hatalı gözlem tespiti
- Bilgi içeriği az olan ve bilgi sağlamayan değişkenlerin veriden silinmesi
- Kategorik ve sürekli değişken incelemeleri
- Tekilleştirme: Gözlem ve değişken bazında çoklama problemlerinin giderilmesi

- Kategorik değişkenlerin sayısal hale getirilmesi
- Ölçeklendirme

TM ve FBRef sitelerinden elde edilen verilerin eşleştirilmesi için önce, futbolcuların hem TM'deki hem de FBRef'teki futbolcu kimlik bilgileri eşleştirilmiştir. Eşleştirme sonucunda beş büyük Avrupa liginde oynayan 3948 futbolcu iki veri setinde de kullanabilir hale gelmiştir.

Keşifçi veri analizi yöntemleri ve özet istatistikler yardımıyla değişkenlerin dağılımı, içerdikleri faydalı ve faydasız bilgi, değişkenlerdeki anomaliler ve eksiklikler hakkında bilgi edinilmiştir. FBRef'ten alınan performans istatistiklerinde aykırı gözlem veya hata içeren gözlemlere rastlanılmamıştır. TM'den alınan sakatlık geçmişi ve kaçırılan maçlar, cezalı olma durumu veri setlerinde anomali olan durumlar veri setinden çıkartılmıştır. Futbolcuların kişisel ve karakteristik bilgileri hem TM hem de FBRef'ten alınarak veri doğrulaması yapılmıştır.

Lig, kupa, uluslararası maçlar ve milli takım performans istatistiklerinde çoğunlukla değişken dağılımları sağa çarpıktır (Şekil 5.1).

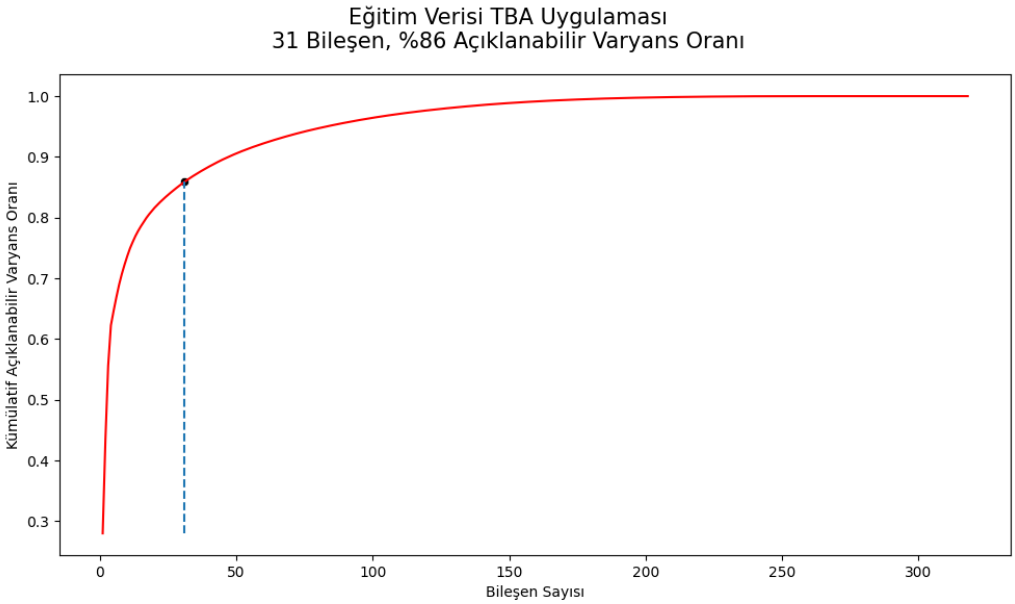


Şekil 5.1 Performans istatistikleri veri görselleştirme

5.2 Temel Bileşen Analizi

Makine öğrenmesinde model performansını ve başarısını arttırmak için veri ön işleme, öznitelik mühendisliği, özellik seçimi ve parametre ayarı yapmak gerekli-dir. Veri setindeki değişken sayısı arttıkça modelin karmaşıklığı artmaktadır ve makine öğrenmesinin en büyük problemlerinden biri olan modelin aşırı ezberle-mesine (overfitting) neden olmaktadır. Ayrıca, karmaşıklığı yüksek bir modelin canlıya alındığında sorun yaşaması da muhtemeldir. Bu nedenlerden dolayı mo-deller, temsiliyeti en iyi ve en kolay olacak şekilde tasarlanmalıdır. Bu yüzden, büyük veri setlerinde yapılan çalışmalarda boyut indirgeme veya özellik seçimi yöntemleri sıklıkla kullanılmaktadır.

Veri setinde bulunan 344 değişkenin 318 tanesi futbolcuların lig, kupa, uluslara-rası turnuva ve milli maç performans istatistiklerinden oluşmaktadır. 318 değiş-kenin temsiliyetini daha az değişkenle yapabilmek için boyut indirgeme yöntemi olan TBA kullanılmıştır.



Şekil 5.2 Kümülatif açıklanabilir varyans oranı ve bileşen sayısı grafiği

TBA uygulaması sonucunda 318 değişkenin 31 bileşen ve %86'lık açıklanabilirlik ile temsil edilebileceği görülmüştür. Modelleme sırasında bileşenlerin modele yap-tıkları katkılar ve model performansları incelenerek 10 bileşenin yeterli olacağı görülmüştür.

Tablo 5.1 Modelleme kullanılan bileşen sayısı ve kümülatif açıklanabilir varyans oranı

Bileşen Sayısı	Kümülatif AVO
10	%74

5.3 Modelleme

Araştırma kapsamında 4 adet makine öğrenmesi algoritması kullanılmıştır. Topluluk öğrenme yöntemi olan Random Forest (Rasgele Orman), Bagging (Torbalama), XGBoost ve LGBM algoritmaları kullanılmıştır.

Modeller Python programlama dili içerisindeki varsayılan ayarlarıyla kurulmuştur. Kurulan 4 model ile araştırmada kullanılabilecek olan referans modele karar verilmiştir.

Tablo 5.2 Referans model karşılaştırmaları

MODEL	SET	RMSE	MAE	R^2
RF	Eğitim	403,51	285,67	0,9504
RF	Test	973,51	721,38	0,7028
Bagging	Eğitim	484,01	322,23	0,9286
Bagging	Test	1011,56	742,61	0,6791
XGBoost	Eğitim	441,03	317,66	0,9407
XGBoost	Test	950,30	695,85	0,7168
LGBM	Eğitim	718,87	533,46	0,8425
LGBM	Test	916,84	658,56	0,7364

Kurulan 4 model içerisinde en iyi performansı sergileyebilecek olan modelin LGBM olduğu Tablo 5.2'de görülmektedir.

LGBM modelinin performansını iyileştirmek için parametre ayarı ve öznitelik mühendisliği denemeleri yapılmıştır. Tablo 5.3'te referans model olan 1. Model ile geliştirmelerin yapıldığı modellerin kıyaslamaları eğitim ve test sonuçları olarak verilmiştir.

Tablo 5.3 3.LGBM modelinin performansı

MODEL PERFORMANSI						
LGBM	EĞİTİM			TEST		
	RMSE	MAE	R^2	RMSE	MAE	R^2
1. Model	718,87	533,46	0,8425	916,84	658,56	0,7364
2. Model	884,21	637,29	0,7548	901,26	647,69	0,7401
3. Model	708,14	496,57	0,8471	694,21	493,82	0,8489

3 model arasındaki farklılığın nedeni kısaca Tablo 5.4'te gösterilmiştir. İlk model referans model olup herhangi bir parametre ayarı ve öznitelik mühendisliği yapılmamıştır. İkinci modelde eğitim ve test veri setleri kullanılarak model performansını iyileştiren model parametrelerini ayarlama işlemi yapılmıştır. Ayarlama işlemi sonucu tahminci sayısı 100, ağacın maksimum derinliği 9, yaprak sayısı 10, öğrenme oranı 0.1 ve L1 düzenleme oranı 0.05 olarak bulunmuştur. Son olarak üçüncü modelde ise veri seti içerisinde bulunan bazı değişkenler üzerinden yeni değişkenler yaratarak modele katkıları incelenmiştir ve model performansını iyileştiren değişkenler final modelde kullanılmıştır. Oluşturulan yeni değişkenlerde modele en büyük katkı yapan özellik futbolcuların bir sezon önceki market değerleri olmuştur.

Tablo 5.4 Model farklılıklarının nedenleri

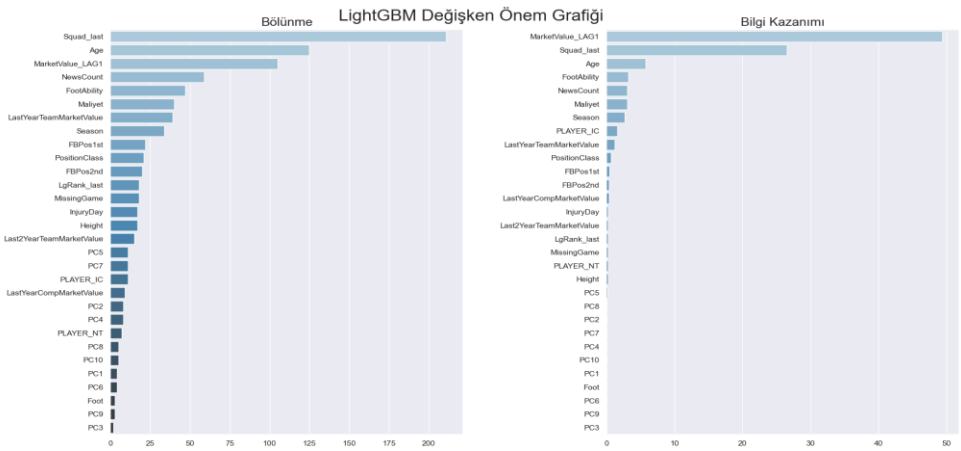
Model	Parametre Ayarı	Öznitelik Mühendisliği
1. Model	Yapılmadı	Yapılmadı
2. Model	Yapıldı	Yapılmadı
3. Model	Yapıldı	Yapıldı

Model performansı ayrıca pozisyon sınıfları incelenerek de değerlendirilebilir.

Tablo 5.5 Pozisyon sınıflarına göre ortalama mutlak hata

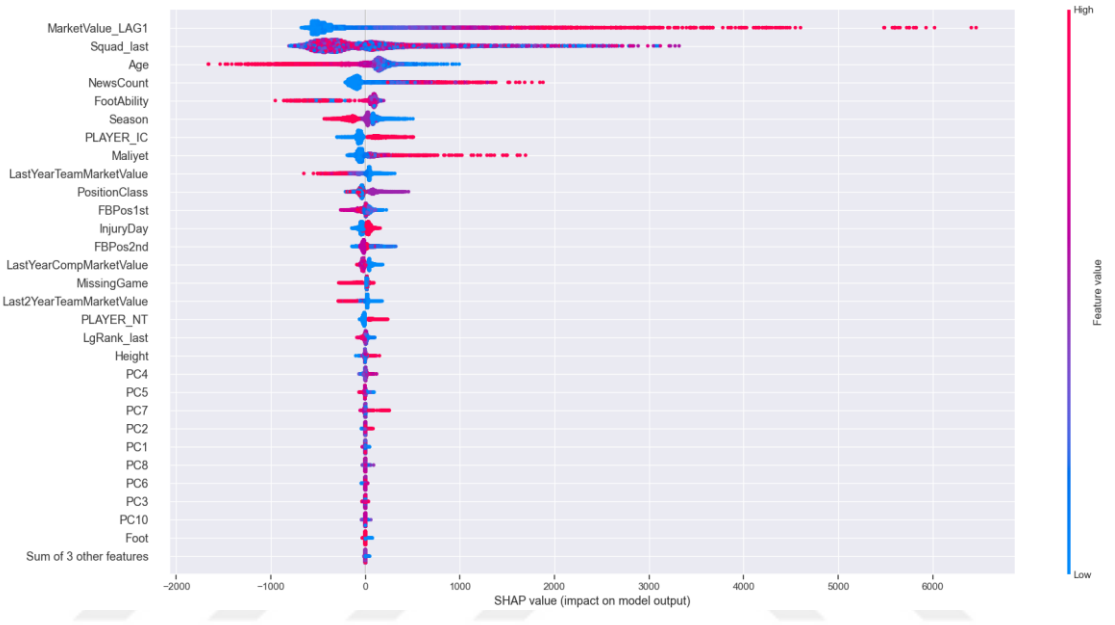
Pozisyonlara Göre Ortalama Mutlak Hata		
Pozisyon Sınıfı	Eğitim MAE	Test MAE
Defans	465,67	480,92
Orta Saha	478,38	490,56
Forvet	548,70	512,18

5.4 Değişken Önem Grafiği ve SHAP Analizi



Şekil 5.3 LGBM değişken önem grafikleri

Şekil 5.3'te LGBM değişken önem grafikleri incelendiğinde piyasa değerine etki eden değişkenler gösterilmiştir. Bilgi kazanımı grafiği incelendiğinde modele en çok katkı sağlayan ilk 3 özellik sırasıyla futbolcuların piyasa değerinin bir sezon önceki değeri, oynadıkları takım ve yaş bilgisi olduğu görülmektedir. Performans istatistikleri ise TBA ile oluşturulan 10 bileşen PC olarak kodlanmış olup, bilgi kazanımı grafiğinde son sıralarda yer almaktadır.

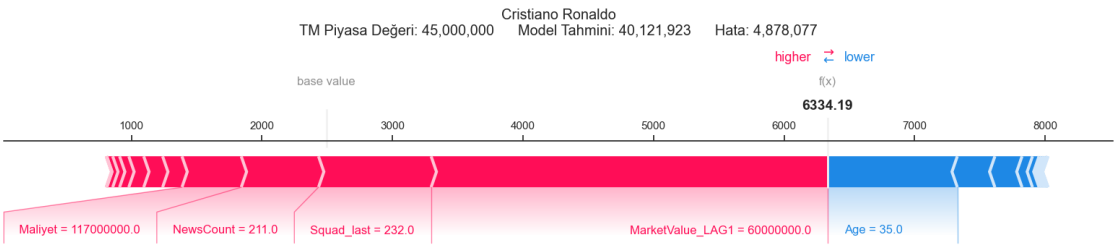


Şekil 5.4 SHAP değişken önem grafiği

Şekil 5.4'te ise SHAP analizi ile oluşturulan değişken önem grafiği verilmiştir. Grafikteki kırmızı noktalar değişken içerisinde bulunan yüksek değerleri, mavi noktalar ise düşük değerleri ifade etmektedir. LGBM grafiğine benzer olarak performans istatistiklerinin önem düzeyi son sıralarda yer alırken, futbolcuların bir sezon önceki piyasa değeri, oynadıkları takım ve yaş bilgisi en önemli ilk üç değişken olarak tespit edilmiştir.

SHAP analizi sayesinde gözlem bazında da model tahminlerinin açıklanabilirliğini incelemek mümkündür. Şekil 5.5'te Cristiano Ronaldo isimli futbolcunun model tahminine etki eden değişkenler görselleştirilmiştir. Oyuncunun 2020-2021 sezonu sonundaki TM piyasa değeri 45 milyon Euro iken model oyuncunun piyasa değerini 40.1 milyon Euro olarak belirlemiştir. SHAP grafiğinde ise model tahminini düşüren en önemli etkenin yaş bilgisi olduğu, model tahminini yükselten en

önemli etkenlerin ise oyuncunun bir sezon önceki piyasa değeri, oynadığı takım, haberlere çıkma frekansı ve maliyeti olduğunu belirtmektedir.



Şekil 5.5 Cristiano Ronaldo'nun SHAP ile değerlendirilmesi

5.5 Bonservis Bedelleri ile Tahmin Edilen Piyasa Değerlerinin Kıyaslanması

TM piyasa değeri, model tahmini ve bonservis bedelleri arasında inceleme yaparak piyasa değerinin transfer süreçlerinde ne kadar gerçekçi olduğu hakkında bilgi edinilebilir. Bunun için özet istatistikler ve hipotez testleri kullanılabilir.

Test verisi olarak kullanılan 2020-2021 sezonu verisi içerisinde 2817 futbolcudan 453 tanesinin bonservis bedeli ile birlikte transfer olduğu tespit edilmiştir. Tablo 5.6'da beş büyük Avrupa liginde 2020-2021 sezonu içerisinde transfer olan futbolcuların TM piyasa değeri, model tahmini ve bonservis bedeli hakkında özet istatistikler bulunmaktadır.

Tablo 5.6 TM piyasa değeri, model tahmini, bonservis bedeli özet istatistikleri

Değişken	Ortalama (Euro)	Standart Sapma (Euro)
TM Piyasa Değeri	10.357.947,01	12.273.541,70
Model Tahmini	9.267.815,34	10.782.091,96
Bonservis Bedeli	7.639.626,93	11.097.023,44

TM piyasa değeri ve model tahmininin bonservisi bedeline olan yakınlıkları ayrıca MAPE metriği ile de ölçülebilir. Tablo 5.7'de hem TM piyasa değeri üzerinden kurulan model tahminlerinin hem de gerçek TM piyasa değerlerinin bonservis bedelini açıklamada yüksek hata içerdiği görülmektedir.

Tablo 5.7 TM piyasa değeri ve model tahmininin bonservis bedeli ile MAPE üzerinden kıyası

Değişkenler	MAPE
TM Piyasa Değeri Bonservis Bedeli	54.16
Model Tahmini Bonservis Bedeli	86.94

Üç değişken arasındaki ilişkinin incelenmesi için ayrıca hipotez testleri kullanılmıştır. Hipotez testleri yapılmadan önce verilerin dağılımının normal olup olmadığı incelenir. Eğer veriler normal dağılma sahipse parametrik testler, normal dağılıma sahip değilse parametrik olmayan istatistiksel testler uygulanmaktadır. Araştırma uygulanan testler sırasıyla Shapiro normallik testi, Mann Whitney U testi ve Spearman korelasyon testidir. Test sonuçları Tablo 5.8, Tablo 5.9 ve Tablo 5.10'da gösterilmiştir.

Tablo 5.8 Shapiro normallik testi

Shapiro Normal Dağılım Testi				
Değişken	p-değeri	p-değeri < 0.05	Hipotez	Dağılım
TM Piyasa Değeri	1,05E-10	Küçük	Reddedilir	Normal dağılmamaktadır.
Model Tahmini	9,35E-12	Küçük	Reddedilir	Normal dağılmamaktadır.
Bonservis Bedeli	3,43E-14	Küçük	Reddedilir	Normal dağılmamaktadır.

- H_0: Değişken normal dağılmaktadır.
- H_1: Değişken normal dağılmamaktadır.

Shapiro testi sonucunda üç değişkende de H_0 hipotezi reddedilmiş olup, değişkenlerin istatistiksel olarak normal dağılmadığı tespit edilmiştir. Bu yüzden parametrik olmayan hipotez testleri tercih edilmelidir.

Tablo 5.9 Mann Whitney U testi

Mann Whitney U Testi				
Değişkenler	p-değeri	p-değeri < 0.05	Hipotez	Yorum
TM Piyasa Değeri Model Tahmini	0,2672	Büyük	Reddedilemez	İki değişkenin ortalamalarında istatistiksel olarak anlamlı bir fark yoktur.
TM Piyasa Değeri Bonservis Değeri	3,11E+05	Küçük	Reddedilir	İki değişkenin ortalamalarında istatistiksel olarak anlamlı bir fark vardır.
Model Tahmini Bonservis Değeri	6,50E+06	Küçük	Reddedilir	İki değişkenin ortalamalarında istatistiksel olarak anlamlı bir fark vardır.

- H_0: İki değişkenin ortalamaları arasında istatistiksel olarak anlamlı bir fark yoktur.
- H_1: İki değişkenin ortalamaları arasında istatistiksel olarak anlamlı bir fark vardır.

Mann Whitney U testi yardımıyla TM piyasa değerinin ve model tahmininin, bonservis bedeli ile ortalamaları arasında anlamlı bir farkın olup olmadığı istatistiksel olarak test edilmiştir. Test sonucunda ise H_0 hipotezi reddedilmiş ve bonservis bedeli ile anlamlı bir farkın olduğu görülmüştür.

Tablo 5.10 Spearman korelasyon testi

Spearman Korelasyon Testi					
Değişkenler	Korelasyon Değeri	p-değeri	p-değeri < 0.05	Hipotez	Yorum
TM Piyasa Değeri Model Tahmini	0,8630	7,61E-121	Küçük	Reddedilir	İki değişkenin arasında istatistiksel olarak anlamlı güçlü bir ilişki vardır.
TM Piyasa Değeri Bonservis Değeri	0,6939	2,49E-51	Küçük	Reddedilir	İki değişkenin arasında istatistiksel olarak anlamlı orta düzey bir ilişki vardır.
Model Tahmini Bonservis Değeri	0,5713	1,33E-25	Küçük	Reddedilir	İki değişkenin arasında istatistiksel olarak anlamlı orta düzey bir ilişki vardır.

- H_0: İki değişken arasında istatistiksel olarak anlamlı bir ilişki yoktur.
- H_1: İki değişken arasında istatistiksel olarak anlamlı bir ilişki yoktur.

TM piyasa değeri, model tahmini ve bonservis değeri arasındaki ilişkinin gücünü test etmek için Spearman korelasyon katsayısı kullanılmıştır.

TM piyasa değeri ile model tahmini arasında 0.86 korelasyon değeriyle güçlü bir ilişki bulunmaktadır. Buna karşın TM piyasa değeri ve model tahmininin bonservis değeri ile ilişki incelendiğinde aralarında 0.69 ve 0.57 korelasyon değerleriyle orta düzey bir ilişki olduğu tespit edilmiştir.

Hipotez testleri, özet istatistikler, model sonuçları ve model hataları incelendiğinde piyasa değeri ile bonservis bedelleri arasında bir ilişkinin olduğu ancak piyasa değerinin tam anlamıyla bonservis bedelini açıklayamadığı bilgisine ulaşılmıştır.

Tablo 5.11 2020-2021 sezonunda en yüksek bonservis değeriyle transfer olan ilk 10 futbolcu ve piyasa değerleri

Futbolcu	Pozisyon	Kulüp	TM Piyasa Değeri (Euro)	Model Tahmini (Euro)	Bonservis Değeri (Euro)
Kai Havertz	FW	Chelsea	70.000.000	87.352.319	80.000.000
Victor Osimhen	FW	Napoli	50.000.000	23.062.162	70.000.000
Ruben Dias	DF	Manchester City	75.000.000	34.762.833	68.000.000
Miralem Pjanic	MF	Barcelona	20.000.000	37.771.878	60.000.000
Timo Werner	FW	Chelsea	65.000.000	61.588.448	53.000.000
Ben Chilwell	DF	Chelsea	45.000.000	38.904.720	50.200.000
Mauro Icardi	FW	Paris S-G	40.000.000	42.701.862	50.000.000
Thomas Partley	MF	Arsenal	40.000.000	30.026.875	50.000.000
Nathan Ake	DF	Manchester City	32.000.000	26.675.087	45.300.000
Leroy Sane	FW	Bayern Munich	60.000.000	70.898.708	45.000.000

6
SONUÇ VE ÖNERİLER

Çalışma kapsamında beş büyük Avrupa liginden (Almanya, Fransa, İngiltere, İspanya, İtalya) toplam 3948 futbolcunun 2017-2021 sezonları arasındaki verileri iki farklı kaynaktan toplanarak makine öğrenmesi yöntemi olan regresyon ile bir tahmin modeli kurulmuştur. Geliştirilen model, transfer süreçlerinde bir futbolcuya ödenecek olan bonservis bedeline referans olan piyasa değerini tahmin etmeye yaramaktadır.

Çalışma için toplanan veri beş büyük Avrupa ligi ile sınırlandırılmıştır. Bunun nedeni ise bu liglerin futbol dünyasında en çok takip edilen ligler olmasıdır. Bu yüzden bu liglerde oynayan futbolcuların çeşitli istatistiklerine daha kolaylıkla ulaşılabilir. Farklı iki kaynaktan veri toplanıldığı için de yeterli ve zengin içerikli veri seti elde etmek amacıyla bu beş lig tercih edilmiştir.

Bu çalışma kapsamında problemin çözümü için takip edilen adımlar sırasıyla; problemin belirlenmesi, probleme uygun örneklem ihtiyacının tespiti, web kazıma ile veri toplama, veri ön işleme, boyut indirgeme ve öznitelik seçimi, modelleme, model performansının değerlendirilmesi ve araştırma sorusunun değerlendirilmesi şeklindedir.

Çalışma içerisinde dört farklı makine öğrenimi algoritması kıyaslanmıştır. Model performansını değerlendirmek ve modelin aşırı ezberlemesinin önüne geçmek için veri seti eğitim ve test seti olarak ikiye ayrılmıştır. Kıyaslamalar regresyon yönteminde kullanılan RMSE, MAE ve R^2 metrikleri ile yapılmıştır. Ölçümler sonucunda dört model arasından LGBM modeli referans model olarak seçilmiştir ve LGBM modeli geliştirilmiştir. LGBM modelinin performansını arttırmak için hiperparametre ayarı ve öznitelik mühendisliği yapılmıştır ve kurulan modeller arasında en iyi performansı sergileyen LGBM modeli final model olarak tercih edilmiştir. Bu çalışmadan sonra yapılacak olan futbolcuların piyasa değerinin tahminlenmesinde ve benzer çalışmalara karşılaştırma imkanı sağlayacak bilgiler sunmaktadır.

Geliştirilen modele etki eden değişkenleri incelemek için LGBM modelinin değişken önem düzeyleri grafiği oluşturulmuş ve SHAP analizi yapılmıştır. Model performansına etki eden ve TM piyasa değerini açıklamada en önemli ilk beş özellik sırasıyla; futbolcunun bir önceki sezondaki piyasa değeri, oynadığı takım, yaşı, kullandığı ayak becerisi ve haber olma sayısı olarak bulunmuştur. Modelleme öncesi futbolcuların 318 performans istatistikleri TBA yardımıyla 10 adet bileşene indirgenmiştir. Değişken önem grafiğinde bu 10 bileşenin son sıralarda yer aldığı görülmektedir. Performans istatistiklerinin model içerisinde etkisinin düşük çıkması TM piyasa değerlerinin, performansa dayalı olmadan ve spekülatif bir şekilde hesaplandığı yönünde şüpheler uyandırmaktadır.

SHAP analizi makine öğrenmesi algoritmalarını daha şeffaf bir şekilde açıklamada kullanılan bir yöntem olup çalışmada futbolcuların piyasa değeri tahmininde tahmini yükselten veya düşüren değişkenlerin ve gözlemlerin incelenmesinde kullanılmıştır.

TM piyasa değerinin gerçekçiliğini ölçmek adına 2020-2021 sezonunda bonservis bedeli ile birlikte transfer olan 453 futbolcu ele alınmıştır. Burada transfer olan futbolcuların TM piyasa değeri, model tahmini ve bonservis bedeli incelenmiştir. Ölçümler özet istatistikler, MAPE ve hipotez testleri kullanılarak yapılmıştır. TM piyasa değeri ve model tahmininin bonservis bedeli ile bir ilişkisi olduğu ancak piyasa değerinin tam olarak bonservis bedeli olarak kullanılamayacağı yargısına ulaşılmıştır.

Sonuç olarak, piyasa değerleri ile bonservis bedelleri arasında bir ilişkinin bulunduğu ancak tam olarak birbirinin yerine geçmedikleri bu çalışmada ispatlanmıştır. İki değişken arasındaki ilişkiden dolayı transfer süreçlerinde bir referans noktası olarak kullanılabilir. Çalışma, makine öğrenmesi algoritmaları kullanılarak hem piyasa değeri tahmini yapmak hem de piyasa değerinin bonservis bedeli ile arasındaki ilişkiyi incelemede kullanılabileceğini göstermiştir.

Spor alanında yapılan veri analitiği çalışmalarında yaşanılan en büyük zorluk ücretsiz ve zengin içerikli kaynaklarına sahip olamamaktır. Bu yüzden araştırmacılar ilgilendikleri problem hakkında farklı kaynaklardan veri toplama sürecine girmek

durumundadır. Farklı kaynaklardan veri toplamak araştırma kapsamını daraltmanın yanında bilgi kaybına, veri setlerini eşleştirme sorunlarına neden olabilir. Çalışma kapsamında da bir futbolcunun piyasa değerini etkileyebilecek çeşitli pek çok değişken kullanılan iki veri kaynağında bulunmamaktadır. TM ve FBRef sitelerinde bulunmayan ancak modele katkı sağlayabilecek diğer değişkenler için başka veri kaynaklarından veri toplanmalıdır. Bu da çalışmanın karmaşıklığını ve zorluğunu arttıran bir faktördür. Fayda sağlayabilecek ek değişkenlerin ve kaynakların veri setine eklenmesi model performansını iyileştirebilir.

Çalışma sonucunda makine öğrenmesi modelleri spor dallarında bir karar verme mekanizması olarak kullanılabileceği düşünülmektedir. Ancak kullanılacak veri setleri mümkün olduğunca probleme hizmet edecek şekilde zengin ve çeşitli olmalıdır. Böylelikle model performansları artacak ve analizlerin güvenirlikleri iyileşecektir. Bu da spor kulüplerindeki yöneticilerin ve teknik ekibin kararlarını daha doğru almalarını sağlayarak kulüplere minimum zararı verecek şekilde hareket etmelerine yardımcı olacaktır. Transfer süreçlerinde bonservis bedelinin belirlenmesinde piyasa değeri ve tahmin modelleri kullanılabilir. Ancak piyasa değerinin spekülatif olabileceği ve performansa dayalı olmayabileceği unutulmamalıdır. Çalışma içerisinde değişken önem grafiklerinde performans istatistiklerinin alt sıralarda yer alması TM piyasa değeri hesaplama yönteminde performansın etkisinin düşük olduğu yönünde sinyaller vermektedir. Bu yüzden piyasa değeri tahminleme çalışmalarında TM dışında kuruluşlardan elde edilen piyasa değerleri ile de çalışmalar yapılmalıdır. Alternatif bir yöntem olarak piyasa değerleri bonservis bedelleri üzerinden kurulan makine öğrenmesi algoritmalarıyla da belirlenebilir. Ayrıca TM piyasa değeri, modelinin tahmini ve bonservis bedeli arasında bulunan ilişki makine öğrenmesi yöntemlerinin bu alanda gelişime açık ve fayda sağlayabileceğini de göstermektedir.

KAYNAKÇA

[1] C. Hayashi, K. Yajima, H.H. Bock, N. Ohsumi, Y. Tanaka and Y. Baba, “Data Science, Classification, and Related Methods”, Proceedings of the Fifth Conference of the International Federation of Classification Societies (IFCS-96), Kobe, Japan, March 27–30, 1996.

[2] P. Bruce, A. Bruce, and P. Gedeck, “Practical statistics for data scientists: 50+ essential concepts using R and Python”, O'Reilly Media, 2020.

[3] J.W. Tukey, “The future of data analysis”, The annals of mathematical statistics, 33(1), 1-67, 1962.

[4] Forbes, https://bit.ly/3HZ0lh1, 30.04.2022.

[5] A. L. Samuel, “Some studies in machine learning using the game of checkers”. II—Recent progress, IBM Journal of research and development, 11(6), 601-617, 1967.

[6] A.L. Fradkov, “Early history of machine learning”, IFAC-PapersOnLine, 53(2), 1385-1390, 2020.

[7] Forbes, https://bit.ly/3A0Ntoz, 30.04.2022.

[8] M. Lewis, “Moneyball: The art of winning an unfair game”, WW Norton & Company, 2004.

[9] IMDB, https://www.imdb.com/title/tt1210166/, 30.04.2022.

[10] Harvard Business School, https://hbs.me/3bsWn49, 30.04.2022.

[11] F. Carmichael, “The player transfer system in soccer”, Chapters, 2006.

[12] O. Müller, A. Simons, and M. Weinmann, “Beyond crowd judgments: Data-driven estimation of market value in association football”, European Journal of Operational Research, 263(2), 611-624, 2017.

[13] A. E. Aydemir, T. Taskaya Temizel, and A. Temizel, “A Machine Learning Ensembling Approach to Predicting Transfer Values”, SN Computer Science, 3(3), 1-14, 2022.

[14] M. Ezzeddine, "Pricing football transfers: determinants, inflation, sustainability, and market impact: finance, economics, and machine learning approaches”, PhD Thesis, Universite Pantheon-Sorbornne-Paris I, 2020.

[15] I. Behravan, and S.M. Razavi, “A novel machine learning method for estimating football players’ value in the transfer market”, Soft Computing, 25(3), 2499-2511, 2021.

[16] L. Richau, F. Follert, M. Frenger, and E. Emrich, “Performance indicators in football: The importance of actual performance for the market value of football players”, Sciamus–Sport und Management, 10, 41-67, 2010.

[17] A. T. Yiğit, B. Samak, and T. Kaya, “Football player value assessment using machine learning techniques”, In International Conference on Intelligent and Fuzzy Systems (pp. 289-297). Springer, Cham, 2019.

[18] Y. Geurkink, J. Boone, S. Verstockt, and J. G. Bourgois, “Machine learning-based identification of the strongest predictive variables of winning and losing in Belgian professional soccer”, Applied Sciences, 11(5), 2378, 2019.

[19] FIFA, https://bit.ly/3OCOg3e, 30.04.2022.

[20] J. He, “An Analysis of Zhan Guo Ce”, Peking University Press. ss. 59-82, 2001.

[21] FIFA, https://bit.ly/3bsWaOp, 30.04.2022.

[22] M.S. Pala, Scout Açısı, İthaki Yayınları ss. 25, 2020.

[23] Football Observatory, https://bit.ly/3A6eOWI, 30.04.2022.

[24] B. Frick, “The football players’ labor market: empirical evidence from the major European leagues”, Scottish Journal of Political Economy, 54 (3), 422–446, 2007.

[25] Alexa, https://bit.ly/3y0B0yS, 30.04.2022.

[26] N. Seeman, “Don't Mistake 'Likes' on Facebook For Real Social Change”, Huffington Post, 2013.

[27] Wikipedia, https://bit.ly/3QKZ9Sp, 30.04.2022.

[28] Football Observatory, https://bit.ly/3bpGzPy, 30.04.2022.

[29] TRT Spor, https://bit.ly/3OkWey1, 30.04.2022.

[30] Transfermarkt, https://bit.ly/3Op3sRV, 30.04.2022.

[31] Webtekno, https://bit.ly/3Nnt2W7, 30.04.2022.

[32] Valamis.com, https://bit.ly/3A5ujOC, 30.04.2022.

[33] E. Alpaydin, “Introduction to machine learning”, MIT press, 2020.

[34] T. Hastie, R. Tibshirani, J.H. Friedman, “The elements of statistical learning: data mining, inference, and prediction (Vol. 2, pp. 1-758)”, 2009.

[35] Google Trends Machine Learning, Artificial Intelligence, Data Science, Deep Learning, https://bit.ly/3tY5hgk, 30.04.2022.

[36] R. Alpar, Uygulamalı Çok Değişkenli İstatistiksel Yöntemler, 4.Baskı, Detay Yayıncılık, Ankara, 2013.

[37] A. M. Legendre, “ Nouvelles méthodes pour la détermination des orbites des comètes”, chez Firmin Didot, libraire pour lew mathematiques, la marine, l'architecture, et les editions stereotypes, rue de Thionville, 1806.

[38] F. Galton, “Presidential address, section H, anthropology”, Report of the British Association for the Advancement of Science, 55, 1206-14, 1885.

[39] Microsoft LGBM, https://lightgbm.readthedocs.io/en/latest/, 30.04.2022.

[40] Microsoft LGBM, https://github.com/microsoft/LightGBM, 30.04.2022.

[41] Github LGBM ML Challenge Winning Solutions, https://bit.ly/3NqPVaV, 30.04.2022.

[42] XGBoost, https://xgboost.readthedocs.io/en/stable/, 30.04.2022.

[43] Github XGBoost ML Challenge Winning Solutions, https://bit.ly/3HYZfSr, 30.04.2022.

[44] G. S. Linoff, and M. J. Berry, "Data mining techniques: for marketing, sales, and customer relationship management", IN: Wiley Publishing, Indianapolis, 2011.

[45] L. Breiman, "Machine Learning", Kluwer Academic Publishers Hingham, 45(1), 5-32, 2001.

[46] T. K. Ho, "The Random Subspace Method for Constructing Decision Forests", IEEE Transaction on Pattern Analysis and Machine Intelligence, 20(8), 832-844, 1998.

[47] I. H. Witten, E. Frank, and M. A. Hall, "Data Mining: Practical Machine Learning Tools and Techniques", Second Edition, Elsevier, 2011.

[48] S. M. Lundberg, and S. I. Lee, "A unified approach to interpreting model predictions", Advances in neural information processing systems, 30, 2017.

[49] G. A. Tataroğlu, G. Özbulak, and K. K. Eren, "Determination of the Genetic Variant Reliability Using SHAP Approach", In 2020 28th Signal Processing and Communications Applications Conference (SIU) (pp. 1-4), IEEE, 2020.

[50] SHAP, https://shap.readthedocs.io/en/latest/index.html, 30.04.2022.

[51] Python, https://docs.python.org/3/faq/general.html, 30.04.2022.

[52] TIOBE, https://www.tiobe.com/tiobe-index/, 30.04.2022.

[53] Zhao, B. (2017). Web scraping. Encyclopedia of big data, 1-3.

[54] Transfermarkt, https://bit.ly/3A4LRu7, 30.04.2022.

[55] Football Reference, https://bit.ly/39U4Z3o, 30.04.2022.

[56] C. V. G. Zelaya, "Towards explaining the effects of data preprocessing on machine learning", In 2019 IEEE 35th international conference on data engineering (ICDE) (pp. 2086-2090), IEEE, 2019.

A
EK

Tablo A.1 Transfermark ve Football Reference web sitelerinden toplanan çeşitli veriler.

Veri Kaynağı	Değişken Kategorisi	Değişken	Değişken Açıklaması
Transfermarkt	Sakatlık ve Kaçırılan Maç Bilgisi	Sakat olduğu gün sayısı	Toplam sakat olduğu gün sayısı
		Kaçırdığı maç sayısı	Toplam kaçırdığı maç sayısı
	Piyasa Değeri	Piyasa Değeri	Oyuncunun belli tarihlerdeki piyasa değerini verir
		Tarih	Piyasa değerinin Transfermarkt'ta ne zaman güncellendiği
		Oynadığı Takım	Oyuncunun oynadığı takım bilgisi
	Transfer Geçmişi	Bonservis Değeri	Oyuncuya verilen toplam ücret
		Veren Kulüp	Oyuncuyu satan kulüp
		Alan Kulüp	Oyuncuyu satın alan kulüp
		Transfer Durumu	Oyuncunun hangi tür transfer olduğu bilgisini verir
	Cezalı Olma Durumu	Ceza Durumu	Oyuncunun hangi tür cezalı olduğu bilgisi
		Başlangıç Tarihi	Cezanın başladığı tarih
		Bitiş Tarihi	Cezanın bittiği tarih
		Cezalı Olduğu Gün Sayısı	Toplam cezalı olduğu gün sayısı
		Kaçırdığı Maç Sayısı	Toplam kaçırdığı maç sayısı
	Futbolcu Kişisel Bilgileri	Boy	Oyuncunun boyu (cm)
		Pozisyon	Hangi mevkide oynuyor
		Ayak	Kullandığı ayak
Football Reference	Popülerlik	Haber Sayısı	Tüm sezonlar için FBRef içerisinde bulunan haberlerin frekansı
	Futbolcu Kişisel Bilgileri	Yaşı	Oyuncunun kaç yaşında olduğu bilgisi
		Boy	Oyuncunun boyu (cm)
		Kilo	Oyuncunun kilosu (kg)
		Pozisyon	Hangi mevkide oynuyor
		Ayak	Kullandığı ayak
		Kullandığı Ayak Yeteneği	Kullandığı ayağı kullanma oranı
		Uluslararası Tanınırlık	Oyuncunun başarıları

Tablo A.2 Football Reference performans istatistikleri

Değişken Kategorisi	Değişken Alt Kategorisi	Değişken İsmi	Değişken Açıklaması (TR)
Genel	#	#	Forma numarası
		Nation	Oyuncunun milliyeti
	Pozisyon (Pos)	GK	Kaleciler
		DF	Defans
		MF	Orta saha
		FW	Forvet
		FB	Bekler
		LB	Sol Bekler
		RB	Sağ Bekler
		CB	Orta Sırtlar
		DM	Defansif orta saha
		CM	Merkez Orta Saha
		WM	Geniş Orta Saha
		LW	Sol Kanatlar
		RW	Sağ Kanatlar
		AM	Ofansif orta saha
Oyuncu İstatistikleri (Player Stats)	Oynama Süresi (Playing Time)	Min	Dakika
	Performans (Performance)	Gls	Goller (Atılan veya izin verilen goller)
		Ast	Asist
		PK	Penaltı Atışları Yapıldı
		Pkatt	Penaltı Atışları Denendi
		Sh	Toplam Şut
		SoT	Kaleye yapılan şutlar (Not: Kaleye yapılan atışlar penaltı vuruşlarını içermez)
		CrdY	Sarı Kartlar
		CrdR	Kırmızı kartlar
		Touches	Bir oyuncunun topa dokunma sayısı. Not: Bir pas almak, ardından top sürmek, ardından pas göndermek tek dokunuş olarak sayılır
		Press	Baskılar (Topu alan, taşıyan veya bırakan rakip oyuncuya baskı uygulama sayısı)
		Tkl	Ele geçirilen oyuncu sayısı
		Int	Durdurmalar
		Blocks	Yolunda durarak topa blok yapma sayısı

Tablo A.2 Football Reference performans istatistikleri (devamı)

Oyuncu İstatistikleri (Player Stats)	Beklenen (Expected)	xG	Beklenen Goller (xG toplamları penaltı vuruşlarını içerir, ancak penaltı atışlarını içermez (aksi belirtilmedikçe). StatsBomb tarafından sağlanır. Altı çizili bir eşleşmenin eksik veri olduğunu, ancak mevcut olduğunda güncelleneceğini gösterir.)
		npxG	Penaltı Dışı Beklenen Hedefler (StatsBomb tarafından sağlanır. Alt çizgi, verilerin eksik olduğu ancak mevcut olduğunda güncelleneceği bir eşleşme olduğunu gösterir.)
		xA	xG Destekli
		xG	Şut asist yapan bir pasın ardından gelen pas (StatsBomb tarafından sağlanır. Alt çizgi, eksik veri olan bir eşleşme olduğunu, ancak uygun olduğunda güncelleneceğini gösterir.)
	Şut Yaratma Aksiyonları (SCA)	SCA	Şut Yaratma Hareketleri Paslar, driplingler ve beraberlik faulleri gibi doğrudan şutla sonuçlanan iki hücum hareketi. Not: Tek bir oyuncu birden fazla eylem için kredi alabilir ve şut atan oyuncu da kredi alabilir.
		GCA	Gol Yaratma Hareketleri Paslar, driplingler ve beraberlik faulleri gibi doğrudan hedefe giden iki hücum hareketi. Not: Tek bir oyuncu birden fazla eylem için kredi alabilir ve şut atan oyuncu da kredi alabilir.
	Paslaşma (Passes)	Cmp	Geçişler Tamamlandı
		Att	Geçilmeye çalışıldı
		Cmp%	Geçiş Tamamlama Yüzdesi Lider olmaya hak kazanmak için takım başına en az 30 dakika oynandı
		PrgDist	Aşamalı Mesafe Tamamlanan pasların rakibin kalesine doğru katedilen yarda cinsinden toplam mesafe. Not: Rakibin kalesinden yapılan paslar, sıfır ilerlemeli yard olarak sayılır.
	Top Taşıma (Carries)	Carries	Oyuncunun topu ayaklarıyla kontrol etme sayısı
		PrgDist	Aşamalı Mesafe Toplam mesafe, yarda, bir oyuncu topu kontrol ederken topu rakibinin kalesine doğru ayaklarıyla hareket ettirdi
	Top Sürme (Dribbles)	Succ	Driplingler Başarıyla Tamamlandı
		Att	Dripling denemesi
Paslaşma (Passing)	Toplam (Total)	Cmp	Geçişler Tamamlandı
		Att	Geçilmeye çalışıldı
		Cmp%	Geçiş Tamamlama Yüzdesi Lider olmaya hak kazanmak için takım başına en az 30 dakika oynandı
		TotDist	Tamamlanan pasların herhangi bir yönde kat ettiği toplam mesafe, yard olarak
		PrgDist	Aşamalı Mesafe Tamamlanan pasların rakibin kalesine doğru katedilen yarda cinsinden toplam mesafe. Not: Rakibin kalesinden yapılan paslar, sıfır ilerlemeli yard olarak sayılır.
	Kısa (Short)	Cmp	Geçişler Tamamlandı
		Att	5 yarddan daha kısa olan Pas Denemelerini Geçer

Tablo A.2 Football Reference performans istatistikleri (devamı)

Paslaşma (Passing)	Orta (Medium)	Cmp	Paslar Tamamlanan Paslar 5 ila 25 yard arasında
		Att	5 ve 25 yarda arasında denenen pasları geçer
		Cmp%	Pas Tamamlama Yüzdesi 5 ve 25 yard arasında paslar Lider olarak kalifiye olmak için takım oyunu başına minimum 30 dakika oynanır
	Uzun (Long)	Cmp	Paslar Tamamlandı 25 yarddan uzun paslar
		Att	Geçer Denenmiş Paslar 25 yarddan daha uzun
		Cmp%	Pas Tamamlama Yüzdesi Lider olmak için takım oyunu başına minimum 30 dakika oynanan 25 yarddan uzun paslar
	Performans (Performance)	Ast	Asist
		xA	xG Destekli
		xG	Bu, StatsBomb tarafından sağlanan bir şuta asist yapan bir pası takip eder. Alt çizgi, verilerin eksik olduğu ancak mevcut olduğunda güncelleneceği bir eşleşme olduğunu gösterir.
		KP	Önemli Paslar Doğrudan şutla sonuçlanan paslar (destekli atışlar)
		1/3	Hedefe en yakın olan sahanın 1/3'üne giren Son Üçüncü Tamamlanan paslara paslar Duran toplar dahil değil
		PPA	Penaltı Alanına Geçer Tamamlanan paslar 18 yarda alanına geçer Sabit parçalar dahil değildir
		CrsPA	Ceza Alanına girer Tamamlanan 18 yarda kutusuna geçer Sabit parçalar dahil değildir
		Prog	Aşamalı Paslar Son altı pasta en uzak noktasından en az 10 yarda topu rakibin kalesine doğru hareket ettiren tamamlanmış paslar veya tamamlanmış paslar ceza alanına girer. Sahanın %40'ını savunan pasları hariç tutar
Pas Tipleri (Passing Types)	Toplam	Att	Geçilmeye çalışıldı
	Pas Tipleri (Pass Types)	Live	Canlı top pasları
		Dead	Ölü top pasları Serbest vuruşları, köşe vuruşlarını, başlama vuruşlarını, taç atışlarını ve kale vuruşlarını içerir
		FK	Serbest vuruşlardan pas girişimi
		TB	Aradan Toplar Tamamlandı pas, arka defans oyuncuları arasında açık alana gönderildi
		Press	Rakibin baskısı altındayken yapılan paslar
		Sw	Saha genişliğinin 40 yardından daha fazla seyahat eden Geçişler
		Crs	Orta
		CK	Köşe vuruşları
	Kornerler (Corner Kicks)	In	İlerleyen Köşe Vuruşları
		Out	Dışa Açılan Köşe Vuruşları
		Str	Düz Köşe Vuruşları
	Yükseklik (Height)	Ground	Zemin geçişleri
		Low	Alçak paslar Yerden ayrılan ancak omuz seviyesinin altında kalan paslar
		High	Yüksek paslar Zirve yükseklikte omuz seviyesinin üzerinde olan paslar

Tablo A.2 Football Reference performans istatistikleri (devamı)

Pas Tipleri (Passing Types)	Vücut Kısımları (Body Parts)	Left	Sol ayağını kullanarak pas girişiminde bulundu
		Right	Sağ ayağını kullanarak pas girişiminde bulundu
		Head	Kafa kullanılarak paslar denendi
		TI	Alınan taçlar
		Other	Oyuncunun kafası veya ayakları dışındaki vücut kısımlarını kullanmaya teşebbüs eden paslar
	Sonuç (Outcomes)	Cmp	Geçişler Tamamlandı
		Off	Ofsaytlar
		Out	Sınırların dışında
		Int	Ele geçirilen oyuncu sayısı
		Blocks	Engellenen Paslar Yolun üzerinde duran rakip tarafından engellendi
Defensive Actions	Müdaheleler (Tackles)	Tkl	Ele geçirilen oyuncu sayısı
		TklW	Müdahaleler Kazanılan Müdahaleyi yapanın takımının topa sahip olduğu müsabakalar
		Def 3rd	Savunmada mücadele ediyor 1/3
		Mid 3rd	Ortadaki mücadeleler 1/3
		Att 3rd	1/3 hücumda mücadele ediyor
	İkili Mücadeleli Top Sürme (Vs Dribbles)	Tkl	Mücadele edilen top sürme sayısı
		Att	Top Sürme Yarışması Top sürme sayısı artı müdahale sayısı
		Tkl%	Topla oynanan driplingcilerin yüzdesi Toplanan top sürenlerin sayısı bölünenlere bölünür artı dripling yapılan süreler lider olarak kalifiye olmak için takım başına mücadele edilen minimum .625 driplingci
		Past	Rakip bir oyuncu tarafından dripling yapma sayısı
	Baskılar (Pressures)	Press	Baskılar Topu alan, taşıyan veya bırakan rakip oyuncuya baskı uygulama sayısı
		Succ	Başarılı Baskılar Takımın baskı uyguladıktan sonraki beş saniye içinde ele geçirme sayısı
		%	Başarılı Basınç Yüzdesi Takımın baskı uyguladıktan sonraki beş saniye içinde sahip olduğu zamanın yüzdesi Lider olarak kalifiye olmak için takım oyunu başına minimum 6,44 baskı
		Def 3rd	Savunmada topu alan, taşıyan veya bırakan rakip oyuncuya baskı uygulama sayısı 1/3
		Mid 3rd	Ortada topu alan, taşıyan veya bırakan rakip oyuncuya baskı uygulama sayısı 1/3
		Att 3rd	Hücumda topu alan, taşıyan veya bırakan rakip oyuncuya baskı uygulama sayısı 1/3
	Engellemeler (Blocks)	Blocks	Yolunda durarak topa blok yapma sayısı
		Sh	Yolunda durarak bir şutu engelleme sayısı
		ShSv	Hedefe ulaşan bir şutu yolunda durarak engelleme sayısı
		Pass	Yolunda durarak bir geçişi engelleme sayısı
		Int	Durdurmalar
		Tkl+Int	Ele alınan oyuncu sayısı artı müdahale sayısı

Tablo A.2 Football Reference performans istatistikleri (devamı)

Defensive Actions	Engellemeler (Blocks)	Clr	Boşluklar
		Err	Hatalar (Rakibin şutuna neden olan hatalar)
Topla Oynama (Possessions)	Topla Buluşma (Touches)	Touches	Bir oyuncunun topa dokunma sayısı. Not: Bir pas almak, ardından top sürmek, ardından pas göndermek tek dokunuş olarak sayılır
		Def Pen	Savunma ceza sahasındaki dokunuşlar
		Def 3rd	Savunmada dokunuşlar 1/3
		Mid 3rd	Ortadaki dokunuşlar 1/3
		Att 3rd	Hücumda dokunuşlar 1/3
		Att Pen	Hücum ceza sahasındaki dokunuşlar
		Live	Canlı top dokunuşları. Köşe vuruşlarını, serbest vuruşları, taç atışlarını, başlama vuruşlarını, kale vuruşlarını veya penaltı vuruşlarını içermez
	Top Sürme (Dribbles)	Succ	Driplingler Başarıyla Tamamlandı
		Att	Dripling denemesi
		Succ%	Başarıyla Tamamlanan Dribble Yüzdesi Lider olmak için takım başına minimum 0,5 dripling
		#Pl	Geçmişte Top Sürülen Oyuncu Sayısı
		Megs	Nutmegs Bir oyuncunun topu rakip oyuncunun bacaklarına sürme sayısı
	Top Taşıma (Carries)	Carries	Oyuncunun topu ayaklarıyla kontrol etme sayısı
		TotDist	Toplam mesafe, yard olarak, bir oyuncu topu herhangi bir yönde ayaklarıyla kontrol ederken topu hareket ettirdi
		PrgDist	Aşamalı Mesafe Toplam mesafe, yarda, bir oyuncu topu kontrol ederken topu rakibinin kalesine doğru ayaklarıyla hareket ettirdi
	Alınan Pas (Receiving)	Targ	Pas Hedefleri Bir oyuncunun pas girişiminin hedefi olma sayısı
		Rec	Alınan Paslar Bir oyuncunun başarılı bir şekilde pas alma sayısı
		Rec%	Alınan Pas Yüzdesi Bir oyuncunun başarılı bir şekilde pas aldığı sürenin yüzdesi Lider olarak kalifiye olmak için takım oyunu başına minimum 30 dakika oynadı
		Miscon	Yanlış Kontroller Bir oyuncunun bir topun kontrolünü ele geçirmeye çalışırken başarısız olma sayısı
		Dispos	Dispossessed Bir oyuncunun, rakip bir oyuncu tarafından müdahale edildikten sonra topun kontrolünü kaybetme sayısı. Top sürme girişimini içermez
Çeşitli İstatistikler (Miscellaneous Stats)	Performans (Performance)	CrdY	Sarı Kartlar
		CrdR	Kırmızı kartlar
		2CrdY	İkinci Sarı Kart
		Fls	Yapılan Faulller

Tablo A.2 Football Reference performans istatistikleri (devamı)

Çeşitli İstatistikler (Miscellaneous Stats)	Performans (Performance)	Fld	Alınan Fauller
		Off	Ofsaytlar
		Crs	Ortalar
		Int	Durdurmalar
		TklW	Müdahaleler Kazanılan Müdahaleyi yapanın takımının topa sahip olduğu müsabakalar
		Pkwon	Penaltı Atışları Kazanıldı
		Pkcon	Penaltı Atışları Kabul Edildi
		OG	Kendi kalesine attığı gol
		Recov	Toplanan top sayısı
	Hava Topu Mücadeleleri (Aerial Duels)	Won	Hava topu mücadelesi kazandı
		Lost	Hava topu mücadelesi kaybetti
		Won%	Bir lider olarak kalifiye olmak için takım oyunu başına minimum .97 hava düellosu kazanan antenlerin yüzdesi

TEZDEN ÜRETİLMİŞ YAYINLAR

Konferans Bildirileri

1. H. E. Bayar, E. Çene, “Futbolcuların Piyasa Değerinde Performans Her Şey Midir? Bir Makine Öğrenimi Yaklaşımı”, Uluslararası Bilimsel Gelişmeler Kongresi - ICONSAD21, 2021. ISBN: 978-605-74234-9-8