Bölüm 6

Banka Sektöründe Otomatik Makine Öğrenme Yöntemi (DataBlender) ile Vadeli Mevduat Talep Tahmini

Gizem Aydın[1]

Mehmet Yalçın[2]

Özet

Günümüzde şirketlerin, müşteri havuzlarından doğru kişilere ulaşarak ürünleri ve hizmetleri pazarlaması oldukça önemli hale gelmiştir. Bu durum, bankacılık ve finans kuruluşları için de geçerlidir, çünkü büyük müşteri kitlesi içinden doğru kişilere ulaşmak maliyetli ve zaman alıcı bir süreç olabilmektedir. Bu doğrultuda kurumlar, ürün ve hizmetleri satın alan müşterileri içerisinden demografik ve finansal verilerini kullanarak makine öğrenmesi tabanlı modellerini arttırmışlardır. Bu modellerin hem doğru kişileri tespit ederek başarılı sonuçlar çıkarması hem de zaman ve para maliyetini azaltması bu teknolojiye olan ilgiyi arttırmıştır. Makine öğrenmesi tabanlı modeller bu doğrultuda sıklıkla kullanım imkânı bulmaktadır.

Bu çalışmada, Portekiz'deki finans kuruluşunu tercih eden müşterilerin vadeli mevduatı alıp almayacağı üzerine tahmin modelleri kurulmuştur. Böylece kuruluş, vadeli mevduata abone olma şansı daha yüksek olan mevcut müşterileri belirleyecek ve pazarlama çabaları ile bu müşterilere odaklanacaktır. Çalışmada; Portekiz'deki finans kuruluşunun, müşterilerini arayarak vadeli mevduat hesabı isteyip istemediği bilgisinin tutulduğu veri seti kullanılmıştır. Veri ön işleme adımları ile modelleme işlemleri otomatik bir şekilde yapılmış ve bunun için "DataBlender" otomatik makine öğrenmesi uygulaması kullanılmıştır. Çalışma sonucunda modeller içerisinden "LightGBM" modeli en başarılı model çıkmıştır. Sonuçlar neticesinde ağaç temelli modellerin başarı ölçütlerinin birbirine yakın olduğu gözlenmiştir. Vadeli mevduat hesabı isteyip istemediği hedef değişkeni üzerindeki etkili değişkenler, model sonucunda bulunarak müşteriye ait özellikler sıralanmıştır.

1 Veri Bilimci, ERETEAM, gaydin@ereteam.com, 0000-0001-6353-0648

2 Yıldız Teknik Üniversitesi, myalcinmh@gmail.com, 0000-0002-8162-6085

https://doi.org/10.58830/ozgur.pub518.c2128

1. Giriş

Günümüzde, hızla değişen ve gelişen teknolojik ortam işletmelerin pazarlama stratejilerini ve müşteri odaklı yaklaşımlarını temelden değiştirmektedir. Özellikle bankacılık ve finans sektörü, dijitalleşme ve teknolojik ilerlemelerin etkisiyle müşterilere daha özelleştirilmiş ve hedef odaklı hizmetler sunmaya çalışmaktadır. Bu sektörlerdeki kurumlar için, müşterilerle daha derin ve anlamlı bağlantılar kurmak, doğru zamanda doğru mesajları iletmek ve uygun ürün ve hizmetleri sunmak giderek önem kazanmaktadır. Müşteri havuzundan doğru kişilere ulaşmak ve onların gereksinimlerini anlamak, finansal kurumlar için kritik bir görev haline gelmiştir. Ancak, geniş müşteri tabanları arasından doğru hedef kitlenin belirlenmesi ve bu kişilere uygun ürün veya hizmetleri sunmak, zaman alıcı ve maliyetli bir süreç olabilir. Bu nedenle, finans kuruluşları ve benzeri sektörler, veri odaklı yaklaşımlarla müşteri davranışlarını analiz etmeye ve makine öğrenmesi gibi teknolojileri kullanarak daha etkin bir hedefleme stratejisi geliştirmeye yönelik çabalarını artırmaktadır (Moro, Laureano, & Cortez, 2011; Moro vd., 2014).

Makine öğrenmesi, finans sektöründe önemli bir araç haline gelmiştir. Müşteri profilleri, finansal veriler ve dijital izler üzerinde yapılan analizler sayesinde, vadeli mevduat hesaplarından kredi riskine kadar geniş bir yelpazede modeller geliştirilebilmektedir. Özellikle, vadeli mevduat ürünlerine olan ilgiyi öngörmek ve bu alanda müşteri hedefleme stratejilerini iyileştirmek için makine öğrenmesi tabanlı modellerin kullanımı artmaktadır (A.Elsalamony, 2014).

Literatürdeki önemli çalışmalarla birlikte değerlendirildiğinde, finans sektöründe makine öğrenmesi tabanlı modellerin kullanımının çeşitliliği ve başarı potansiyeli hakkında kapsamlı bir anlayış sunmayı amaçlamaktadır. Ghatasheh ve arkadaşlarının yaptığı çalışma, dengesiz veri setlerinin etkisini hafifletmek ve müşterilerin vadeli mevduat hesabı açıp açmayacağını öngörmek amacıyla çok katmanlı bir algılayıcı (MLP) sınıflandırıcı kullanarak derin öğrenme uygulamıştır (Ghatasheh vd., 2020). Okur ve Çetin'in çalışması da Microsoft Azure makine öğrenmesi platformunda regresyon ve ikili sınıflandırma türünde makine öğrenmesi algoritmaları kullanarak kredi risk değerlendirmesi yapmıştır. Bu çalışma, büyük veri setlerinin etkin bir şekilde analiz edilebileceğini ve kredi riski tahmininde makine öğrenmesinin etkili bir araç olduğunu vurgulamıştır (Okur & Cetin, 2019). Coşkun ve Turanlı'nın yaptığı çalışma da ise kullanıcıların kredi riski değerlendirmesi için CatBoost, XGBoost ve LightGBM gibi topluluk öğrenme yöntemlerini karşılaştırarak, kredi skorlamasında makine öğrenmesi tabanlı modellerin etkinliğini ortaya koymaktadır (Coşkun & Turanli, 2023).

Bu çalışmada, Portekiz'deki bir finans kuruluşunun müşterilerinin vadeli mevduat ürünlerini tercih etme olasılığını öngörmek amacıyla makine öğrenmesi tabanlı tahmin modelleri oluşturulmuştur. Çalışma, kurumun mevcut müşteri tabanında vadeli mevduata abone olma olasılığı yüksek olan müşterileri tespit ederek pazarlama çabalarını optimize etmeyi hedeflemektedir. Veri seti üzerinde yapılan veri ön işleme ve modelleme işlemleri, DataBlender otomatik makine öğrenimi aracıyla gerçekleştirilmiştir.

Çalışmanın sonucunda, en başarılı modelin LightGBM modeli olduğu belirlenmiş ve ağaç tabanlı modellerin benzer başarı kriterlerine sahip olduğu gözlemlenmiştir. Bu modeller, vadeli mevduat hesabı isteme olasılığı üzerinde etkili olan değişkenleri belirlemiş ve müşterilerin özelliklerini sıralamıştır. Bu sonuçlar, finans kuruluşlarının hedefleme stratejilerini iyileştirmek ve pazarlama çabalarını daha etkili bir şekilde yönlendirmek için makine öğrenmesi modellerini kullanmalarını teşvik edebilir. Ayrıca bu çalışmanın DataBlender ile gerçekleştirilen ilk çalışma olması sebebiyle literatüre bu anlamda da katkı sunacaktır.

2. Metodoloji

Firmaların hizmetlerini pazarlamak için doğru kişilere ulaşması önemli bir durumdur. Müşteri tabanından doğru kişilere ulaşmak maliyetli ve zaman alıcı bir süreç olabilmektedir. Bu sebeple makine öğrenimi algoritmalarına olan ilgi ve kaynak artmıştır. Çalışma kapsamında da makine öğrenmesi algoritmaları kullanılarak kullanıcı odaklı olarak kod yazmadan otomatik olarak veri ön işleme ve modelleme yapılmasını sağlayan DataBlender ile problemin çözümü gerçekleştirilmiştir. Metodoloji kapsamında makine öğrenimi yöntemlerinden ve DataBlender uygulamasından faydalanılmıştır.

2.1. Makine Öğrenimi

Makine öğrenimi, bilgisayar sistemlerinin insanlar gibi öğrenme yeteneğini kazanmaları için geliştirilen algoritmalar, modeller ve tekniklerin incelendiği bir bilim dalıdır. Bu alanda, "makine" olarak adlandırılan yazılım ve istatistiksel algoritmalardan oluşan sistemler, verileri analiz ederek öğrenme sürecini gerçekleştirirler (Alpaydın, 2010). Sade bir açıklama ile, geçmiş verilerden yola çıkarak hem gelecekteki olayları tahmin etmek hem de mevcut veri setindeki örüntüleri ve farklılıkları keşfetmek için kullanılırlar. İstatistiksel ve matematiksel modellere dayalı bu süreçler, çeşitli programlama dilleri kullanılarak yürütülür. Makine öğrenimi, esas olarak üç ana kategori altında incelenir: Denetimli öğrenme (Supervised Learning), Denetimsiz öğrenme (Unsupervised Learning) ve Pekiştirmeli öğrenme (Reinforcement Learning). Çalışmada denetimli öğrenme modelleri kullanılmıştır. Denetimli

öğrenme; hem girdi (input) hem de çıktı (output) değerlerinin yer aldığı veri setlerinde girdi ve çıktılar arasındaki ilişkiyi açıklamak için fonksiyonların elde edilmesidir. Sınıflandırma ve Regresyon modelleri denetimli öğrenme modelleri altında yer almaktadır.

2.2. Veri Seti

Çalışma kapsamında kullanılan veriler, UC Irvine Machine Learning Repository'den alınmıştır. Portekizli bir bankacılık kurumunun doğrudan pazarlama kampanyalarına (telefon görüşmeleri) ilişkindir. Veri içerisinde, 41188 gözlem ve 21 değişken içererek kaggle.com üzerinde yetkili kuruluşun izni ile yayınlanmıştır. Veri seti; 11 kategorik ve 10 numerik değişkenden oluşmaktadır. Vadeli Mevduatı İsteme Durumu hedef değişken olarak y değişkeni altında bu bilgi mevcuttur.

Tablo 1. Veri İçerisindeki Değişkenler ve Açıklamaları

Değişken	**Açıklama**
Age	Kullanıcı yaşı
Job	İşin türü
Marital	Medeni durumu
Education	Eğitimi
Default	Kredinin temerrüde düşmesi var mı?
Housing	Konut kredisi var mı?
Loan	Bireysel krediniz var mı?
Contact	İletişim türü
Month	Yılın son iletişim ayı
Day_of_week	Haftanın son iletişim günü
Duration	Son iletişim süresi
Campaign	Bu kampanya sırasında ve bu müşteri için gerçekleştirilen iletişim sayısı
Pdays	Önceki bir kampanyadan müşteriyle son iletişime geçildikten sonra geçen gün sayısı
Previous	Bu kampanyadan önce ve bu müşteri için gerçekleştirilen iletişim sayısı
Poutcome	Önceki pazarlama kampanyasının sonucu
Emp.var.rate	İstihdam değişim oranı
Cons.price.idx	Tüketici fiyat endeksi- aylık gösterge
Cons.conf.idx	Tüketici güven endeksi- aylık gösterge
Euribor3m	Euribor 3 aylık oran- günlük gösterge
Nr.employed	Çalışan sayısı- üç aylık gösterge
y	Müşteri vadeli mevduata abone oldu mu?

Tablo 2'de numerik değişkenlere ait istatistiksel tanımlayıcı bilgiler yer almaktadır. Değişkenlerin değer aralıkları oldukça birbirinden farklılık göstermektedir.

Tablo 2. Numerik Değişkenlerin İstatistiksel Tanımlayıcı Bilgileri

	gözlem sayısı	ortalama	standart sapma	minimum	25%	50%	75%	maximum
age	41188.0	40.024.060	10.421.250	17.000	32.000	38.000	47.000	98.000
duration	41188.0	258.285.010	259.279.249	0	102.000	180.000	319.000	4.918.000
campaign	41188.0	2.567.593	2.770.014	1.000	1.000	2.000	3.000	56.000
pdays	41188.0	962.475.454	186.910.907	0	999.000	999.000	999.000	999.000
previous	41188.0	172.963	494.901	0	0	0	0	7.000
emp.var.rate	41188.0	81.886	1.570.960	-3.400	-1.800	1.100	1.400	1.400
cons.price.idx	41188.0	93.575.664	578.840	92.201	93.075	93.749	93.994	94.767
cons.conf.idx	41188.0	-40.502.600	4.628.198	-50.800	-42.700	-41.800	-36.400	-26.900
euribor3m	41188.0	3.621.291	1.734.447	634	1.344	4.857	4.961	5.045
nr.employed	41188.0	5.167.035.911	72.251.528	4.963.600	5.099.100	5.191.000	5.228.100	5.228.100

Tablo 3'de kategorik değişkenlere ait istatistiksel tanımlayıcı bilgiler yer almaktadır. En çok benzersiz sınıf sayısına sahip job değişkeni olurken en az sınıf sayısına sahip değişken y olmuştur.

Tablo 3. Kategorik Değişkenlerin İstatistiksel Tanımlayıcı Bilgileri

	gözlem sayısı	benzersiz sınıf sayısı	en çok gözlenen sınıf
job	41188	12	admin
marital	41188	4	married
education	41188	8	university.degree
default	41188	3	no
housing	41188	3	yes
loan	41188	3	no
contact	41188	2	cellular
month	41188	10	may
day_of_week	41188	5	thu
poutcome	41188	3	nonexistent
y	41188	2	no

2.3. Metot

Çalışma kapsamında "DataBlender" uygulaması altında makine öğrenmesi modellerinden denetimli öğrenme modelleri uygulanmıştır. Makine öğrenmesi çalışma döngüsü altında yer alan veri ön işleme adımları ve modelleme adımları "DataBlender" üzerinde gerçekleştirilmiştir. Modelleme için "Python Sklearn" kütüphanesi, veri ön işleme için de yine

"Python" altında yer alan kütüphaneler uygulama tarafından otomatik olarak kullanılmıştır. Modelleme için Lojstik Regresyon, Destek Vektör Makineleri (Support Vector Machines, SVM), Torbalama (Bagging) ve Boosting (Hızlandırma) modelleri kullanılmıştır.

2.3.1. Lojistik Regresyon

Lojistik Regresyon, bağımlı ve bağımsız değişkenler arasındaki ilişkiyi inceleyen ve yaygın olarak kullanılan bir istatistiksel analiz yöntemidir. Lojistik Regresyon, bağımlı değişkenin kategorik olması ve sınıflandırma işlemlerinde kullanılması bakımından Doğrusal Regresyondan ayrılır. Bu analiz türü, bağımlı değişkene uygulanan logit dönüşümünden adını alır ve amacı, bağımlı ve bağımsız değişkenler arasındaki ilişkiyi açıklayacak bir model oluşturmaktır. Lojistik Regresyon, Doğrusal Regresyonda gereken varsayımları gerektirmediği için daha esnek bir şekilde kullanılabilir (Mertler & Reinhart, 2005).

Lojistik Regresyon Modeli:

$$L = \left[\frac{P_i}{1 - P_i} \right] = Z_i = b_0 + b_1 X_i + e_i \qquad (1)$$

Burada olma ihtimalini P_i ve olmama ihtimali $1 - P_i$, $\frac{1}{1 + e^{-z}}$ denkliği ile hesaplanır.

$Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots.. + \beta_n X_n$ şeklinde ifade edilir. β'lar regresyon katsayısını göstermektedir (Özer, 2004).

2.3.2. Destek Vektör Makineleri (Support Vector Machines, SVM)

Destek Vektör Makineleri (SVM), hem regresyon hem de sınıflandırma işlemleri için kullanılan popüler bir makine öğrenmesi yöntemidir. Ancak, özellikle sınıflandırma amacıyla daha yaygın olarak tercih edilmektedir. Bu yöntem hem doğrusal hem de doğrusal olmayan veri setlerini etkili bir şekilde sınıflandırabilir. Genellikle, verileri doğrusal olarak sınıflandırmaya odaklanır ve az hesaplama gücüyle yüksek doğruluk oranları sağlaması nedeniyle tercih edilir. Destek Vektör Makineleri'nin temel işlevi, iki sınıfa ait verileri en uygun şekilde ayırmaktır. Bu, karar sınırları veya hiper düzlemler belirleyerek gerçekleştirilir ve sınıfları en iyi şekilde ayıran çizgiyi bulmaya çalışırlar (James vd., 2013).

Destek Vektör Makineleri, iki sınıfa ait verileri ayırmak için etkin bir sınıflandırma sağlamayı hedefler. Bu yöntem, iki sınıf arasındaki en uygun

karar fonksiyonunu, yani hiper düzlemi belirlemeye odaklanır. İki sınıflı verilerin sınıflandırılmasında, mevcut sonsuz sayıdaki hiper düzlemler arasından en etkili ayrımı yapacak olan hiper düzlemi seçmek önemlidir. Bu seçim, sınıflandırma işleminin doğruluğunu ve verimliliğini artırır (Vapnik, 1995).

2.3.3. Rassal Ormanlar (Random Forest)

Rassal orman algoritması makine öğrenmesinde karar ağaçlarını ana model olarak kullanan ve yaygın bir torbalama tekniğidir. Bu yöntemde, rastgele orman, daha önceden belirtilen torbalama prosedürü kullanılarak elde edilen bootstrap örneklerine dayalı bir dizi karar ağacından oluşturulur. Torbalama metodunun etkililiği, temel alınan modellerin çeşitliliğine büyük oranda bağlıdır. Eğer temel modeller arasında yüksek korelasyon varsa, birleşik model muhtemelen benzer sonuçlar üretecektir (Coşkun & Turanli, 2023; Breiman, 2001).

Rassal Ormanlar algoritmasında, Bagging yöntemi gibi ağaçlar için gözlemler, bootstrap ile rastgele örnek seçimi kullanılarak belirlenir. Ancak, Rassal Ormanlar yönteminde farklı olarak, değişkenler arasından Random Subspace (Rastgele Alt Uzay) metodu ile rastgele bir alt grup seçilir. Bu seçilen alt grup, ilgili düğümün dallanmasında kullanılır. Bu rastgele seçim, orman içindeki ağaçların birbirine benzeme olasılığını azaltır. Bu nedenle, Rassal Ormanlar, Bagging metodunun geliştirilmiş bir versiyonu olarak kabul edilir (Breiman, 2001).

Rassal Ormanlar algoritmasının oluşturulması ve uygulanması kolay olması, bu yöntemin temel avantajlarından biridir. Son zamanlarda yapılan araştırmalar, birçok veri yapısı ve problemde karşılaşılan regresyon ve sınıflandırma hatalarının Rassal Ormanlar yöntemi kullanılarak azaltılabileceğini göstermiştir. Ayrıca, daha az parametre içerdiği için bu algoritmanın eğitimi daha hızlıdır. Minimum veri hazırlığı gereksinimiyle büyük veri setlerinde bile etkili ve verimli çalışabilirler. Ayrıca, veri setinde bulunabilecek aykırı gözlemlere ve aşırı öğrenmeye karşı dirençlidirler (Yalçin & Kalkan, 2022;Lewis, 2017).

2.3.4. Gradyan Artırma Makineleri (Gradient Boosting Machines)

Gradyan Artırma, tarafından önerilen ve daha güçlü bir model elde etmek amacıyla birden fazla karar ağacını birleştiren popüler bir ensemble yöntemidir. Bu yöntemde, her bir tahminci, önceki tahmincinin hatalarını etiket olarak kullanarak eğitilir. Gradyan destekli ağaçlar makine öğrenmesinde yaygın olmasına karşın, bu ağaçlar oldukça sığ yapıdadır. Örneğin, ağaçların

derinliği genellikle bir ila beş arasında değişir, bu da modelin bellek açısından daha az yer kaplamasını sağlar ve tahmin süreçlerini hızlandırır. Bu sığ ağaçlar, zayıf öğreniciler olarak görev yapar ve tahmin ediciye daha fazla sığ ağaç eklenerek performans sürekli olarak iyileştirilir(Coşkun & Turanli, 2023;Friedman, 2001).

2.3.5. Ekstrem Gradyan Artırma (Extreme Gradient Boost, XGBoost)

Ekstrem Gradyan Artırma (XGBoost) algoritması, performans ve işlem hızı açısından Gradyan Artırma algoritmasının ileri bir versiyonudur. Büyük veri setlerini içeren makine öğrenmesi problemlerinde kullanılmak üzere tasarlanmıştır. Tianqi Chen tarafından 2014 yılında geliştirilen bu algoritma, açık kaynak kodlu makine öğrenmesi algoritmalarının yer aldığı DMLC (Dağıtık Makine Öğrenimi Topluluğu) kütüphanesine eklenmiştir(Chen & Guestrin, 2016).

XGBoost, milyonlarca veri setini kısa sürede, diğer algoritmalara kıyasla daha az işlemci gücü ve geçici bellek (RAM) kullanarak analiz edebilir. Günümüzde veri miktarının artması ve veri analizi için gerekli işlem gücünün maliyetlerini azaltma ihtiyacı nedeniyle bu modele olan ilgi artmıştır. XGBoost, veriyi tamamen incelemek yerine parçalara ayırarak işler ve bu yüzden Gradyan Artırma algoritmasına kıyasla daha hızlı bir performans sunar.

2.3.6. Kategorik Artırma (Category Boosting, CatBoost)

CatBoost (Kategorik Artırma), Yandex mühendisleri tarafından geliştirilen ve karar ağaçlarına dayalı bir gradyan artırma algoritmasıdır.

CatBoost, kategorik verileri işleyebilen bir gradyan artırma kütüphanesidir. Kategorik değerlerin ikili yerine konulması yerine, bu yöntem veri setinin rastgele bir permutasyonunu gerçekleştirir ve aynı kategori değerine sahip örneğin verilen sıradan önce yer alan ortalama etiket değerini hesaplar(Jhaveri vd., 2019). Bu, CatBoost'ta tanıtılan ve yeni bir artırma şeması olan sıralı artırma olarak adlandırılan önemli bir ilerlemedir. Bu yöntem, gradyan yanlılığı nedeniyle oluşan tahmin kaymasını aşabilir ve modelin genelleme yeteneğini daha da artırabilir(Coşkun & Turanli, 2023;Zhang vd., 2020).

2.3.7. Hafif Gradyan Arttırma Makinesi (LightGBM)

LightGBM, Microsoft tarafından geliştirilmiş ve XGBoost algoritmasının eğitim süresi performansını artırmak için tasarlanmış hafif bir gradyan artırıcıdır. LightGBM, dört açıdan geliştirilmiş bir XGBoost sürümüdür.

İlk olarak, LightGBM algoritması gradyan tabanlı tek taraflı örnekleme algoritmasını içerir. İkinci olarak, optimal segmentasyon noktasını belirlemek için bir histogram kullanır ve özel özellik paketlemesi yoluyla özelliği belirli bir ölçüde azaltır. Son olarak ise, geleneksel seviye bazında yerine derinlik sınırlaması ile yaprak bazında bir algoritma kullanır, bu da hem doğruluk artışı sağlar hem de aşırı öğrenmeyi önler(Ke vd., 2017).

2.4. Uygulama

Uygulama kapsamında kullanılan veri seti içerisinde, 41188 gözlem ve 21 değişken içererek kaggle.com üzerinde yetkili kuruluşun izni ile yayınlanmıştır. Veri seti; 11 kategorik ve 10 numerik değişkenden oluşmaktadır. Vadeli Mevduatı İsteme Durumu hedef değişken olarak y değişkeni altında bu bilgi mevcuttur.

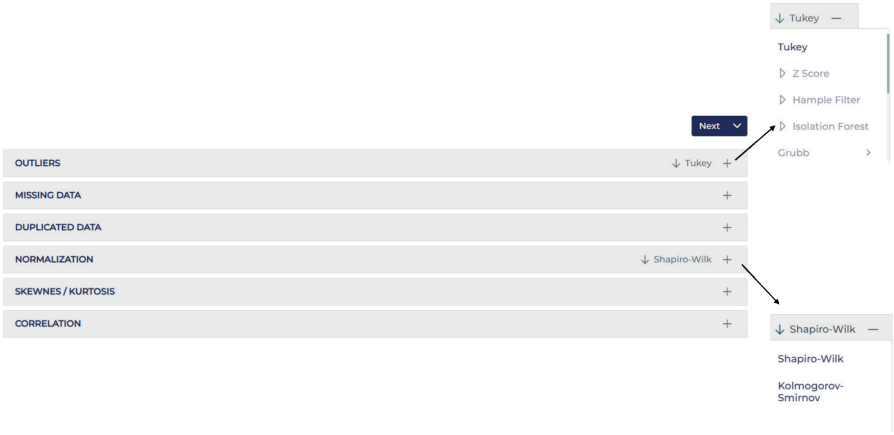
Bağımlı (hedef) değişken olan Vadeli Mevduatı Kabul Etme Durumu (y) incelendiğinde imbalance (dengesiz) olduğu görülmektedir. Vadeli Mevduat isteme durumuna 36548 Hayır (%89) ve 4640 Evet (% 11) diyen müşteri mevcuttur.

DataBlender uygulaması altında makine öğrenmesi modellerinden denetimli öğrenme modelleri uygulanmıştır. Makine öğrenmesi çalışma döngüsü altında yer alan veri ön işleme adımları ve modelleme adımları DataBlender üzerinde gerçekleştirilmiştir. Modelleme için "Python Sklearn" kütüphanesi, veri ön işleme için de yine "Python" altında yer alan kütüphaneler uygulama tarafından otomatik olarak kullanılmıştır.

Modelleme kapsamında, kategorik değişkenler otomatik olarak "Label Encoder" ve "One Hot Encoder" ile sayısallaştırılmışlardır. Veri ön işleme adımları tamamlandıktan sonra model kurma işlemi gerçekleştirilmiştir.

Veri seti, bağımlı değişkeni, Vadeli Mevduatı İsteme Durumu olmak üzere %80 eğitim seti, %20 test seti şeklinde ayrılmıştır. 5'li "cross validation" uygulanmıştır. Çalışma kapsamında kullanılan modeller "DataBlender" altında yer alan "Sklearn" üzerinden alınan sınıflandırma modellerdir. Modeller AUC metriğine göre eğitilmiştir.

DataBlender uygulaması Resim 1'deki gibi veri ön işleme adımlarında kullanılan istatistiksel yöntemlerini sunmaktadır. Kullanıcının kararına göre bu işlemler seçilerek ön işleme adımları hızlı bir şekilde geçilmektedir. Veri ön işleme ve veri hazırlığı adımlarının makine öğrenmesi modellemelerinde oldukça zaman alması göz önüne alındığında, DataBlender uygulamasının bu yardımı oldukça verimli olmaktadır.



Resim 1. "DataBlender" Uygulaması İçerisindeki "Preprocessing" Ara Yüzü

Outlier yöntemlerinden (Tukey, Z-Score, Hampel Filter, Isolation Forest, DBScan, Grubb, Standard Deviation, Median Absolute Deviation, Double Median Absolute Deviation olmak üzere toplam 9 yöntem bulunmaktadır) Tukey seçilmiştir. Tukey yöntemi, veri setindeki sürekli değişkenlerin dağılımı hakkında varsayımlar yapmaksızın medyan, alt ve üst çeyrekler ile alt ve üst uç gibi temel istatistiksel bilgileri görselleştirmek için kullanılan etkili bir araçtır (Seo & Gary M. Marsh, 2006). Bu kural, aykırı değerlerin, veri setinin çeyrekler arası aralığının 1,5 katından fazla olan değerleri tanımlar.

Veri setinde eksik veri bulunmadığından eksik veri yöntemleri içerisinde yer alan Silme, KNN, Simple Imputer, Iterative Imputer, 0 ile doldurma yöntemleri seçilmeden devam edilmiştir.

Normalization sekmesi altında veri setindeki değişkenlerin Shapiro-Wilk ve Kolmogrov Smirnov testine göre normallik testleri yapılmaktadır. Çarpıklık ve Basıklık değerleri gösterilmektedir.

Correlation altında Spearman korelasyon testine göre pozitif ve negatif en yüksek değişken çiftleri gösterilmektedir. Korelasyon katsayısı, iki değişken arasındaki ilişkinin yönünü ve şiddetini ölçen bir istatistiksel değerdir. Bu katsayı, -1 ile +1 arasında bir değer alır. Pozitif değerler, aralarında pozitif yönlü bir doğrusal ilişki bulunan değişkenleri gösterirken, negatif değerler negatif yönlü bir doğrusal ilişkiyi işaret eder. İlişkinin şiddeti, genellikle şu şekilde yorumlanır: 0 ile 0.29 arasındaki korelasyon katsayıları düşük düzeyde ilişkiyi, 0.30 ile 0.70 arasındakiler orta düzeyde ilişkiyi ve 0.71 ile 1 arasındakiler ise yüksek düzeyde ilişkiyi temsil eder (James vd., 2013).

Tablo 4'de numerik değişkenler arasındaki ilişki ve ilişki gücü gösterilmektedir. Emp.var.rate olarak adlandırılan istihdam değişim oranı ve günlük bir gösterge olan Euribor 3 aylık oranı (euribor3m) arasında belirgin bir şekilde yüksek pozitif bir ilişki bulunmaktadır. Bu ilişki, günlük Euribor oranı aracılığıyla temsil edilen kısa vadeli ekonomik değişimlerin, üç aylık istihdam değişim oranı üzerinde önemli bir etkisi olduğunu göstermektedir. Pdays' değişkeni, bir önceki kampanyadan beri müşterinin son temas edilme tarihini ifade eder. 'Previous' ise bu kampanyadan önce ve bu müşteri için yapılan temasların sayısını gösterir. Bu iki değişken arasında belirgin bir şekilde negatif bir ilişki bulunmaktadır. Bu güçlü negatif ilişki, önceki temas sayısı arttıkça son temasın üzerinden geçen sürenin azaldığını gösterir, bu da kampanyalar arasında daha sık ve kısa süreli etkileşim olasılığını işaret eder.

Tablo 4. Değişkenler Arasındaki Pozitif ve Negatif Yönlü Korelasyon Tabloları

Pozitif Yönlü En Yüksek Korelasyon Listesi			Negatif Yönlü En Yüksek Korelasyon Listesi		
degisken1	**degisken2**	**korelasyon**	**degisken1**	**degisken2**	**korelasyon**
emp.var.rate	euribor3m	0.972245	pdays	previous	-0.587514
euribor3m	nr.employed	0.945154	previous	nr.employed	-0.501333
emp.var.rate	nr.employed	0.906970	previous	euribor3m	-0.454494
emp.var.rate	cons.price.idx	0.775334	previous	emp.var.rate	-0.420489
cons.price.idx	euribor3m	0.688230	previous	cons.price.idx	-0.203130
cons.price.idx	nr.employed	0.522034	pdays	cons.conf.idx	-0.091342
pdays	nr.employed	0.372605	campaign	previous	-0.079141
pdays	euribor3m	0.296899	duration	campaign	-0.071699
cons.price.idx	euribor3m	0.277686	previous	cons.conf.idx	-0.050936
pdays	emp.var.rate	0.271004	duration	pdays	-0.047577

3. Bulgular

Sınıflandırma modellerinin değerlendirme kriterleri arasında genellikle Karmaşıklık Matrisi (Confusion Matrix) ve ROC Eğrisi değerleri kullanılmaktadır(Han vd., 2012).

Tablo 5. İki Sınıflı Karmaşıklık Matrisi

Karmaşıklık Matrisi (Confusion Matrix)		*Tahmin Edilen Sınıf*	
		0	*1*
Gerçek Sınıf	*0*	Doğru Negatif (DN)	Yanlış Negatif (YN)
	1	Yanlış Pozitif (YP)	Doğru Pozitif (DP)

- Doğru Pozitif (DP) durumu, bir sınıflandırıcının pozitif olarak sınıflandırılması gereken verilerden kaç tanesini doğru şekilde pozitif olarak sınıflandırdığını gösterir.
- Doğru Negatif (DN) ise, negatif sınıfa ait verilerin kaç tanesinin sınıflandırıcı tarafından doğru şekilde negatif olarak sınıflandırıldığını ifade eder.
- Yanlış Negatif (YN) durumu, gerçekte pozitif sınıfa ait olan ancak sınıflandırıcı tarafından yanlışlıkla negatif sınıf olarak işaretlenen verileri tanımlar.

Yanlış Pozitif (YP), gerçekte negatif sınıfa ait bir verinin, sınıflandırıcı tarafından yanlışlıkla pozitif sınıf olarak etiketlenmesi durumudur (Özdemir, 2021; ALAN & KARABATAK, 2020).

Modellerin karşılaştırılması için Karmaşıklık Matrisi'nden elde edilen değerler ile Doğruluk, Kesinlik, Hassasiyet, F1 skoru, AUC, Kappa, MCC metrikleri göz önünde bulundurulmuştur.

- Doğruluk (Accuracy): Tahminlerin genel doğruluk oranıdır.

$$Doðruluk = \frac{DP + DN}{DP + DN + YP + YN} \tag{2}$$

- Kesinlik (Precision): Pozitif olarak tahmin edilen örneklerin gerçekte pozitif olma oranıdır.

$$Kesinlik = \frac{DP}{DP + YP} \tag{3}$$

- Hassasiyet (Recall): Gerçekte pozitif olan örneklerin pozitif olarak tahmin edilme oranıdır.

$$Hassasiyet = \frac{DP}{DP + YN} \tag{4}$$

- F1 Ölçütü: Kesinlik ve Hassasiyet değerlerinin harmonik ortalamasıdır. Hedef değişkenin dengesiz olduğu durumlarda özellikle incelenen bir ölçüttür.

$$F1\ Ölçütü = \frac{2*K*G}{K+G} \quad (5)$$

- Eğri Altında Kalan Alan (Area Under the Curve, AUC): Pozitif örnekleri negatif örneklerden ayırma yeteneğini ölçer. ROC eğrisinin altında kalan alan değeridir.
- KAPPA Ölçütü: Accuracy'nin beklenen doğruluk oranından sapmasını ölçer.
- Matthews Korelasyon Katsayısı (Matthews Correlation Coefficient, MCC): Gerçek sınıflandırmanın tahmin edilen sınıflandırmaya yakınlığını ölçer (Burkov, 2019)

"DataBlender" ile birden fazla model kullanıcı kararıyla oluşturulabilmektedir. Çalışma kapsamında ağaç temelli modeller ve klasik modeller seçilerek performansları kıyaslanmıştır. Oluşturulan modellerin hata metrikleri aşağıdaki tabloda gösterilmektedir.

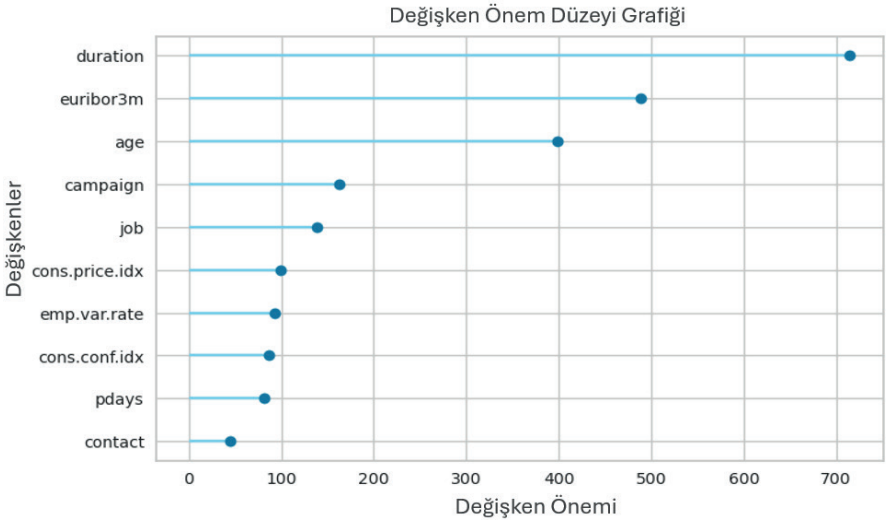
Tablo 6. Model Performansları

Model	Doğruluk	Eğri Altında Kalan Alan	Hassasiyet	Kesinlik	F1 Ölçütü	Kappa Ölçütü	MCC Ölçütü
Light Gradient Boosting Machine	0.9179	0.9502	0.5493	0.6653	0.6010	0.5558	0.5594
Catboost	0.9161	0.9492	0.5576	0.6495	0.5995	0.5530	0.5553
Gradient Boosting Machine	0.9165	0.9464	0.5339	0.6604	0.5900	0.5441	0.5482
Extreme Gradient Boosting(XgBoost)	0.9129	0.9460	0.5425	0.6332	0.5837	0.5355	0.5378
Rassal Ormanlar (Random Forest)	0.9135	0.9410	0.4701	0.6644	0.5502	0.5039	0.5134
Lojistik Regresyon	0.9105	0.9328	0.4129	0.6671	0.5093	0.4631	0.4797
Destek Vektör Makineleri	0.8207	0.0000	0.4758	0.5421	0.3987	0.3372	0.3710

Topluluk öğrenme modelleri (Bagging ve Boosting), genel olarak başarılı sonuçlar vermiştir. Tablodaki performans ölçütleri incelendiğinde, modellerin sınıflandırma performanslarını değerlendirebilmek adına çeşitli metriklerin dikkate alındığı görülmektedir. En yüksek AUC skoru (%95) ile LightGBM modeli, verilen veri seti üzerinde en iyi sınıflandırma yeteneğini

sergilemiştir. Veri dengesiz olduğundan F1 skoru, veri desenine hiç müdahale edilmemesine rağmen %60,1 'dir. Bu da modellemenin iyi bir başarı oranı yakaladığını göstermektedir.

Diğer modeller arasında, Catboost Classifier ve Gradient Boosting Classifier modelleri de oldukça yüksek başarı oranlarına sahiptir. Bu modeller, Accuracy (Doğruluk) ve AUC değerlerinde diğer modellerle yakın performans gösterirken, Recall (Geri Çağırma) ve Precision (Hassasiyet) ölçütlerinde de dengeli sonuçlar elde etmişlerdir. Ayrıca, model performansı yanında işlem süreleri de dikkate alındığında, LightGBM gibi yüksek performanslı modellerin yanı sıra Gradient Boosting ve Random Forest gibi modellerin de oldukça kabul edilebilir işlem sürelerinin olduğu görülmektedir. Bu sonuçlar, topluluk öğrenme modellerinin genel olarak iyi performans gösterdiğini ve bu tür modellerin bu belirli veri seti üzerinde daha etkili sınıflandırma yetenekleri sergilediğini vurgulamaktadır.



Resim 2. Değişken Önem Düzeyi Grafiği

En başarılı çıkan model üzerinden hedef değişken üzerinde etkili olan değişkenleri önem sırasına göre sıralamak mümkündür. "LightGBM" modeli en başarılı model çıktığı için bu model üzerinden, rastgele seçilen değişkenler arasında kurulan modeller sonucunda bağımlı değişken üzerinde en fazla etkisi olan faktörler belirlenmiştir. Bu analizler neticesinde 'Duration' (Son iletişim süresi) en etkili değişken olarak belirlenmiştir. Yani, müşterilerin vadeli mevduat alma davranışı üzerinde en önemli değişken olarak öne

çıkmıştır. Ardından sırasıyla euribor3m (Euribor 3 aylık oran- günlük gösterge) ve age (yaş) değişkenleri gelmektedir.

4. Sonuç ve Tartışma

Bu çalışma, finansal hizmetlerde müşteri tercihlerini anlama ve vadeli mevduat talebini öngörme amacıyla yapılmıştır. "DataBlender AutoML" ile veri seti detaylıca incelenmiş ve uygun ön işleme adımları ve modelleme çalışması kod yazmadan yapılmıştır. İncelenen veri seti, dengesizlik içeren bir dağılım göstermiştir bu da doğru sınıflandırmayı zorlaştırmıştır. Ancak, yapılan modelleme çalışmaları, topluluk öğrenme modellerinin, özellikle "LightGBM" gibi yüksek performanslı modellerin, bu dengesizlikle başa çıkabilme yeteneğini ortaya koymuştur. Özellikle AUC skoru üzerinden değerlendirildiğinde, "LightGBM" modeli en yüksek sınıflandırma yeteneği sergilemiştir. Çalışma, veri setindeki dengesizliği göz önünde bulundurarak F1 skoru üzerinden modele daha dengeli bir değerlendirme yapma imkânı sunmuştur. Topluluk öğrenme modelleri genel olarak yüksek başarı oranları göstermiş ve özellikle "Catboost Classifier" ve "Gradient Boosting Classifier" modelleri, dengeli sonuçlar elde etmiştir. Ayrıca, işlem süreleri de dikkate alındığında, yüksek performanslı modellerin yanı sıra "Gradient Boosting Classifier" ve "Random Forest Classifier" gibi modellerin de kabul edilebilir işlem süreleriyle etkileyici sonuçlar verdiği gözlemlenmiştir.

Çalışmanın sonucunda, başarılı çıkan "LightGBM" modeli üzerinden, vadeli mevduat hesabı isteme olasılığı üzerinde etkili olan değişkenleri belirlemiş ve müşterilerin özelliklerini sıralamıştır.

Bu çalışma ile müşterilerin model sonuçlarına göre atanacağı tahmin olasılık skorlarına göre sıralanıp en yüksek vadeli mevduatı alacak müşterilere odaklanılması sağlanılacaktır. Böylece hedef kitle daraltılarak zaman ve maliyet yönünden tasarruf sağlanıp doğru strateji ile müşteriye ulaşılması sağlanılacaktır. Bu sonuçlar, finans kuruluşlarının hedefleme stratejilerini iyileştirmek ve pazarlama çabalarını daha etkili bir şekilde yönlendirmek için makine öğrenmesi modellerini kullanmalarını teşvik edebilir.

Kaynakça

A.Elsalamony, H. (2014). Bank Direct Marketing Analysis of Data Mining Techniques. *International Journal of Computer Applications*, *85*(7), 12–22. https://doi.org/10.5120/14852-3218

ALAN, A., & KARABATAK, M. (2020). Veri Seti - Sınıflandırma İlişkisinde Performansa Etki Eden Faktörlerin Değerlendirilmesi. *Fırat Üniversitesi Mühendislik Bilimleri Dergisi*, *32*(2), 531–540. https://doi.org/10.35234/fumbd.738007

Alpaydın, E. (2010). *Introduction to Machine Learning*. USA: Massachusetts Institute of Technology Press.

Breiman, L. (2001). Random Forests. In *Machine Learning*. Springer.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-August-2016*, 785–794. https://doi.org/10.1145/2939672.2939785

Coşkun, S. B., & Turanli, M. (2023). Credit risk analysis using boosting methods. *Journal of Applied Mathematics, Statistics and Informatics*, *19*(1), 5–18. https://doi.org/10.2478/jamsi-2023-0001

Friedman, J. H. (2001). *Greedy function approximation: A gradient boosting machine. The Annals of Statistics*.

Ghatasheh, N., Faris, H., AlTaharwa, I., Harb, Y., & Harb, A. (2020). Business Analytics in Telemarketing: Cost-Sensitive Analysis of Bank Campaigns Using Artificial Neural Networks. *Applied Sciences*, *10*(7), 2581. https://doi.org/10.3390/app10072581

Han, Jiawei Kamber, Micheline Pei, J. (2012). *Data mining: Concepts and techniques*. Morgan Kaufmann Publishers.

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

Jhaveri, S., Khedkar, I., Kantharia, Y., & Jaswal, S. (2019). Success Prediction using Random Forest, CatBoost, XGBoost and AdaBoost for Kickstarter Campaigns. *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, 1170–1173. IEEE. https://doi.org/10.1109/ICCMC.2019.8819828

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., … Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *NIPS*.

Lewis, D. (2017). *Machine Learning Made Easy with R: An Intuitive Step by Step Blueprint for Beginners*. CreateSpace Independent Publishing Platform.

Mertler, C. A., & Reinhart, R. V. (2005). *Advanced and multivariate statistical methods: practical application and interpretation*.

Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, *62*, 22–31. https://doi.org/10.1016/j.dss.2014.03.001

Moro, S., Laureano, R. M. S., & Cortez, P. (2011). Using data mining for bank direct marketing: An application of the CRISP-DM methodology. *ESM 2011 - 2011 European Simulation and Modelling Conference: Modelling and Simulation 2011*, (Figure 1), 117–121.

Okur, H., & Cetin, A. (2019). Credit Risk Estimation With Machine Learning. *2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, 1–6. IEEE. https://doi.org/10.1109/ISMSIT.2019.8932917

Özdemir, R. (2021). *Makine öğrenmesindeki sınıflandırma yöntemlerinin karşılaştırılması ve e-ticaret üzerinde bir uygulama*. İstanbul Ticaret Üniversitesi.

Özer, H. (2004). *Nitel Değişkenli Ekonometrik Modeller*. Ankara: Nobel Yayın Dağıtım.

Seo, S., & Gary M. Marsh, P. D. (2006). A review and comparison of methods for detecting outliersin univariate data sets. *Department of Biostatistics, Graduate School of Public Health*, 1–53. Retrieved from http://d-scholarship.pitt.edu/7948/

Vapnik, V. N. (1995). *The nature of statistical learning theory*. New York, USA: Springer-Verlag.

Yalçin, M., & Kalkan, S. B. (2022). DETERMINING THE BEST ESTIMATION MODEL WITH TREE-BASED MACHINE LEARNING METHODS: IMPLEMENTATION ON CUSTOMER SPENDINGS FOR E-COMMERCE WEBSITES. *Advances and Applications in Statistics*, *75*, 91–109. https://doi.org/10.17654/0972361722029

Zhang, Y., Zhao, Z., & Zheng, J. (2020). CatBoost: A new approach for estimating daily reference crop evapotranspiration in arid and semi-arid regions of Northern China. *Journal of Hydrology*, *588*(May), 125087. https://doi.org/10.1016/j.jhydrol.2020.125087