ELM472 Makine Öğrenmesinin Temelleri

Çeşitli Veri Dengeleme Yöntemleri Ve Ekstrem Gradyan Artırma Yöntemi Kullanılarak Kaçak Elektrik Kullanımı Tespiti

Kağan Fıkırkoca
k.fikirkoca2019@gtu.edu.tr
[2]Elektronik Mühendisliği Bölümü, GTÜ, Kocaeli, Türkiye

I. GİRİŞ

Bu proje, özellikle akıllı şebekelerin bağlamında, Gelişmiş Ölçüm Altyapısı'ndaki (GÖA) elektrik hırsızlığının ortaya koyduğu zorlukları ele almaktadır. GÖA'nın temel bir bileşeni olan akıllı sayaçlar, özellikle gelişmiş cihazlar veya siber saldırı teknikleri kullanılarak kötü niyetli kullanıcılar tarafından manipüle edilme riskine sahiptir, bu da dünya genelinde elektrik şirketleri için önemli finansal kayıplara neden olabilir. Elektrik hırsızlığını tespit etmek için geleneksel yöntemler kullanılmaktadır. Bu yöntemler yetkililerin sahaya giderek sayaçlardan aylık ölçüm verilerinin el ile analizlerini yapmasıdır, bu da zaman alıcıdır ve uzman bilgisine dayanır. Bu soruna çözüm olarak, özellikle yapay zeka (YZ) ve ekstrem gradyan artırma (XGBoost) gibi makine öğrenimi yöntemleri etkili araçlar olarak önerilmektedir. Proje, elektrik hırsızlığına karşı bir YZ çözümü önermektedir. Bu çözümde, veri seti ön işlemesi, çeşitli veri seti dengeleme yöntemleri ve ekstrem gradyan artırma yöntemi mevcuttur. Mevcut YZ temelli yöntemlerde, aşırı uyum problemi ve gürültülü verilerde zayıf performans gösterme problemleri vardır ve bu da XGBoost'u umut vadeden bir çözüm olarak ortaya çıkarır. Önerilen XGBoost tabanlı elektrik hırsızlık tespiti, düzgün kullanıcıların ve kaçak kullanıcıların ölçüm verileri içeren bir eğitim aşamasını ve kaçak kullanıcıların tespiti için bir test aşamasını içerir. Çin Devlet Şebeke Kurumu'nun verileri [5], kullanılarak yapılan simulasyonlar, XGBoost tabanlı yöntemin diğer algoritmalarla karşılaştırıldığında üstün performansını gösterir. Proje, önerilen yöntemin sonuçlarını özetlerken, yaklaşımın sınırlamalarını da sunar.

II. İLGİLİ ÇALIŞMALAR

Elektrik güç kaybı, üretim ve dağıtım arasındaki birim farkını ifade eder. Son yıllarda, araştırmacılar Teknik Olmayan Kaybın (TOK) tespiti sorunlarına odaklanarak, makine öğrenimi ve derin öğrenme algoritmalarını kullanarak bu kaybı tespit etmeye çalışmıştır. Literatürde bulunan çözümler genellikle donanım tabanlı ve veri tabanlı olmak üzere iki ana kategoride sıralanır. Donanım tabanlı çözümler pahalı olabilir, bu nedenle genellikle veri tabanlı yaklaşımlara odaklanılmaktadır.

Veri tabanlı yöntemler genellikle sınıflandırma teorisi üzerine odaklanmaktadır. Örneğin, Evrişimsel Sinir Ağları (CNN) [1] makalesinde, güç tüketim verilerinin periyodik olmayan özelliklerini yakalamak için kullanılmıştır. Ancak, bu yöntemde oldukça dengesiz bir veri kullanılmıştır. Başka bir örnekte, uzun kısa vadeli bellek (LSTM) ve çok katmanlı algılayıcı (MLP) kombinasyonu [2] makalesinde, TOK tespiti için kullanılmıştır. Ancak, bu yöntemde parametre optimizasyonu için şebeke araması kullanıldığından hesaplama karmaşıklığı ortaya çıkar ve optimum parametreler bilinemez.

Makine öğrenimi algoritmalarının performansını artırmak amacıyla, [3], [2] ve [4] makalelerinde siyah delik algoritması, CNN ve Maksimal Örtüşmeli Sürekli Dalga Paketi Dönüşümü (MODWPT) kullanılarak veri setinden özellik çıkarımı yapılmıştır. Ancak, bu yöntemlerin her biri kendi zorluklarına sahiptir.

Sonuç olarak, literatürdeki çeşitli yöntemlerin avantajları ve dezavantajları bulunmaktadır ve her biri kendi bağlamında değerlendirilmelidir.

III. ÖNERİLEN YÖNTEM

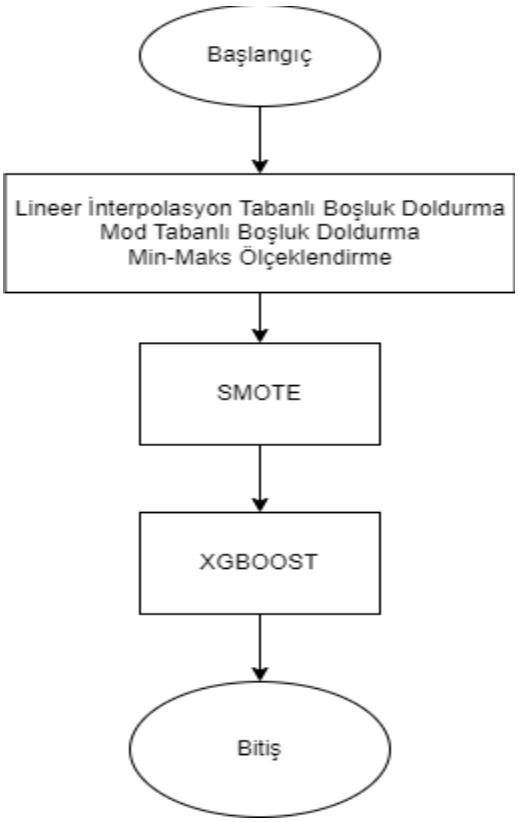
Bu projede belirlenen TOK tespiti için kullanılan yöntem Sentetik Azınlık Aşırı Örnekleme Tekniği (SMOTE) ve XGBoost [6] yöntemidir. SMOTE yönteminin kullanılabilmesi için ise öncelikle bahsedilen veri setine bir ön işleme işlemi yapılmıştır. Yapılan veri seti ön işleme yönteminden bahsetmek gerekirse, öncelikle veri setindeki modelin işine yaramayan müşteri numarası kısmı çıkarılmıştır, sonrasında veri setinde pek çok boşluk bulunduğu için bu boşluklar mod yöntemi ve lineer interpolasyon yöntemi ile doldurulmuştur. Ardından elde edilen veri seti minimum-maksimum ölçeklendirme ile veri setindeki değerler 0 ile 1 arasında sınırlandırılmıştır. Bu işlemlerin yapılmasının sebebi veri kalitesini artırmak ve aykırı değerlerin yapay zeka modeli üzerindeki etkisini azaltmaktır. Veri seti ön işlemesi tamamlandıktan sonra çeşitli veri seti dengeleme yöntemleri kullanılmıştır. Bunlar, SMOTE [7], Adaptif Sentetik (ADASYN) [7],

rastgele örnek azaltma [7] ve SMOTEEN [7] yöntemleridir. Bu yöntemler tek tek denenerek ve gerekli parametre ayarlamaları yapılarak en yüksek doğruluk değerini veren yöntem olarak SMOTE öne çıkmaktadır. SMOTE yöntemi eğitim veri setine uygulandıktan sonra veri seti XGBoost modeli ile eğitilmiştir. XGBoost kullanılmasının sebepleri ise hızlı çalışması, eksik veri işleme yeteneğine sahip olması, çeşitli regülarizasyon teknikleri içermesi ve dengesiz veri kümesiyle başa çıkabilmesidir.

Sonuç olarak, TOK tespiti için SMOTE ve XGBoost yöntemleri kullanılmıştır. Veri seti ön işleme adımları, müşteri numarası temizleme, boşluk doldurma ve ölçeklendirme gibi süreçleri içermektedir. Veri setini dengelemek için SMOTE, ADASYN, rastgele örnek azaltma ve SMOTEEN yöntemleri denenmiş, en yüksek doğruluk değeri SMOTE ile elde edilmiştir. SMOTE uygulandıktan sonra XGBoost modeli ile veri seti eğitilmiştir. Bu yöntemlerin birleşimi, etkili bir TOK tespiti modeli oluşturmuştur.

A. *Sistem Tanımı*

Bu yazıda kullanılan sistemde TOK tespitinin başarılı bir şekilde yapılabilmesi için Şekil 1'de görülen sistem kullanılmıştır.



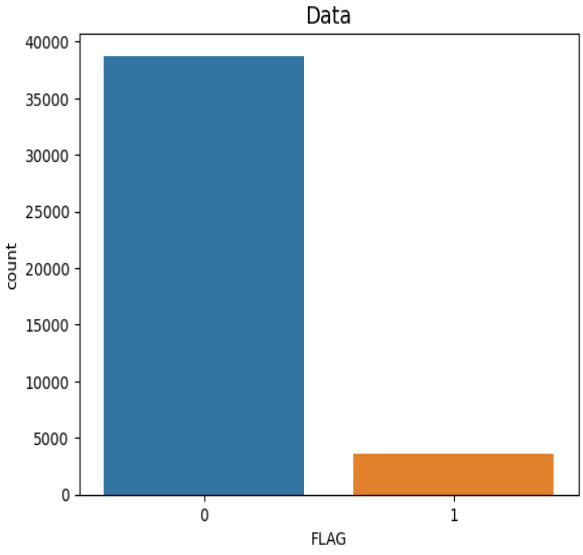
Şekil 1

B. *Veri Seti*

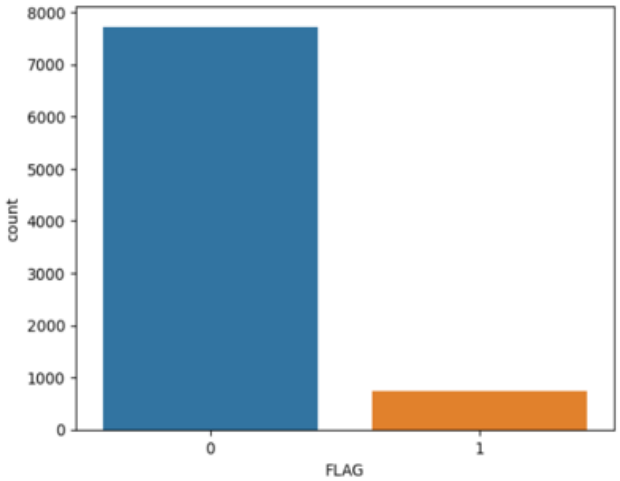
- Veri Seti olarak Çin Devlet Şebeke Kurumu'nun verileri[5] kullanılmıştır. Bu veri seti 42372 adet kullanıcının, 3 yıl boyunca her gün harcadığı verileri içermektedir. Dolayısıyla kullanıcı başına 1080 adet veri bulunmaktadır. Şekil 2'de veri setinin bir kısmı görülebilmektedir. Ayrıca Şekil 3'de 0 ile ifade edilen düzgün kullanıcıların ve 1 ile ifade edilen kaçak kullanıcıların sayısı da görülebilmektedir. Bu aşamada Şekil 4'de görüldüğü üzere test verisi ayrılmıştır. Eğitim verisinin gösterilmemesinin sebebi her veri seti dengeleme yöntemi uygulandığında eğitim veri setinin değişmesidir. Ancak test veri seti Şekil 4'de görüldüğü gibi sabittir. Bu şekilde modelin sağlıklı bir şekilde test edilebilmesi sağlanmıştır.
-

1/22/2014	1/23/2014	1/24/2014	1/25/2014	1/26/2014	CONS_NO	FLAG
2419	2462	1177	340	2272	A0E79140(	1
3840	1422	1956	1380	96	B415F931I	1
1188,6	1362,6	1286	1267,6	1444	DE8E1EAE	1
1090,2	1101	1079,4	1121,4	1145,4	2952491E9	0
1463,1	1405,5	1240,05	1190,25	1455,45	DBEED8FD	1
14,85	16,8	4,35	4,35	2,85	9495BA31I	1
727,92	726,59	671,9	679,77	629,39	58A9AA93	0
0	0	0	0	0	1BC425A7	1

Şekil 2



Şekil 3



Şekil 4

C. Denklemler

Şekil 1'de bahsedilen yöntemlerin uygulanabilmesi için pek çok denklem kullanılmıştır. Lineer interpolasyon için (1), Mod değerinin bulunabilmesi için (2) ve Min-maks ölçeklendirme için (3) denklemleri kullanılmıştır. Mod değerinin formülünde (2) "n" değeri belirli bir değerin toplam tekrar sayısını, "N" değeri ise veri setindeki toplam örnek sayısını ifader eder. Dolayısıyla bu denklem tüm veriler için uygulanarak en sık kullanılan değer bulunmuştur. Ayrıca modelin doğruluğunu göstermek için de çeşitli metrikler kullanılmıştır bunlar, kesinlik değeri (4), duyarlılık değeri (5), F1-Skoru (6) ve doğruluk değeridir (7).

Kullanılan Denklemler:

$$y = \frac{(y2 - y1)}{(x2 - x1)} \times (x - x1) + y1 \qquad (1)$$

$$maks\left(\frac{n}{N}\right) \times 100 \qquad (2)$$

$$x' = \frac{x - \min(x)}{maks(x) - \min(x)} \qquad (3)$$

$$Kesinlik = \frac{TP}{TP + FP} \qquad (4)$$

$$Duyarlılık = \frac{TP}{TP + FP} \qquad (5)$$

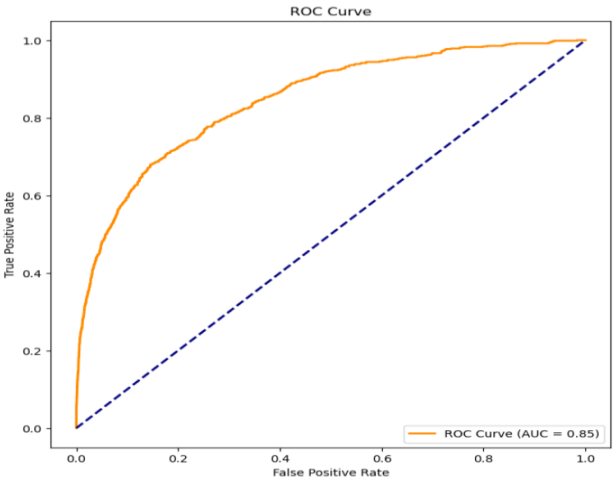
$$F1\ Skoru = 2 \times \frac{Kesinlik \times Duyarlılık}{Kesinlik + Duyarlılık} \qquad (6)$$

$$Doğruluk\ Değeri = \frac{TP + TN}{TP + FP + TN + FN} \qquad (7)$$

IV. DENEY SONUÇLARI

A. XGBoost Kullanılarak Elde Edilen Sonuçlar

Bu aşamada deney yapılırken III. Önerilen Yöntem kısmında bahsedilen veri seti üzerinde herhangi bir veri dengeleme işlemi yapılmadan XGBoost modeli denenmiştir ve çeşitli metrikler kullanılarak model test edilmiştir. Şekil 5'de ROC eğrisi ve eğrinin altında kalan alan görülmektedir, Tablo 1'de ise modelin doğruluk değeri, hem düzgün kullanıcılar hem de kaçak kullanıcılar için kesinlik, duyarlılık ve F1-skoru değerleri görülebilmektedir. Bu metriklere bakıldığında sadece XGBoost kullanıldığı zaman kaçak elektrik kullanıcılarının tespitinin yetersiz olduğu söylenebilir.



Şekil 5

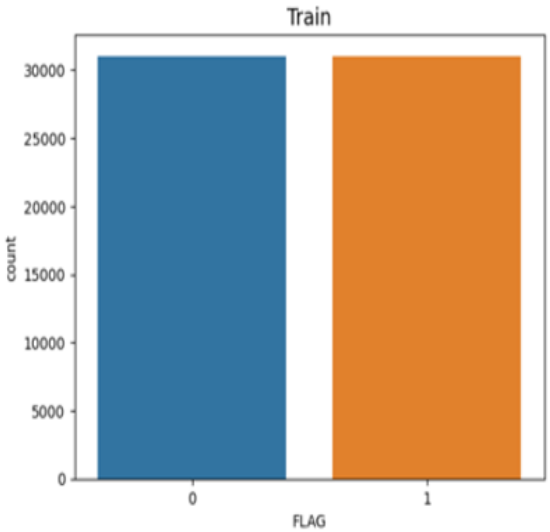
Tablo 1

XGBoost	Kesinlik	Duyarlılık	F1-Skoru	Veri Sayısı
0 (Düzgün Kullanıcılar)	0.93	0.99	0.96	7725
1 (Kaçak Kullanıcılar)	0.76	0.23	0.36	750
Doğruluk Değeri	0.9255			

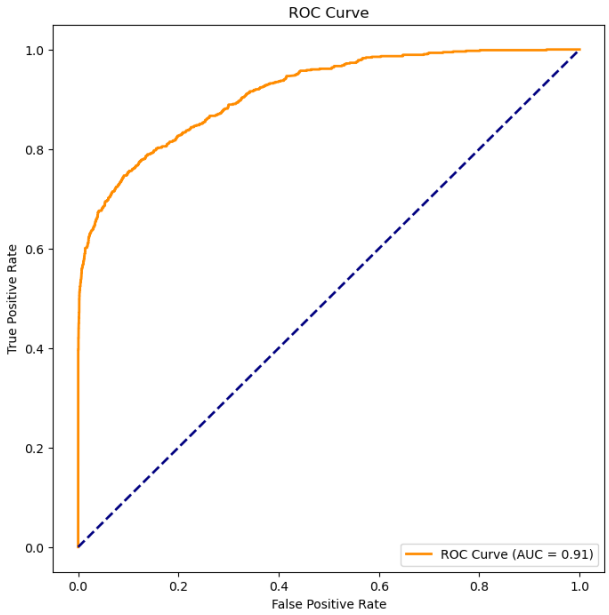
B. SMOTE ve XGBoost Kullanılarak Elde Edilen Sonuçlar

Bu aşamada deney yapılırken III. Önerilen Yöntem kısmında bahsedilen veri seti üzerinde SMOTE veri seti dengeleme yöntemi kullanılmasının ardından XGBoost modeli denenmiştir ve çeşitli metrikler kullanılarak model test edilmiştir. SMOTE yöntemi kullanılırken parametreler en yüksek doğruluk oranını verecek şekilde ayarlanmıştır

ve Şekil 6'da görüldüğü gibidir. Şekil 7'de ROC eğrisi ve eğrinin altında kalan alan görülmektedir, Tablo 2'de ise modelin doğruluk değeri, hem düzgün kullanıcılar hem de kaçak kullanıcılar için kesinlik, duyarlılık ve F1-skoru değerleri görülebilmektedir. Bu metriklere bakıldığında SMOTE ve XGBoost yöntemleri kullanıldığı zaman kaçak elektrik kullanıcılarının tespitinin sadece XGBoost kullanılan yönteme kıyasla çok daha iyi bir seviyede olduğu söylenebilir.



Şekil 6



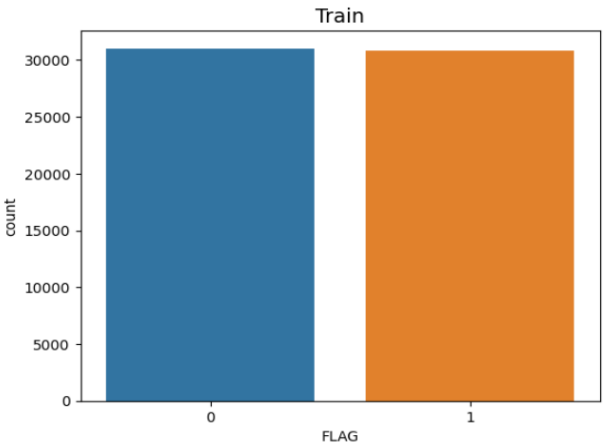
Şekil 7

Tablo 2

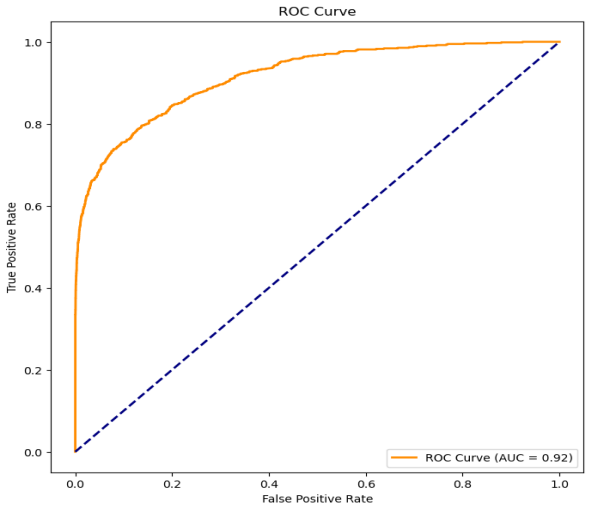
SMOTE + XGBoost	Kesinlik	Duyarlılık	F1-Skoru	Veri Sayısı
0 (Düzgün Kullanıcılar)	0.97	0.97	0.97	7725
1 (Kaçak Kullanıcılar)	0.66	0.65	0.65	750
Doğruluk Değeri	0.9389			

C. ADASYN ve XGBoost Kullanılarak Elde Edilen Sonuçlar

Bu aşamada deney yapılırken III. Önerilen Yöntem kısmında bahsedilen veri seti üzerinde ADASYN veri seti dengeleme yöntemi kullanılmasının ardından XGBoost modeli denenmiştir ve çeşitli metrikler kullanılarak model test edilmiştir. ADASYN yöntemi kullanılırken parametreler en yüksek doğruluk oranını verecek şekilde ayarlanmıştır ve Şekil 8'de görüldüğü gibidir. Şekil 9'da ROC eğrisi ve eğrinin altında kalan alan görülmektedir, Tablo 3'de ise modelin doğruluk değeri, hem düzgün kullanıcılar hem de kaçak kullanıcılar için kesinlik, duyarlılık ve F1-skoru değerleri görülebilmektedir. Bu metriklere bakıldığında ADASYN ve XGBoost yöntemleri kullanıldığı zaman kaçak elektrik kullanıcılarının tespitinin sadece XGBoost kullanılan yönteme kıyasla çok daha iyi bir seviyede olduğu söylenebilir. Ancak SMOTE ve XGBoost yöntemlerine kıyasla biraz daha kötü sonuç verdiği söylenebilir.



Şekil 8



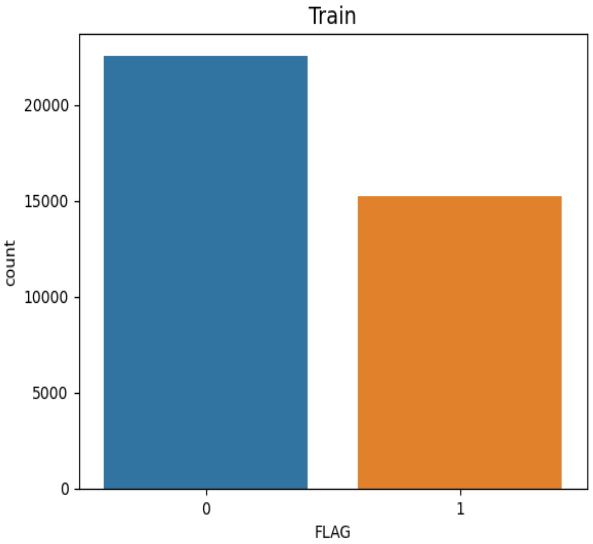
Şekil 9

Tablo 3

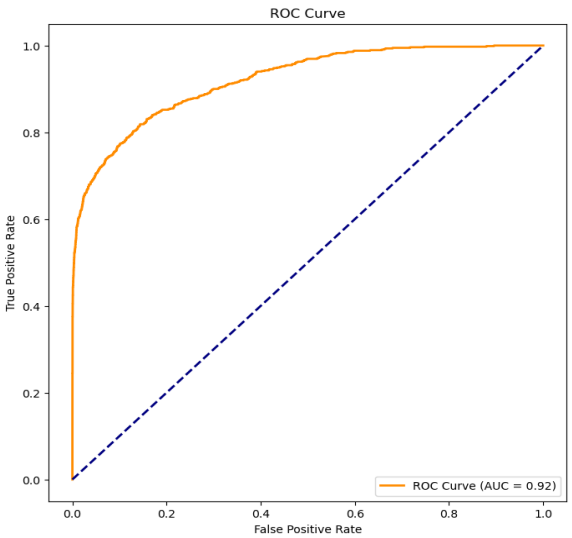
ADASYN + XGBoost	Kesinlik	Duyarlılık	F1-Skoru	Veri Sayısı
0 (Düzgün Kullanıcılar)	0.97	0.97	0.97	7725
1 (Kaçak Kullanıcılar)	0.66	0.65	0.65	750
Doğruluk Değeri	0.9389			

D. SMOTEEN ve XGBoost Kullanılarak Elde Edilen Sonuçlar

Bu aşamada deney yapılırken III. Önerilen Yöntem kısmında bahsedilen veri seti üzerinde SMOTEEN veri seti dengeleme yöntemi kullanılmasının ardından XGBoost modeli denenmiştir ve çeşitli metrikler kullanılarak model test edilmiştir. SMOTEEN yöntemi kullanılırken parametreler en yüksek doğruluk oranını verecek şekilde ayarlanmıştır ve Şekil 10'da görüldüğü gibidir. Şekil 11'de ROC eğrisi ve eğrinin altında kalan alan görülmektedir, Tablo 4'de ise modelin doğruluk değeri, hem düzgün kullanıcılar hem de kaçak kullanıcılar için kesinlik, duyarlılık ve F1-skoru değerleri görülebilmektedir. Bu metriklere bakıldığında SMOTEEN ve XGBoost yöntemleri kullanıldığı zaman kaçak elektrik kullanıcılarının tespitinin sadece XGBoost kullanılan yönteme kıyasla çok daha iyi bir seviyede olduğu söylenebilir. Ancak SMOTE ve XGBoost yöntemlerine kıyasla biraz daha kötü sonuç verdiği söylenebilir.



Şekil 10



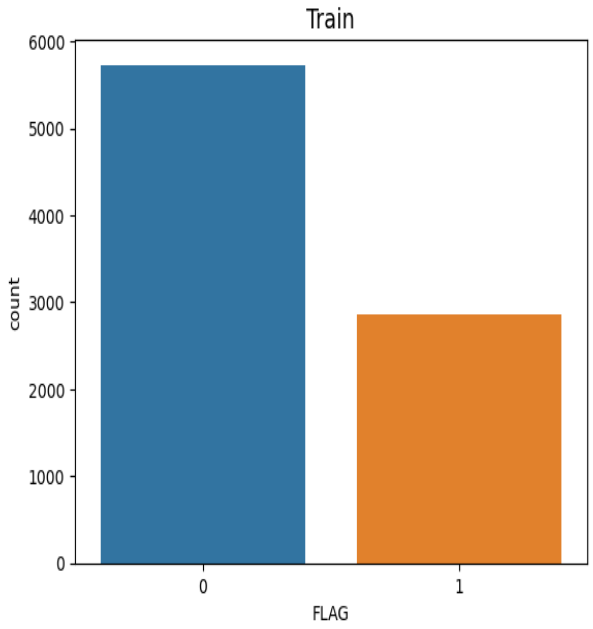
Şekil 11

Tablo 4

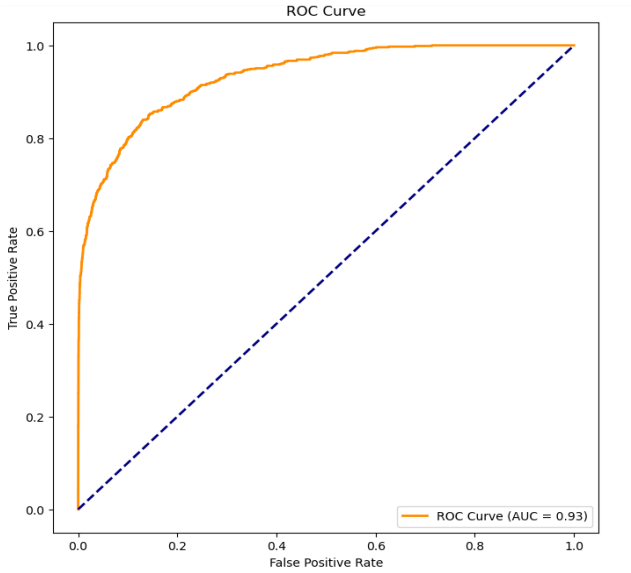
SMOTEEN + XGBoost	Kesinlik	Duyarlılık	F1-Skoru	Veri Sayısı
0 (Düzgün Kullanıcılar)	0.97	0.94	0.96	7725
1 (Kaçak Kullanıcılar)	0.55	0.71	0.62	750
Doğruluk Değeri	0.9233			

E. Rastgele Örnek Azaltma ve XGBoost Kullanılarak Elde Edilen Sonuçlar

Bu aşamada deney yapılırken III. Önerilen Yöntem kısmında bahsedilen veri seti üzerinde rastgele veri azaltma veri seti dengeleme yöntemi kullanılmasının ardından XGBoost modeli denenmiştir ve çeşitli metrikler kullanılarak model test edilmiştir. Rastgele veri azaltma yöntemi kullanılırken parametreler en yüksek doğruluk oranını verecek şekilde ayarlanmıştır ve Şekil 12'de görüldüğü gibidir. Şekil 13'de ROC eğrisi ve eğrinin altında kalan alan görülmektedir, Tablo 5'de ise modelin doğruluk değeri, hem düzgün kullanıcılar hem de kaçak kullanıcılar için kesinlik, duyarlılık ve F1-skoru değerleri görülebilmektedir. Bu metriklere bakıldığında rastgele veri azaltma ve XGBoost yöntemleri kullanıldığı zaman kaçak elektrik kullanıcılarının tespitinin sadece XGBoost kullanılan yönteme kıyasla çok daha iyi bir seviyede olduğu söylenebilir. Ancak rastgele veri azaltma ve XGBoost yöntemlerine kıyasla biraz daha kötü sonuç verdiği söylenebilir.



Şekil 12



Şekil 13

Tablo 5

Rastgele Veri Azaltma + XGBoost	Kesinlik	Duyarlılık	F1-Skoru	Veri Sayısı
0 (Düzgün Kullanıcılar)	0.97	0.95	0.96	7725
1 (Kaçak Kullanıcılar)	0.56	0.71	0.63	750
Doğruluk Değeri	0.9255			

F. Sonuçların Karşılaştırılması

Tablo 6'da görüldüğü üzere bahsedilen yöntemlerin hepsi kıyaslanmıştır. Kıyaslama yapılırken projenin amacı kaçak elektrik kullanımının tespiti olduğu için kıyaslama, test veri setindeki kaçak elektrik kullanıcılarının metrik değerleri üzerinden yapılmıştır. Buradaki metrik değerlerine dikkatli bir şekilde bakıldığında doğruluk değerinin bütün yöntemler için oldukça yüksek elde edilirken F1-Skoru'nun bütün yöntemlerde farklı olması ve doğruluk değerine göre düşük elde edilmiş olmasıdır. Bunun sebebi, doğruluk değerinin bulunurken hem düzgün kullanıcıların hem de kaçak kullanıcıların dikkate alınarak bu sonuca varılmasıdır. Veri setinde düzgün kullanıcılar yüksek çoğunlukta olduğu için herhangi bir veri seti dengeleme yöntemi kullanılmadan düzgün kullanıcılar için doğru tahminler yapılabilmektedir. Ancak aynı durum azınlık sınıfı olan kaçak kullanıcılar için söylenemez. Doğruluk değeri genel olarak tahminlerin doğruluğuna baktığı için bu projede bizi yanıltmaktadır. Dolayısıyla kaçak

kullanıcıların bulunması amacıyla F1-Skoru'na bakılması doğru bir yaklaşım olacaktır.

Sonuç olarak, denenen yöntemler arasında en iyi sonuç veren yöntem Şekil 1'de görülen SMOTE ve XGBoost yöntemlerinin kullanılması olmuştur.

Tablo 6

	Kesinlik	Duyarlılık	F1-Skoru	Doğruluk Değeri
XGBoost	0.76	0.23	0.36	0.9255
SMOTE + XGBOOST	0.66	0.65	0.65	0.9389
ADASYN + XGBoost	0.62	0.67	0.64	0.9335
SMOTEEN + XGBoost	0.55	0.71	0.62	0.9233
Rastgele Örnek Azaltma + XGBoost	0.41	0.78	0.54	0.8820

SONUÇ

Bu projede, Çin Devlet Şebeke Kurumu'nun verileri [5] üzerinden gerçekleştirilen kaçak elektrik kullanıcılarının tespiti başarılı bir şekilde tamamlanmıştır. Özellikle, çalışmanın odak noktası, veri seti dengeleme yöntemlerinin uygulanması öncesi ve sonrasındaki model performansındaki önemli değişiklikler üzerinedir.

Çalışmanın öne çıkan katkısı, veri seti dengeleme yöntemlerinin kullanılmadan önceki modelin performansı ile dengeleme yöntemleri uygulandıktan sonraki performans arasında belirgin bir iyileşme gözlemlenmiş olmasıdır. Bu iyileşme, özellikle Tablo 6'da detaylı bir şekilde açıklanmıştır. Tablo 6'da görüldüğü üzere, dengeleme yöntemleri kullanılmadan önce model, üç kaçak kullanıcıdan birini doğru bir şekilde tespit edebilirken, dengeleme yöntemleri kullanıldığında bu oranın iki kaçak kullanıcıyı doğru bir şekilde tespit edebilecek seviyeye yükseldiği gözlemlenmiştir.

Veri seti dengeleme yöntemlerinin kullanımının, modelin duyarlılığını ve özellikle kaçak elektrik tespitindeki doğruluğunu artırmada kritik bir rol oynadığı sonucuna varılmıştır. Özellikle, çalışmada kullanılan dengeleme yöntemleri arasında en üst düzey performansı, önerilen SMOTE ve XGBoost'un birleşik kullanımının sağladığı belirlenmiştir. Şekil 1'de görsel olarak vurgulanan bu yöntem kombinasyonunun kaçak elektrik tespiti uygulamalarında tercih edilebilecek bir etkinlik sunabileceği kanıtlanmıştır.

Bu bağlamda, veri seti dengeleme yöntemleriyle güçlendirilen modelin, enerji sektöründeki kurumların kaçak elektrikle mücadelesine önemli bir katkı sağlayabileceği ve daha etkili sonuçlar elde edilebileceği vurgulanmıştır.

KAYNAKÇA

[1] Zheng, Z., Yang, Y., Niu, X., Dai, H.N. and Zhou, Y.,"Wide and deep convolutional neural networks for electricity-theft detection to secure smart grids." 2017, IEEE Transactions on Industrial Informatics, 14(4), pp.1606-1615. Zheng,
[2] Z. Xiao, Y. Xiao and D. H. Du, March 2013.Exploring Malicious Meter Inspection in Neighborhood Area 376 Smart Grids, in IEEE Transactions
on Smart Grid, vol. 4, no. 1, pp. 214-226.
[3] Maamar, A. and Benahmed, K., A Hybrid Model for Anomalies Detection in AMI System Combining K-means Clustering and Deep Neural Network.",2019, vol.60, no.1, pp.15-39.
[4] N. F. Avila, G. Figueroa, and C.-C. Chu,NTL detection in electric distribution systems using the maximal overlap discrete wavelet-packet transform and random undersampling boosting, 2015, IEEE Transactions on Power Systems, vol. 33, pp. 7171-7180.
[5] https://www.kaggle.com/datasets/bensalem14/sgcc-dataset
[6] https://en.wikipedia.org/wiki/XGBoost
[7]https://en.wikipedia.org/wiki/Oversampling_and_undersampling_in_data_analysis