Cilt: 2, Sayı: 1, 2026, Sayfa: 27-45

Yönetim Bilişim Sistemleri Derneği Dergisi

ISSN: 3108-6454

Journal of Management Information Systems Association

https://www.ybs.org.tr

METİNSEL VE METADATA ÖZELLİKLERİNİ KULLANAN SAHTE İŞ İLANLARINI TESPİT ETMEYE YÖNELİK HİBRİT MAKİNE ÖĞRENMESİ ÇERÇEVESİ

Ayşe Gökçen ÇATAK [a] ve Oğuz KAYNAR [b]

Makale Bilgisi

Makale Türü
Araştırma Makalesi

Gönderim Tarihi:
16/04/2026

Kabul Tarihi:
26/06/2026

Anahtar Kelimeler:
Sahte İş İlanı Tespiti, Topluluk Öğrenmesi, XGBoost, SMOTE, Metin Madenciliği.

Özet

Çevrimiçi istihdam platformlarının hızla yaygınlaşması, sahte iş ilanlarının tespitini hem iş arayanlar hem de kurumlar açısından önemli bir sorun hâline getirmiştir. Bu çalışmada, TF-IDF tabanlı metinsel temsiller ile şirket logosu varlığı, maaş bilgisi durumu ve eksiklik sinyalleri gibi metaveri özelliklerini birleştiren hibrit bir makine öğrenmesi çerçevesi önerilmiştir. Sahte ilan oranının yalnızca %5 olduğu dengesiz veri setinde Lojistik Regresyon, Random Forest ve XGBoost modelleri; Maliyet Duyarlı Öğrenme ve SMOTE stratejileri altında karşılaştırılmıştır. Bulgular, topluluk öğrenmesi modellerinin lineer modellere göre daha başarılı olduğunu göstermiştir. Random Forest + SMOTE modeli 0.993 AUC değeriyle en yüksek performansa ulaşmıştır. Sonuçlar, hibrit özellik kullanımının yanlış pozitif oranını azaltırken yüksek duyarlılığı koruduğunu ve güvenilir tespit sağladığını ortaya koymaktadır.

[a]Yüksek Lisans Öğrenci, Sivas Cumhuriyet Üniversitesi, Yönetim Bilişim Sistemleri Bölümü, Sivas/Türkiye, 20249348006@cumhuriyet.edu.tr , ORCID: 0009-0008-0712-853X (Sorumlu Yazar)

[b]Prof. Dr., Sivas Cumhuriyet Üniversitesi, Yönetim Bilişim Sistemleri Bölümü, Sivas/Türkiye, okaynar@cumhuriyet.edu.tr, ORCID: 0000-0003-2387-4053 (Sorumlu Yazar)

A HYBRID MACHINE LEARNING FRAMEWORK FOR FAKE JOB POSTING DETECTION USING TEXTUAL AND METADATA FEATURES

Article Information

Article Type
Research Article

Submission Date:
16/04/2026

Acceptance Date:
26/06/2026

Keywords:
Fake Job Posting Detection, Ensemble Learning, XGBoost, SMOTE, Text Mining

Abstract

The rapid growth of online employment platforms has made fake job posting detection a critical challenge for both job seekers and organizations. This study proposes a hybrid machine learning framework that combines TF-IDF based textual representations with structural metadata features such as company logo presence, salary disclosure status, and missingness signals. Using the "Real or Fake Job Posting Prediction" dataset, where fraudulent postings constitute only 5% of all records, Logistic Regression, Random Forest, and XGBoost models were evaluated under Cost-Sensitive Learning and SMOTE strategies. Experimental findings indicate that ensemble-based models outperform linear classifiers. In particular, Random Forest combined with SMOTE achieved the highest AUC score of 0.993, demonstrating near-perfect discrimination between genuine and fraudulent postings. Results show that the hybrid feature set reduces false positive rates while preserving high recall and provides reliable detection performance for online recruitment fraud tasks.

1. GİRİŞ

Dijital iş arama platformlarının kullanımındaki hızlı artış, iş ilanlarının doğruluğunu ve güvenilirliğini kritik bir problem hâline getirmiştir. Özellikle çevrim içi platformların kolay erişilebilirliği ve denetimsiz doğası, kötü niyetli aktörlerin sahte iş ilanları oluşturarak iş arayanları dolandırmasına imkân tanımaktadır. Bu tür sahte ilanlar, kişisel bilgi hırsızlığı, haksız ücret talebi, phishing (oltalama) e-postaları ve sahte şirket profilleri gibi çeşitli güvenlik riskleri yaratmaktadır (Holt ve Bossler, 2015; Vidros vd., 2017). Europol ve ABD Federal Trade Commission (FTC) raporlarına göre, çevrim içi iş dolandırıcılığı son yıllarda önemli ölçüde artış göstermiş ve kullanıcı güvenliğini tehdit eden yaygın bir sosyal mühendislik pratiği hâline gelmiştir (Europol, 2021; FTC, 2023).

Makine öğrenmesi temelli otomatik tespit sistemleri, sahte ilanların yapısal ve dilsel farklılıklarını modelleyerek bu tür dolandırıcılık faaliyetlerinin önlenmesinde etkili bir çözüm olarak öne çıkmaktadır. Bu çalışmada, "Real or Fake Job Posting Prediction" veri seti (Vidros vd., 2017) temel alınarak metinsel ve yapısal özellikleri bütünleştiren hibrit bir sınıflandırma çerçevesi geliştirilmiştir. TF-IDF tabanlı metin temsilleri, kategorik meta-bilgiler ve eksik bilgi oranlarını (missingness patterns) içeren genişletilmiş özellik seti, tek bir boru hattı (pipeline) üzerinde birleştirilmiştir. Çalışmanın odağında, veri setindeki belirgin sınıf dengesizliği (%5 sahte ilan oranı) yer almaktadır. Bu problemi ele almak amacıyla geleneksel Lojistik Regresyon modeline ek olarak, karmaşık veri örüntülerini yakalamada başarılı olan Random Forest (Breiman, 2001) ve XGBoost (Chen ve Guestrin, 2016) algoritmaları da

araştırmaya dâhil edilmiştir. Çevrimiçi istihdam platformlarındaki sahte ilanların büyük çoğunluğu gerçek ilanlarla iç içe geçtiğinden, veri setinde ciddi bir sınıf dengesizliği ortaya çıkmaktadır. Çalışmada kullanılan veri setinde sahte ilanların oranı yalnızca %5 düzeyinde kalmaktadır; bu durum, standart eğitim yaklaşımlarıyla geliştirilen modellerin çoğunluk sınıfını (gerçek ilanları) öğrenmeye odaklanarak sahte ilanları gözden kaçırmasına zemin hazırlamaktadır. Söz konusu problemi ele almak amacıyla iki farklı yaklaşım bir arada değerlendirilmiştir. Maliyet Duyarlı Öğrenme yönteminde modelin kayıp fonksiyonuna müdahale edilerek azınlık sınıfındaki hatalara daha yüksek ceza puanı atanmış, böylece modelin sahte ilanları daha dikkatli sınıflandırması sağlanmıştır. SMOTE yönteminde ise eğitim seti içindeki dengesizlik, sentetik azınlık örnekleri üretilerek veri düzeyinde giderilmeye çalışılmıştır (Chawla vd., 2002). Her iki stratejinin lineer ve ağaç tabanlı modeller üzerindeki etkileri karşılaştırmalı biçimde analiz edilerek hangi kombinasyonun sahte ilan tespitinde daha güvenilir sonuçlar verdiği belirlenmiştir.

Bu çalışma literatüre üç temel açıdan katkı sağlamaktadır. İlk olarak, yalnızca metin içeriğine dayanmak yerine eksik veri örüntülerini ve yapısal meta-bilgileri de kapsayan hibrit bir özellik mühendisliği mimarisi geliştirilmiştir. İkinci olarak, Lojistik Regresyon, Random Forest ve XGBoost algoritmaları aynı koşullar altında sistematik biçimde karşılaştırılmış; sınıf dengesizliğini gidermek amacıyla uygulanan Maliyet Duyarlı Öğrenme ve SMOTE tekniklerinin bu modeller üzerindeki destekleyici etkileri ayrıca incelenmiştir. Üçüncü olarak, yüksek doğruluk oranı elde etmek yerine sahte ilanların gözden kaçırılma riskini en aza indirmeyi ön plana alan güvenlik odaklı bir değerlendirme çerçevesi benimsenmiş ve bu çerçevede en güvenilir model belirlenmiştir. Çalışmanın ortaya koyduğu bulgular, hem sahte iş ilanlarının yapısal ve dilsel özelliklerinin daha iyi anlaşılmasına hem de gerçek platformlara entegre edilebilecek, yüksek duyarlılıklı bir tespit sisteminin tasarımına katkı sunmaktadır.

2. KAVRAMSAL ÇERÇEVE/LİTERATÜR

Sahte iş ilanı tespiti, dijital işgücü piyasalarındaki güven sorunlarının giderek derinleşmesiyle birlikte araştırmacıların ilgisini çeken bir alan hâline gelmiştir. Literatürdeki çalışmalar, bu probleme yaklaşmak için genellikle metinsel özellikler, n-gram temsilleri veya temel meta-bilgiler üzerine kurulu sınıflandırma modelleri geliştirme yolunu tercih etmektedir.

Örneğin, Alghamdi ve Alharby (2019), SVM tabanlı özellik seçimi ile Random Forest sınıflandırıcısını birleştirerek şirket profili ve logo varlığının sahte ilan tespitinde en ayırt edici özellikler olduğunu göstermiştir. Lal vd. (2019), topluluk öğrenmesi yaklaşımının tekil sınıflandırıcılara kıyasla daha başarılı sonuçlar ürettiğini raporlamıştır. Benzer şekilde, Amaar vd. (2022), doğal dil işleme ve makine öğrenmesi tekniklerini birleştirerek sahte iş ilanlarının genellikle belirsiz dil, eksik maaş bilgisi ve aşırı cazip vaatler içerdiğini ortaya koymuştur. Kim vd. (2019) ise hiyerarşik kümeleme tabanlı derin sinir ağları aracılığıyla iş ilanı sahteciliğini tespit etmiş ve model mimarisinin sınıf dengesizliğine karşı direncinin başarımı doğrudan etkilediğini vurgulamıştır.

Metin tabanlı yaklaşımların yanında, meta-özelliklerin de sahte ilan tespitinde kritik rol oynadığı güncel çalışmalarda belirtilmiştir. Özellikle uzaktan çalışma (telecommuting) durumu, şirket logosunun yokluğu, standart dışı çalışma şekilleri ve tutarsız lokasyon bilgilerinin dolandırıcılık göstergeleri olduğu ifade edilmiştir (Mahbub vd., 2022). Ullah ve Jamjoom (2023), topluluk öğrenmesi yöntemlerini aynı veri seti üzerinde karşılaştırmalı olarak değerlendirmiş ve AdaBoost ile XGBoost modellerinin sınıf dengesizliği koşullarında güvenilir sonuçlar ürettiğini göstermiştir.

Son yıllarda derin öğrenme ve transformer tabanlı yaklaşımların sahte ilan tespitine uygulanması giderek yaygınlaşmaktadır. Akram vd. (2024), derin öğrenme yöntemlerini kullanarak çevrimiçi işe alım dolandırıcılığını tespit etmiş ve geleneksel makine öğrenmesi yöntemlerine kıyasla anlamlı performans artışları elde etmiştir. Taneja vd. (2025), BERT tabanlı Fraud-BERT modelini geliştirerek bağlam duyarlı bir tespit yaklaşımı önermiş; Varshni vd. (2025) ise transformer modellerini SHAP açıklanabilirlik yöntemiyle birleştirerek model kararlarının yorumlanabilirliğini artırmıştır. Kumar vd. (2025) ise makine öğrenmesi ve doğal dil işleme tabanlı tespit yaklaşımlarını kapsamlı biçimde derleyerek alandaki güncel eğilimleri ortaya koymuştur. Bu çalışmada transformer tabanlı modeller kullanılmamış olmakla birlikte, söz konusu yaklaşımlar gelecek araştırmalar için önemli bir yön göstermektedir. Ancak mevcut çalışmaların büyük bir bölümü ya tekil algoritmalar üzerine yoğunlaşmış ya da sınıf dengesizliğinin farklı model mimarileri üzerindeki etkisini karşılaştırmalı olarak incelemekte sınırlı kalmıştır. Bu nedenle, metinsel ve yapısal özellikleri bütünleştiren, farklı algoritmaları ve dengesiz veri stratejilerini birlikte değerlendiren hibrit yaklaşımlara ihtiyaç duyulmaktadır.

3. YÖNTEM

Bu çalışmada, çevrimiçi platformlardaki sahte iş ilanlarının yüksek doğrulukla tespiti için metinsel ve yapısal (metadata) verilerin birlikte işlendiği hibrit ve karşılaştırmalı bir makine öğrenmesi çerçevesi sunulmuştur. Çalışmanın metodolojisi, ham verinin temizlenmesinden farklı algoritmaların performans kıyaslamasına (benchmarking) kadar uzanan uçtan uca bir boru hattı (pipeline) yaklaşımını benimsemektedir (Pedregosa vd., 2011). Önerilen çerçeve dört temel aşamadan oluşmaktadır. İlk aşamada metinsel gürültüler filtrelenerek eksik veri desenleri bilgi sinyali olarak yönetilmiştir. İkinci aşamada TF-IDF tabanlı metin temsilleri ile yapısal özellikler tek bir vektör uzayında birleştirilmiştir. Üçüncü aşamada Lojistik Regresyon, Random Forest ve XGBoost algoritmaları Maliyet Duyarlı Öğrenme ve SMOTE stratejileri altında eğitilmiştir. Son aşamada ise modellerin performansı güvenlik odaklı metrikler ve ROC analizi üzerinden karşılaştırmalı olarak değerlendirilmiştir.

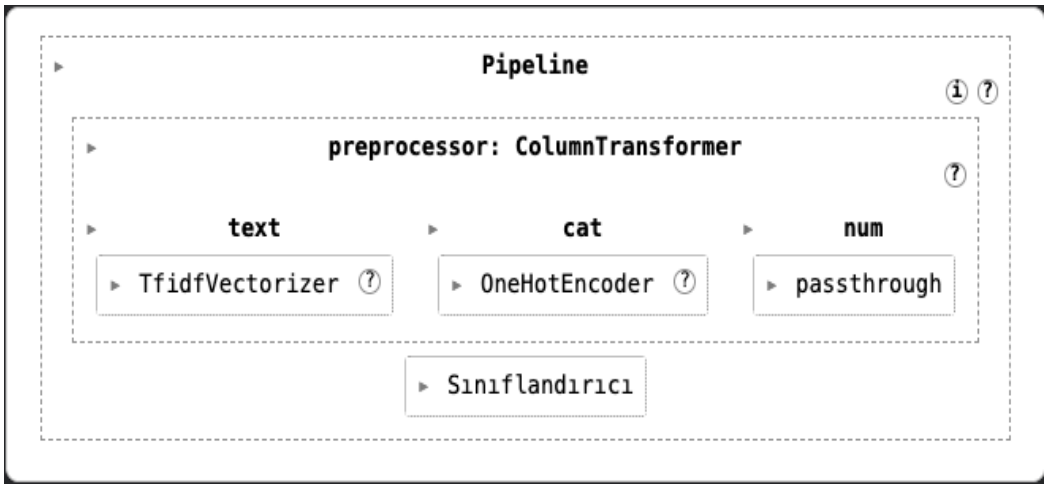
3.1. Veri Önişleme ve Özellik Yapılandırması

Analiz edilen veri seti; yapılandırılmamış metinsel içerikler, kategorik meta-bilgiler ve çeşitli eksik veri desenlerini bir arada barındıran heterojen bir yapıya sahiptir. Bu nedenle, ham veriyi analize hazırlamak için veri türüne özgü çok katmanlı bir ön işleme stratejisi izlenmiştir.

Veri temizleme aşamasında, eksik verilerin rastgele oluşmadığı, aksine sahte ilanlarda birer sinyal niteliği taşıyabileceği varsayımından hareket edilmiştir. Bu doğrultuda title, description, requirements ve company_profile gibi serbest metin alanlarındaki eksik değerler, TF-IDF vektörleştirmesinin teknik bütünlüğünü korumak amacıyla boş karakter dizisiyle doldurulmuştur. Kategorik değişkenlerdeki veri kaybı ise modelin bu durumu ayrı bir bilgi olarak öğrenmesine olanak tanımak için "Unknown" etiketiyle yeni bir sınıf olarak kodlanmıştır. Telecommuting ve has_company_logo gibi sayısal özelliklerdeki eksiklikler, yokluk durumunu temsil etmek üzere 0 değeriyle tamamlanmıştır.

Özellik mühendisliği aşamasında, veri setindeki parçalı metin alanları (başlık, açıklama, gereksinimler) tek bir text_all sütununda birleştirilerek modelin ilan içeriğini bütünsel biçimde değerlendirmesi sağlanmıştır. TfidfVectorizer aracılığıyla metinler küçük harfe dönüştürülmüş, İngilizce etkisiz kelimelerden arındırılmış ve vektör uzayına aktarılmıştır. Elde edilen metin vektörleri ile yapısal kategorik ve sayısal veriler, ColumnTransformer mimarisi aracılığıyla tek bir birleşik öznitelik uzayına dönüştürülmüştür (Pedregosa vd., 2011). Hazırlanan bu hibrit veri yapısı, çalışmada kullanılan üç farklı sınıflandırma algoritmasına (Lojistik Regresyon, Random Forest ve XGBoost) ayrı ayrı girdi olarak beslenmiştir. Şekil 1, oluşturulan pipeline mimarisini göstermektedir. Bu hibrit yapı; TF-IDF vektörlerinin sağladığı dilsel ayrım gücü ile One-Hot kodlanmış meta bilgilerin taşıdığı yapısal sinyalleri tek bir potada eriterek model başarısını artırmayı hedeflemektedir.

Şekil 1'de görüldüğü üzere, önerilen pipeline üç paralel dönüşüm kolundan oluşmaktadır. Birinci kolda (text) ham metin verisi TfidfVectorizer ile sayısal vektörlere dönüştürülmektedir. İkinci kolda (cat) employment_type, industry gibi kategorik değişkenler OneHotEncoder ile ikili vektörlere kodlanmaktadır. Üçüncü kolda (num) telecommuting, has_company_logo ve salary_provided gibi sayısal değişkenler herhangi bir dönüşüm uygulanmadan (passthrough) modele aktarılmaktadır. ColumnTransformer bu üç kolun çıktısını tek bir öznitelik vektöründe birleştirmekte, ardından bu birleşik vektör ilgili sınıflandırıcıya (Lojistik Regresyon, Random Forest veya XGBoost) girdi olarak verilmektedir.



Şekil 1. Çalışmada Kullanılan Hibrit Pipeline Mimarisi ve Sınıflandırma Süreci.

3.2. Özellik Çıkarımı

Modelin sahte ilanları yüksek başarıyla ayırt edebilmesi için, metinsel içeriklerin semantik gücü ile yapısal verilerin (metadata) ayırt ediciliğini birleştiren hibrit bir özellik seti türetilmiştir.

3.2.1. Metin Bazlı Özellikler

Birleştirilen metin verisi (text_all), sahte ilanlarda sıkça rastlanan dolandırıcılık odaklı kalıp ifadeleri (örn. "fast money", "no experience", "guaranteed income") yakalayabilmek amacıyla TF-IDF (Term Frequency-Inverse Document Frequency) yöntemiyle vektörleştirilmiştir. Vektör uzayı oluşturulurken hesaplama maliyeti ile gürültü dengesi gözetilerek korpus içerisindeki en ayırt edici 10.000 özellik seçilmiştir. Yalnızca tekil kelimeler değil kelime öbekleri de kapsanacak şekilde ngram_range=(1,2) parametresi belirlenmiş; anlamsal değeri düşük İngilizce etkisiz kelimeler (stop-words) ise vektörleştirme öncesinde listeden çıkarılmıştır.

3.2.2. Metadata ve Yapısal Özellikler

Çalışmada metinsel özelliklere ek olarak, literatürde sahtecilik sinyali taşıdığı belirtilen yapısal meta-bilgiler modele entegre edilmiştir. employment_type, required_experience, required_education, industry ve function değişkenleri, makine öğrenmesi algoritmalarının işleyebilmesi için One-Hot Encoding yöntemiyle ikili (binary) vektörlere dönüştürülmüştür. Telecommuting ve has_company_logo gibi mevcut değişkenlerin yanı sıra, ilan metinlerinin uzunluğunun dolandırıcılık tespiti için önemli bir ipucu olabileceği düşüncesiyle her bir metin alanı için kelime sayısı (word count) hesaplanarak özellik setine eklenmiştir. Bunlara ek olarak, sahte ilanlarda maaş bilgisinin genellikle gizlendiği gözleminden yola çıkılarak salary_range alanının doluluk durumuna göre salary_provided (0/1) adında ikili bir değişken türetilmiştir. Bu yaklaşım, eksik veriyi bir bilgi sinyali olarak ele alan missingness-as-information ilkesine dayanmaktadır (Little ve Rubin, 2019).

3.3. Sınıf Dengesizliği Probleminin Giderilmesi

Veri setindeki fraudulent sınıfı toplam verinin yalnızca yaklaşık %5'ini oluşturmakta olup, bu durum ciddi bir sınıf dengesizliğine (class imbalance) yol açmaktadır. Modellerin çoğunluk sınıfına (gerçek ilanlar) bias geliştirmesini önlemek ve "sahte" etiketini öğrenmelerini kolaylaştırmak için iki temel strateji karşılaştırmalı olarak uygulanmıştır:

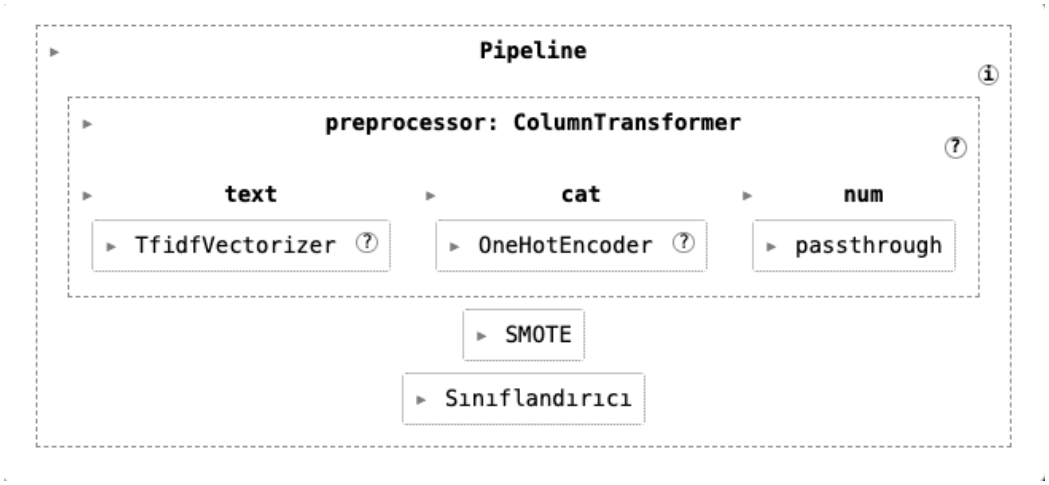
3.3.1. Maliyet Duyarlı Öğrenme (Class-Weight Adjustment)

Bu yaklaşımda, veri setini değiştirmek yerine algoritmaların maliyet fonksiyonlarına müdahale edilmiştir. Eğitim sürecinde azınlık sınıfa yapılan tahmin hatalarına daha yüksek ceza puanı atanarak modellerin duyarlılığı artırılmıştır. Bu strateji kapsamında Lojistik Regresyon ve Random Forest (Breiman, 2001) için class_weight='balanced' parametresi kullanılırken, XGBoost (Chen ve Guestrin, 2016) için eğitim setindeki pozitif ve negatif sınıf oranından hesaplanan scale_pos_weight=19.53 değeri atanmıştır.

3.3.2. Sentetik Örnekleme (SMOTE):

Eğitim setindeki dengesizliği veri düzeyinde gidermek amacıyla Synthetic Minority Oversampling Technique (SMOTE) kullanılmıştır (Chawla vd., 2002). SMOTE, azınlık sınıfına ait mevcut örneklerin k en yakın komşusu (k=5) arasında sentetik örnekler üreterek eğitim setini dengeler. Veri sızıntısını (data leakage) önlemek amacıyla SMOTE işlemi yalnızca eğitim setine uygulanacak şekilde imbalanced-learn kütüphanesinin pipeline yapısına entegre edilmiştir (Pedregosa vd., 2011). Şekil 2, SMOTE katmanının pipeline içindeki konumunu göstermektedir.

Şekil 2, SMOTE entegreli genişletilmiş pipeline mimarisini göstermektedir. Temel mimariden (Şekil 1) farklı olarak, ColumnTransformer'ın çıktısı doğrudan sınıflandırıcıya verilmemekte; araya eklenen SMOTE katmanı sentetik azınlık örnekleri üreterek eğitim setini dengelemektedir. Bu yapı sayesinde SMOTE yalnızca eğitim verisine uygulanmakta, test verisi dönüşümden etkilenmemekte ve veri sızıntısı (data leakage) riski ortadan kalkmaktadır. Dengeleme işleminin ardından elde edilen veri sınıflandırıcıya aktarılmaktadır.



Şekil 2. SMOTE Entegreli Genişletilmiş Pipeline Mimarisi.

3.4. Model Geliştirilmesi ve Algoritmik Yapı

Özellik mühendisliği aşamasında oluşturulan hibrit vektör uzayı, sınıflandırma algoritmalarına girdi olarak verilmiştir. Bu çalışmada, problemin karmaşıklığını ve dengesiz yapısını en iyi şekilde modelleyebilmek adına, farklı matematiksel temellere dayanan üç algoritma karşılaştırmalı (benchmark) olarak kullanılmıştır:

Lojistik Regresyon

Lojistik Regresyon, yüksek boyutlu ve seyrek veri yapılarında hesaplama açısından verimli sonuçlar üretmesi ve elde edilen katsayıların yorumlanabilirliği nedeniyle bu çalışmada temel referans modeli (baseline) olarak belirlenmiştir. TF-IDF tabanlı metin temsilleri doğası gereği seyrek matrisler ürettiğinden, Lojistik Regresyon bu tür verilerle teknik açıdan uyumlu bir seçimdir. Literatürde de

benzer metin sınıflandırma problemlerinde sıklıkla başlangıç noktası olarak tercih edilmekte ve daha karmaşık modellerin performansını değerlendirmek için karşılaştırma zemini oluşturmaktadır (Amaar vd., 2022).

Random Forest

Random Forest, birbirinden bağımsız karar ağaçlarının bir araya getirilmesiyle oluşturulan bir topluluk öğrenmesi yöntemidir (Breiman, 2001). Her ağacın veri setinin rastgele bir alt kümesi üzerinde eğitilmesi, modelin aşırı öğrenmeye (overfitting) karşı direncini artırmakta ve tahmin varyansını belirgin biçimde düşürmektedir. Sahte iş ilanı tespiti gibi gürültülü ve heterojen veri yapılarında bu özellik önemli bir avantaj sağlamaktadır. Alghamdi ve Alharby (2019) da aynı problemde Random Forest'ın yüksek ayırt edici güç sergilediğini göstermiştir.

XGBoost

XGBoost, gradyan artırma (gradient boosting) ilkesine dayanan ve zayıf öğrenicileri sıralı biçimde birleştirerek güçlü bir sınıflandırıcı oluşturan bir topluluk yöntemidir (Chen ve Guestrin, 2016). Metin özellikleri ile yapısal metadata arasındaki doğrusal olmayan ilişkileri modellemedeki başarısı ve sınıf dengesizliğine karşı scale_pos_weight parametresi aracılığıyla özelleştirilebilir yapısı, bu algoritmayı çalışma için uygun kılmaktadır. Ullah ve Jamjoom (2023) da aynı veri seti üzerinde XGBoost'un sınıf dengesizliği koşullarında güvenilir sonuçlar ürettiğini raporlamıştır.

Her bir algoritma için modelin başarısını ve dayanıklılığını ölçmek adına iki farklı eğitim senaryosu kurgulanmıştır. Birinci senaryoda veri setine herhangi bir müdahale yapılmadan algoritmaların kendi ağırlıklandırma parametreleri (class_weight veya scale_pos_weight) aracılığıyla azınlık sınıf hatalarına daha yüksek ceza uygulanmıştır. İkinci senaryoda ise pipeline içerisine entegre edilen SMOTE katmanı ile eğitim seti sentetik olarak dengelenerek modeller bu yeni veri dağılımı üzerinde eğitilmiştir (Chawla vd., 2002). Bu tasarım sonucunda toplamda 6 farklı deney (3 Model × 2 Senaryo) gerçekleştirilerek hangi stratejinin sahte ilan tespitinde en güvenilir sonuçları verdiği analiz edilmiştir.

3.5. Tekrar Edilebilirlik ve Hiperparametreler

Çalışmanın bağımsız olarak doğrulanabilmesi amacıyla veri bölme stratejisi, rastgelelik kontrolü ve model hiperparametreleri bu bölümde ayrıntılı biçimde açıklanmaktadır. Veri seti, sınıf oranlarını korumak amacıyla tabakalı örnekleme (stratified split) yöntemiyle %80 eğitim ve %20 test olarak ayrılmıştır. Bölme işleminde random_state=42 parametresi sabitlenmiş; SMOTE uygulamasında da aynı değer kullanılmıştır. Çalışmada tek bir sabit test seti üzerinde değerlendirme yapılmış olup çapraz doğrulama (cross-validation) uygulanmamıştır. Bu tercih, SMOTE'un yalnızca eğitim verisine uygulanması gerekliliğinden kaynaklanmakta; aksi hâlde her katlamada veri sızıntısı (data leakage) riski ortaya çıkmaktadır. Tüm modellere ait hiperparametre değerleri EK-1'de sunulmuştur.

3.6. Değerlendirme Metrikleri

Veri setindeki belirgin sınıf dengesizliği (%95 gerçek, %5 sahte ilan) göz önünde bulundurulduğunda, yalnızca doğruluk (accuracy) oranına odaklanmak yanıltıcı sonuçlar doğurabilmektedir. Zira bir model tüm ilanları "gerçek" olarak etiketlese dahi %95 doğruluk elde edebilir; oysa bu durumda sahte ilanların tamamı gözden kaçmış olur. Bu nedenle değerlendirmede güvenlik odaklı metriklere ağırlık verilmiştir.

Karışıklık matrisinden (confusion matrix) türetilen temel metrikler şu şekilde tanımlanmaktadır: Gerçekte sahte olan ve doğru tespit edilen ilanlar Doğru Pozitif (TP), gerçekte sahte olduğu halde gözden kaçan ilanlar Yanlış Negatif (FN), gerçek ilanların yanlışlıkla sahte olarak işaretlenmesi Yanlış Pozitif (FP) ve gerçek ilanların doğru sınıflandırılması Doğru Negatif (TN) olarak adlandırılmaktadır.

Bu temel değerlerden hareketle hesaplanan Recall (Duyarlılık) = TP / (TP + FN) formülüyle ifade edilmekte olup sahte ilanların ne kadarının sisteme sızdığını göstermesi bakımından güvenlik uygulamaları için en kritik metrik olarak kabul edilmektedir. Precision (Kesinlik) = TP / (TP + FP) ise sistemin meşru ilanları haksız yere engelleyip engellemediğini ölçmekte ve işveren memnuniyeti açısından önem taşımaktadır. F1-Score = 2 × (Precision × Recall) / (Precision + Recall) formülüyle hesaplanan harmonik ortalama, sınıf dengesizliği koşullarında Precision ve Recall arasındaki dengeyi tek bir değerle özetlemektedir. Son olarak ROC-AUC skoru, modelin karar eşiğinden (threshold) bağımsız olarak pozitif ve negatif sınıfları birbirinden ayırma yeteneğini karşılaştırmalı biçimde ortaya koymakta; 1'e yakın değerler neredeyse mükemmel bir ayrım gücüne işaret etmektedir.

4. BULGULAR VE TARTIŞMA

Geliştirilen hibrit tespit çerçevesinin etkinliği, ayrılan test veri seti üzerinde üç farklı algoritma (Lojistik Regresyon, Random Forest, XGBoost) kullanılarak karşılaştırmalı olarak değerlendirilmiştir. Deneysel analizler iki temel senaryo üzerine kurgulanmıştır. Birinci senaryoda veri setine herhangi bir müdahale yapılmadan algoritmalar maliyet parametreleri (Class Weight) ile eğitilmiştir. İkinci senaryoda ise eğitim aşamasında SMOTE uygulanarak sentetik veri ile dengeleme yapılmıştır. Bu bölümde her iki senaryonun model performansı üzerindeki etkileri sayısal metrikler, karışıklık matrisleri ve ROC eğrileri üzerinden detaylandırılmıştır.

4.1. Temel Modellerin Performansı

Geliştirilen hibrit çerçevenin performansı, %80 eğitim ve %20 test olarak ayrılan veri seti üzerinde altı farklı model konfigürasyonu için değerlendirilmiştir. Tablo 1, tüm modellere ait Precision, Recall, F1-Score ve AUC değerlerini karşılaştırmalı biçimde sunmaktadır.

Tablo 1. Altı Model Konfigürasyonuna Ait Karşılaştırmalı Performans Sonuçları

Model	**Precision**	**Recall**	**F1-Score**	**AUC**
Random Forest + SMOTE	1.0000	0.6994	0.8231	0.9934
Random Forest + Maliyet Duyarlı	0.9902	0.5838	0.7345	0.9914
XGBoost + Maliyet Duyarlı	0.9079	0.7977	0.8492	0.9892
XGBoost + SMOTE	0.9650	0.8671	0.8734	0.9922
Lojistik Regresyon + SMOTE	0.5869	0.8786	0.7037	0.9822
Lojistik Regresyon + Maliyet Duyarlı	0.4845	0.9017	0.6303	0.9812

Tablo 1 incelendiğinde, ağaç tabanlı topluluk modellerinin lineer modele kıyasla belirgin biçimde üstün performans sergilediği görülmektedir. En yüksek AUC değerine 0.9934 ile Random Forest + SMOTE modeli ulaşmıştır. Bu modelin Precision değerinin 1.0000 olması dikkat çekicidir; model test setinde hiçbir gerçek ilanı yanlışlıkla sahte olarak işaretlememiştir. Bununla birlikte Recall değerinin 0.6994 düzeyinde kalması, sahte ilanların yaklaşık %30'unun gözden kaçırıldığına işaret etmektedir.

Maliyet Duyarlı yaklaşımda eğitilen Random Forest modeli ise 0.9914 AUC ile ikinci sıraya yerleşmiştir. Bu modelde Recall 0.5838'e gerilemiş, ancak Precision yüksek kalmaya devam etmiştir. XGBoost + Maliyet Duyarlı ve XGBoost + SMOTE modelleri, Precision ve Recall arasında daha dengeli bir dağılım sergilemiştir. Özellikle XGBoost + SMOTE kombinasyonu 0.8671 Recall ve 0.9650 Precision değerleriyle her iki metrikte de kabul edilebilir düzeyde kalmayı başarmıştır.

Lojistik Regresyon modelleri ise her iki senaryoda da yüksek Recall sağlamasına karşın düşük Precision değerleri üretmiştir. Maliyet Duyarlı yaklaşımda Precision 0.4845'e kadar gerilemiş, bu da modelin gerçek ilanların yaklaşık yarısını yanlışlıkla sahte olarak etiketlediğine işaret etmektedir. SMOTE uygulaması Precision'ı 0.5869'a yükselterek belirli bir iyileşme sağlamış olsa da bu oran güvenlik odaklı bir sistem için hâlâ yüksek yanlış alarm riski taşımaktadır.

Confusion matrix görselleri ayrıntılı hata dağılımı açısından EK-1'de sunulmuştur.

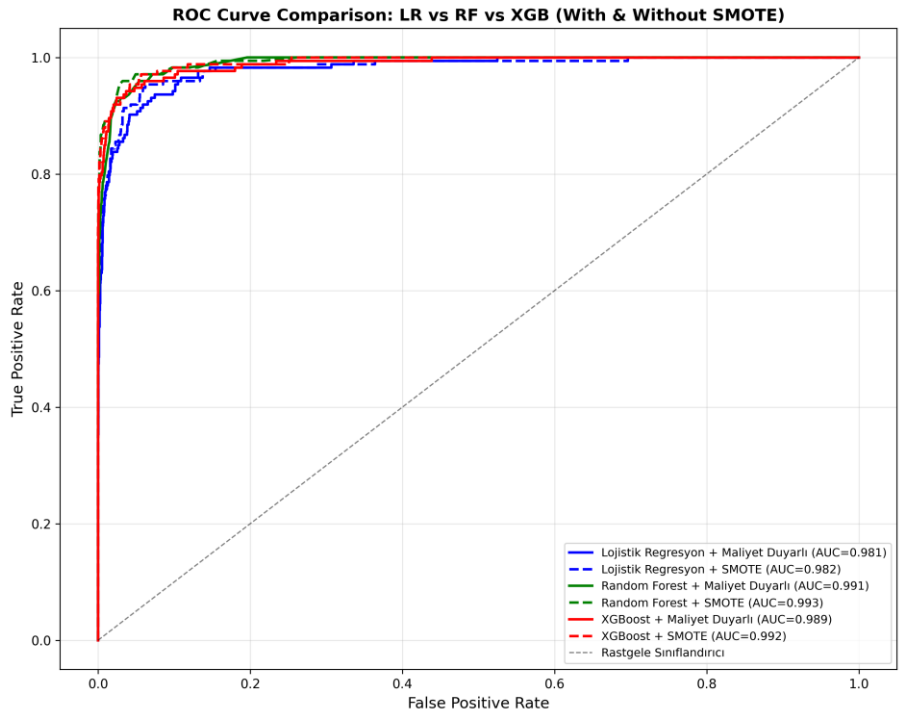
4.2. SMOTE Entegreli Modellerin Performansı

SMOTE'un model performansına etkisi algoritma mimarisine göre farklılık göstermiştir. Lojistik Regresyon modelinde etki oldukça belirgin biçimde gözlemlenmiş; Yanlış Pozitif sayısı 166'dan 107'ye gerilerken Precision değeri 0.49'dan 0.59'a yükselmiştir. Bununla birlikte bu modelin Recall değeri 0.90 düzeyinde kalmış, yüksek duyarlılık korunmuştur. Ağaç tabanlı modellerde ise SMOTE daha ölçülü ancak anlamlı katkılar sağlamıştır. Random Forest modelinde AUC 0.9914'ten 0.9934'e yükselmiş, Precision değeri 1.0000'e ulaşarak test setinde hiç yanlış alarm üretilmemiştir. XGBoost modelinde ise SMOTE ile Precision 0.9079'dan 0.9650'ye yükselmiş, Recall 0.7977 düzeyinde korunmuş; böylece yüksek kesinlik ve kabul edilebilir duyarlılık dengesi sağlanmıştır. Genel olarak SMOTE, modellerin yalnızca çoğunluk sınıfını ezberlemesini engelleyerek karar sınırlarını gerçek ve sahte ilanlar arasında daha net bir konuma taşımıştır. Confusion matrix görselleri EK-1'de sunulmuştur.

4.3. ROC Analizi ve Genel Ayrım Gücü

Geliştirilen modellerin pozitif (sahte) ve negatif (gerçek) sınıfları birbirinden ayırma yeteneği, sınıflandırma eşik değerinden bağımsız olarak ROC (Receiver Operating Characteristic) eğrileri üzerinden karşılaştırmalı olarak analiz edilmiştir. Şekil 3'te sunulan ROC eğrileri incelendiğinde, analize dahil edilen tüm modellerin uygulanan senaryodan bağımsız olarak 0.98'in üzerinde AUC skoru elde ettiği görülmektedir. Bu durum, oluşturulan hibrit öznitelik uzayının sahtecilik tespitinde güçlü bir ayrım gücü sağladığının göstergesidir.

Modeller arasında en yüksek genel ayrım gücüne 0.9934 AUC değeri ile Random Forest + SMOTE modeli ulaşmıştır. Onu 0.9914 AUC ile Maliyet Duyarlı Random Forest modeli takip etmektedir. Ağaç tabanlı modellerin (RF ve XGB) ROC eğrilerinin sol üst köşeye lineer modellere kıyasla daha dik bir açıyla yakınsadığı, yani yanlış pozitif oranını düşük tutarken yüksek duyarlılık sağladığı gözlemlenmiştir. Lojistik Regresyon modelleri ise daha geniş bir eğri çizerek daha fazla yanlış alarm üretme eğiliminde olduğunu ortaya koymuştur.



Şekil 3. Karşılaştırmalı ROC Eğrileri.

4.4. Bulguların Değerlendirilmesi / Tartışma

Elde edilen bulgular bir bütün olarak değerlendirildiğinde, önerilen hibrit mimarinin sahte iş ilanı tespitinde güçlü bir performans sergilediği görülmektedir. Literatürde genellikle göz ardı edilen missingness-as-information prensibinin bu çalışmada da geçerliliği doğrulanmıştır. Maaş bilgisinin eksikliği veya meta-verilerdeki boşluklar, metin madenciliği çıktılarıyla birleştirildiğinde modellerin

ayırt ediciliğini artırmıştır. Tüm modellerin 0.98 üzerinde AUC skoruna ulaşması, oluşturulan hibrit öznitelik setinin dilden bağımsız güçlü sinyaller taşıdığını ortaya koymaktadır.

Çalışmanın en çarpıcı bulgusu, ağaç tabanlı topluluk algoritmalarının (RF ve XGBoost) lineer modele kıyasla belirgin biçimde daha kararlı sonuçlar üretmesidir. Lojistik Regresyon yüksek Recall sağlasa da karar sınırını çizerken çok sayıda gerçek ilanı sahte olarak etiketlemiş ve düşük Precision değerleri üretmiştir. Buna karşılık Random Forest ve XGBoost, metin ve metadata arasındaki karmaşık doğrusal olmayan ilişkileri daha iyi modelleyerek yanlış alarm sayısını minimize etmiştir. Özellikle Random Forest, güvenlik sistemleri için kritik olan düşük gürültü ve yüksek tespit dengesini en iyi sağlayan algoritma olmuştur.

Sınıf dengesizliğinin giderilmesinde SMOTE tekniği model mimarisine göre farklı düzeylerde katkı sağlamıştır. Lojistik Regresyon gibi doğrusal karar sınırlarıyla çalışan modellerde bu etki oldukça belirgin biçimde gözlemlenmiş; Yanlış Pozitif sayısı 166'dan 107'ye gerilerken Precision değeri 0.49'dan 0.59'a yükselmiştir. Dengesiz bir veri setinde bu iyileşme, sistemin meşru ilanları haksız yere engellemesinin önüne geçmesi bakımından pratik önem taşımaktadır. Zaten yüksek performans sergileyen Random Forest modelinde ise SMOTE daha mütevazı ancak anlamlı bir katkı sunmuş; AUC skoru 0.9914'ten 0.9934'e yükselmiş ve modelin genelleme yeteneği pekişmiştir. Bu bulgular, SMOTE'un yalnızca azınlık sınıfını ezberletmekle kalmayıp karar sınırlarını gerçek ve sahte ilanlar arasında daha net bir konuma taşıdığına işaret etmektedir.

Elde edilen sonuçlar, alana yakın zamanda katkı sağlayan güncel çalışmalarla da karşılaştırmalı biçimde değerlendirilmiştir. Akram vd. (2024), derin öğrenme tabanlı yaklaşımlarla aynı veri seti üzerinde yüksek tespit performansı elde etmiş olmakla birlikte, bu çalışmada önerilen hibrit çerçeve geleneksel makine öğrenmesi algoritmalarıyla karşılaştırılabilir AUC değerlerine ulaşmıştır. Taneja vd. (2025) tarafından geliştirilen Fraud-BERT modeli, transformer tabanlı bağlam duyarlı bir yaklaşım sunmakta ve güçlü sınıflandırma performansı sergilemektedir; ancak bu tür modeller hesaplama maliyeti ve yorumlanabilirlik açısından ek gereksinimler doğurmaktadır. Varshni vd. (2025) ise transformer modellerini SHAP açıklanabilirlik yöntemiyle birleştirerek model kararlarının yorumlanabilirliğini artırmış; bu yaklaşım özellikle gerçek dünya uygulamalarında önemli bir avantaj sağlamaktadır. Kumar vd. (2025) tarafından gerçekleştirilen kapsamlı derleme çalışması, alandaki yöntemleri sistematik biçimde ele almakta ve makine öğrenmesi ile derin öğrenme tabanlı yaklaşımların birbirini tamamlayıcı nitelikte olduğunu ortaya koymaktadır. Bu karşılaştırma, önerilen çerçevenin hesaplama açısından verimli ve yorumlanabilir bir alternatif sunduğunu; transformer tabanlı modellerin ise özellikle yüksek kaynak ortamlarında daha ileri performans hedefleri için gelecek araştırmalara yön gösterdiğini teyit etmektedir.

5. SONUÇ VE ÖNERİLER

Bu çalışmada, çevrim içi istihdam platformlarının güvenilirliğini tehdit eden sahte iş ilanlarının tespiti amacıyla metinsel ve yapısal öznitelikleri bütünleştiren hibrit bir makine öğrenmesi mimarisi önerilmiş ve üç farklı algoritma (Lojistik Regresyon, Random Forest, XGBoost) üzerinde kapsamlı bir kıyaslama yapılmıştır.

Gerçek dünya verileri üzerinde gerçekleştirilen deneysel analizler, salt metin madenciliği yerine eksik veri örüntülerinin (missingness patterns) ve meta-verilerin modele dahil edilmesinin ayırt edici gücü artırdığını kanıtlamıştır. Çalışmanın en çarpıcı bulgusu, ağaç tabanlı modellerin (özellikle Random Forest) lineer modellere göre sahtecilik tespitinde çok daha kararlı olduğudur. SMOTE tekniği, Lojistik Regresyon modelinde yanlış alarm sayısını 166'dan 107'ye düşürerek Precision değerini anlamlı biçimde iyileştirirken; Random Forest modelinde 0.9934 AUC skoruna ulaşılmasını sağlayarak sistemin yüksek bir ayrım gücüne sahip olduğunu doğrulamıştır.

Çalışmanın bazı sınırlılıkları da göz önünde bulundurulmalıdır. Elde edilen yüksek AUC değerleri veri setinin yapısına özgü güçlü korelasyonlardan kaynaklanıyor olabilir; bu nedenle sonuçların farklı veri setleri üzerinde doğrulanması önem taşımaktadır. Gelecek çalışmalarda metin temsili için BERT tabanlı derin dil modellerinin mimariye entegre edilmesi, transformer tabanlı yaklaşımlarla karşılaştırmalı analizlerin genişletilmesi ve sistemin dilden bağımsız (language-agnostic) yeteneklerinin geliştirilmesi hedeflenmektedir.

KAYNAKÇA

Akram, N., Irfan, R., Al-Shamayleh, A. S., Kousar, A., Qaddos, A., Imran, M., & Akhunzada, A. (2024). Online recruitment fraud (ORF) detection using deep learning approaches. *IEEE Access, 12*, 109388–109408. https://doi.org/10.1109/ACCESS.2024.3435670

Alghamdi, B. M., & Alharby, F. (2019). An intelligent model for online recruitment fraud detection. *Journal of Information Security*, *10*(3), 155–166. https://doi.org/10.4236/jis.2019.103009

Amaar, A., Aljedaani, W., Rustam, F., Ullah, S., Rupapara, V., & Ludi, S. (2022). Detection of fake job postings by utilizing machine learning and natural language processing approaches. *Neural Processing Letters*, *54*(3), 2219–2247. https://doi.org/10.1007/s11063-021-10727-z

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357. https://doi.org/10.1613/jair.953

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

Europol. (2021). *Internet organised crime threat assessment (IOCTA) 2021*. Europol. https://www.europol.europa.eu

Federal Trade Commission (FTC). (2023). *Consumer Sentinel Network data book 2022*. Federal Trade Commission. https://www.ftc.gov

Holt, T. J., & Bossler, A. M. (2015). *Cybercrime in progress: Theory and prevention of technology-enabled offenses*. Routledge.

Kim, S. H., Kim, H., & Kim, H. (2019). Fraud detection for job placement using hierarchical clusters-based deep neural networks. *Applied Intelligence, 49*(8), 2842–2858. https://doi.org/10.1007/s10489-019-01419-2

Kumar, S., Yasmeen, S., & Kumar, P. (2025). The growing threat of fake job postings: A review of machine learning and NLP-based detection approaches. *Revolutionary Advances in Computing and Electronics: An International Journal*, 41–51.

Lal, S., Jiaswal, R., Sardana, N., Verma, A., Kaur, A., & Mourya, R. (2019). ORFDetector: Ensemble learning based online recruitment fraud detection. In *Proceedings of the 2019 International Conference on Computing, Communication, and Intelligent Systems (IC3)* (pp. 1–6). IEEE. https://doi.org/10.1109/IC3.2019.8844879

Little, R. J. A., & Rubin, D. B. (2019). *Statistical analysis with missing data* (3rd ed.). Wiley. https://doi.org/10.1002/9781119482260

Mahbub, S., Pardede, E., & Kayes, A. S. M. (2022). Online recruitment fraud detection: A study on contextual features in Australian job industries. *IEEE Access, 10*, 82776–82788. https://doi.org/10.1109/ACCESS.2022.3197225

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830. https://doi.org/10.5555/1953048.2078195

Taneja, K., Vashishtha, J., & Ratnoo, S. (2025). Fraud-BERT: Transformer based context aware online recruitment fraud detection. *Discover Computing, 28*(1),9 https://doi.org/10.1007/s10791-025-09502-8

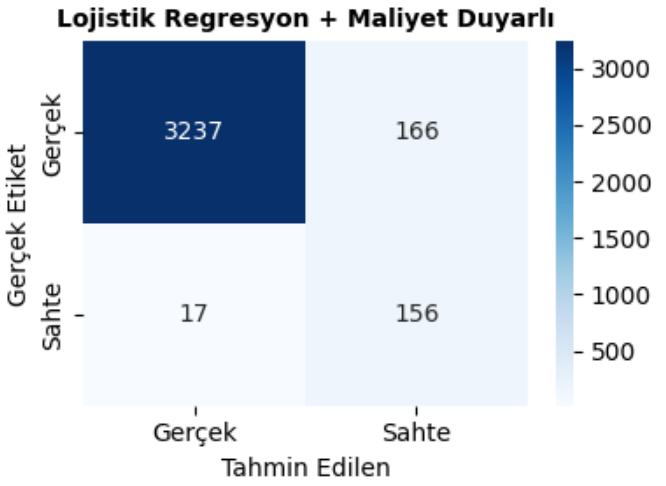
Ullah, Z., & Jamjoom, M. (2023). A smart secured framework for detecting and averting online recruitment fraud using ensemble machine learning techniques. *PeerJ Computer Science*, *9*, e1234. https://doi.org/10.7717/peerj-cs.1234

Varshni, I., Dharshini, N., & Kaviya, M. (2025). Online recruitment fraud (ORF) detection using transformer models with SHAP explainability. In *Proceedings of the IEEE International Conference on Signal Processing, Computation, Electronics, Power and Telecommunication (IConSCEPT)* (pp. 1–8). IEEE.

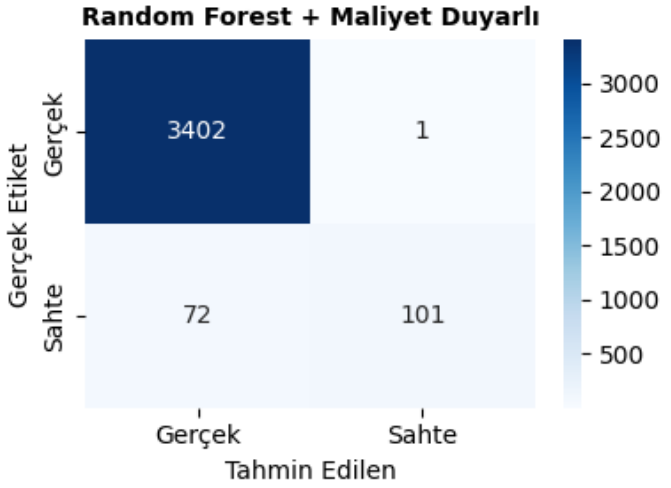
Vidros, S., Kolias, C., Kambourakis, G., & Akoglu, L. (2017). Automatic detection of online recruitment frauds: Characteristics, methods, and a public dataset. *Future Internet*, *9*(1), 6. https://doi.org/10.3390/fi9010006

EK-1. MODEL HİPERPARAMETRELERİ VE KARIŞIKLIK MATRİSLERİ

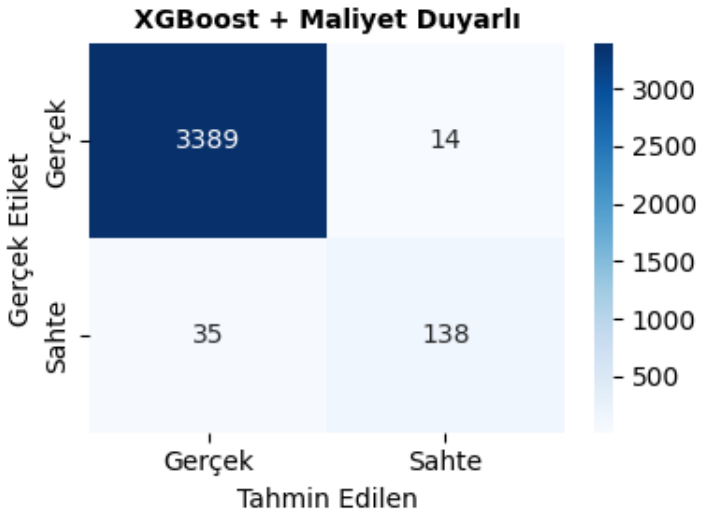
Bileşen	**Parametre**	**Değer**
Veri Bölme	test_size	0.20
Veri Bölme	random_state	42
Veri Bölme	stratify	y
TF-IDF	max_features	10.000
TF-IDF	ngram_range	(1, 2)
TF-IDF	stop_words	english
Lojistik Regresyon	class_weight	balanced
Lojistik Regresyon	max_iter	1000
Lojistik Regresyon	random_state	42
Random Forest	class_weight	balanced
Random Forest	n_estimators	100
Random Forest	random_state	42
XGBoost	scale_pos_weight	19.53
XGBoost	eval_metric	logloss
XGBoost	random_state	42
SMOTE	k_neighbors	5
SMOTE	random_state	42



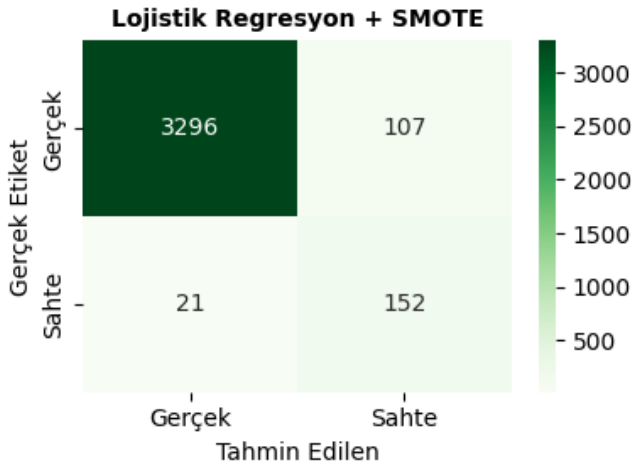
Şekil EK-1.1. Maliyet Duyarlı Lojistik Regresyon Modeline Ait Karışıklık Matrisi.



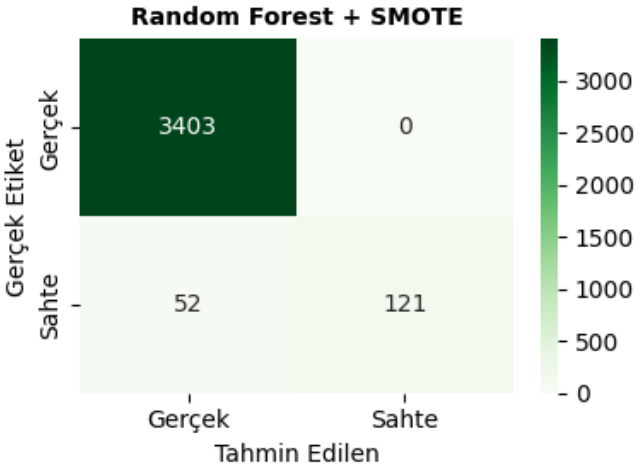
Şekil EK-1.2. Maliyet Duyarlı Random Forest Modeline Ait Karışıklık Matrisi.



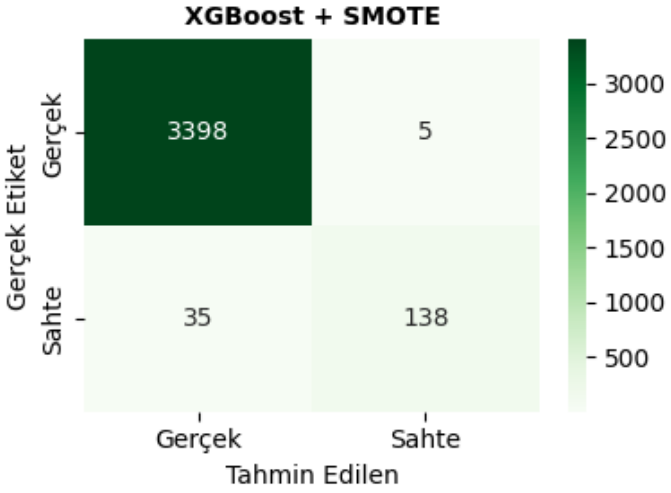
Şekil EK-1.3. Maliyet Duyarlı XGBoost Modeline Ait Karışıklık Matrisi.



Şekil EK-1.4. SMOTE Destekli Lojistik Regresyon Modeline Ait Karışıklık Matrisi.



Şekil EK-1.5. SMOTE Destekli Random Forest Modeline Ait Karışıklık Matrisi.



Şekil EK-1.6. SMOTE Destekli XGBoost Modeline Ait Karışıklık Matrisi.

EXTENDED ABSTRACT

A HYBRID MACHINE LEARNING FRAMEWORK FOR FAKE JOB POSTING DETECTION USING TEXTUAL AND METADATA FEATURES

The rapid expansion of online recruitment platforms has significantly transformed hiring processes by increasing accessibility and efficiency for both employers and job seekers. However, this transformation has also created new opportunities for fraudulent actors who publish deceptive job advertisements for financial gain, identity theft, or phishing purposes. Fake job postings have become an important cybersecurity and trust issue in digital labor markets. Such advertisements often imitate legitimate recruitment announcements and may be difficult for users to distinguish manually. Therefore, developing automated and reliable detection systems has become increasingly important for protecting applicants and preserving the credibility of online employment platforms.

Previous studies on fake job posting detection have mainly focused on textual analysis, keyword extraction, linguistic patterns, or limited metadata variables. Although these approaches provide useful insights, relying only on textual content may overlook structural signals embedded in job advertisements. Features such as missing salary information, absence of company logos, incomplete firm profiles, unusual employment categories, or inconsistent metadata may also indicate suspicious activity. In this context, combining textual and structured information may improve classification performance and increase model robustness.

This study proposes a hybrid machine learning framework that integrates textual and metadata based features for fake job posting detection. The publicly available “Real or Fake Job Posting Prediction” dataset was used as the empirical basis of the study. One of the main challenges of this dataset is severe class imbalance, since fraudulent postings represent only a small proportion of the total observations. Imbalanced data structures may lead models to favor the majority class and fail to identify fraudulent cases accurately. For this reason, imbalance handling strategies were incorporated into the proposed framework.

During the preprocessing stage, textual fields such as job title, description, requirements, and company profile were combined into a single text corpus. TF-IDF vectorization was then applied to transform textual information into numerical representations. In addition, several metadata variables were included in the model, such as company logo presence, telecommuting status, salary disclosure, employment type, required experience, and education level. Missing data patterns were not ignored; instead, they were treated as potentially informative signals and converted into usable indicators.

To evaluate classification performance, three machine learning algorithms were compared: Logistic Regression, Random Forest, and XGBoost. These algorithms were selected because they represent different modeling perspectives, including linear learning and ensemble tree-based methods. Two imbalance management strategies were tested. First, cost-sensitive learning was implemented by

assigning higher penalties to minority class errors. Second, the Synthetic Minority Oversampling Technique (SMOTE) was used to generate synthetic minority class samples within the training process.

The empirical findings demonstrate that ensemble-based methods outperform the linear baseline model in terms of discrimination power and overall stability. Logistic Regression showed high sensitivity but produced relatively higher false positive rates. Random Forest and XGBoost provided more balanced results by reducing false alarms while maintaining strong recall levels. Among all tested configurations, Random Forest combined with SMOTE achieved the best overall performance with an AUC value of 0.993. This result indicates that the proposed hybrid framework can distinguish genuine and fraudulent postings with near-perfect accuracy.

The results also highlight the practical importance of metadata engineering and missingness indicators. Structural variables that may appear secondary at first glance contributed substantially to model success when combined with textual representations. This suggests that fraud detection systems should not rely exclusively on language patterns but instead adopt multidimensional feature architectures.

From a managerial perspective, the proposed model can be integrated into online recruitment platforms as an early warning or decision support system. Suspicious advertisements can be flagged automatically before publication or routed to human moderators for further review. Such applications may reduce reputational risks for platforms and protect job seekers from potential scams.

In conclusion, this study shows that hybrid machine learning architectures offer a highly effective solution for fake job posting detection under imbalanced data conditions. Future studies may extend this framework by incorporating transformer-based language models such as BERT, multilingual datasets, and real-time streaming environments.