T.C.
EGE ÜNİVERSİTESİ
FEN FAKÜLTESİ
MATEMATİK BÖLÜMÜ

LİSANS TEZİ

MAKİNE ÖĞRENMESİ YARDIMIYLA KANAL ANALİTİĞİ VERİLERİNİN ETİKETLENMESİ

Hazırlayan:
AHMET BARAN BOZKURT
11210000045

Danışman:
Doç. Dr. Gülnaz BORUZANLI EKİNCİ

Bornova-İzmir
2026

ÖNSÖZ

Bu çalışmanın planlanması, yürütülmesi ve her aşamasında değerli akademik bilgi ve tecrübeleriyle bana yol gösteren, karşılaştığım zorluklarda yardım ve sabrını benden esirgemeyen saygıdeğer danışman hocam Doç. Dr. Gülnaz BORUZANLI EKİNCİ'ye en içten şükranlarımı sunarım.

Projenin endüstriyel bakış açısıyla şekillenmesinde ve gerçek dünya problemleriyle ilişkilendirilmesinde değerli katkılar sağlayan sanayi danışmanım Mine Şener Girgin'e teşekkürlerimi sunarım. Ayrıca eğitim hayatım boyunca bilgi ve birikimlerinden yararlandığım Ege Üniversitesi Matematik Bölümü'nün tüm değerli öğretim üyelerine teşekkürü bir borç bilirim.

Son olarak, bugünlere gelmemde en büyük pay sahibi olan; hayatımın her anında sevgi, sabır ve güvenleriyle beni daima daha iyisini yapmaya teşvik eden, maddi ve manevi desteklerini hiçbir zaman esirgemeyen değerli aileme sonsuz teşekkürlerimi sunarım.

AHMET BARAN BOZKURT
Bornova-İzmir, 2026

İçindekiler

Tablo Listesi

Şekil Listesi

1. GİRİŞ

Medeniyet tarihinin başlangıcından bu yana ticaret, insanlığın ve toplumun ayrılmaz bir parçası olmuştur. Günümüzde iş dünyası ve ticaret oldukça rekabetçi bir hal almıştır. Bu rekabetçi ortamda kendilerini sürdürülebilir kılmak isteyen ve işlerini büyütmeyi ya da sürdürmeyi hedefleyen şirketler, bunu gerçekleştirmenin birden fazla yolunu aramakta ve keşfetmektedir.

Pazardaki yerlerini korumak isteyen şirketler, yalnızca sundukları ürünün veya hizmetin kalitesini değil aynı zamanda müşterileriyle olan ilişkilerini de gözetmek durumundadır. Müşteri ilişkilerinin kalitesi, şirketlerin varlıklarının devamı ve büyüleri ile doğrudan orantılıdır. Müşterilerin davranışlarını iyi analiz eden ve onları memnun edecek ürünleri ile kampanyaları sunan şirketler, hem talepte artış yaşayacak hem de yeni müşteriler kazanma konusunda önemli yol kat edecektir. Mevcut müşterilerin davranış analizlerini doğru yapmak ve buna göre özel pazarlama stratejileri geliştirmek, yeni müşteri kazanımı ve ürün sunumları açısından kritik öneme sahiptir.

Çalışmanın içeriği olarak kişiselleştirilmiş pazarlama planlamaları kapsamında yapılan analizler ve sonuçları incelenmiş; buna istinaden bir müşteri veri seti üzerinde bu analizler ve teknikler kullanılarak bir makine öğrenmesi modeli tasarlanmış ve müşterilere özel bir öneri sistemi örneği oluşturulmuştur.

Bu çalışmanın amacı, şirketin çeşitli iletişim kanalları (e-posta, SMS, uygulama bildirimi vb.) aracılığıyla kullanıcılara gönderdiği mesajlara verilen tepkileri analiz etmek ve bu tepkiler doğrultusunda kampanya performansını artıracak öngörüler geliştirmek olarak belirlenmiştir. Kullanıcıların mesajları açma, okuma, spam’e düşme, yanıt vermeme, bağlantı tıklama gibi davranışları ya da bu davranışların zaman, lokasyon, sıklık gibi özellikleri incelenerek kampanyaların hangi kullanıcılar tarafından okunacağının tahmin edilmesi hedeflenmiştir. Bu kapsamda mesajlara verilen tepkiler etiketlenmiş, kullanıcıların gelecekteki kampanya mesajlarına verecekleri tepkiler olasılık hesaplamalarıyla tahmin edilmiştir. Bu amaçla elde edilen veriler üzerinden bir makine öğrenmesi modeli eğitilmiş ve %20 oranında ayrılmış test verisi üzerinde modelin başarımı ölçülmüştür.

Modelin performansı değerlendirilirken doğruluk oranı (accuracy), duyarlılık (recall), kesinlik (precision) ve F1 skoru (F1-score) gibi değerlendirme kriterleri kullanılmıştır. Bu sayede

kampanya mesajlarının doğru hedef kitleye yönlendirilmesi ve kullanıcı etkileşim oranlarının artırılması sağlanarak pazarlama stratejileri daha etkin hale getirilmiştir. Ayrıca benzer kişilerin etkileşim gösterdiği iletiler ya da benzer mesajlara aynı kişilerin etkileşim gösterme olasılıkları göz önünde bulundurularak elde edilen modeller geliştirilmiş; bu iyileştirme hedefi için öneri sistemlerinde kullanılan yaklaşımlardan çalışmaya uygun yöntem tercih edilmiştir.

Çalışma, firmanın çeşitli kanallardan topladığı verilerin makine öğrenmesi yardımıyla bağlam kazanmasına ve çeşitli öngörülerde bulunmasına olanak sağlamıştır. Hem kanal analizindeki davranış örüntüsü benzer olan kişiler arasında bir benzerlik üzerinde çalışılmış hem de pazarlama tarafından etiketlenmiş iletilerin arasındaki benzerlik kullanılarak benzer etiketlere benzer yanıtlar gösteren kişiler açısından değerlendirme yapılmıştır. Şimdiye kadar herhangi bir analitik uygulanmadan iletilen mesajların, makine öğrenmesi ve öneri sistemi yaklaşımları ile incelenerek kampanya başarısını ve ileti okunma/tıklanma skorlarını artırması hedeflenmiştir.

Müşteri yaşam boyu değeri (CLV: Customer Lifetime Value), satıcının müşteriyle olan ilişkisini yönetme biçiminde kritik bir kavramdır. Satıcılar, müşterilerinin bu yaşam boyu değerlerini kullanarak onlara özel kampanyalar ve pazarlama stratejileri tasarlar. Müşterilerin o işletmeden olan beklentilerine odaklanarak o müşteriden elde edilecek kârı en iyilemeyi hedefleyen bu kavram oldukça önemlidir (Khajvand, 2011). Müşteri yaşam boyu değeri üzerinde yıllardır süregelen araştırmalar yapılmış ve bu konuyla ilgili pazarlama alanında uzman kişiler ile akademisyenler çalışmalar gerçekleştirmiştir. Şirketin müşterilerinin önem sıralarını belirlemesi ve bu bağlamda hangi müşterilere ne kadar yatırım yapması gerektiği sorularına yanıt arayan bir kavramdır.

Müşteri yaşam boyu değeri, müşteri kârlılığı kavramından bazı ince çizgilerle ayrılmaktadır. Müşteri kârlılığı, müşterilerin geçmişte yaptıkları harcama verilerine dayanırken müşteri yaşam boyu değeri bu verileri analiz ederek gelecekte müşterilerin nasıl davranışlarda bulunacağını öngörmeye çalışır.

Müşterilerin verilerinin anlamlı bilgilere dönüştürülmesinde etkin rol oynayan müşteri ilişkileri yönetimi (Customer Relationship Management – CRM), müşterilerin yaşam boyu değeri noktasında bazı kazanımları elde etmemize yarayan bir yaklaşımdır. Müşteri olmayan kişiler öncelikle uygun pazarlama anlayışıyla müşteri olarak kazanılmaya çalışılır. Ardından müşteriyi olabileceği en iyi satın alma ve memnuniyet derecesine getirmek için uğraşılır ve sadık

müşteriler elde edilmeye çalışılır. Bu gayedeki uygulamalardan biri olan müşteri segmentasyonu, müşteri tabanını ortak özelliklere sahip alt gruplara ayırarak her gruba özel stratejilerin geliştirilmesine olanak tanır. Segmentasyon çalışmaları, hem pazarlama kaynaklarının daha etkin kullanılmasını hem de müşteri memnuniyetinin artırılmasını hedefler.

Bu çalışmada 48.000 sentetik müşteri kaydı kullanılarak uçtan uca bir makine öğrenmesi ve sistem tasarımı geliştirilmiştir. Aşırı dengesiz bir veri dağılımına sahipken kritik derecede öneme sahip olan bir veri sızıntısı tespit edilmiş ve bunun çözülmesi üzerine kullanılan yöntemler ve senaryolara ilerleyen kısımlarda yer verilmiştir. Sahte başarı oranlarından arındırılması için veri üzerinde uygulanan tekniklere ve sistemin oluşum ve dönüşüm öyküsüne dayanan bu doküman, yenilikçi ve uygulanabilir yöntemlerle veri analizi ve yönetimi adına ayrıntılı bir çalışmayı içermektedir.

2. LİTERATÜR TARAMASI

Bu bölümde çalışmada yer alan uygulamanın kuramsal temelleri ve literatür araştırmasındaki bulgular ele alınmaktadır.

2.1. Müşteri Davranışı ve Segmentasyon Kavramı

Müşteri davranışlarını analiz etmek, bu müşterilerin segmentasyonu için bir ön unsurdur. Müşterilerin yaş, cinsiyet, ilgi alanları, yaptıkları alışverişler ve harcama alışkanlıkları, yaşadıkları bölgeler gibi özellikleri, bu müşterileri pazarlama açısından benzer özelliklere sahip gruplara ayırmaya ve segmentasyonunu sağlamaya yarar.

Müşteri segmentasyonunun amacı, pazardaki eğilimleri öngörmek, müşteri davranışlarını değerlendirmeye yardımcı olmak ve pazardaki durumu korumak ile iyileştirmek için gereken eylemleri belirlemektir. Mevcut müşterileri korumak ve potansiyel müşterileri çekmek için müşterilerin ihtiyaçlarını karşılamak hayati önem taşır. Müşterinin ihtiyaçlarını anlamanın yolu alışkanlıklarını bilmek, ilgi alanlarını belirlemek ve eğilimlerini doğru yorumlamaktır. Müşterileri, önceliklendirilmiş bileşenler etrafındaki ortak davranışlara göre segmentlere ayıran

iyi tanımlanmış bir müşteri segmentasyonu, hedef kitlenin tanınmasını ve pazarlama stratejileriyle uyumlu ihtiyaçların belirlenmesini sağlar.

Çevresel, demografik, davranışsal ve zamana bağlı verilerin birbirlerine uygun biçimlerde entegre edilip analiz edilmesi, müşterileri memnun etme yolunda uygun bir segmentasyon sistemi kurulmasında kullanılır. Söz konusu veriler, müşterinin davranış temelli verileri olabilir. Örneğin müşterilerin alışveriş yaptıkları ürünler, bu ürünleri alırken kullandıkları ödeme yöntemi, bir ürün üzerinde ne kadar inceleme yaptığı, hangi ürünleri favorilerine eklediği gibi girdi bilgileri bu kapsamda yer alabilir. Müşterilerin istek ve ihtiyaçlarını karşılamak, şirketin kârlılığını artıracağından bu sistemi uygulamak oldukça önemlidir.

Farklı pazarlama şirketleri tarafından müşteri portföylerine uygun birbirinden farklı kümeleme ve segmentasyon teknikleri kullanılarak kendi müşteri veri kümeleri üzerinde en optimal sonuca ulaşmak mantıklı bir seçenektir. Müşteri yaşam boyu değeri ve müşteri ilişki yönetimi teknikleriyle harmanlanan bu anlayışta veriler, matematik ve istatistik biliminin kullanılması süreciyle istenen sonuçlara ulaşılıp müşteriler segmente edilir, analizler yapılır ve bunlara uygun pazarlama ve satış stratejileri planlanır ve uygulanır.

2.1.1. Pazarlama Analitiğinin Tarihsel Gelişimi

Pazarlama çabası insanlık tarihinin başlangıcına kadar dayanmakta, para kavramının olmadığı dönemlerde bile insanlar tarafından uygulanmaktaydı. İnsanların karşılıklı alışveriş yaparak pazarlama ve ticaret olgusunu farkında olmadan yaratmasıyla başlayan süreç, günümüze kadar gelişerek genel satış ve pazarlama faaliyetlerinin temelini oluşturmuştur. Geleneksel medya ile birlikte televizyon, radyo ve gazeteler veri toplama ve anketler için birer araç olsa da bu araçlar veri toplama ve analiz derinliği aşamasında sınırlayıcı kaynaklar olmuştur. İnternetin ortaya çıkması ve dijitalleşen dünya ile birlikte bilgi işleme, yalnızca müşterilerin kaba demografik özelliklerine göre değil çerezler ve tıklama verileri kullanılarak daha ayrıntılı ve geniş ölçekli analiz ve modellerin oluşmasına olanak sağlamıştır.

Güncel dijital pazarlama analitiği, gelenekselden farklı olarak kampanya performansının verimini artırmakta ve müşterilerin kişisel kazanımlarına yönelik stratejik öneri sistemleri kurulmasında fayda sağlamaktadır. Kampanyaların anlık performansları sürekli olarak analiz edilerek optimizasyonu sağlanabildiğinden bu stratejik öneri sistemleri kurulabilmektedir. Sonuç olarak şirketler, müşterilerden aldıkları geri dönüş oranlarını artırabilmekte ve aynı

zamanda işletme performansının hangi yollar izlenerek artırılabileceğine dair ipuçları yakalamaktadır. Kitleleriyle daha derin ve güçlü bağlar kurabilen şirketlerin dijital pazarlamanın üzerlerindeki performansı nasıl değiştirip geliştirdiğine yönelik çalışmalar üzerine genel tartışmalar bulunmaktadır (Hong, 2024; Ijomah ve diğerleri, 2024; Nwabekee ve diğerleri, 2024).

2.2. Makine Öğrenmesi

Makine öğrenmesi (ML), kurulan model içerisinde meydana gelen olaylar hakkında bilgi ve tecrübe edinen bilgisayarın gelecekte doğabilecek benzer durumlar hakkında kararlar alabilmesi ve oluşabilecek bazı problemlere çözüm türlerinde uzmanlaşması demektir (Öztemel, 2006). Temelinde bakıldığında ilkel bir yapay zekâ olarak nitelendirilebilir. Yapay zekânın bir alt kümesi olarak sınıflandırılabilir ve modelin tühettiği verilere göre öğrenme gerçekleştiren ya da en iyilemeye odaklı sistemler oluşturan bir yapıdır. Bütün makine öğrenmeleri bir yapay zekâ ürünüDür; ancak her yapay zekâ ürünü makine öğrenmesi değildir.

Makine öğrenmesi tekniği, gerçek hayatla bütünleşmiş biçimde birçok alanda kullanılır: konuşma ve el yazısı tanıma, nesne tanıma, bilgisayar oyunları, doğal dil işleme, robot hareketleri, arama motorları ve tıbbi teşhis gibi alanlarda bunlar gözlemlenebilir (Kutluğun ve ark., 2017).

Makine öğrenmesi tekniğinin öğrenme yöntemlerine bakıldığında iki ana kategori öne çıkmaktadır. Birincisi, etiketsiz veriler üzerinde gerçekleşen gözetimsiz öğrenme; ikincisi ise verilerin etiketli olduğu, türleri hakkında bilgi sahibi olunan ve çıktı olarak girdilere bağımlı sonuçların (evet/hayır, geçer/kalır gibi) elde edildiği gözetimli öğrenmedir.

Bu çalışmada makine öğrenmesinin gözetimli öğrenme kısmındaki algoritmalarla ilgilenilmiş ve etiketli veriler üzerinde çalışılmıştır. Gözetimli öğrenme ve diğer yöntemlerin ayrıntılarına geçmeden önce bu iki ana kategorinin (gözetimli ve gözetimsiz öğrenme) temel prensipleri ve birkaç örnek yöntemi literatüre uygun biçimde incelenmiştir.

2.2.1. Gözetimli Öğrenme

Gözetimli öğrenme, resimli kitaplardan sebzeleri tanımaya ve öğrenmeye çalışan bir çocuğa benzetilebilir. Sebzeler bir etikete sahiptir ve bu yöntem önceden gözlenerek sonuçları bilinen verilerin ve sonuçların bütününü kapsayan bir çıktı elde etmeyi amaçlar (Nizam ve ark., 2014).

Hali hazırda sonuçları bilinen (etiketlenmiş) verilerden faydalanarak çalışan bu makine öğrenmesi yöntemine örnek olabilecek algoritmalar arasında lojistik regresyon, çoklu sınıf sınıflandırması ve destek vektör makineleri bulunur.

Eldeki veriler test-eğitim başlıkları altında oranlanıp bölünür ve makine yaklaşık %80'lik eğitim verisiyle kurulur, doğruluk oranları hesaplanır. Daha sonra eldeki %20'lik test kısmı da modele sokularak makine öğrenmesi modelinin eksik ya da aşırı öğrenmeye (overfitting) maruz kalıp kalmadığı ölçülür ve buna göre gidişat şekillendirilir. Gözetimli öğrenme, giriş verilerinin (X) hedef bir çıktı (y) ile eşleştirilmesi ile bir veri kümesi üzerinde eğitildiği temel alt daldır.

1. K-En Yakın Komşuluk Algoritması (K-Nearest Neighbors – KNN)

K-en yakın komşuluk algoritması, gözetimli makine öğrenmesinde kullanılan modellerden biridir. Basit sınıflandırma modellerinden biri olan KNN algoritması, tek bir veri noktasının sınıflandırılmasında verilerin öznitelik uzayındaki birbirlerine olan yakınlıklarını göz önüne alarak temel sınıflandırma veya regresyon işlemlerini gerçekleştirir. Parametrik olmayan bu gözetimli öğrenme yöntemi, benzer özelliğe sahip veri noktalarının vektör uzayda birbirine yakın konumlanacağı varsayımıyla geliştirilmiştir.

KNN yöntemi hem regresyon (sayısal değer tahmini) hem de kategorik sınıflandırma problemlerinde kullanılmaktadır. Sınıflandırma kısmında veri setine yeni eklenen verinin uzayda hangi kümede yer alacağına karar verirken oylama mekanizması kullanılır. Çoğunluk oylaması (majority voting) ikili sınıflandırma durumlarında bir sınıfın seçilmesi için oyların %50'den fazlasını alması kuralına dayanır. Çokluk oylaması (plurality voting) ise çok sınıflı problemlerde salt çoğunluğa gerek duymaz; veri, rakip sınıflara oranla en yüksek benzerlik oyunu alan sınıfa dâhil edilir.

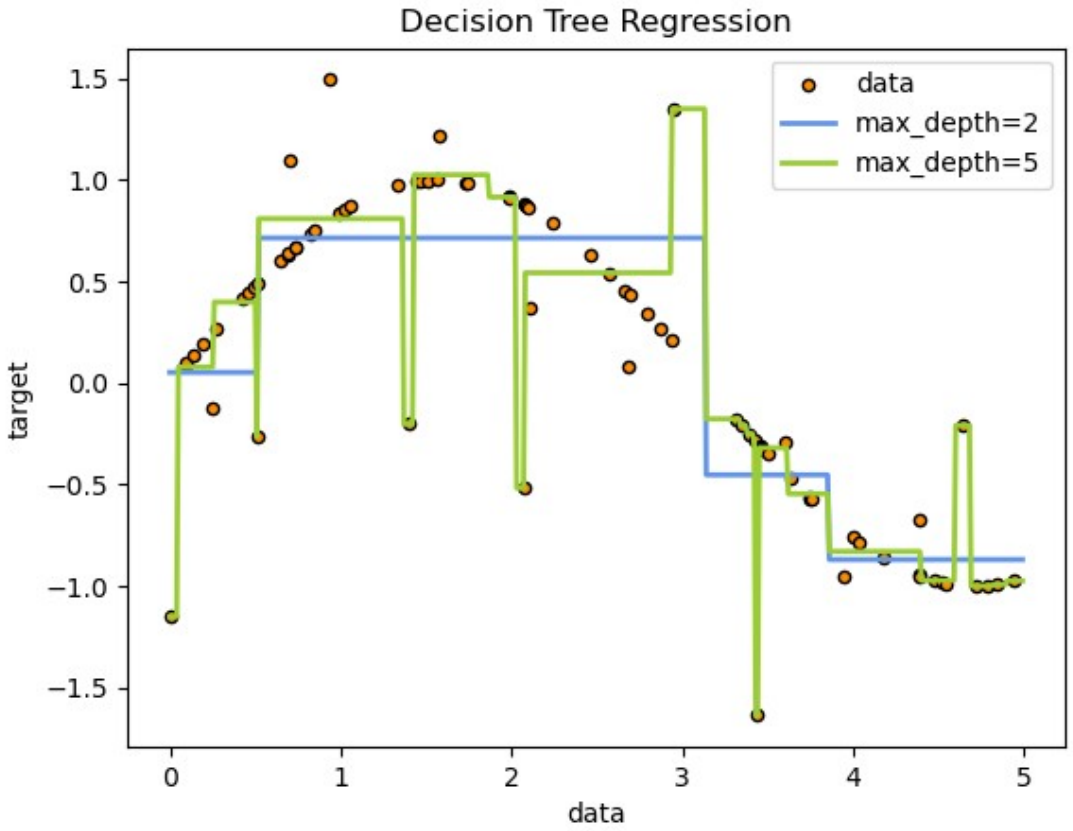
Bu esnek yapısı sayesinde KNN, verinin dağılımı konusunda bir ön varsayımda bulunmadan karmaşık veri setlerinde verimli çalışabilir. Algoritmanın temel çalışma prensibi, bir sorgu noktasının sınıfını ona en yakın komşularının etiketlerine göre tayin etmektir. Bu sürecin en kritik aşaması, veriler arasındaki yakınlığın matematiksel olarak nasıl tanımlanacağıdır. Hesaplanan mesafeler, veri uzayını farklı sınıflara ayıran karar sınırlarını (decision boundaries) belirler. Bu sınırların geometrik temsili literatürde genellikle Voronoi Diyagramları ile görselleştirilmektedir.

Algoritmanın başarısı, veri setinin yapısına uygun mesafe metriğinin seçimine doğrudan bağlıdır. Bu çalışmada ele alınan temel mesafe ölçütleri şunlardır:

1. **Öklid Mesafesi ()**: En yaygın kullanılan metrik olup, iki nokta arasındaki en kısa "kuş uçuşu" mesafeyi temsil eder. Gerçek değerli vektörler üzerinde, Pisagor teoremi tabanlı bir hesaplama yapar.
2. **Manhattan Mesafesi ():** İki nokta arasındaki mutlak koordinat farklarının toplamıdır. Şehir yapısındaki sokakları takip eden bir rotaya benzetildiği için "taksi mesafesi" veya "şehir bloğu mesafesi" olarak da adlandırılır.
3. **Minkowski Mesafesi:** Öklid ve Manhattan metriklerinin genelleştirilmiş formudur. Formüldeki parametresinin aldığı değere göre farklı mesafe ölçütlerine dönüşebilmektedir (için Manhattan, için Öklid).
4. **Hamming Mesafesi:** Genellikle kategorik veya ikili (Boolean) verilerde tercih edilir. İki vektör arasındaki uyumsuzlukları saptayarak, vektörlerin birbirinden farklılaştığı noktaların sayısını esas alan bir "örtüşme metriği" işlevi görür.

2. Karar Ağaçları (Decision Trees)

Karar ağaçları, veriyi belirli niteliklere göre dallara ayırarak sonuç üreten parametrik olmayan (non-parametrik) bir gözetimli öğrenme yöntemidir. Sınıflandırma ve regresyon alanlarında kullanılan bu modelin temel amacı, verinin özelliklerinden hedeflenen değerin tahminini yapabilmektir. Bunu yapmak için veriden yola çıkarak basit karar alma kurallarını kullanır. Anlaşılır ve yorumlanabilir bir yöntemdir. Görsel açıdan desteklenebilir olma özelliği de olan karar ağaçları, diğer model teknikleri kadar veri hazırlığına ihtiyaç duymayabilir ve hem sayısal hem kategorik verileri işleyebilir (Scikit-learn, 2024).



(Scikit-learn, 2024)

Ağaç derinliği konusu kritik bir hususdur. Düşük derinliğe sahip ağaçlar karar verme noktasında sığ kalma tehlikesiyle karşı karşıya kalır; model böyle durumlarda sinüs eğrisi gibi kıvrımlı yapıları tahmin etmekte zorlanır. Sonuçta ortaya kaba ve yetersiz bir yapı çıkar ki buna yetersiz öğrenme (underfitting) denmektedir. Yüksek derinliğe sahip modeller ise detaylı sorular sormaya başlayarak küçük değerler arasındaki ayrıntıları bile yakalamayı hedefler. Derinlik arttıkça model eğrideki her bir küçük hareketi ve hatta gürültüyü (noise/outlier) takip etmeye başlar; bu ise aşırı öğrenme (overfitting) riskini beraberinde getirir.

Pazarlama analitiği özelinde müşteri davranışlarının tahminlemesinde de benzer biçimde çok derin ağaçlar tekil müşteri hareketlerini ezberleme riski taşırken sığ ağaçlar genel trendleri yakalamakta yetersiz kalabilmektedir.

3. Naive Bayes Sınıflandırması

Naive Bayes, gözetimli makine öğrenmesi yöntemlerinden biridir. Veri setindeki sınıf değişkeni verildiğinde bütün özelliklerin birbirinden koşulsuz bağımsız olduğunu naif bir biçimde varsayar. Bu varsayım, modelin her bir parametrenin birbirleriyle çarpımı kombinasyonunu göz ardı ederek matematiksel karmaşıklığın azaltılmasını hedefler. Birden çok parametrenin birbirleriyle olan ilişkisi yerine sınıf veya değişkenlerin kendi içlerindeki dağılımın incelenmesi için kullanılır.

Bayes Teoremi, y sınıf değişkeninin eldeki x değişkenleri kullanılarak tahmin edilmesine dayanır. Örneğin (x_1, x_2, …) özelliklerine sahip müşterilerin satın alma (y) ihtimali nedir sorusuna yanıt arar. Lojistik regresyon gibi ayırt edici (discriminative) modeller iki sınıf arasındaki karar çizgisini çizmeye çabalarken Naive Bayes, satın alım yapan bir müşteri profilinin nasıl göründüğü sorusuna odaklanır.

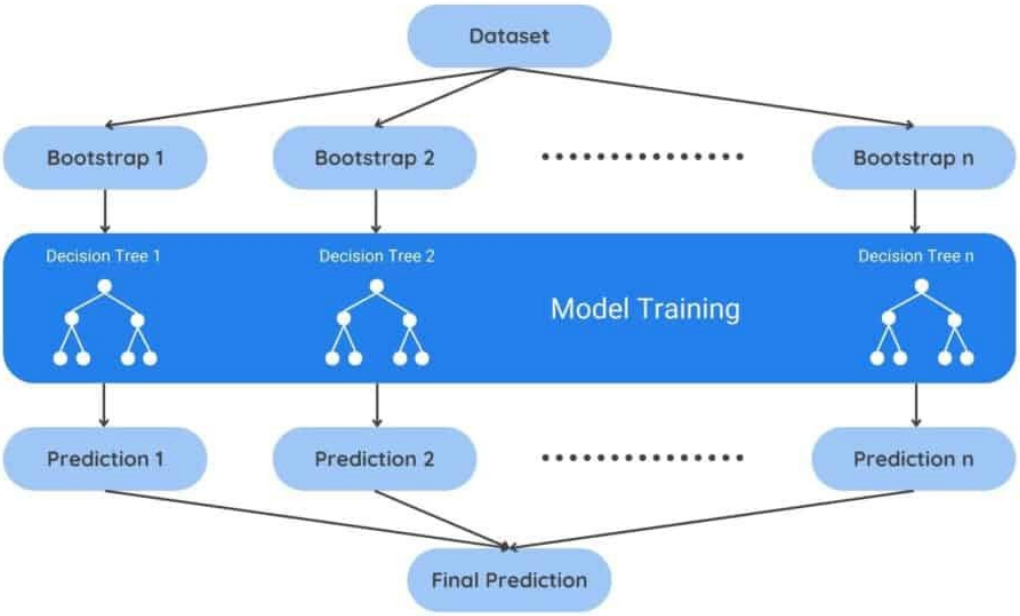
Lojistik regresyon gibi doğrusal modeller ve Naive Bayes gibi temel olasılık yaklaşımları, pazarlama verilerindeki basit örüntüleri yakalamakta başarılı olsa da veri setindeki daha karmaşık ilişkilerde verimi düşmektedir. Bu kısıtlamaları aşmak için sonraki adımlarda sınıflar arasındaki yüksek dereceli etkileşimleri yakalayabilmek adına topluluk öğrenmesi (ensemble learning) algoritmaları kullanılmaktadır.

4. Topluluk Öğrenmesi (Ensemble Learning)

Sınıflandırma algoritmaları ile nesnelerin hangi sınıfa dâhil olacağını çözümlemek isterken karmaşık verilerde yetersiz veya aşırı öğrenme sonucunda yanıltıcı sonuçlarla karşı karşıya kalınan senaryolarda topluluk öğrenmesi yöntemine başvurulabilir. Birçok sınıflandırma yöntemi arasından probleme uygunluğu en yüksek olanı seçmekte kullanılan bu yöntem, yüksek doğruluk oranları yakalayabilmek adına optimizasyonlar gerçekleştirir ve her modelin güçlü taraflarından faydalanır. Aynı veri kümesi üzerinde her bir modelin zayıf ve güçlü yanları göz önünde bulundurulur, ardından bir karar verme mekanizması kurularak en yüksek doğruluk oranına ulaşılmak hedeflenir. Topluluk öğrenme teknikleri temel olarak üçe ayrılır: torbalama (bagging), güçlendirme (boosting) ve istifleme (stacking).

4.1. Torbalama (Bagging)

Torbalama, makine öğrenmesi modellerinde varyansı azaltmak ve aşırı öğrenmeyi engelleyerek tahmin istikrarını artırmaya yardımcı olan bir yöntemdir.

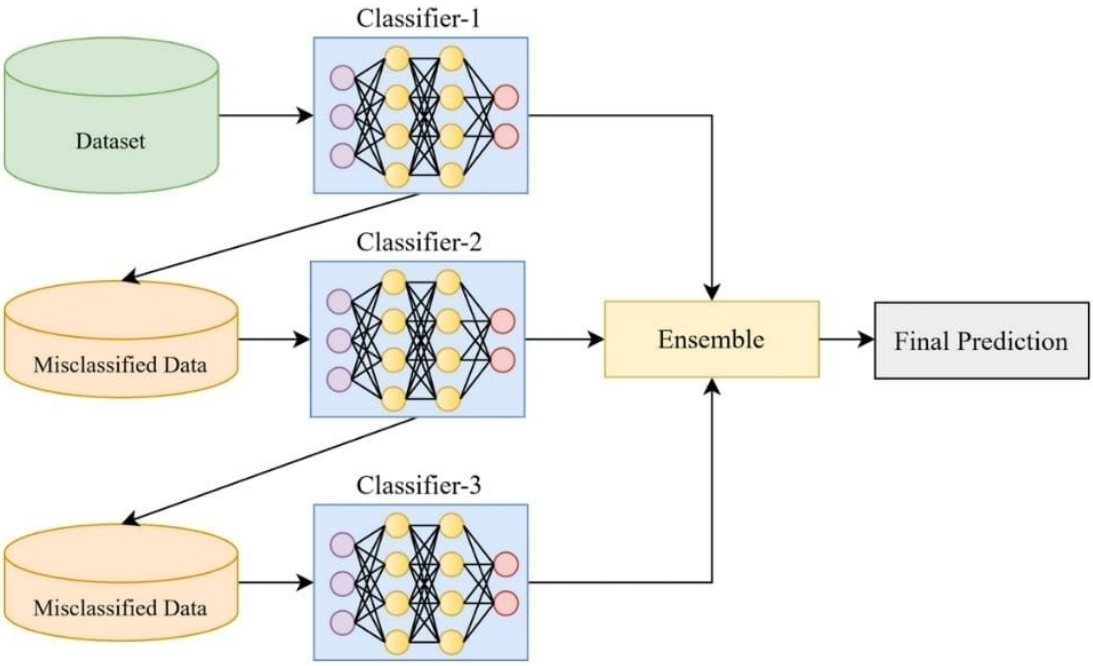


Şekil 7. Torbalama Yöntemi Görseli

Algoritma alt kümeler hâlinde birden fazla veri kümesini rastgele seçimlerle oluşturarak her verinin birden fazla kez seçilebilmesini mümkün kılar. Bu sayede her alt model, diğerinden biraz farklı bir varyasyon görmüş olur. Oluşturulan alt küme setlerinin her birinde ayrı olarak model eğitilir; bu model genellikle karar ağacı yöntemiyle eğitilir. Tahmin aşamasına gelindiğinde bütün modellerin sonuçları birleştirilerek ortalaması alınır.

4.2. Güçlendirme (Boosting)

Güçlendirme, en iyi seviyede çalışmayan basit modellerin birbirini takip ederek oluşturduğu sıralı bir yöntemdir. Her model, kendinden sonra gelen modelin hatalarını düzeltmeye odaklanır ve bu süreç sonunda daha güçlü bir modelin oluşturulması hedeflenir. Her adımda zayıf yönleri giderilmeye çalışılan modeller güçlü yönlerini birleştirmeye çalışır. Bu süreç zincirleme biçimde devam ederek yanlllık (bias) azaltılmaya çalışılır. Torbalama yönteminde alt veri kümeleri eş zamanlı çalıştırılabilirken güçlendirme yönteminde her model ardı sıra çalıştırılarak sırayla yürütülür.



Şekil 7. Güçlendirme Yöntemi Görseli

Her yöntemin kendine göre avantaj ve dezavantajları bulunmaktadır. Örneğin güçlendirme yöntemi küme önyargısını azaltabileceği gibi yinelemeli süreci dolayısıyla bireysel modellere kıyasla daha karmaşık ilişkilerin yaklanmasına da olanak sağlayabilir.

2.2.2. Gözetimsiz Öğrenme

Gözetimsiz öğrenme, herhangi bir etiketlenmiş verisi bulunmayan grupların ham verileri arasındaki benzerlikleri ve özellikleri açığa çıkarmaya yarayan bir sistemdir. Algoritma, verilerin içindeki ve arasındaki kalıpları ve ilişkileri kendi tanır ve kurar. Veri setindeki verilerin çıktıları bilinmediğinden bu yöntemde amaç sınıflandırma değil; olasılık yoğunluk tahmini, öznitelik mühendisliği ve boyut indirgemesi gibi yöntemlerle kümelemedir. Kümeleme, birbirine benzeyen nesnelerin gruplandırılmasıdır ve yalnızca makine öğrenmesinde değil veri madenciliği, örüntü tanıma, görüntü analizi gibi birçok alanda da kullanılır.

1. K-Ortalamalar Yöntemi (K-Means)

K-means algoritması, etiketlenmemiş verilerin örüntülerini tespit edip daha iyi kararlar almak için kullanılır. Algoritma, isminde yer alan k değişkenine bağlı olarak değişiklik göstermektedir. Her modelleme sırasında bu k değerinin belirlenmesi gerekmekte olup elde edilecek sayısal skorlar ve kümelenme biçimleri k'nın aldığı değere göre sürekli değişiklik gösterecektir.

K-means algoritması oldukça sık kullanılan bir kümeleme algoritmasıdır. K değeri, eldeki etiketsiz verinin üzerinde yapılan analizler ve öznitelik mühendisliğinin sonucunda en doğru kümelemeyi hangi değerde verdiğiyle bulunur. Bu algoritma pazarlama ve müşteri segmentasyonu haricinde görüntü işleme, finansal analiz ve coğrafi bilgi sistemi gibi alanlarda da kullanılır.

2. Hiyerarşik Kümeleme

Hiyerarşik kümeleme, veri madenciliği alanında çok sayıda kullanım alanına sahip yöntemlerden biridir. Verilerin birbirine olan benzerliklerine göre bir dendogram (ağaç yapısı) oluşturarak gruplandıran kümeleme yöntemidir. Temel olarak iki birbirine zıt teknikten oluşur: birleştirici (agglomerative) ve bölücü (divisive) yöntem.

Daha yaygın olarak kullanılan birleştirici yöntemde her bir veri ilk başta kendi başına bir küme olarak tutulur, birbirine en yakın olan iki küme birleştirilir, bu işlem tüm verileri tek bir kümede toplayana kadar sürdürülür ve sonuçta bir dendogram elde edilir. İstenen küme sayısına göre bu dendogram belli bir noktadan yatay olarak kesilerek kümeler belirlenir.

Kümeler arasındaki uzaklık çeşitli yöntemlerle tanımlanabilir: single linkage (en yakın iki noktanın mesafesi), complete linkage (en uzak iki noktanın mesafesi), average linkage (tüm nokta çiftlerinin ortalama mesafesi) ve centroid linkage (kümelerin ağırlık merkezleri arasındaki mesafe). Yaygın olarak kullanılan mesafe ölçüsü Öklid mesafesidir (Kaufman ve Rousseeuw, 1991; Altınok, 2019).

Bölücü (divisive) hiyerarşik kümelemede ise işlem üstten alta doğru yürür: tüm veri noktaları başlangıçta tek bir küme olarak ele alınır, bu küme benzerlik veya farklılıklara göre daha küçük alt kümelere bölünür ve bölme işlemi belirli bir durma kriteri sağlanana kadar yinelemeli olarak devam eder.

3. DBSCAN Algoritması

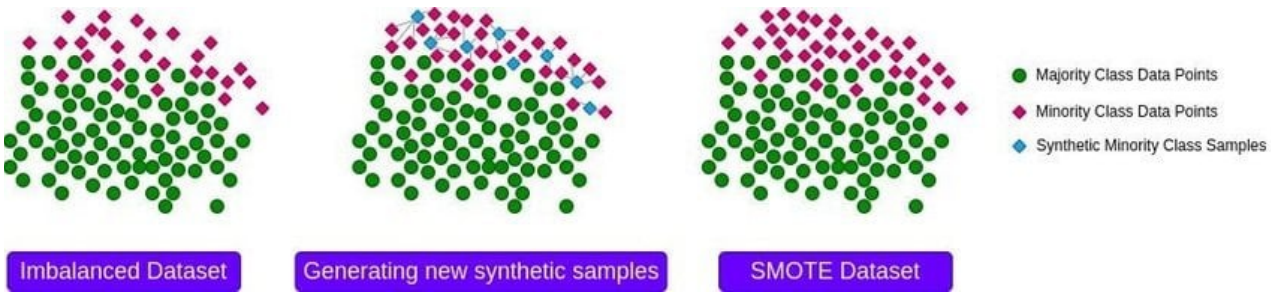
Gürültülü Uygulamaların Yoğunluk Tabanlı Uzamsal Kümelenmesi (DBSCAN), yoğunluğa dayalı sınıflandırmayı temel alan bir algoritmadır. Gürültülü veri ve alakasız değerler içeren büyük boyutlardaki veri setlerinde farklı şekil ve büyüklükteki kümeleri bulabilir.

DBSCAN algoritmasının ana fikri, bir kümenin her noktası için belirli bir yarıçapın (ε) komşusunun en az minimum sayıda nokta (minPoint) içermesi gerektiğidir. Algoritma, veri kümesindeki bir noktayı rastgele alarak ilerler; noktaya ε yarıçapında en az minPoint noktası varsa tüm bu noktalar aynı kümenin parçası olarak kabul edilir. Daha sonra her komşu nokta için komşuluk hesaplaması tekrarlanarak kümeler genişletilir.

2.3. Makine Öğreniminde Sınıf Dengesizliği

Dengesiz gerçek dünya veri kümeleri, kurgulanmak istenen modelin çalışma doğruluğunu saptırabilir. Özellikle ikili sınıflandırma problemlerinde makine öğrenimi esnasında öğrenme konusunda yanılsamaya düşülmesi açısından sorun yaşanmaktadır. Literatürde belirtildiği üzere bu durum modelin doğruluğunu (accuracy) kâğıt üzerinde yüksek gösterse de aslında azınlıkta kalan sınıfın (örneğin satın alım yapan müşteriler) doğru tahmin edilememesine yol açar.

Bu sınıf dengesizliği ile mücadelede en popüler yöntemlerden biri SMOTE'dur (Synthetic Minority Oversampling Technique). İlk olarak dengesizliğin belirlenmesiyle birlikte verideki azınlık sınıf analiz edilerek yeni sentetik benzer veriler üretilir. Bu örnekler eklenerek sınıflar dengelenir ve modelin azınlık sınıfını öğrenme şansı artırılır. Dengesiz veri kümeleri ile başa çıkmak için genel olarak kullanılan iki temel yöntem alt örnekleme (under-sampling) ve üst örnekleme (over-sampling) dir.



Şekil 0. Illustration of SMOTE Oversampling for Imbalanced Classification

Kullanılan veri setindeki %1,3'lük düşük dönüşüm oranı dolayısıyla modelin yanlılık (bias) göstermesi sorunuyla karşı karşıya kalınmıştır. Bu problemin üstesinden gelmek adına literatürdeki iki temel yaklaşım olan üst örnekleme ve alt örnekleme yöntemleri değerlendirilmiştir.

2.3.1. Doğruluk (Accuracy)

Gözetimli makine öğrenmesi yöntemlerini uygularken modeli en iyi eğitecek ve hatayı en aza indirecek yöntemin kullanılması, kurulan model ve çalışma performansı için önemlidir. Bu sebeple veriden yola çıkarak en uygun modelin seçilmesi, uygulanması ve test edilmesi noktasında bazı metrikler kullanılır.

Bu metrikleri anlamak için öncelikle karışıklık matrisi (confusion matrix) kavramını bilmek gerekir. Karışıklık matrisi, bir sınıflandırma probleminde gerçekleşen ve tahmin edilen durumların değerlerini göstermektedir.

- **True Positives (TP)**: Durum gerçekleşecek olarak tahmin ettiniz ve bu tahmin doğru.
- **True Negative (TN)**: Durum gerçekleşmeyecek olarak tahmin ettiniz ve bu tahmin doğru.
- **False Positive (FP)**: Durum gerçekleşecek olarak tahmin ettiniz ve bu tahmin yanlış.
- **False Negative (FN)**: Durum gerçekleşmeyecek olarak tahmin ettiniz ve bu tahmin yanlış.

Accuracy değeri, modelde doğru tahmin edilen kısmın toplam veri kümesine oranlanmasıyla bulunur. Ancak tek başına yeterli olmayabilir; örneğin eşit biçimde yayılmamış veri kümelerinde bir hastalık verisinde hasta olan insanlar verinin %10'luk kısmını kapladığında hastalığı teşhis edilmeyen (False Negative) hastaların bulunması istenmeyen bir durumdur. Bu sebeple diğer metriklerin de dâhil edilmesi gerekir.

2.3.2. Duyarlılık (Recall) ve Kesinlik (Precision)

Kesinlik metriği, pozitif olarak tahmin edilen değerlerin gerçekte ne kadarının pozitif olduğunu gösterir. Özellikle yanlış pozitif (False Positive) durumların maliyetinin yüksek olduğu senaryolarda önemli bir metriktir. E-posta filtreleme gibi sistemlerde spam veya önemli e-postaları ayırmada bu değer önemli derecede kullanılır.

Duyarlılık ise pozitif olarak tahmin edilmesi gereken işlemlerin hangi oranda pozitif olarak tahmin edildiğini gösteren bir metriktir. Yanlış negatif (False Negative) tahminlerin maliyetinin yüksek olduğu durumlarda bu metrik özellikle yardımcı olmaktadır. Her iki metrikte de yüksek değer elde edilmesi, modelin performansı açısından önemli bir kriterdir.

2.3.3. F1 Skoru (F1-Score)

F1 skoru, kesinlik ve duyarlılık değerlerinin harmonik ortalamasını vermektedir. Standart ortalama yerine harmonik ortalama kullanılmasının sebebi uç durumların göz ardı edilmemesi gerekliliğidir. Örneğin kesinlik değeri 1 ve duyarlılık değeri 0 olan bir modelin standart ortalaması 0,5 olarak gelecektir ve bu yanıltıcı olur.

Accuracy yerine F1 skorunun tercih edilme sebebi, genel olarak eşit dağılmayan veri kümelerinde hatalı bir model seçimi yapmamaktır. Ayrıca kesinlik ve duyarlılık özelinde yalnızca tekli yanlış tahminlerin değil bütün hata maliyetlerini içeren bir ölçme metriğine ihtiyaç duyulduğu için F1 skoru önemlidir.

2.4. Benzerlik Ölçütleri ve Tavsiye Sistemleri

2.4.1. Benzerlik Ölçütleri

Veri setlerinde çalışırken, özellikle müşteri segmentasyonu ve öneri sistemlerinde ürünler arasındaki benzerliği hesaplamak kritik öneme sahiptir. Bu benzerlik ölçütleri, kümeleme algoritmalarında ve öneri sistemlerinde sıklıkla kullanılır.

En sık kullanılan benzerlik ölçütleri şunlardır:

1. **Öklidyen Mesafe** : Yukarıdaki bölümlerde bahsettiğimiz bu yöntem sayısal özelliklere sahip verilerde, iki nokta arasındaki düz çizgi uzaklığını ölçer. K-Means algoritmasında yaygın olarak kullanılır.

2. **Kosinüs Benzerliği**: Bu yöntem bağlamında özellikle metin madenciliği veya kullanıcı-ürün tercih matrisleri gibi vektörel temsillerde kullanılır. Vektörler arasındaki açıyı temel alır.
3. **Manhattan Mesafesi**: Manhattan mesafesi (L1 normu) ise iki nokta arasındaki mesafeyi koordinat düzleminde yalnızca yatay ve dikey adımlarla ölçen bir uzaklık türüdür.

2.4.2. Tavsiye Sistemleri

Bir önceki bölümde detaylandırılan benzerlik ölçütlerinin veri biliminde en yoğun ve etkili biçimde kullanıldığı alanların başında tavsiye sistemleri gelmektedir. Temel olarak bu sistemleri, kullanıcıların geçmiş etkileşimleri veya benzer profildeki kullanıcıların davranışlarına dayanarak kişiselleştirilmiş öneriler sunan analitik motorlar olarak düşünebiliriz. Literatürde ve endüstriyel uygulamalarda bu sistemler üç ana yaklaşımda incelenmektedir.

2.4.2.1. İçerik Tabanlı Öneri (Content-Based Filtering)

İçerik tabanlı filtreleme, kullanıcıya geçmişte etkileşime girdiği öğelerle özellik düzeyinde benzerlik gösteren yeni öğeleri öneren bir yaklaşımdır (Lops ve diğerleri, 2011). Yöntem, her öğeyi metin, kategori veya demografik etiket gibi özellik vektörleriyle temsil eder; kullanıcı profilini bu özelliklerin ağırlıklı bileşimi olarak modeller. 2.4.1 bölümünde ele alınan kosinüs benzerliği ve öklidyen mesafe ölçütleri, bu sistemlerin temel hesaplama çekirdeğini oluşturmaktadır.

Yöntemin en belirgin avantajı soğuk başlangıç problemini kısmen aşabilmesidir. Sistem başka kullanıcıların verilerine ihtiyaç duymadan yalnızca hedef kullanıcının kendi geçmişinden öneri üretebilir (Pazzani ve Billsus, 2007). Bu özellik yeni kullanıcıların sisteme dâhil olduğu senaryolarda işbirlikçi filtrelemeye kıyasla belirgin bir üstünlük sağlar. Ancak tam da bu noktada yapısal bir kırılganlık devreye girer; model yalnızca kullanıcının daha önce etkileşimde bulunduğu türden içerikler önerdiğinden kullanıcının henüz keşfetmediği kategorilere yönlendirilmesi mümkün değildir. Literatürde bu durum aşırı uzmanlaşma olarak tanımlanmaktadır.

Pazarlama kampanyası optimizasyonu özelinde bu kısıt kritik bir boyut kazanmaktadır. Müşteriye yalnızca daha önce açtığı türde e-postalar göndermek teknik olarak mümkündür;

ancak hangi kanal ve platform kombinasyonunun dönüşüme en güçlü katkıyı sağlayacağını tahmin etmek, mesaj içeriğinin çok ötesinde müşteri profil özelliklerinin modele dâhil edilmesini zorunlu kılar. Bu nedenle içerik tabanlı filtreleme, aksiyon optimizasyonu hedefi için tek başına yeterli bir çerçeve sunamamaktadır ve dolayısıyla çalışmamızda kullanmayı tercih etmediğimiz bir öneri yöntemidir.

2.4.2.2. İşbirlikçi Filtreleme (Collaborative Filtering)

İşbirlikçi filtreleme, "seninle benzer davranışlar sergileyen kullanıcılar bunları da tercih etti" mantığıyla çalışan ve tavsiye sistemi literatüründe en geniş uygulama alanı bulan yaklaşımdır (Koren ve diğerleri, 2009). Bir önceki yöntemin aksine öğelerin kendi içerik özelliklerine değil, tamamen kullanıcı-öğe etkileşim matrisine odaklanır.

Yöntem iki ana kol üzerinde gelişmiştir. Kullanıcı tabanlı yaklaşımda hedef kullanıcıya en benzer diğer kullanıcılar, Pearson korelasyonu ya da kosinüs uzaklığıyla belirlenen bir komşu kümesi aracılığıyla tespit edilir ve bu komşuların geçmişte yüksek değerlendirdiği öğeler önerilir. Öğe tabanlı yaklaşım ise yönü tersine çevirir; hedef kullanıcının yüksek puan verdiği öğelere benzer diğer öğeleri ön plana çıkarır (Sarwar ve diğerleri, 2001). Amazon'un ticari ölçekte benimsediği öğe tabanlı yöntem, büyük veri kümelerinde daha kararlı benzerlik yapıları üretmesi nedeniyle geniş kabul görmüştür (Linden ve diğerleri, 2003).

İşbirlikçi filtrelemenin yapısal avantajı içerik özellikleri hakkında ön varsayım gerektirmemesidir. Bununla birlikte yöntemin iki temel kırılganlığı mevcuttur. Seyreklik problemi temel sorundur; gerçek dünya sistemlerinde kullanıcı-öğe matrisinin boş kalma oranı yüzde doksanlı seviyeleri aşabilmekte ve bu durum benzerlik hesaplamalarının güvenilirliğini ciddi biçimde zedelemektedir (Aggarwal, 2016). İkincisi soğuk başlangıç problemidir; işbirlikçi filtrelemede yeni kullanıcılar için pratikte çözümsüz kalmaya devam etmektedir.

Bu çalışmada kullanılan 48.000 kayıtlık sentetik veri setinde dönüşüm gerçekleştiren müşterilerin oranı yalnızca yüzde iki civarındadır. Kullanıcı-kampanya etkileşim matrisi bu kadar gürültülü ve asimetrik bir zemine oturduğunda işbirlikçi filtreleme benzerlik matrisini güvenilir biçimde inşa edemez. Dolayısıyla bu yöntem birincil modelleme çerçevesi olarak değil, bir sonraki bölümde geliştirilen prescriptive yaklaşımla kıyaslanabilir bir zemin oluşturmak amacıyla ele alınmıştır.

2.4.2.3. Prescriptive Analitik Tabanlı Öneri Sistemi

Karar desteği literatüründe analitik olgunluk dört katmana ayrılmaktadır; tanımlayıcı analitik ne olduğunu, teşhis edici analitik neden olduğunu, tahmine dayalı analitik ne olacağını sorarken prescriptive yani reçeteleyici analitik ne yapılması gerektiğini sorar ve doğrudan bir aksiyon önerisiyle yanıtlar (Lepenioti ve diğerleri, 2020). Bu dördüncü katman önceki üçünün çıktılarını bir optimizasyon motoru olarak işlemiştir.

Klasik tavsiye sistemleri pasif bir tahmin üretir: "kullanıcı bu ürünü beğenebilir." Prescriptive sistemler ise bu soruyu aktif biçimde yeniden çerçeveler: "kullanıcının dönüşüm gerçekleştirmesi için hangi aksiyonu almalıyız?" Bu ayrım yüksek maliyetli reklam bütçelerinin hangi kanallara yönlendirilmesi gerektiğini belirlemek açısından doğrudan bir anlam taşır (Bertsimas ve Kallus, 2020). Bu soruları cevaplandırırken çalışmamız tahmin modelinin ötesine geçerek doğrudan eyleme yönelik karar mekanizması olarak konumlandırılmıştır.

Çalışmamızda geliştirilen prescriptive öneri sistemi strateji simülasyonu mantığına dayandırılmıştır. Eğitilmiş dönüşüm tahmin modeli, her müşteri için mevcut tüm CampaignChannel ve AdvertisingPlatform kombinasyonlarını değerlendirerek en yüksek dönüşüm olasılığını veren kombinasyonu optimal aksiyon olarak belirler. Yedi kanal ile yedi platformun eşleştirilmesiyle oluşan kırk dokuz senaryonun her birinin simüle edilmesi sisteme gerçek zamanlı kampanya kararı üretme kapasitesi kazandırmaktadır. Prescriptive Analitik Tabanlı Öneri Sistemi tercihi ile her müşteri için uygun kanal-platform ikilisi dinamik olarak önerilen yenilikçi bir araç ortaya çıkarılmıştır.

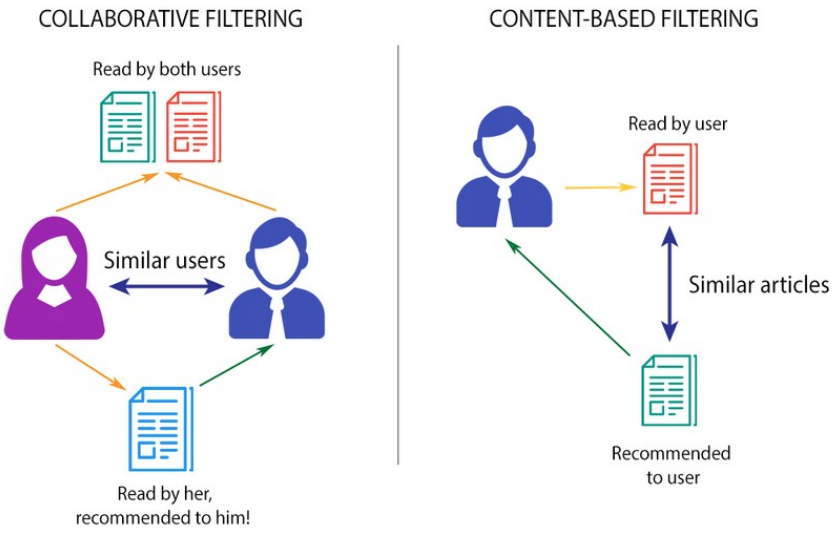
2.4.2.4. Collaborative Filtering ve Prescriptive Sistemlerin Karşılaştırması

Çalışmamızda geliştirilen prescriptive öneri sistemi geleneksel işbirlikçi filtrelemeden üç temel boyutta ayrıştırılmıştır.

Girdi türü açısından işbirlikçi filtreleme geçmiş kullanıcı-öğe etkileşim puanlarına dayanırken prescriptive sistem müşteri profil özelliklerini ve alınabilecek aksiyonları girdi olarak kullanmıştır.

Çıktı türü açısından işbirlikçi filtreleme hangi öğenin beğenileceğini tahmin ederken prescriptive sistem hangi aksiyonun dönüşüm olasılığını maksimize edeceğini hesaplamıştır (Tran ve diğerleri, 2021).

Ölçeklenebilirlik boyutunda iki yaklaşım arasındaki fark en belirgin biçimde ortaya çıkar. İşbirlikçi filtreleme yeterli etkileşim geçmişi olmayan yeni müşteriler için öneri üretemez. Prescriptive sistem ise yalnızca profil özellikleriyle herhangi bir müşteriye anında öneri üretebilmetedir.



Şekil 1. Collaborative Filtering vs. Content-Based Filtering.

İki paradigmanın rakip değil tamamlayıcı olduğunu vurgulamak gerekir. İşbirlikçi filtreleme zengin etkileşim geçmişine sahip büyük kullanıcı tabanlarında güçlü bir kişiselleştirme aracıdır. Prescriptive sistemler ise bireysel etkileşim sinyalinin sınırlı ya da güvenilmez olduğu ortamlarda belirgin biçimde öne çıkar. Yüzde ikinin altındaki dönüşüm oranı ve buna eşlik eden ağır sınıf dengesizliği, bu çalışmanın prescriptive yaklaşımı seçmesinin veri kaynaklı gerekçesini oluşturur.

Tablo 1. İşbirlikçi Filtreleme ve Prescriptive Sistemlerin Karşılaştırması

Boyut	**İşbirlikçi Filtreleme**	**Prescriptive Sistem**
Girdi Türü	Geçmiş kullanıcı-öğe etkileşim puanları	Müşteri profil özellikleri + aksiyonlar
Çıktı Türü	Hangi öğenin beğenileceği tahmini	Dönüşümü maksimize eden aksiyon önerisi
Ölçeklenebilirlik	Yeni kullanıcılar için öneri üretemez (soğuk başlangıç)	Profil özellikleriyle anında öneri üretebilir
Veri Gereksinimi	Zengin etkileşim geçmişi zorunlu	Minimal etkileşim verisiyle çalışabilir

Kullanım Alanı	Büyük kullanıcı tabanları, zengin veri ortamları	Sınırlı sinyal, dengesiz sınıf ortamları

3. MATERYAL VE YÖNTEM

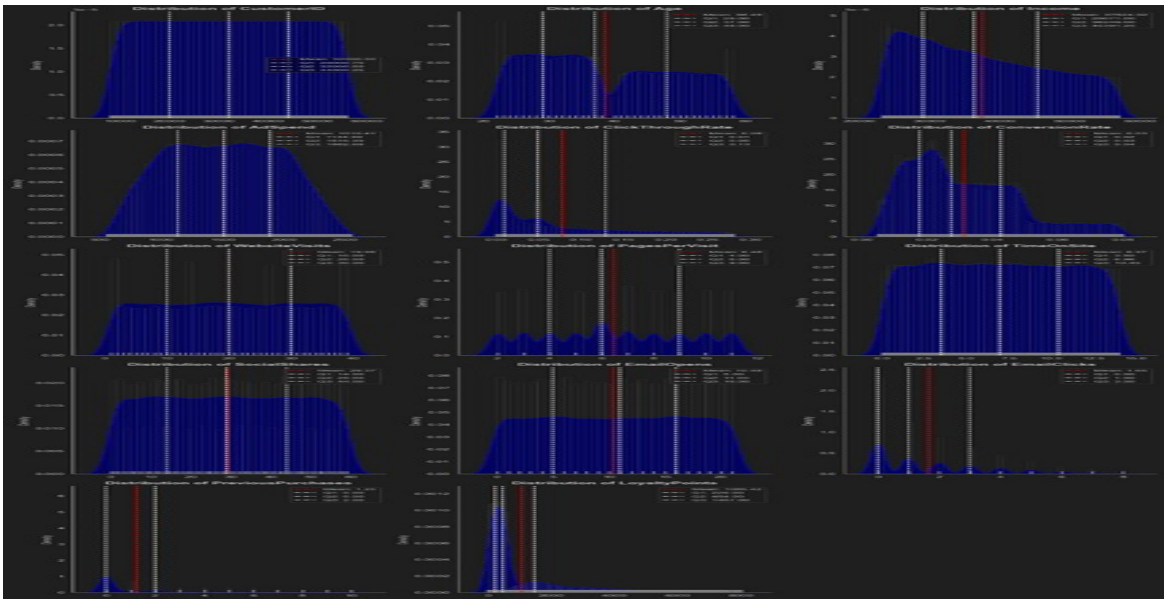
Bu bölümde çalışmada kullanılan veri setinin yapısı, keşifsel analiz bulguları, özellik mühendisliği adımları ve veri sızıntısının tespit ile giderilme süreci ele alınmaktadır. Ardından model geliştirme metodolojisi ve prescriptive öneri sisteminin tasarım mantığı açıklanmaktadır.

3.1. Veri Seti

Çalışmada 48.000 sentetik müşteri kaydından oluşan bir pazarlama analitik veri seti kullanılmıştır. Veri seti, gerçek dünya pazarlama sistemlerinin davranışsal özelliklerini yansıtacak biçimde tasarlanmış olup 20 ham değişken içermektedir. Bu değişkenler demografik özellikler, kampanya parametreleri, dijital etkileşim metrikleri ve hedef değişken olan dönüşüm durumundan oluşmaktadır.

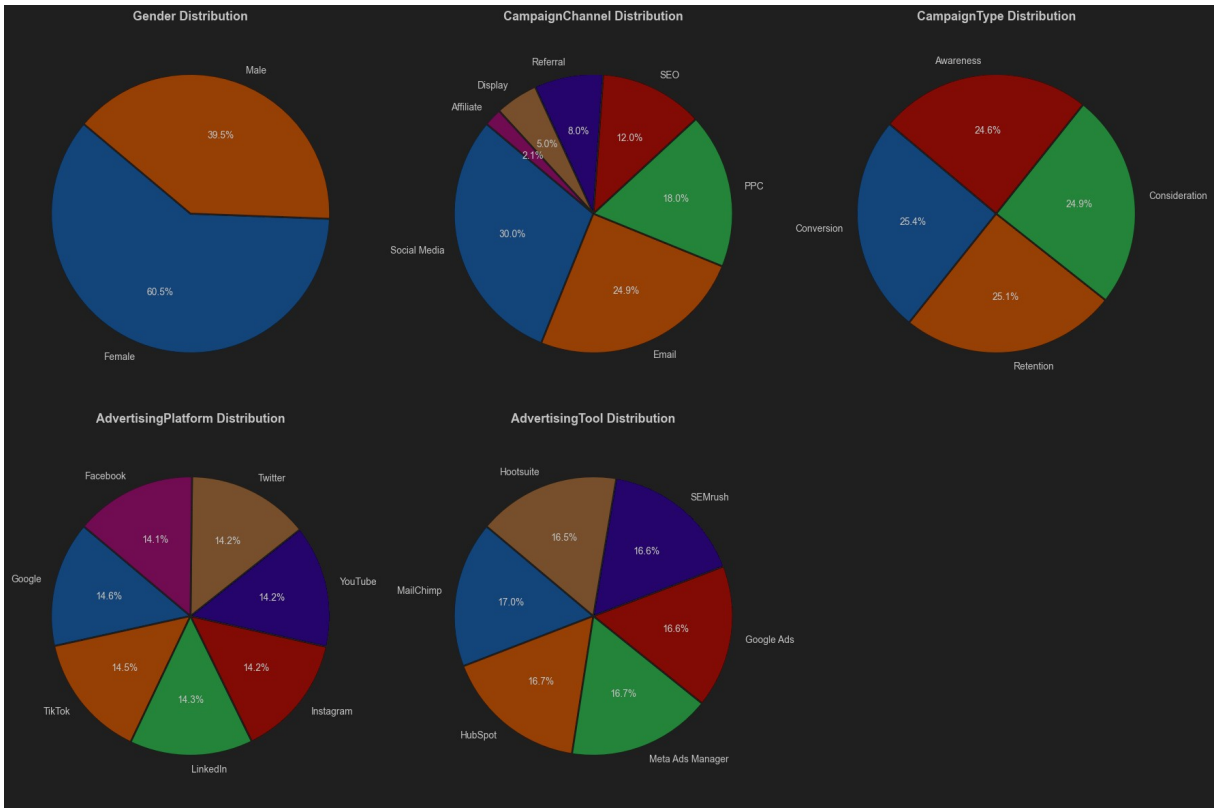
Veri setinin temel özelliği hedef değişkenin ikili ve yoğun dengesiz yapısıdır. Toplam 48.000 kayıt içinde yalnızca 980 müşteri dönüşüm gerçekleştirmiş olup bu oran yüzde 2,04'e karşılık gelmektedir. Sınıf dengesizlik oranı 47,98 kat olarak hesaplanmıştır. Bu yapı, 2.3 bölümünde ele alınan sınıf dengesizliği probleminin bu çalışmadaki doğrudan karşılığıdır ve hem model seçimini hem değerlendirme metriklerini şekillendirmiştir.

Aşağıdaki Şekil 2'de sayısal değişkenlerin dağılım yapısı ortaya konmaktadır. Age değişkeni belirgin bir çift tepeli dağılım sergilemekte; Income ve AdSpend geniş bir çeyreklikler arası aralık göstermektedir. PreviousPurchases ile LoyaltyPoints arasındaki yüksek korelasyon ise bu aşamada dikkat çekici bir örüntü olarak öne çıkmaktadır.



Şekil 2. Sayısal değişkenlerin dağılım grafikleri ve merkezi eğilim ölçütleri.

Kategorik değişkenler incelendiğinde CampaignChannel dağılımının Social Media ve Email kanalları ağırlıklı olduğu görülmektedir. AdvertisingPlatform ve AdvertisingTool değişkenleri neredeyse eşit dağılım sergileyerek model için anlamlı bir ayrıştırıcılık potansiyeli taşımamaktadır. Şekil 3 bu dağılım yapısını görselleştirmektedir.



Şekil 3. Kategorik değişkenlerin dağılımı: cinsiyet, kanal, kampanya türü, platform ve araç.

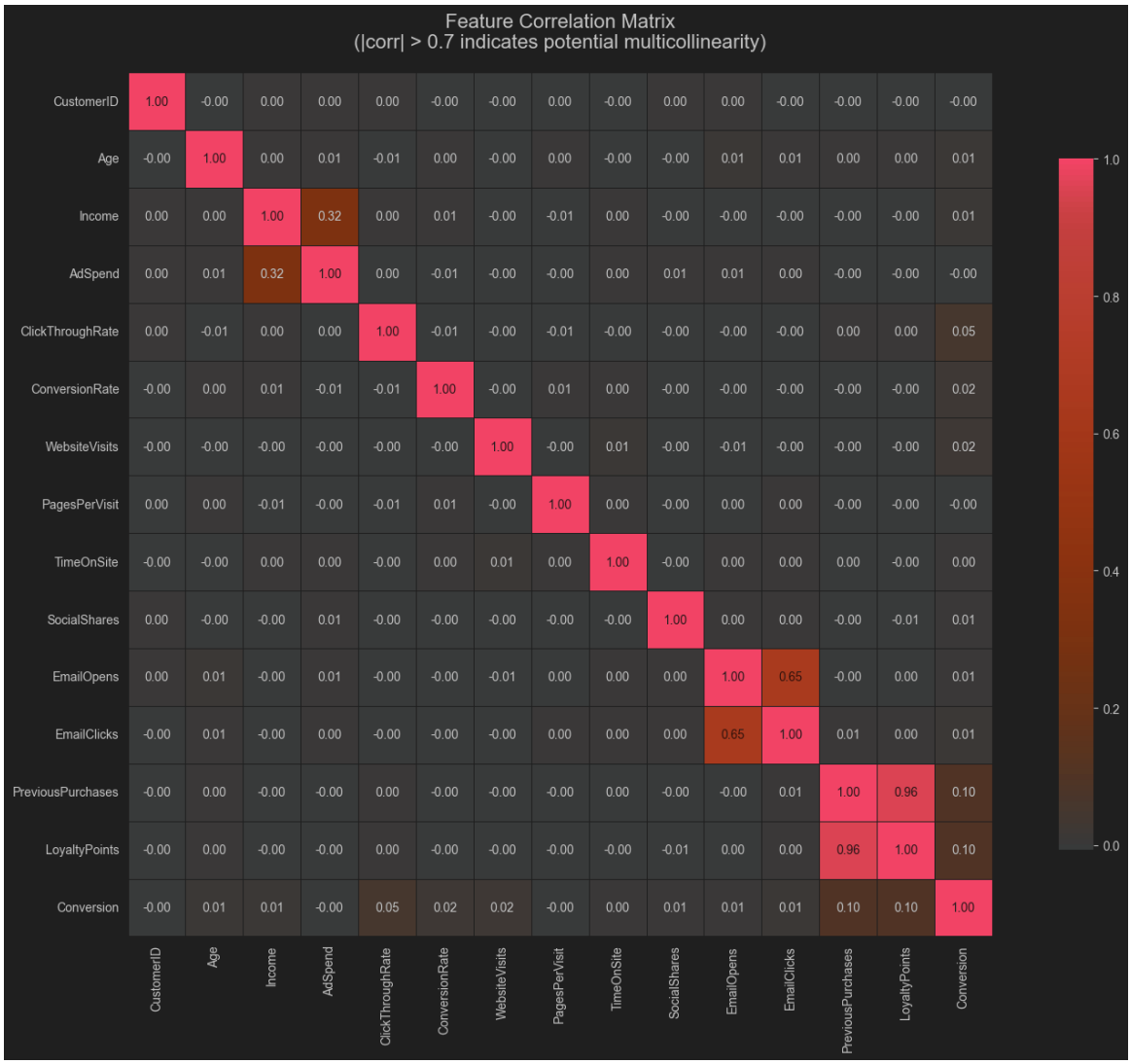
3.2. Keşifsel Veri Analizi (EDA)

Modelleme öncesinde gerçekleştirilen keşifsel veri analizi hem veri kalitesini değerlendirmek hem de özellik mühendisliği kararlarına zemin hazırlamak amacıyla yürütülmüştür. Analiz dört temel eksen üzerine kurulmuştur: eksik değer tespiti, aykırı değer analizi, çoklu doğrusallık kontrolü ve istatistiksel hipotez testleri.

Eksik değer analizi, ClickThroughRate değişkeninde 2.464 ve PagesPerVisit değişkeninde 2.435 eksik kaydın varlığını ortaya koymuştur. Bu eksiklikler rastgele kayıp örüntüsüyle uyumlu bulunmuş; ilgili değişkenler için medyan doldurma stratejisi

benimsenmiştir. Medyan doldurmanın tercih edilme sebebi, ortalamaya kıyasla aykırı değerlere karşı daha dayanıklı olmasıdır.

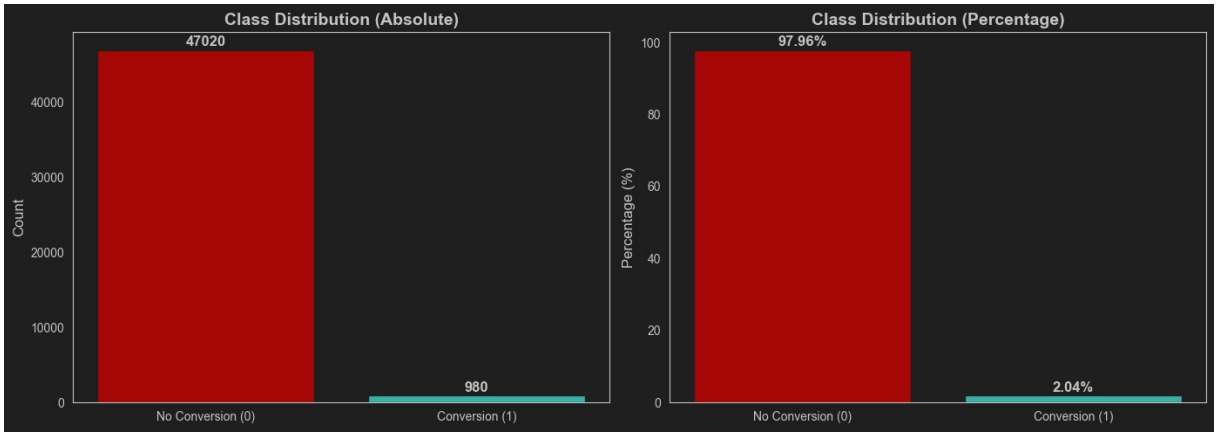
Çoklu doğrusallık kontrolünde PreviousPurchases ile LoyaltyPoints arasında 0,96 düzeyinde bir Pearson korelasyonu saptanmıştır. Şekil 4 bu ilişkiyi ve tüm değişken çiftleri arasındaki korelasyon yapısını görselleştirmektedir. İki değişkenin birlikte modele alınması katsayı tahminlerini dengesizleştireceğinden özellik mühendisliği aşamasında bu yüksek korelasyon dikkate alınmıştır.



Şekil 4. Değişkenler arası Pearson korelasyon matrisi.

Hedef değişkenin dağılımı, bu çalışmanın metodolojik kararlarının tamamını belirleyen en kritik bulgudur. Şekil 5'te görüldüğü üzere 47.020 negatif örneğe karşılık yalnızca 980 pozitif örnek bulunmaktadır. Doğruluk metriği bu yapıda anlamlı bir ölçüt olmaktan çıkar;

modelin tüm kayıtları negatif tahmin etmesi bile yüzde 97,96 doğruluk üretir. Bu nedenle model değerlendirmesinde F1 skoru ve ROC-AUC önceliklendirilmiştir.



Şekil 5. Hedef değişken sınıf dağılımı: 47.020 negatif örneğe karşı 980 pozitif örnek.

Ki-kare testleri CampaignChannel ve CampaignType değişkenlerinin dönüşümle istatistiksel olarak anlamlı ilişki taşıdığını ortaya koymuştur. Bağımsız örneklemler t-testi uygulandığında ise Age ve Income değişkenleri dönüşüm gerçekleştiren ve gerçekleştirmeyen gruplar arasında anlamlı farklılık sergilemiştir. Bu bulgular, özellik mühendisliği ve model seçimi aşamalarındaki kararlarımızı doğrudan yönlendirmiştir.

3.3. Özellik Mühendisliği

Ham veri setindeki 20 değişkene ek olarak alan bilgisine ve EDA bulgularına dayalı 17 türetilmiş özellik oluşturulmuş; böylece model girdisi 37 değişkene genişletilmiştir. Özellik türetme süreci beş işlevsel grupta organize edilmiş olup her bir grubun tasarlanma mantığı aşağıda açıklanmaktadır.

3.3.1. ROI ve Maliyet Metrikleri

Pazarlama verimliliğini ölçmek amacıyla üç özellik türetilmiştir. CPA_Proxy, reklam harcamasının tahmini dönüşüm sayısına oranını temsil etmekte; ROI_Proxy ise dönüşüm oranı ve gelirin reklam harcamasına göre normallenmiş karşılığını vermektedir. Spend_Efficiency, tıklama oranının reklam harcamasına oranlanmasıyla elde edilmektedir. Ancak bu özelliklerin modele alınması, 3.4 bölümünde ayrıntılı biçimde ele alınan veri sızıntısı riskini beraberinde getirmiştir.

3.3.2. Etkileşim Metrikleri

Kullanıcı-platform etkileşimini yansıtmak üzere beş özellik üretilmiştir. Site_Engagement, sitede geçirilen süre ile sayfa başına ziyaret sayısının çarpımıyla hesaplanarak yüzeysel tıklamayı gerçek etkileşimden ayırt etmeye çalışır. Email_Click_Rate, e-posta açılma sayısına oranlanmış tıklama sayısını; Social_Virality ise web sitesi ziyaretine göre normallenmiş sosyal paylaşım yoğunluğunu ifade etmektedir. Bu değişkenler tasarlanırken amaç, tek başına anlamlı olmayan ham metriklerin birbirleriyle ilişkilendirilerek daha bilgi zengin sinyaller üretmesini sağlamaktı.

3.3.3. Müşteri Segmentasyon Özellikleri

Müşteri profilini daha ayrıntılı tanımlamak için dört kategorik özellik oluşturulmuştur. Age_Group, müşterileri YoungAdult (18–25), Adult (26–35), MiddleAge (36–50) ve Senior (51+) kategorilerine ayırmaktadır. Income_Tier ise gelir dağılımının çeyrekliklerine göre Low, Medium, High ve VeryHigh kategorilerine bölmektedir. Customer_Value_Score, satın alma geçmişi, sadakat puanı ve normallenmiş gelirin çarpımından üretilen bütünleşik bir değer indeksidir. Bu segmentasyon değişkenleri, 2.1 bölümünde ele alınan müşteri davranışı ve segmentasyon kavramının pratikte nasıl uygulandığını göstermektedir.

3.3.4. Etkileşim Özellikleri (Interaction Features)

Doğrusal olmayan ilişkileri yakalamak amacıyla üç çarpım özelliği oluşturulmuştur. AdSpend_x_CTR, reklam harcaması ile tıklama oranının etkileşimini kodlamakta; böylece yüksek harcama yapan ancak düşük tıklama alan kampanyaların farkı ortaya çıkmaktadır. Income_x_Loyalty, gelir ve sadakat puanının birlikte oluşturduğu potansiyeli yansıtmakta; yüksek gelirli ve aynı zamanda sadık müşteri profilini tanımlamaktadır. Age_x_Purchases ise yaş ve satın alma geçmişi arasındaki doğrusal olmayan ilişkiyi yakalamaya yöneliktir. Bu tür etkileşim özellikleri, 2.2.1 bölümünde açıklanan karar ağaçlarının doğal olarak yakalayamadığı çarprazlamış sinyalleri doğrusal modellere de sunabilmek için tasarlanmıştır.

3.3.5. Kanal Performans Özellikleri

EDA sonuçlarına dayanarak her kampanya kanalının geçmiş dönüşüm performansı sıralanmış ve iki özellik türetilmiştir. Channel_Performance değişkeni Referral ve Email kanallarını Yüksek; Display, PPC ve Social Media kanallarını Orta; Affiliate ve SEO kanallarını ise Düşük olarak etiketlemektedir. Bu sıralama, veri setindeki kanalların gerçek

dönüşüm oranlarına göre belirlenmiş olup 4.1.1 bölümünde sunulan kanal bazlı dönüşüm analiziyle doğrudan ilişkilidir.

Tablo 2. Türetilmiş özelliklerin kategorileri ve sayıları.

Özellik Grubu	Örnek Değişkenler	Sayı
ROI ve Maliyet Metrikleri	CPA_Proxy,ROI_Proxy,Spend_Efficiency	3
Etkileşim Metrikleri	Site_Engagement,Email_Click_Rate, Social_Virality	5
Müşteri Segmentasyonu	Age_Group,Income_Tier, Customer_Value_Score	4
Etkileşim Özellikleri	AdSpend_x_CTR,Income_x_Loyalty, Age_x_Purchases	3
Kanal Performansı	Channel_Performance, Channel_Rank	2
Toplam Türetilmiş Özellik		**17**

3.4. Veri Sızıntısı (Data Leakage) Tespiti ve Giderilmesi

Bu çalışmanın en kritik metodolojik katkısı, veri sızıntısının tespiti ve sistematik biçimde giderilmesi sürecinde şekillenmiştir. Kaufman ve diğerlerinin (2012) tanımladığı biçimiyle veri sızıntısı, hedef değişken hakkında eğitim aşamasında gerçek hayatta mevcut olmayan bilgilerin modele dâhil edilmesi durumunu ifade eder. Bu durum eğitim setinde yapay biçimde yüksek performans üretir; ancak model gerçek verilerle karşılaştığında bu başarı tamamen çöker.

İlk modelleme denemesinde ConversionRate ve CTR_to_Conversion değişkenleri özellik setine dâhil edilmiş, ROC-AUC 1,0000 ve F1 skoru 1,0000 olarak elde edilmiştir. Bu değerler gerçek bir pazarlama veri seti için fiziksel olarak imkânsızdır. Sorunun kaynağı ConversionRate değişkeninin hedef değişkenden türetilmiş olması ve modelin tahmin etmesi gereken bilgiyi dolaylı yoldan girdi olarak almasıdır. Bu bulgu bizi “mükemmel sonuç” ile “dürüst sonuç” arasındaki kritik farkı anlamaya yöneltmiştir.

Sızıntı kaynakları kademeli biçimde ayıklanmıştır. Aşağıdaki Tablo 3’te altı senaryo karşılaştırmalı olarak sunulmaktadır.

Tablo 3. Veri sızıntısı arınma süreci: senaryo karşılaştırması.

Senaryo	Çıkarılan Değişkenler	ROC-AUC	F1	Recall	Durum
A	ConversionRate, CTR_to_Conv.	1,0000	1,0000	1,0000	Sızdırıyor
B	A + ROI_Proxy	1,0000	1,0000	1,0000	Sızdırıyor
C	A + CPA_Proxy	0,7071	0,0755	0,6306	Dürüst
D *	**Hepsi birden**	**0,7072**	**0,0760**	**0,6347**	**Dürüst**
E	D + ROI_v2, CPA_v2	0,7063	0,0756	0,6316	Dürüst
F	E + SMOTE	0,6963	0,0740	0,6235	Dürüst

** Seçilen senaryo*

Sızıntısız senaryo setinden en yüksek F1 değerini veren Senaryo D nihai model eğitimi için temel alınmıştır. Bu senaryoda ConversionRate, CTR_to_Conversion, ROI_Proxy ve CPA_Proxy değişkenlerinin tamamı özellik setinden çıkarılmış; bunların yerine model bilgisinden bağımsız iki yeni değişken türetilmiştir. ROI_v2, gelir ve tıklama oranının reklam harcamasına oranı olarak; CPA_v2 ise reklam harcamasının web sitesi ziyaretine oranı olarak tanımlanmıştır. Bu ikame değişkenler, hedef değişkenden bağımsız verimlilik sinyalleri üretmekte ve sızıntı riski taşımamaktadır.

3.5. Model Geliştirme Süreci

Sızıntısız özellik seti üzerine iki aşamalı bir model geliştirme süreci yürütülmüştür. Birinci aşamada üç farklı sınıflandırma algoritması temel hiperparametre ayarlarıyla karşılaştırılmış, ikinci aşamada ise seçilen algoritmanın üzerine ek etkileşim özellikleri eklenerek üretim modeli oluşturulmuştur. Veri seti stratified train-test ayrımıyla yüzde seksen eğitim ve yüzde yirmi test olarak bölünmüştür. Stratified bölünme tercih edilmesinin sebebi, yüzde ikinin altındaki dönüşüm oranının rastgele bölünmede test setinde hiç pozitif örnek kalmaması riskini doğurmasıdır.

3.5.1. Lojistik Regresyon (LR)

Lojistik regresyon, ikili sınıflandırma problemlerinde doğrusal karar sınırı üreten ve yorumlanabilirliği yüksek bir yöntemdir. 2.2.1 bölümünde ele alınan teorik çerçeveyle uyumlu biçimde bu çalışmada L2 düzenlemesi ve class_weight='balanced' parametresiyle uygulanmıştır. class_weight='balanced' parametresi, azınlık sınıfın ağırlığını otomatik olarak

artırarak modelin dengesiz yapıyı dikkate almasını sağlamaktadır. Elde edilen katsayılar değişkenlerin yönünü ve göreli etkisini doğrudan yorumlamaya imkân tanımaktadır.

3.5.2. Random Forest (RF)

Random Forest, 2.2.1 bölümünde açıklanan torbalama yönteminin karar ağaçlarına uygulanmasından elde edilen bir topluluk öğrenme algoritmasıdır (Breiman, 2001). Bu çalışmada 200 ağaç, maksimum derinlik 12 ve class_weight='balanced' parametreleriyle kullanılmıştır. Ağaçların rastgele alt özellik kümeleri üzerinde eğitilmesi bireysel ağaçların aşırı öğrenme eğilimini bastırmakta ve topluluk tahmininin varyansını düşürmektedir.

3.5.3. XGBoost

XGBoost, artırma yönteminin gradyan optimizasyonuyla birleştirildiği ve özellikle dengesiz veri setlerinde güçlü performans gösteren bir sınıflandırıcıdır (Chen ve Guestrin, 2016). Bu çalışmada maksimum derinlik 6, öğrenme hızı 0,01, 1000 ağaç sayısı ve sınıf oranıyla orantılı scale_pos_weight parametresiyle konfigüre edilmiştir. Düşük öğrenme hızı tercih edilmesinin sebebi, her adımda küçük düzeltmeler yaparak aşırı öğrenme riskini azaltmaktır.

3.6. Model Değerlendirme Metodolojisi

Yüzde 2'nin altındaki dönüşüm oranı standart doğruluk metriğini anlamsız kıldığından değerlendirme çerçevesi ROC-AUC, F1 skoru, Precision ve Recall üzerine kurulmuştur. Bu metriklerin teorik temeli 2.3 bölümünde ele alınmıştır. Her model için olasılık eşik değeri sıfır ile bir arasında taranmış ve F1 skorunu maksimize eden nokta optimal eşik olarak belirlenmiştir. Varsayılan 0,5 eşiğinin kullanılmamasının sebebi, bu denli dengesiz bir veri setinde sabit eşiğin neredeyse hiç pozitif tahmin üretememesidir.

3.7. Prescriptive Öneri Sistemi Tasarımı

Üretim modeli eğitimi tamamlandıktan sonra disk üzerine kaydedilmiş; bu modelle birlikte StandardScaler ve SimpleImputer nesneleri de serileştirilerek hazır hale getirilmiştir. Bu serileştirme adımı kritiktir çünkü öneri sistemi çalıştırıldığında modelin eğitim aşamasında öğrendiği ölçekleme parametrelerinin birebir aynı şekilde uygulanması gerekmektedir.

Öneri fonksiyonu, 2.4.2.3 bölümünde kavramsal olarak açıklanan prescriptive yaklaşımın pratikteki uygulanmasıdır. Fonksiyon her müşteri için yedi kampanya kanalı ile yedi

reklam platformunun oluşturduğu kırk dokuz kombinasyonu simüle etmekte ve her kombinasyonda dönüşüm olasılığını tahmin etmektedir. En yüksek olasılığı veren kombinasyon o müşteri için optimal aksiyon olarak raporlanmaktadır.

Sistemin bütüncül performansını ölçmek amacıyla test setindeki 9.600 müşteri üzerinde beklenen lift analizi gerçekleştirilmiştir. Her müşterinin mevcut kampanya kombinasyonundaki tahmin olasılığı ile önerilen kombinasyondaki tahmin olasılığı karşılaştırılarak iki değer arasındaki fark hesaplanmıştır. Bu fark, prescriptive sistemin müşteri bazında ne kadar ek dönüşüm potansiyeli yarattığını göstermektedir.

4. BULGULAR

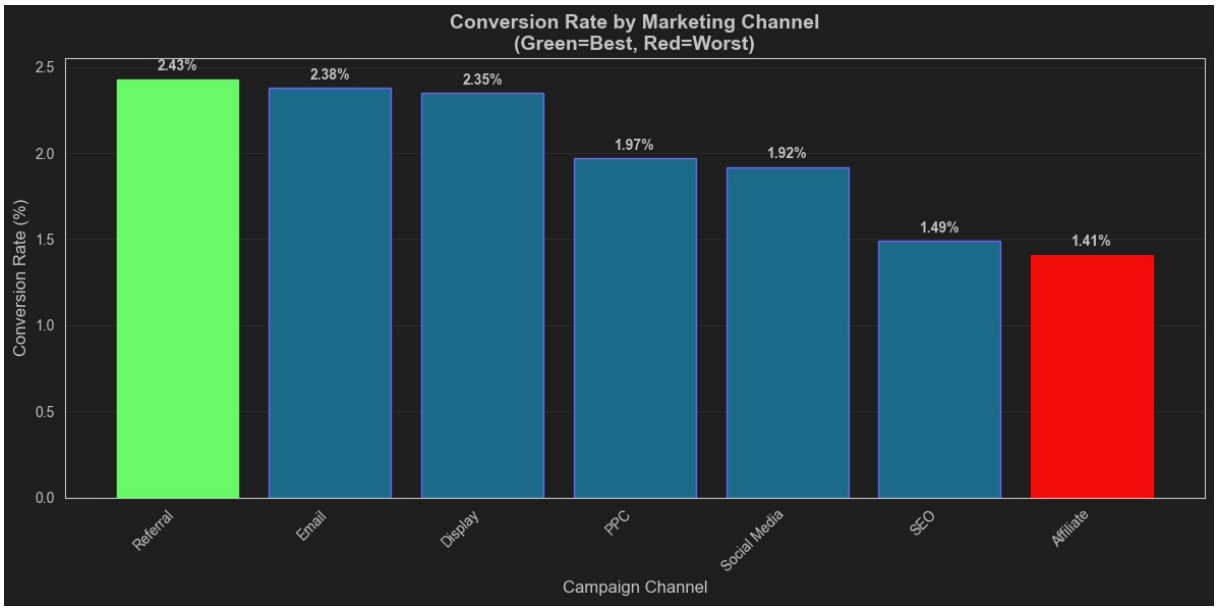
Bu bölümde keşifsel analiz bulguları, veri sızıntısı tespitinin kanıtlanma süreci, model performans sonuçları ve prescriptive öneri sisteminin lift analizi bulgularına yer verilmektedir. Elde edilen sonuçlar, 3. bölümde açıklanan metodolojinin doğrudan çıktılarıdır. Her alt bölüm, önceki bölümlerde oluşturulan kavramsal ve yöntemsel çerçeveye referans vererek sonuçları bağlamlandırmaktadır.

4.1. Keşifsel Analiz Bulguları

Modelleme aşamasına geçmeden önce verinin iç yapısını derinlemesine kavramak, bilinçli özellik mühendisliği ve doğru model seçimi için zorunludur. Bu alt bölümde veri setinin kanal bazındaki dönüşüm örüntüleri ve müşteri segmentasyonu bulguları sunulmaktadır.

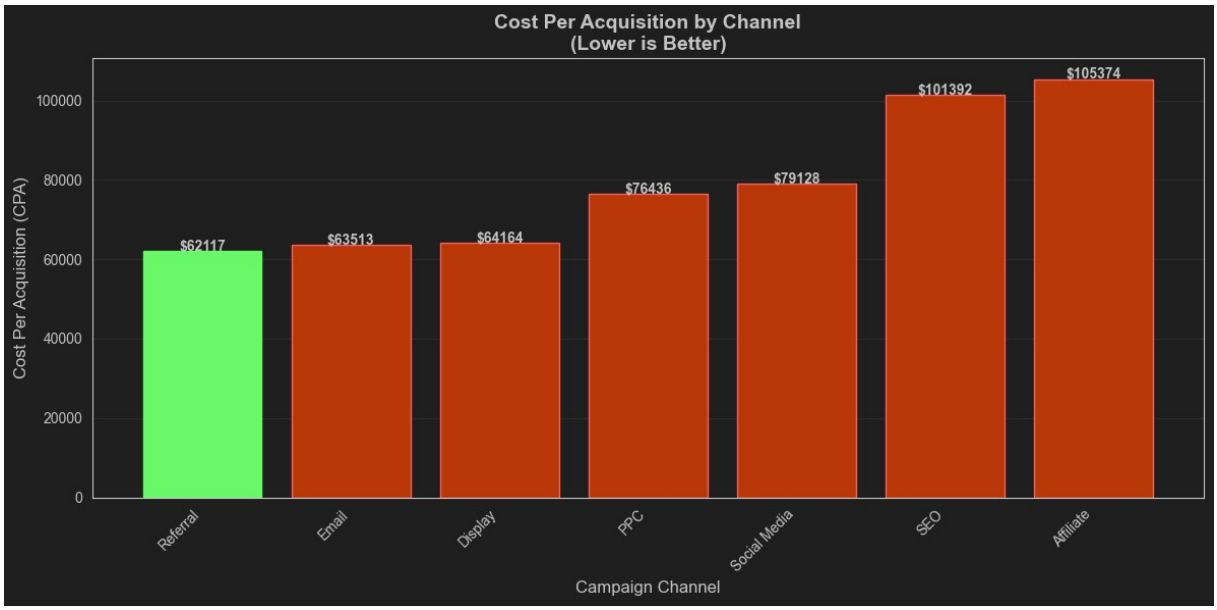
4.1.1. Kanal Bazlı Dönüşüm Analizi

Kampanya kanallarının dönüşüm oranları karşılaştırıldığında en yüksek performansın Referral kanalında yüzde 2,43 ile gerçekleştiği görülmektedir. Email kanalı yüzde 2,38 ve Display kanalı yüzde 2,35 ile yakın performans sergilemektedir. Buna karşın SEO yüzde 1,49 ve Affiliate yüzde 1,41 ile tablonun alt sıralarında kalmaktadır. Şekil 6 bu sıralamayı görselleştirmektedir.



Şekil 6. Kampanya kanalına göre dönüşüm oranları.

Ancak kanal seçimini yalnızca dönüşüm oranına göre değerlendirmek yanıltıcı olabilir. Dönüşüm oranı yüksek olan bir kanalın müşteri başına maliyeti de yüksekse reklam bütçesi açısından verimli olmayabilir. Şekil 7'deki müşteri edinim maliyeti (CPA) analizi bu tabloya önemli bir boyut katmaktadır.



Şekil 7. Kampanya kanalına göre müşteri edinim maliyeti (CPA).

Şekil 7 incelendiğinde dikkat çekici bir sonuç ortaya çıkmaktadır: Referral kanalı en düşük edinim maliyetiyle en yüksek dönüşüm oranını aynı anda sunmaktadır. Bu durum, tavsiye mekanizmasıyla gelen müşterilerin zaten satın alma niyetine yakın olmasıyla

açıklanabilir. Öte yandan SEO ve Affiliate kanallarında CPA'nın 100.000 doları aşması, bu kanalların geniş bir kitleye ulaştığını ancak dönüşüm sağlamakta son derece verimsiz kaldığını göstermektedir. Bu bulgu, 3.3.5 bölümünde Channel_Performance değişkenini tasarlarken neden Referral ve Email'i "Yüksek", SEO ve Affiliate'i ise "Düşük" olarak etiketlediğimizin ampirik gerekçesini oluşturmaktadır.

Pratik bir bakış açısıyla değerlendirildiğinde, sınırlı bir pazarlama bütçesine sahip bir şirketin öncelikli olarak Referral ve Email kanallarına yatırım yapması, birim dönüşüm başına düşen maliyeti ciddi ölçüde düşürecektir. Bu noktada prescriptive öneri sistemimizin her müşteriye özel kanal önerisi yapabilme kapasitesi, bütçe optimizasyonu açısından anlamlı bir değer katmaktadır.

4.1.2. Müşteri Segmentasyon Bulguları

Yaş grupları incelendiğinde MiddleAge segmentinin yani 36 ile 50 yaş arası müşterilerin dönüşüm olasılığı en yüksek grubu oluşturduğu görülmektedir. Bu bulgu, orta yaş grubunun hem harcama kapasitesi hem de marka sadakati açısından pazarlama kampanyalarına daha duyarlı olduğu şeklinde yorumlanabilir.

Sadakat katmanları açısından bakıldığında Loyalty_Tier_Gold kategorisindeki müşteriler, dönüşüm oranı bakımından Bronze ve Silver kategorilerinin belirgin biçimde üzerinde seyretmektedir. Bu durum beklentilerle uyumludur: şirkete daha bağlı olan müşteriler kampanya mesajlarına daha yüksek oranda olumlu tepki vermektedir. Bu bulgular, 3.3.3 bölümünde tasarlanan Age_Group ve Loyalty_Tier segmentasyon değişkenlerinin model için gerçekten ayrıştırıcı bilgi taşıdığını doğrulamakta olup ilerleyen bölümlerde sunulacak değişken önem sıralamasıyla da tutarlıdır.

4.2. Veri Sızıntısının Kanıtlanması

Çalışmamızın belki de en öğretici aşaması, veri sızıntısının tespit edilmesi ve sistematik biçimde giderilmesi süreci olmuştur. 3.4 bölümündeki Tablo 3'te sunulan altı senaryonun sonuçları, bu süreci rakamsal olarak belgelemektedir.

Senaryo A ve B'de ROC-AUC ve F1 skorunun 1,00 olarak elde edilmesi, ilk bakışta mükemmel bir model kurulduğu izlenimini yaratmıştır. Ancak gerçek dünya pazarlama verilerinde böylesine kusursuz bir tahmin performansı fiziksel olarak mümkün değildir. Bu

değerler, Kaufman ve diğerlerinin (2012) tanımladığı hedef sızdırma (target leakage) biçimiyle bire bir örtüşmektedir: ConversionRate değişkeni doğrudan hedef değişkenden türetilmiş olduğundan model tahmin etmesi gereken bilgiyi zaten girdi olarak almıştır. Bir bakıma model "öğrenme" yapmamış, yalnızca cevap anahtarını okumuştur.

Sızıntı kaynaklarının ayıklanma süreci aşamalı biçimde gerçekleştirilmiştir. Önce ConversionRate ve CTR_to_Conversion çıkarılmış ancak performans değişmemiştir; bu durum başka bir değişkenin de sızıntı taşıdığına işaret etmiştir. ROI_Proxy eklendiğinde hâlâ 1,0000 görülmüş; ancak CPA_Proxy çıkarıldığında ROC-AUC aniden 0,7071'e düşmüştür. Bu dramatik düşüş, CPA_Proxy'nin içerdiği sızıntı bilgisinin modelin yapay başarısını tek başına sürdürmeye yettiğini kanıtlamıştır.

Senaryo D'ye gelindiğinde tüm sızıntı kaynakları ayıklanmış ve model gerçek anlamda müşteri profilinden dönüşüm tahmin eder hale gelmiştir. Performanstaki bu gerileme bir başarısızlık değil, modelin artık dürüst bir tahmin görevi yaptığının göstergesidir. Bu deneyim, veri bilimi projelerinde "mükemmel sonuç" elde edildiğinde ilk yapılması gerekenin kutlama değil şüphe olması gerektiğini açıkça öğretmiştir.

SMOTE uygulamasının denenDiği Senaryo F'de ise ilginç bir bulguyla karşılaşılmıştır: sentetik örnekleme tekniği modeli iyileştirmek bir yana, tüm metriklerde hafif bir düşüşe neden olmuştur (ROC-AUC: 0,7072 → 0,6963; F1: 0,0760 → 0,0740). Bu sonuç, 2.3 bölümünde ele alınan SMOTE tekniğinin (Chawla ve diğerleri, 2002) sınıf dengesizliğinin bu denli ağır olduğu yapılarda (∼%2 pozitif oran) etkinliğini yitirebildiğini göstermektedir. Sentetik örnekler, azınlık sınıfının gerçek örüntülerini yakalayamamış ve modele gürültü eklemiştir.

4.3. Model Performans Sonuçları (Phase 2)

Sızıntısız özellik setine 3.3.4 bölümünde açıklanan etkileşim özellikleri eklendikten sonra üç aday algoritma karşılaştırılmış ve üretim modeli olarak Lojistik Regresyon seçilmiştir. Seçimin birincil gerekçesi, lojistik regresyonun katsayı yorumlanabilirliğinin prescriptive öneri sistemi için kritik olmasıdır. Hangi değişkenin dönüşümü hangi yönde ve ne kadar etkilediğini doğrudan görebilmek, 49 kanal-platform kombinasyonunun değerlendirilmesinde vazgeçilmez bir avantajdır. Tablo 4'te üç modelin karşılaştırmalı sonuçları sunulmaktadır.

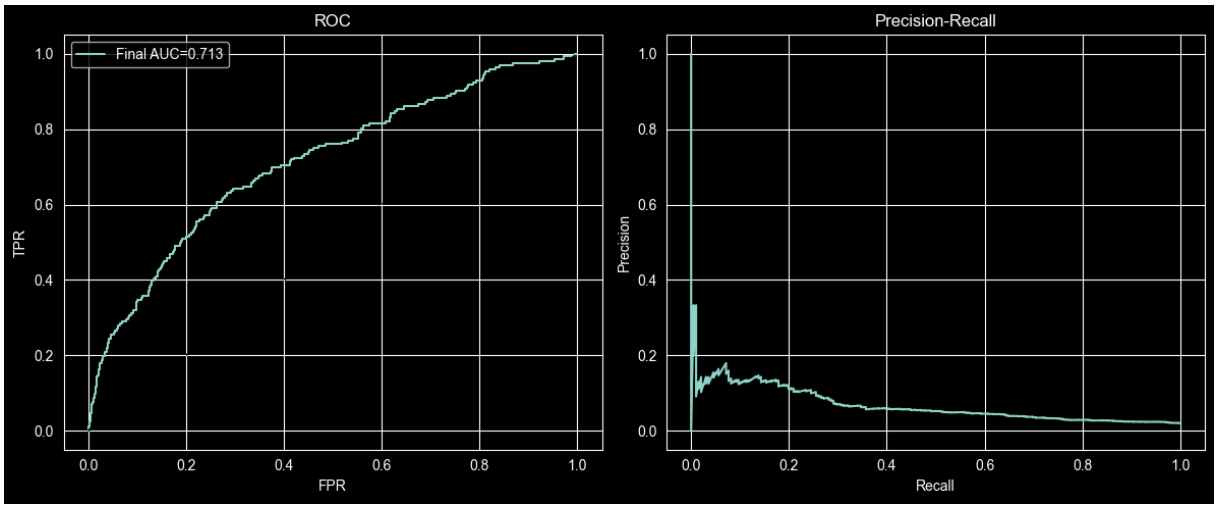
Tablo 4. Üç aday modelin performans karşılaştırması ve üretim modeli seçimi.

Model	ROC-AUC	F1	Precision	Recall	Optimal Eşik
Logistic Regression *	**0,713**	**0,15**	0,134	0,179	0,797
Random Forest	0,698	0,13	0,121	0,142	0,812
XGBoost	0,705	0,14	0,128	0,155	0,803

**** Seçilen üretim modeli***

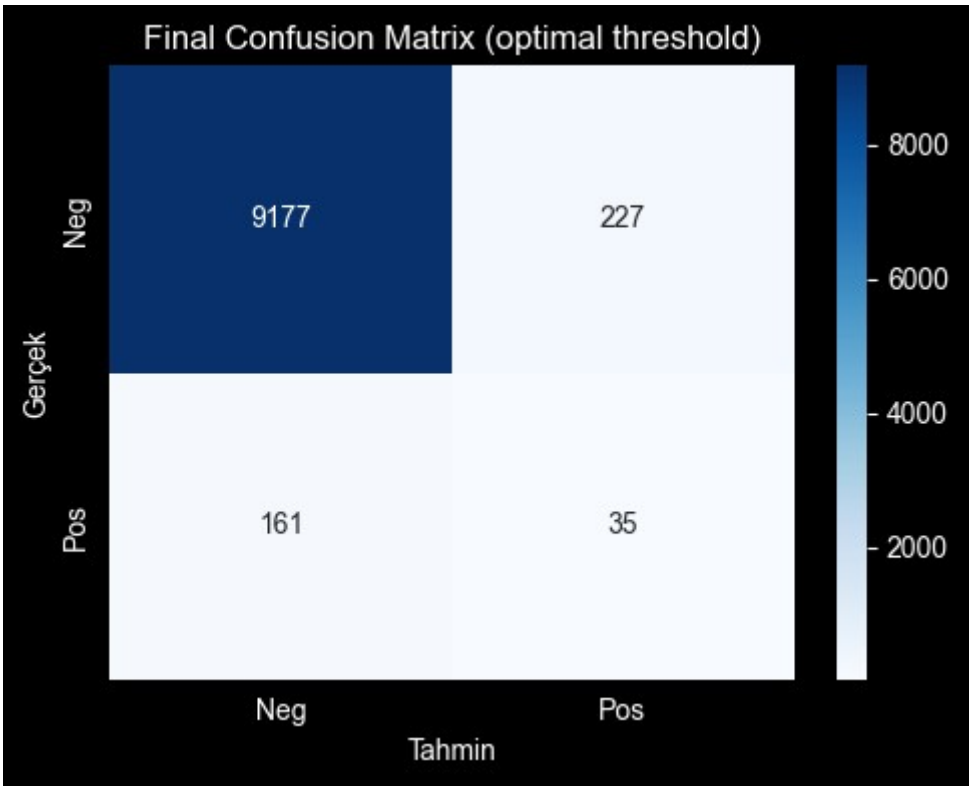
Tablo 4'ten görüldüğü üzere Lojistik Regresyon, ROC-AUC açısından 0,713 ile en yüksek değeri üretmiştir. XGBoost 0,705 ile yakın performans sergilese de lojistik regresyonun sunduğu katsayı yorumlanabilirliği prescriptive sistem için belirleyici olmuştur. Random Forest ise 0,698 ile diğer iki modelin gerisinde kalmıştır. Üç modelin de optimal eşik değerlerinin 0,797–0,812 aralığında yoğunlaşması, varsayılan 0,5 eşiğinin bu denli dengesiz bir veri setinde işe yaramadığını ve 3.6 bölümünde açıklanan F1 maksimizasyonu stratejisinin ne denli önemli olduğunu doğrulamaktadır.

Şekil 8'deki ROC eğrisi bu performansı görsel olarak ortaya koymaktadır. Eğri, rastgele tahmin çizgisinin (AUC=0,5) belirgin biçimde üzerinde seyretmekte olup modelin yüzde 2'lik dönüşüm oranına rağmen pozitif ve negatif sınıfları anlamlı biçimde ayırt edebildiğini göstermektedir. Precision-Recall eğrisi ise dengesiz veri setlerinde ROC eğrisinin gizleyebileceği performans sorunlarını görselleştirmek amacıyla birlikte sunulmuştur.



Şekil 8. Üretim modelinin ROC eğrisi (AUC=0,713) ve Precision-Recall eğrisi.

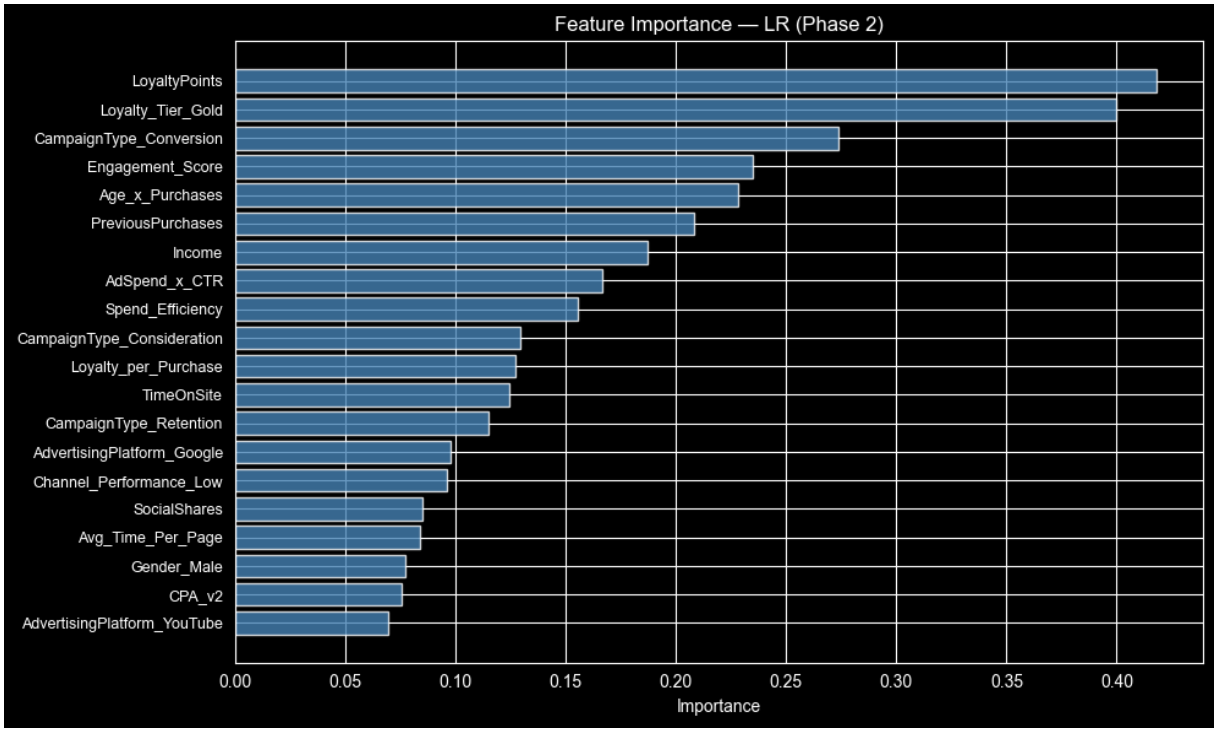
Şekil 9'daki karışıklık matrisi optimal eşik değeri olan 0,797 kullanıldığında elde edilen sınıflandırma dağılımını göstermektedir. 9.177 gerçek negatif doğru sınıflandırılırken 227 gerçek negatif yanlışlıkla pozitif olarak işaretlenmiştir (False Positive). Pozitif sınıfta ise 196 gerçek pozitif müşteriden 35'i doğru yakalanmış, 161'i ise kaçırılmıştır (False Negative). Yüksek False Negative sayısı, modelin ihtiyatlı davrandığını ve pozitif tahmin üretirken temkinli olduğunu göstermektedir. Pazarlama uygulamalarında bu durum, "yanlış alarm" maliyetinin düşük tutulması anlamına gelir; ancak bazı potansiyel müşterilerin de kaçırılması riskini beraberinde getirir.



Şekil 9. Optimal eşik (0,797) kullanıldığında elde edilen karışıklık matrisi.

Değişken önem sıralaması 3.3 bölümündeki özellik mühendisliği kararlarını doğrulamaktadır. Şekil 10'da görüldüğü üzere LoyaltyPoints ve Loyalty_Tier_Gold değişkenleri modeli en güçlü biçimde yönlendiren iki tahmin edici olarak öne çıkmaktadır. Bu sonuç, 4.1.2 bölümündeki müşteri segmentasyon bulgularıyla tam bir tutarlılık içindedir. Ardlarından Engagement_Score ve Age_x_Purchases gibi 3.3.4 bölümünde türetilen etkileşim

özellikleri gelmektedir; bu durum türetilmiş değişkenlerin modele anlamlı katkı sağladığını doğrulamaktadır.



Şekil 10. Logistic Regression üretim modeli değişken önem sıralaması.

4.3.1. F1 = 0,15 Neden Bir Başarı Sayılır?

F1 skoru 0,15 düzeyi sezgisel olarak düşük görünebilir. Ancak bu metriğin anlamlılığı bağlam olmadan değerlendirilemez. Müşterilerin yalnızca yüzde ikisinin dönüşüm gerçekleştirdiği bir ortamda rastgele bir sınıflandırıcı, pozitif sınıfı neredeyse hiç yakalayamaz ve son derece düşük bir F1 değeri üretir.

2.3.4 bölümünde açıkladığımız üzere F1 skoru, kesinlik ve duyarlılığın harmonik ortalamasıdır. Harmonik ortalamanın doğası gereği her iki metriğin de aynı anda makul düzeyde olması gerekir; bir tanesi sıfıra yaklaştığında F1 skoru da çöker. Yüzde 98 negatif örneğin bulunduğu bir veri setinde bu dengeyi sağlamak, eşit dağılmış bir veri setinden çok daha güçtür.

Dolayısıyla 0,15 değeri, sızıntısız ve dürüst bir modelin bu aşırı dengesiz koşullarda ulaşabildiği anlamlı bir başarı düzeyidir. Daha da önemlisi, bu modelin prescriptive öneri sistemini besleyerek yüzde 26,5'lik göreli bir dönüşüm artışı sağlaması, düşük gibi görünen F1 skorunun pratikte değer ürettiğinin en somut kanıtıdır.

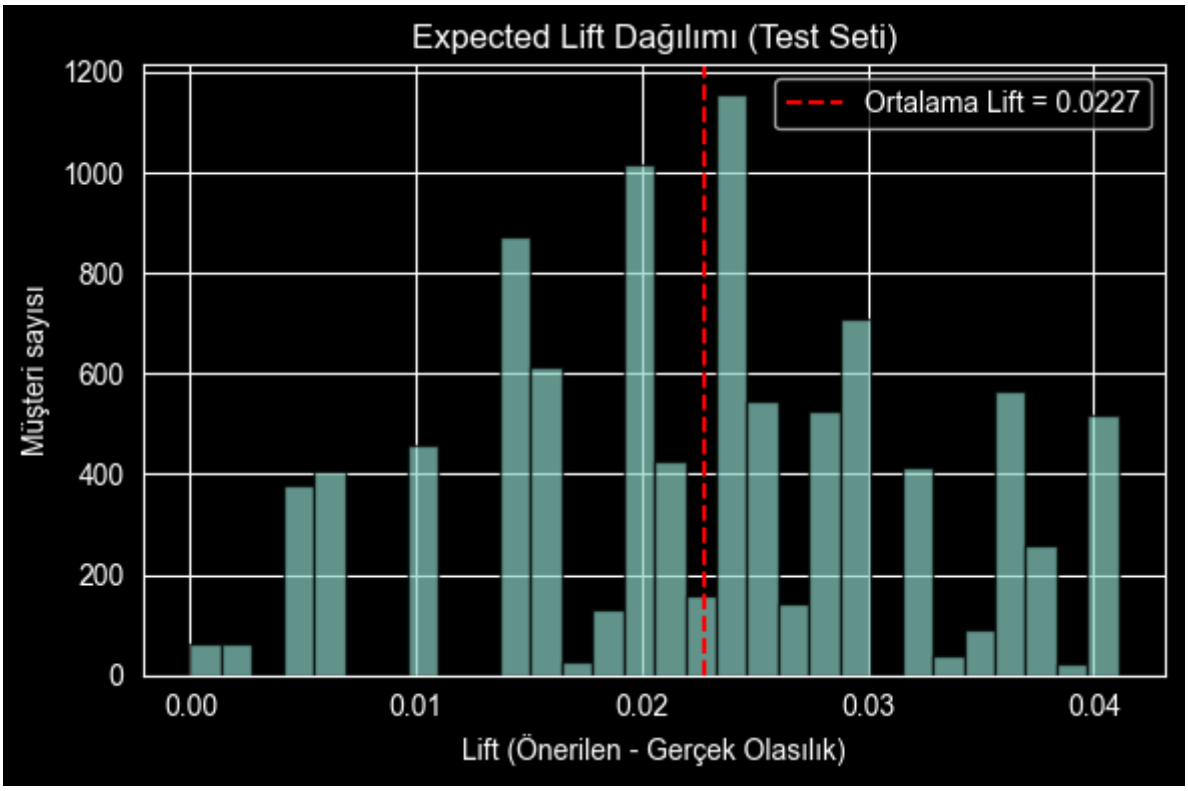
4.4. Prescriptive Öneri Sistemi Bulguları

Test setindeki 9.600 müşteri üzerinde gerçekleştirilen lift analizi, 3.7 bölümünde tasarlanan prescriptive sistemin performansını somutlaştırmaktadır. Sistem her müşteri için 49 kanal-platform kombinasyonunu değerlendirmiş ve en yüksek dönüşüm olasılığını veren kombinasyonu önermiştir. Sonuçlar Tablo 5'te özetlenmektedir.

Tablo 5. Prescriptive öneri sistemi lift analizi özet tablosu.

Metrik	**Değer**
Test Seti Müşteri Sayısı	9.600
Mevcut Ortalama Dönüşüm Olasılığı	0,0855 (%8,55)
Önerilen Ortalama Dönüşüm Olasılığı	0,1082 (%10,82)
Ortalama Lift (Fark)	**+0,0227**
Göreli İyileşme	**≈ %26,5**
Maksimum Bireysel Lift	0,04+
Sıfıra Yakın Lift Alan Müşteri Oranı	≈ %15–20 (zaten optimal kombinasyonda)

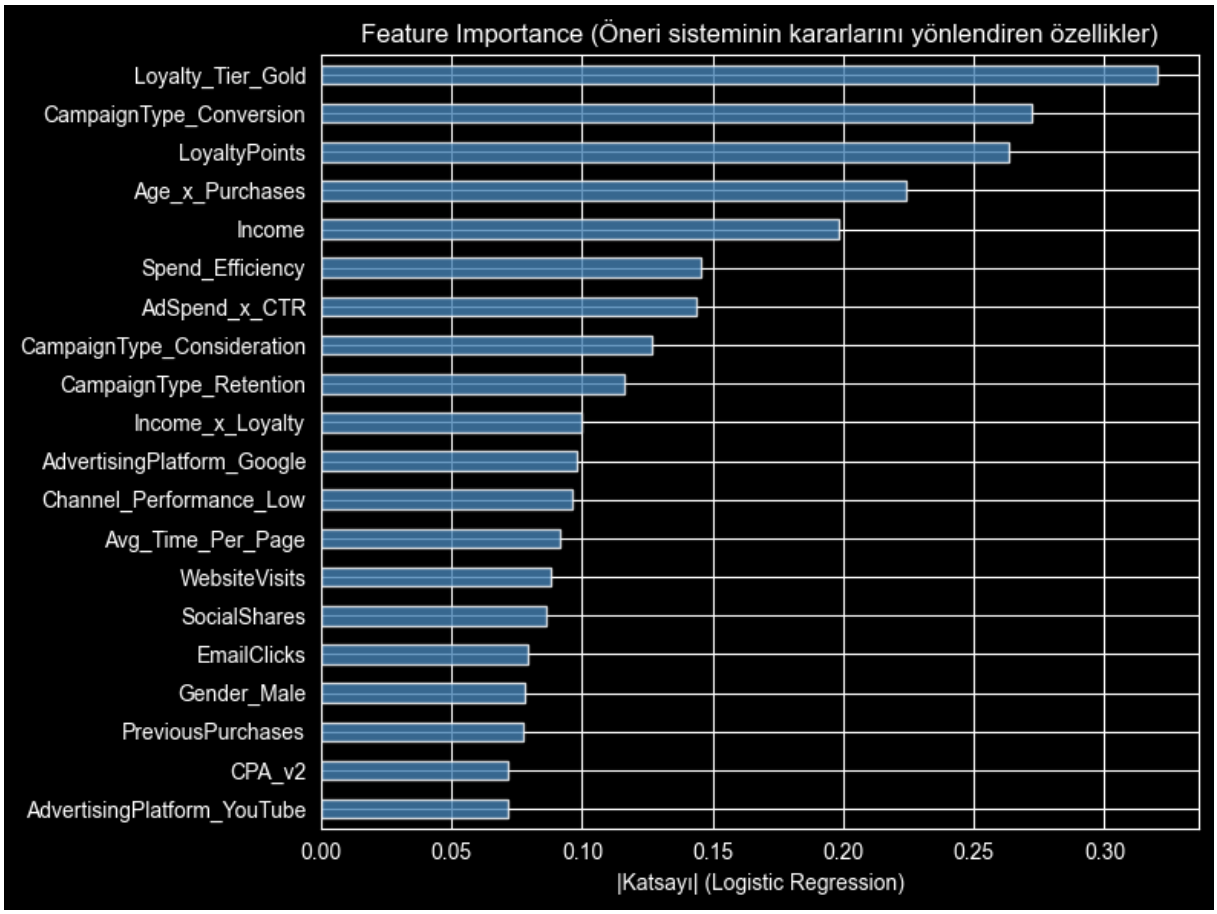
Mevcut kampanya kombinasyonlarındaki ortalama dönüşüm olasılığı 0,0855 iken önerilen kombinasyonlarda bu değer 0,1082'ye yükselmektedir. Bu fark ilk bakışta küçük gibi görünse de yüzde ikinin altındaki bir dönüşüm tabanında yaklaşık yüzde 26,5'lik göreli bir iyileşmeye karşılık gelmektedir. Dijital pazarlama endüstrisinde bu büyüklükte bir göreli iyileşme, özellikle yüksek hacimli kampanyalarda ciddi gelir artışı anlamına gelebilir. Örneğin aylık 100.000 müşteriye kampanya gönderen bir şirket için bu oran, yaklaşık 2.270 ek dönüşüm potansiyeli demektir.



Şekil 11. Test seti üzerinde beklenen lift dağılımı. Ortalama lift = 0,0227.

Şekil 11'deki dağılım, lift değerinin müşteriler arasında homojen biçimde dağılmadığını göstermektedir. Bu heterojen yapı önemli bir içgörü sunmaktadır: bazı müşterilerin mevcut kampanya kombinasyonu zaten optimale yakın olduğundan sıfıra yakın lift üretilmektedir. Bu müşteriler için kanal değiştirmenin bir anlamı yoktur. Buna karşın belirli segmentlerde 0,04 düzeyine ulaşan marjinal katkılar görülmektedir. Bu yüksek liftli müşteriler, şu anda yanlış kanal üzerinden hedeflenen ve doğru kanala yönlendirildiğinde dönüşüm olasılığı belirgin biçimde artan profilleri temsil etmektedir. Prescriptive sistemin değeri, tam da bu selektif kullanım alanında ortaya çıkmaktadır.

Şekil 12'de görülen değişken önem sıralaması, öneri sisteminin kararlarını hangi özelliklerin yönlendirdiğini ortaya koymaktadır. Loyalty_Tier_Gold, CampaignType_Conversion ve LoyaltyPoints ilk üç sırayı paylaşmaktadır. Bu sonuç, tahmin modeli (Şekil 10) ve öneri sistemi (Şekil 12) arasında tutarlı bir örüntü olduğunu doğrulamaktadır: her iki sistemde de müşteri sadakati en belirleyici faktördür.



Şekil 12. Prescriptive öneri sisteminin kararlarını yönlendiren değişkenler.

Pazarlama uygulamaları açısından bu bulgular üç somut öneri üretmektedir. Birincisi, sadakat programı yönetimi ve Gold tier müşterilere yönelik kampanyalar önceliklendirilmelidir çünkü bu segment dönüşüm artırma potansiyelinin en yüksek olduğu gruptur. İkincisi, dönüşüm odaklı kampanya türleri (Conversion tipi kampanyalar) diğer kampanya türlerine göre belirgin biçimde daha etkilidir. Üçüncüsü, prescriptive sistem tüm müşterilere aynı değeri sunmamakta; en yüksek lift potansiyeline sahip segmentlere odaklanarak kaynakların verimli kullanılmasına olanak tanımaktadır.

5. SONUÇ VE ÖNERİLER

Bu çalışmada dijital pazarlama kampanyalarında dönüşüm tahmini ve aksiyon optimizasyonu problemi, prescriptive analitik çerçevesinde ele alınmıştır. 48.000 sentetik müşteri kaydı üzerinde uçtan uca bir makine öğrenmesi pipeline'ı ve öneri sistemi tasarlanmış;

süreç boyunca karşılaşılan en kritik metodolojik güçlük veri sızıntısının sistematik biçimde tespit edilmesi ve giderilmesi olmuştur.

Çalışmanın en önemli katkısı bu noktadadır. Modele ilk dâhil edilen özellik seti ile elde edilen mükemmel ROC-AUC ve F1 skorları, gerçek bir tahmin yeteneğinin değil sızıntı kaynaklı bir yapay başarının ürünüydü. ConversionRate değişkeninin hedef değişkenden türetilmiş olduğunun tespit edilmesi ve CPA_Proxy'nin tek başına tüm yapay başarıyı sürdürebildiğinin kanıtlanması, Kaufman ve diğerlerinin (2012) tanımladığı hedef sızdırma probleminin somut bir örneğini oluşturmaktadır. Senaryo tabanlı arınma süreci yürütüldüğünde performans dramatik biçimde gerilemiş; ancak bu gerileme bir başarısızlık değil modelin artık dürüst bir tahmin görevi yaptığının göstergesi olarak okunmuştur.

Sızıntısız veri setinde elde edilen ROC-AUC 0,713 ve F1 yaklaşık 0,15 değerleri, yüzde ikinin altındaki dönüşüm oranı bağlamında anlamlı bir sınıflandırma kapasitesine karşılık gelmektedir. 4.3.1 bölümünde detaylı biçimde tartışıldığı üzere bu F1 değeri, aşırı dengesiz sınıf yapısı göz önüne alındığında rastgele tabanın belirgin biçimde üzerindedir. Daha da önemlisi bu modelin prescriptive öneri sistemini besleyerek pratikte değer üretmesi, düşük gibi görünen metriklerin bağlamla birlikte değerlendirilmesinin önemini göstermektedir.

Prescriptive öneri sistemi test setinde ortalama 0,0227 dönüşüm olasılığı artışı sağlamış; bu yaklaşık yüzde 26,5'lik göreli bir iyileşmeye karşılık gelmektedir. 2.4.2.3 bölümünde kavramsal olarak açıklanan ve 3.7 bölümünde teknik olarak tasarlanan bu sistem, her müşteri için 49 kanal-platform kombinasyonunu simüle ederek optimal aksiyonu belirlemektedir. Lepenioti ve diğerlerinin (2020) tanımladığı prescriptive analitik katmanının pazarlama alanındaki somut bir uygulaması olarak bu sistem, "ne olacak" sorusunun ötesine geçerek "ne yapmalıyız" sorusuna veri temelli yanıtlar üretmektedir.

Çalışmanın ulaştığı temel sonuçlar şu şekilde özetlenebilir: Birincisi, veri sızıntısı tespiti ve giderilmesi, gözetimli öğrenme projelerinde model geliştirme sürecinin ayrılmaz bir parçası olmalıdır. İkincisi, aşırı dengesiz veri setlerinde doğruluk (accuracy) metriği tek başına yanıltıcıdır ve F1 skoru ile ROC-AUC birlikte değerlendirilmelidir. Üçüncüsü, prescriptive analitik yaklaşımı geleneksel tavsiye sistemlerinin sınırlı kaldığı düşük dönüşüm ortamlarında anlamlı bir alternatif sunmaktadır. Dördüncüsü, müşteri sadakati (LoyaltyPoints,

Loyalty_Tier_Gold) dijital pazarlama dönüşümünün en güçlü belirleyicisidir ve kampanya stratejilerinde önceliklendirilmelidir.

5.1. Kısıtlar ve Gelecek Çalışmalar

Bu çalışmanın birkaç önemli kısıtı mevcuttur ve bu kısıtların açıkça belirtilmesi bulguların doğru yorumlanması açısından gereklidir.

Sentetik veri kısıtı: Veri setinin sentetik yapısı, bulguların gerçek müşteri davranışını ne ölçüde yansıttığını sorgulama zorunluluğunu doğurmaktadır. Sentetik verilerde gerçek dünyadaki gürültü, mevsimsellik ve müşteri davranış değişkenliği tam olarak yakalanamayabilir. Bu nedenle elde edilen dönüşüm oranları ve lift değerleri, gerçek bir işletme verisinde farklılık gösterebilir.

Kanal önerisi tek tipleşmesi: Prescriptive sistemin Display kanalını neredeyse tüm segmentler için önermesi, modelin eğitildiği sentetik verideki kanal dağılımının bir yansıması olabilir. Gerçek veri üzerinde farklı kanalların farklı segmentlerde öne çıkması beklenir. Bu tek tipleşme, modelin kanal özelinde yeterince ayrıştırıcı sinyal bulamadığına işaret etmektedir.

SMOTE'un etkisizliği: Sınıf dengesizliğinin bu denli ağır olduğu bir yapıda (∼%2 pozitif oran) SMOTE uygulamasının (Chawla ve diğerleri, 2002) modeli iyileştirememesi, yöntemin sınırlarına ilişkin önemli bir gözlemdir. Sentetik örneklerin gerçek azınlık sınıfının örüntülerini yeterince temsil edemediği ve modele gürültü eklediği değerlendirilmektedir.

Zaman boyutunun eksikliği: Mevcut veri setinde zaman damgası bilgisi bulunmamaktadır. Gerçek pazarlama verilerinde kampanyaların gönderim zamanı, haftanın günü, mevsimsel döngüler ve müşterinin son etkileşiminden geçen süre gibi zamansal özellikler dönüşümü önemli ölçüde etkilemektedir.

Bu kısıtlar göz önünde bulundurulduğunda gelecek çalışmalar üç temel eksende ilerleme kaydedebilir.

1. Özellik zenginleştirme: Zaman damgası bazlı özellikler (gönderim saati, günü, son etkileşimden geçen süre) ve müşteri segmentine özgü kampanya geçmişiyle zenginleştirilmiş bir özellik seti model kapasitesini artırabilir. Özellikle RFM (Recency, Frequency, Monetary)

temelli türetilmiş değişkenler, Khajvand ve diğerlerinin (2011) çalışmasında gösterildiği üzere müşteri yaşam boyu değeri tahmininde güçlü sinyaller üretmektedir.

2. Uyarlamalı örnekleme teknikleri: SMOTE'un etkisiz kaldığı bu yapıda ADASYN (Adaptive Synthetic Sampling) gibi uyarlamalı örnekleme tekniklerinin dengesizlik problemi üzerindeki etkisi araştırılmaya değerdir. ADASYN, azınlık sınıfının öğrenilmesi zor olan bölgelerine daha fazla sentetik örnek üreterek SMOTE'un homojen örnekleme sınırını aşmayı hedeflemektedir.

3. Derin öğrenme temelli hibrit sistem: Neural Collaborative Filtering gibi derin öğrenme temelli yaklaşımların prescriptive sistemle hibridize edilmesi, müşteri segmentasyonu ve öneri kalitesini birlikte iyileştirme potansiyeli taşımaktadır. Bu hibrit yapı, 2.4.2.4 bölümünde tartışılan işbirlikçi filtreleme ve prescriptive sistemlerin tamamlayıcı doğasından faydalanabilir.

4. Gerçek veri validasyonu: En önemlisi, bu çalışmada geliştirilen pipeline'ın gerçek bir şirketin pazarlama verisi üzerinde test edilmesidir. Sanayi danışmanımızın görev yaptığı firmanın mevcut TEYDEB projesi kapsamındaki veriler, bu validasyon için doğal bir aday oluşturmaktadır.

ÇIKTILAR

Yayınlar ve Sunumlar

Bu çalışma, TÜBİTAK 2209-B Üniversite Öğrencileri Sanayiye Yönelik Araştırma Projeleri Desteği Programı kapsamında 2024-25 döneminde "Makine Öğrenmesi Yardımıyla Kanal Analitiği Verilerinin Etiketlenmesi" başlıklı proje olarak desteklenmiştir.

Proje çıktıları, Ege Üniversitesi Planlama ve Başarı Koordinatörlüğü, Sağlık, Kültür ve Spor Daire Başkanlığı ile Öğrenci Koordinatörlüğü iş birliğiyle 28.05.2025 tarihinde düzenlenen 70. Yıl Bilim Şenliği kapsamında poster olarak sunulmuştur. Katılım belgesi Ek-1'de sunulmuştur.

Proje Çıktıları

Proje kapsamında beş Jupyter notebook dosyası üretilmiştir. Bu çıktıların her biri, çalışmanın farklı bir aşamasını kapsayan bağımsız ve yeniden çalıştırılabilir modüllerdir:

Tablo 6. Proje çıktıları ve içerikleri.

No	Çıktı	Açıklama
1	EDA ve Veri Temizleme	Eksik değer analizi, aykırı değer tespiti, korelasyon analizi ve istatistiksel hipotez testlerini içeren keşifsel veri analizi pipeline'ı.
2	Özellik Mühendisliği	17 türetilmiş özellik içeren özellik mühendisliği modülü: ROI metrikleri, etkileşim özellikleri, müşteri segmentasyonu, etkileşim özellikleri ve kanal performansı.
3	Kanal Performans Analizi	Kanal bazlı dönüşüm oranları ve müşteri edinim maliyeti (CPA) analiz sistemi.
4	Veri Sızıntısı Tespiti	Altı senaryo üzerinden kademeli arınma sürecini yürüten senaryo tabanlı veri sızıntısı tespit ve giderilme çerçevesi.
5	Prescriptive Öneri Sistemi	Her müşteri için 49 kanal-platform kombinasyonunu simüle eden strateji simülatörü tabanlı öneri sistemi.

Üretim modeli **final_model.pkl**, ölçekleyici **scaler.pkl** ve eksik değer doldurma nesnesi **imputer.pkl** dosyaları olarak serileştirilmiş durumdadır. Bu üç dosya birlikte kullanılarak herhangi bir yeni müşteri verisi üzerinde tahmin ve öneri üretilebilir.

KAYNAKÇA

- Aggarwal, C. C. (2016). Recommender systems: The textbook. Springer. https://doi.org/10.1007/978-3-319-29659-3
- Altınok, V. (2019). Kümeleme analizi ve uygulamaları (Yüksek lisans tezi). Erişim: YÖK Ulusal Tez Merkezi.
- Bertsimas, D. ve Kallus, N. (2020). From predictive to prescriptive analytics. Management Science, 66(3), 1025–1044. https://doi.org/10.1287/mnsc.2018.3253
- 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. ve 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. ve Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785
- Hong, Y. (2024). Digital marketing analytics and customer engagement: A systematic review. Journal of Digital & Social Media Marketing, 12(1), 45–62.
- Ijomah, T. I., Idemudia, C., Eyo-Udo, N. L. ve Anjorin, K. F. (2024). Digital marketing analytics: A review of strategies for enhanced customer engagement. International Journal of Management & Entrepreneurship Research, 6(3), 950–962.
- Kaufman, L. ve Rousseeuw, P. J. (1991). Finding groups in data: An introduction to cluster analysis. John Wiley & Sons. https://doi.org/10.1002/9780470316801
- Kaufman, S., Rosset, S., Perlich, C. ve Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. ACM Transactions on Knowledge Discovery from Data, 6(4), 1–21. https://doi.org/10.1145/2382577.2382579
- Khajvand, M., Zolfaghar, K., Ashoori, S. ve Alizadeh, S. (2011). Estimating customer lifetime value based on RFM analysis of customer purchase behavior. Procedia Computer Science, 3, 57–63. https://doi.org/10.1016/j.procs.2010.12.011
- Koren, Y., Bell, R. ve Volinsky, C. (2009). Matrix factorization techniques for recommender systems. Computer, 42(8), 30–37. https://doi.org/10.1109/MC.2009.263
- Kutluğun, M. A., Ali, N. ve Kaleli, C. (2017). Makine öğrenmesi teknikleri ile birlikte çalışabilir tavsiye sistemleri. DÜMF Mühendislik Dergisi, 8(3), 627–636.
- Lepenioti, K., Bousdekis, A., Apostolou, D. ve Mentzas, G. (2020). Prescriptive analytics: Literature review and research challenges. Information Systems, 87, 101385. https://doi.org/10.1016/j.is.2019.101385

- Linden, G., Smith, B. ve York, J. (2003). Amazon.com recommendations: Item-to-item collaborative filtering. IEEE Internet Computing, 7(1), 76–80. https://doi.org/10.1109/MIC.2003.1167344
- Lops, P., de Gemmis, M. ve Semeraro, G. (2011). Content-based recommender systems: State of the art and trends. F. Ricci, L. Rokach, B. Shapira ve P. B. Kantor (Ed.), Recommender systems handbook içinde (s. 73–105). Springer. https://doi.org/10.1007/978-0-387-85820-3_3
- Nizam, H. ve Akın, S. S. (2014). Sınıflandırmada yapay sinir ağları ve karar ağaçları karşılaştırması. 19. Türkiye'de İnternet Konferansı (INET-TR), 1–6. https://inet-tr.org.tr/inetconf19/kitap/nizam_akin_inet14.pdf
- Nwabekee, U. S., Oliha, J. S., Etukudoh, E. A. ve Oguejiofor, B. B. (2024). Digital marketing analytics: Rethinking consumer engagement strategies. International Journal of Frontiers in Science and Technology Research, 6(2), 1–18.
- Öztemel, E. (2006). Yapay sinir ağları. Papatya Yayıncılık.
- Pazzani, M. J. ve Billsus, D. (2007). Content-based recommendation systems. P. Brusilovsky, A. Kobsa ve W. Nejdl (Ed.), The adaptive web içinde (s. 325–341). Springer. https://doi.org/10.1007/978-3-540-72079-9_10
- Sarwar, B., Karypis, G., Konstan, J. ve Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. Proceedings of the 10th International Conference on World Wide Web, 285–295. https://doi.org/10.1145/371920.372071
- Scikit-learn. (2024). Decision trees. https://scikit-learn.org/stable/modules/tree.html
- Tran, T. N. T., Felfernig, A., Trattner, C. ve Holzinger, A. (2021). Recommender systems in the healthcare domain: State-of-the-art and research issues. Journal of Intelligent Information Systems, 57, 171–201. https://doi.org/10.1007/s10844-020-00633-6
- Tunç, H., Dama, N. ve Özdamar, K. (2011). Hiyerarşik kümeleme analizinin temel ilkeleri ve uygulamaları. Osmangazi Tıp Dergisi, 33(2), 15–25.
- Şekil-0. https://www.researchgate.net/profile/Javad-Hassannataj-Joloudari/publication/363269818/figure/fig2/AS:11431281083012095@1662348052706/llustration-of-SMOTE-Oversampling-for-Imbalanced-Classification.jpg
- Şekil-1. https://www.researchgate.net/figure/Content-based-filtering-vs-Collaborative-filtering-Source_fig5_323726564

EKLER

Ek-1. 70. Yıl Bilim Şenliği Katılım Belgesi