YALOVA ÜNİVERSİTESİ
MÜHENDİSLİK FAKÜLTESİ
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

- MAKİNE ÖĞRENMESİ -

MÜŞTERİ SATIN ALMA DAVRANIŞI TAHMİNİ PROJESİ

İLYAS SÜZMEN - 190101086
FARUK NAMAL - 190101035

YALOVA , 2025

1.VERİ SETİNİN İNCELENMESİ

"Customer Personality Analysis" veri seti, bir şirketin müşterileri hakkında kapsamlı bilgiler sunarak, pazarlama stratejilerinin kişiselleştirilmesine ve müşteri segmentasyonuna olanak tanır. Bu veri seti, müşterilerin demografik özelliklerinden alışveriş alışkanlıklarına, kampanyalara verdikleri tepkilerden web sitesi etkileşimlerine kadar geniş bir yelpazeyi kapsar.

Veri setinde yer alan sütunlar; müşterilerin yaşam tarzlarını, harcama davranışlarını ve markayla olan etkileşim düzeylerini anlamamıza yardımcı olur. Böylece pazarlama ekipleri, hedef kitlelerini daha iyi tanıyabilir, müşteri memnuniyetini artıracak kişiselleştirilmiş kampanyalar geliştirebilir ve müşteri yaşam döngüsünü optimize edebilir.

Veri setindeki sütunlar, müşterilerin davranışsal ve demografik özelliklerini yedi ana kategori altında toplamaktadır. Demografik bilgiler; yaş, medeni durum, eğitim düzeyi ve gelir gibi temel kişisel nitelikleri içerir. Hane bilgileri, evde yaşayan çocuk sayısı ve müşteriyle kurulan ilk temas tarihi gibi aile yapısı ve ilişki süresiyle ilgili verileri sunar. Harcama alışkanlıkları; et, balık, tatlı, meyve, şarap ve altın gibi farklı ürün kategorilerine yönelik tüketim düzeylerini yansıtır. Alışveriş kanalları, müşterinin web, katalog ya da fiziksel mağaza gibi kanallar aracılığıyla gerçekleştirdiği alışveriş tercihlerini analiz etmeye olanak tanır. Pazarlama tepkileri, müşterilerin geçmiş kampanyalara verdiği yanıtları ortaya koyar. Web etkileşimleri, aylık site ziyaret sayısı gibi dijital davranışları değerlendirir. Son olarak, şikayet ve sadakat göstergeleri, müşterilerin memnuniyet düzeylerini ve markaya olan bağlılıklarını anlamada kritik rol oynayan metrikleri içerir.

Bu veri seti, hem istatistiksel analizler hem de makine öğrenmesi modelleri için sağlam bir temel oluşturur. Müşteri davranışlarının tahmin edilmesi ve segmentasyon gibi farklı analizlerde etkili şekilde kullanılabilir.

Bu çalışmada, müşterilerin satın alma eğilimlerini tahmin etmeye yönelik bir veri setinden yararlanılmıştır. Aşağıdaki tablo veri setindeki önemli sütunları, veri tiplerini ve kısa

açıklamalarını içermektedir:

Sütun Adı	Veri Tipi	Açıklama
ID	Integer	Müşteri benzersiz kimlik numarası
Year_Birth	Integer	Doğum yılı
Education	Categorical	Eğitim durumu (Bachelor, Master, PhD, vs.)
Marital_Status	Categorical	Medeni durum (Evli, Bekar, vb.)
Kidhome	Integer	Evdeki çocuk sayısı (0,1,2)
Teenhome	Integer	Evdeki ergen çocuk sayısı
Income	Float	Yıllık gelir (eksik değerler mevcut)
Dt_Customer	Date	Müşterinin kayıt tarihi
Recency	Integer	Son alışverişten sonra geçen gün sayısı
MntWines	Integer	Yıllık şarap harcaması
MntFruits	Integer	Yıllık meyve harcaması
MntMeatProducts	Integer	Yıllık et ürünleri harcaması
MntFishProducts	Integer	Yıllık balık ürünleri harcaması
MntSweetProducts	Integer	Yıllık tatlı harcaması
MntGoldProds	Integer	Yıllık altın ürünleri harcaması
NumDealsPurchases	Integer	Kampanya sırasında yapılan indirimli alışveriş sayısı
NumWebPurchases	Integer	İnternet üzerinden yapılan alışveriş sayısı

Sütun Adı	Veri Tipi	Açıklama
NumCatalogPurchases	Integer	Katalog üzerinden yapılan alışveriş sayısı
NumStorePurchases	Integer	Mağaza üzerinden yapılan alışveriş sayısı
NumWebVisitsMonth	Integer	Son bir ayda web sitesi ziyaret sayısı
AcceptCmp1	Binary	İlk kampanyaya katılım durumu
AcceptCmp2	Binary	İkinci kampanyaya katılım durumu
AcceptCmp3	Binary	Üçüncü kampanyaya katılım durumu
AcceptCmp4	Binary	Dördüncü kampanyaya katılım durumu
AcceptCmp5	Binary	Beşinci kampanyaya katılım durumu
Complain	Binary	Müşterinin son 2 yılda herhangi bir şikayette bulunma durumu
Z_CostContact	Integer	Pazarlama kampanyaları kapsamında müşteriye ulaşmak için yapılan maliyet
Z_Revenue	Integer	Müşterinin şirkete sağladığı gelir düzeyi
Response	Binary	Kampanya tepkisi (1 = olumlu, 0 = olumsuz)

2. VERİ ÖN İŞLEME VE ÖZELLİK MÜHENDİSLİĞİ

Veri seti üzerinde yapılan ön işlemler ve özellik mühendisliği adımları, model başarısını artırmak ve verinin kalitesini iyileştirmek amacıyla gerçekleştirilmiştir.

2.1 Eksik Değerlerin Doldurulması

Income sütununda bazı kayıtların eksik olduğu tespit edilmiştir. Eksik değerlerin uygun şekilde işlenmemesi, model performansını olumsuz etkileyebileceği için doldurma işlemi yapılması gerekmektedir.

```
📉 Eksik Veri Sayıları:
ID                     0
Year_Birth             0
Education              0
Marital_Status         0
Income                24
Kidhome                0
Teenhome               0
Dt_Customer            0
Recency                0
MntWines               0
MntFruits              0
MntMeatProducts        0
MntFishProducts        0
MntSweetProducts       0
MntGoldProds           0
NumDealsPurchases      0
NumWebPurchases        0
NumCatalogPurchases    0
NumStorePurchases      0
NumWebVisitsMonth      0
AcceptedCmp3           0
AcceptedCmp4           0
AcceptedCmp5           0
AcceptedCmp1           0
AcceptedCmp2           0
Complain               0
Z_CostContact          0
Z_Revenue              0
Response               0
dtype: int64
```

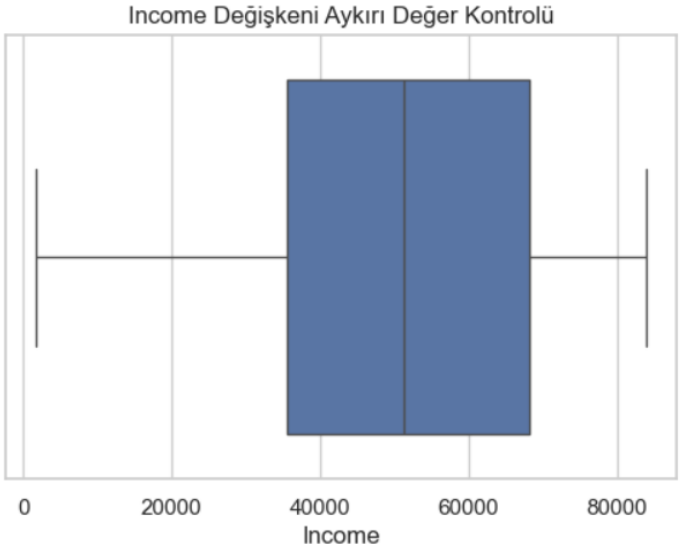
Income değişkeni incelendiğinde, veride ciddi oranda çarpıklık ve uç değerlerin bulunduğu gözlemlenmiştir. Bu tür durumlarda ortalama değerinin kullanılması, aşırı yüksek veya düşük uç değerlerden etkilenerek yanıltıcı sonuçlar doğurabilir. Bu nedenle, merkezi eğilim ölçüsü olarak medyan tercih edilmiştir. Medyan, verinin ortanca değeri olması sebebiyle uç değerlerin

etkisini minimize eder ve gelir dağılımındaki dengesizliği daha sağlam şekilde temsil eder.

Böylece, Income sütunundaki eksik veriler, veri setinin genel yapısına daha uygun ve tutarlı bir değerle doldurularak, modelin doğruluğunu ve genellenebilirliğini artırmak hedeflenmiştir.

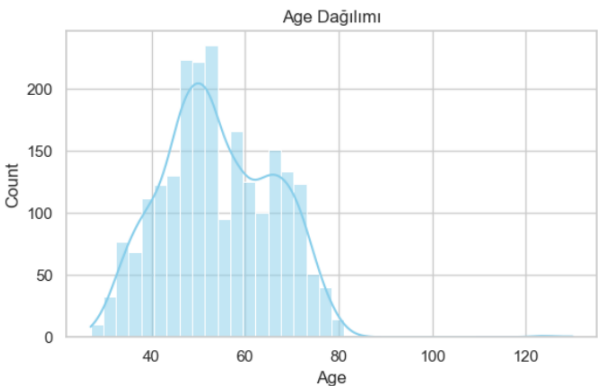
2.2 Aykırı Değer Sınırlaması

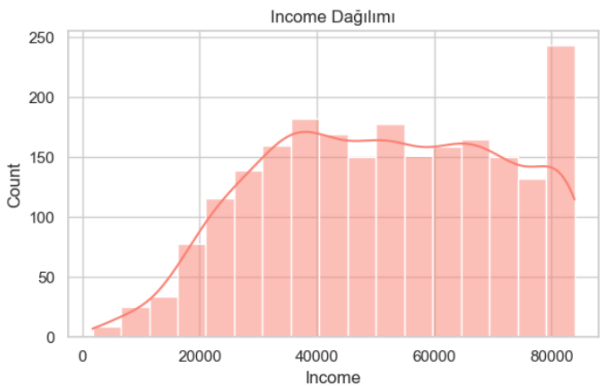
Aykırı değerlerin modellenme sürecine olumsuz etkisini azaltmak için, gelir değişkenindeki aşırı yüksek değerler %95'lik üst sınır değerine sabitlenmiştir. Bu sayede uç değerlerin model performansını bozmasının önüne geçilmiştir.



2.3 Yeni Özelliklerin Eklenmesi

Özellik mühendisliği kapsamında, müşterilerin yaşları doğum yılı bilgilerinden hesaplanarak modele anlamlı bir değişken olarak eklenmiştir. Evdeki toplam çocuk sayısını temsil eden yeni bir değişken oluşturulmuş (Kidhome ve Teenhome toplamı), ayrıca müşterinin kayıt tarihinden veri setindeki son tarihe kadar geçen süre gün bazında hesaplanarak "üyelik süresi" şeklinde bir özellik elde edilmiştir.





2.4 Kategorik Değişkenlerin İşlenmesi

Veri setinde yer alan bazı kategorik değişkenler, modelin öğrenme sürecine dahil edilebilmesi için sayısal formata dönüştürülmüştür. Özellikle Education ve Marital_Status gibi değişkenler, doğrudan sayısal anlam taşımadıkları için makine öğrenmesi algoritmaları tarafından işlenemez. Bu nedenle, bu tür nominal kategorik değişkenler one-hot encoding yöntemiyle işlenmiştir. One-hot encoding, her kategori için ayrı bir sütun oluşturarak, bu sütunlara ilgili kategoriye ait olup olmadığını belirten ikili değerler atar. Bu yöntem, algoritmaların kategorik değişkenler arasında yapay büyüklük veya sıralama ilişkisi varsayımına girmesini engeller ve böylece model her kategoriye eşit ve tarafsız yaklaşarak önyargısız öğrenme sağlar.

One-hot encoding'in tercih edilmesinin temel nedeni, kategorik değişkenlerin birbirleriyle sıralı veya aritmetik bir ilişki içinde olmadığı durumlarda bu yöntemin en doğru ve etkili temsil biçimi olmasıdır. Alternatif yöntemler, örneğin etiket kodlama (label encoding), kategorilere sayısal sıralama atayarak modelin yanlışlıkla kategoriler arasında hiyerarşi algılamasına neden olabilir. Bu da modelin performansını olumsuz etkileyebilir. Bu nedenle, kategorik verilerin doğru ve tarafsız biçimde modellenmesi için one-hot encoding kullanılmıştır.

Ayrıca, Marital_Status değişkeninde çok az sayıda gözleme sahip olan 'Alone', 'Absurd' ve 'YOLO' gibi nadir kategoriler 'Single' sınıfı altında birleştirilmiştir. Bu sayede sınıf

dengesizliği azaltılmış ve modelin nadir kategorilere aşırı duyarlı olması önlenerek daha istikrarlı bir öğrenme süreci sağlanmıştır.

2.5.Gereksiz Sütunların Çıkarılması

Modelin doğruluğunu ve genellenebilirliğini artırmak amacıyla, hedef değişkenle doğrudan ilişkili olan veya model için anlamsız kabul edilen bazı sütunlar veri setinden çıkarılmıştır.

Buna göre; Z_CostContact ve Z_Revenue gibi hedef değişkenle bağlantılı sütunlar, ayrıca ID, Year_Birth, Dt_Customer, Kidhome ve Teenhome gibi analiz sürecinde artık kullanılmayan sütunlar kaldırılmıştır. Böylece modelin gereksiz bilgiyle karmaşıklaşması önlenmiştir.

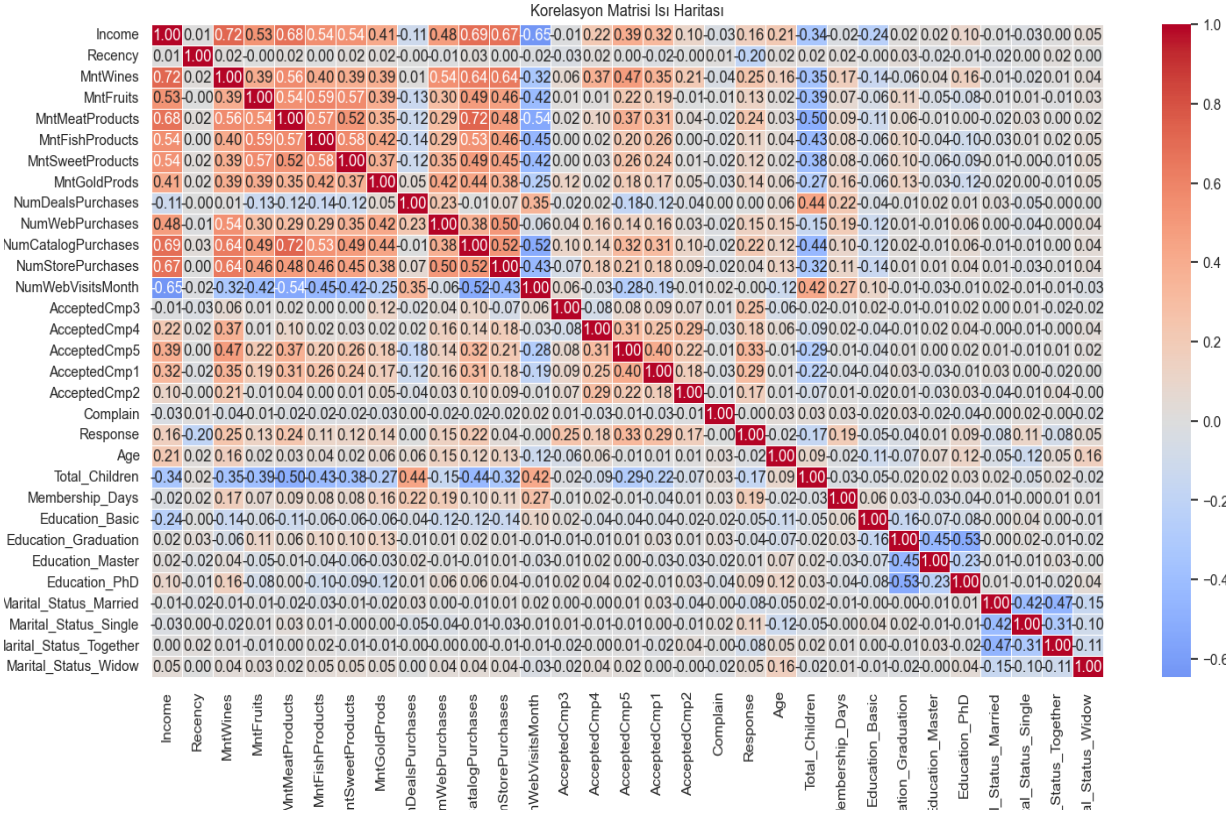
3. Keşifsel Veri Analizi

Modelleme sürecine başlamadan önce, veri setindeki özelliklerin birbirleriyle ve hedef değişkenle olan ilişkilerini anlamak amacıyla keşifsel veri analizi yapıldı. Bu aşama, veri setinin yapısını anlamak, olası anormallikleri tespit etmek ve model performansını etkileyebilecek ilişkileri değerlendirmek için kritik öneme sahiptir.

Özellikler arasındaki doğrusal ilişkileri incelemek için Pearson korelasyon katsayısı kullanılarak bir korelasyon matrisi oluşturuldu. Korelasyon matrisi, her bir değişken çifti arasındaki ilişkinin yönünü ve gücünü -1 ile +1 arasında bir değerle ifade eder. Pozitif değerler değişkenlerin birlikte arttığını, negatif değerler ise ters yönlü ilişki olduğunu gösterir.

Bu matrisi daha kolay yorumlayabilmek için seaborn kütüphanesi ile bir ısı haritası görselleştirildi. Isı haritasında koyu kırmızı tonlar güçlü pozitif korelasyonu, koyu mavi tonlar ise güçlü negatif korelasyonu temsil etmektedir. Renk skalası, korelasyonların görsel olarak

hızlı bir şekilde anlaşılmasını sağlamaktadır.



Analiz sonucunda, bazı harcama kalemlerinin birbirleriyle ve bazı satın alma davranışları ile güçlü pozitif korelasyona sahip olduğu görüldü. Örneğin, yıllık şarap harcaması (MntWines) ile et ürünleri harcaması (MntMeatProducts) arasında anlamlı bir pozitif ilişki bulundu. Benzer şekilde, internet üzerinden alışveriş sayısı (NumWebPurchases) ile katalog ve mağaza alışveriş sayıları da belirgin bir şekilde ilişkilendirildi.

⚠️ Yüksek Korelasyonlu Özellikler (|r| > 0.85): ['Income', 'Recency', 'MntWines', 'MntFruits', 'MntMeatProducts', 'MntFishProducts', 'MntSweetProducts', 'MntGoldProds', 'NumDealsPurchases', 'NumWebPurchases', 'NumCatalogPurchases', 'NumStorePurchases', 'NumWebVisitsMonth', 'AcceptedCmp3', 'AcceptedCmp4', 'AcceptedCmp5', 'AcceptedCmp1', 'AcceptedCmp2', 'Complain', 'Age', 'Total_Children', 'Membership_Days', 'Education_Basic', 'Education_Graduation', 'Education_Master', 'Education_PhD', 'Marital_Status_Married', 'Marital_Status_Single', 'Marital_Status_Together', 'Marital_Status_Widow']

Hedef değişken olan Response ile korelasyon incelendiğinde, belirli alışveriş ve kampanya katılım özelliklerinin olumlu tepkiyi artırmada rol oynadığı belirlendi. Ancak, Response

değişkeni ikili yapıda olduğundan korelasyon katsayısı genellikle düşük değerlerde kaldı; bu da sınıflandırma problemlerinde sık karşılaşılan bir durumdur.

📈 Response ile En Yüksek Korelasyonlar:

Response	1.000000
AcceptedCmp5	0.326634
AcceptedCmp1	0.293982
AcceptedCmp3	0.254258
MntWines	0.247254
MntMeatProducts	0.236335
NumCatalogPurchases	0.220810
Membership_Days	0.194481
AcceptedCmp4	0.177019
AcceptedCmp2	0.169293

Name: Response, dtype: float64

📉 Response ile En Düşük Korelasyonlar:

Recency	-0.198437
Total_Children	-0.169163
Marital_Status_Married	-0.079378
Marital_Status_Together	-0.075770
Education_Basic	-0.049451
Education_Graduation	-0.040217
Age	-0.021325
NumWebVisitsMonth	-0.003987
Complain	-0.001707
NumDealsPurchases	0.002238

Name: Response, dtype: float64

Çoklu Bağlantı (Multicollinearity) Değerlendirmesi

EDA sürecinde yüksek korelasyona sahip bazı özellikler tespit edildi. Özellikle, birbirleriyle oldukça ilişkili değişkenlerin modele birlikte dahil edilmesi, çoklu bağlantı sorunlarına yol açabilir. Çoklu bağlantı, modelin parametre tahminlerinin kararsızlaşmasına, bazı özelliklerin gereğinden fazla veya az önemsenmesine ve genel performansın olumsuz etkilenmesine neden olur.

Bu nedenle, yüksek korelasyonlu özellik çiftleri dikkatle değerlendirildi. Bazı durumlarda, benzer bilgiyi taşıyan değişkenlerden birinin çıkarılması veya özellik indirgeme tekniklerinin uygulanması önerilir. Ancak, sınıflandırma problemlerinde decision trees ve random forest gibi modeller çoklu bağlantıdan daha az etkilenmektedir.

4. Veri Setinin Eğitim ve Test Olarak Ayrılması

Veri seti, model performansının objektif değerlendirilmesi amacıyla %80 eğitim ve %20 test

olacak şekilde ayrılmıştır. Bu ayrımda, hedef değişkenin sınıf dağılımının eğitim ve test setlerinde benzer şekilde korunabilmesi için stratify yöntemi kullanılmıştır. Stratify işlemi, özellikle dengesiz sınıf dağılımlarında modelin öğrenme sürecinin ve sonuçların tutarlılığı açısından kritik öneme sahiptir.

Eğitim setindeki sınıf dağılımı incelendiğinde, hedef sınıflardan birinin diğerine kıyasla çok daha az gözlem içerdiği görülmüştür. Bu dengesizlik, modelin nadir sınıfı yeterince öğrenememesi ve tahminlerde başarısız olması riskini doğurmaktadır. Bu nedenle, eğitim setinde sınıf dengesini sağlamak üzere SMOTE yöntemi uygulanmıştır. SMOTE, azınlık sınıfına ait sentetik örnekler üreterek sınıf sayısını artırır ve böylece modelin her iki sınıf için de yeterince veriyle eğitilmesini sağlar.

Bu ön işleme adımları sonucunda, eğitim verisi hem sınıf dağılımı açısından dengelenmiş hem de modelin genel performansını artırmaya uygun hale getirilmiştir. Test seti ise orijinal sınıf dağılımını koruyarak, modelin gerçek dünya performansını yansıtacak şekilde bırakılmıştır.

```
Veri Seti Boyutu: (2240, 31)

Eğitim Seti (orijinal): (1792, 30), Sınıf Dağılımı: {0: 1525, 1: 267}
Eğitim Seti (SMOTE sonrası): (3050, 30), Sınıf Dağılımı: {0: 1525, 1: 1525}
Test Seti: (448, 30), Sınıf Dağılımı: {0: 381, 1: 67}
```

5. Özelliklerin Ölçeklendirilmesi

Makine öğrenmesi modellerinin performansını artırmak ve sonuçların tutarlılığını sağlamak amacıyla, sayısal özellikler StandardScaler kullanılarak standartlaştırılmıştır. Standartlaştırma işlemi, her bir özelliğin ortalamasını 0, standart sapmasını ise 1 olacak şekilde dönüştürülmesini sağlar. Bu sayede, farklı ölçeklerde ve birimlerde olan değişkenler karşılaştırılabilir hale gelir.

Özellikle mesafe temelli algoritmalar (kNN, SVM) ve gradyan tabanlı optimizasyon kullanan modeller için özelliklerin aynı ölçeğe getirilmesi önemli bir adımdır. Ölçeklendirme

yapılmadığında, büyük ölçekli değişkenler modelin öğrenme sürecinde daha baskın hale gelebilir ve bu da modelin genel performansını olumsuz etkileyebilir.

Bu çalışmada, eğitim verisi önce SMOTE yöntemiyle dengelendikten sonra, elde edilen sentetik örnekler de dahil olmak üzere StandardScaler ile ölçeklendirilmiştir. Test verisi ise yalnızca eğitim setinde öğrenilen ölçeklendirme parametreleri (ortalama ve standart sapma) kullanılarak dönüştürülmüştür. Böylece veri sızıntısının önüne geçilmiş ve modelin gerçekçi performans ölçümü sağlanmıştır.

6. Model Seçimi, Eğitimi ve Hiperparametre Optimizasyonu

Bu çalışmada, müşteri satın alma davranışını tahmin etmek amacıyla sınıflandırma problemini çözebilecek çeşitli makine öğrenimi modelleri seçilmiş ve performansları karşılaştırılmıştır. Kullanılan modeller arasında Lojistik Regresyon, Random Forest, SVM, kNN, Decision Tree ve Naive Bayes yer almaktadır.

Her model için performansın artırılması ve en uygun parametrelerin belirlenmesi amacıyla Grid Search yöntemi uygulanmıştır. Grid Search ile belirlenen hiperparametre aralıklarında kapsamlı bir tarama yapılmış ve 5 katlı çapraz doğrulama (cross-validation) kullanılarak her parametre kombinasyonunun başarımı değerlendirilmiştir. Bu sayede modellerin aşırı öğrenme yapmadan genellenebilirliği sağlanmıştır.

Örnek olarak, Lojistik Regresyon modelinde C parametresi 0.01’den 10’a kadar farklı değerlerle, ceza terimleri olarak L1 ve L2 normları denenmiştir. Random Forest modelinde ise ağaç sayısı (n_estimators), maksimum derinlik (max_depth) ve dallanma için gerekli minimum örnek sayısı (min_samples_split) parametreleri optimize edilmiştir. Diğer modeller için de benzer şekilde temel hiperparametreler belirlenmiş ve optimize edilmiştir.

Her model, eğitim verisi üzerinde 5 katlı çapraz doğrulama ile eğitilmiş ve F1 skoru optimizasyonu hedeflenmiştir. En iyi hiperparametre kombinasyonları seçilerek, modellerin performansları karşılaştırmaya hazır hale getirilmiştir.

7. Model Performans Değerlendirmesi

Bu çalışmada, eğitim ve hiperparametre optimizasyonu tamamlanan modellerin sınıflandırma başarısı, test veri seti üzerinde çeşitli performans metrikleri kullanılarak kapsamlı şekilde değerlendirilmiştir. Modellerin accuracy, precision, recall, F1 skoru, AUC-ROC skoru, PR AUC ve MCC gibi temel metrikleri hesaplanmıştır. Bu metrikler, modellerin farklı yönlerden sınıflandırma performanslarını değerlendirmede kullanılmıştır. Bu metrikler, modellerin sınıflandırma performansını farklı açılardan ölçmek için kullanılır. Doğruluk, doğru tahminlerin oranını gösterirken, dengesiz verilerde yanıltıcı olabilir. Kesinlik, pozitif tahminlerin doğruluğuna, duyarlılık ise gerçek pozitiflerin ne kadarının yakalandığına bakar. F1 skoru, kesinlik ve duyarlılığın dengeli ortalamasıdır. AUC-ROC, modelin pozitif ve negatifleri ayırt etme başarısını ifade ederken, PR AUC özellikle dengesiz veri setlerinde pozitif sınıf başarısını ölçer. MCC ise pozitif ve negatif tahminlerin dengeli olup olmadığını gösterir ve +1 mükemmel sınıflandırmayı temsil eder.

Model performanslarının genellenebilirliği ve overfitting riski, eğitim verisi üzerinde uygulanan 5 katlı çapraz doğrulama (cross-validation) ile incelenmiş, F1, AUC-ROC ve PR AUC metriklerinin ortalamaları elde edilmiştir. Böylece, modellerin hem test hem de eğitim verisi üzerindeki tutarlılıkları detaylı şekilde analiz edilmiştir.

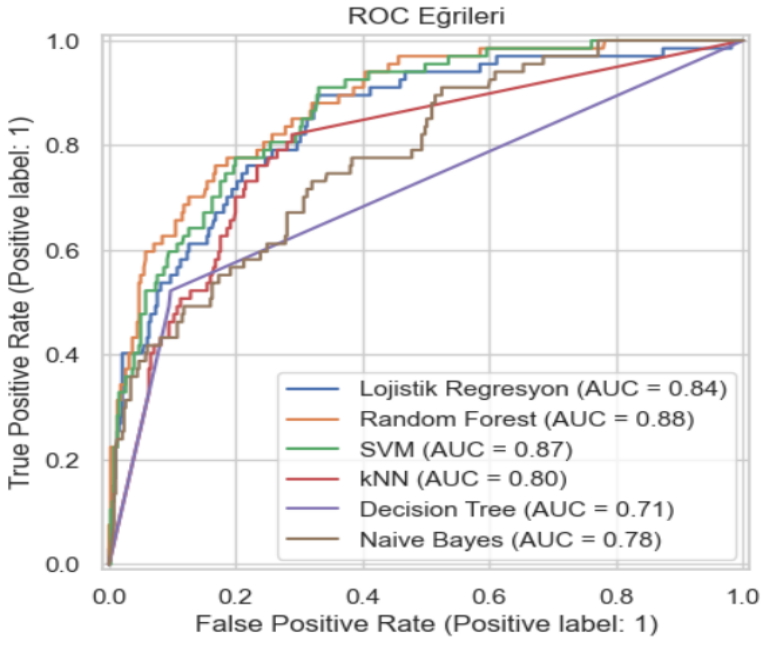
Test veri seti üzerindeki sonuçlar tabloda özetlenmiştir. Tablo, modellerin performansını birden fazla açıdan karşılaştırmakta olup, özellikle F1 skoru dikkate alınarak modellerin dengeli başarıları ortaya konmuştur. Örneğin, Random Forest modeli 0.886 doğruluk, 0.638 kesinlik ve 0.592 F1 skoru ile en yüksek genel performansı göstermiştir. Bu modelin AUC-ROC skoru 0.878, PR AUC değeri ise 0.628 olarak hesaplanmıştır. MCC ise 0.528 olup, bu değer modelin

hem pozitif hem negatif sınıfları dengeli biçimde tahmin ettiğini göstermektedir. Diğer modeller ise genel olarak Random Forest'in ardından SVM ve Lojistik Regresyon modelleri olarak sıralanmıştır.

Model Performansları:

Model	Accuracy	Precision	Recall	F1	AUC-ROC	PR AUC	MCC	CV F1	CV AUC	CV PRC
Random Forest	0.886	0.638	0.552	0.592	0.878	0.628	0.528	0.935	0.984	0.984
SVM	0.877	0.603	0.522	0.560	0.866	0.564	0.491	0.913	0.975	0.976
Lojistik Regresyon	0.842	0.476	0.582	0.523	0.844	0.550	0.433	0.846	0.933	0.934
Decision Tree	0.844	0.478	0.493	0.485	0.712	0.322	0.393	0.854	0.859	0.811
kNN	0.795	0.374	0.552	0.446	0.796	0.386	0.335	0.901	0.946	0.920
Naive Bayes	0.728	0.296	0.597	0.396	0.780	0.450	0.270	0.687	0.789	0.767

Model performanslarının görsel karşılaştırması için Şekil 1'de modellerin ROC eğrileri çizilmiştir. ROC eğrisi, sınıflandırıcının farklı eşik değerlerinde false-positive oranına karşı true-positive oranını gösterir. Eğrilerin altında kalan alan (AUC-ROC) ise modelin ayırt etme gücünü nicel olarak ifade eder. Şekilde Random Forest modeli en yüksek AUC-ROC değerine sahip olup, diğer modellerle kıyaslandığında daha iyi sınıflandırma performansı sergilemiştir. ROC eğrileri, modellerin pozitif ve negatif sınıflar arasındaki ayrım kabiliyetlerini net biçimde görselleştirerek, karar verme aşamasında önemli bilgiler sunmaktadır.



Sonuç olarak, farklı metrikler ve çapraz doğrulama sonuçları dikkate alınarak Random Forest modeli, hem test hem de eğitim verisi üzerinde dengeli ve yüksek performans göstermiştir.

8. Sonuçların Değerlendirilmesi ve En İyi Modelin Analizi

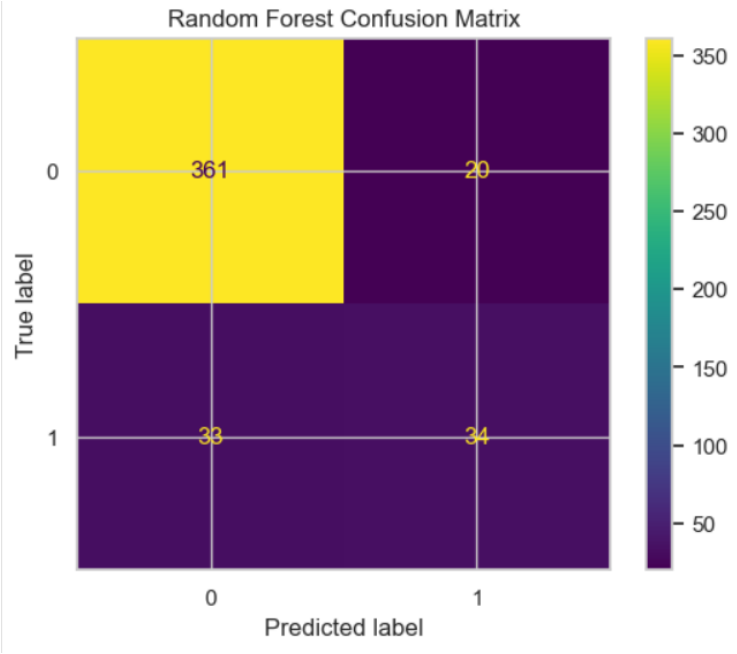
Bu bölümde, modelin performansını daha iyi kavrayabilmek ve karar verme sürecini şeffaflaştırmak için iki temel görsel araç kullanılmıştır: confusion matrix ve feature importance grafiği.

Confusion Matrix:

Karmaşıklık matrisi, modelin gerçek sınıflarla tahmin ettiği sınıflar arasındaki ilişkiyi detaylı olarak sunar. Pozitif ve negatif sınıflar için doğru ve yanlış sınıflandırmalar bu matriste açıkça gösterilir. Bu sayede, modelin özellikle hangi sınıflarda güçlü performans sergilediği, hangi sınıflarda ise hata oranlarının daha yüksek olduğu anlaşılır. Dolayısıyla karmaşıklık matrisi, modelin sadece genel doğruluk oranından daha fazlasını ortaya koyarak, müşteri satın alma davranışını doğru yorumlamaya dayalı uygulama bazlı kararların alınmasına imkan tanır.

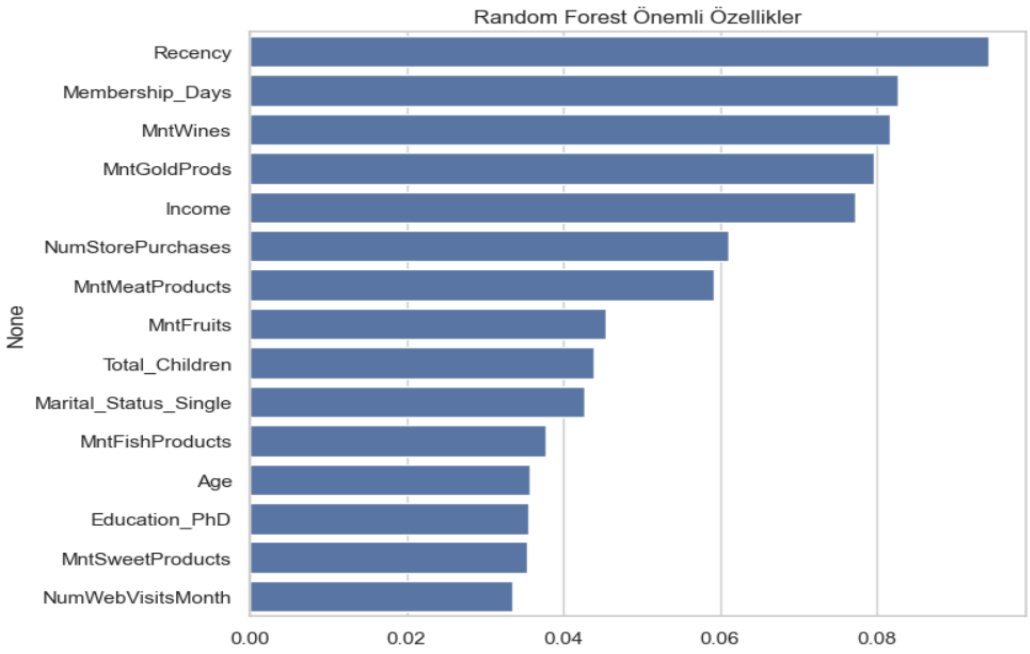
Aşağıdaki confusion matrix görseli, modelin sınıflandırma performansını detaylı şekilde ortaya koymaktadır. Sol üstte yer alan sarı bölge (361), modelin negatif sınıfı doğru şekilde negatif olarak tahmin ettiği örnekleri temsil eder (true negatives). Bu, modelin negatif sınıfları ayırt etmede oldukça başarılı olduğunu göstermektedir. Sol alttaki mor bölge (33), pozitif sınıfın yanlışlıkla negatif olarak tahmin edildiği örneklerdir (false negatives); bu durum, modelin bazı pozitif örnekleri gözden kaçırdığını gösterir. Sağ üstteki mor bölge (20), negatif sınıfın yanlışlıkla pozitif olarak tahmin edildiği örnekleri temsil eder (false positives), yani model bazı negatifleri yanlışlıkla pozitif olarak sınıflandırmıştır. Sağ alttaki mor bölge (34) ise modelin pozitif sınıfı doğru tahmin ettiği örnekleri (true positives) göstermektedir. Bu dağılım, modelin negatif sınıfları yüksek doğrulukla tanıyabildiğini ancak pozitif sınıf

tahminlerinde nispeten daha düşük başarı sergilediğini ortaya koymaktadır.



Feature Importance:

Random Forest algoritmasının yapısı gereği, model karar ağaçlarında hangi değişkenlere ne kadar ağırlık verdiğini ölçmek mümkündür. Bu çalışma kapsamında, modelin tahmin sürecinde en etkili olan ilk 15 özellik görsel olarak sunulmuştur. Bu özellikler, müşteri satın alma davranışlarını tahmin etmede en fazla bilgi taşıyan faktörleri temsil eder. Pazarlama stratejileri açısından, bu veriler hedef kitlenin alışkanlıklarını ve tercihlerini daha iyi anlamak, kaynakları etkin kullanmak ve kampanyaların geri dönüşümünü maksimize etmek için önemli yol göstericiler sağlar. Ayrıca, önemli özelliklerin belirlenmesi, veri setindeki gereksiz veya düşük etkili değişkenlerin elenmesiyle modelin basitleştirilmesine ve hızlandırılmasına da olanak tanır.



Özellik önemi grafiği, modelin karar verme mekanizmasında en kritik rolü oynayan değişkenleri sıralar. Bu sayede, iş analistleri ve pazarlama uzmanları, müşterilerin satın alma davranışlarını ve kampanya tepkilerini şekillendiren temel faktörleri önceliklendirebilir, stratejik planlamalarını daha bilinçli ve etkili yapabilirler.

Çalışmanın sonunda yapılan kapsamlı performans analizleri, Random Forest modelinin test ve eğitim verileri üzerinde dengeli ve yüksek başarı sergilediğini göstermiştir. Karmaşıklık matrisi ile modelin güçlü ve zayıf yönleri ayrıntılı olarak tespit edilmiş, önemli özellikler analizi ise müşteri satın alma davranışlarını etkileyen kritik faktörlerin belirlenmesini sağlamıştır.