

Makine Öğrenmesi Yöntemleriyle Müşteri Terk Tahmini ve Veri Analizi

Bu projenin temel amacı; geçmiş müşteri verilerini analiz ederek, hangi müşterilerin abonelik iptali riski taşıdığını belirleyen bir Makine Öğrenmesi modeli geliştirmektir. Proje kapsamında ham veri işlenmiş, görselleştirilmiş ve anlamlı sonuçlar elde edilmiştir.

Veri Seti ve Özellikleri

Proje kapsamında, veri bilimi literatüründe güvenilirliği kabul görmüş olan ve Kaggle platformunda paylaşılan Telco Customer Churn veri seti kullanılmıştır.

Veri Kaynağı: IBM Kaggle

Veri Büyüklüğü: 7043 Satır, 21 Sütun.

Hedef Değişken (Target): Churn

Özellikler : Veri setinde müşteriye ait demografik bilgiler (Cinsiyet, Yaşlılık durumu), hizmet detayları (İnternet türü, Telefon servisi) ve hesap bilgileri (Kontrat tipi, Ödeme yöntemi, Aylık ücret) bulunmaktadır.

Bu veri setinin seçilme nedeni; hem yazışal hem de nümerik verileri bir arada barındırması ve gerçek hayat problemlerini simülle etmek için ideal bir yapıya sahip olmasıdır.

Veri Ön İşleme Yöntemleri

Ham veri seti, makine öğrenmesi algoritmalarının doğrudan işleyebileceği formatta değildi. Bu nedenle, model başarısını artırmak ve hataları önlemek adına Python programlama dili ve Pandas kütüphanesi kullanılarak aşağıdaki ön işleme adımları uygulanmıştır:

Veri Tipi Düzeltmesi (TotalCharges): Veri seti incelendiğinde, müşterilerin toplam ödemesini gösteren TotalCharges sütununun sayısal olması gereklidir, içeriği bazı boşluk (" ") karakterleri nedeniyle Object olarak algılandığı tespit edilmiştir. Bu sütun pd.to_numeric fonksiyonu ile sayısal formata zorlanmış, sayıya dönüştürülemeyen 11 satır veri setinden temizlenmiştir.

Gereksiz Özniteliklerin Çıkarılması: Her müşteriye özel olan customer ID sütunu, modelin tahminleme yeteneğine matematiksel bir katkısı olmadığı ve ezberlemeye yol açabileceği için veri setinden çıkarılmıştır.

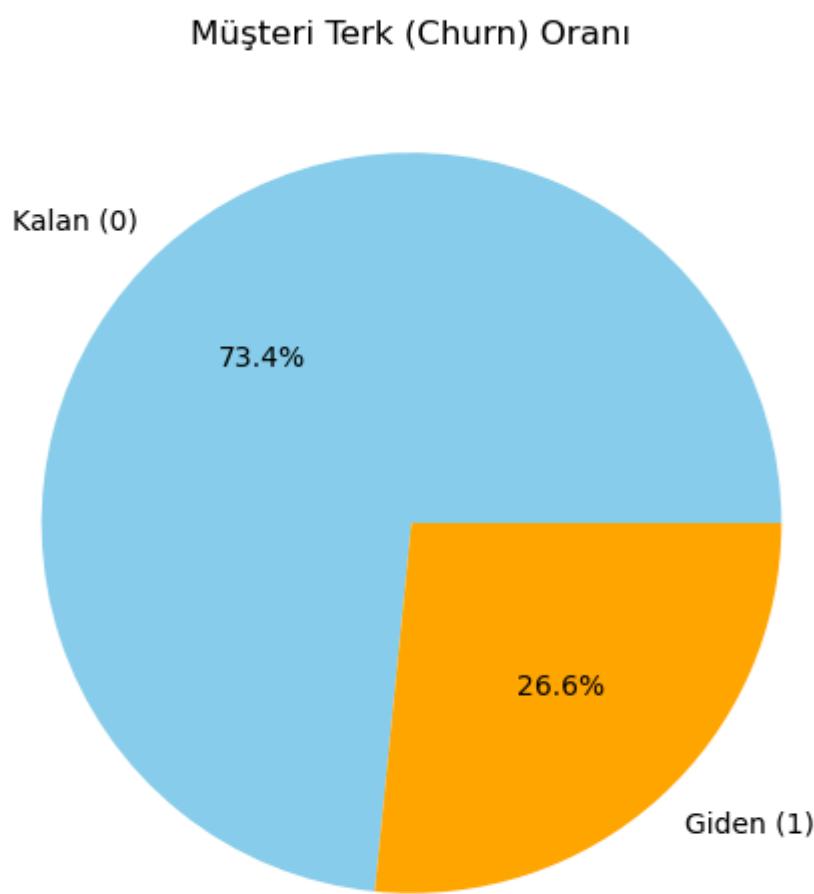
Label Encoding (Sayısal Dönüşüm): Hedef değişkenimiz olan Churn sütunu "Yes" ve "No" değerlerini içermektedir. Bilgisayarın bu durumu işleyebilmesi için bu değerler 1 (Yes - Terk Eden) ve 0 (No - Kalan) olarak dönüştürülmüştür.

Veri Analizi (EDA)

Veri setini daha iyi anlamak ve değişkenler arasındaki ilişkileri çözmek adına görselleştirme çalışmaları yapılmıştır.

Hedef Değişken Dağılımı

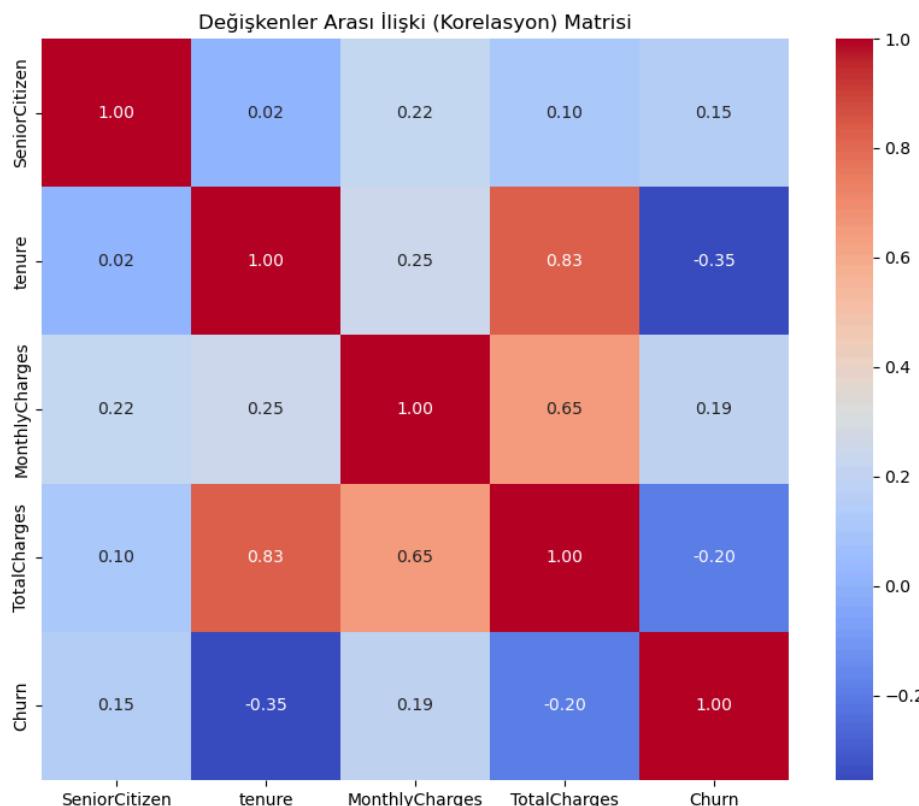
Veri setindeki müşterilerin terk etme oranları incelenmiştir.



Şekil 1'de görüldüğü üzere, veri setindeki müşterilerin yaklaşık **%26.6'sı** hizmeti terk etmiş, **%73.4'ü** ise hizmet almaya devam etmektedir. Bu oran, veri setinde hafif bir dengesizlik olduğunu gösterse de, model kurulumu için kabul edilebilir değerlerdir.

Değişkenler Arası İlişki

Sayısal değişkenlerin birbirleriyle ve hedef değişken (Churn) ile olan ilişkisi incelenmiştir.



Yorum: Korelasyon matrisi incelendiğinde

MonthlyCharges (Aylık Ücret) ile Churn arasında pozitif bir ilişki olduğu görülmüştür. Yani aylık ödeme miktarı arttıkça müşterinin gitme eğilimi artmaktadır.

Tenure (Abonelik Süresi) ile Churn arasında negatif bir ilişki vardır. Müşteri şirkette ne kadar uzun süre kalırsa, terk etme ihtimali o kadar azalmaktadır. Sadık müşterilerin daha az risk taşıdığını doğrulamaktadır.

Veri Ön İşleme

Ham veri setinin makine öğrenmesi algoritmaları tarafından işlenebilmesi ve modelin başarısının artırılması amacıyla aşağıdaki ön işleme adımları uygulanmıştır:

Kategorik Verilerin Dönüşürtlürülmesi

Veri setinde yer alan "Gender", "Partner", "InternetService" gibi metinsel değişkenler, matematiksel modellere uygun hale getirilmek üzere sayısal değerlere dönüştürülmüştür. Bu işlem için Pandas get_dummies fonksiyonu kullanılmıştır. Çoklu bağlantı problemini önlemek adına, oluşturulan ikili değişkenlerden ilki veri setinden çıkarılmıştır

Veri Setinin Bölünmesi

Modelin eğitimi ve performansının objektif olarak değerlendirilebilmesi için veri seti, literatürde yaygın olarak kabul göre 80/20 kuralı doğrultusunda ikiye ayrılmıştır:

- **Eğitim Seti** : Verinin %80'i modelin öğrenmesi için kullanılmıştır.
- **Test Seti** : Verinin %20'si modelin daha önce hiç görmediği veriler üzerindeki tahmin başarısını ölçmek için ayrılmıştır.

Bu oran modelin veri setindeki genel kalıpları kavraması ile test sonucunun istatistiksel güvenilirliği arasındaki dengeyi sağlamak amacıyla tercih edilmiştir.

Özellik Ölçeklendirme

Veri setindeki sayısal değişkenlerden Aylık Ücret (MonthlyCharges) ve Abonelik Süresi (Tenure) gibi farklı büyüklükteki değerlerin modeli yaniltmasını engellemek amacıyla StandardScaler yöntemi kullanılmıştır. Bu işlemle tüm sayısal veriler ortalaması 0 ve standart sapması 1 olacak şekilde standartlaştırılmıştır.

Model Kurulumu ve Yöntem

Bu çalışmada, sınıflandırma problemi için Random Forest algoritması tercih edilmiştir.

Algoritma Seçimi: Random Forest

Random Forest, birden fazla karar ağacının bir araya gelerek oluşturduğu bir topluluk öğrenme yöntemidir. Tek bir karar ağacının veriyi ezberleme (overfitting) riskine karşı, Random Forest yüzlerce ağacın sonucunun ortalamasını alarak daha kararlı, genellenebilir ve yüksek doğruluklu sonuçlar üretir.

Model Parametreleri

Model parametresi olan karar ağaçları sayısını (`n_estimators`) belirlenirken, Scikit-learn kütüphanesinin önerdiği güncel varsayılan değerler ve işlem maliyeti/performans dengesi dikkate alınmıştır.

Kütüphane Standartları: Scikit-learn kütüphanesi, versiyon 0.22 güncellemesiyle birlikte, model kararlılığını (stability) artırmak amacıyla varsayılan ağaç sayısını 100 olarak belirlemiştir. Projede bu endüstri standartına sadık kalınmıştır.

Deneysel Gözlem ve Maliyet: Ağaç sayısının gereksiz yere artırılması (Örn: 500 veya 1000), modelin eğitim süresini uzatırken, başarı oranında marginal bir katkı sağlamamaktadır. Bu nedenle, donanım kaynaklarını verimli kullanmak ve modelin aşırı karmaşıklaşmasını önlemek adına 100 ağaçlık yapı tercih edilmiştir.

Performans Değerlendirme

Modelin tahmin başarısını ölçmek ve değerlendirmek amacıyla temel olarak Doğruluk puanı kullanılmıştır.

- **Doğruluk:** Modelin doğru tahmin ettiği (hem Churn eden hem etmeyen) müşteri sayısının toplam müşteri sayısına oranıdır. Genel performansı görmek adına birincil ölçüt olarak seçilmiştir.

Ancak, dengesiz veri setlerinde tek başına doğruluk oranı yaniltıcı olabileceğinden, modelin başarısı Hata matrisi ile de desteklenmiştir. Hata matrisi sayesinde modelin sadece genel başarısı değil; kaç müşteriye yanlışlıkla gidecek dediği veya kaç giden müşteriyi kalacak sanarak kaçındığı analiz edilmiştir.

Sonuçlar

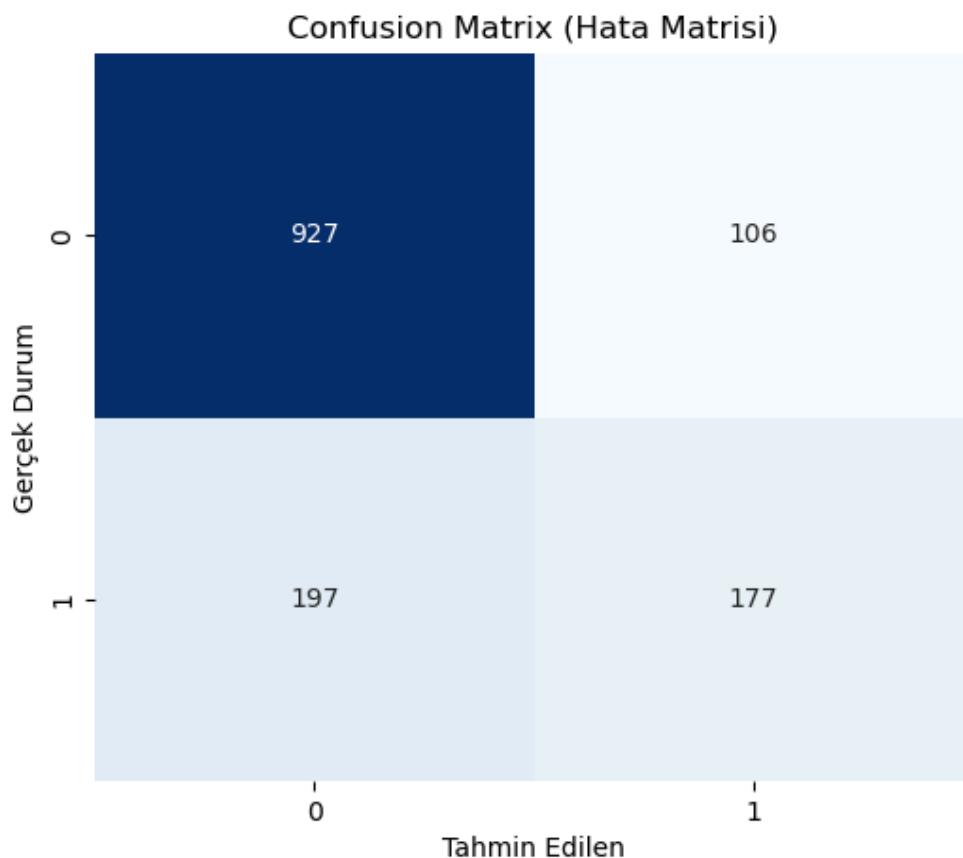
Geliştirilen model ayrılan test veri seti üzerinde değerlendirilmiş ve aşağıdaki sonuçlar elde edilmiştir.

Model Başarısı

Modelin genel doğruluk oranı %78.46 olarak ölçülüştür. Bu oran, modelin müşterilerin davranışlarını yaklaşık %80 güvenilirlikle doğru tahmin ettiğini göstermektedir ve telekomünikasyon sektörü için kabul edilebilir bir başarı seviyesidir.

Hata Matrisi

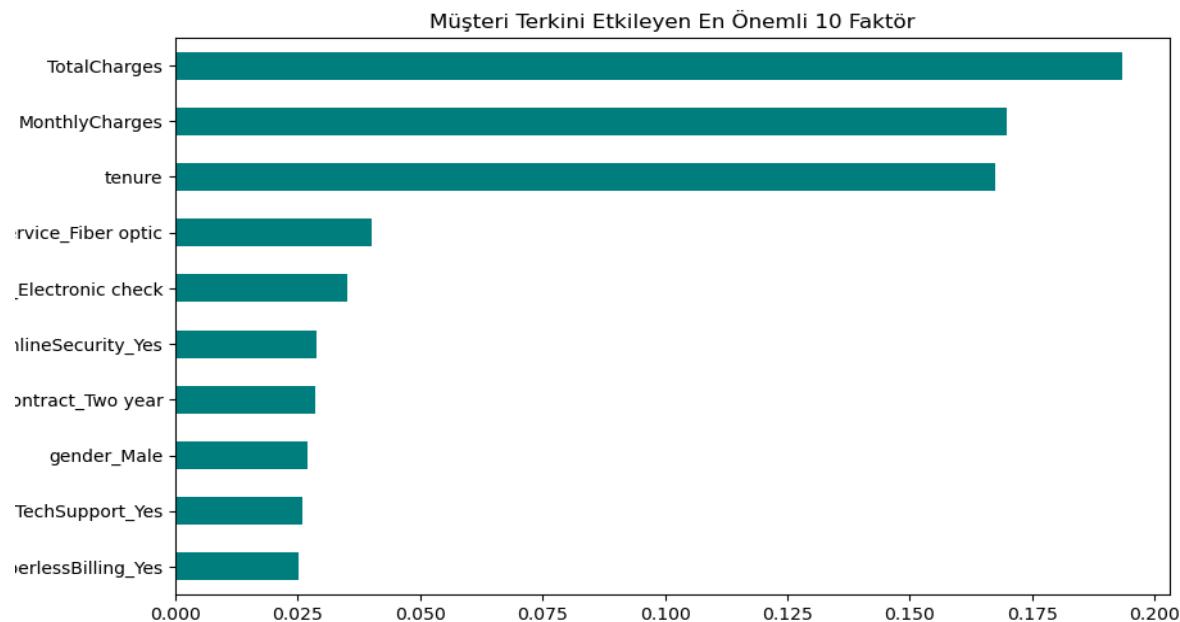
Modelin tahminleri ile gerçek değerlerin karşılaştırması aşağıdaki hata matrisinde gösterilmiştir.



Matris incelendiğinde, modelin hizmet almaya devam eden müşterileri tespit etmede yüksek başarı gösterdiği, ancak terk eden müşterileri yakalamada hataya daha açık olduğu görülmektedir.

Önem Düzeyleri

Random Forest algoritmasının sağladığı nitelik önem analizi sonucunda müşteri kaybını etkileyen en kritik faktörler belirlenmiştir.



Grafikte görüldüğü üzere;

- Finansal Etkenler:** TotalCharges ve MonthlyCharges değişkenlerinin en üst sıralarda yer alması, müşteri kaybında en belirleyici faktörün fiyatlandırma olduğunu göstermektedir.
- Sadakat Süresi:** Tenure (Abonelik Süresi) değişkeninin yüksek öneme sahip olması, yeni müşterilerin risk grubunda olduğunu, hizmet süresi uzadıkça müşterinin sadıklaştığını kanıtlamaktadır.
- Kontrat Tipi:** Aylık sözleşme yapan müşterilerin ayrılma eğilimi, taahhütlü müşterilere göre daha belirleyicidir.

kaynakça

Scikit-learn Developers. *RandomForestClassifier Documentation*. Retrieved from <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>

Veri Bilimi Okulu. *Makine Öğrenmesi ve Model Doğrulama Yöntemleri*.

Kaggle. *Telco Customer Churn Dataset*. Retrieved from <https://www.kaggle.com/bblastchar/telco-customer-churn>

Scikit-learn Developers. *Model Evaluation: The Scoring Parameter*. Scikit-learn 1.3 Documentation. Retrieved from https://scikit-learn.org/stable/modules/model_evaluation.html#accuracy-score

Google Developers. . *Machine Learning Crash Course: Classification - Accuracy*. Retrieved from <https://developers.google.com/machine-learning/crash-course/classification/accuracy>

<https://github.com/erdemkrhn/musteri-terk-analizi>

Mehmet erdem karahan 23430070053