İlerleme Raporu

Web Uygulamaları için Kullanıcı Ayrılma (Churn) Tahmini ve Önleme Sistemi

Hazırlayan: Hasan Burak Songur
Ders Yürütücüsü: Gözde Çay
17.11.2025

1. ÖZET

Bu projenin amacı, KKBox müzik servisinin kullanıcı verilerini analiz ederek, müşterilerin aboneliklerini iptal etme (churn) olasılığını tahmin eden bir makine öğrenmesi sistemi geliştirmektir. Proje kapsamında, Kaggle'dan temin edilen ve kullanıcıların demografik bilgilerini, işlem geçmişlerini ve dinleme alışkanlıklarını içeren dört ayrı veri seti kullanılmıştır. Kapsamlı veri ön işleme, özellik mühendisliği ve modelleme aşamaları sonucunda, Lojistik Regresyon ve XGBoost gibi farklı makine öğrenmesi modelleri karşılaştırılmıştır. Şu anki en iyi modelimiz olan XGBoost, churn tahmininde ~0.58 F1-Skoru elde ederek, temel modele göre önemli bir başarı sağlamıştır.

2. Giriş ve Arka Plan

Problem Tanımı:
Abone tabanlı dijital servislerde müşteri kaybı (churn), şirket gelirleri üzerinde doğrudan olumsuz bir etkiye sahiptir. Risk altındaki müşterileri proaktif olarak tespit etmek, onlara özel kampanyalar veya destek sunarak elde tutma (retention) stratejileri geliştirmek için kritik öneme sahiptir. Bu proje, KKBox veri setini kullanarak bu tespit işlemini otomatize etmeyi hedeflemektedir.

Problem için Yapay Zeka Çözümü:
Proje, farklı karmaşıklıktaki modelleri karşılaştırarak en iyi tahmin performansını bulmayı amaçlamaktadır. Bu doğrultuda, yorumlanabilirliği yüksek bir temel model olarak Lojistik Regresyon ve daha yüksek tahmin gücü için ağaç tabanlı bir model olan XGBoost kullanılmıştır.

3. Yapay Zeka Modelleme

Projede Kullanılan Veri Seti:
Kaggle "KKBox Churn Prediction Challenge" veri seti kullanılmıştır. Bu set, kullanıcıların temel bilgilerini (members_v3.csv), işlem geçmişlerini (transactions.csv), günlük dinleme loglarını (user_logs.csv) ve hedef değişkeni (is_churn) içeren train.csv dosyasından oluşmaktadır.

Veri Ön İşleme ve Özellik Mühendisliği:
Ham veri, modellemeye uygun değildi ve çeşitli sorunlar içeriyordu. Bu nedenle veri birleştirme, eksik değer doldurma, aykırı değer temizleme ve kategorik veri dönüşümü gibi standart adımlar uygulanmıştır. En önemli adım, 'transactions' ve 'user_logs' gibi işlem kayıtlarından, her kullanıcı için 'sum', 'mean', 'count', 'min', 'max' gibi istatistiksel özetler çıkarılarak yeni özellikler türetilmesi olmuştur.

Kullanılan Yapay Zeka Modelleri:

- Kullanılan Modeller:
    1. Lojistik Regresyon: Yorumlanabilirliği yüksek, hızlı ve basit bir temel model (baseline) oluşturmak için seçildi.
    2. XGBoost (Extreme Gradient Boosting): Doğrusal olmayan ilişkileri yakalayabilen, yüksek performanslı ve endüstride sıkça kullanılan bir model olduğu için seçildi.

- Kullanılan Kütüphaneler: pandas, numpy, scikit-learn, xgboost, imbalanced-learn, matplotlib.

4. İLERLEME

Proje planına uygun olarak aşağıdaki ana adımlar tamamlanmıştır:

1. **Veri Analizi ve Birleştirme:** Tüm veri setleri incelendi, temizlendi ve tek bir ana veri setinde birleştirildi.
2. **Kapsamlı Özellik Mühendisliği:** Özellikle 'user_logs' verisinden, kullanıcı davranışlarını daha iyi temsil eden ('ortalama dinleme süresi', 'aktif gün sayısı' gibi) 35 yeni özellik türetildi.
3. **Lojistik Regresyon Modellemesi:** Temel model olarak eğitildi ve ~0.42 F1-Skoru elde edildi.
4. **XGBoost Modellemesi:** Lojistik Regresyon'a göre belirgin bir üstünlük sağlayarak, F1-Skorunu ~0.58'e yükseltti. Eşik ayarlaması (threshold tuning) ile en iyi F1-Skorunun 0.8 eşik değerinde elde edildiği bulundu. Özellik önem analizi, 'is_auto_renew_max' ve 'is_cancel_sum' gibi işlem özelliklerinin en kritik faktörler olduğunu doğruladı.

Karşılaşılan Problemler ve Çözümler:

- Problem: Basit özellik mühendisliği ile eğitilen Lojistik Regresyon modelinin performansının artmaması.

- Analiz ve Çözüm: İlk olarak, 'user_logs' verisinden sadece toplam değerler ('sum') alınarak özellikler türetilmiş, ancak bu durum model performansını iyileştirmemiştir. Bunun üzerine, Lojistik Regresyon'un doğrusal yapısının, karmaşık kullanıcı davranışlarını yakalamada yetersiz kalabileceği hipotezi kurulmuştur. Bu hipotezi test etmek ve daha güçlü özellikler yaratmak için, 'user_logs' verisi 'mean', 'count', 'min', 'max' gibi daha zengin istatistiklerle yeniden işlenmiştir. Bu yeni özellikler, Lojistik Regresyon'da yine büyük bir artış sağlamasa da, XGBoost modelinin F1-Skorunu ~0.42'den ~0.58'e çıkarmasını sağlamıştır. Bu süreç, doğru modelin doğru özelliklerle birleştirilmesinin önemini ve problemin daha derinine inmenin gerekliliğini ortaya koymuştur.

- Problem: 'user_logs.csv' dosyasının büyüklüğü.

- Çözüm: Veriyi 'chunksize' ile parça parça işlemek ve sadece eğitim setinde bulunan kullanıcıların loglarını işleyerek işlem süresini optimize etmek.

5. SONRAKİ AŞAMALAR

Projenin kalan kısmında, model performansını daha da artırmak ve sistemi tamamlamak için çok yönlü bir yaklaşım izlenecektir:

1. **Model Çeşitliliğini Artırma (Hafta 5):** Ağaç tabanlı XGBoost modeline alternatif olarak, özellikler arasındaki farklı türde ilişkileri yakalama potansiyeli olan bir MLP (Multi-Layer Perceptron) derin öğrenme modeli de denenecektir. Bu, en iyi model mimarisini bulma konusunda kapsamlı bir karşılaştırma sağlayacaktır.
2. **XGBoost Optimizasyonuna Devam (Hafta 5-6):** Mevcut en iyi modelimiz olan XGBoost'un performansını daha da artırmak için, daha geniş bir parametre aralığında ve daha fazla iterasyonla (`GridSearchCV` veya daha kapsamlı bir `RandomizedSearchCV`) hiperparametre optimizasyonu tekrarlanacaktır. Ayrıca, özellik önem analizinden yola çıkarak yeni etkileşim özellikleri (interaction features) oluşturulması denenecektir.
3. **API ve Arayüz Geliştirme (Hafta 6-8):** En iyi XGBoost modeli ve yorumlama için Lojistik Regresyon modeli, FastAPI ile bir API servisi haline getirilecektir. React ile geliştirilecek yönetici paneli, bu API'yi kullanarak bir kullanıcının churn skorunu ve churn'e etki eden faktörleri gösterecektir.
4. **Entegrasyon, Test ve Raporlama (Hafta 9-10):** Uçtan uca sistem testleri yapılacak, performans optimizasyonları gerçekleştirilecek ve final proje raporu ile sunum hazırlanacaktır.