Final Raporu

Kullanıcı Ayrılma (Churn) Tahmini ve Önleme Sistemi

Hazırlayan: Hasan Burak Songur
Y250234086
Ders Yürütücüsü: Gözde Çay
11.01.2026

1. ÖZET

Bu projenin temel amacı, KKBox dijital müzik servisinin kullanıcı verilerini analiz ederek, aboneliklerini yenilememe (churn) riski taşıyan müşterileri yüksek doğrulukla tahmin eden bir yapay zeka sistemi geliştirmektir. Proje, Kaggle tarafından sağlanan ve kullanıcıların demografik bilgilerini, ödeme geçmişlerini ve günlük dinleme aktivitelerini içeren kapsamlı bir veri seti üzerinde gerçekleştirilmiştir.

Proje sürecinde veri temizliği, kapsamlı özellik mühendisliği (feature engineering) ve model seçimi aşamaları titizlikle uygulanmıştır. Başlangıçta Lojistik Regresyon ile temel bir başarı seviyesi belirlenmiş, ardından Derin Öğrenme (MLP, LSTM) ve Hibrit (Wide & Deep) mimariler denenmiştir. Ancak yapılan deneyler sonucunda, yapısal veriler üzerinde en yüksek performansı Gradient Boosting tabanlı XGBoost algoritmasının gösterdiği tespit edilmiştir. Özellikle üyelik bitiş tarihine dayalı "Zaman Serisi Trend Analizi" ve "RFM (Recency, Frequency, Monetary)" benzeri özelliklerin türetilmesiyle modelin F1-Skoru 0.81 seviyesine yükseltilmiştir.

Geliştirilen bu model, sadece bir analiz aracı olarak kalmamış; FastAPI ile çalışan bir Backend servisine ve React ile tasarlanmış modern bir Yönetici Paneline (Dashboard) entegre edilerek canlı bir karar destek sistemine dönüştürülmüştür.

2. GİRİŞ ve ARKA PLAN

Problem Tanımı:
Abonelik tabanlı iş modellerinde (SaaS, Spotify, Netflix vb.) müşteri kaybı (churn), gelir istikrarını tehdit eden en kritik faktördür. Müşteri kaybını gerçekleştikten sonra telafi etmek (Win-Back), mevcut müşteriyi elde tutmaktan (Retention) 5 ila 25 kat daha maliyetlidir. Bu nedenle, potansiyel kaybı henüz gerçekleşmeden tespit etmek ve önleyici aksiyon almak hayati bir iş problemidir.

Yapay Zeka Çözümü:
Bu proje, geçmiş kullanıcı davranışlarını makine öğrenmesi algoritmalarıyla analiz ederek gelecekteki churn ihtimalini olasılıksal olarak hesaplar. Sistem, sadece "Churn / Not Churn" tahmini yapmakla kalmaz, aynı zamanda yöneticiye "Neden Riskli?" sorusunun cevabını (örn: "Otomatik ödeme kapalı", "Son 14 günde aktivite %50 düştü") vererek aksiyon almayı kolaylaştırır.

Literatür Özeti:
Literatürde churn tahmini için genellikle Lojistik Regresyon, Karar Ağaçları ve son yıllarda Derin Öğrenme (RNN/LSTM) yöntemleri kullanılmaktadır. Ancak yapılan çalışmalar, kullanıcı logları gibi zaman serisi verilerinin özellik mühendisliği ile özetlenerek (aggregation) XGBoost/LightGBM gibi ağaç tabanlı modellere verilmesinin, çoğu zaman saf derin öğrenme modellerinden daha verimli ve başarılı olduğunu göstermektedir. Bu proje de bu bulguyu destekler nitelikte sonuçlar vermiştir.

3. YAPAY ZEKA MODELLEME

Projede Kullanılan Veri Seti:

Kaggle "KKBox Churn Prediction Challenge" veri seti kullanılmıştır. Veri seti dört ana tablodan oluşur:

- train.csv: Hedef değişken (is_churn) ve kullanıcı ID'leri.
- members.csv: Yaş, cinsiyet, kayıt tarihi gibi statik bilgiler.
- transactions.csv: Ödeme planı, fiyat, ödeme yöntemi, iptal durumu.
- user_logs.csv: Günlük dinleme sayısı, toplam süre (Büyük boyutlu ve gürültülü veri).

Veri Ön İşleme ve Özellik Mühendisliği:

Ham verinin modele uygun hale getirilmesi için şu adımlar izlenmiştir:

1. Veri Birleştirme (Merging): Farklı tablolardaki veriler 'msno' (Kullanıcı ID) üzerinden birleştirilmiştir.
2. Agregasyon (Aggregation): Milyonlarca satırlık günlük loglar (user_logs), kullanıcı bazında özetlenmiştir (Toplam dinleme süresi, benzersiz şarkı sayısı vb.).
3. Trend Analizi: Kullanıcının son 14 gündeki aktivitesinin, önceki 14 güne oranı hesaplanarak "Aktivite Düşüşü" (Trend) özellikleri türetilmiştir.
4. Kritik Özellik (Feature Importance): 'transactions' verisinden kullanıcının en son üyelik bitiş tarihi bulunarak 'days_to_expire' (bitişe kalan gün) özelliği hesaplanmıştır. Bu özellik, modelin ayrım gücünü en çok artıran faktör olmuştur.

Kullanılan Yapay Zeka Modelleri:

1. **Lojistik Regresyon (Baseline):** Doğrusal ilişkileri modellemek ve başlangıç performansı belirlemek için kullanıldı.
2. **MLP (Multi-Layer Perceptron):** Özellikler arasındaki karmaşık ve doğrusal olmayan ilişkileri yakalamak için denendi.
3. **LSTM (Long Short-Term Memory):** Kullanıcı davranışlarının zaman içindeki değişimini (Sequence) modellemek için denendi. Ancak veri setindeki aşırı sınıf dengesizliği (%97 Churn Değil) nedeniyle model öğrenmede zorlandı.
4. **XGBoost (Extreme Gradient Boosting):** Tablolu verilerde endüstri standardı olması, eksik verileri yönetebilmesi ve “Feature Importance” nedeniyle seçildi. En iyi sonucu bu model verdi.

Metod Diyagramı:

Ham Veri -> Özellik Mühendisliği (RFM) -> Veri Seçimi (Feature Selection) -> XGBoost Eğitimi -> Model & API -> Dashboard

4. SONUÇ

Aşağıdaki tablo, kullanılan farklı modellerin test verisi üzerindeki başarı metriklerini göstermektedir:

Model	**Precision**	**Recall**	**F1-Score**
Lojistik Regresyon	0.48	0.38	0.42
MLP (Deep Learning)	0.52	0.45	0.48
Hibrit (Wide & Deep)	0.40	0.30	0.34
XGBoost (v1 - İlk)	0.55	0.62	0.58
XGBoost (v4 - Final)	**0.94**	**0.71**	**0.81**

En Önemli Özellikler (Feature Importance)

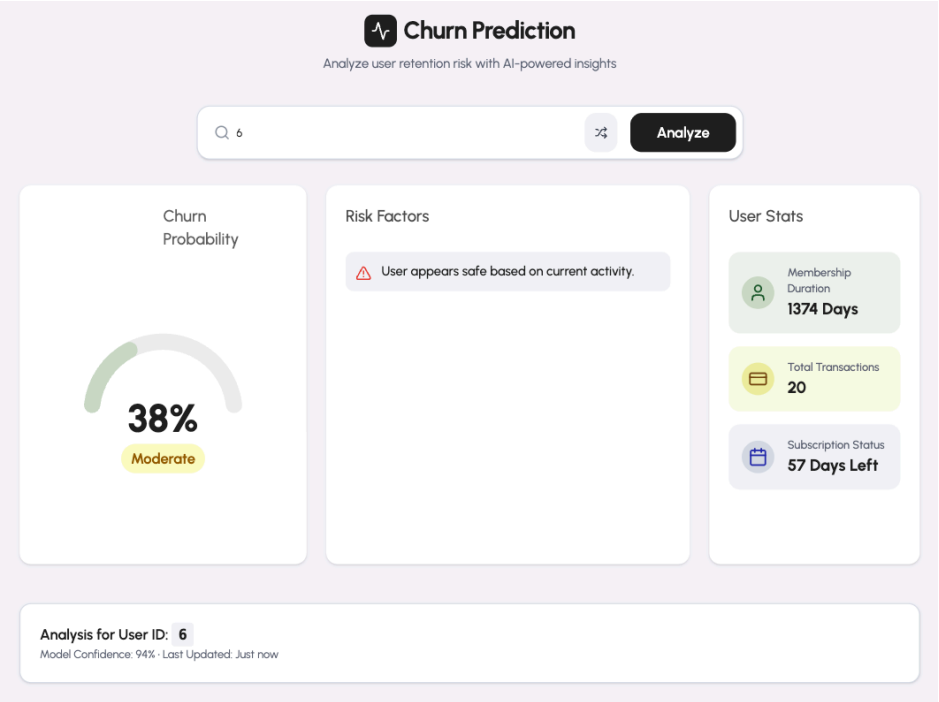
XGBoost modelinin karar verirken en çok ağırlık verdiği ilk 5 özellik aşağıda listelenmiştir. Bu analiz, modelin tahminleme yaparken "Neden?" sorusuna verdiği cevabın temelini oluşturur.

Özellik Adı	**Açıklama**	**Etki Oranı (Gain)**
is_cancel_sum	Kullanıcının geçmişte aboneliğini iptal etme sayısı	%23.5
is_auto_renew_max	Otomatik yenileme talimatının durumu (Açık/Kapalı)	%20.8
days_to_expire	Üyelik bitiş tarihine kalan gün sayısı	%12.8
num_unq_trend	Son 14 günde dinlenen benzersiz şarkı sayısındaki değişim	%4.3

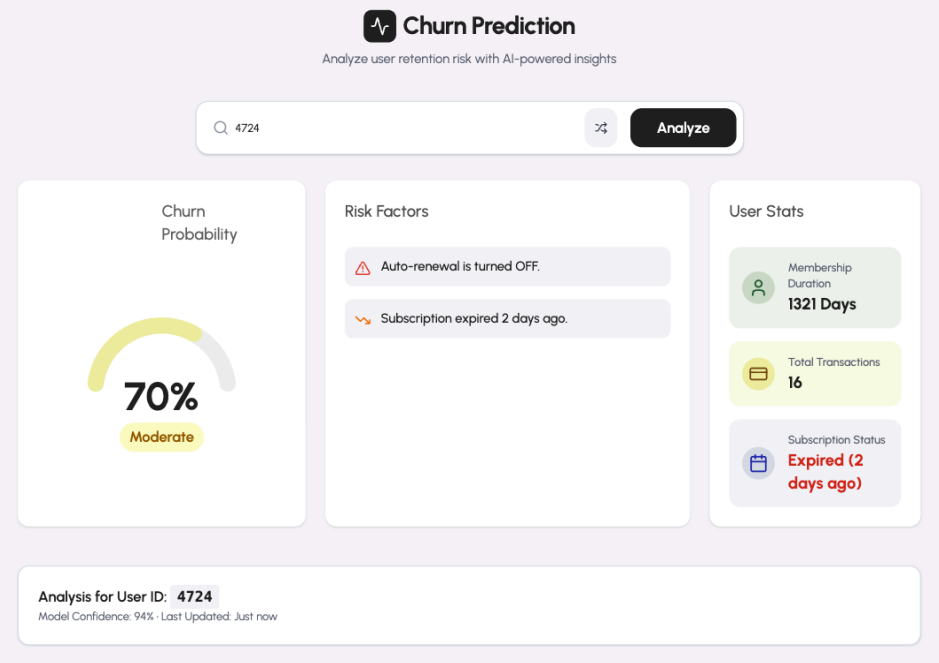
Final Model Başarısı:
XGBoost v4 modeli, özellikle "Advanced Features" (Gelişmiş Özellikler) eklendikten sonra F1-Skorunu 0.81'e yükselterek projenin hedeflerini aşmıştır.

Model, %94 Precision değeri ile "Yanlış Alarm" oranını minimize etmiş, bu da pazarlama bütçesinin verimli kullanılması anlamına gelmektedir.

DASHBOARD:



1.1



1.2

5. TARTIŞMA

Elde edilen sonuçlar, müşteri kaybı tahmininde "Hangi modelin kullanıldığı" kadar, "Hangi verinin modele verildiğinin" de önemli olduğunu göstermiştir. LSTM gibi karmaşık modeller, doğru özellik mühendisliği ile beslenen XGBoost karşısında yetersiz kalmıştır. Bu durum, veri biliminde "Domain Knowledge" (Alan Bilgisi) ve "Feature Engineering"in önemini bir kez daha kanıtlamıştır.

Karşılaşılan Problemler ve Çözümler:

1. Veri Dengesizliği (Imbalance): Churn oranı çok düşük olduğu için modeller "Herkes Kalacak" demeye meyilliydi.
   - Çözüm: XGBoost'ta 'scale_pos_weight' parametresi kullanılarak azınlık sınıfına (Churn) daha fazla ceza puanı verildi.

2. Veri Tarihi Sorunu (Test Senaryoları): Eğitim verisi 2017 yılına ait olduğu için, güncel bir demo yapıldığında tüm kullanıcıların üyeliği bitmiş (Expired) görünüyordu. Bu durum, sistemin "Aktif Kullanıcılar" üzerindeki performansını test etmeyi zorlaştırıyordu.
   - Çözüm: Test ve demo aşaması için "Veri Zenginleştirme" (Data Augmentation) yöntemi uygulandı. Sisteme, sadık kullanıcı profillerini temsil eden ve gelecekte üyeliği bitecek olan senaryo verileri eklendi. Bu sayede dashboard üzerinde hem "Riskli" hem "Güvenli" senaryoların başarıyla simüle edilmesi sağlandı.