

Youtube Trending Video Prediction

Grup adı: Trend Hunters

22040301135 – Muhammed Chreiki

22040301003 – Emad Alkasabli

22040301092 – Abdelrahman Abdelhalim

22040301145 – Omar mokhtar abdo Boghdady

Youtube Link: <https://youtu.be/bRsEUWgdVLY>

Problem Tanımı

YouTube platformunda her gün milyonlarca video yayınlanmaktadır. Bu videoların çok küçük bir kısmı «Trending» listesine girebilmekte ve bu listeye giren videolar, hem içerik üreticileri hem de platform algoritmaları açısından yüksek görünürlük ve etkileşim kazanmaktadır. Ancak bir videonun trend olup olmayacağı önceden tahmin etmek, birçok değişkenin bir arada değerlendirilmesini gerektiren karmaşık bir problemdir.



Veriseti Tanımı

```
Veri Seti Boyutu (satır, sütun)  
(390043, 15)

Sütun İsimleri
- video_id
- title
- publishedAt
- channelId
- channelTitle
- categoryId
- trending_date
- tags
- view_count
- likes
- comment_count
- thumbnail_link
- comments_disabled
- description
- is_trending
```

Bu projede kullanılan veri seti, YouTube platformunda 2022-2025 yılları arasında yayınlanmış videolara ait meta verileri içermektedir. Veri seti, bir video'nun trend olup olmama durumunu (`is_trending`) tahmin etmeye yönelik bir sınıflandırma problemi için hazırlanmıştır.

Bu veri seti **sınıf dengesizliği (class imbalance)** içermektedir; çünkü trend olan videolar, trend olmayan videolara kıyasla sayıca oldukça azdır. Bu durum, modellerin çoğunluk sınıfına (trend olmayan videolar) eğilim göstermesine neden olabileceğinden, değerlendirme metrikleri ve modelleme aşamasında özel yöntemler kullanılmasını gerektirir.

Veri Seti Boyutu

Toplam Satır Sayısı: 390.043

Toplam Sütun Sayısı: 15

Veri Seti Veri Tipleri

Numerical Feature Sayısı: 6

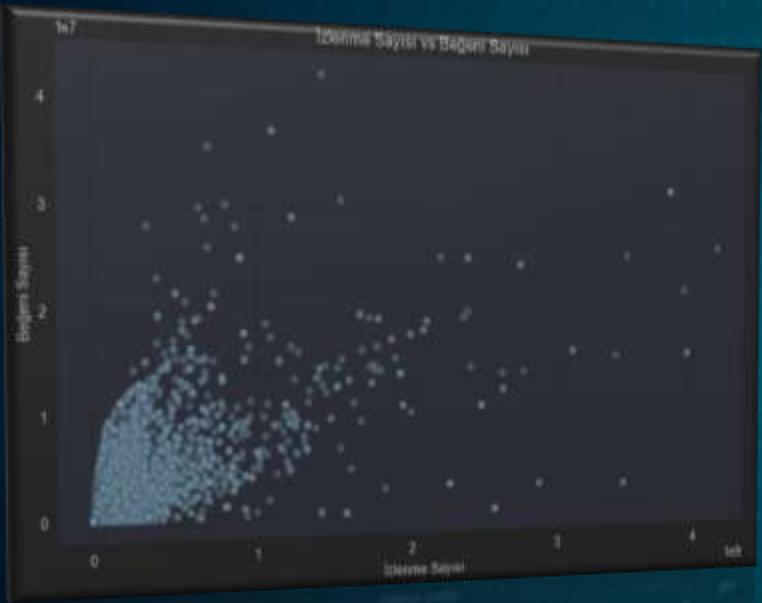
Categorical Feature Sayısı: 9

Veriseti linki: [YouTube Video Trends & Non-Trends Dataset](#)



>>>>

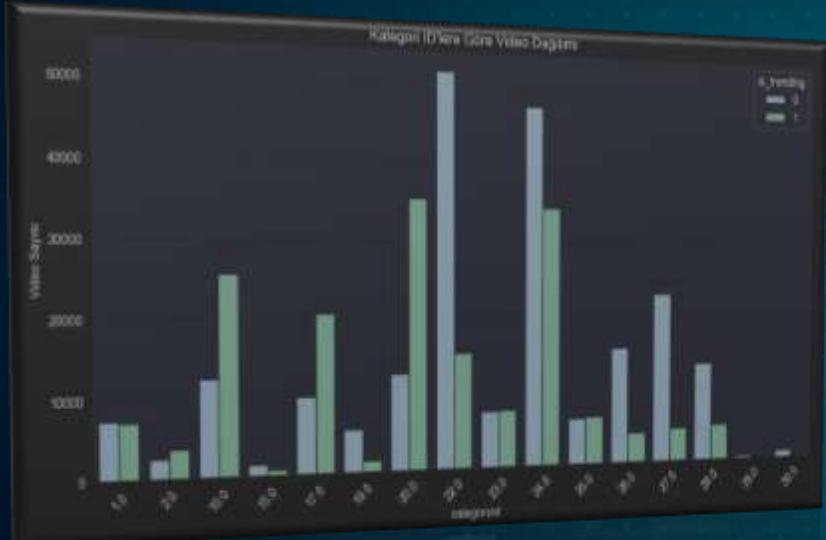
EDA (Data Analysis)



Bu scatter plot, izlenme sayısı ile beğeni sayısı arasındaki ilişkiyi göstermektedir. Grafikte yoğunluğun düşük izlenme ve düşük beğeni bölgesinde toplandığı, ancak bazı videoların **çok yüksek izlenme ve beğeni değerlerine sahip aykırı noktalar** (outlier) oluşturduğu görülmektedir.

Bu durum, etkileşim metriklerinin **son derece çarpık** (skewed) bir dağılıma sahip olduğunu göstermektedir. Ayrıca yüksek izlenme sayısının her zaman orantılı bir şekilde beğeniye dönüşmediği de gözlemlenmektedir. Bu nedenle izlenme ve beğeni gibi metrikler, modelde doğrudan belirleyici olmaktan ziyade **normalize edilmiş veya oran bazlı** (engagement) özellikler olarak kullanılmalıdır.

<<<<



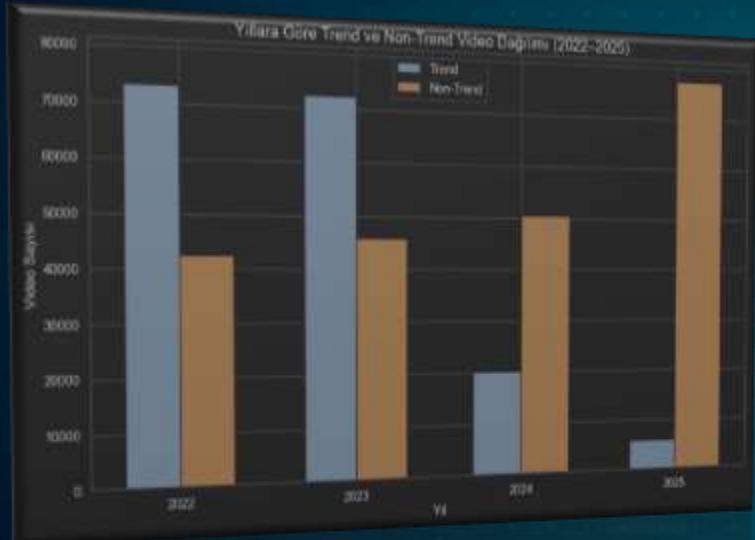
EDA (Data Analysis)

Bu grafik, videoların kategori bazında trend olup olmama durumlarını göstermektedir. Bazı kategorilerde (örneğin Music, Entertainment, Gaming gibi) trend olan video sayısının diğer kategorilere kıyasla daha yüksek olduğu açıkça görülmektedir. Buna karşılık bazı kategorilerde trend videolar o dukça sınırlıdır.

Bu dağılım, kategori bilgisinin trend olma olasılığı üzerinde güçlü bir belirleyici olduğunu göstermektedir. Dolayısıyla categoryld değişkeni modelde yüksek ağırlıkla ele alınmalı ve kategoriler arası farklılıklar öğrenilebilir şekilde temsil edilmelidir (örneğin embedding veya hedefe yönelik encoding).



EDA (Data Analysis)



Bu grafik, yıllar bazında trend olan ve olmayan videoların dağılımını göstermektedir. 2022 ve 2023 yıllarında trend olan video sayısı görece yüksekken, 2024 ve özellikle 2025 yılında trend video sayısında belirgin bir düşüş gözlemlenmektedir. Buna karşın non-trend video sayısı yıllar ilerledikçe artış göstermektedir.

Bu durum, veri setinin zamansal olarak dengesiz olduğunu ve özellikle son yıllarda trend videoların daha nadir hale geldiğini göstermektedir. Modelleme sürecinde zaman faktörünün ve sınıf dengesizliğinin dikkate alınması gerektiğini ortaya koymaktadır.



Feature Engineering & Feature Extraction

Metin Özellikleri (Text Features):

Başlık (Title) ve Etiketler (Tags) tek bir metin halinde birleştirildi.

Bu metin, TF-IDF teknigi kullanilarak sayisal verilere dönüştürüldü (en sık geçen 5000 kelime kullanıldı).

Zaman Özellikleri (Time Features):

Yayınlanma Ayı (Month): Trend videoların belirli bir mevsimi olup olmadığını anlamak için.

Haftanın Günü (Day of Week): Hafta sonu veya tatil günlerinin etkisini görmek için.

Yayınlanma Saati (Hour): Video yüklemek için en iyi saatı belirlemek için.

Kategorik Özellikler (Categorical Features):

Kategori ID (Category ID): Videonun türü (Eğlence, Eğitim, Spor vb.).

Türetilmiş Özellikler (Engineered Features):

Başlık Uzunluğu (Title Length): Başlıktaki kelime sayısı.

Büyük Harf Oranı (Caps Ratio): Başlıktaki büyük harflerin oranı (Genellikle dikkat çekici "Clickbait" başlıklarında kullanılır).

Noktalama İşaretleri (Punctuation): Başlıktaki ünlem (!) veya soru işaretçi (?) bulunup bulunmadığı.

Train-Test Split Oranı

Train-Test Split Veri seti, `test_size=0.2` kullanılarak %80 eğitim ve %20 test olarak ayrılmıştır. Eğitim verileri modelin öğrenmesi için, test verileri ise model performansının değerlendirilmesi için kullanılmıştır.

Ayrıca `stratify=y` parametresi ile trend ve trend olmayan videoların oranı her iki veri kümesinde de korunmuştur.



Modellerin karşılaştırıldığı tablo



Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
XG-BOOST	0.8623	0.8454	0.8655	0.8253	0.8996
MLP	0.9072	0.9287	0.8590	0.8925	0.9658
LightGBM	0.8934	0.8842	0.8425	0.8304	0.9152
Tabnet	0.90453	0.8796	0.8508	0.8864	0.9167
STACKING ENSEMBLE Models: XGBoost + LightGBM + CatBoost	0.8654	0.8763	0.8064	0.8506	0.8803





Sonuç ve değerlendirme



Projede, YouTube videolarının trend olup olmayacağı tahmin etmek için videoların başlık, kategori, etiketler, açıklama ve yayınlanma zamanı gibi özelliklerini analiz ettik. Yeni yüklenen videoların henüz izlenme, beğeni ve yorum verisi olmadığı için, modelimiz sadece yüklenme anında bilinebilecek bilgilerle çalışmaktadır. Bu durum tahmin gücünü sınırlamakta olup, gelecek çalışmalarda erken dönem verilerin de modele dahil edilmesi planlanmaktadır. Gerçekleştirdiğimiz model karşılaştırmalarında, base modeller (Random Forest, Gradient Boosting) ve advanced modeller (XGBoost, LightGBM, Stacking) test edilmiştir. Veri setimizdeki yapıyı en iyi öğrenen ve en başarılı sonucu veren algoritma XGBoost olmuştur.

Geliştirdiğimiz model, kullanıcıdan alınan video bilgilerine göre trend olma potansiyelini başarıyla tahmin edebilmekte. Başlık, etiketler, açıklama ve yayınlanma zamanı gibi özellikler üzerinde yapılan detaylı feature engineering ile model performansı optimize ettik. Gerçekleştirdiğimiz testlerde model, gerçekçi ve tutarlı sonuçlar üretmekte olup, kullanıma hazır duruma getirilmiştir. Sistem, interaktif test modülüyle yeni videolar için anlık tahmin yapabilmektedir..

