**Film Gişe Hasılatı Tahmini**

Movie Gross Prediction

Yazar: Barış Ceylan 202803012

# **ÖZET**

Film gişe hasılatı tahmini projesinde başarı ölçütlerinde yüksek sonuçlar alabilmek adına düzenlemeler yaparak ölçüt sonuç değerleri yükseltildi. Bu değerler, modelin gerçek verilere ne kadar iyi uyum sağladığını ve tahminlerin ne kadar doğru olduğunu göstermektedir. Ayrıca, film türleri ve distribütörler üzerinden merkezi eğilim ve merkezi dağılım ölçümleri de hesaplandı. Bu ölçümler ile, farklı film türleri ve distribütörler arasındaki gelir dağılımının genel eğilimleri analiz edilmiş oldu.

Projenin temel amacı, film endüstrisinde veri madenciliği tekniklerini kullanarak bilgi çıkarmaktı. Elde edilen başarı ölçütleri, projenin etkili bir şekilde gerçekleştirildiğini ve film geliri tahmininde önemli bir doğruluk sağladığını göstermektedir. Bu proje, veri madenciliği alanında pratik bir uygulama sunarak film endüstrisine değerli bilgiler katmayı amaçlamıştır. Proje, sadece gelir tahmini değil, aynı zamanda sektördeki eğilimleri ve faktörleri anlamak adına film endüstrisine derinlemesine bir bakış sunmayı amaçlamıştır. Bu, gelecekteki film projeleri için stratejik kararlar almak isteyen endüstri profesyonelleri ve karar alıcıları için değerli bir kaynak oluşturabilir.

Anahtar Kelimeler: Film Endüstrisi, Film Geliri, Sayısal Tahmin

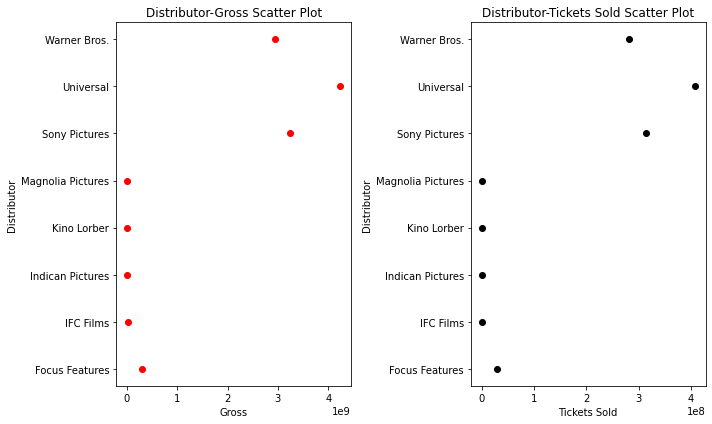
1. Giriş

Film endüstrisi, yüksek karlılığı ve eğlence niteliği nedeniyle hem ekonomistler hem de halk için büyük bir ilgi konusudur. Filmin gişe başarısını önceden tahmin etmek, film piyasasındaki yatırımcıların bilinçli kararlar almak istemesi nedeniyle önemli bir sorundur. Geleneksel olarak, insanlar gişeyi, film türü, MPAA derecesi, bütçe, yönetmen, ilk haftadaki sinema sayısı vb. özel karakteristiklere dayalı olarak tahmin ederler, ancak bu tahminler genellikle sınırlı başarıya sahiptir [1]. Farklı insanlar, film gişe tahminine farklı perspektiflerden yaklaşırlar. Önceki çalışmaların çoğu IMDB (Internet Movie Database) verileri üzerinden regresyon veya stokastik modeller kullanarak film gişe tahmininde bulunur. Ancak, bu modeller ya zayıf performans gösterir ya da makul tahminler yapabilmek için sonrasında veriye ihtiyaç duyarlar, ki bu da pratikte kabul edilemez. Örneğin, Sawhney ve Eliashberg [2], modelinin gişe verilerinin ilk üç haftasını girdi olarak kullanarak oldukça iyi çalıştığını iddia ettiler, ancak erken aşama gişe verileri olmadan model parametreleri veya gişe için şekil tahmini yapmanın çok daha zor olduğunu kabul ettiler. Sonrasında yapılan modeller bazı durumlarda kullanışlı olsa da, önceki dönem modelleri daha fazla pratik öneme sahiptir. Bu projede “The Numbers” film sitesinden verileri çekip film gelirini, film türü ve distribütör verileri üzerinden tahmin etmeye çalıştım. Film türünü kullanmamdaki amaç ise insanların genel algı ve psikolojik tercih sonuçlarının film geliri üzerinde büyük etkisini gösterebilmekti. Böylelikle tahmin başarım sonucu izleyici kitlesi gibi etken faktör üzerinden de değerlendirilebildi. Filmin geniş kitlelere yayılabilmesindeki temel faktör olan reklam için de distribütör şirketleri oldukça kritik öneme sahiptir. Reklamı yüksek olan filmin de izlenme sayısı dolayısıyla geliri yüksek olacağından film geliri tahmininde distribütör etkisini göstermek istedim.

1. Verilerin Anlatımı

The Numbers [3] web sitesinden 2020, 2021, 2022 ve 2023 yıllarında sinemada gösterime girmiş tüm filmlerin film adı, yayınlanma tarihi, distribütör, film türü, gişe hasılatı ve son olarak satılan bilet sayısı verileri ayrı ayrı excel tablosu olarak kaydedecek şekilde uygun kodlarla çekilip tüm veriler tek tabloda birleştirildi. Daha sonra, veri setini daha anlamlı ve odaklı hale getirmek amacıyla film adı ve yayınlanma tarihi kolonlarına özellik seçimiyle tablodan silip ek olarak distribütör filtrelemesi uygulandı. Böylelikle, projenin belirli bir alt kümesi üzerinde odaklanarak daha etkili sonuçlar elde etme hedeflendi. Film distribütörlerinin son 4 yıldaki toplam gişe hasılatı değerleri Şekil 1’de verilmiştir. Bu grafikten yola çıkarak son 4 yılda en yüksek gişe hasılatına Universal’ın ulaştığını söyleyebiliriz.

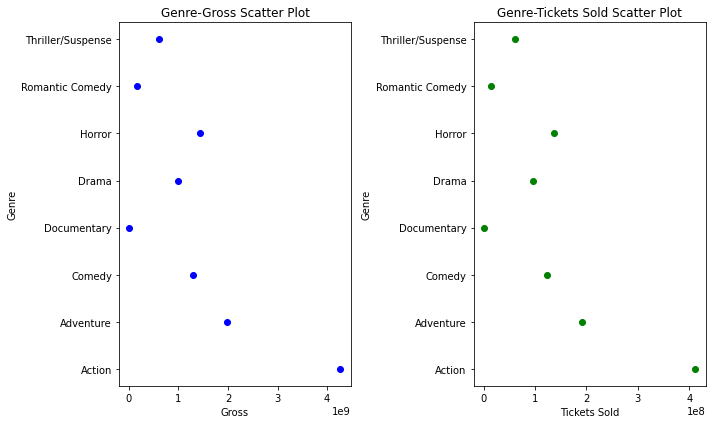
a b



**Şekil 1** Her Distribütörün Son 4 Yıldaki Toplam Gişe Hasılatı ve Satabildiği Bilet Sayıları Değerleri (Kaynak: [The Numbers – Top Grosssing Movies])

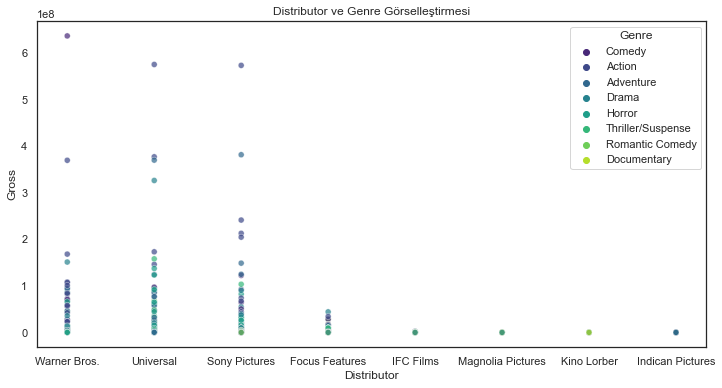
Gişe hasılatı tahmini için kullanılacak olan diğer etken özellik ise film türleridir. Dolayısıyla son 4 yılda elde edilen gişe hasılatı ve satılan bilet sayısı değerlerinden yola çıkarak halk tarafından en çok tercih edilen türün Aksiyon türündeki filmler olduğunu söyleyebiliriz [Şekil 2].

***c d***



**Şekil 2** Son 4 Yılda Gişe Hasılatının ve Satılan Bilet Sayılarının Film Türlerine Göre Değerleri (Kaynak: [The Numbers – Top Grossing Movies])

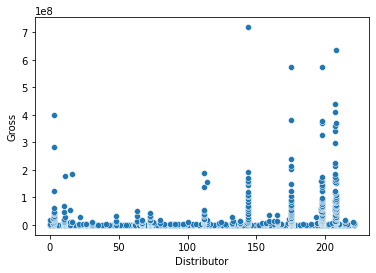
Film distribütörlerinin son 4 yılda hangi film türünden ne miktarda gişe geliri elde ettiğini tespit etmek amacıyla Şekil 3 oluşturulmuştur.



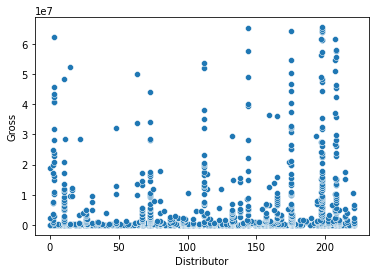
**Şekil 3** Film Distribütörlerinin Film Türlerine Göre Son 4 Yıldaki Toplam Gişe Hasılatı

Veri seti projeye dahil edildikten sonra distribütörlerin gişe hasılatı değerlerini daha iyi inceleyip anlayabilmek adına görselleştirmeler yapıldı. Verilerin dağılımını düzenlemek adına aykırı değerler silindi ve son hali tekrardan görselleştirildi [Şekil 4].

a (Veri setinin normal hali)



b (veri setinin aykırı değerleri çıkartılmış hali)



**Şekil 4** Veri Setinin Versiyonlarının Görsel Halleri

1. Önerilen Yöntemler

3.1. Veri Toplama Süreci

Veri toplama sürecini başlatmak için, Python programlama dilinde otomasyon kütüphanesi olan Selenium'un webdriver'ını kullandım. Bu kütüphane, web tarayıcısını otomatik olarak kontrol etme yeteneği sağlar. Veri toplama sürecinde kullanmak üzere, öncelikle Selenium'u Python ortamına kurarak projemde kullanılabilir hale getirdim. Daha sonra, The Numbers film sitesinden ilgili verilere erişebilmek için Selenium'un sağladığı tarayıcı kontrolünü kullanarak Python kodu yazdım. Bu kodlar, web tarayıcısını otomatik olarak açma, belirli bir URL'ye gitme, sayfa içinde gezinme ve gerekli verileri çekme işlemlerini içeriyordu. Selenium'un webdriver'ı, tarayıcı otomasyonunu sağlamak için kullanıldı. Tarayıcı otomasyonu, siteye bağlanma, sayfada dolaşma ve belirli öğeleri tanıma gibi görevleri otomatikleştirmeme olanak tanıdı. Özellikle, Python dilinde yazılmış bu otomasyon kodları, The Numbers sitesine bağlanarak belirli film verilerini çekmek üzere özelleştirildi.

3.1.1. Veri Analiz Süreci

Feature selection ve distribütör filtrelemesi uyguladıktan sonra metin olarak çekilen distribütör ve film türü kolonlarını sayısallaştırdım. Python' da Pandas kütüphanesini kullanarak merkezi eğilim ölçülerini ve merkezi yayılım ölçülerini hesaplayarak verilerdeki eğilimi ve yayılım durumunu inceledim.

3.1.2. Kaba Kod

1. Veriyi 'movienumericversion2.xlsx' dosyasından oku

2. Distributor kolonundaki NaN değerlere sahip satırları sil

3. Scatter plot çizdir: x="Distributor", y="Gross"

4. Veriyi Gross değerine göre sırala

5. Verinin üst ve alt yüzde 10'unu al

6. Gross değerlerinin ortalamasını hesapla ve yazdır

7. Scatter plot çizdir: x="Distributor", y="Gross" (filtrelenmiş veri üzerinde)

8. Veriyi özellikler (X) ve hedef değişken (y) olarak ayır

9. Veriyi eğitim ve test setlerine ayır

10. Veriyi ölçeklendir

11. Random Forest Regressor modelini eğit

12. Modelin performansını değerlendir: MAE, RMSE, R2

13. Veri setinde belirli bir satırı seç

14. Seçilen satırı modeli kullanarak tahmin et

15. Gerçek ve tahmin edilen değerler arasında scatter plot çizdir

1. Deneysel Sonuçlar ve Tartışma

Veri setini elde edip düzenlemeler yaptıktan sonra merkezi eğilim ve merkezi yayılım ölçüm sonuçlarının gösterdiğine göre son 4 yıl içerisinde ortalama en yüksek gişe hasılatını yakalayan distribütör Universal olurken tek bir filmde en yüksek hasılatı elde eden ise Warner Bros. olmuştur [Tablo 1]. Ek olarak film türleri üzerinden ölçümleri uygulayınca en yüksek ortalama hasılata sahip tür aksiyonken tek filmde en yüksek hasılata sahip tür komedi olmuştur [Tablo 2].

**Tablo 1** Distribütörlerin Son Dört Yıldaki Gişe Gelirlerinin Eğilim ve Yayılım Ölçüm Değerleri ($)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Distributor | Avg | SS | Max | Min | Median | Mod |
| Warner Bros. | 55357274.83 | 102145661 | 636210056 | 38615 | 23009311 | 38615 |
| Universal | 60569443.87 | 99976510 | 574934330 | 14350 | 22609647.5 | 14350 |
| Sony Pictures | 43902528.64 | 90409274.71 | 572984769 | 539 | 10297617 | 539 |
| Focus Features | 6242786.408 | 9751319.835 | 44141550 | 10085 | 1860050 | 10085 |
| IFC Films | 154431.3696 | 348891.7311 | 2418546 | 1250 | 34178.5 | 1250 |
| Magnolia Pictures | 91632.80851 | 157222.2409 | 667114 | 180 | 34502 | 180 |
| Kino Lorber | 41784.60656 | 51316.48868 | 213960 | 1052 | 22313 | 1052 |
| Indican Pictures | 10433.1875 | 22611.65902 | 161313 | 200 | 5500.5 | 200 |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

**Tablo 2** Film Türlerine Göre Son Dört Yıldaki Gişe Gelirlerinin Eğilim ve Yayılım Ölçüm Değerleri ($)

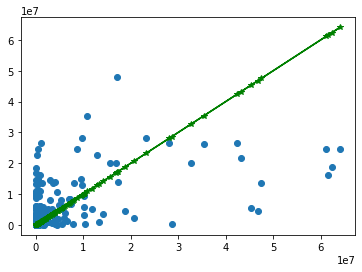
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Genre | Avg | SS | Max | Min | Median | Mod |
| Drama | 6655704.879 | 30865705.85 | 326034265 | 180 | 76718 | 180 |
| Comedy | 16979281.01 | 75246449.36 | 636210056 | 572 | 56687 | 572 |
| Romantic Comedy | 10705045.73 | 18423502.78 | 68229160 | 1988 | 1018532 | 1988 |
| Thriller/Suspense | 9788053.587 | 25513217.65 | 157901466 | 684 | 87344 | 684 |
| Documentary | 186529.6935 | 714747.2644 | 5354970 | 1787 | 27517.5 | 1787 |
| Action | 80510623.53 | 132058697.2 | 574934330 | 974 | 23791725 | 974 |
| Adventure | 60011610.94 | 92518693.02 | 381311319 | 1006 | 26234535 | 1006 |
| Horror | 24268674.71 | 35305915.31 | 137216390 | 1242 | 8702060 | 1242 |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

Modeli geliştirirken Karar Ağacı, K-En Yakın Komşular, Lineer Regresyon, Polinom Regresyon, Ridge Regresyon, Lasso Regresyon, Elastic Net Regresyon, Poisson Regresyon, Support Vector Regresyon ve son olarak Rassal Ormanlar Regresyon yöntemlerinin her biri için öncelikle X\_train ve y\_train için 1304 adet veri ayrılırken X\_test ve y\_test için 306 adet veri ayrılarak veri setinin %19’lik kısmı test aşaması için ayrılmıştır. Modelin performansını ölçebilmek adına bazı metriklere başvurulup yukarıda belirtilen her bir regresyon modeli için sonuçlar üretilmiştir. Yöntemler arasından Rassal Ormanlar Regresyon yöntemiyle geliştirilen model diğer yöntemlerinkinden daha iyi sonuç üretmiştir [Tablo 3].

**Tablo 3** Rassal Ormanlar Regresyon ile Geliştirilen Modelin Metrik Sonuçları

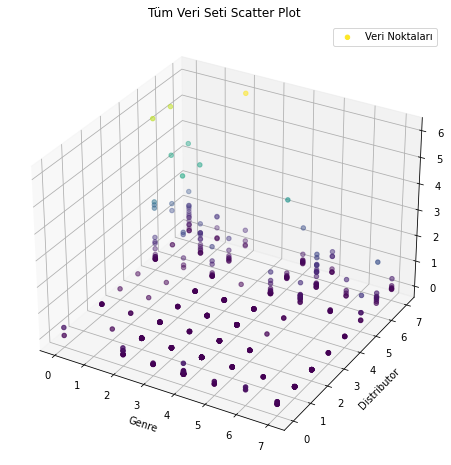
|  |  |
| --- | --- |
| Metrik | Sonuç |
| R-Sqaured Skor | 0.266 |
| Mean Absolute Error | 3,835,678.76 |
| Root Mean Squared  Error | 8,589,221.52 |
|  |  |

Rassal Ormanlar Regresyon ile geliştirilen modelin tahminlerinin gerçek değerlerle ne derecede örtüştüğünü göstermek amacıyla scatter plot kullanıldı [Şekil 5]. Bu grafiğe göre modelin tahmin ettiği değerlerle gerçek değerler arasında gözle görülür bir fark olsa da veri setinin özellik sayısının ve veri sayısının artırılmasıyla daha iyi sonuçlar üretme potansiyeli taşımaktadır.



**Şekil 5** Gerçek Değerler ve Tahmin Değerleri

Metrik sonuçlarının daha iyi anlaşılması ve karşılaştırılmasının yapılabilmesi için 3 boyutlu görsel oluşturuldu. [Şekil 6]. Bu görselle beraber verinin hangi özelliklere göre ne şekilde dağıldığını daha iyi gözlemleyebiliriz.



**Şekil 6** Veri Setinin 3 Boyutlu Hali

1. Sonuç

Bu projede web sitesinden uygun kodlarla çekilen veriler öncelikle temiz hale getirildi ardından ayrı ayrı oluşturulan tablolar birleştirilip farklı regresyon yöntemlerine uygun hale getirildi. Bu yöntemler ile geliştirilen modellerde başarım metrikleri hesaplanarak performans değerlendirmeleri yapıldı. Metrik sonuçları diğer yöntemlerden daha iyi olup daha iyi performans göstermesinden dolayı Rassal Ormanlar Regresyon yöntemi seçildi. Merkezi eğilim ölçümleri ve merkezi yayılım ölçüleri kullanılarak film distribütörleri ve film türlerinin istatistiki değerleri

gösterildi. Bu sayede film endüstrisi için film türü ve film distribütöründen yola çıkılarak elde edilebilecek gişe hasılatı hesaplanmış oldu.

Yazarların Katkıları

Bu projeyi tek kişi olarak gerçekleştirdim. Öncelikle Python’da Selenium kütüphanesini dahil ederek WebDriver ile tarayıcıya bağlanabilmek için kodları yazıp düzenledim. İlgili film sitesindeki Top Movies Gross verilerini 2020, 2021, 2022 ve 2023 yılları için çekip ayrı ayrı excel tabloları halinde kaydettim. Topladığım verileri tek tabloda birleştirdikten sonra distribütör sayısının fazlalığından dolayı oluşan karmaşıklığı önlemek adına en çok tekrar eden ilk 8 distribütör dışındakileri veri setinden çıkarıp ardından model eğitimi sırasında kullanmayacağım yayınlanma tarihi ve film adı kolonlarını veri setinden sildim. Verilerin son hali üzerinden ortalama, standart sapma, en yüksek değer, en düşük değer, medyan ve son olarak mod merkezi ölçüm sonuçlarını Python’da Pandas kütüphanesini kullanarak hesapladım. Daha sonra kategorik verileri sayısallaştırıp verileri normalize ederek 10 farklı regresyon yöntemiyle modeller oluşturdum. Regresyon modeline hazır hale getirip modellerin performansını ölçebilmek için R-Sqaured Skor, Mean Absolute Error ve Root Mean Squared Error başarım metriklerini kullandım. Bu başarım metriklerinden en iyi sonuçları üreten Rassal Ormanlar Regresyon yöntemini seçtim.

Kaynaklar

[1] M. S. Sawhney and J. Eliashberg, “A parsimonious model for forecasting gross box-office revenues of motion pictures,” Marketing Science, vol. Vol. 15, No. 2, pp. 113–131, 1996.

[2] W. Zhang and S. Skiena, "Improving Movie Gross Prediction through News Analysis," 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, Milan, Italy, 2009, pp. 301-304, doi: 10.1109/WI-IAT.2009.53.

[3] The Numbers. Top Grossing Movies. https://www.the-numbers.com/market/2023/top-grossing-movies