Gişe Verilerine Göre Hasılat Tahmini

Miray Gürbüz
Bilişim Sistemleri Mühendisliği
Kocaeli Üniversitesi
Kocaeli, Türkiye
221307031

Zeynep Yılmaz
Bilişim Sistemleri Mühendisliği
Kocaeli Üniversitesi
Kocaeli, Türkiye
221307012

***Özet*— Bu rapor, Yazılım Geliştirme Laboratuvarı – I dersi kapsamında gerçekleştirilen gişe verilerine göre hasılat tahmini projesinin detaylarını içermektedir. İlk aşamada, Box Office Mojo web sitesinden günlük gişe verileri toplanmıştır. Veri toplama Selenium ve BeautifulSoup4 kullanılarak gerçekleştirilmiştir. Veri ön işleme aşamasında veri setindeki eksik bilgiler doldurulmuş; zaman serisinin yeniden örneklenmesi, trend ve mevsimsellik bileşenlerinin ayrıştırılması, otokorelasyon durumunun incelenmesi adımları uygulanmıştır. Bu adımlar sonrasında çıkartılan sonuçlar doğrultusunda özellik mühendisliği yapılarak projenin ikinci aşamasındaki modelleme adımında yarar sağlayacak yeni sütunlar eklenmiştir. İkinci aşamada, zaman serisi tahmini için Transformer tabanlı modeller üzerinde çalışılmıştır. Autoformer, TFT, Informer, FEDformer ve Vanilla Transformer olmak üzere beş farklı model uygulanmış ve performansları karşılaştırılmıştır.**

Anahtar Kavramlar—zaman serisi tahmini, web kazıma, veri ön işleme, yeniden örnekleme, trend, mevsimsellik, otokorelasyon, özellik mühendisliği, transformer modelleri

I. GİRİŞ

Zaman serileri, belirli bir veri ögesinin zaman içinde düzenli aralıklarla ölçülmesiyle elde edilen gözlemlerden oluşur[1]. Proje kapsamında günlük olarak ölçülen gişe verileri kullanılarak hasılat tahmini yapmak amaçlanmaktadır.

II. VERİ TOPLAMA

A. *Veri Kaynağı Seçme*

Veri kaynağı seçilirken verilerin düzenli aralıklarla kayıt altına alınmış olması, eksik verilerin olmaması ve kaynağın güvenilirliği önemlidir. Tüm bunlardan dolayı Amazon şirketi olan IMDb'ye 2008 yılından beri bağlı olan Box Office Mojo[2] web sitesi kaynak olarak seçilmiş ve sitedeki Daily sayfasındaki günlük verilerden yararlanılmıştır.

B. *Web Kazıma Aşaması*

Box Office Mojo web sitesinden 2009-2024 yılları arasında kaydedilmiş 5781 günlük gişe verisi çekilmiştir. Selenium ve BeautifulSoup4 kullanılarak 15 yıla ait 15 adet sayfanın her birindeki günlük satır kayıtları çekildikten sonra, BeautifulSoup4 ve Requests kütüphaneleri kullanılarak 5781 güne ait 5781 adet sayfanın her birinden o gün vizyonda olan filmlerin toplam hasılat değerleri toplanmış ve toplanmış olan günlük toplam hasılat değeri çekilmiştir. Aynı işlem 2004-2019 *(pandemi etkisi görülmeyen yıllar)* yılları arasında kaydedilmiş 15 yıla ait 5844 günlük veri için de yapılmıştır. Bu işlemler concurrent.futures kütüphanesindeki ThreadPoolExecutor[3] sınıfı kullanılarak paralel olarak gerçekleştirilmiştir. Çekilen veriler iki ayrı veri seti olarak kaydedilmiştir.

C. *Veri Sütunlarının İçeriği*

- *url:* O günün bireysel sayfasının URL bilgisini içerir.
- *date:* O güne ait tarihi içerir.
- *day_of_year:* Yılın kaçıncı günü olduğunu belirtir.
- *releases:* O gün vizyonda olan film sayısını gösterir.
- *top_1_gross:* O gün hasılat bakımından birinci olan filmin günlük hasılatını belirtir.
- *top_1_release:* O gün hasılat bakımından birinci olan filmin adını içerir.
- *top_10_gross:* O gün hasılat bakımından ilk on film arasına giren filmlerin günlük toplam hasılatını gösterir.
- *total_gross:* O gün vizyonda olan tüm filmlerin günlük toplam hasılatını belirtir.

III. VERİ ÖN İŞLEME

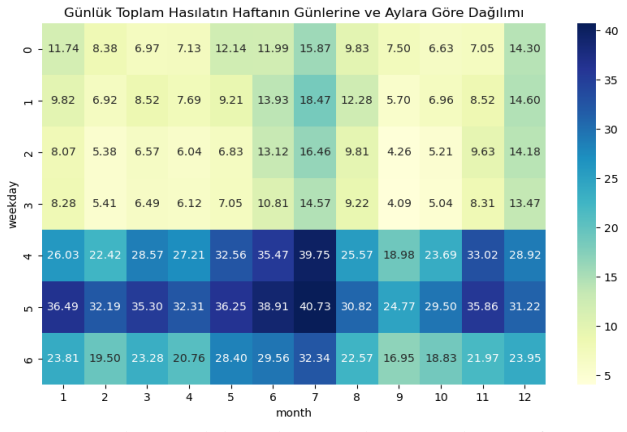
Veri ön işleme aşamasında; veri analizi, zaman değişkenlerinin eklenmesi, zaman serisinin yeniden örneklenmesi, trend ve mevsimsellik bileşenlerinin ayrıştırılması, otokorelasyon ve özellik mühendisliği adımları uygulanmıştır.

A. *Veri Analizi*

Bu adımda günlük toplam hasılat verileri üzerinde çeşitli analizler yapılarak filmlerin hasılat performansları gözlemlenmiştir.

1) Günlük Toplam Hasılatın Haftanın Günlerine ve Aylara Göre Dağılımı

Seaborn kütüphanesi kullanılarak haftanın günlerine ve aylara göre günlük toplam hasılat verilerinin dağılımı görselleştirilmiştir. Elde edilen sonuçlar, hangi günlerin ve ayların daha yüksek gelir sağladığını ortaya koymuştur.

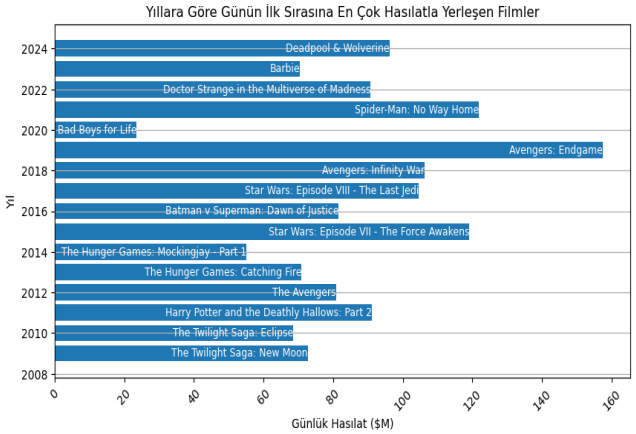


Görsel 1: Günlük Toplam Hasılatın Dağılım Grafiği

Grafikte toplam hasılat değerinin en yüksek olduğu ayın temmuz, günün ise cumartesi olduğu gözlemlenmektedir.

2) Yıllara Göre Günün İlk Sırasına En Çok Hasılatla Yerleşen Filmler

Yıllara göre her günün en yüksek hasılatını elde eden filmler incelenmiş ve yıllık bazda hangi filmlerin daha yüksek hasılatlarla zirveye yerleştiği analiz edilmiştir.



Görsel 2: Yıllara Göre İlk Sıraya Yerleşen Filmler Grafiği (2009–2024)

Grafikte 2009-2024 yılları arasında günün ilk sırasına en çok hasılatla yerleşen filmler listelenmiştir. Bu bağlamda en başarılı yılın 2019, filmin ise *Avengers: Endgame* olduğu yorumu yapılabilmektedir.

B. Zaman Serisinin Yeniden Örneklenmesi

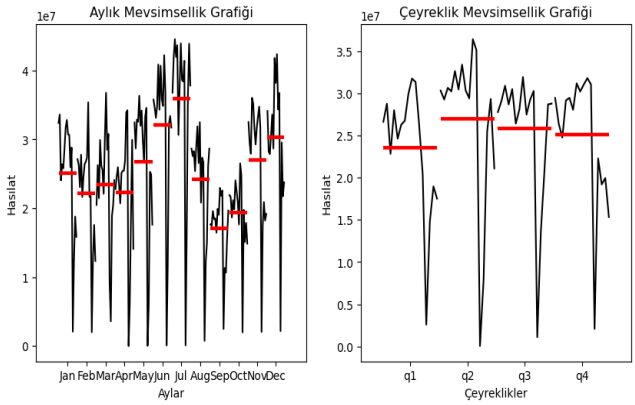
Zaman serisi yeniden örneklenerek verilerin tutarlı bir sıklıkta dağıtılması sağlanılır. Yeniden örnekleme orijinal verinin zaman frekansını değiştirerek gerçekleştirilir.[4] İki adet yeniden örnekleme yöntemi vardır:

- *Yukarı Örnekleme:* Zaman frekansını arttırır. Örneğin; günlerden saatlere ayırmak. Bu sayede veri sayısı artar.
- *Aşağı Örnekleme:* Zaman frekansını azaltır. Örneğin; günlerden haftalara ayırmak. Bu sayede veri sayısı azalır ve daha yumuşak bir gözlem yapılabilir.

Proje kapsamında çekilen veriler günlük frekansta olduğundan aşağı örnekleme tercih edilmiştir. Günlük veri; haftalık, aylık, çeyreklik ve yıllık olarak yeniden örneklenmiştir.

C. Trend ve Mevsimsellik Bileşenleri

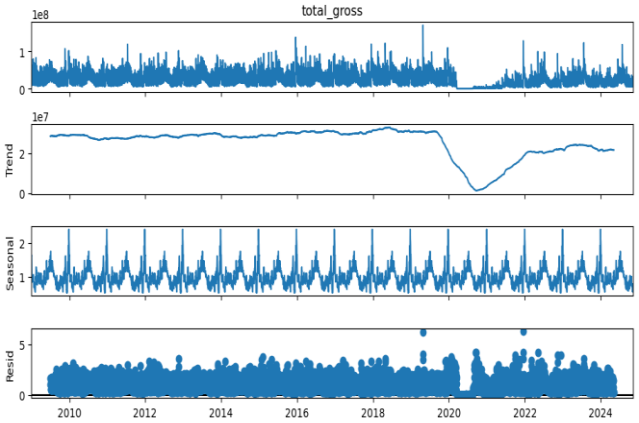
- *Trend:* Verideki uzun vadede gözlemlenen genel hareketi ifade eder.
- *Mevsimsellik:* Verinin belirli bir periyodik döngüde değişen ve bu değişikliklerin önceden tahmin edilebilir olduğu durumu ifade eder.



Görsel 3: Aylık ve Çeyreklik Mevsimsellik Grafiği(2009-2024)

Yukarıdaki grafikler incelendiğinde veride aylık ve çeyreklik olarak mevsimsellik gözlemlenmektedir. Kırmızı çizgiler toplam hasılat verisinin ilgili ay ve çeyrekliklerdeki ortalama değerlerini gösterirken, siyah çizgiler toplam hasılat verisinin yıllar boyu olan değişimini gösterir. Kırmızı çizgilerin, yani ortalama değerlerin, artıp azalıyor olması mevsimselliğe işarettir.

1) Trend ve Mevsimsellik Bileşenlerinin Ayrıştırılması



Görsel 4: Mevsimsel Ayrışım Grafiği (2009-2024)

Statsmodels kütüphanesinin seasonal_decompose fonksiyonu kullanılarak toplam hasılat sütununun trend ve mevsimsellik bileşenleri ayrıştırılmıştır. Verinin yıllık döngü yaptığı varsayılmış ve bileşenler multiplicative model ile ayrıştırılmıştır.

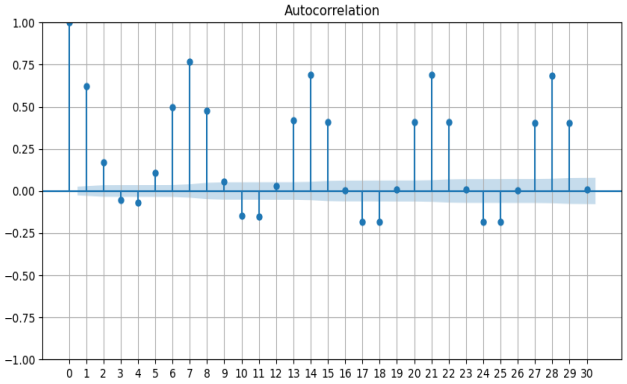
Grafik incelendiğinde trend bileşeninin 2020 sonrasında pandemi etkisiyle düşüş yaşamış olabileceği yorumu yapılabilmektedir. Mevsimsellik grafiğinde örüntü gözlemlenmesi verinin mevsimselliğinin olduğuna işaret etmektedir.

D. Otokorelasyon

Otokorelasyon, zaman serilerinde serinin gecikmeli değerleriyle şu anki değerleri arasındaki ilişkiyi ölçer. Değerlerin kendi kendisiyle olan korelasyonunu gösterir.

Otokorelasyon katsayısı 1 ile -1 arasında değişiklik gösterir:

- *1'e yakın değerler:* Pozitif otokorelasyon anlamına gelir. Önceki değerlerle benzer yönde bir değişim eğilimi gösterir.
- *0'a yakın değerler:* Otokorelasyonun zayıf olduğunu gösterir. Geçmiş verilerin mevcut veriler üzerindeki etkisi yok sayılır.
- *-1'e yakın değerler:* Negatif otokorelasyon anlamına gelir. Önceki değerlerin tersine hareket eder.



Görsel 5: Otokorelasyon Grafiği

Yukarıdaki grafik, otokorelasyonun gözlenmesi için Statsmodels'in ACF fonksiyonuyla 30 gecikmeye kadar oluşturulan otokorelasyon grafiğidir.

Grafiğe bakıldığı zaman otokorelasyon değerlerinin 7 günlük aralıklarla yükseldiği gözlemlenmektedir. Buradan yola çıkılarak günlük toplam hasılat değerlerinin *haftalık* olarak mevsimsellik gösterdiği şeklinde bir yorum yapılabilmektedir.

E. Özellik Mühendisliği

- *Özellik:* Verideki gözlemler hakkında bilgi veren her bir değişken veya özniteliktir.
- *Özellik mühendisliği:* Ham veriyi modellemede kullanılabilecek özelliklere dönüştürmek için seçme, manipüle etme ve dönüştürme sürecidir[5].

Özellik mühendisliği adımında veri setine; 7 günlük gecikmeli toplam hasılat değeri, 7 ve 30 günlük toplam hasılat hareketli ortalamaları, günlük getirileri, ilgili günün resmi tatil olup olmadığı gibi bilgileri içeren sütunlar eklenmiştir. Döngüsel sütunlara sinüs ve kosinüs dönüşümleri uygulanmıştır.

IV. MODELLEME

A. Veri Seti ve Hazırlık

Modelleme aşamasında veri seti olarak 2004-2019 yılları arasında kaydedilmiş 5844 günlük veri seti seçilmiştir. Hedef değişken olarak toplam gişe hasılatı (*total_gross*) belirlenmiştir.

Veri ön işleme aşamasında yapılan özellik mühendisliğine ek yeni özellikler eklenmiştir.

- Gecikmeli özellikler: *top_10_gross_[lag7]*, *top_1_gross_[lag7]*.
- Trend özellikleri: Harmonik trend *(harmonic_trend)*, momentum, ortalama değişim *(avg_change)*, kümülatif değişim *(cumulative_change)*.

Veri seti; eğitim ve test setlerine %80-%20 oranında ayrılmış, ardından eğitim setinin %20 oranında validasyon boyutu belirlenmiştir.

Tüm modellerin eğitimi öncesinde zaman serisi verileri *local_scaler_type* parametresi kullanılarak ölçeklenmiştir. *Standard* ölçekleme yöntemi (sıfır ortalama ve birim varyans) tercih edilmiş, bu sayede tüm zaman serilerinde tutarlı bir ölçek sağlanmıştır. Bu işlem; modelin eğitim sürecinde daha hızlı yakınsama, kararlı öğrenme ve genelleme performansını artırmak amacıyla gerçekleştirilmiştir. Tahminler, ölçekleme sonrası orijinal birimlere kütüphane tarafından otomatik olarak geri dönüştürülmektedir.

B. Kullanılan Modeller

Zaman serisi tahmini için 5 farklı transformer modeli test edilmiştir. Bu modellerin her biri NeuralForecast kütüphanesi[6] ile uygulanmıştır:

- *Autoformer*
- *TFT (Temporal Fusion Transformer)*
- *Informer*
- *FEDformer*
- *Vanilla Transformer*

Her model için hiperparametreler seçilmiştir, örneğin:

- Tahmin ufku (*h*): 30 gün ve
- Giriş veri uzunluğu (*input_size*): 60 gün

olacak şekilde belirlenmiştir.

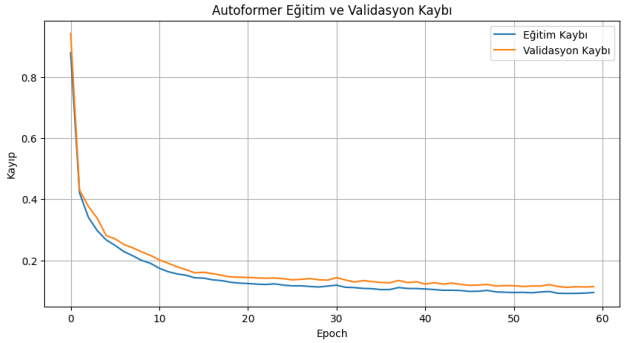
1) Autoformer

Autoformer modeli, uzun vadeli tahminlerde karmaşık zaman desenlerini anlamaya çalışır[7].

- Trend ve sezonluk bileşenleri hareketli ortalama filtresiyle ayırır.
- Dönemsel bağımlılıkları otokorelasyon mekanizması ile bulur.
- Çok başlı dikkat mekanizması içeren klasik encoder-decoder yapısını kullanır.

a) Model Eğitimi ve Performansı

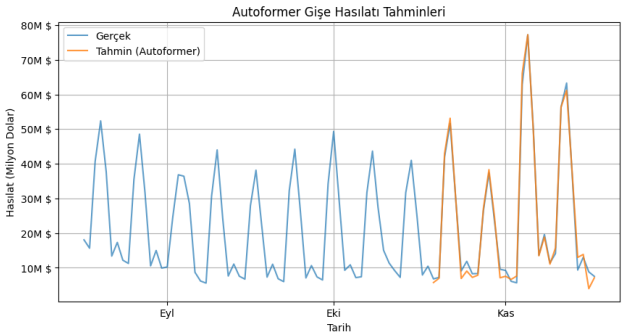
Eğitim süresi; 185,11 saniye, yani yaklaşık 3,09 dakika olarak hesaplanmıştır.



Görsel 6: Autoformer Eğitim ve Validasyon Kaybı

b) Tahminler

Çıkarım süresi; 0,30 saniye olarak hesaplanmıştır.



Görsel 7: Autoformer Gişe Hasılatı Tahminleri

c) Autoformer Model Performans Metrikleri

- *MAPE:* %10,80
- *MAE:* 1,30 milyon
- *MSE:* 3,03 trilyon
- *RMSE:* 1,74 milyon
- *R-squared:* %99,29

Autoformer modeli, %10,80 MAPE ve 1,30 milyon MAE ile kabul edilebilir bir performans göstermiştir. 3,03 trilyon MSE ve 1,74 milyon RMSE değerleri, tahminlerin doğruluğunu desteklerken, %99,29 R-squared değeri modelin veriyi iyi bir şekilde açıkladığını göstermektedir.

2) TFT (Temporal Fusion Transformer)

Temporal Fusion Transformer (TFT), birden fazla adım ileriye tahmin yaparken şu yapıların birleşimini kullanır[8]:

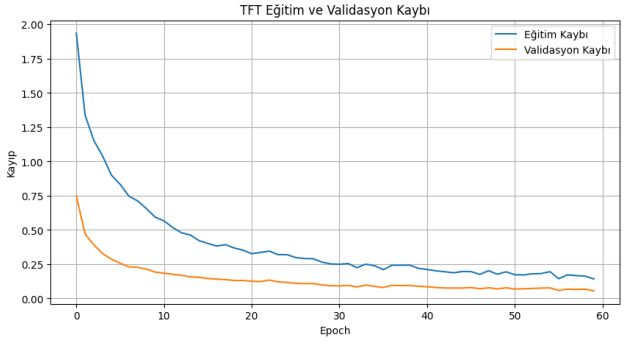
- Gating katmanları (veri seçimi için)
- LSTM tabanlı bir encoder (zaman dizilerini işlemek için)
- Çok başlı dikkat katmanları (önemli bilgileri seçmek için)

TFT'nin girdileri şunlardır:

- *Sabit özellikler:* Değişmeyen veriler (örneğin, ürün türü)
- *Geçmiş bilgiler:* Geçmişteki ek veriler
- *Tahmin anındaki ek bilgiler:* O sırada bilinen diğer veriler
- *Oto-regresif özellikler:* Daha önceki tahmin edilen hedef veriler

a) Model Eğitimi ve Performansı

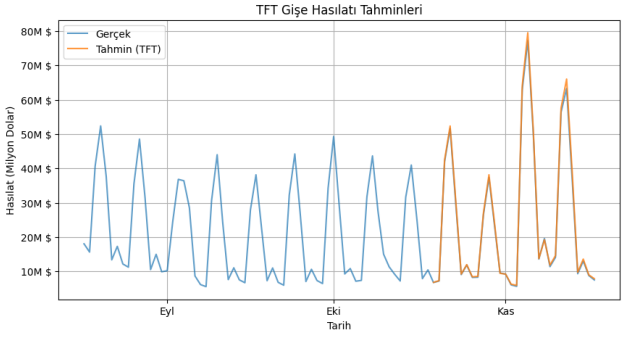
Eğitim süresi; 144,54 saniye, yani yaklaşık 2,41 dakika olarak hesaplanmıştır.



Görsel 8: TFT Eğitim ve Validasyon Kaybı

b) Tahminler

Çıkarım süresi; 0,16 saniye olarak hesaplanmıştır.



Görsel 9: TFT Gişe Hasılatı Tahminleri

c) TFT Model Performans Metrikleri

- *MAPE:* %2,59
- *MAE:* 0,64 milyon
- *MSE:* 0,86 trilyon
- *RMSE:* 0,93 milyon
- *R-squared:* %99,80

TFT modeli, %2,59 MAPE ve 0,64 milyon MAE ile düşük hata oranları sunarak başarılı bir performans göstermiştir. 0,86 trilyon MSE ve 0,93 milyon RMSE değerleri; tahminlerin doğruluğunu desteklerken, %99,80 R-squared değeri modelin veriyi etkili bir şekilde açıkladığını göstermektedir.

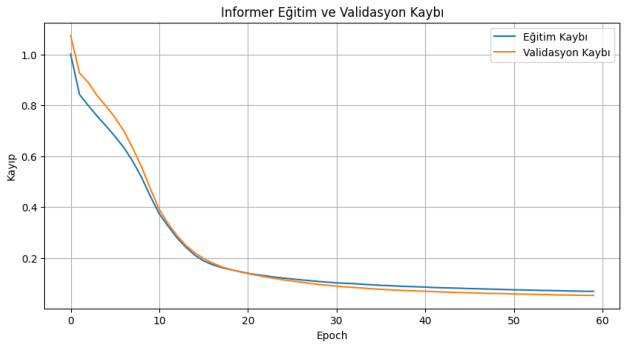
3) Informer

Informer modeli, uzun vadeli tahminlerdeki karmaşıklığı azaltmayı amaçlar[9].

- Daha hızlı ve verimli bir dikkat mekanizması (ProbSparse) kullanır.
- Uzun giriş dizilerini özetleyen dikkat azaltma sürecini uygular.
- Çok adımlı tahminleri tek bir işlemde gerçekleştiren bir çıkış katmanına sahiptir.

a) Model Eğitimi ve Performansı

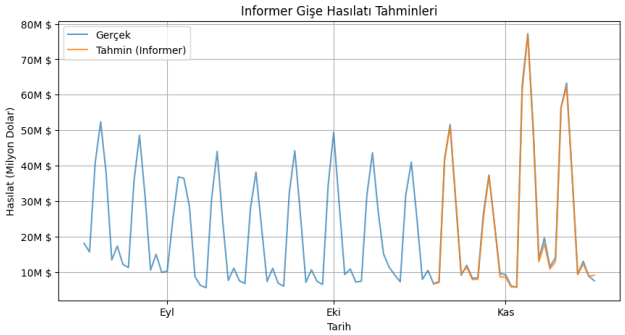
Eğitim süresi; 75,97 saniye, yani yaklaşık 1,27 dakika olarak hesaplanmıştır.



Görsel 10: Informer Eğitim ve Validasyon Kaybı

b) Tahminler

Çıkarım süresi; 0,25 saniye olarak hesaplanmıştır.



Görsel 11: Informer Gişe Hasılatı Tahminleri

c) Informer Performans Metrikleri

- *MAPE:* %4,05
- *MAE:* 0,60 milyon
- *MSE:* 0,58 trilyon
- *RMSE:* 0,76 milyon
- *R-squared:* %99,86

Informer modeli, %4,05 MAPE ve 0,60 milyon MAE ile düşük hata oranlarına sahip olup başarılı bir performans sergilemiştir. %99,86 R-squared değeri, modelin veriyi yüksek doğrulukla açıkladığını gösterirken RMSE ve MSE değerleri de tahminlerin güvenilir olduğunu göstermektedir.

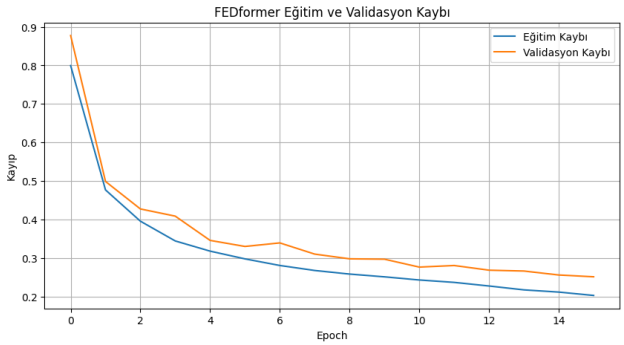
4) FEDformer

FEDformer modeli, uzun vadeli tahminlerde karmaşık zaman desenlerini anlamaya çalışır[10].

- Trend ve sezonluk bileşenleri hareketli ortalama filtresiyle ayırır.
- Dikkat mekanizmasını daha verimli hale getirmek için Fourier dönüşümüne dayalı Frekans Geliştirilmiş Blok ve Dikkat mekanizması kullanır.
- Klasik encoder-decoder yapısı ve çok başlı dikkat mekanizması içerir.

a) Model Eğitimi ve Performansı

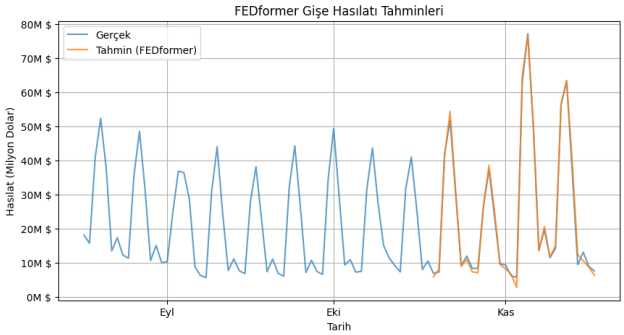
Eğitim süresi; 86,43 saniye, yani yaklaşık 1,44 dakika olarak hesaplanmıştır.



Görsel 12: FEDformer Eğitim ve Validasyon Kaybı

b) Tahminler

Çıkarım süresi; 0,18 saniye olarak hesaplanmıştır.



Görsel 13: FEDformer Gişe Hasılatı Tahminleri

c) FEDformer Performans Metrikleri

- *MAPE:* %8,75
- *MAE:* 1,11 milyon
- *MSE:* 2,11 trilyon
- *RMSE:* 1,45 milyon
- *R-squared:* %99,51

FEDformer modeli, %8,75 MAPE ve 1,11 milyon MAE ile kabul edilebilir bir hata oranı sergilemiştir. %99,51 R-squared değeri, modelin veriyi açıklama gücünün yüksek olduğunu göstermektedir. Ayrıca, RMSE ve MSE değerleri tahminlerin tutarlı ve güvenilir olduğunu vurgulamaktadır. Ancak, Informer ve TFT modellerine kıyasla daha yüksek hata oranlarıyla çalışmış olması; FEDformer modelinin performans açısından Informer ve TFT modellerinin gerisinde kaldığını göstermektedir.

5) VanillaTransformer

VanillaTransformer[11]:

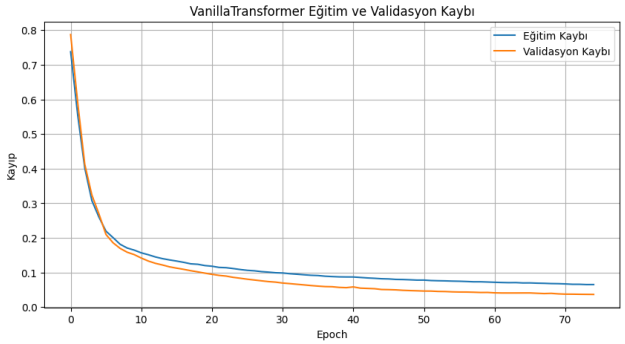
- Tam dikkat mekanizmasıyla çalışır, ancak bu mekanizma zaman ve bellek kullanımı açısından daha yoğundur ($O(L^2)$).
- Encoder-decoder yapısı, farklı veri noktalarını anlamak ve ilişkilerini kurmak için çok başlı dikkat mekanizmasını kullanır.
- Uzun vadeli tahminleri adım adım yapmak yerine, tüm tahminleri tek seferde gerçekleştiren bir çıkış katmanına sahiptir.

Tablo 1: Model Performans Karşılaştırma Metrikleri

	Autoformer	TFT	Informer	FEDformer	VanillaTransformer
MAPE	%10,80	%2,59	%4,05	%8,74	%3.03
MAE *(milyon)*	1,30	0,64	0,60	1,11	0,44
MSE *(trilyon)*	3,03	0,86	0,58	2,11	0,30
RMSE *(milyon)*	1,74	0,93	0,76	1,45	0,55
R-squared	%99,29	%99,8	%99,73	%99,51	%99,93

a) Model Eğitimi ve Performansı

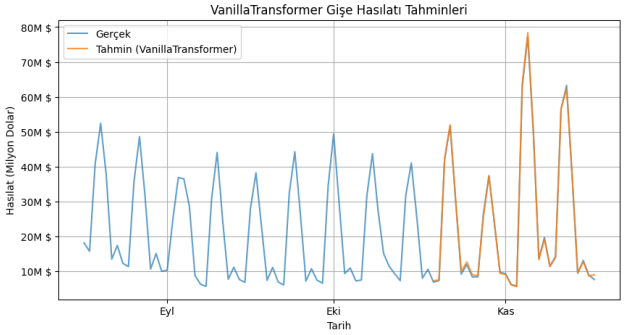
Eğitim süresi; 59,78 saniye olarak hesaplanmıştır.



Görsel 14: VanillaTransformer Eğitim ve Validasyon Kaybı

b) Tahminler

Çıkarım süresi; 0,22 saniye olarak hesaplanmıştır.



Görsel 15: VanillaTransformer Gişe Hasılatı Tahminleri

c) VanillaTransformer Performans Metrikleri

- *MAPE:* %3,03
- *MAE:* 0,44 milyon
- *MSE:* 0,30 trilyon
- *RMSE:* 0,55 milyon
- *R-squared:* %99,93

VanillaTransformer modeli, %3,03 MAPE ve 0,44 milyon MAE ile en düşük hata oranlarına ulaşarak güzel bir performans göstermiştir. 0,30 trilyon MSE, 0,55 milyon RMSE ve %99,93 R-squared değeri, modelin tahminlerde yüksek doğruluk ve güçlü açıklayıcılık sunduğunu göstermektedir.

V. SONUÇ

Model performans karşılaştırma metrikleri incelendiğinde, VanillaTransformer modelinin %3,03 MAPE ve 0,44 milyon MAE ile gişe hasılatı tahmininde en başarılı model olduğu görülmektedir. TFT ve Informer modelleri de oldukça yakın performans sergilemiştir. Autoformer modeli ise diğer modellerden düşük performans göstermiştir. Genel başarı sıralaması VanillaTransformer > Informer ≥ TFT > FEDformer > Autoformer şeklindedir.

REFERANSLAR

[1] "Time Series Analysis: The Basics", Australian Bureau of Statistics. [Çevrimiçi] Mevcut: https://www.abs.gov.au/websitedbs/D3310114.nsf/home/Time+Series+Analysis:+The+Basics Erişim Tarihi: [10 Kasım 2024]

[2] Box Office Mojo. [Çevrimiçi] Mevcut: https://www.boxofficemojo.com/ Erişim Tarihi: [4 Kasım 2024]

[3] "concurrent.futures — Launching parallel tasks", python.org. [Çevrimiçi] Mevcut: https://docs.python.org/3/library/concurrent.futures.html Erişim Tarihi: [10 Kasım 2024]

[4] "How to Resample Time Series Data in Python?", GeeksforGeeks. [Çevrimiçi] Mevcut: https://www.geeksforgeeks.org/how-to-resample-time-series-data-in-python/ Erişim Tarihi: [10 Kasım 2024]

[5] "Feature Engineering Explained", Built In. [Çevrimiçi] Mevcut: https://builtin.com/articles/feature-engineering Erişim Tarihi: [12 Kasım 2024]

[6] Olivares, K. G., Challú, C., Garza, F., M., Mergenthaler Canseco, M., ve Dubrawski, A., "NeuralForecast: User friendly state-of-the-art neural forecasting models," *PyCon*, Salt Lake City, Utah, ABD, 2022. Erişim: https://github.com/Nixtla/neuralforecast

[7] "Autoformer", Nixtla. [Çevrimiçi] Mevcut: https://nixtlaverse.nixtla.io/neuralforecast/models.autoformer.html Erişim Tarihi: [2 Ocak 2025]

[8] "TFT", Nixtla. [Çevrimiçi] Mevcut: https://nixtlaverse.nixtla.io/neuralforecast/models.tft.html Erişim Tarihi: [2 Ocak 2025]

[9] "Informer", Nixtla. [Çevrimiçi] Mevcut: https://nixtlaverse.nixtla.io/neuralforecast/models.informer.html Erişim Tarihi: [2 Ocak 2025]

[10] "FEDformer", Nixtla. [Çevrimiçi] Mevcut: https://nixtlaverse.nixtla.io/neuralforecast/models.fedformer.html Erişim Tarihi:[2 Ocak 2025]

[11] "VanillaTransformer", Nixtla. [Çevrimiçi] Mevcut: https://nixtlaverse.nixtla.io/neuralforecast/models.vanillatransformer.html Erişim Tarihi: [2 Ocak 2025]