Yapay Zeka 1:

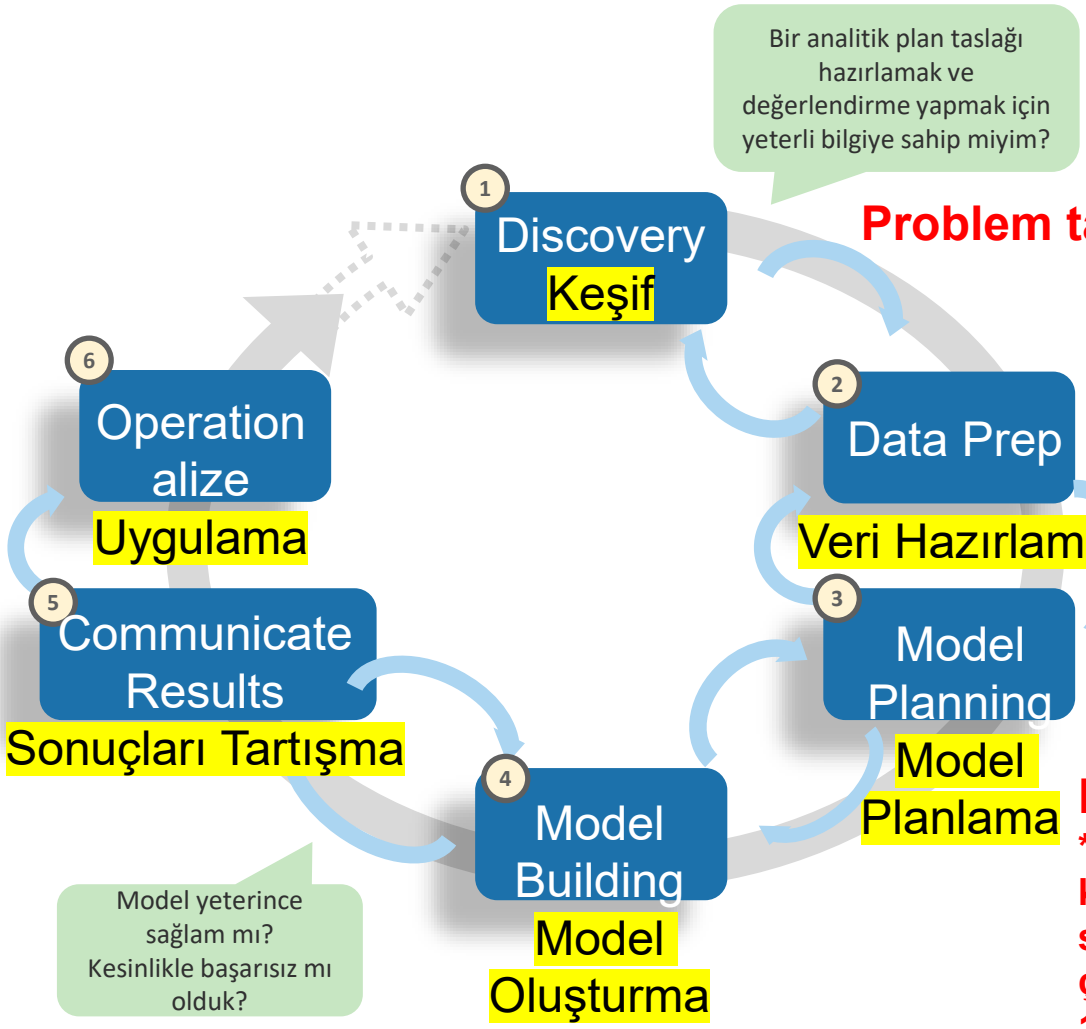
İleri Analik – Yöntemler
Zaman Serisi Analizi

Doç. Dr. Taner Arsan, Ph.Dc. H. Fuat Alsan, Ph.Dc. Sena Kılınç

Bilgisayar Mühendisliği Bölümü

Mühendislik ve Doğa Bilimleri Fakültesi - Kadir Has Üniversitesi

Data Analytics Lifecycle – Veri Analitiği Yaşam Döngüsü



Modeli planlama:

*** Sahip olduğum veri ve kaynaklarla insanlar bu sorunu genel olarak nasıl çözüyorlar?**

1. **Yeterince işe yarıyor mu? Yoksa yeni bir şey mi bulmam gerekiyor?**
2. **Benzer problemler neler? Nasıl çözüldüler?**
3. **Bunu yapabilir miyim?**

Ne Tür Bir Sorunu Çözmem Gerekiyor? Nasıl çözerim?

Çözülmesi Gereken Sorun	Kategoriler	Yöntemler
Öğeleri benzerliğe göre gruplandırmak istiyorum. Verilerdeki benzerlikleri (ortaklıkları) bulmak istiyorum	Clustering – Kümeleme	K-means clustering
Eylemler veya öğeler arasındaki ilişkileri keşfetmek istiyorum	Association Rules	Apriori
Sonuç ve girdi değişkenleri arasındaki ilişkiyi belirlemek istiyorum	Regresyon	Linear Regression Logistic Regression
Nesnelere (bilinen) etiketler atamak istiyorum	Classification - Sınıflandırma	Naïve Bayes Decision Trees
Zamansal bir süreçteki yapıyı bulmak istiyorum Zamansal bir sürecin davranışını tahmin etmek istiyorum	Time Series Analysis – Zaman Serisi Analizi	ACF, PACF, ARIMA
Metin verilerimi analiz etmek istiyorum	Text Analysis – Metin Analizi	Regular expressions, Document representation (Bag of Words), TF-IDF

- En popüler, sık kullanılan yöntemler.
- Veri Bilimine yeni başlayanlar için anlaması ve kavraması nispeten kolay yöntemler.
- Çeşitli sektörlerdeki geniş bir problem yelpazesine uygulanabilir.

Zaman Serisi Analizi

Bu derste aşağıdaki konular ele alınmaktadır:

- Zaman Serisi Analizi ve tahminde uygulamaları
- ARMA and ARIMA Modelleri
- Box-Jenkins yönteminin R Kullanılarak Uygulanması
- Zaman Serisi Analizi seçme nedenleri (+) ve dikkat edilecek noktalar (-)

Zaman Serisi Analizi

Satış (Ciro) Tahmini – Sales Forecasting (İleride ne olacak?)

- Yatırımlarını planlamak,
- Yeni ürünleri piyasaya sürmek,
- Ürünleri ne zaman kapatacaklarına veya geri çekeceklerine karar vermek ve benzeri amaçlar.

Önemli Noktalar:

- Çoğu işletme için kritik bir süreç,
- Satış tahmini sürecinin bir kısmı geçmişi incelemektir.
  - Son birkaç ayda ne kadar iyi performans gösterdik
  - Son birkaç yılın aynı döneminde satışlarımız ne kadardı?

Zaman Serisi Analizi

Sonuç olarak,

- Zaman Serisi Analizi, satış tahmini için bilimsel bir metodoloji sağlar.
- Zaman Serisi Analizi, eşit aralıklı zaman birimleri boyunca sıralı verilerin analizidir.
- Zaman Serisi, farklı zaman dilimlerindeki birçok gözlem için bir veya daha fazla değişkene ilişkin verilerin toplandığı temel bir araştırma metodolojisidir.

Zaman Serisi Analizi

Zaman Serisi Analizindeki ANA HEDEFLER şunlardır:

- Zaman serisinin temel yapısını bileşenlerine ayırarak anlamak.
- Matematiksel bir model geliştirin ve ardından geleceği kestirmek

Özellikleri

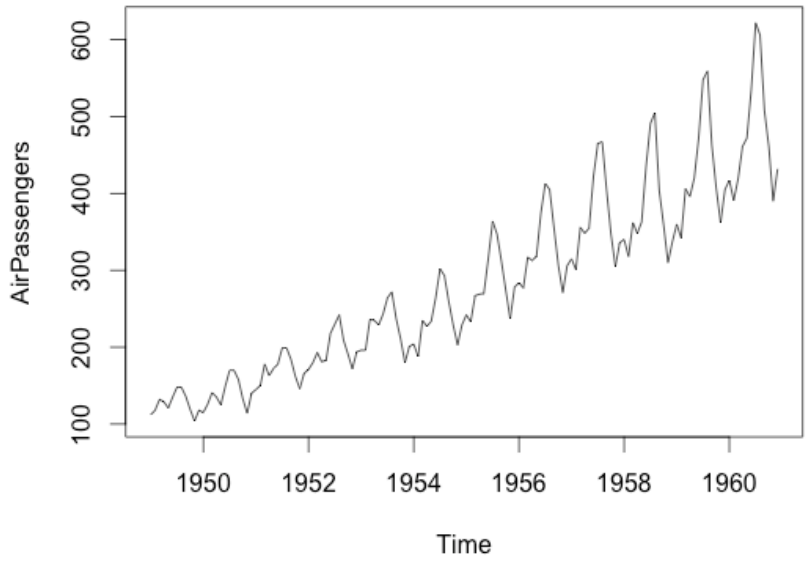
- Zaman periyotları genellikle düzenli aralıklıdır.
- Gözlemler tek değişkenli veya çok değişkenli olabilir.
  - Tek değişkenli zaman serileri zaman içinde yalnızca bir değişkenin ölçüldüğü seriler,
  - Çok değişkenli zaman serileri ise birden fazla değişkenin aynı anda ölçüldüğü serilerdir.
- Verilerin iç yapısı bir eğilimi (**TREND**) , mevsimselliği (**SEASONALITY**), ve döngüleri veya rastgeleliği (**CYCLIC/RANDOM**) belirtebilir.

Zaman Serisi Analizi

- **Zaman Serisi:** Zaman içinde eşit aralıklı değerlerin sıralı dizisi
- **Zaman Serisi Analizi:** Zaman içinde alınan gözlemlerin iç yapısını açıklar
  - **Trend** (Eğilim) – bir zaman serisinde uzun vadeli hareket
  - **Seasonality** (Mevsimsellik) – yılın zamanına bağlıdır
  - **Cycles** (Döngüler) – mevsimsel olmayan tabita bağlı
  - **Random** (Rastgele veya Kaotik)– Serinin diğer bileşenleri hesaba katıldığında geriye kalan rastgele veya kaotik değerle**r.**
- **Hedefler**
  - Zaman serisinin iç yapısını tanımlamak
  - Gelecekteki olayları tahmin etmek
    - Örnek: Satış geçmişine göre önümüzdeki Aralık ayı satışları ne olacak?
    - Ciro Tahmini ???
- **Yöntem: Box-Jenkins (ARMA – Auto Regressive Moving Averages)**

Box-Jenkins Yöntemi:

- Geleceği tahmin etmek için tarihsel geçmiş davranışı modeller



- ARMA (Autoregressive Moving Averages) Otoregresif Hareketli Ortalamalar yöntemini uygular
  - Giriş: Zaman Series
    - *Trend (Eğilim) ve Seasonality (Mevsimsellik) bileşenlerini kullanır*
  - Çıkış: Zaman serisinin beklenen gelecekteki değeri

Use Cases

Forecast - Kestirim:

- Gelecek ayın satışları
- Yarınki hisse senedi fiyatı
- Saatlik güç talebi

- Ekonomi/Finans
- Sosyoloji / Çevre / Meteoroloji
- Epidemiyoloji:
  - SARS 2003 / MERS 2012 / COVID-19
- Tıp: Hipertansiyon

Zaman Serinin Modellenmesi

- Zaman serisini aşağıda gibi modelleyebiliriz

$$Y_t = T_t + S_t + R_t, \qquad t=1,\ldots,n.$$

- T_t: Trend (Eğilim) terimi
  - Hava yolculuğu son birkaç yılda istikrarlı bir şekilde arttı
- S_t: Seasonal (Mevsimsellik) terimi
  - Hava yolculuğu bir yıl boyunca düzenli bir şekilde dalgalanıyor
- R_t: Random (Rastgele) bileşen
  - ARMA ile modellenecek

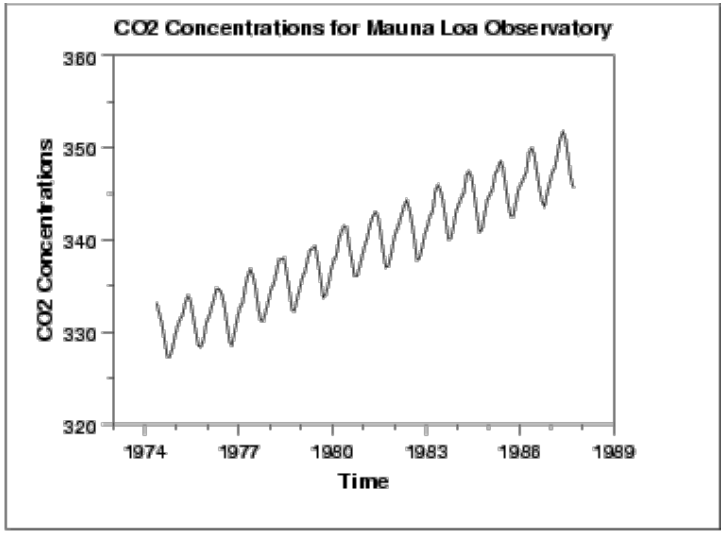
Stationary Sequences – Durağan Dizi

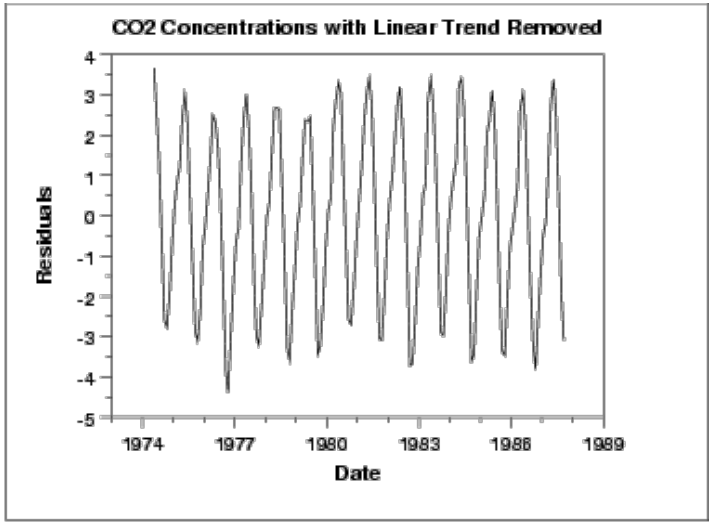
- Box-Jenkins (ARMA) yöntemi rastgele bileşenin durağan bir dizi olduğunu varsayar
  - Sabit Ortalama
  - Sabit varyans
  - Otokorelasyon zamanla değişmez
    - Bir değişkenin kendisiyle farklı zamanlarda sabit korelasyonu.

- Uygulamada durağan bir dizi elde etmek için veriler şu şekilde olmalıdır:
  - Trend'den arındırılmış
  - Seasonally (mevsimsellikten) arındırılmış

De-trending – Trend'den Arındırma

- Bu örnekte doğrusal bir eğilim görüyoruz, dolayısıyla doğrusal bir modele uyuyoruz
  - $T_t = m \cdot t + b$
- Bu durumda trend'den arındırılmış (de-trended) seri
  - $Y^1_t = Y_t - T_t$
- Bazı durumlarda doğrusal olmayan bir modele uymak gerekebilir
  - Quadratic – ikinci dereceden
  - Exponential – üstel

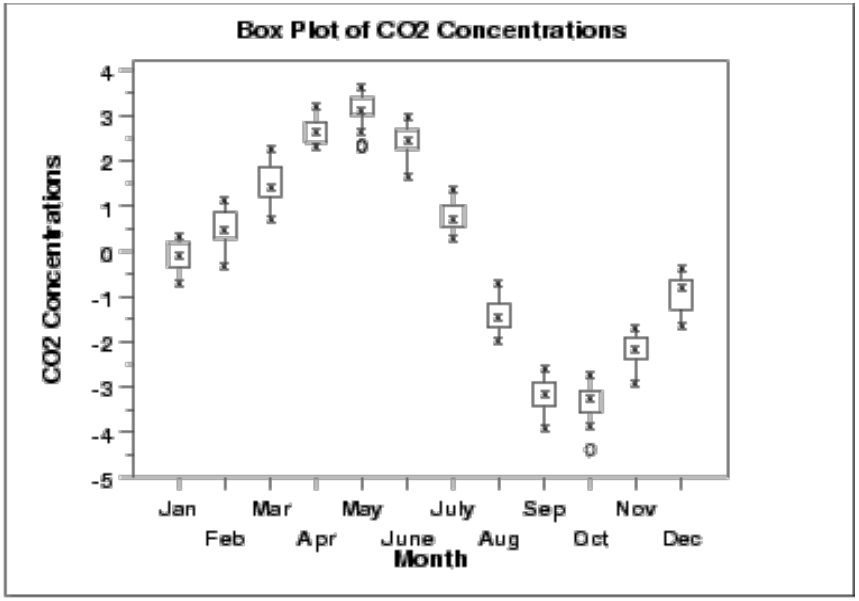




İlk farklar az çok sabitse Doğrusal Trend Modeli kullanın
[(y2-y1) = (y3-y2) = ……. = (yn-yn-1)]
İkinci farklar az çok sabitse İkinci Dereceden Trend Modelini kullanın.
[(y3-y2) – (y2-y1) = ……= (yn-yn-1)-(yn-1-yn-2)]
Yüzde farkları daha fazla veya sabitse Üstel Trend Modeli kullanın.
[((y2-y1) /y1) * %100 = …….((yn-yn-1)/yn-1) * %100

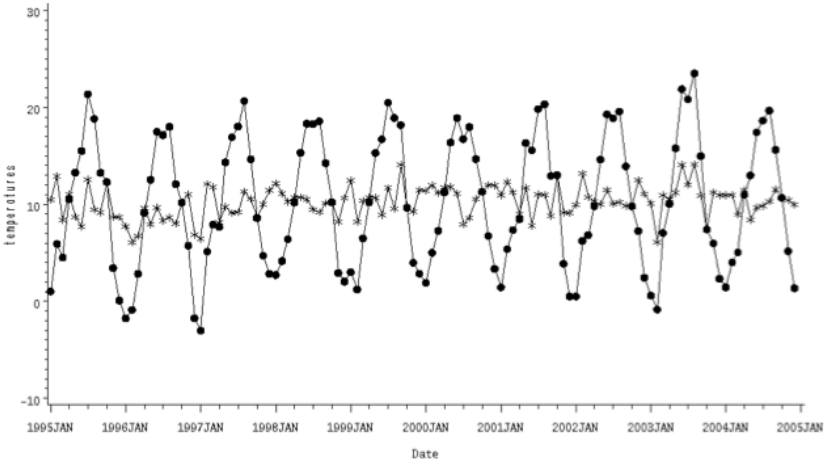
Seasonal Adjustment – Mevsimsellikten Arındırma

- Trend'den arındırılmış serilerin grafiğinin çizilmesi mevsimleri tanımlar
  - CO2 konsantrasyonu için dönemi, ay seviyesinde değişiklikle birlikte bir yıl olarak modelleyebiliriz.



- Basit anlık ayarlama: Birkaç yıllık verileri alın, her ay için ortalama değeri hesaplayın ve bunu Y^1_t'den çıkarın

$$Y^2_t = Y^1_t - S_t$$



ARMA(p, q) Model – Auto Regressive Moving Averages

$$Y_t = \delta + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \ldots + \theta_q \epsilon_{t-q}$$

- En basit Box-Jenkins Modeli
  - Y_t trend'den ve mevsimsellikten arındırılmış
- İki süreç modelinin kombinasyonu
  - **Autoregressive**: Y_t son p değerlerinin lineer kombinasyonudur.
  - **Moving average**: Y_t sabit bir değer ve buna ilave olarak son q zaman değerleri (gecikmeler) üzerindeki sönümlenmiş beyaz gürültü sürecinin etkileridir

– AR (Otomatik Regresyon) terimlerinin sayısı (p): Otoregresyon. Bir gözlem ile bazı gecikmeli gözlemler arasındaki bağımlı ilişkiyi kullanan bir model.
p: Modeldeki bağımlı değişkenin gecikme gözlemlerinin sayısı, gecikme sırası olarak da adlandırılır.
– MA (Hareketli Ortalama) terimlerinin sayısı (q): Hareketli Ortalama. Bir gözlem ile bağımlılık arasındaki gecikme gözlemlerine uygulanan hareketli bir ortalama modelden kalan bir hata arasındaki bağımlılığı kullanan bir model.
q: Hareketli ortalama sırası olarak da adlandırılan hareketli ortalama penceresinin boyutu.

ARIMA(p, d, q) Model-Auto Regressive Integrated Moving Averages
Otoregresif Entegre Hareketli Ortalamalar

- ARIMA, ARMA modeline bir fark terimi (differencing term) olan d'yi ekler:
  - Otoregresif Entegre Hareketli Ortalamalar
  - Modelin bir parçası olarak trend'den ayrıştırmayı (de-trending) içerir
    - doğrusal eğilim d=1 ile kaldırılabilir
    - ikinci dereceden eğilim için *d*=2
    - ve benzeri daha yüksek dereceli eğilimler için

- Sezonluk olmayan genel model ARIMA (*p, d, q*) olarak bilinir:
  - *p:* Otoregresif terimlerin sayısı
  - *d:* Farklılaşmaların sayısı
  - *q:* Hareketli ortalama penceresinin boyutu

p: Modeldeki bağımlı değişkenin gecikme gözlemlerinin sayısı, gecikme sırası olarak da adlandırılır.
q: Hareketli ortalama sırası olarak da adlandırılan hareketli ortalama penceresinin boyutu.
d: Ham gözlemlerin farklılaşma sayısı, aynı zamanda farklılaşma derecesi de denir.

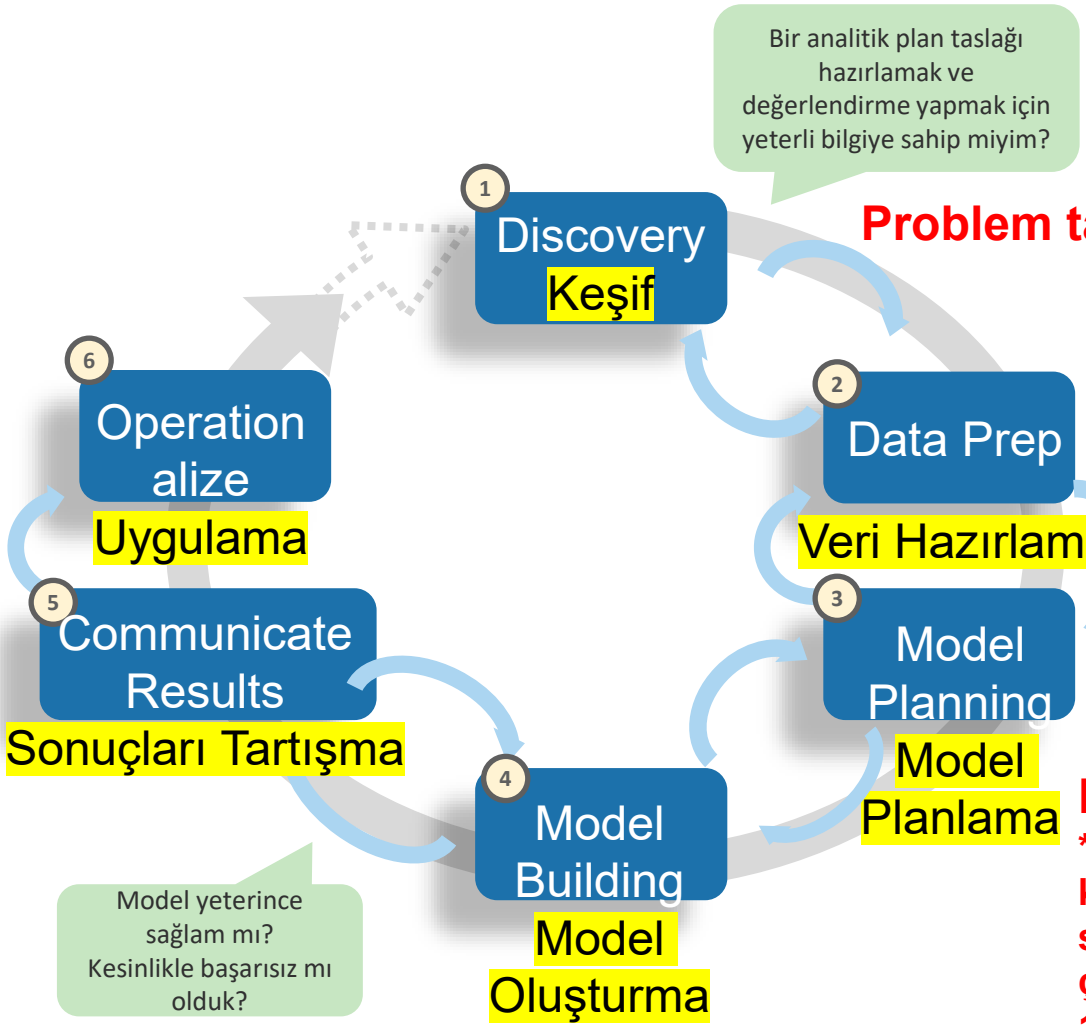
ACF & PACF

- Auto Correlation Function (ACF)
  - Zaman serisinin değerlerinin kendisiyle korelasyonu
  - Otokorelasyon "devam ediyor"
  - MA modelinin sırasını (q) belirlemeye yardımcı olur
    - ACF nerede sıfıra gider?

- Partial Auto Correlation Function (PACF)
  - Önceki terimlerin doğrusal bağımlılığı kaldırıldıktan sonra hesaplanan bir otokorelasyon
  - AR modelinin sırasını (p) belirlemeye yardımcı olur
    - PACF nerede sıfıra gidiyor?

.

Data Analytics Lifecycle – Veri Analitiği Yaşam Döngüsü



Kullanılacak yöntemi belirleme

Modeli planlama:

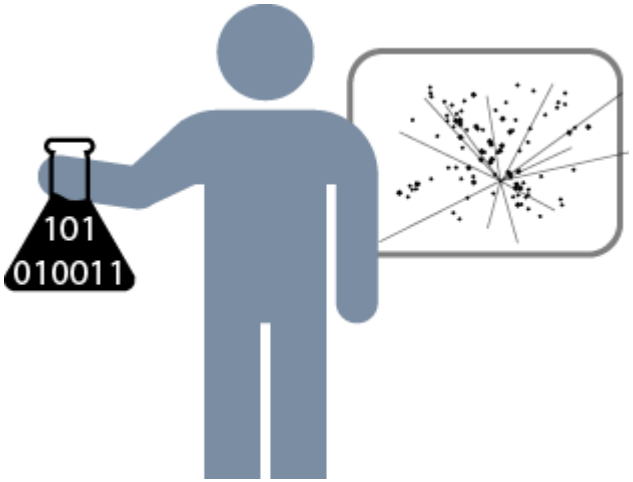
*** Sahip olduğum veri ve kaynaklarla insanlar bu sorunu genel olarak nasıl çözüyorlar?**

1. **Yeterince işe yarıyor mu? Yoksa yeni bir şey mi bulmam gerekiyor?**
2. **Benzer problemler neler? Nasıl çözüldüler?**
3. **Bunu yapabilir miyim?**

Model Seçimi

- Veri Bilimcisi verilere dayanarak p, d ve q'yu seçer
  - Alan bilgisi, modelleme deneyimi ve birkaç yineleme gerekebilir
  - Mümkün olduğunda basit bir model kullanın
    - AR modeli (q = 0)
    - MA modeli (p = 0)

- ACF ve PACF kullanılarak birden fazla modelin oluşturulması ve karşılaştırılması gerekir



Zaman Serisi Analizi – Seçme Nedenleri (+) ve Dikkat Edilecek Noktalar (-)

Seçme Nedenleri (+)	Dikkat edilecek noktalar (-)
Minimum veri toplama Sadece serinin kendisini toplayın	Kestirim yalnızca geçmiş performansa dayalı Açıklayıcı değer yok "Ya şöyle olursa" senaryoları yapılamıyor Stres testi yapılamıyor
Gecikmeli zaman serilerinin doğal otokorelasyonunu ele alacak şekilde tasarlanmıştır	Uygun parametreleri seçmek bir sanat gerektirir
Trendleri ve mevsimselliği hesaba katar	Yalnızca kısa vadeli tahminler için uygundur

.

Time Series Analysis with R

Important R functions and commands we will be using are listed here.

- The function "*ts*" is used to create time series objects
  - **mydata<- ts(mydata,start=c(1999,1),frequency=12)**
- Visualize data
  - **plot(mydata)**
- De-trend using differencing
  - **diff(mydata)**
- Examine ACF and PACF
  - **acf(mydata)**: It computes and plots estimates of the autocorrelations
  - **pacf(mydata)**: It computes and plots estimates of the partial autocorrelations

Other Useful R Functions in Time Series Analysis

- **ar()**: Fit an autoregressive time series model to the data
- **arima()**: Fit an ARIMA model
- **predict()**: Makes predictions
  - *"predict"* is a generic function for predictions from the results of various model fitting functions. The function invokes particular methods which depend on the *class* of the first argument
- **arima.sim()**: Simulate a time series from an ARIMA model
- **decompose()**: Decompose a time series into seasonal, trend and irregular components using moving averages
  - Deals with additive or multiplicative seasonal component
- **stl():** Decompose a time series into seasonal, trend and irregular components using loess

Useful ARIMA model in Python for Time Series Analysis

statsmodels.tsa.arima.model.ARIMA

```
class statsmodels.tsa.arima.model.ARIMA(
    endog,
    exog=None,
    order=(0, 0, 0),
    seasonal_order=(0, 0, 0, 0),
    trend=None,
    enforce_stationarity=True,
    enforce_invertibility=True,
    concentrate_scale=False,
    trend_offset=1,
    dates=None,
    freq=None,
    missing='none',
    validate_specification=True
) ¶                                                          [source]
```

Autoregressive Integrated Moving Average (ARIMA) model, and extensions

This model is the basic interface for ARIMA-type models, including those with exogenous regressors and those with seasonal components. The most general form of the model is SARIMAX(p, d, q)x(P, D, Q, s). It also allows all specialized cases, including

- autoregressive models: AR(p)
- moving average models: MA(q)
- mixed autoregressive moving average models: ARMA(p, q)
- integration models: ARIMA(p, d, q)

Yapay Zeka 1:

İleri Analik – Yöntemler
Anomaly Detection – Anomali Tespiti

Doç. Dr. Taner Arsan, Ph.Dc. H. Fuat Alsan, Ph.Dc. Sena Kılınç

Bilgisayar Mühendisliği Bölümü

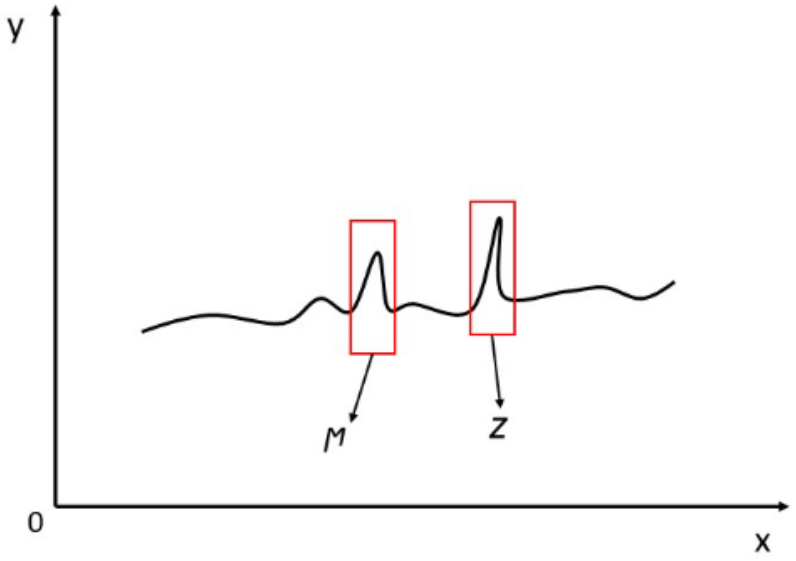
Mühendislik ve Doğa Bilimleri Fakültesi - Kadir Has Üniversitesi

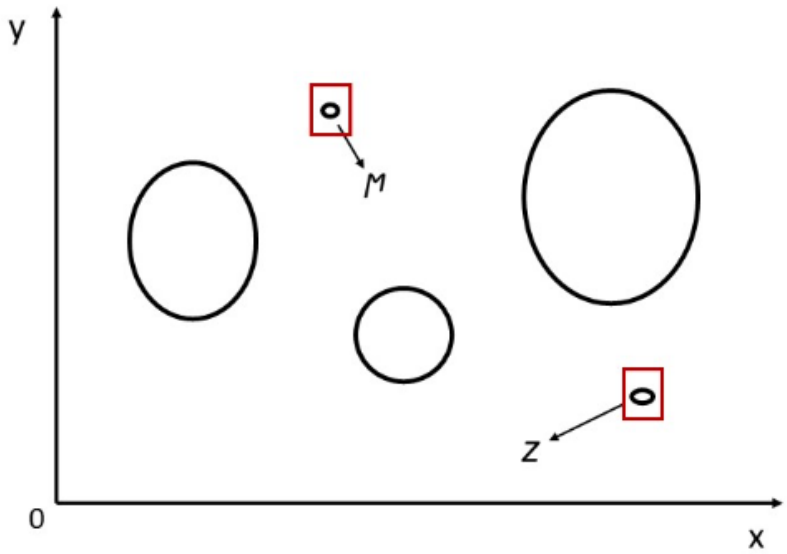
.

Anomali Tespiti

Anomali Tespiti *(*Anomaly Detection); bir veri kümesindeki aykırı değerlerin, farklı öğelerin ve olayların belirlenmesi olarak ifade edilmektedir.

Bu değer/öğe/durumların literatür karşılığı ise ***Outlier*** (Aykırı Değer) ya da kimi yazılım dili kullanımlarında sürekli karşımıza çıkan ***Exception*** (Exception Handling – İstisnai Durum) şeklinde adlandırılmaktadır.



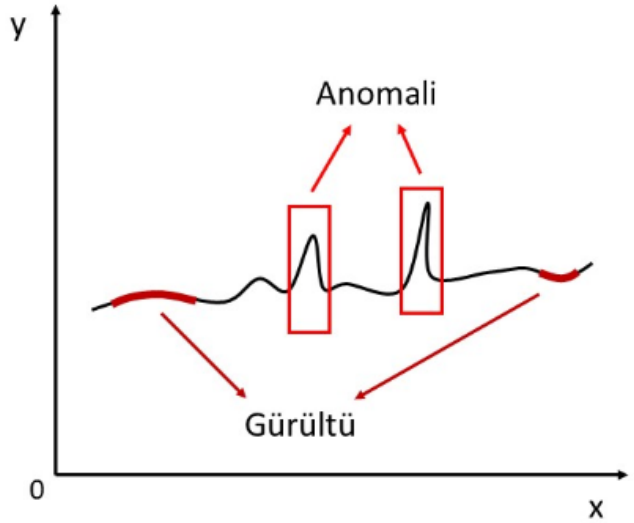


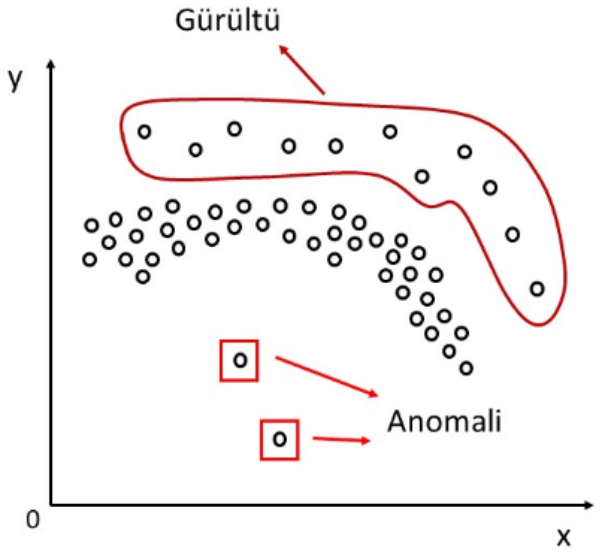
Gürültü (Noise) ve Anomali (Anomaly)

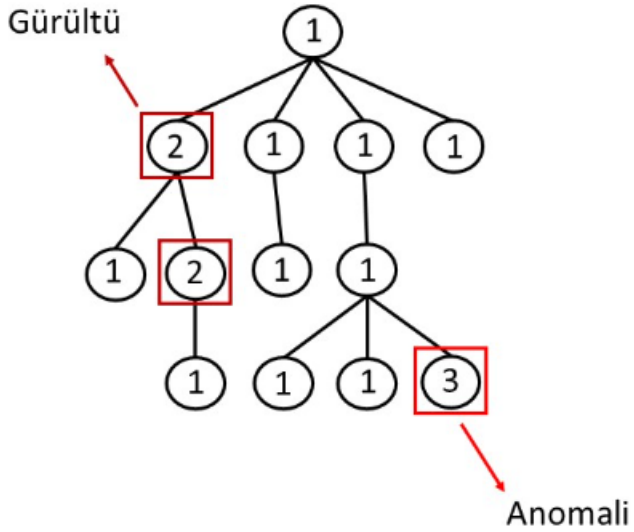
Anomali (Anomaly); bir veri kümesinde elde edilen normlardan oluşan aykırı değerler, farklı öğeler ve olaylar olarak ifade edilmektedir.

Gürültü (Noise); veri girişi, veri işlenmesi ya da veri toplanması esnasından oluşan/saptanan hatalar olarak ifade edilmektedir. (Yanlış giriş, yanlış etiketlenme vb. durumlar.)

Ölçüm/Analiz = **Temiz Veri** + **Gürültü**







Uygulama Alanları

Internet Teknolojileri (Siber Güvenlik, E-Ticaret)

- Ağ Davranışı Anomalileri (*Network Behavior Anomalies*)
- Dalgalı Band Anomalileri (*Wavy Band Anomalies*)
- Web Sitesi Trafik Anomalileri (*Website Traffic Anomalies*)
- Web Satış Performansı Anomalileri (*Web Sales Performance Anomalies*)
- Kullanıcı Profili Anomalileri / Güvelik İhlali Gerçekleşmiş (*User Profile Anomalies*)

Sahtekarlık Girişimleri (Fraud Detection)

- Müşteri Davranış Anomalileri / E-Ticaret (*Customer Behavior Anomalies*)
- Kredi Kartı, Online Bankacılık Anomalileri (*Credit Card, Online Banking Anomalies*)

Tıbbı-Medikal Problemler

- Hasta Biyolojik Durum Anomalileri (*Patient Biological Status Anomalies*)
- Klinik Ortam Anomalileri (*Clinical Environment Anomalies*)

Uygulama Alanları

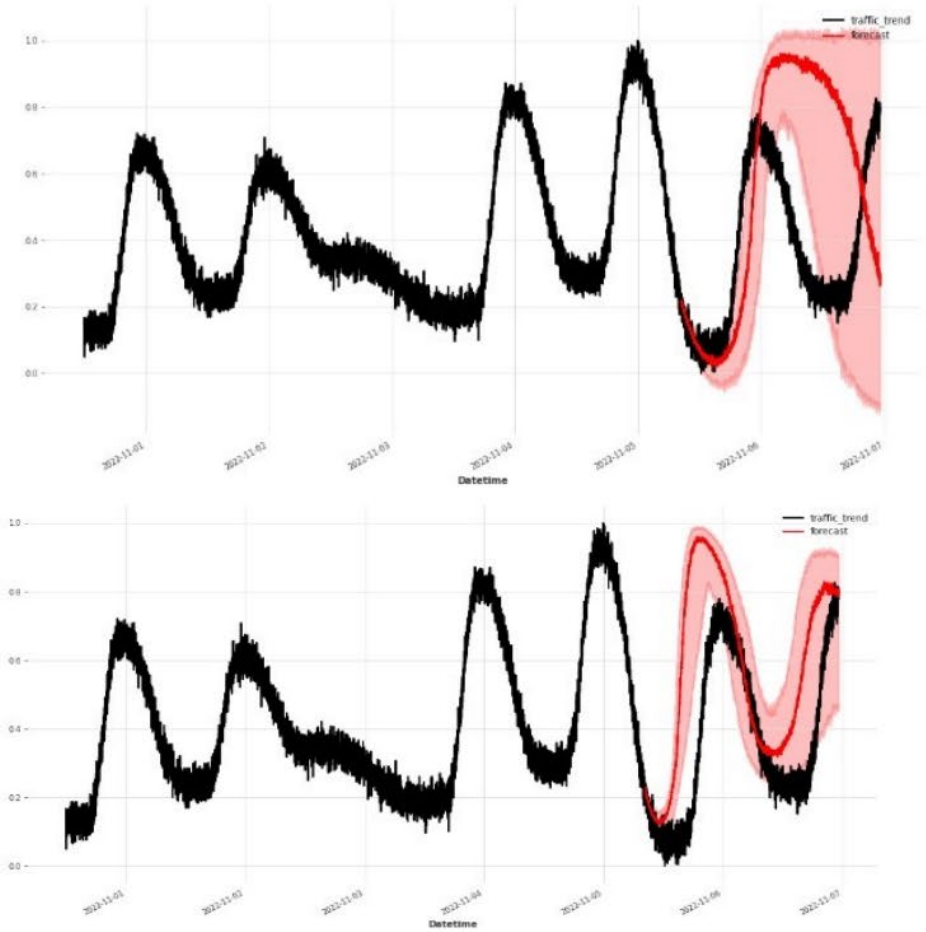
Finansal Problemler

- Hisse Senedi Ticareti Anomalileri (*Stock Trading Anomalies*)
- Borsa Alım-Satım Anomalileri (*Exchange Trading Anomalies*)

Endüstriyel Problemler (Üretim ve ARGE)

- Endüstriyel Otomasyon İşlem Anomalileri (*Industrial Automation Process Anomalies*)
- Üretim Anomalileri (*Production Anomalies*)
- Arge Analiz Anomalileri (*R&D Analysis Anomalies*)
- Akıllı Ev Enerji Tüketim Anomalileri (*Smart Home Energy Consumption Anomalies*)

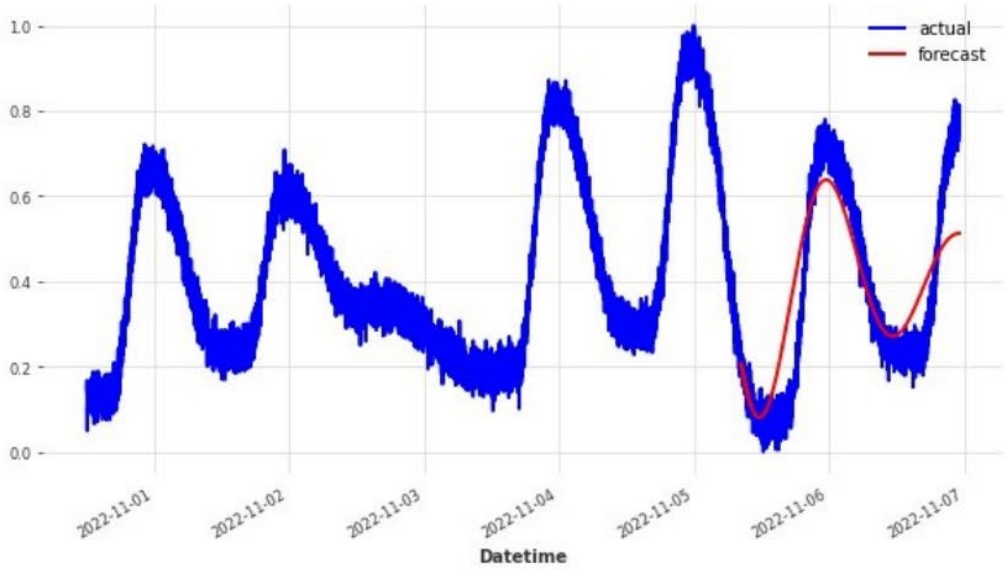
Prediction



- Here are the predictions with different parameters using the ARIMA model.
- The result found by training 75% of the seven-day data

.

Prediction (Cont.)



- The estimation result here was found with the Regression Model.
- The result found by training 75% of the seven-day data

Artificial Intelligence I: Introduction to Data Science and Machine Learning

Chapter 4: Time Series & Anomaly Detection

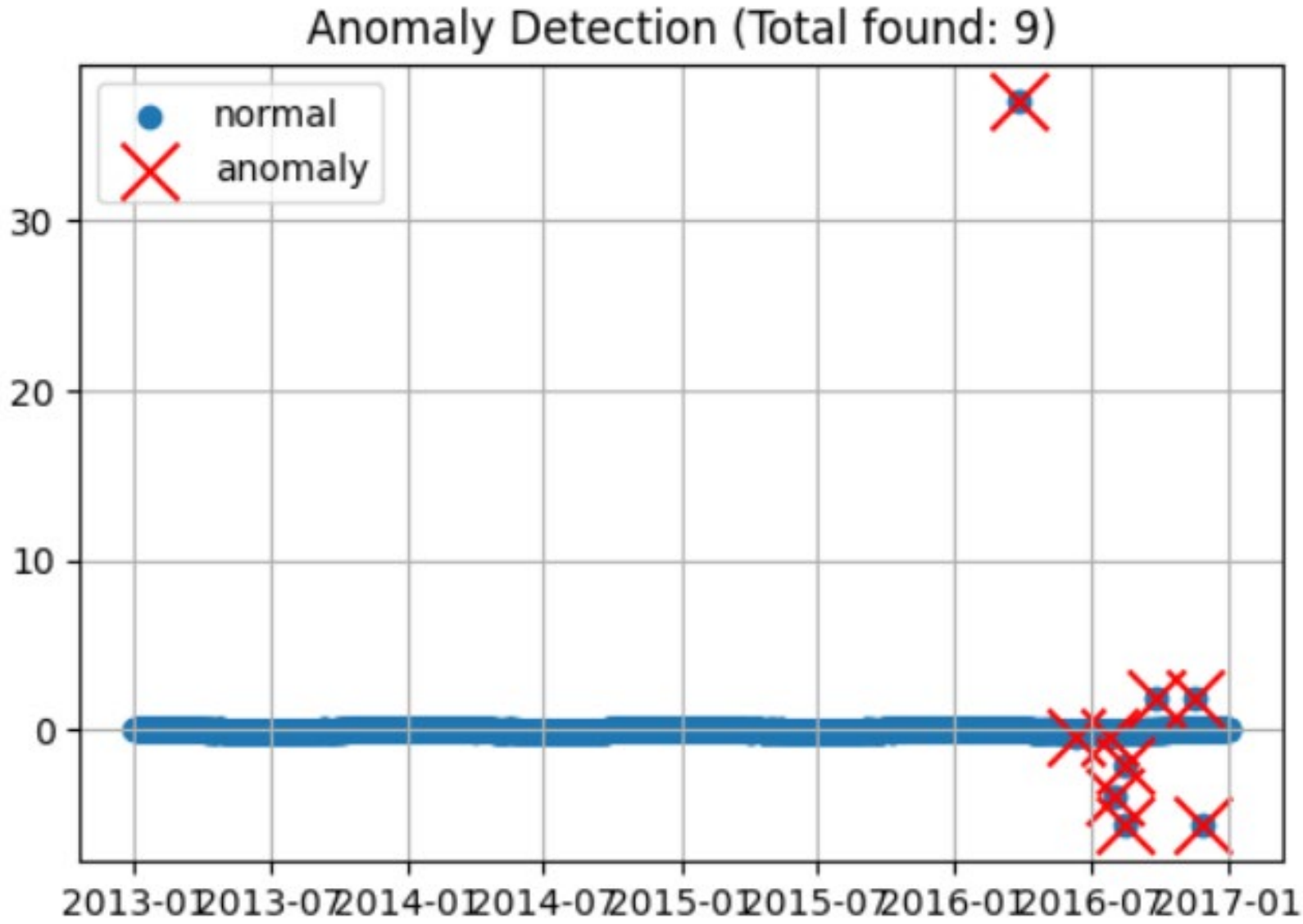
Online Retail Fraud Detection

- https://archive.ics.uci.edu/dataset/352/online+retail

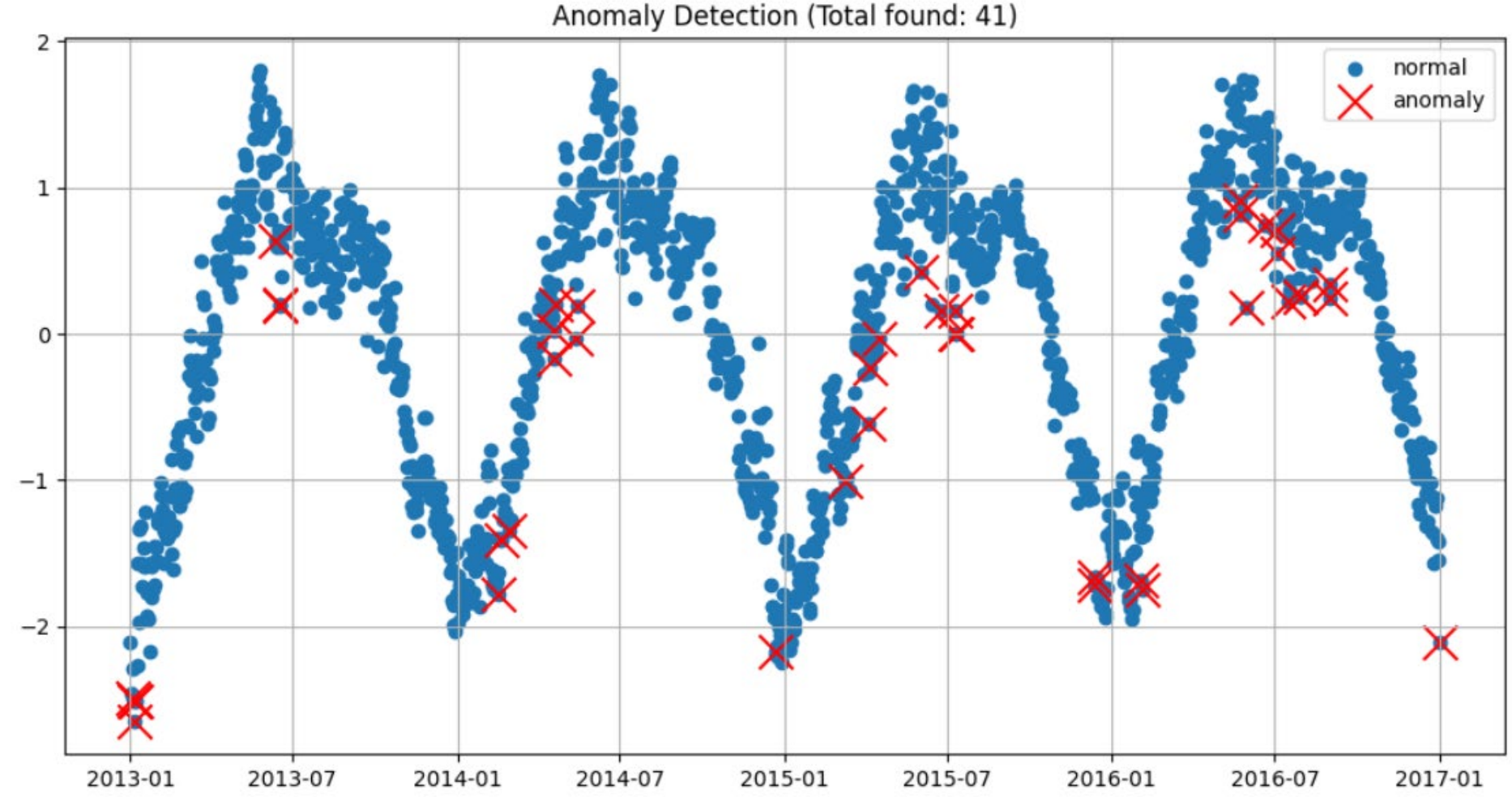
In [1]:

```
import numpy as np
import pandas as pd

import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
```

.


Anomaly Detection (Total found: 9)
normal
anomaly
30
20
10
0
2013-01
2013-07
2014-01
2014-07
2015-01
2015-07
2016-01
2016-07
2017-01

.


Anomaly Detection (Total found: 41)
normal
anomaly
2
1
0
−1
−2
2013-01
2013-07
2014-01
2014-07
2015-01
2015-07
2016-01
2016-07
2017-01

.

Derin Öğrenme ile Anomali Tespiti: Knowledge Distillation

- Anomalileri tespit etmek için model tahminlerini gerçek verilerle karşılaştırırız (artık analiz)
- Yeniden yapılandırma hatası ve anormallik: model tahmin hataları anormallik olarak karıştırılmamalıdır
- Modellerin sağlamlığının ve verimliliğinin arttırılması: Birden fazla model (farklı ufuklar) bir bütün olarak birlikte çalışır
- Kullanım verimliliği: Transfer Learning
- Eğitimde verimlilik: Knowledge Distillation

Anomali Tespiti – Seçme Nedenleri (+) ve Dikkat Edilecek Noktalar (-)

Seçme Nedenleri (+)	Dikkat edilecek noktalar (-)
Veriseti üzerinde uyarlanacak yapıda özelliklerin normalleştirilmesi gerekli değildir, (*Geniş zaman bazlı planlama gerektirmez.*)	Analiz sonuçlarnıı görselleştirmek karmaşıktır; kontrol noktaları iyi planlanıp probleme yönelik uygun görselleştirmelerde bulunulmalıdır,
Veriseti üzerindeki verilere ait özelliklerin dağılımının tahmin edilemediği durumlarda daha etkili bir yöntemdir, (Çıkarım oluşturur.)	Kontrol noktalarına ya da vakaya uygun olarak optimize edilmezse uygulama ve uyarlama uzun olabilir. Bunun sonucu olarak daha yüksek hesaplama gücü ve hesaplama zamanı gerektirebilir,
Birkaç parametre kullanılması; modelin optimize edilmesi ve uyarlanabilirliğinin kolay olmasını sağlar.	Sonuç güvenilirliği birçok durumda belirsizdir. (Uzman görüşünde dahi ihtimal yer almaktadır.)

.

Teşekkürler...

Doç. Dr. Taner ARSAN
arsan@khas.edu.tr
0532 651 65 87