Anomaly Detection

Data Science & Business Analytics

Lesson 5

Tiago Cabo

Lesson Plan

- Anomaly detection
- Novelty detection
- Time Series
- Recommender systems
- Test time

Lesson goals

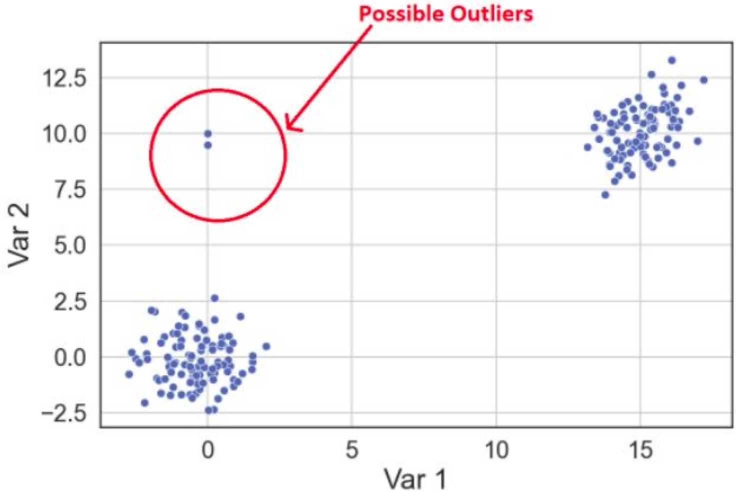
- Understand the difference between anomaly detection and novelty detection
- Be able to work with time series data
- Have positive grade in the final test.

Anomaly detection

Anomaly detection is a subfield in machine learning that develops specific models for anomaly detection. You at this point must be think, but haven't already applied methods to remove outliers? Why do we need specific models for this?

This field is important because, because in specific use cases outliers are not something we might want to eliminate but is something we might want to find. Typical models:

- Isolation forest
- Single class SVM

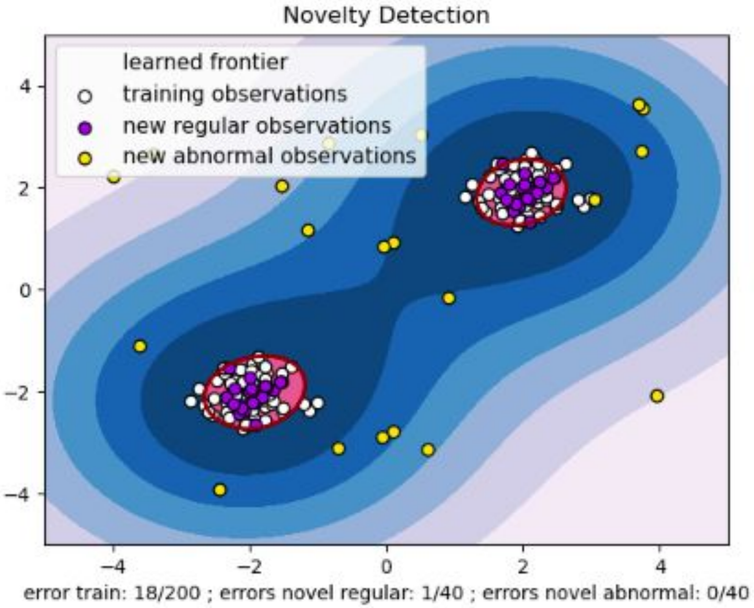


We want the Multivariate Methods to detect the 2 major outliers — Image by Author

Novelty detection

This field is very similar to outlier detection, but there is a conceptual difference. While in the outlier detection algorithm we want to find data that does not belong to the data. In novelty detection we want to find new data points that didn't exist in the trained observations.

The models used are the same so depending on the approach you have one or the other.



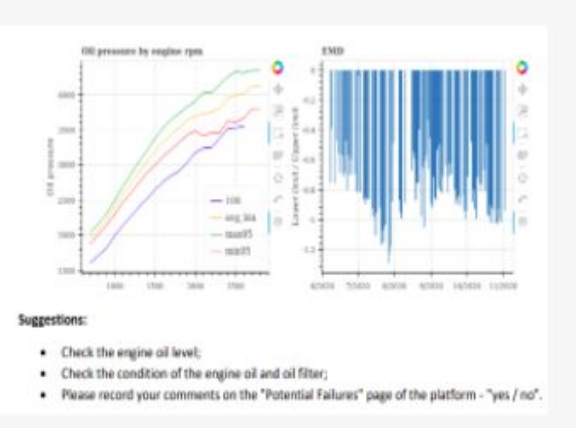
Stratio Use Case

Article

https://medium.com/stratio/stratio-ai-cortex-a-self-supervised-deep-learning-framework-for-anomaly-detection-in-time-series-f2c3c89a8224

In this case you can see how based on the prediction error it was possible to find bus malfunctions.

In the image of the right, you can see how a simple frequency plot was designed in order to find when the oil was not inside the working boundaries.



Suggestions:

- Check the engine oil level;
- Check the condition of the engine oil and oil filter;
- Please record your comments on the "Potential Failures" page of the platform - "yes / no".

Time Series

Data Science & Business Analytics

Lesson 5

Tiago Cabo

Time series

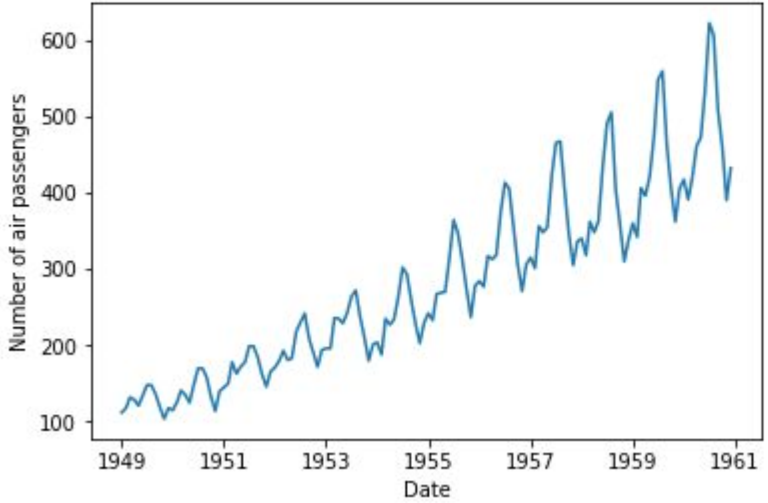
Time series data normally is treated as a regression problem.

Most problems have one variable but they can have multiple.

However, the same data, can be used in standard classification models. For that features need to be extracted from the dataset. For example, average by week.

Let’s exercise using the air passenger dataset.

- Plot the following graph.

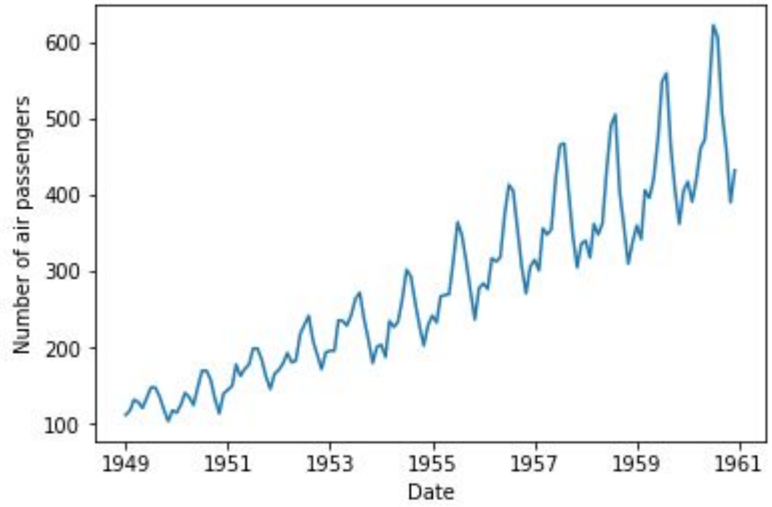


Time series - components

What can you notice from the plot?

- It has a trend
- It has some kind of period

What can we do with this information?



Let’s decompose

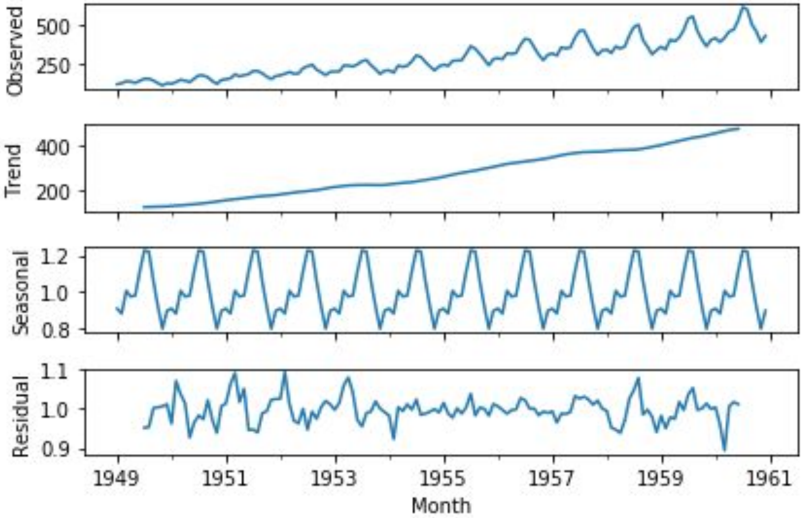
From we saw in the previous slide, we can separate the data in different components. The most common model is the addictive, where

data = T + S + Residual

This can be done using

```
import statsmodels.api as sm

decompose = sm.tsa.seasonal_decompose(X,
model="addictive", freq=12)
```



ARIMA Model

ARIMA is a kind of autoregressive model where the next value in the timeseries is a linear combination of the previous ones.

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \ldots + \phi_p x_{t-p} + w_t$$

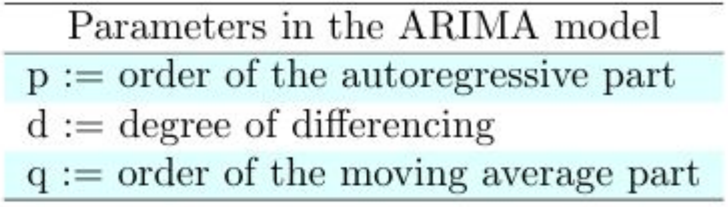
Moving-average (MV) model

A very similar model called moving-average is defined as followed

$$x_t = w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \ldots + \theta_q w_{t-q}$$

In this case wt element represent the previous error terms

ARIMA results from the combination of the AR and MA models. Also a differentiation method is applied in order to remove non-stationary elements of a series.



Parameters in the ARIMA model
p := order of the autoregressive part
d := degree of differencing
q := order of the moving average part

```
from statsmodels.tsa.arima.model import ARIMA

model = ARIMA(differenced, order=(7,0,1))
```