Data Analysis for Business Analytics

CIS621

Week 7 - Time-Series Analysis

Lecture Goals

1. Understand the main concepts of time-series analysis.
2. Understand the unique challenges of time-series models.
3. Understand ways to build time-series predictive models.

Time-Series Data

- Time-series data captures a value of interest over time.
- Values may be recorded at regular intervals or non-regular intervals.
- Examples:
  - Quarterly GDP
  - Monthly revenue
  - Daily electricity consumption
  - Heart rate reported via a smart watch

What Makes Time-Series Problems Unique?

- We cannot break the sequence. That is, we have to make sure to only analyze data that simulates when it was collected.
- For example, we want to build a predictive model to assess the likelihood a sales agent will close a deal.
- A feature in our model could be the agent’s historical close rate.
- We cannot simply slap on her close rate as it stands today! This is called feature leakage.

WRONG!!!!

user_id	acquisition_date	closed	close_rate	channel	cost
112	3/2/22	1	0.57	social	23.09
147	4/17/22	0	0.57	social	26.00
178	5/1/22	1	0.57	display	17.89

Close rate on day of training (5/15/22): 0.57

Right!!!

user_id	acquisition_date	closed	close_rate	channel	cost
112	3/2/22	1	0.47	social	23.09
147	4/17/22	0	0.51	social	26.00
178	5/1/22	1	0.55	display	17.89

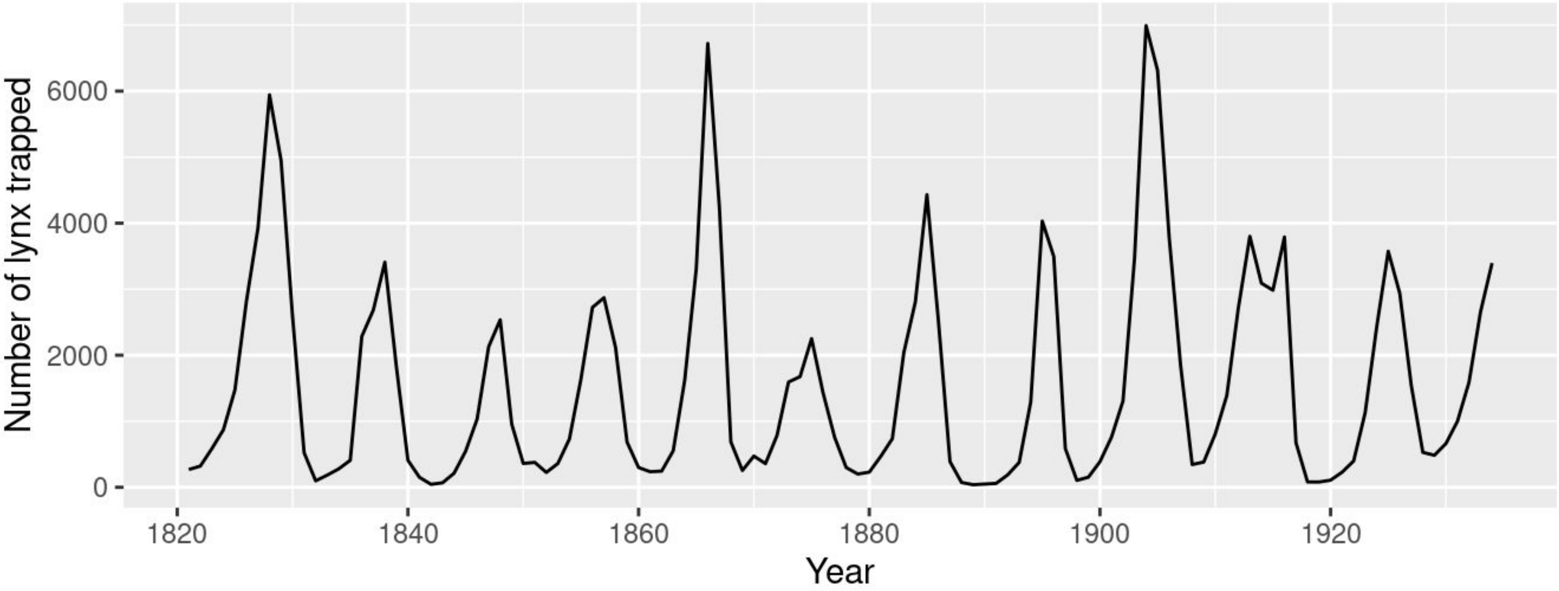
Close rate on day of training (5/15/22): 0.57

Qualities of a Time-Series

- Trend - is the series moving upward or downward
- Seasonality - are there timed repeated patterns
- Cycles - are there longer-range, self-reinforcing patterns
- Randomness - not every movement has a “cause”

Time-Series with Seasonality and Cycles

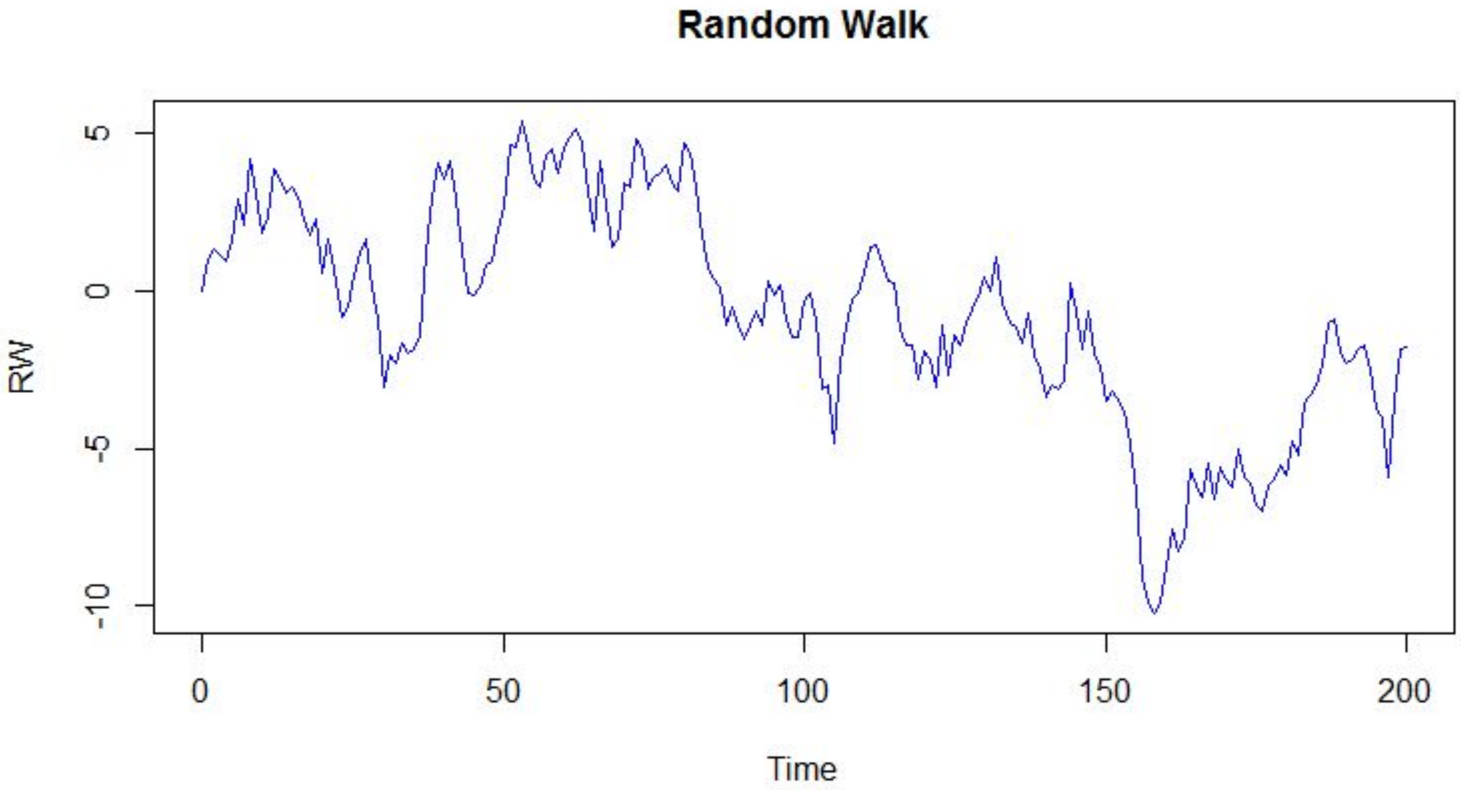
Image Take From: https://robjhyndman.com/hyndsight/cyclicts/



Demo: Decomposing a time-series into components

Random Walks

- A succession of random steps.
- Humans find patterns where none exist. The on the next page chart is just a random walk - there is no underlying data-generating process!
- Some people hypothesize that the stock market follows a random walk, more or less.
- Image Credit for next slide:
  - https://financetrain.com/simulate-random-walk-rw-in-r


Random Walk
RW
5
0
-5
-10
Time
0
50
100
150
200

Stationarity

- In a stationary dataset, values do not depend on time.
- In other words, stationary datasets do not have trend or seasonality.
- Example:
  - We are analyzing data from a call center.
  - We can only properly compare counts of call topics over time if the data is stationary.
  - That is, we might be getting more calls about locked accounts simply because we have more calls in total!

How Can we Make a Series Stationary?

- A classic way is to take the *difference.*

sales	first_difference	second_difference
10	-	-
13	13 - 10 = 3	-
19	19 - 13 = 6	6 - 3 = 3
27	27 - 19 = 8	8 -6 = 2
37	37 - 27 = 10	10 -8 = 2

How Can we Make a Series Stationary?

- We might also employ seasonal differencing.
- For example, we subtract last January’s value from this January’s value.

How Can we Make a Series Stationary?

- We could also perform a log transformation, which may help in some cases. This basically 1) spreads out smaller values and 2) “moves in” larger values.
- In some cases, a power transformation to make a series more normal could help with stationarity.

How Can we Tell If a Series is Stationary?

- In addition to the eyeball test, we can use the Augmented Dickey-Fuller test.

Baseline Model: Moving Average

- The simplest model is the moving average. The main parameter is the window.
- For example, we predict the next value as the average of the past three values.

10
15
17
14

Baseline Model: Moving Average

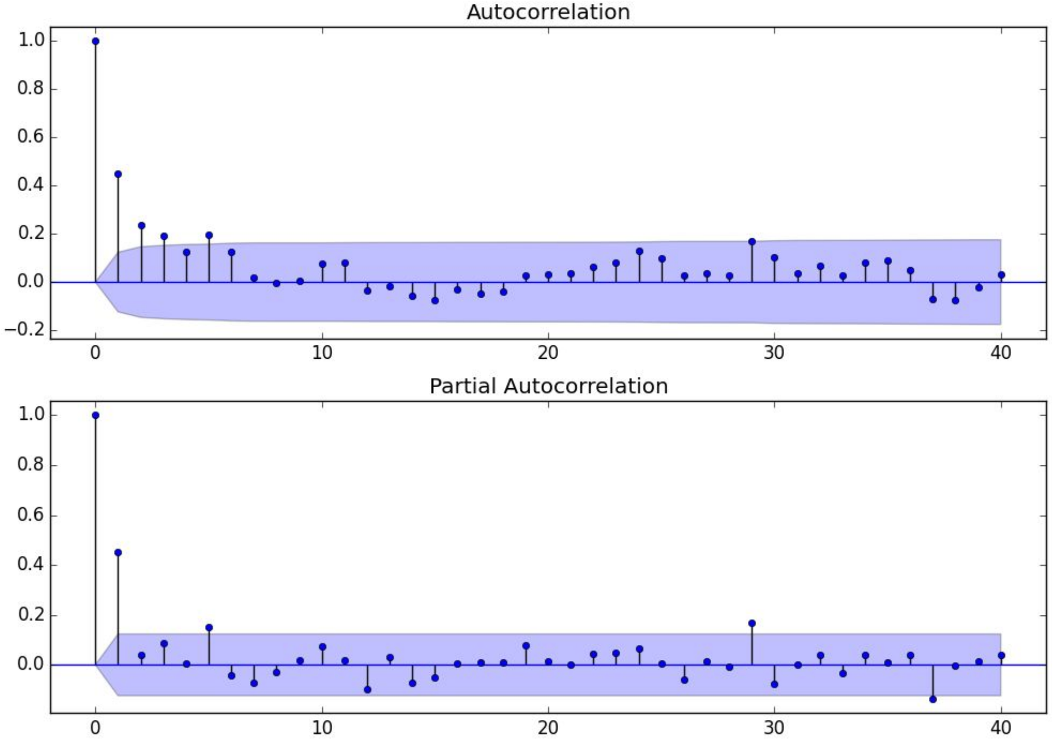
- We can also weight observations.

value	weight
10	1
15	2
17	3
15.1	

ACF and PACF Plots

- ACF: shows correlation coefficients between a value and past values of itself.
- PACF: shows correlations with a value and past values of itself after removing the effects already explained by previous lags.

Images Taken From: https://stats.stackexchange.com/questions/134487/analyse-acf-and-pacf-plots



The Classic Time-Series Model: ARIMA

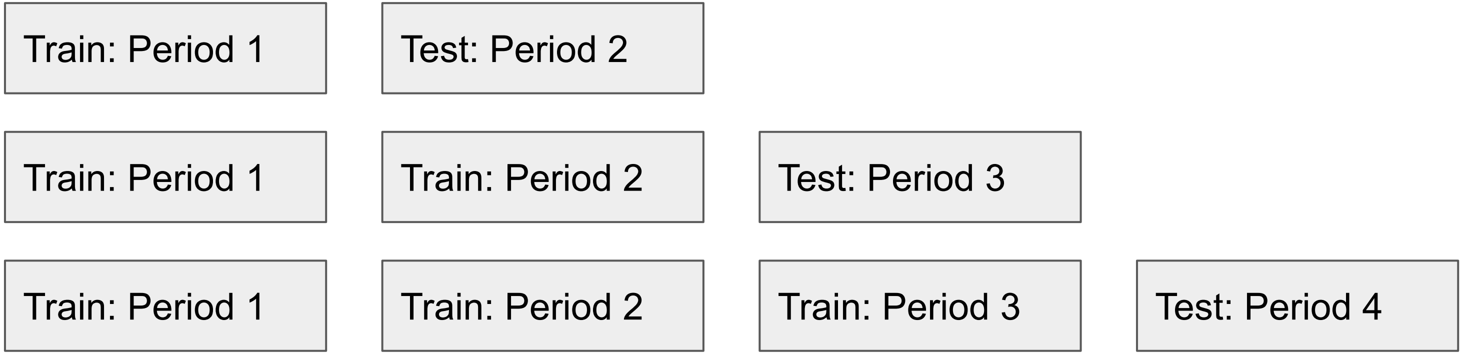
- The classic time series model is an ARIMA, which stands for autoregressive integrated moving average. It’s a bit of a world salad that basically means we difference the series to make it stationary and then build model components based on lagged values.
- An ARIMA model has three parameters that control its three components:
  - P - number of lags of a value to include in the autoregressive component
  - D - number of differences to take to make the series stationary
  - Q - number of past forecast errors to include for the moving average component
- We can use Python to automatically find the best values for each.

For the Future: Cross Validation

- Cross validation is a core concept in machine learning. Simply, cross validation splits our training data into new training and testing sets in a rotating fashion.
- Three-fold cross validation works like the following, with each fold getting assigned a random set of observations:
  - Fit 1: folds 1 and 2 are used for training, and fold 3 is used for testing.
  - Fit 2: folds 2 and 3 are used for training, and fold 1 is used for testing.
  - Fit 3: folds 1 and 3 are used for training, and fold 2 is used for testing.

For the Future: Cross Validation for Time Series

- With cross validation for a time-series problem, we cannot perform random assignment. We have to keep the lineage of our data. Otherwise, we might test on data that came before our training data!



Train: Period 1	Test: Period 2		
Train: Period 1	Train: Period 2	Test: Period 3	
Train: Period 1	Train: Period 2	Train: Period 3	Test: Period 4

Prophet

- A popular univariate time-series model is called Prophet, developed by Facebook. It essentially builds a generalized linear model (GLM) under the hood.

LSTM

- Recurrent neural networks are also useful for time-series prediction tasks. A popular choice is an LSTM, otherwise known as long-short term memory.

XGBoost

- We can also use any old machine learning model on a time-series problem.
- Some people default to ARIMA, but you can use any machine learning model in your arsenal…your target is just a future time-series value.
- For example, on 8/8/22 we want to predict the value for electricity consumption on 8/9/22.
- We set our target to be the electricity consumption on 8/9/22.
- To make the prediction, *we only use data we had on 8/8/22 or earlier.*

Linear (Regularized) Regression

- We could also use a linear regression model. In practice, you will almost never use plain linear regression.
- Rather, you will use a regularized version, such as Ridge or Lasso. Depending on which flavor, these models shrink coefficients towards or to zero, effectively eliminating features the model deems unimportant. This helps to prevent overfitting and handles outliers.

Let’s build some time-series models live and on-the-fly!