4 Statistical Paradoxes every Data Scientist needs to master

1 ACCURACY PARADOX

You have a dataset of 100 potential customer. Only 5% are interested in buying.

A model that always predict that a lead will not purchase will be right 95% of the time.

However, it will fail to identify the 5 potential buyers in the dataset.

ACCURACY PARADOX

Takeaway

high accuracy can be misleading when trying to identify rare events or occurrences within a dataset

2 FALSE POSITIVE PARADOX

A company wants to identify potential fraudulent transactions in their online store.

They create a fraud detection system that flags any transaction with an IP address from a high-risk country.

This results in a large number of false positives, as legitimate transactions from customers traveling abroad also get flagged.

FALSE POSITIVE PARADOX

While the fraud detection system may have a high accuracy rate, it is not effective in catching actual instances of fraud.

The False Positive Paradox, aka *the Base Rate Fallacy*, occurs when the probability of a true positive is low and the number of false positives is high.

This can lead to incorrect conclusions despite high accuracy rates.

FALSE POSITIVE PARADOX

Takeaway

When interpreting results, it's important to take into account the frequency of the condition or event being tested, instead of just relying on accuracy rates.

3 SIMPSON'S PARADOX

A company is analyzing the effectiveness of a new marketing campaign across different regions.

The analyst notices that the overall sales have increased since the launch, concluding that the campaign was successful.

However, when the data is broken down by region, it appears that some regions have experienced a decrease in sales, while others have seen an increase.

SIMPSON'S PARADOX

What's going on in here?

The regions with a higher sales volume before the campaign also had a higher increase in sales, while the regions with lower sales volume had a lower increase or even a decrease in sales.

If the company were to rely solely on the overall sales data, they may conclude that the campaign was successful, which might lead to suboptimal results.

SIMPSON'S PARADOX

Takeaway

When analyzing data, overall summary statistics and aggregate may mask important variations within subgroups.

4 BERKSON'S PARADOX

A company is analyzing the relationship between employee productivity and job satisfaction.

The analyst observes a positive correlation between the two, suggesting that happier employees are more productive.

However, the positive correlation disappears when controlled for employee tenure.

BERKSON'S PARADOX

The analyst soon realizes that employee tenure is a confounding variable for job satisfaction (independent var.) and productivity (dependent var.). In other words, employee tenure is positively associated with both.

Therefore, the positive correlation between job satisfaction and productivity is driven by the common association to employee tenure, not a direct cause-effect relationship between the two variables.

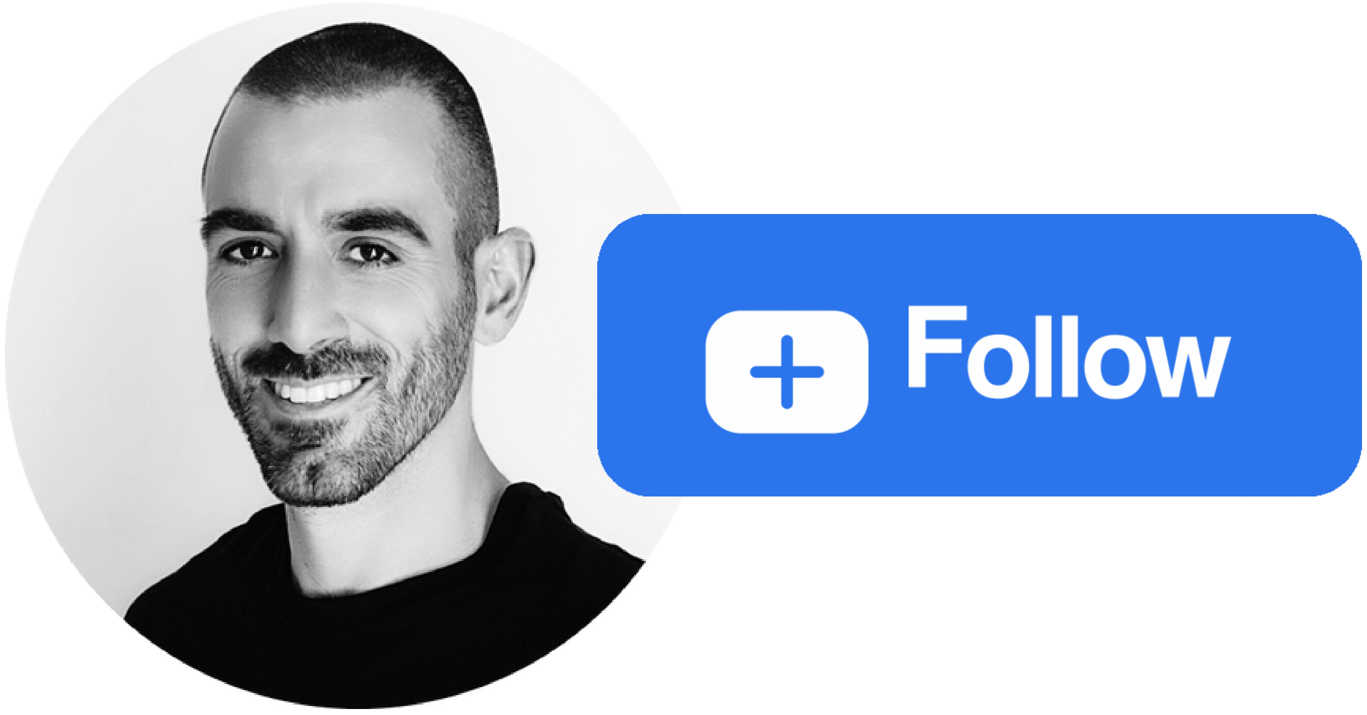
SIMPSON'S PARADOX

Takeaway

when studying the relationship between two variables, it's important to consider and control for potential confounding variables that might lead to drawing incorrect conclusions

RESHARE
THIS POST

IT'S THE BEST THING YOU CAN DO TO HELP OTHERS ON LINKEDIN



FOR MORE INSIGHTS AT THE INTERSACTION OF
DATA SCIENCE & SOLOPRENEURSHIP