Analytics Vidhya · Follow publication

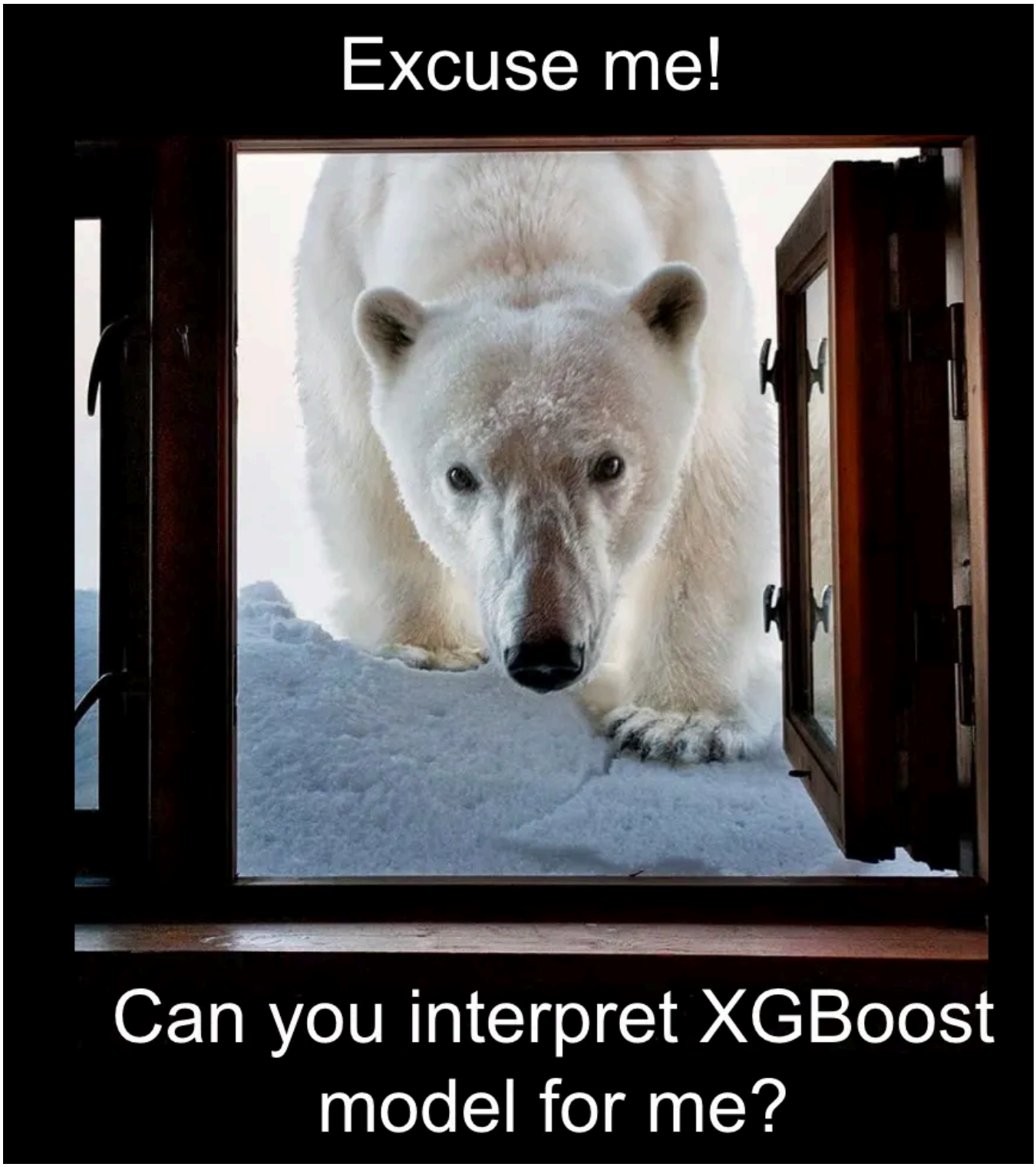
What is multicollinearity and how to remove it?

6 min read · Mar 12, 2020

Sharoon Saxena Follow

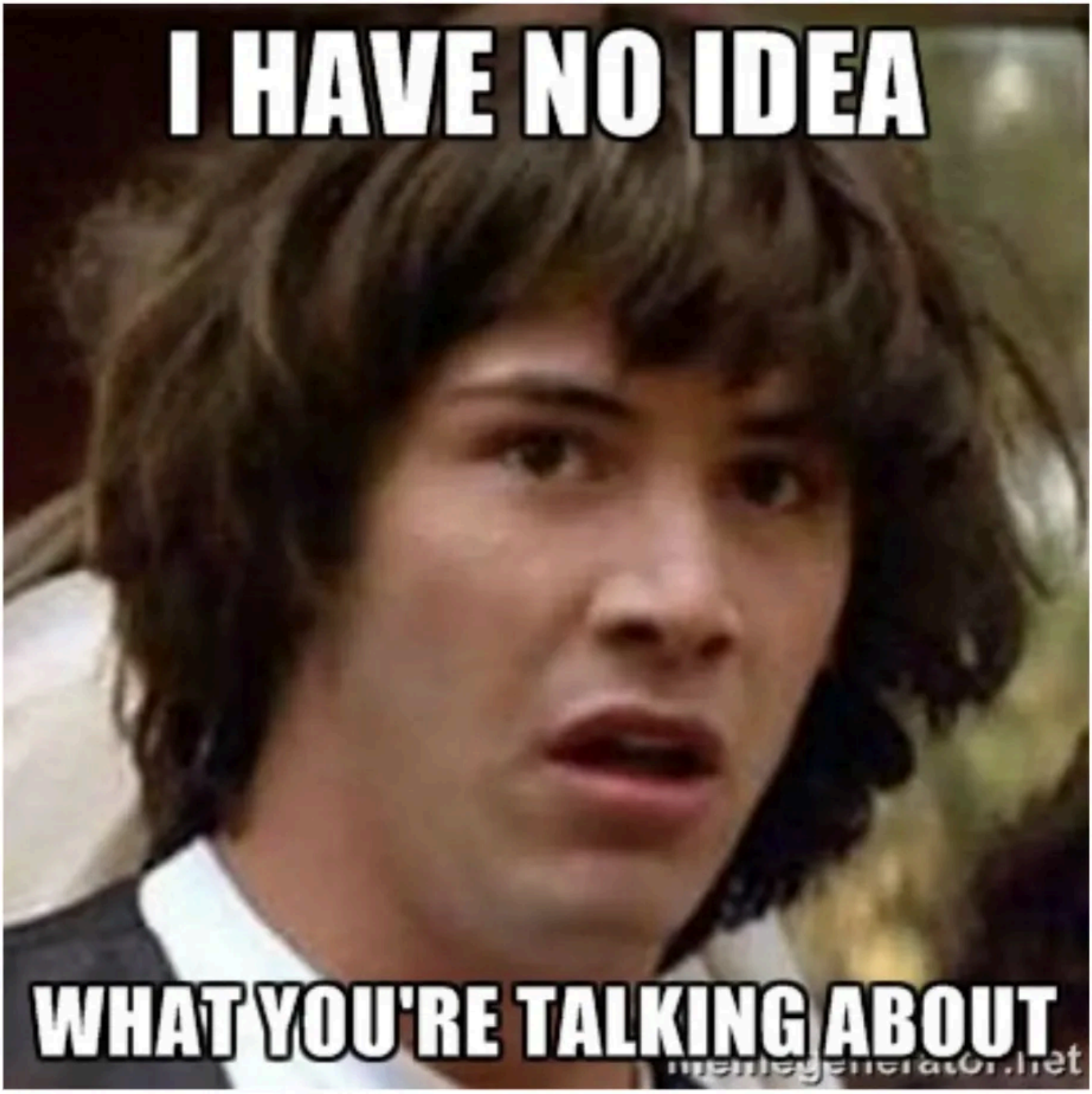
Introduction

With the advancements in Machine Learning and Deep Learning, we now have an arsenal of Algorithms that can handle any problem we throw at them. But there is an issue with most of these advanced and complex Algorithms, They are not easily interpretable.



When it comes to the interpretability of the Machine Learning Models, nothing comes close to the simplicity and interpretability of the Linear Regression. But there can be certain issues with the interpretability of the Linear Regression, especially when the assumptions of Linear regression known as Multicollinearity violated.

I am assuming you are familiar with assumptions of Linear Regression.



If you are like him, refer to the link below to know more about “Assumptions of Linear Regression”.

https://www.analyticsvidhya.com/blog/2016/07/deeper-regression-analysis-assumptions-plots-solutions/

We will be looking at the following questions regarding Multicollinearity:

1. What is Multicollinearity?
2. How Multicollinearity affects the Interpretation?
3. How do we detect and remove it?

So let's begin answering these questions one by one.

1. What is multicollinearity?

Multicollinearity is a condition when there is a *significant dependency or association between the independent variables or the predictor variables*. A significant correlation between the independent variables is often the first evidence of presence of multicollinearity.

Let's understand this through an Example:

Consider that I am working with the subset of the BigMart dataset as shown in the image

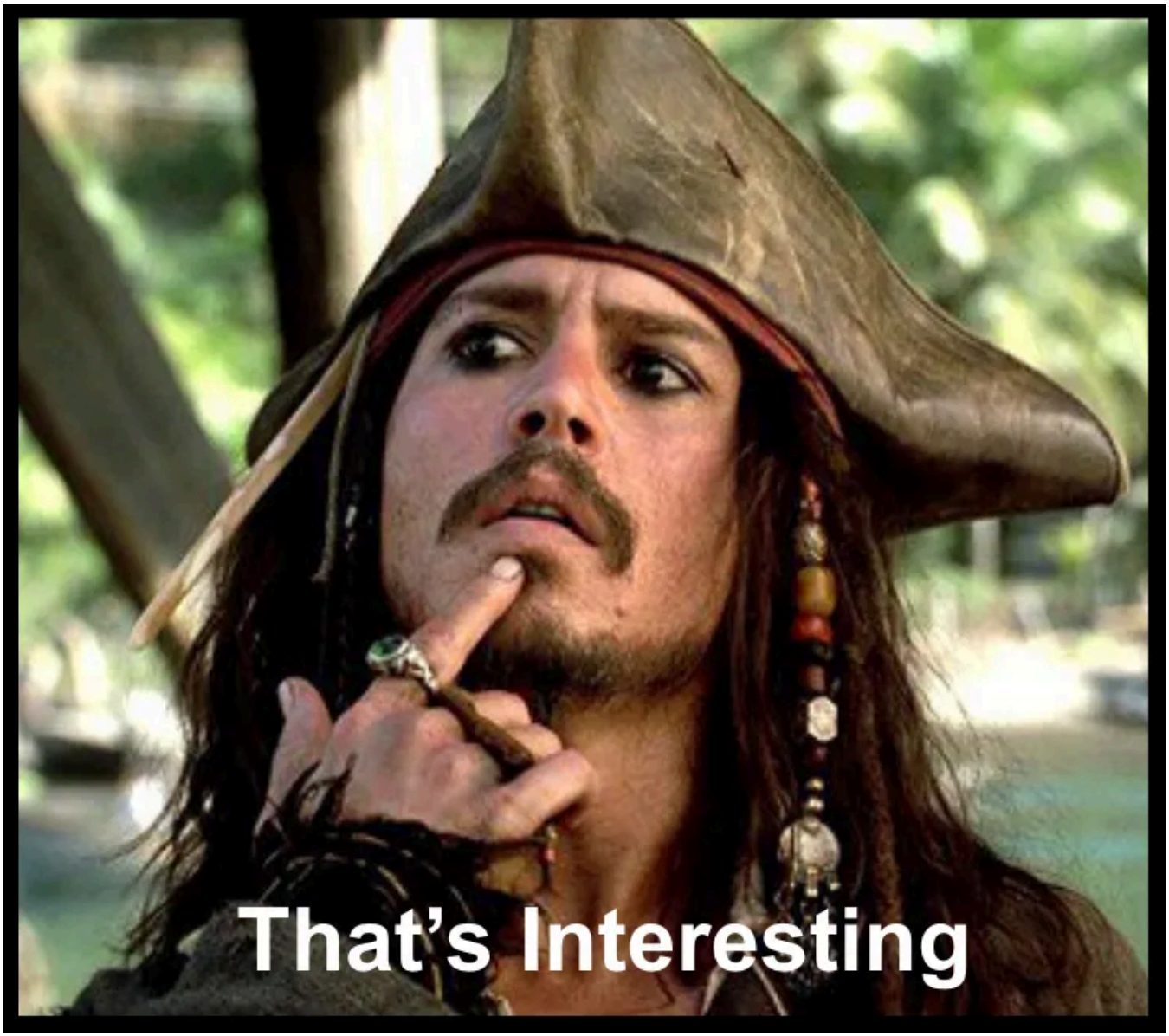
	Item_Weight	Item_Visibility	Item_MRP	Outlet_Establishment_Year	Item_Outlet_Sales
0	9.30	0.016047	249.8092	1999	3735.1380
1	5.92	0.019278	48.2692	2009	443.4228
2	17.50	0.016760	141.6180	1999	2097.2700
3	19.20	0.000000	182.0950	1998	732.3800
4	8.93	0.000000	53.8614	1987	994.7052

As we discussed before, multicollinearity occurs when there is a high correlation between the independent or predictor variables.

So let's have a look at the correlation matrix

	Item_Weight	Item_Visibility	Item_MRP	Outlet_Establishment_Year
Item_Weight	1.000000	-0.077522	0.022859	0.520561
Item_Visibility	-0.077522	1.000000	-0.001315	-0.074834
Item_MRP	0.022859	-0.001315	1.000000	0.005020
Outlet_Establishment_Year	0.520561	-0.074834	0.005020	1.000000

We can see in the correlation table that there is a significant correlation between the variables Outlet_Establishment_Year and the Item_Weight. This is our first clue that multicollinearity may be present.



2. How Multicollinearity affects the Interpretation

Consider the following Following Regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

In this model we can clearly see that there are 4 independent variables as X and the corresponding coefficients are given as β. Now consider a situation where *all the variables are independent except X3 and X4.*

Or in other words, X3 and X4 have significant correlation between them.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \boxed{\beta_3 X_3 + \beta_4 X_4}$$

Now to estimate the β coefficient of each independent variable with respect to Y, we observe the *change in the magnitude of Y variable when we slightly change the magnitude of any one independent variable at a time.*

Case 1:

Considering the Variables X1 and X2, they are independent of every other variable. If we try to *change the magnitude of the either X1 or X2 , they will not cause any other independent variable to change its value or by some negligible amount.* As a result we can clearly observe the influence of independent Variable X over Y.

$$Y\uparrow = \beta_0 + \beta_1 X_1\uparrow + \overbrace{\beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4}^{\text{Negligible change}}$$

Case 2:

In case of variables X3 and X4, they are significantly correlated. Can you guess what will happen if we apply the same procedure as Case 1?

Image Below illustrates exactly the same.

$$\overbrace{Y\uparrow}^{\text{Resultant Change}} = \beta_0 + \overbrace{\beta_1 X_1 + \beta_2 X_2}^{\text{Negligible change}} + \beta_3 X_3\uparrow + \beta_4 X_4\uparrow$$

According this Image, *If we try to change the magnitude of X3 (as shown in red) to observe the change in Y(red) , there will also be a significant difference in the value of X4(orange).* As a result, the change that we observe in Y is due to the change in both X3(red) and X4(orange). The resultant change(blue) is greater than the Actual change(orange).

Now you might ask, is that even a problem?

Yes, as we are trying to estimate the Coefficient corresponding to X3, the contribution of variable X4 causes the coefficient to be overestimated. And because of this, the coefficients are overestimated. As a result, our interpretations can be misleading.

Removing independent variables only on the basis of the correlation can lead to a valuable predictor variable as they correlation is only an indication of presence of multicollinearity.

But we are determined to eliminate it. Let's find out how we do it.

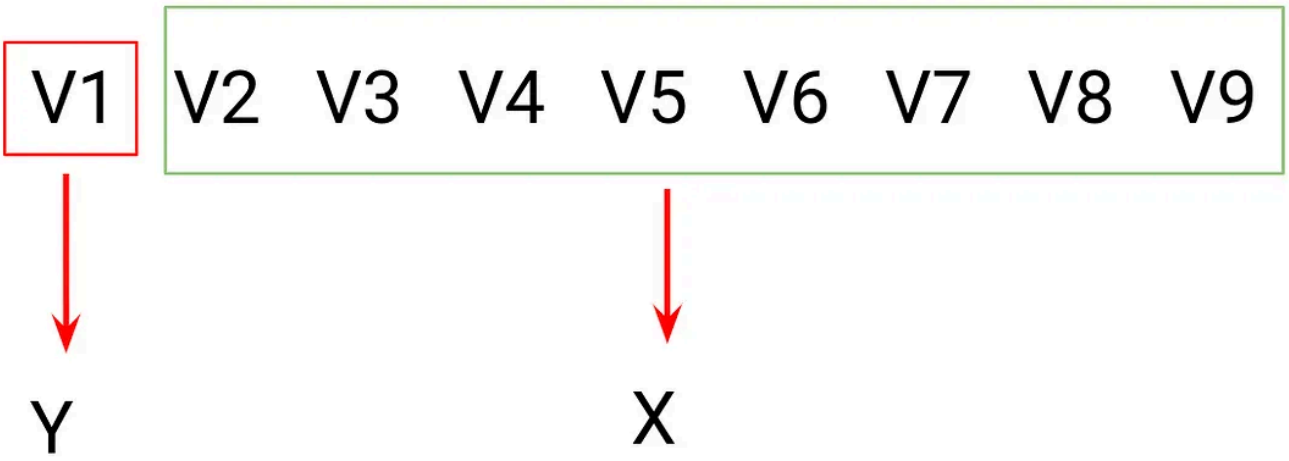
3. How do we detect and remove multicollinearity?

The best way to identify the multicollinearity is to calculate the **Variance Inflation Factor (VIF)** corresponding to every independent Variable in the Dataset.

VIF tells us about how well an independent variable is predictable using the other independent variables. Let's understand this with the help of an example.

V1 V2 V3 V4 V5 V6 V7 V8 V9

Consider that we have 9 independent variables as shown. To calculate the VIF of variable V1, we isolate the variable V1 and consider as the target variable and all the other variables will be treated as the predictor variables.



We use all the other predictor variables and train a regression model and find out the corresponding R2 value.

Using this R2 value, we compute the VIF value gives as the image below.

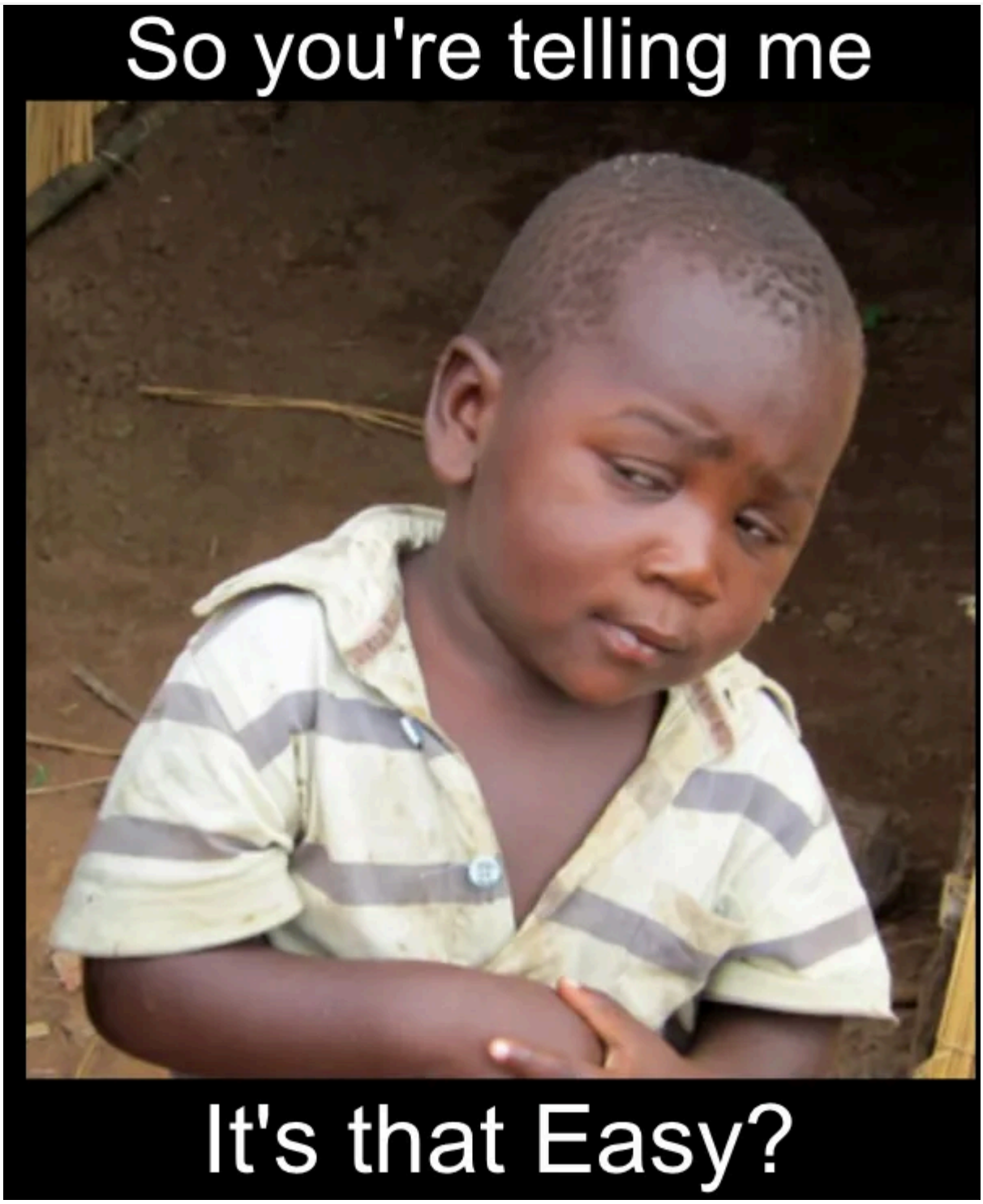
$$R^2 = 1 - \frac{\text{MSE(model)}}{\text{MSE(baseline)}} \qquad \text{VIF} = \frac{1}{1 - R^2}$$

Looking at the formulation we can clearly see that as the R2 value increases, the VIF value also increases. A higher R2 value signifies that:

"the target independent variable is very well explained by the other independent variables"

Now what should be the VIF threshold value to decide whether the variable should be removed or not?

It is always desirable to have VIF value as small as possible, but it can lead to many significant independent variables to be removed from the dataset. Therefore a VIF = 5 is often taken as a threshold. Which means that any independent variable greater than 5 will have to be removed. Although the ideal threshold value depends upon the problem at hand.



Power of Linear regression lies in the simple interpretation of the model. Missing out on multicollinearity will definitely kill the purpose of using the linear regression in the first place. I will wrap this up assuming that you have understood the concept of multicollinearity, issues caused by multicollinearity and how to detect and remove the multicollinearity in any given problem.

Machine Learning | Multicollinearity | Linear Regression | Model Interpretation

Follow

Published in Analytics Vidhya

74K Followers · Last published Mar 6, 2025

Analytics Vidhya is a community of Generative AI and Data Science professionals. We are building the next-gen data science ecosystem https://www.analyticsvidhya.com

Follow

Written by Sharoon Saxena

225 Followers · 6 Following

https://www.linkedin.com/in/sharoon-saxena-0539a0126/

Responses (9)

Kathir G

What are your thoughts?

Shwetank
Sep 25, 2024

> Actual change(orange).

should it be red and not orange? X3 change and corresponding y change is red.

Reply

Infinity
May 28, 2024

I dont think this guy understands multicollinearity

Reply

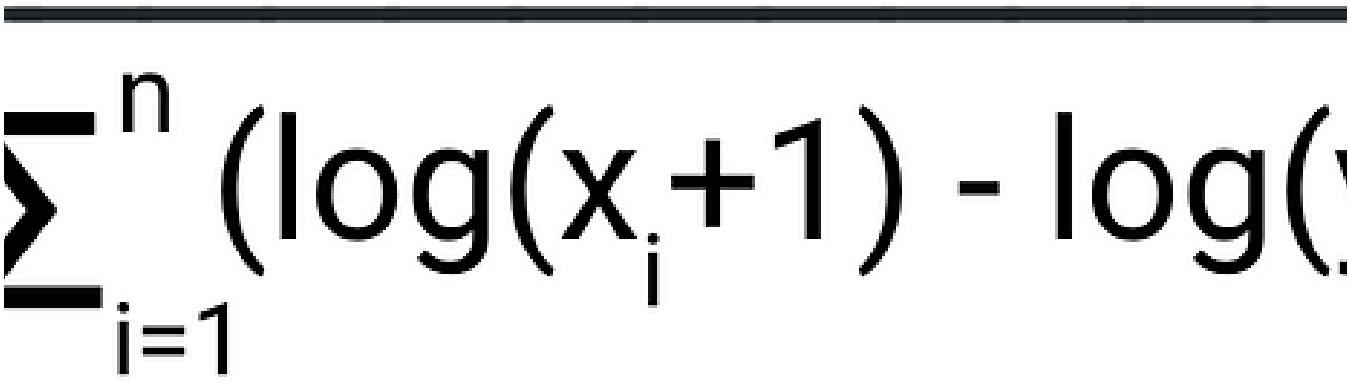
Infinity
May 28, 2024

what would you do in case of a classification task?

Reply

See all responses

More from Sharoon Saxena and Analytics Vidhya



In Analytics Vidhya by Sharoon Saxena

RMSE vs RMLSE—What's the Difference? When Should You use Them?

Introduction

Jun 26, 2019 1.2K 13

In Analytics Vidhya by Harikrishnan N B

Confusion Matrix, Accuracy, Precision, Recall, F1 Score

Binary Classification Metric

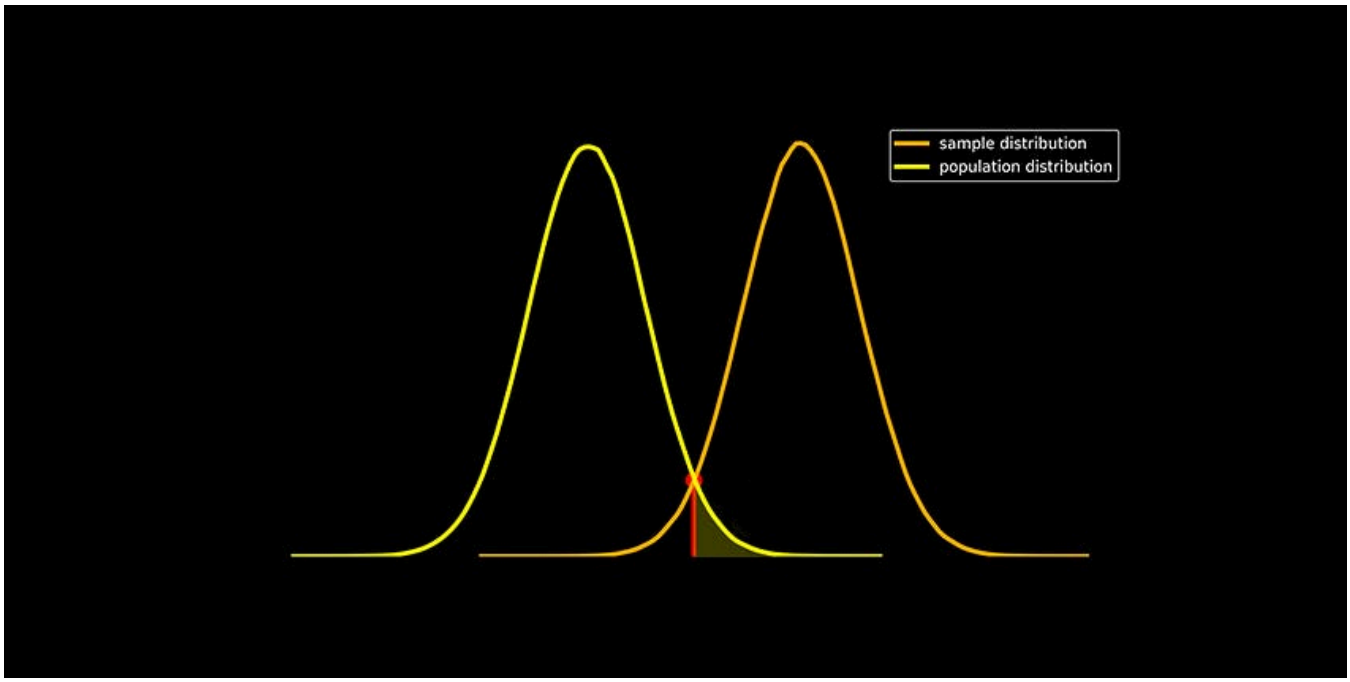
Dec 10, 2019 1.2K 6

In Analytics Vidhya by Kia Eisinga

How to create a Python library

Ever wanted to create a Python library, albeit for your team at work or for some open source project online? In this blog you will learn...

Jan 27, 2020 2.8K 33



In Analytics Vidhya by Sharoon Saxena

Everything you Should Know about p-value from Scratch for Data Science

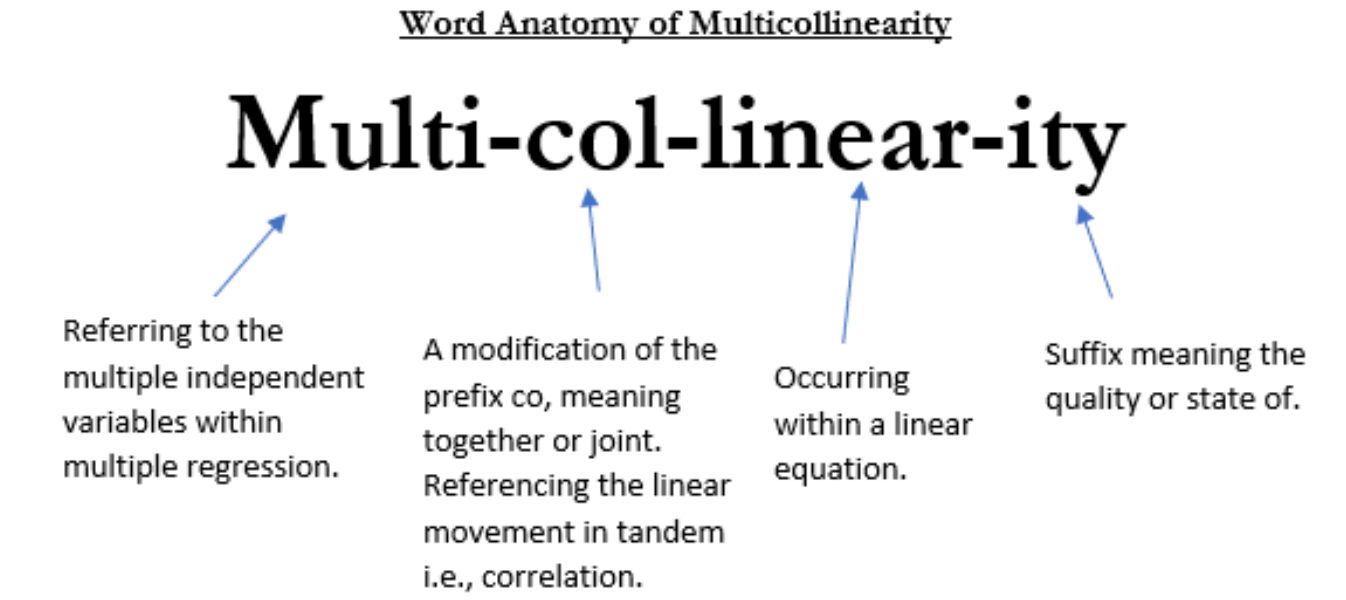
What is p-value? Where is it used in data science? And how can we calculate it? We answer all these questions in this article and more.

Sep 5, 2019 857 11

See all from Sharoon Saxena

See all from Analytics Vidhya

Recommended from Medium



In Academy Team by Mustafa Erboga, Ph.D.

Multicollinearity in Data Science and Machine Learning: The Hidden Threat and How to Tackle It

In data science and machine learning, understanding the relationships between variables is essential for building accurate and...

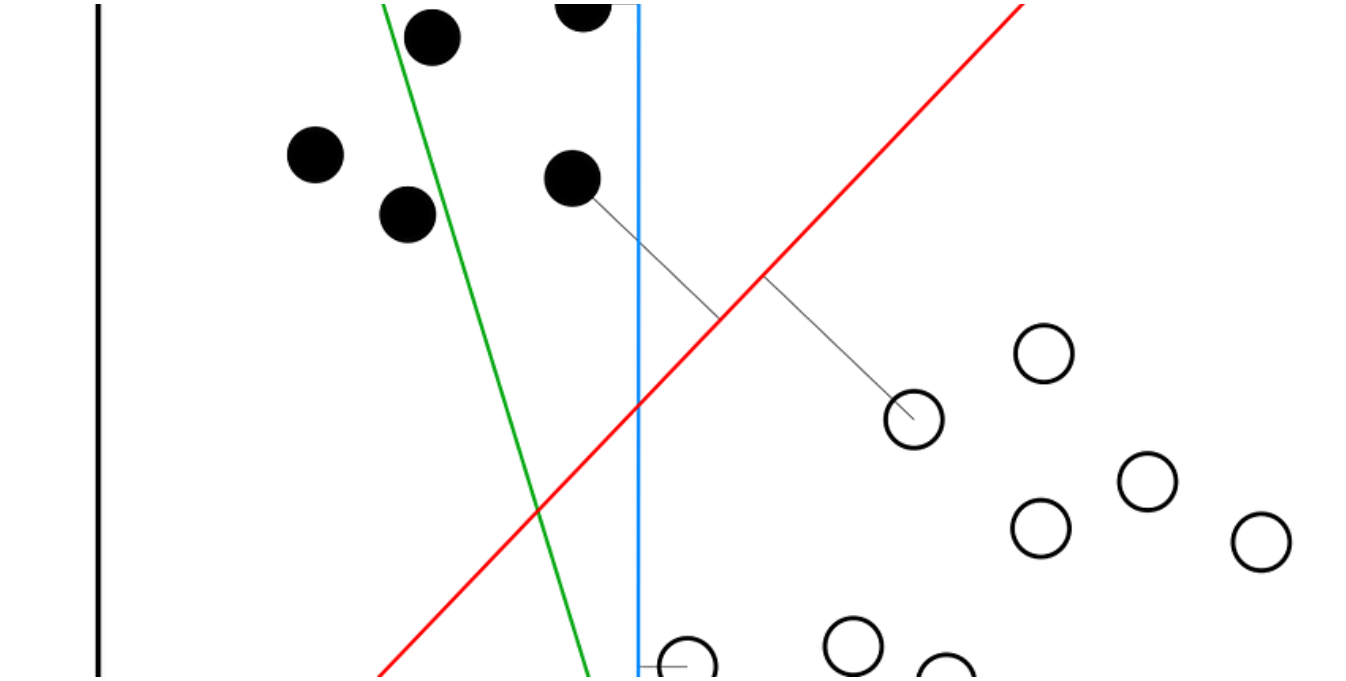
Nov 9, 2024 381 5

UATeam

Machine Learning Models for Time Series Analysis: A Comprehensive Guide

Time series data, consisting of observations collected over time, is prevalent across various domains such as finance, healthcare, and IoT...

Nov 19, 2024 1



In Artificial Intelligence in Plain English by Ritesh Gupta

Data Science All Algorithm Cheatsheet 2025

Stories, strategies, and secrets to choosing the perfect algorithm.

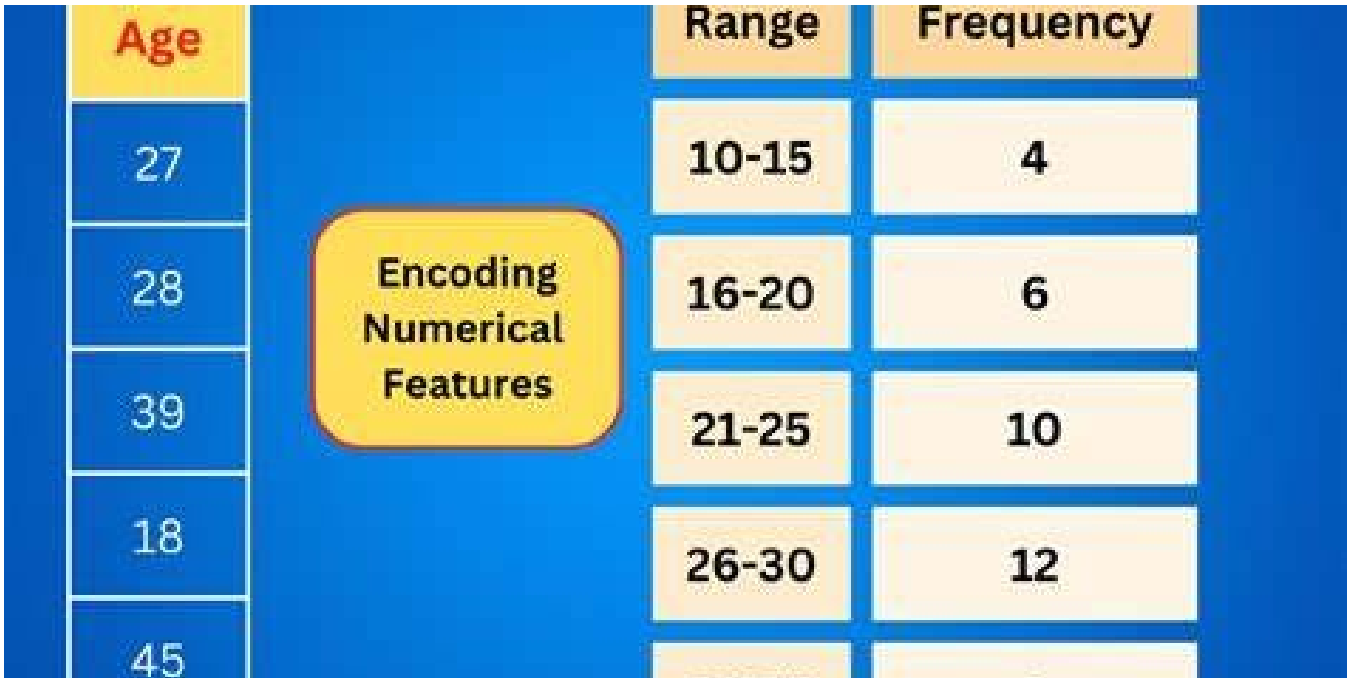
Jan 5 1.6K 42

In Learn AI for Profit by Nipuna Maduranga

You Can Make Money With AI Without Quitting Your Job

I'm doing it, 2 hours a day

Mar 25 8.4K 389

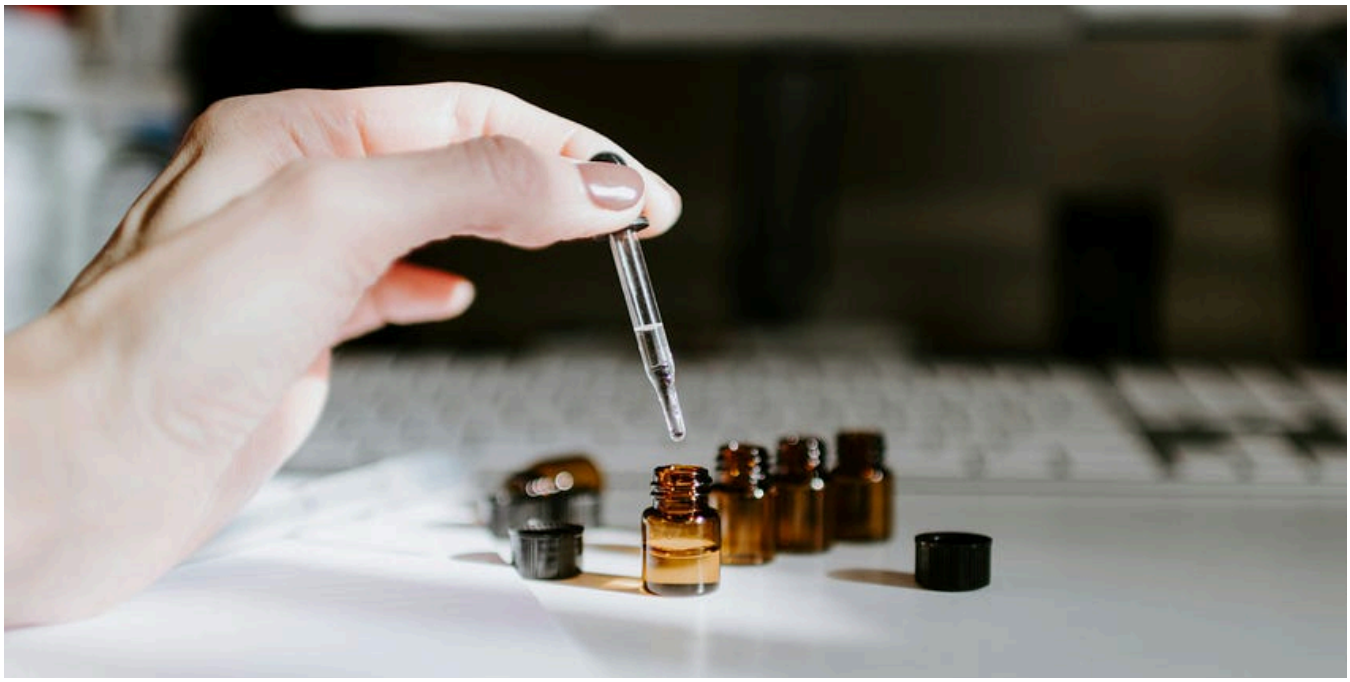


ajaymehta

"Exploring Different Types of Binning and Discretization Techniques in Data Preprocessing Part"

disclaimer: read it below blog first

Jan 2 2



Sunghyun Ahn

How to Deal with Multiple Hypothesis Testings: FWER and FDR

A/B Testing for Data Science Series (9): Bonferroni Correction and False Discovery Rates

Nov 24, 2024 3 2