Variance Inflation Factor (VIF) method works, both conceptually and in terms of this code

What is VIF?

Variance Inflation Factor (VIF) measures multicollinearity in regression analysis. It quantifies how much the variance of a regression coefficient is inflated due to correlation with other predictor variables.

The Formula

For each feature, VIF is calculated as:

$VIF = 1 / (1 - R^2)$

Where R^2 is the coefficient of determination when that feature is regressed against all other features.

Interpretation

- **VIF = 1**: No correlation with other features
- **VIF = 1-5**: Moderate correlation (generally acceptable)
- **VIF = 5-10**: High correlation (concerning)
- **VIF > 10**: Severe multicollinearity (problematic)

How the Code Works

Let me break down your function step by step:

```
def cal_variance_inflation_factor(df):

    # 1. Create empty DataFrame to store results
    vif_data = pd.DataFrame()

    # 2. Extract only numeric column names
    vif_data["feature"] = df.select_dtypes(include=[np.number]).columns

    # 3. Calculate VIF for each numeric column
    vif_data["VIF"] = [
        variance_inflation_factor(
            df.select_dtypes(include=[np.number]).values,  # All numeric data as array
            i  # Index of the feature to calculate VIF for
        )
        for i in range(len(df.select_dtypes(include=[np.number]).columns))
    ]

    return vif_data
```

What happens internally:

For each feature at index `i`:

1. The function treats that feature as the **dependent variable**
2. All other numeric features become **independent variables**
3. It performs a regression: Feature_i ~ Feature_1 + Feature_2 + ... + Feature_n
4. Calculates R^2 from this regression
5. Computes VIF = 1 / (1 - R^2)

Example

If you have features [Age, Income, Debt]:

- For "Age": Regresses Age ~ Income + Debt, gets $R^2 = 0.75$, so VIF = 1/(1-0.75) = **4**
- For "Income": Regresses Income ~ Age + Debt, gets $R^2 = 0.90$, so VIF = 1/(1-0.90) = **10**

- For "Debt": Regresses Debt ~ Age + Income, gets $R^2 = 0.60$, so VIF = 1/(1-0.60) = **2.5**

This would suggest "Income" has high multicollinearity and might need removal.