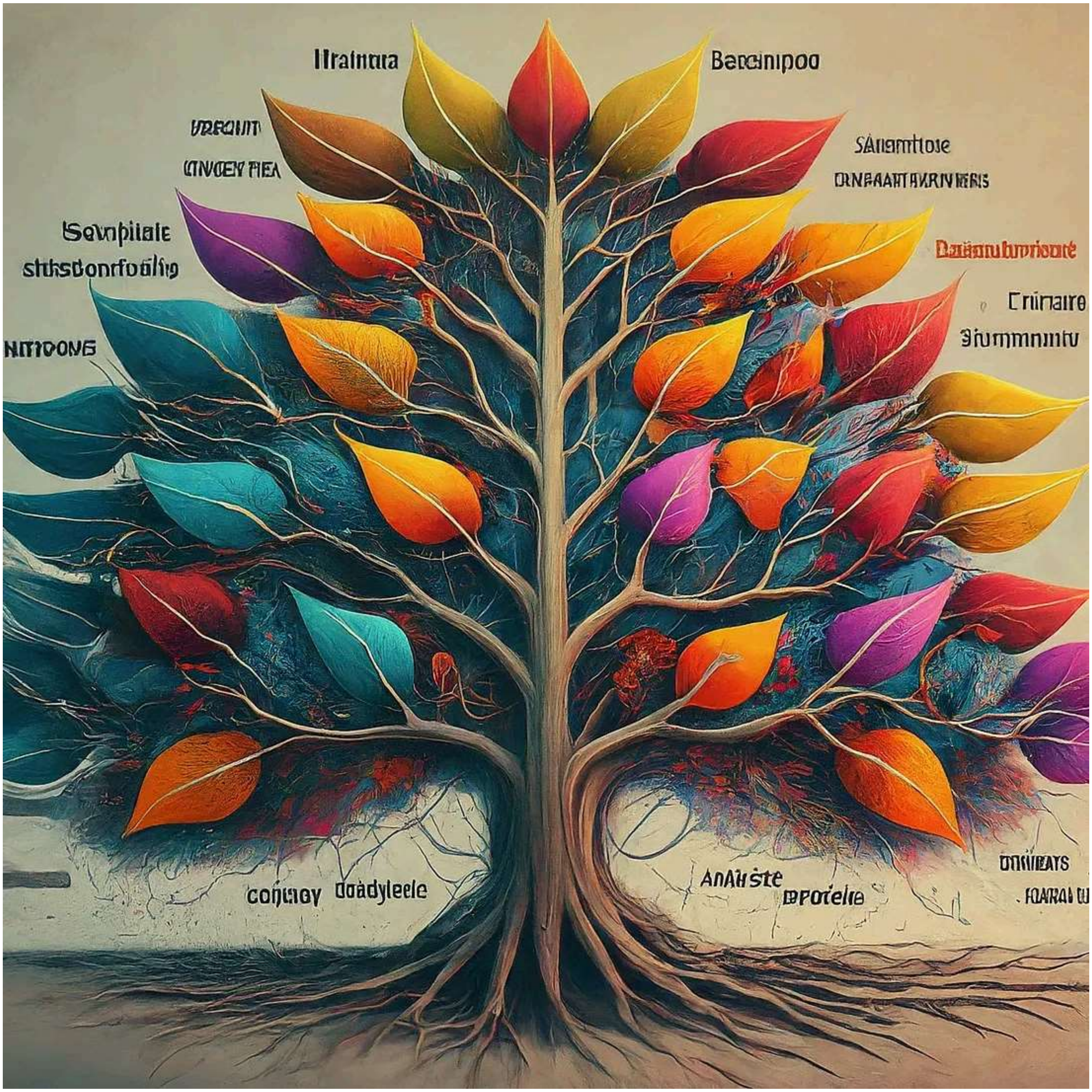
High Correlation in Decision Tree Models: Why It Matters

James John Adison · Follow

6 min read · Jun 13, 2024



Introduction:

Machine learning models for classification are designed to predict discrete categories, such as "Yes" or "No," rather than continuous numerical values. One widely used classification algorithm is the Decision Tree, which splits data into subsets based on feature values to form a tree-like model of decisions. While the impact of high feature correlation (multicollinearity) in regression models is well-documented, its effects on decision tree models are less frequently addressed.

High Correlation in Decision Tree Models:

The common perception is that multicollinearity does not pose a significant issue in decision tree algorithms because these algorithms implicitly handle it by selecting only one of the highly correlated features. However, this study aims to demonstrate that high multicollinearity between features in decision tree-based models can negatively impact the model's predictive accuracy. High correlation between features leads to redundancy and increases the risk of overfitting, thereby reducing the model's ability to make accurate predictions.

Methodology:

Defining Covariance and Correlation

COVARIANCE

Covariance is a measure of how much two random variables vary together. For example, height and weight of giraffes have positive covariance because when one is big the other tends also to be big.

Definition: Suppose X and Y are random variables with means μX and μY . The covariance of X and Y is defined as $Cov(X, Y) = E((X - \mu - \mu Y))$. ** (2)

Where 'E' is the Expectation function and its defined for discrete values as:

$$E[X] = \sum_{i=1}^{n} x_i P(X = x_i)$$

Where P is the probability for the xi values of the 'X' feature.

CORRELATION

Correlation The units of covariance Cov(X, Y) are 'units of X times units of Y '. This makes it hard to compare covariances: if we change scales then the covariance changes as well. Correlation is a way to remove the scale from the covariance.

Definition: The Pearson correlation coefficient between X and Y is defined by

$$Cor(X, Y) = \rho = \frac{Cov(X, Y)}{\sigma X \, \sigma Y} \qquad ** \ (3)$$

There are other correlation coefficients but we'll use the most common (Pearson)

Where

- Cov(*x*,*y*) is the covariance of *xx* and *yy*,
- *σx* is the standard deviation of *xx*,
- *σy* is the standard deviation of *yy*.

Geometrically deviations are defined as the distance between an ‘xi’ point and the average x

So for all the ’n’ x i cases and yi cases

$$\sigma x \ \sigma y = \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

And Covariance is defined as

$$\mathrm{Cov(x,y)} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Finally the Person Coefficient correlation ‘r’ is defined as

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

HIGH CORRELATION BEETWEEN FEATURES

For a Decision Tree model that has features ; X, Y, A, B………Z

Variable to predict : P and ’n’ samples

If there’s a very high correlation between X and Y (very close to 1)

Then we could approximate the person coefficient to : r →1 , then

$$1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Then

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad \text{.... (1)}$$

For the equation to be solved the root square has to be raised in the below way

$$(\sqrt{a})^2 = a$$

$$\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2} = (\sqrt{a})(\sqrt{a})$$

So the deviations of xi and yi must be proportional by a constant 'k'

$$(x_i - \bar{x}) = k(y_i - \bar{y}) \qquad \text{...... (2)}$$

Replacing (2) in (1)

$$\sum_{i=1}^{n} k(y_i - \bar{y}) \quad (y_i - \bar{y}) = k\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$\sum_{i=1}^{n} k(y_i - \bar{y}) \quad (y_i - \bar{y}) = k\left|\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\right|^2$$

$$\sum_{i=1}^{n} k(y_i - \bar{y}) \quad (y_i - \bar{y}) = k\sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$k\sum_{i=1}^{n}(y_i - \bar{y})^2 = k\sum_{i=1}^{n}(y_i - \bar{y})^2$$

So the equation becomes an identity.

Now we can proceed to use the equation (2)

$$(x_i - \bar{x}) = k(y_i - \bar{y})$$

$$yi - \bar{y} = (1/k)^* (xi - \bar{x})$$

$$yi = (1/k)\, xi + \underbrace{\bar{y} - (1/k)\, \bar{x}}_{\text{Is an scalar that can be named “ a”}}$$

$$yi = b\, xi + a \quad \ldots\ldots\ (3)$$

It shows the linear relation of yi and xi

Working in the Decision Tree model

For the below dataset for a model for binomial classification (n or p)

N	X	Y	K	L		Z	P
1	x1	y1	k1	l1	…….	z1	p
2	x2	y2	k2	l2	…….	z2	p
3	x3	y3	k3	l3	…….	z3	n
4	x4	y4	k4	l4	…….	z4	p
5	x5	y5	k5	l5	…….	z5	n
6	x6	y6	k6	l6	…….	z6	n
	…….	…….	…….	…….	…….	…….	…….
	…….	…….	…….	…….	…….	…….	…….
N	xn	yn	kn	ln		zn	p

Where X, Y, A……Z are the features and P is the variable to predict

In Decision Tree algorithm the objective is to choose the feature that would provide the best split using “Information Gain”

GAIN = Entropy(S) — I(Atribute) . . . . (4)

Entropy(S) = f(pi)

Is the total entropy of the system and will be the same in every evaluation for each attribute(feature)

Evaluating the models the objective is to choose the feature depending on the Average information for this feature:

- Calculate **Average Information**:

$$I(Attribute) = \sum \frac{p_i + n_i}{p+n} Entropy(A) \qquad \text{. . . . . (5)}$$

And the Entropy(A) of every attribute will be calculated as ;

- Calculate **Entropy** (Amount of uncertainity in dataset):

$$Entropy = \frac{-p}{p+n} log_2(\frac{p}{p+n}) - \frac{n}{p+n} log_2(\frac{n}{p+n}) \qquad \text{. . . . . (6)}$$

Therefore the Average Information I(X) depends of and the relation between 'X' and 'P'

$$I(X) = f\,(X,P) = \sum \frac{p_i + n_i}{p+n} \times \frac{-p}{p+n} log_2(\frac{p}{p+n}) - \frac{n}{p+n} log_2(\frac{n}{p+n}) \qquad \text{. . . . . (7)}$$

Hence for every feature F , its Average Information depends of the relation between ‘F’ and ‘P’

I (F) = f (F, P) . . . . (8)

And for the feature Y

I (Y) = f (Y,P) = -Σ p(y) log p(y) . . . . (9)

Because there’s a very high correlation between X and Y , I showed that in (3) that that for every ‘yi’ that E Y

Or Y = bX + a . . . . . (4)

So in (9)

I (Y) = I (bX + a)

The case is referred to discrete variables the

The entropy of Y, denoted as I(Y), can be calculated as:

Since Y is a linear transformation of X, we can substitute y = ax + b into equation (9):

I(Y) = -Σ p(ax + b) log p(ax + b)

Applying the change of variable technique

Y = g(X),

from probability theory

P(y) = P(g^-1(y)) * d[g^-1(y)]/dy, where g^-1(y) is the inverse function of g(x). . . . . (10)

In our case, g(x) = ax + b, and its inverse is

g^-1(y) = (y — b)/a.

Applying the derivative

d[g^-1(y)]/dy = d[(y-b)/a]/dy = d[(1/a)y -b/a}/dy = d[(1/a)y/dy -d(b/a)/dy = 1/a

Substituting this into the entropy equation, we get:

I(Y) = -Σ p((y — b)/a) log p((y — b)/a) * (1/a) . . . . (11)

From (4) y = ax = b then x = (y-b)/a , then (11) can be re-expressed as

I(Y) = -(1/a) Σ p(x) log p(x) = (1/a) I(X)

Since the entropy is invariant to a constant multiplicative factor, we can conclude that:

I(aX + b) = I(X) = I(Y)

Results:

The findings indicate that features with very high correlation (approaching 1) in decision tree classifier models provide redundant information, resulting in identical Information Gain. This redundancy introduces several critical issues:

1. **Redundancy:** High correlation means that features carry similar information, adding unnecessary complexity to the model.
2. **Overfitting:** Redundant features increase the likelihood of overfitting, where the model learns noise and specifics of the training data rather than the underlying patterns, thus performing poorly on unseen data.
3. **Reduced Predictive Performance**

Conclusion:

The study shows that in the presence of multicollinearity, although decision trees may select only one of the highly correlated features during splitting, the presence of such features still leads to redundancy and overfitting, impairing predictive performance.

You could contact me at:

yot181@hotmail,com

REFERENCES

(1) C. Shah, *A Hands-On Introduction to Machine Learning*. University of Washington, 2023

(2) Jeremy Orloff and Jonathan Bloom — MIT Mathematics

(3) Jeremy Orloff and Jonathan Bloom — MIT Mat

Decision Tree Algorithm | Machine Learning | Correlation | Multicollinearity

Written by James John Adison

0 Followers · 1 Following

Follow

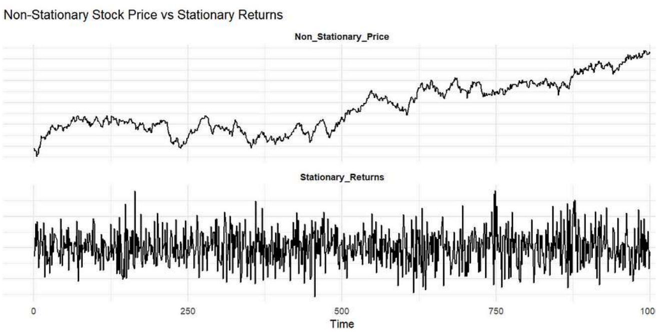
Electronic Engineer, Machine Learning researcher.

No responses yet

John Bennett

What are your thoughts?

Recommended from Medium

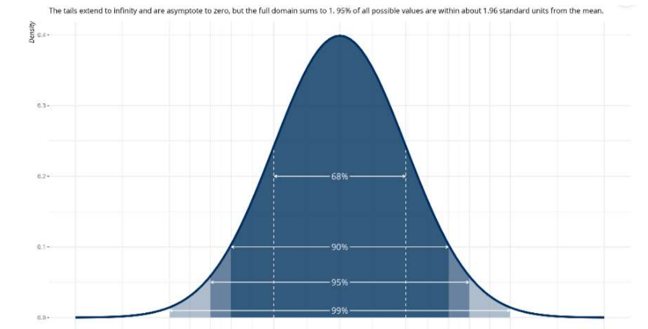


Jhalyl Mason

Mastering Time Series Stationarity

Understanding Stationarity for Improved Forecasting

Feb 20 · 2



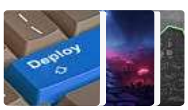
ishan

CLT and Confidence Intervals

Central Limit Theorem (CLT)

Nov 1, 2024

Lists

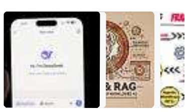


Predictive Modeling w/ Python

20 stories · 1845 saves

Practical Guides to Machine Learning

10 stories · 2217 saves

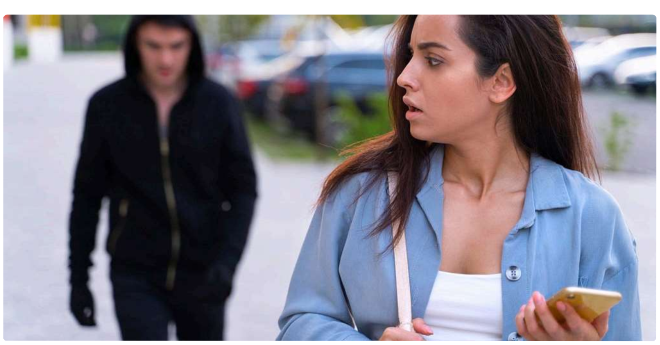


Natural Language Processing

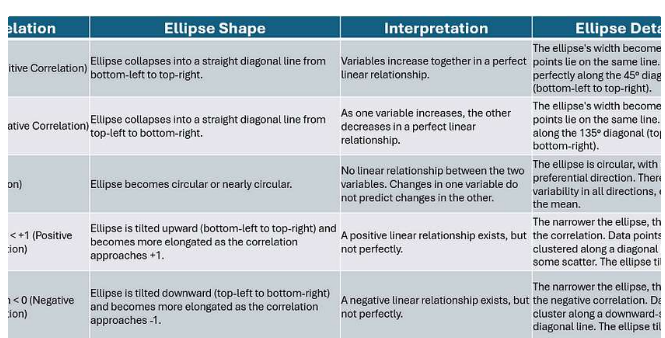
1962 stories · 1606 saves

The New Chatbots: ChatGPT, Bard, and Beyond

12 stories · 559 saves



Damini Vadrevu



elation	Ellipse Shape	Interpretation	Ellipse Deta
itive Correlation)	Ellipse collapses into a straight diagonal line from bottom-left to top-right.	Variables increase together in a perfect linear relationship.	The ellipse's width become points lie on the same line. perfectly along the 45° diag (bottom-left to top-right).
ative Correlation)	Ellipse collapses into a straight diagonal line from top-left to bottom-right.	As one variable increases, the other decreases in a perfect linear relationship.	The ellipse's width become points lie on the same line. along the 135° diagonal (to bottom-right).
on)	Ellipse becomes circular or nearly circular.	No linear relationship between the two variables. Changes in one variable do not predict changes in the other.	The ellipse is circular, with preferential direction. There variability in all directions, the mean.
< +1 (Positive ion)	Ellipse is tilted upward (bottom-left to top-right) and becomes more elongated as the correlation approaches +1.	A positive linear relationship exists, but not perfectly.	The narrower the ellipse, th the correlation. Data points clustered along a diagonal some scatter. The ellipse ti
< 0 (Negative ion)	Ellipse is tilted downward (top-left to bottom-right) and becomes more elongated as the correlation approaches -1.	A negative linear relationship exists, but not perfectly.	The narrower the ellipse, th the negative correlation. Da cluster along a downward- diagonal line. The ellipse ti

Rayan Yassminh

Predicting Customer Behaviour with Propensity Modeling

A Step-By-Step Guide on How to Build a Propensity Model

Feb 1 · 3

Covariance and Correlation in Machine Learning: Practical…

In machine learning, everything revolves around variables — the features we use to…

Sep 13, 2024 · 27 · 1

Julian Wang

Ordinal regression

To predict ranked levels such as “very low,” “low,” “median,” “high,” and “very high” based…

Dec 16, 2024

$$z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

Lu Jie

Null hypothesis, Alternative hypothesis and p-value

Alternative hypothesis(Ha): predicts there are significant effect or difference.

6d ago

See more recommendations