Mutual Information Reduction Techniques and its Applications in Feature Engineering

Ruixin Chen
Katz School of Science and Health
Yeshiva University
New York, NY
rchen4@mail.yu.edu

David Li
Katz School of Science and Health
Yeshiva University
New York, NY
david.li@yu.edu

***Abstract*—Feature engineering is a vital aspect of machine learning model development, as it involves selecting and transforming relevant data features to improve predictive accuracy. Traditional feature selection methods focus on maximizing mutual information (MI) between features and the target variable. In contrast, this paper introduces novel mutual information reduction techniques aimed at minimizing redundant information between features. By reducing mutual information among features, these methods enhance feature selection and creation, allowing machine learning models to benefit from less correlated and more informative variables. We provide examples and an implementation to demonstrate how mutual information reduction improves model performance, particularly in classification tasks. Additionally, the integration of Weight of Evidence (WOE) transformation further boosts predictive power by capturing the unique information from each feature.**

***Index Terms*—Entropy, Mutual Information, Feature Engineering**

I. Introduction

Feature engineering is a critical component of the machine learning process, as it involves transforming raw data into features that improve the predictive power of models. The objective is to provide machine learning algorithms with the most relevant and informative variables, which can significantly enhance the model's ability to make accurate predictions or classifications. [1], [3], [4] One effective tool in optimizing this process is Mutual Information (MI), a statistical technique that helps evaluate the relationships between features and the target variable. [2]

A. Mutual Information (MI)

Mutual information is a statistical measure that quantifies the amount of information one variable contains about another. In essence, it evaluates the dependency between two variables, capturing both linear and nonlinear relationships. Unlike simpler methods such as correlation, mutual information provides a more comprehensive assessment of how much knowing one variable reduces uncertainty about another.

In the context of machine learning, mutual information is particularly useful for assessing the relevance of features with respect to the target variable. Features that have higher mutual information scores are considered to be more informative and are likely to contribute to better model predictions.

B. Mutual Information for Feature Selection

Feature selection is a key part of feature engineering, where the goal is to identify the most relevant features that contribute to the model's performance. Mutual information offers a robust criterion for feature selection by capturing both linear and nonlinear dependencies between features and the target variable.

By selecting features with the highest mutual information scores, practitioners can ensure that their models are built on the most informative and predictive variables. This is particularly useful in situations where relationships between features and the target are complex or non-obvious.

C. Mutual Information for Feature Creation

In addition to feature selection, mutual information can also inform the creation of new features. By analyzing mutual information between various feature combinations and the target variable, it is possible to discover important interactions or transformations that may not be apparent in the raw data.

New features generated through mutual information analysis may capture additional information that enhances model performance. For example, nonlinear transformations or interaction terms identified via mutual information can help the model capture intricate patterns in the data, improving its predictive accuracy.

D. The Contribution of this Study

In this paper, we examine the relationship between feature engineering and mutual information. Unlike most research, which focuses on using mutual information between features and the target for feature selection, we propose a novel approach that reduces mutual information between features. This method can be applied to both feature selection and the creation of new features, ultimately enhancing the performance of machine learning models.

II. Compute the mutual information between two numerical variables

In machine learning and data science, assessing the association between variables is a critical aspect of the modeling process. Traditional methods, such as the correlation coefficient, often fail to detect complex, nonlinear relationships between variables. Mutual information offers a more flexible and robust metric for univariate screening compared to correlation. While mutual information can capture both linear and nonlinear dependencies, correlation may overlook significant relationships between variables.

A. Entropy

The entropy of a random variable measures the average uncertainty or amount of information associated with its possible states or outcomes. [5]

Let X be a discrete random variable. The entropy of X is defined as,

$$H(X) = \sum_{x \in X} p(x) \log \frac{1}{p(x)} = -\sum_{x \in X} p(x) \log p(x) \tag{1}$$

B. Conditional Entropy

Conditional entropy measures the uncertainty of a random variable Y given that some information about another random variable X is already known.

$$H(Y|X) = -\sum_{x \in X, y \in Y} p(x, y) \log \frac{p(x, y)}{p(x)} \tag{2}$$

If X and Y are independent, then the conditional entropy degenerates to the self entropy of Y,

$$H(Y|X) = H(Y) \tag{3}$$

C. Mutual Information

As its name implies, mutual information quantifies how much information is shared between two variables, rather than simply measuring their correlated ”movement.” Mutual information (MI) between two random variables is a non-negative measure of their dependency. It is zero if and only if the two variables are independent, with higher values indicating a greater level of dependency between them.

Consider X and Y again. In equation (3), Y is completely independent of X. If the two variables are not completely independent, the remaining information is the entropy shared between X and Y.

$$I(Y; X) = H(Y) - H(Y|X) \tag{4}$$

- Mutual information for two continuous variables,

$$I(X; Y) = \int_y \int_x p_{X,Y}(x, y) \log \frac{p_{X,Y}(x, y)}{p_X(x) p_Y(y)} dxdy \tag{5}$$

- Mutual information for two discrete variables,

$$I(X; Y) = \sum_{y \in Y} \sum_{x \in X} p_{X,Y}(x, y) \log \frac{p_{X,Y}(x, y)}{p_X(x) p_Y(y)} \tag{6}$$

D. Numerical Techniques for Mutual Information

The equations (5) and (6) rely on the underlying probabilities distributions, which are usually unavailable. The main idea of computing mutual information is to cut the continuous variables into bins, then the probability of bins can be estimated numerically.

Rewrite the equation (5) into,

$$\begin{aligned} I(X; Y) &= \int_y \int_x p_{X,Y}(x, y) \log \frac{p_{X,Y}(x, y)}{p_X(x) p_Y(y)} dxdy \\ &\approx I_{\text{binned}}(X; Y) \\ &= \sum_{i,j} p(i, j) \log \frac{p(i, j)}{p_X(i) p_Y(j)} \end{aligned} \tag{7}$$

where

- N is the number of total points
- i, j are the respective bins of X and Y
- $p_X(i) \approx \frac{n(i)}{N}$ is the fraction of the number of points in bin i to the total number of points
- $p_Y(j) \approx \frac{n(j)}{N}$ is the fraction of the number of points in bin j to the total number of points
- $p(i, j) \approx \frac{n(i,j)}{N}$ is the fraction of the number of points in the intersection of bin i and j to the total number of points

Similarly, the entropy of a continuous variable in (1) can be computed using bins.

$$\begin{aligned} H(X) &= \sum_{x \in X} p(x) \log \frac{1}{p(x)} \\ &\approx H_{\text{binned}}(X) \\ &= -\sum_i p_X(i) \log p_X(i) \end{aligned} \tag{8}$$

In practical applications, the study applies binning methods to estimate the MI value of continuous variables. While the number of bins can influence MI calculation accuracy, the study found that the impact is minimal when an appropriate bin count is selected. In feature engineering, the MI values from binning closely align with those from the actual distribution. Thus, using binned MI provides a reliable estimation of mutual information without significantly affecting the results.

	bins	AUC	F1 Score	Precision	Recall
1	without bins	0.650087	0.375581	0.275504	0.589841
2	5	0.645166	0.379959	0.281623	0.58381
3	7	0.651583	0.378524	0.27627	0.600952
4	10	0.650283	0.376031	0.273687	0.600635
5	20	0.650177	0.37597	0.275439	0.592063
6	100	0.6494	0.376134	0.275546	0.592381

TABLE I
BINS INFLUENCE COMPARISON OF MODEL PERFORMANCE METRICS.

III. COMPUTE THE MUTUAL INFORMATION MATRIX OF FEATURES

To find the mutual information between any two features, we will need to compute the mutual information matrix. There are several technical issues to resolve.

A. Normalize the Mutual Information

Mutual information between two variables can be very large if the marginal entropies are also large, which may misleadingly imply a strong relationship. To standardize mutual information, we can consider it analogous to covariance. Just as large variances in two variables can result in a large covariance, this does not inherently indicate a strong linear relationship unless the covariance is scaled by the marginal variances.

One way to normalize the mutual information [6] between the interval $[0, 1]$ is:

$$\widehat{I_{\text{binned}}}(X;Y) = \frac{2 \times I_{\text{binned}}(X;Y)}{H_{\text{binned}}(X) + H_{\text{binned}}(Y)} \tag{9}$$

Because of the symmetry of mutual information,

$$I_{\text{binned}}(X;Y) = I_{\text{binned}}(Y;X) \tag{10}$$

We have the symmetry of normalized mutual information,

$$\widehat{I_{\text{binned}}}(X;Y) = \widehat{I_{\text{binned}}}(Y;X) \tag{11}$$

B. Determine the Number of Bins

There is no ”best” number of bins, and different bin sizes can reveal different features of the data. In this study we take the square root of one fifth of the number of data points in the sample and rounds to the next integer. [7]

$$k = \lfloor \sqrt{N/5} \rfloor \tag{12}$$

where N is the number of data points.

C. The Mutual Information Matrix

Consider a data set with n features,

$$\{X_1, X_2, X_3, \ldots, X_n\}$$

A mutual information matrix can be defined as,

$$\widehat{\mathcal{I}}_n = \begin{bmatrix} \widehat{I_{\text{binned}}}(X_1;X_1) & \ldots & \widehat{I_{\text{binned}}}(X_1;X_n) \\ \vdots & \ddots & \\ \widehat{I_{\text{binned}}}(X_n;X_1) & & \widehat{I_{\text{binned}}}(X_n;X_n) \end{bmatrix} \tag{13}$$

Because of the symmetry as shown in equation (11), this mutual information matrix must be symmetric, similar to correlation matrix. In addition, because of the normalization, the diagonal of the matrix are ones. Therefore, the mutual information matrix will look like,

$$\widehat{\mathcal{I}}_n = \begin{bmatrix} 1 & \ldots & \widehat{I_{\text{binned}}}(X_1;X_n) \\ \vdots & \ddots & \\ \widehat{I_{\text{binned}}}(X_1;X_n) & & 1 \end{bmatrix} \tag{14}$$

IV. Mutual Information Reduction Techniques

In feature engineering, when two features exhibit high mutual information, it indicates that they contain redundant information. The techniques discussed in this section aim to eliminate mutual information from one feature, ensuring that the modified feature has minimal mutual information with the other. This modification can reveal hidden information to the machine learning models. When feature selection and creation are executed effectively, this approach can enhance the model's performance.

A. Correlation and Mutual Information

Mutual information measures the dependency between random variables and is frequently compared to linear correlation, as it can also capture nonlinear dependencies. In this section, we will explore a simplified example to examine the relationship between correlation and mutual information. While the conclusion is based on this simplified case, it provides a useful guideline for reducing mutual information between two features without losing critical information.

Consider a univariate Gaussian random variable X, the entropy is,

$$H(X) = \frac{1}{2}\ln(2\pi e \sigma_x^2). \tag{15}$$

The joint entropy of two jointly Gaussian random variables (X, Y) is,

$$H(X,Y) = \frac{1}{2}\ln((2\pi e)^2(\sigma_x^2\sigma_y^2 - \sigma_{xy}^2)), \tag{16}$$

where σ_{xy} is the covariance of X and Y.

Then the mutual information of X and Y is,

$$\begin{aligned} I(X;Y) &= H(X) + H(Y) - H(X,Y) \\ &= \frac{1}{2}\ln\left(\frac{\sigma_x^2\sigma_y^2}{\sigma_x^2\sigma_y^2 - \sigma_{xy}^2}\right) \\ &= \frac{1}{2}\ln\left(\frac{\sigma_y^2}{\sigma_y^2 - \sigma_{xy}^2/\sigma_x^2}\right) \\ &= -\frac{1}{2}\ln\left(\frac{\sigma_y^2 - \sigma_{xy}^2/\sigma_x^2}{\sigma_y^2}\right) \\ &= -\frac{1}{2}\ln\left(1 - \left(\frac{\sigma_{xy}}{\sigma_x\sigma_y}\right)^2\right) \\ &= -\frac{1}{2}\ln\left(1 - \rho_{(X,Y)}^2\right) \end{aligned} \tag{17}$$

If $\rho_{(X,Y)} = 0$, X and Y are independent and the mutual information is 0 . If $\rho_{(X,Y)} = \pm 1$, X and Y are perfectly correlated and the mutual information is infinite.

B. Control the Correlation

The equation (17) indicates that one can control the correlation between two variable X and Y to adjust the mutual information.

Consider two variables X and Y. Let's create a new feature

$$X' = X - \alpha Y \tag{18}$$

where α is a constant.

The expression $X - \alpha Y$ represents a linear transformation of the variable Y. The term $-\alpha Y$ indicates that the influence of Y on the correlation with X is scaled by the constant α and reversed in direction (if α is positive).

The correlation coefficient $\rho(Y, X - \alpha Y)$ can be calculated as follows,

$$\rho(Y, X - \alpha Y) = \frac{\text{Cov}(Y, X - \alpha Y)}{\sigma_Y \sigma_{X-\alpha Y}} \tag{19}$$

where

- $\text{Cov}(Y, X - \alpha Y)$ is the covariance between Y and $X - \alpha Y$,
- σ_Y is the standard deviation of Y,
- $\sigma_{X-\alpha Y}$ is the standard deviation of $X - \alpha Y$.

In equation (19), we can derive,

$$\begin{aligned}\text{Cov}(Y, X - \alpha Y) &= \text{Cov}(Y, X) - \alpha \text{Cov}(Y, Y) \\ &= \text{Cov}(Y, X) - \alpha \sigma_Y^2\end{aligned} \tag{20}$$

$$\sigma_{X-\alpha Y} = \sqrt{\sigma_X^2 + \alpha^2 \sigma_Y^2 - 2\alpha \text{Cov}(X, Y)} \tag{21}$$

Rewrite equation (19) to,

$$\begin{aligned}\rho(Y, X - \alpha Y) &= \frac{\text{Cov}(Y, X - \alpha Y)}{\sigma_Y \sigma_{X-\alpha Y}} \\ &= \frac{\text{Cov}(Y, X) - \alpha \sigma_Y^2}{\sigma_Y \sqrt{\sigma_X^2 + \alpha^2 \sigma_Y^2 - 2\alpha \text{Cov}(X, Y)}}\end{aligned} \tag{22}$$

This equation indicates that the correlation $\rho(Y, X - \alpha Y)$ is generally influenced by the covariance between X and Y, the value of α, and the variances of X and Y.

C. Mutual Information Reduction Techniques

The objective is to reduce the mutual information between features. Given a variable X, the study creates a new feature X' that minimizes mutual information with Y.

The parameter α controls the correlation, thus affecting the mutual information. The optimal value of α can be determined by:

$$\alpha^* = \arg\min_{\alpha} \widehat{I_{\text{binned}}}(X - \alpha Y; Y) \tag{23}$$

For Gaussian variables (X, Y), the analytical solution is:

$$\alpha^* = \frac{\text{Cov}(X, Y)}{\sigma_Y^2} \tag{24}$$

In practice, datasets often deviate from Gaussian distribution. Nevertheless, the analytical solution can serve as an efficient initial estimate, reducing mutual information (MI) significantly, although it is not guaranteed to be globally optimal.

Example

The dataset is the loan default dataset at https://github.com/MarcyChen-ruixin/MI_Reduction.

The dataset includes a variety of features related to loan applicants, such as demographic details, educational background, credit card usage, spending behavior, and loan inquiry frequency. The primary objective is to analyze the relationship between these features and loan default, aiming to predict the likelihood of a borrower defaulting on their loan.

Logistic regression models will be built to predict the loan default. Performance will be evaluated at the end. In this example, we select 5 columns [CD162, CD164, CD166, CD167, CD172] to demonstrate the idea. The initial mutual information matrix is shown below:

	CD162	CD164	CD166	CD167	CD172
CD162	1.0000	0.5998	0.1969	0.1787	0.3377
CD164	0.5998	1.0000	0.1986	0.1982	0.3495
CD166	0.1969	0.1986	1.0000	0.6145	0.2662
CD167	0.1787	0.1982	0.6145	1.0000	0.2310
CD172	0.3377	0.3495	0.2662	0.2310	1.0000

TABLE II
Mutual information matrix for 5 features.

Using the analytical solution α^*, we generate a new feature:

$$\text{CD167}' = \text{CD167} - \alpha^* \text{CD166} \tag{25}$$

The new feature successfully reduces the mutual information as shown below:

	CD162	CD164	CD166	CD167	CD167'
CD162	1.0000	0.5998	0.1969	0.1787	0.0826
CD164	0.5998	1.0000	0.1986	0.1982	0.0571
CD166	0.1969	0.1986	1.0000	0.6145	0.0582
CD167	0.1787	0.1982	0.6145	1.0000	0.0135
CD167'	0.0826	0.0571	0.0582	0.0135	1.0000

TABLE III
Updated mutual information matrix with new feature CD167'.

Similarly, the study creates another new feature:

$$\text{CD164}' = \text{CD164} - \alpha^* \text{CD162} \tag{26}$$

The updated mutual information matrix confirms the effectiveness of the analytical solution.

	CD162	CD164	CD166	CD167	CD164'
CD162	1.0000	0.5998	0.1969	0.1787	0.0715
CD164	0.5998	1.0000	0.1986	0.1982	0.0248
CD166	0.1969	0.1986	1.0000	0.6145	0.0251
CD167	0.1787	0.1982	0.6145	1.0000	0.0125
CD164'	0.0715	0.0248	0.0251	0.0125	1.0000

TABLE IV
Updated mutual information matrix with new feature CD164'.

V. Combine with WOE in Feature Engineering

A critical component of this study is the application of Weight of Evidence (WOE) transformation [8] on the new feature by mutual information reduction. WOE is a technique used to convert categorical variables into numerical values, capturing the predictive power of each category more effectively. This transformation is particularly useful for logistic regression models, which are employed in this study for their suitability in binary classification tasks.

WoE was originally developed for the credit and financial sectors to improve the predictive power of models assessing loan default risk. In other words, it helps predict the likelihood that a loan made to an individual or institution will result in a loss. WoE quantifies the effectiveness of a grouping method in distinguishing between good and bad credit risks (i.e., default vs. non-default).

Since the loan data contains many features with overlapping information, mutual information reduction techniques are used to isolate the shared information among features. This ensures that each feature retains unique information, which can enhance model performance.

Another reason to use WOE is that while mutual information reduction creates new features with less redundancy, these features are linear combinations of the originals. This won't add value in linear or logistic regression since the scaling factors merge into model coefficients. WOE transforms these features through binning, capturing the relationship between feature ranges and the target variable.

A. Create WOE-transformed Features

There are steps to create WOE features,

- Bin the Continuous Variables (or Group Categories for Categorical Variables):
  For continuous variables, the first step is to group the data into discrete bins or intervals. For categorical variables, group the categories into meaningful segments, if necessary.
- Calculate the Distribution of the Target Variable for Each Bin:
  For each bin (or category), calculate the proportion of events (e.g., defaults) and non-events (e.g., non-defaults) in the dataset. This gives the event rate and the non-event rate for each bin.

$$\begin{aligned}&\text{Event Rate}(P(1|\text{Bin}))\\&=\frac{\text{Number of Events in the Bin}}{\text{Total Number of Events in the Dataset}}\\&\text{Non-Event Rate}(P(0|\text{Bin}))\\&=\frac{\text{Number of Non-Events in the Bin}}{\text{Total Number of Non-Events in the Dataset}}\end{aligned} \tag{27}$$

- Calculate Weight of Evidence (WOE) for Each Bin:
  WOE measures the strength of the separation between the event (e.g., defaults) and non-event (e.g., non-defaults) distributions in each bin. The formula for calculating WOE is:

$$\text{WOE} = \ln\left(\frac{\text{Event Rate}(P(1|\text{Bin}))}{\text{Non-Event Rate}(P(0|\text{Bin}))}\right) \tag{28}$$

  This formula calculates the natural logarithm of the ratio of the event rate to the non-event rate for each bin. Then the WOE features can be used in the machine learning model.

B. Performance Improvement

With binning and WOE transformation for the mutual information reduced feature, the logistic regression shows an improvement in the prediction.

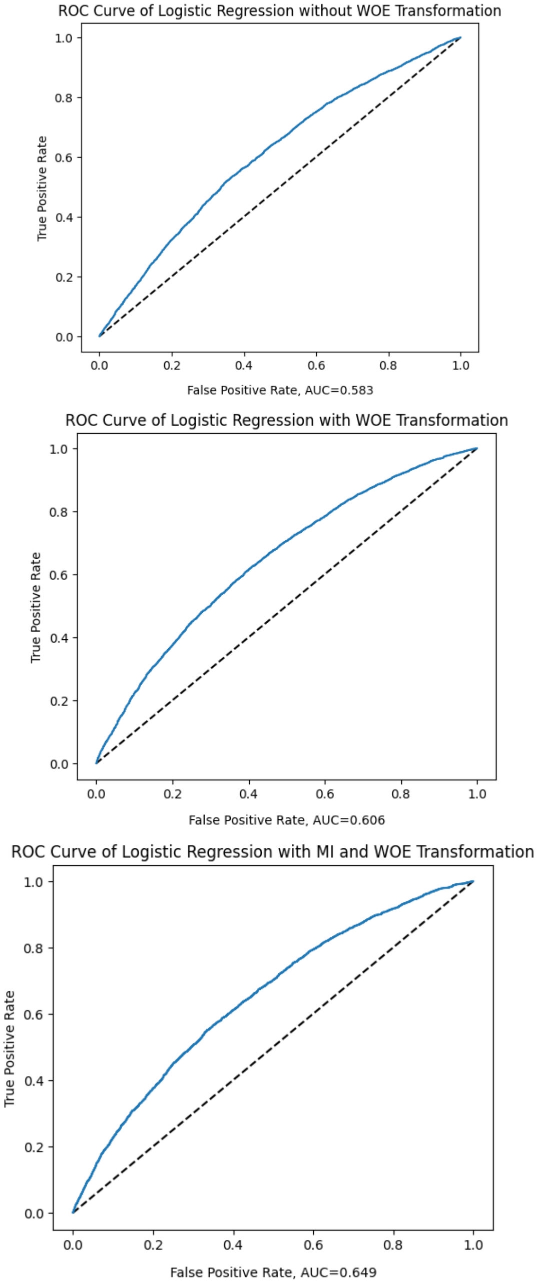

Fig. 1. Model Performance Improvement

	Logistic Regression	AUC	F1 Score	Precision	Recall
1	With MI and WOE	0.6487	0.3772	0.2774	0.5889
2	With WOE	0.6059	0.3751	0.2745	0.5918
3	Without WOE	0.5828	0.3485	0.2645	0.5107

TABLE V
COMPARISON OF MODEL PERFORMANCE METRICS.

The logistic regression model with WOE transformation on mutual information reduced features outperforms the model without mutual information reduction across all metrics (AUC, F1 Score, Precision, and Recall).

Feature 1	Feature 2	MI
PA022	PA023	0.749863
PA022	PA029	0.270738
PA023	PA029	0.205334
TD001	TD005	0.253645
TD005	TD009	0.332073
TD006	TD010	0.414127
TD006	TD014	0.233298
TD010	TD014	0.455980

TABLE VI
HIGH MI FEATURE PAIRS

Feature	Rank (No MI)	Importance (No MI)	Rank (With MI)	Importance (With MI)
CR009	1	0.156821	1	0.136308
AP001	2	0.133107	2	0.117804
CR019	3	0.094143	4	0.082965
TD009	4	0.076094	7	0.060663
AP008	5	0.064646	9	0.058718
TD005	6	0.060509	5	0.080534
TD014	7	0.059823	13	0.043442
TD001	8	0.058256	3	0.085047
PA029	9	0.052356	12	0.043738
PA022	10	0.050463	10	0.051020
TD010	11	0.046147	8	0.059726
TD006	12	0.043018	6	0.073186
CR015	13	0.041674	14	0.036396
PA023	14	0.040845	11	0.050556
AP003	15	0.022098	15	0.019898

TABLE VII
FEATURE IMPORTANCE COMPARISON WITH AND WITHOUT MI

Overall, the mutual information reduction improves the model's ability to correctly classify defaults and non-defaults, making it a better choice for this specific dataset. Further improvements can still be made by experimenting with additional feature engineering, model tuning, and balancing techniques.

By reducing MI, the changes in feature importance rankings reflect an enhanced separation of feature contributions. This process not only mitigates redundancy but also boosts the significance of the newly engineered features, making them more predictive in nature. As a result, the performance gains are especially evident in complex, nonlinear models like Random Forests, where the reduction of feature overlap leads to better accuracy, improved generalization, and lower risk of overfitting. This method proves particularly advantageous for datasets with high feature interdependence. As it systematically reduces feature overlap and ensures that each feature contributes uniquely to the prediction of the target variable.

Incorporating above-mentioned future enhancements will help to further improve the accuracy, robustness, and interpretability of the loan default prediction models. By leveraging advanced techniques and additional data sources, the study can provide more actionable insights and contribute to better risk management in the financial sector.

VI. Conclusions and Discussions

In this paper, we explored the application of mutual information (MI) in feature engineering, particularly focusing on feature selection and creation. MI offers a robust mechanism for evaluating the dependency between features and the target variable, capturing both linear and nonlinear relationships. While traditional approaches mainly focus on maximizing MI for selecting relevant features, our study introduced a novel technique that reduces mutual information between features. This reduction approach allows for the removal of redundant information, ensuring that each feature contributes unique and valuable information to the predictive model.

We demonstrated that applying mutual information reduction techniques can enhance model performance by minimizing redundant information across features. This was illustrated with a loan default dataset, where our brute-force method was able to determine the optimal parameter to minimize mutual information between features. The results showed a significant reduction in redundancy, as reflected by the updated mutual information matrix.

Furthermore, we integrated the Weight of Evidence (WOE) transformation to enhance the predictive power of the newly created features. WOE is particularly effective for binary classification problems, such as loan default prediction, and its combination with MI reduction leads to improved model performance, particularly in logistic regression models. This integration provides better risk management insights by allowing models to focus on non-redundant, highly informative features.

Overall, mutual information reduction techniques, when combined with WOE transformation, show great promise in improving machine learning model accuracy and interpretability. Future work may involve exploring more advanced methods for optimizing MI reduction, as well as extending these techniques to other domains and datasets.

References

[1] Hall, M. A., & Smith, L. A. (1999, May). Feature selection for machine learning: comparing a correlation-based filter approach to the wrapper. In FLAIRS conference (Vol. 1999, pp. 235-239).

[2] Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis & Machine Intelligence, (8), 1226-1238.

[3] Yu, L., & Liu, H. (2003). Feature selection for high-dimensional data: A fast correlation-based filter solution. In Proceedings of the 20th international conference on machine learning (ICML-03) (pp. 856-863).

[4] Zhao, Z., & Liu, H. (2007, June). Spectral feature selection for supervised and unsupervised learning. In Proceedings of the 24th international conference on Machine learning (pp. 1151-1157). ACM.

[5] Shannon, Claude E. (July 1948). ”A Mathematical Theory of Communication”. Bell System Technical Journal. 27 (3): 379–423. doi:10.1002/j.1538-7305.1948.tb01338.x. hdl:10338.dmlcz/101429.

[6] Guo X, Zhang H, Tian T (2018) Development of stock correlation networks using mutual information and financial big data. PLoS ONE 13(4): e0195941. https://doi.org/10.1371/journal.pone.0195941

[7] Estimating number of bins when computing mutual information, https://stats.stackexchange.com/questions/179674/number-of-bins-when-computing-mutual-information/181195#181195.

[8] Weed, D.L. (2005), Weight of Evidence: A Review of Concept and Methods. Risk Analysis, 25: 1545-1557. https://doi.org/10.1111/j.1539-6924.2005.00699.x