Department of Statistics

LUND UNIVERSITY

Weight of evidence transformation in credit scoring models: How does it affect the discriminatory power?

Rickard Persson

PAPER INFO

Level: Master thesis, 15 ETCS
Received: 29 September 2021
Revised: 18 October 2021
Supervisors: Peter Gustafsson, Lund University & Ola Jönsson, Nordea
Keywords: Weight of evidence, WOE

ABSTRACT

Weight of evidence (WOE) transformation has been used for several decades in the credit industry. However, despite its widespread use, it has, surprisingly, been an overlooked approach in published literature. In this paper, we, therefore, investigate what effect WOE transformation has on the discriminatory power of a credit-scoring model. Our results suggest that using WOE transformation with logistic regression decreased the discriminatory power across a majority of the evaluation metrics compared to the models that did not use WOE transformed variables. Moreover, using an information value for variable selection did not provide any benefits over using the backward selection technique. However, applying support vector machine, we found mixed results depending on the preferred evaluation metric. Using an information value seems to provide some benefits regarding variable selection compared to the recursive feature elimination technique.

1. Introduction

In financial institutions, one of the main business activities is to create loans by granting credit. In the past, credit decisions were made based on qualitative judgments made by experts, whereas today, they are based mostly on statistical **credit-scoring models** [1] (Marques et al. 2013 p.1384; Thomas 2000 p.151).

A statistical method that has been used for a long time in the credit-scoring model industry is the **weight of evidence** [2] (WOE) transformation of variables (Abdou 2009 p.11403). WOE transformation deals, for example, with missing values in the data and allows for variable selection through an information value (IV) that can be calculated after variables have undergone WOE transformations.

The credit-scoring research has, for the most part, focused on comparing the **discriminatory power**, *a model's ability to discriminate between good and bad loan applicants*, of different classification techniques (Marvi et al. 2008; Blochwitz et al 2005). Some earlier research suggests that neural networks are superior compared to other techniques (Tam & Kiang 1992; Desai et al. 1996). However, Yobas et al. 2000 compare the model performance of the linear discriminant analysis, neural networks, decision trees, and genetic algorithms. In their study, the linear discriminant analysis performs the best. On the other hand, Ong et al. (2005) use two datasets and compare the performance of logistic regression (LogR), neural networks, and genetic-programming. Based on their dataset they find that genetic algorithms outperform LogR and neural networks, but the difference is negligible (Ong et al. 2005 p.46). Finlay (2009) compares the performance of generic algorithms to LogR and finds that they produce similar results (Finlay 2009 p.9069). Finally, Zurada et al. (2014) use five datasets and considers methods like neural networks, decisions trees, LogR, support vector machine (SVM), and k-nearest neighbor. Overall, Zurada (2014) finds that no single method consistently outperformed any other method, which is consistent when one summarizes the overall results across several published papers.

Another feature in credit-scoring modeling that has been studied is variable selection (e.g., Wang & Huang 2009; Chi & Hsu 2012). For example, Bernhardsen & Larsen (2007) used a genetic-programming technique to find the optimal explanatory variables. Other statistical methods like forward and backward stepwise selection techniques have also been used (Abdou & Pointon 2011 p.66).

Despite that WOE transformation has been used for several decades in the credit industry, it has, surprisingly, been an overlooked approach in published literature (Abdou 2009 p.11402). It is reasonable to assume WOE may impact the probability of default and the model's discriminatory power because it allows for variable selection through an IV. It may also cause overfitting of the sample data because variable coding is based on the dependent variable and, as a result, cause poor performance on the out-of-sample data (Gool et al. 2012 p.107). Furthermore, banks play a very critical role in the economy and it is of importance for them to have sound credit-rating systems (Elbannan 2017 p. 225).

The frequent use of WOE in the industry, the importance of banks to have sound models, and the lack of research on the effect on the discriminatory power highlights a relevant research area that needs more attention. In this paper, **we investigate what effect WOE transformation has on the discriminatory power of a credit-scoring model**.

The dataset used in this study is available at the UCI machine-learning repository. We found that using WOE transformation with LogR decreased the discriminatory power on a majority of the evaluation metrics compared to models that did not use WOE transformed variables. Moreover, using an IV for variable selection did not provide any benefits over using the backward selection technique. However, applying SVM, we found mixed results depending on the preferred evaluation metric, and using an information value seems to provide some benefits regarding variable selection compared to the recursive feature elimination (RFE) technique.

The remainder of this paper is organized as follows. Section 2 presents WOE. Section 3 presents the data and methods, Section 4 gives the results and Section 5 presents the conclusions.

2. Weight of evidence

The underlying theory of WOE was provided by Good (1950), and the expression describes whether the evidence in favor or against some hypothesis is more or less strong (Bernardo et al. 1985 p.249; Good 1984).

The use of WOE involves a transformation of data that requires binning, which is a process that transforms a continuous or a categorical variable into set groups or bins (Zeng 2014 p.3229). According to Siddiqi (2006 p.80), a good binning strategy should follow these guidelines:

1. Missing values should be grouped separately.
2. A bin should contain ≥ five percent of all observations.
3. No bin should have zero good or bad loans.

[1] A **credit-scoring model** is defined as a risk management tool that assesses the creditworthiness, i.e., the ability to repay the loan, of a loan applicant by estimating her probability of default based on historical data (Frenandez Vidal & Barbon 2019 p.4).

[2] **WOE** transforms the values of a variable into discrete categories and assigns to each category a unique WOE. If any of the categories has a large proportion of defaulters compared to non-defaulters, the WOE value will be large which in turn tells us that the category separates the defaulters from non-defaulters well (Lin & Hsieh 2014 p.1).

It should also, as stated by Good (1969 p.141) and Baesens (2016 p.116), establish monotonic relationships between the independent and dependent variables. However, according to Siddiqi (2006 p.81) it is more about establishing a logical (not necessary linear) relationship that makes an operational sense. Experimenting with different groupings mostly eliminates illogical relationships (Ibid.).

The calculation process carried out as follows. Let's say that we have a dataset of *N* independent observations, where *Y* is a binary dependent variable that takes values *1 = default* and *0 = not default*. Let $X_1,..,X_p$ be a set of independent variables. Let $B_1,..,B_k$ be bins for the variable X_j. The WOE for the variable X_j in bin *i* is then defined as

$$Y = \begin{cases} 1 \; if \; default \\ 0 \; if \; not \; default \end{cases} \tag{1}$$

$$WOE_{ij} = \log\left(\frac{P(X_j \epsilon B_i | Y=1)}{P(X_j \epsilon B_i | Y=0)}\right)$$

where

$$P\left(X_j \in B_i \middle| Y = 1\right) = \frac{N_{X_j \epsilon B_i | Y=1}}{N_{X_j | Y=1}} =$$

$$\frac{Total\ number\ of\ defaulters\ in\ variable\ X_j\ in\ bin\ B_i}{Total\ number\ of\ defaulters\ in\ variable\ \ X_j}$$

$$P\left(X_j \in B_i \middle| Y = 0\right) = \frac{N_{X_j \epsilon B_i | Y=0}}{N_{X_j | Y=0}} =$$

$$\frac{Total\ number\ of\ non-defaulters\ of\ variable\ X_j\ in\ bin\ B_i}{Total\ number\ of\ non-defaulters\ of\ variable\ \ X_j}$$

WOE measures the strength of each grouped attribute, in separating defaulters and non-defaulters, where high negative values are equivalent to a high risk of default and vice versa (Siddiqi 2006 p.81).

After the WOE transformation, an IV for variable X_j is calculated as

$$IV_{X_j} = \sum_{i=1}^{k} \left(P\left(X_j \in B_i \middle| Y = 1\right) - P\left(X_j \in B_i \middle| Y = 0\right)\right) * WOE_{ij} \tag{2}$$

This IV measures the strength between the dependent and independent variables and it is used for variable selection. One rule of thumb is that a value < 0.02: is unpredictive, 0.02 to 0.1: is weak, 0.1 to 0.3: is medium, 0.3 to 0.5: is strong and whereas > 0.5: is a very strong prediction (Siddiqi 2006 pp.81-82). In (2), we see that, whereas the absolute values of the WOE are important, the difference between the WOE values of the groups is crucial to establishing differentiation (Ibid). The larger the difference between subsequent groups, the higher the differentiation ability of that characteristic becomes (Ibid.).

There are several reasons for doing WOE. First, it should establish a monotonic relationship to the dependent variable (Good 1969 p.141; Baesens et al. 2016 p.116). However, non-monotonic relationship can occur, which can be kept as long as the relationship can be explained (Siddiqi 2006 p.84). It also deals with missing values and outliers conveniently.

There are also some drawbacks. First, there may be a loss of information (variation) due to the binning procedure. Second, correlations between the independent variables are not taken into account. For example, there may be a strong correlation between some independent variables, which highlights the importance of doing data exploration before applying the technique.

3. Data & methods

3.1 Data

We use the Taiwan dataset that is available at the UCI machine-learning repository. A total of 30,000 observations are included in the dataset, along with 24 variables, out of which 23,363 are non-defaulters. We divide the dataset randomly into two groups, a training set with 70 percent of the observations, and a validation set with 30 percent of the observations. The models are fitted on the training set and then evaluated on the validation set. We repeat the aforementioned process 50 times and averaging the evaluation metrics to mitigate the impact of random sampling on the outcome (James et al. 2017 p.178).

3.2 Model classification techniques & models

LogR and SVM are used as classification techniques. The purpose of choosing two classification techniques is to see if WOE transformation behaves differently for each technique.

LogR models the probability that the response variable *Y* belongs to a particular category. The probability of default, π, is defined as

$$\pi = \frac{1}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}} \tag{3}$$

where $X_1, \ldots, X_p$ are independent variables and $\beta_0, \beta_1, \ldots, \beta_p$ are the regression coefficients.

The SVM is a generalization of the maximal margin classifier (James et al. 2017 p.337). The maximal margin classifier constructs a hyperplane that is the farthest from the training observations (James et al. 2017 p.341). That is, we can compute the smallest distance for the training observations; the smallest distance is the minimal distance from the observations to the hyperplane and it is known as the *margin* (Ibid.).

The support vector classifier (SVC) classifies test observations based on which side of a hyperplane they lie on (James et. al 2017 p.345). The hyperplane is chosen where it separates most of the observations correctly, but some observations would be misclassified (Ibid). If *M* is the width of the margin, it is the solution to the following optimization problem

$$\underset{\beta_0, \beta_1, \ldots, \beta_p, \epsilon_1, \ldots, \epsilon_n}{\text{maximize}} \; M \tag{4}$$

$$subject\ \text{to} \sum_j^p \beta_j^2 = 1\,, \tag{5}$$

$$y_i\left(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}\right) \geq M(n - \epsilon_i), \tag{6}$$

$$\epsilon_i \geq 0, \sum_{i=1}^{n} \epsilon_i \leq C \tag{7}$$

where C is a nonnegative tuning parameter, $\epsilon_1, \ldots \epsilon_n$ are slack variables (James et al. 2017 p.346).

The solution to the problem (4) - (7) involves the inner products of two observations. The inner products of two r - vectors a and b is defined as $\langle a, b\rangle = \sum_{i=1}^{r} a_i b_i$ (James et al. 2017 p.351). Thus the inner product of two observations is given by $\langle x_i, x_k\rangle = \sum_{j=1}^{r} x_{ij}, x_{kj}$ (Ibid). It can be shown that the SVC can be written as:

$$f(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i \langle x, x_i\rangle \tag{8}$$

Because α_i is zero, if the observation is not on the support vector the solution function (8) can be written as

$$f(x) = \beta_0 + \sum_{i \epsilon \mathcal{S}} \alpha_i \langle x, x_i\rangle, \tag{9}$$

where $\mathcal{S}$ is the number of indices on the support points. In the SVM, we replace the inner products with a generalization, which takes the form $K(x_i, x_k)$, where *K* is some function that we call a kernel (James et al. 2017 p.352). When a SVC is combined with a non-linear kernel it is called a SVM (support vector machine) and the solution function becomes:

$$f(x) = \beta_0 + \sum_{i \epsilon \mathcal{S}} \alpha_i K\,(x, x_i) \tag{10}$$

Table 1 presents the models that we apply in this paper. In the first model, we do a WOE transformation of the variables and select variables with an IV value ≥ 0.3. In the second model, we select variables with the backward selection (BS) method based on Akaike information criterion (AIC) without using WOE transformed variables. In model 3 the variable selection is based on IV ≥ 0.3, but the model is estimated without WOE transformed variables. In the fourth model, the variable selection is based on BS with WOE transformed variables. Models 1-4 are estimated with LogR.

Model 5 and 7 resemble models 1 and 3, the only difference is that they are estimated with the SVM technique. Model 6 is an RFE-SVM without WOE transformed variables and model 8 is an RFE-SVM with WOE transformed variables. RFE is usually used when performing variable selection with SVM (Sanz et al. 2018)

Table 2 summarizes the models that will be compared and the effects we can evaluate by comparing them. By comparing Model 1 to Model 3, we can evaluate the effect of using WOE transformed variables because the only difference between the two will be the WOE transformation. Model 2 would be compared to Model 4

Table 1
Models

Model	Estimation technique	Variable selection methodology	WOE transformed variables
Model 1	Log R	IV	Yes
Model 2	Log R	Backward selection	No
Model 3	Log R	IV	No
Model 4	Log R	Backward selection	Yes
Model 5	SVM	IV	Yes
Model 6	SVM	RFE	No
Model 7	SVM	IV	No
Model 8	SVM	RFE	Yes

Table 2
Model comparisons

Model	Evaluates
Model 1 vs. Model 3	The effect of WOE transformation when IV is used as a variable selection method
Model 2 vs. Model 4	The effect of WOE transformation when BS is used as a variable selection method
Model 1 vs. Model 4	Variable selection method IV vs. BS with WOE transformed variables
Model 2 vs. Model 3	Variable selection method IV vs. BS without WOE transformed variables
Model 5 vs. Model 7	The effect of WOE transformation when IV is used as a variable selection method
Model 6 vs. Model 8	The effect of WOE transformation when RFE is used as a variable selection method
Model 5 vs. Model 8	Variable selection method IV vs. RFE with WOE transformed variables
Model 6 vs. Model 7	Variable selection method IV vs. RFE without WOE transformed variables

Table 3
Measures of discriminatory power

Measure	Definition	Optimal value
Accuracy rate (AR)	(TP+TN) / (TP+TN+FP+FN)	1
AUC		1
Sensitivity (SE)	TP / (TP+FN)	1
Specificity (SP)	TN / (TN+FP)	1
F1-score	2TP / (2TP+FP+FN)	1

Note: TP = true positives, TN = true negatives, FP = False positives, FN = false negative

because we can examine the effect of WOE transformed variables when BS is used as a variable selection method. Moreover, the setup of models allows us to examine the use of WOE in terms of variable selection as well, because the difference between Model 1 and Model 4 and Model 2 and Model 3 is the variable selection technique. Model 5 – 8 will be compared in a similarly way as Model 1 – 4.

3.3 Measures of discriminatory power & evaluation

The most common metrics in practice and academia of measuring a model's discriminatory power are the accuracy ratio (AR), and the area under the receiving operating curve (AUC) (Al Marques et al. 2013 p.1393; Lingo & Winkler 2008 p.2). Two other common metrics are sensitivity (SE), which measures the accuracy of true defaulters, and specificity (SP), which measures the accuracy of true non-defaulters (Ibid.).

Empirical and theoretical evidence suggests that the AR is strongly biased with an imbalanced dataset. In light of the fact that we are dealing with an unbalanced dataset, we can also use the F1-score, because it was created to deal with imbalanced datasets (Guo et al. 2018 p.5). The F1-score is a harmonic mean between SE and precision (positive predictive value). Table 3 summarizes the metrics that we use to investigate how WOE transformation affects the discriminatory power.

We want to do a pairwise comparison that examines the effect of the WOE transformation (see Table 2). One common method used in the literature for comparing the results from different methods of variable transformation, selection, or classification techniques in datasets is the Wilcoxon signed rank test (Benavoli et al. 2016; Demsar 2006).

The Wilcoxon signed-rank test compares the mean differences of an evaluation metric between two models for each test dataset, and ranks both the positive and negative differences.

The calculation of the Wilcoxon signed-rank test is carried out as follows. Let d_i be the difference in an evaluation metric of the ith out of N sample dataset. The differences are ranked according to their absolute values, if there is a tie, then, average ranks are assigned. Let R^+ be the sum of ranks for the nonnegative differences, and R^- the sum of ranks for the non-positive differences:

$$R^+ = \sum_{d_i>0} rank(|d_i|) + \frac{1}{2}\sum_{d_i=0} rank(|d_i|) \quad (11)$$

$$R^- = \sum_{d_i<0} rank(|d_i|) + \frac{1}{2}\sum_{d_i=0} rank(|d_i|)$$

Let W be the larger of the sums, $W = \max(R^+, R^-)$ and for a larger number of dataset, the statistics

$$z = \frac{W - \frac{1}{4}N(N+1)}{\sqrt{\frac{1}{24}N(N+1)(2N+1)}} \quad (12)$$

is distributed approximately normally.

4. Results & discussion

In this section, we present the results and discussion. Table 4 shows the mean values of the evaluation metrics of the discriminatory power. Table 5 shows the summary of the Wilcoxon-signed rank tests for the AR and the F1-score. The Wilcoxon-sign ranked tests for the other evaluation metrics can be provided upon request, but the results of these will not change the overall results and conclusions.

As seen in Table 4, Model 2 and Model 3 the very models without WOE transformed variables perform better, on average, than Model 1 and Model 4 the models with WOE transformed variables across a majority of the evaluation metrics. Further, analyzing Table 5 suggests that the differences between Model 1 and Model 3 and between Model 2 and Model 4 are significant for the AR and F1-score.

Table 4 and 5 suggest that the BS method performs better, on average, in a majority of the evaluation metrics than the IV method in terms of variable selection when Model 1 is compared to Model 4. Further, comparing Model 2 and Model 3 suggest that using an IV for variable selection gives no significant benefit over the BS technique in terms of increased model performance.

The overall combined results of Table 4 and 5 indicate that WOE transformation of variables decreased the discriminatory power of a model in a majority of the evaluation metrics when LogR is applied as a classification technique. Moreover, using an IV for variable selection provides no significant benefit in terms of increased model performance.

Table 4 shows that Model 5 is the best model, on average, across most evaluation metrics of discriminatory power when SVM is used as an estimation technique. Inspecting Table 5, we can observe when comparing Model 5 to Model 7, that, in terms of the AR, there is no significant benefit of using WOE. But when the F1-score is preferred, there is an advantage. Further, when comparing Model 6 to Model 8, we notice that the difference for the AR is only significant at the 10 percent level and insignificant for the F1-score.

The SVM models also provide mixed results regarding the use of an IV as a variable selection method compared to the RFE method. Evaluating Model 6 and Model 7 we notice that there is a significant difference in terms of the AR ratio, but not in the F1-score. Using the AR ratio as an evaluation metric suggests that there is a small benefit of using an IV as a variable selection method compared to RFE. Further, comparing Model 5 to Model 8 suggests that there is no significant difference between the models when the AR ratio is used, but there is a difference when the F1-score is used, indicating that using IV as a variable selection method is preferred over the RFE.

Overall, considering the AR as an evaluation metric, our results suggest that using WOE transformation does not improve the discriminatory power when SVM is used as a classification technique. However, using the F1-score as an evaluation metric appears to be beneficial. The general results regarding variable selection, considering all evaluation metrics, suggest that using an IV for

Table 4
Result mean values of measures of discriminatory power

Measures	Model 1 (IV,WOE,LogR)	Model 2 (BS,LogR)	Model 3 (IV, LogR)	Model 4 (BS,WOE,LogR)	Model 5 (IV,WOE,SVM)	Model 6 (RFE,SVM)	Model 7 (IV, SVM)	Model 8 (RFE,WOE,SVM)
Accuracy rate	0,8186	0,8202	0,8201	0,8191	0,8197	0,8191	0,8196	0,8194
AUC	0,6485	0,6544	0,6546	0,6511	0,6509	0,6461	0,6441	0,6470
Sensitivity	0,3442	0,3579	0,3586	0,3505	0,3489	0,3367	0,3301	0,3386
Specificity	0,9528	0,9509	0,9506	0,9516	0,9528	0,9556	0,9581	0,9554
F1-score	0,4555	0,4674	0,4677	0,4607	0,4600	0,4501	0,4461	0,4522

Note: The values are the means of 50 observations. The same sample training test sets are used across all models i.e. sample 1,..,50 have been used for each of the respective models. WOE values/intervals in the training sets have been used in the test sets.

Model 1 uses IV for variable selection, is estimated with LogR, and uses WOE transformed variables. Model 2 uses backward selection (BS) for variable selection, is estimated with LogR, and uses untransformed variables. Model 3 uses IV for variable selection, is estimated with LogR, and uses untransformed variables. Model 4 uses BS for variable selection, is estimated with LogR, and uses WOE transformed variables.

For all of the SVM models, we use a 3-fold cross-validation technique on the training set to determine the optimal parameters for the SVM. The optimal values for C that we test are 0.25,0.5,1,2 together with a polynomial kernel with 1 and 2 degrees. Model 5 uses IV for variable selection, is estimated with SVM, and uses WOE transformed variables. Model 6 uses RFE for variable selection, is estimated with SVM, and uses untransformed variables. Model 6 uses IV for variable selection, is estimated with SVM, and uses untransformed variables. Model 8 uses RFE for variable selection, is estimated with SVM, and uses WOE transformed variables.

Table 5
Summary of the Wilcoxon signed-ranks test

Model	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Model 2 $^{1.AR\ 2.F1\text{-}score}$	1. -5.7 ***	-					
	2. -6.1 ***						
Model 3 $^{1.AR\ 2.F1\text{-}score}$	1. -5.7 ***	1. 0.5	-				
	2. -6.1 ***	2. -0.9					
Model 4 $^{1.AR\ 2.F1\text{-}score}$	1. -3.0 ***	1. 5.7 ***	1. 4.5 ***	-			
	2. -5.7 ***	2. 6.0 ***	2. 5.8 ***				
Model 5 $^{1.AR\ 2.F1\text{-}score}$	1. -4.1 ***	1. 2.2 **	1. 1.2	1. -2.5 **	-		
	2. -1.9 *	2. 1.7	2. 1.4	2. 0.7			
Model 6 $^{1.AR\ 2.F1\text{-}score}$	1. -1.8 *	1. 5.1 ***	1. 4.5 ***	1. 0.0	1. 3.0 ***	-	
	2. 1.8 *	2. 4.6 ***	2. 4.8 ***	2. 3.9 ***	2. 3.8 ***		
Model 7 $^{1.AR\ 2.F1\text{-}score}$	1. -3.5 ***	1. 2.7 ***	1. 2.3 **	1. -2.1 **	1. -0.1	1. -3.8 ***	-
	2. 3.0 ***	2. 5.2 ***	2. 5.3 ***	2. 5.0 ****	2. 4.8 ***	2. 1.5	
Model 8 $^{1.AR\ 2.F1\text{-}score}$	1. -2.8 ***	1. 3.4 ***	1. 2.9 ***	1. -1.2	1. 1.2	1. -1.8 *	1. 1.6 *
	2. 1.0	2. 4.8 ***	2. 4.9 ***	2. 3.3 ***	2. 2.5 ***	2. -1.2	2. -2.9 ***

Note: *, **, *** significant at 10, 5 and 1 percent level. 1. Accuracy ratio. 2. F1-score.
The reported values are the z values from the normal approximation (see (12)).

variable selection increased the model performance compared to the RFE technique.

Evaluating the classification techniques, we notice that Model 2, Model 3, and Model 5 perform the best across all the evaluation metrics, but the difference in magnitude is small amongst them and only significant for the AR ratio between Model 2 and Model 5. Overall, our results indicate that using a more sophisticated classification technique does not increase the discriminatory power of a model significantly. Instead, we notice that LogR outperformed SVM, on average, in three out of four cases, i.e., if Model 1|2|3|4| is compared to Model 5|6|7|8 and considering the significance of the evaluation measures AR and F1-score. However, Abdou & Pointon (2011) point out that most studies that made a comparison found that more sophisticated methods performed better when AR is preferred as an evaluation measure. On the other hand, as pointed out by Marques et al. (2013), most articles do not contain any formal statistical test, which makes it difficult to assess whether the differences are significant.

As WOE deals with possibly linearity, we could expect that, if the data is non-linear, the models that are using WOE transformed variables should perform better, and we could also suspect that the SVM should perform better than the LogR technique. However, this is not the case according to the result of this paper. One explanation for this would be that the dataset is linear, and that is why the WOE does not provide any benefits over the regular methods. This is in line with Baesens et al. (2003), who noted that most credit data are only weakly non-linear, that is why, for example, that LogR yields similar performance like SVM.

We notice by analysing Table 4 that all models perform poorly when the sensitivity rate (the accuracy of true defaulters) is preferred as an evaluation metric, which probably depends on the imbalanced dataset. Therefore, the F1-score should be preferred because the AR is biased. Imbalanced datasets can also be dealt with by using, for example, a SMOTE (synthetic minority over-sampling technique) algorithm (Xue & Zhang 2016). But, this has not been done in this paper, and we leave it for future research.

So far, to the best of our knowledge, there are not many studies that have evaluated the effect of WOE transformation in the context of credit-scoring modelling. Banasik et al. (2003) compared a WOE model with a binary model and found that they produced similar results (Banasik et al. 2003 pp. 826- 827,830-831). On the other hand, Adbou (2009) found that WOE models preformed worse than other models when AR was used as an evaluation metric, but in terms of type II errors (One minus sensitivity), the WOE performed better than all other models (Adbou 2009 p.11412). Our results, however, suggest that the models without WOE transformed variables when LogR is used as an estimation technique perform better in terms of type II errors (One minus sensitivity see Table 4), while it is the opposite applying SVM.

5. Conclusion

WOE transformation has been used for several decades in the credit industry, however, it has, surprisingly, been an overlooked approach in published literature. In this paper, we investigate what effect WOE transformation has on the discriminatory power of a credit-scoring model. We use a real-world dataset of 30,000 observations, which we have divided into a training set and one validation set. Eight models are estimated where Model 1-4 applied LogR and Model 5-8 applied SVM as a classification technique. We repeat this procedure 50 times and average the evaluation metrics. The Wilcoxon sign ranked test is used to assess whether there are any significant differences between models that use WOE transformed variables or IV as a variable selection method compared to the models that do not use WOE transformed variables and use other variable selection techniques.

Our results suggest that using WOE transformation decreased the discriminatory power across a majority of the evaluation metrics compared to the models that did not use WOE transformed variables and that there is not any significant benefit of using an IV value for variable selection compared to the BS technique when LogR is applied as a classification technique. However, applying SVM, we found mixed results depending on the preferred evaluation metric. Weak or insignificant differences if the AR is used and significant differences if the F1-score is used. The general result also shows that using an IV is better for variable selection than the RFE technique.

As WOE deals with possible linearity, we could expect that if the data is non-linear, the WOE model should perform better. One explanation would be that the credit data used is already linear, and that is why the WOE does not provide any benefits over the regular methods. This is in line with Baesens et al. (2003), who noted that most credit data are only weakly non-linear, and that is why, for example, LogR yields similar performance to SVM.

There are some limitations to this paper. We have only used one dataset, therefore more studies must be carried out that use real data with all the variables that are normally accessible in real-world applications. Moreover, considering that our dataset is imbalanced, it may have had an impact on our results. Future research needs to take this issue into account.

Acknowledgments: *We thank Ola Jönsson for the research idea. We appreciate comments from Peter Gustafsson and Ola Jönsson have contributed to valuable comments. We are also thankful for comments from seminar participants at Lund University 2020-09-28, George Sikasote and Björn Holmquist.*

References

Abdou, H.A. 2009, "Genetic programming for credit scoring: The case of Egyptian public sector banks", Expert systems with applications, vol. 36, no. 9, pp. 11402-11417.

Abdou, H. & Pointon, J., 2011, “CREDIT SCORING, STATISTICAL TECHNIQUES AND EVALUATION CRITERIA: A REVIEW OF THE LITERATURE”, Intelligent Systems in Accounting, Finance and Management, 18(2-3), pp.59–88.

Baesens, B., Roesch, D. & Scheule, H. 2016, “Credit risk analytics: measurement techniques, applications, and examples in SAS”, Wiley, Hoboken, New Jersey.

Banasik, J., Crook, J. & Thomas, L. 2003, "Sample selection bias in credit scoring models", Journal of the Operational Research Society, vol. 54, no. 8, pp. 822-832.

Benavoli, A., Corani, G., & Mangili, F. 2016, “Should we really use post-hoc tests based on based on mean-ranks?” Journal of Machine Learning Research, 17(5):1–10

Bernardo, J.M., DeGroot, M.H., Lindley, D.V. & Smith, A.F.M. 1983, Bayesian statistics 2: Proceedings of the Second Valencia International Meeting

Bernhardsen, E. & Larsen, K. 2007, "Modelling credit risk in the enterprise sector--further development of the SEBRA model", Economic Bulletin, vol. 78, no. 3, pp. 102.

Blochwitz, S., Hamerle, A., Hohl, S., Rauhmeier, R. & Rousch, D. 2005, "Myth and reality of discriminatory power for rating systems", Wilmott, vol. 2005, no. 1, pp. 78-82.

Chi, B. & Hsu, C. 2012, "A hybrid approach to integrate genetic algorithm into dual scoring model in enhancing the performance of credit scoring model", Expert Systems With Applications, vol. 39, no. 3, pp. 2650-2661.

Demsar, J 2006, “Statistical comparisons of classifiers over multiple data sets”, Journal of Machine Learning Research 7(1)

Desai, V. S., Crook, J. N., & Overstreet, G. A. 1996, “A comparison of neural networks and linear scoring models in the credit union environment”, European Journal of Operational Research, 95(1), 24–37.

ElBannan M., 2017, “The Financial Crisis, Basel Accords and Bank Regulations: An Overview”, International Journal of Accounting and Financial Reporting, Vol. 7, No. 2.

Fernandez Vidal M. & Barbon F. 2019, “Credit Scoring in Financial Inclusion. Technical Guide”, Washington, D.C.: CGAP.

Finlay SM 2009, “Are we modelling the right thing? The impact of incorrect problem specification in credit scoring”, Expert Systems with Applications 36(5): 9065-907

Good, I. J. 1950 , “Probability and the Weighing of Evidence”, Griffin,London and Hafner Publ., New York.

Good, I.J. & Osteyee, D.B. 2008, “Information, Weight of Evidence the Singularity Between Probability Measures and Signal Detection”, Springer, New York.

Good, I.J. 1984, "C197. The best explicatum for weight of evidence", Journal of statistical computation and simulation, vol. 19, no. 4, pp. 294-299.

Good, I.J. 1983, “Good thinking: the foundations of probability and its applications”, N - New edn, University of Minnesota Press, Minneapolis.

Guo, H., Zhou, J. & Wu, C. 2018, "Imbalanced Learning Based on Data-Partition and SMOTE", Information (Basel), vol. 9, no. 9, pp. 238.

James, G., Witten, D., Hastie, T., Tibshirani, R. 2017, “An Introduction to Statistical Learning: with Applications in R”, Springer New York, New York, NY.

Lin, T-Y & Lin, A. 2014, “Expanding the Use of Weight of Evidence and Information Value to Continuous Dependent Variables for Variable Reduction and Scorecard Development”, https://www.lexjansen.com/sesug/2014/SD-20.pdf (2021-09-19)

Lingo, M & Winkler, G 2008, “Discriminatory power: an obsolete validation criterion?” Journal of Risk Model Validation, 2(1), 45–72.

Marqués, A.I., García, V. & Sánchez, J.S. 2013, "A literature review on the application of evolutionary computing to credit scoring", The Journal of the Operational Research Society, vol. 64, no. 9, pp. 1384-1399.

Mavri, M., Angelis, V., Ioannou, G., Gaki, E. & Koufodontis, I. 2008, "A two-stage dynamic credit scoring model, based on customers' profile and time horizon", Journal of Financial Services Marketing, vol. 13, no. 1, pp. 17-27.

Ong, C., Huang, J. & Tzeng, G. 2005, "Building credit scoring models using genetic programming", Expert Systems With Applications, vol. 29, no. 1, pp. 41-47.

Sanz, H., Valim, C., Vegas, E., Oller, J.M. & Reverter, F. 2018, "SVM-RFE: selection and visualization of the most relevant features through non-linear kernels", BMC bioinformatics, vol. 19, no. 1, pp. 432-432.

Siddiqi, N 2006), “Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring”, John Wiley & Sons, New York, NY, USA,

Tam, K. Y., & Kiang, M. Y. 1992, “Managerial applications of the neuralnetworks: the case of bank failure prediction”, Management Science,38(7), 926–947.

Thomas, L.C. 2000, "A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers", International journal of forecasting, vol. 16, no. 2, pp. 149-172.

Van Gool, J., Verbeke, W., Sercu, P. & Baesens, B. 2012, "Credit scoring for microfinance: is it worth it?", International Journal of Finance & Economics, vol. 17, no. 2, pp. 103-123.

Wang, C. & Huang, Y. 2009, "Evolutionary-based feature selection approaches with new criteria for data mining: A case study of credit approval data", Expert Systems With Applications, vol. 36, no. 3, pp. 5900-5908.

Xue, W. & Zhang, J. 2016, "Dealing with Imbalanced Dataset: A Re-sampling Method Based on the Improved SMOTE Algorithm", Communications in statistics. Simulation and computation, vol. 45, no. 4, pp. 1160-1172.

Yobas, M.B., Crook, J.N., Ross, P., 2000. Credit scoring usingneural and evolutionary techniques. IMA Journal of Mathematics Applied in Business and Industry 11, 111–125.

Zurada, J., Kunene, N. & Guan, J. 2014, "The classification performance of multiple methods and datasets: cases from the loan credit scoring domain", Journal of International Technology and Information Management, vol. 23, no. 1, pp. 57.

Zeng , G 2014, “A necessary condition for a good binning algorithm in credit scoring,” Applied Mathematical Sciences, vol. 8 , no. 65, pp. 3229–3242