Ranking Classification Algorithms Using MCDM Techniques for Credit Card Fraud Detection with Imbalanced Perspective

Dr. Maira Anis
Assistant Professor, NUST Business School, National University of Sciences and Technology
maira.anis@nbs.nust.edu.pk

Dr. Mohsin Ali
Director PDC, PNEC, National University of Sciences and Technology

Dr. Asma Gul
Associate Professor, NUST Business School, National University of Sciences and Technology

&

Dr. Sana Aroos Khattak
Senior Assistant Professor, Bahria Business School, Bahria University Karachi Campus

Abstract

Over the last few years, credit card fraud (CCF) has emerged as a serious concern in financial risk detection. Classification algorithms are among the robust methods to detect, analyze, and predict such frauds with low complexity, pre-detection, and loss estimation. However, class imbalance—where the number of legitimate transactions far exceeds the number of fraudulent ones—adversely affects a classifier's performance, and the results become divergent for different performance measures due to their multifaceted outcomes. This poses a key challenge in ranking such algorithms. Moreover, fraud data is inherently imbalanced with non-static behavior, so a single classifier or a group of classifiers cannot provide satisfactory results for all imbalance ratios. Motivated by this observation, we analyze the impact of class imbalance through 10 classifiers and 8 performance measures, which are finally ranked by Multi Criteria Decision Making (MCDM) techniques for a unified approach. Results are analyzed for low skewed datasets, where classification algorithms outperformed for low skewed CCF datasets. Based on our findings, we present a unified approach for ranking the classification algorithms in relation to different imbalance ratios in credit card datasets. This study can be very useful for financial institutions to increase their fraud-catching rate.

Keywords: classification algorithms; MCDM; class imbalance; credit card fraud.

Introduction

Over the past few decades, credit card fraud has emerged as a significant challenge for financial institutions (Maryeme, Hatim, & Mahmoud, 2019; Singh, et al., 2024). The rapid expansion of e-commerce has further heightened the risk of fraud (Cherif, et al., 2023). To mitigate the financial losses associated with fraud, there is a pressing need to adopt advanced techniques from knowledge discovery, machine learning, and pattern recognition (Anis, Ali, Mirza, & Munir, 2020; Maira & Mohsin., 2017; Maira, Mohsin, & Amit, 2015). One promising approach is the use of classification algorithms, which can streamline the fraud detection process.

Classification algorithms are known for their efficiency in detecting fraud by employing various models such as mathematical programming (Chen & Shih., 2006; Frydman, Altma, & Kao, 1985; Peng et.al., 2008; Tseng, Liu, & Ho, 2008), non-parametric statistical analysis (Opricovic & Tzeng, 2004), artificial intelligence (Altman, Avery, & Eisenbeis, 1981; Atiya, 2001; Carter & Catlett, 1987; Desai, Convay, & Crook, 1997; Leonard, 1993; Varetto, 1998), and traditional statistical models (Opricovic & Tzeng, 2004). These algorithms are often considered superior to traditional fraud detection methods due to their higher predictive accuracy and reduced susceptibility to human error (Atiya, 2001).

However, a critical challenge in using classification algorithms for credit card fraud detection is the issue of class imbalance, where legitimate transactions vastly outnumber fraudulent ones (Anis, Ali, Mirza, & Munir, 2020; Maira & Mohsin., 2017; Maira, Mohsin, & Amit, 2015). This imbalance can significantly impair the performance of classifiers, leading to inconsistent results across different performance measures (Feng, Zhou, & Tong, 2021;

Zhu & Zhang, 2024). Consequently, ranking these algorithms becomes a complex task due to the varied outcomes produced by different evaluation metrics.

While numerous studies have focused on the application of classification algorithms in fraud detection, there is limited research on systematically evaluating and ranking these algorithms in the context of class imbalance. Most existing studies do not address how different imbalance ratios affect the performance of various classifiers, nor do they provide a comprehensive method for comparing these algorithms across multiple performance measures.

Addressing this research gap, our study investigates the impact of class imbalance on the performance of 10 classification algorithms using 8 performance evaluation measures. To achieve a robust comparison, we employ Multi-Criteria Decision Making (MCDM) techniques.

MCDM techniques are a set of methods used to evaluate and prioritize multiple competing criteria in decision-making processes. In our study, we use three MCDM techniques to rank the classifiers:

Technique for Order Preference by Similarity to Ideal Solution (TOPSIS): This method ranks alternatives based on their distance from an ideal solution.

Analytic Hierarchy Process (AHP): AHP decomposes a complex decision-making problem into a hierarchy of simpler sub-problems, each of which can be analyzed independently.

VIKOR: This method focuses on ranking and selecting from a set of alternatives, and determines compromise solutions by considering the closeness to the ideal solution.

Our study aims to provide a unified approach for ranking classification algorithms across various imbalance ratios in credit card datasets, ultimately aiding financial institutions in enhancing their fraud detection capabilities.

Literature Review

Literature provides a series studies for the ranking of classification algorithms in various domains. The approach proposed in Kou, Lu, Peng, and Shi (2012) resolved disagreements among MCDM methods based on Spearman's rank correlation coefficient. The conflicting MCDM rankings reached an agreement for the techniques employed for the study. Evaluation of clustering algorithms is fundamentally a difficult task. However, Kou., Peng., and Wang (2014) presented an MCDM approach for the evaluation of clustering algorithms used for financial risk analysis. The results show the effectiveness of MCDM methods and highlight the repeated-bisection method as a strong performer for 2-way clustering solutions in financial risk datasets.

Feature selection is crucial for text classification as it reduces dimensionality, thereby enhancing classification performance and efficiency. Additionally, it improves model interpretability by identifying the most relevant features, which aids in better generalization on small sample data. Kou et al (2020) evaluates feature selection methods for text classification with small sample datasets involves multiple criteria, making it an MCDM problem. The results of study show the effectiveness of MCDM-based methods in evaluating feature selection methods and provide recommendations based on ranked results.

Various classifiers have been proposed for financial risk prediction, but using a single performance metric is inadequate for evaluating imbalanced classifications (Batool, Awais, Rehman, Shafiq, & Dar, 2019), (Yaqub, Rehman, Awais, & Shafiq, 2018). The study Song and Peng (2019) presented a multi-criteria decision-making (MCDM) approach to evaluate imbalanced classifiers in credit and bankruptcy risk prediction by considering multiple performance metrics simultaneously. An experimental study using the TOPSIS method over seven financial datasets from the UCI Machine Learning Repository indicates that SMOTE-based ensemble techniques outperform other imbalanced learning methods. SMOTEBoost-C4.5, SMOTE-C4.5, and SMOTE-MLP were ranked as the top classifiers based on their performance across six criteria.

Classification algorithm performance for learned predictive model is evaluated on the training dataset for the unseen observations (test dataset). The past studies in (Anis, Ali, Mirza, & Munir, 2020; Maira & Mohsin., 2017; Maira, Mohsin, & Amit, 2015) indicate that the classification results optimal with one performance measure may not be best with the use of different performance measures. Also, there are some other factors that contribute to the divergent results obtained by these performance measures such as, class distribution, noise, dataset characteristics etc. Thus, it is not feasible to use a single performance measure. Following are some of the basic evaluation measures used for classification algorithms (Kou et al., 2020; Kou, Lu, Peng & Shi, 2012; Kou., Peng, & Wang., 2014; Song & Peng, 2019).

Overall Accuracy (OA)

Accuracy defines the overall effectiveness of the classifier by giving the %age of correctly classified instances. It is one of the most commonly used classification performance metrics and is given by where TN, TP, FN, and FP are acronyms for True Negative, True Positive, False Negative and False Positive, respectively. Here negative represents good or legitimate instances whereas positive refers to bad or fraudulent instances.

$$\mathrm{OA} = \frac{TN + TP}{TN + FN + TP + FP}$$

TP represents the bad transactions predicted as bad.
TN represents the good transactions predicted as good.
FN represents the bad transactions predicted as good.
FP represents the good transactions predicted as bad.

True Positive Rate

A True Positive Rate (TPR) is the number of correctly classified positive instances or minority instances. In this case positive instances are the fraudulent instances. TPR is also called sensitivity measure given as

$$\mathrm{TPR} = \frac{TP}{TP + FN}$$

True Negative Rate

A True negative rate (TNR) defines the number of correctly classified negative or majority instances. In this case the negative instances belong to the non-fraudulent class. TNR is also called specificity and is defined as

$$\mathrm{TNR} = \frac{TN}{TN + FP}$$

Precision

Precision (P) is the number of classified minority or positive instances that actually are positive instances. In this case precision refer to the fraudulent instances that are actually fraudulent, and given as

$$\mathrm{P} = \frac{TP}{TP + FP}$$

Area under ROC Curve

ROC refers to Receiver Operating Characteristic that represents a tradeoff between the TP and FP. ROC analysis offers more robust evaluation for the predictive performance, of the relative class, of alternative models than traditional comparison of relative error. Key assumption of ROC analysis is that the tradeoff between TP and FP tells the predictive performance of the classifiers independent of their class distributions. Area under ROC curve calculates the accuracy of classifiers. Larger area represents better classifiers with high TP rate.

F-measure

F-measure is defined as the weighted harmonic mean of the precision and recall. This metric was firstly used in information retrieval and is defined as

$$\mathrm{F-measure} = \frac{2(Precision * Recall)}{Precision + Recall}$$

Kappa-Statistics

Kappa statistic measures the similarity or agreement between the qualitative variables. This agreement between variables is defined as

$$\text{Kappa Statistic} = \frac{P(A) + P(E)}{1 - P(E)}$$

Mean Absolute Error

It measures how much predictions are deviated from the true probability. MAE is defined as

$$\text{MAE} = \frac{\sum_{j=1}^{c} \sum_{i=1}^{m} |f(i,j) - P(i,j)|}{m, c}$$

where (i, j) represents the estimated probability of i module to be of class j that take values in [0, 1].

As the classification algorithms are assessed over different performance measures e.g., AUC, F-measure, TPR etc, and the selection of an optimal classifier process involves several criteria (performance measures) so this can be modeled as a MCDM problem.

Multi criteria decision making (MCDM) is a process of selecting optimal alternative under a given set of criteria and their sub criteria if there are any. Over last decade, some remarkable studies have been presented to rank classification algorithms for different real-world problems (Kim et al., 2003; Kou, Lu, Peng, & Shi, 2012; Kou et al., 2005; Peng & Wang, 2010; Visa & Ralescu, 2005). These studies have also ranked classification algorithms for credit risk analysis. However, none of them consider the class imbalance in ranking classification algorithms. This study adopts the analytical hierarchy model (AHM) proposed by Kou and Wenshuai (Peng & Wang, 2010). For this study 10 classifiers will be ranked for the different imbalance ratios of credit card data sets based on their performance measures.

Methodology

This study utilizes two datasets: the Australian Credit Approval (ACA) and the German Credit Dataset (GCD), each with distinct imbalance ratios. The primary objective is to evaluate and select classification algorithms for credit risk analysis across various class distribution levels.

Datasets Selection and Preparation

Australian Credit Approval (ACA)*:* Originally, ACA exhibits a class distribution split of 44.5% for positive cases (approved credit) and 55.5% for negative cases . To align with the study's objectives and to ensure comparability with the GCD dataset, ACA was adjusted through random undersampling of negative cases. This adjustment was made to create balanced splits of 70/30, 75/25, and 80/20 (positive/negative class ratios). Ultimately, this resulted in a dataset where only 166 observations were used for the 70/30 split.

German Credit Dataset (GCD)*:* The original GCD dataset has an imbalance ratio of 70/30 (positive/negative class ratios), making it suitable for direct comparison with the adjusted ACA dataset. No further modifications were required for the GCD dataset.

Implementation of Classification Algorithms

Ten widely recognized classification algorithms in credit risk analysis were implemented using WEKA 3.7.9. The classification algorithms used for rhis study are: CART (Classification and Regression Trees), Bayes network (BNK), Naïve Bayes (NBS), Linear logistic (LL), J48 (C4.5 algorithm), IBK (Instance-Based Learner), SVM (Support Vector Machines), RBF network (Radial Basis Function Network), Voted perceptron (VP).

Experimental Setup

Each dataset variant (70/30, 75/25, and 80/20 class distributions) underwent further division into training and testing sets. For robust evaluation, 10-fold cross-validation was applied across all experiments.These distributions have been taken in to account in accordance with the past studies for the credit card fraud detection (Maira, Mohsin, & Amit, 2015; Maira. & Mohsin., 2017; Anis M. , Ali, Mirza, & Munir, 2020; Gaudreault & Branco, 2024)

Performance Evaluation

The performance of classification algorithms was assessed using a suite of performance/evaluation measures appropriate for imbalanced datasets, as detailed in Section 2. These measures were selected to provide a comprehensive evaluation of each algorithm's effectiveness in fraud detection scenarios.

Primary Ranking Phase

To rank the classification algorithms based on their performance across different class distributions, three MCDM techniques were employed: VIKOR, PROMETHEE II, and Technique for Order Preference by Similarity to Ideal Solution (TOPSIS). These methods were chosen to ensure a robust and objective comparison of algorithmic performance (Baesens et.al, 2003).

Unified Ranking

Finally, Multi-Criteria Decision Analysis (MCDA) was applied to consolidate the rankings generated by the MCDM techniques. This process resulted in a unified ranking of classification algorithms for each class distribution scenario across the ACA and GCD datasets.

Rankings produced after MCDM techniques are taken as an inputfor unified ranking phase using TOPSIS that gives a consolidated ranking for credit scoring datasets. As MCDM does not come under the umbrella of class imbalance learning and is used only as a technique for ranking, therefore for brief introduction of MCDM techniques can be found in (Brans, 1982; Hwang & Yoon, 1981).

Table 1: Summary of the datasets with their respective class ratios

Datasets	Altered class ratios					
	70/30		75/25		80/20	
	Maj	Min	Maj	Min	Maj	Min
GCD	487	212	420	140	386	94
ACA	246	106	257	86	246	62
	* Majority = Maj; Minority = Min					

Results and Discussion

As explained before that this study is designed in three phases. For the first phase i.e. Data Mining phase, classification results for both datasets ACA and GCD are presented in Table 2 and Table 3. To understand the impact of class imbalance thoroughly on credit datasets, 3 imbalance levels are created to measure the impact of degree of skewness on the performance of classification algorithms. The degree of skewness corresponds to the ratios 70/30, 75/25 and 80/20.

Table 2: Performance values of classifiers at different class imbalance level for ACA

Class Distribution	Classifier	OA	TP	TN	Precision	Kappa statistic	ROC	F-measure	MAE
70/30	Bayes Net	0.8606	0.767	**0.9143**	0.836	0.6933	0.905	0.8000	0.1712
	Cart	0.8303	0.733	0.8857	0.786	0.6280	0.878	0.7590	0.213
	IBK	0.8182	0.717	0.8762	0.768	0.6014	0.796	0.7410	0.1836
	J48	0.8364	0.817	0.8476	0.754	0.6526	0.823	0.7840	0.1979
	LL	0.8424	0.767	0.8857	0.793	0.6571	0.904	0.78	0.2467
	NB	0.8424	0.633	0.9619	0.905	0.6361	0.913	0.745	0.1709
	RBF	0.8424	0.667	0.9429	0.87	0.6416	0.895	0.755	0.2326
	RF	0.8303	0.683	0.9143	0.82	0.6198	0.868	0.745	0.2291
	SVM	0.8606	0.833	0.8762	0.794	0.702	0.855	0.813	0.1394
	VP	0.8182	0.85	0.8	0.708	0.6233	0.864	0.773	0.1825
75/25	Bayes Net	0.86	0.767	0.914	0.836	0.693	0.902	0.8	0.1709
	Cart	0.85	0.833	0.867	0.781	0.69	0.885	0.806	0.2075
	IBK	0.836	0.7	0.914	0.824	0.635	0.807	0.757	0.1656
	J48	0.861	0.833	0.876	0.794	0.702	0.885	0.813	0.1793
	LL	0.836	0.717	0.905	0.811	0.637	0.925	0.761	0.2104
	NB	0.83	0.617	0.952	0.881	0.608	0.91	0.725	0.179

	RBF	0.842	0.65	0.952	0.886	0.639	0.916	0.75	0.236
	RF	0.836	0.7	0.914	0.824	0.635	0.884	0.757	0.2309
	SVM	0.818	0.683	0.895	0.788	0.596	0.789	0.732	0.1818
	VP	0.818	0.717	0.876	0.768	0.601	0.873	0.741	0.1818
80/20	Bayes Net	0.8545	0.733	0.9238	0.846	0.6765	0.921	0.786	0.154
	Cart	0.8182	0.683	0.8952	0.788	0.5956	0.877	0.732	0.2212
	IBK	0.8061	0.633	0.9048	0.792	0.5622	0.769	0.704	0.1959
	J48	0.8424	0.767	0.8857	0.793	0.6571	0.886	0.78	0.1994
	LL	0.8424	0.75	0.8952	0.804	0.6546	0.925	0.776	0.2117
	NB	0.8182	0.583	0.9524	0.875	0.5769	0.912	0.7	0.1877
	RBF	0.8303	0.617	0.9524	0.881	0.6081	0.876	0.725	0.2391
	RF	0.8061	0.583	0.9333	0.833	0.5522	0.849	0.686	0.2352
	SVM	0.8182	0.683	0.8952	0.788	0.5956	0.789	0.732	0.1818
	VP	0.8182	0.717	0.8762	0.768	0.6014	0.888	0.741	0.1773

Table 3: Performance values of classifiers at different class imbalance for GCD

Class Distribution	Classifier	OA	TP	TN	Precision	Kappa statistic	ROC	F-measure	MAE
70/30	Bayes Net	0.73	0.284	0.9151	0.581	0.2342	0.749	0.382	0.3366
	Cart	0.7433	0.489	0.8491	0.573	0.3529	0.738	0.528	0.3486
	IBK	0.6667	0.477	0.7453	0.438	0.2168	0.611	0.457	0.3338
	J48	0.68	0.375	0.8066	0.446	0.1905	0.554	0.407	0.3554
	Linear logistic	0.7867	0.489	0.9104	0.694	0.4368	0.818	0.573	0.2989
	NB	0.7667	0.511	0.8726	0.625	0.4056	0.793	0.563	0.2832
	RBF	0.7233	0.409	0.8538	0.537	0.2826	0.744	0.465	0.3489
	RF	0.74	0.341	0.9057	0.6	0.2822	0.735	0.435	0.333
	SVM	0.7367	0.125	0.9906	0.846	0.1539	0.558	0.218	0.2633
	VP	0.7433	0.261	0.9434	0.657	0.2485	0.798	0.374	0.2567
75/25	Bayes Net	0.723	0.273	0.91	0.558	0.215	0.708	0.366	0.3414
	Cart	0.717	0.341	0.873	0.526	0.238	0.7	0.414	0.343
	IBK	0.68	0.386	0.802	0.447	0.196	0.594	0.415	0.3206
	J48	0.713	0.33	0.873	0.518	0.226	0.676	0.403	0.3118
	Linear logistic	0.773	0.375	0.939	0.717	0.365	0.81	0.493	0.2899
	NB	0.76	0.42	0.901	0.638	0.357	0.795	0.507	0.2768
	RBF	0.733	0.341	0.896	0.577	0.269	0.744	0.429	0.322
	RF	0.71	0.239	0.905	0.512	0.171	0.733	0.326	0.3257
	SVM	0.767	0.33	0.948	0.725	0.33	0.639	0.453	0.2333
	VP	0.74	0.205	0.92	0.729	0.21	0.796	0.316	0.26
80/20	Bayes Net	0.7133	0.102	0.967	0.563	0.091	0.732	0.173	0.3224
	Cart	0.71	0.102	0.9623	0.529	0.0845	0.698	0.171	0.3364
	IBK	0.69	0.307	0.8491	0.458	0.1725	0.578	0.367	0.3108
	J48	0.7167	0.318	0.8821	0.528	0.2266	0.602	0.397	0.3016
	Linear logistic	0.7067	0	1	0	0	0.5	0.828	0.5
	NB	0.7367	0.33	0.9057	0.592	0.2702	0.777	0.423	0.2844
	RBF	0.7133	0.159	0.9434	0.538	0.1291	0.713	0.246	0.3202
	RF	0.7333	0.136	0.9811	0.75	0.1545	0.736	0.231	0.3077
	SVM	0.7067	0	1	0	0	0.5	0	0.2933
	VP	0.75	0.205	0.9764	0.783	0.2308	0.803	0.205	0.25

For the original split of GCD i.e. 70/30 LL, NB and VP are on top in terms of ROC while SVM shows promising results for the measures TN and Precision with low values of ROC and TP. LL and CART were the top classifiers in detecting the most fraudulent cases. In ACA dataset, for 70/30, SVM and BN gave highest score for OA and ROC whereas high TP rate was achieved by VP. J48, LL and BN were the top classifiers that were able to predict more fraudulent transactions.

For the distribution 75/25, LL performs well for most performance values in GCD whereas the highest TP value is given by NB. However, in ACA, It was observed that J48 performed well for most of the performance measures, Cart and VP were also the classifiers that performed above average.

In GCD, for the ratio 80/20, VP is best classifier for Precision, OA and ROC whereas LL and SVM are the lowest classifiers in terms of TP rate. However, the highest TP rate and F-measure is attained by NB. RF is the best classifier with high TN rate. However, in ACA, VP is best for TP. Highest ROC values are attained by LL and BN and NB

In general, it is analyzed that for every class distribution in low skewness, classifier achieving good scores on one evaluation measure can be the classifier performing poor for another measure. However, there were some classifiers that consistently performed poorly for the lower skewed distributions e.g., IBK, NB.

Although some similarities are there between the two credit card datasets, no clear conclusion can be drawn which classifier will give optimal results in each class ratio. For this purpose, MCDM techniques are utilized to present a unified approach towards selection of classification algorithms.

MCDM Phase

This phase is further divided to two phases to get ranking of classifiers.

Primary Ranking Phase

In this phase initial ranking of classifiers have been achieved using 3 MCDM techniques i.e. TOPSIS (Amin, Anwar, & Adnan, 2016), VIKOR (Hwang & Yoon, 1981) and PROMETHEE-II (West & Bhattacharya, 2015). Initial ranking of the 10 classifiers used in this study are illustrated in the Table 4.

From Table 4, with little degree of disagreement, MCDM methods show correspondence between the rankings. Weight assigned to the MCDM methods is in accordance with the previous studies (Peng. & Wang., 2010; Visa & Ralescu, 2005). As in credit card fraud, the predictive model with highest TP and ROC values is best so these two measures are given the weight 10 while other performance measures are set to 1. These weights are then normalized such that the sum of all weights is 1. Rankings attained for both datasets in each distribution is discussed below.

Table 4: Primary Ranking for Australian and German credit dataset.

Class Dist	Classifier	Australian Credit Data Set						German Credit Data Set					
		VIKOR		TOPSIS		Promethee		VIKOR		TOPSIS		Promethee	
		V	R	V	R	V	R	V	R	V	R	V	R
70/30	BN	0.1247	3	0.667	5	0.38	6	0.5785	7	0.962	2	-0.62	7
	Cart	0.2890	5	0.842	2	1.07	1	0.0997	2	0.902	10	1.952	1
	IBK	1.0000	10	0.556	8	-1.1	9	0.6264	9	0.926	6	-0.82	9
	J48	0.5728	8	0.659	7	-0.3	7	0.5998	8	0.943	5	-0.79	8
	LL	0.0000	1	0.760	4	0.59	4	0.1993	4	0.925	7	1.211	3
	NB	0.9656	9	0.199	10	-2.6	10	0.0000	1	0.911	9	1.496	2
	RBF	0.4173	7	0.532	9	-0.4	8	0.1214	3	0.919	8	0.879	4
	RF	0.0703	2	0.714	5	0.56	5	0.3846	5	0.945	4	-0.03	5
	SVM	0.2890	5	0.842	2	1.07	1	1.0000	10	0.995	1	-3.00	10
	VP	0.1323	4	0.853	1	0.81	3	0.4846	6	0.955	3	-0.25	6
75/25	BN	0.1347	3	0.715	3	0.77	3	0.6995	6	0.967	2	-0.72	8
	Cart	0.0000	1	0.866	1	1.18	2	0.4806	4	0.947	6	-0.15	4
	IBK	0.8848	9	0.340	8	-0.9	9	0.8147	8	0.934	8	-0.99	10
	J48	0.0279	2	0.864	2	1.22	1	0.6035	5	0.952	5	-0.5	7
	LL	0.2473	4	0.565	4	0.32	4	0.1482	2	0.930	9	1.738	2
	NB	0.8293	8	0.317	9	-0.7	8	0.0000	1	0.918	10	1.907	1
	RBF	0.6761	7	0.375	7	-0.3	7	0.3467	3	0.945	7	0.384	3

	RF	0.4689	6	0.455	6	-0.1	5	0.7725	7	0.971	1	-0.96	9
	SVM	1.0000	10	0.267	10	-1.3	10	0.8843	10	0.954	4	-0.34	5
	VP	0.3367	5	0.499	5	-0.2	6	0.8607	9	0.965	3	-0.36	6
80/20	BN	0.1145	3	0.744	6	0.64	5	0.4983	7	0.910	3	-0.18	7
	Cart	0.3290	6	0.772	5	0.47	6	0.1188	2	0.878	9	2.035	1
	IBK	1.0000	10	0.515	9	-1.86	9	0.1681	3	0.892	7	0.295	5
	J48	0.7717	8	0.654	7	-0.57	7	0.0082	1	0.872	10	1.794	2
	LL	0.0291	2	0.836	4	0.90	4	0.3305	5	0.895	6	1.04	4
	NB	0.9572	9	0.225	10	-3.04	10	0.9839	9	0.917	2	-2.83	9
	RBF	0.4054	7	0.563	8	-0.58	8	0.2185	4	0.891	8	1.102	3
	RF	0.0002	1	0.876	1	1.71	1	0.4728	6	0.908	5	0.064	6
	SVM	0.2481	5	0.857	3	1.15	3	1.0000	10	0.947	1	-2.95	10
	VP	0.2208	4	0.862	2	1.18	2	0.5482	8	0.910	4	-0.37	8

For ACA, in the 70/30 distribution, LL is 1st by VIKOR and is ranked 4th by PROMEETHE and TOPSIS. Similarly, VIKOR ranks LL 1st for GCD while 7th and 3rd by TOPSIS and PROMETHEE respectively. For ACA, VP also stands among the top classifiers.

For 75/25, LL ranked 2nd best for GCD while 4th for ACA by all MCDM techniques. J48 gave the best ranking scores of 1st and 2nd for ACA while performed above average for GCD. It was noticed that IBK attained the worst rankings for both datasets.

For the last distribution i.e., 80/20, LL and Cart stand out with best rankings whereas BN also achieved good ranks of 2nd and 3rd. For this distribution also, SVM has attained conflicting rankings for both datasets. For this distribution SVM is ranked either 8th or 7th by all MCDM techniques for ACA whereas for GCD it is ranked among the top three classifiers. However, VP is performing well in both datasets and is ranked 4th. Among all distribution of lower skewed datasets, it is observed that at first NB was the worst classifier but as the skewness is leveraged for both datasets, NB started to gain higher ranks in both datasets. However, this needs to be explored further for data sets with high skewness. Similarly, SVM was the top classifier but as the imbalance ratios among the datasets increase, SVM either performed good or gave average performance.

Unified Ranking Phase

From the primary ranking phase, results we got are tolerably different from each other in all distributions for ACA and GCD. As the goal of this study is towards finding a unified ranking, we will use the ranking scores found from three MCDM methods to the final ranking phase that uses TOPSIS. In the final ranking phase, all the MCDM techniques i.e. TOPSIS, VIKOR and PROMETHEE have been assigned equal weight of 0.33. Table 5 illustrates the final ranking of the 10 classifiers for each class ratio of ACC and GCC.

For the distribution 70/30, SVM and VP are ranked 1st and 2nd for ACA and GCD respectively whereas LL was ranked 4th and 3rd respectively. However, it is observed that for the other two distributions of low skewness LL attained top ranks being 1st. As the skewness is leveraging, J48 rank has also been improved.

Table 5: Final rank of classification algorithm in accordance with the class imbalance of dataset.

Low skewed distributions	**Classifiers**	**Aus**		**Germ**	
		V	**R**	**V**	**R**
70/30	BN	0.4140	5	0.4849	7
	Cart	0.2399	2	0.4299	4
	IBK	0.8608	9	0.7442	10
	J48	0.7009	7	0.6498	9
	LL	0.2558	4	0.4196	3
	NB	0.9378	10	0.4010	1
	RBF	0.7662	8	0.4557	6
	RF	0.3503	5	0.4083	2
	SVM	0.2399	2	0.5851	8
	VP	0.2176	1	0.4491	5
75/25	BN	0.2222	3	0.4856	5
	Cart	0.0625	1	0.4117	2

	IBK	0.8444	9	0.8257	10
	J48	0.0890	2	0.5180	7
	LL	0.3333	4	0.4177	4
	NB	0.8094	8	0.4149	3
	RBF	0.6667	7	0.3912	1
	RF	0.5182	6	0.5114	6
	SVM	1.0000	10	0.5675	9
	VP	0.4818	5	0.5418	8
80/20	BN	0.4149	5	0.5152	6
	Cart	0.5182	6	0.4008	1
	IBK	0.9107	9	0.4517	4
	J48	0.7008	7	0.4300	2
	LL	0.2726	3	0.4462	3
	NB	0.9379	10	0.5823	8
	RBF	0.7377	8	0.4553	5
	RF	0.0000	1	0.5182	7
	SVM	0.3058	4	0.5851	9
	VP	0.2051	2	0.6088	10

For 75/25, LL represents unified rank for both datasets. Whereas Cart and RF have started to excel in their ranks as the skewness is leveraged.

It is noteworthy to mention that rank of classifiers is shifting when skewness is leveraged i.e. the weak classifiers are progressively achieving good ranks and vice versa. For example, LL has shown good performance in low skewed datasets but this needs to be explored further whether LL will continue to show a similar performance as the class imbalance ratio is increased. Similarly, J48 is the average classifier in datasets with low skewness. However, further research must be conducted to assess the behavior of classifiers for different imbalance ratios of datasets.

Table 5 shows that final ranking demonstrate reduced discrepancies among classifiers, with LL consistently performing well in low-skewed datasets. However, as skewness increases, other classifiers like J48 show improved rankings, indicating variable performance under different imbalance ratios. Further research is needed to explore classifier behavior in datasets with higher skewness levels.

References

Alkhateeb, Z. K., & Maolood, A. T. (2019). Machine Learning-Based Detection of Credit Card Fraud: A Comparative Study. *American Journal of Engineering and Applied Sciences, 12*(4), 535-542.

Altman, E. I., Avery, R. B., & Eisenbeis, R. A. (1981). *Application of classification techniques in business.* Banking and Finance, JAI Press, Inc., CT.

Amin, A., Anwar, S., & Adnan, A. (2016). Comparing over-sampling techniques to handle the class imbalance problem: a customer churn prediction case study . *IEEE Access*.

Anis, M., Ali, M., Mirza, S. A., & Munir, M. M. (2020). Analysis of resampling techniques on predictive performance of credit card classification. *Modern Applied Science*, *14*(7), 92.

Atiya, A. F. (2001). Bankruptcy prediction for credit risk using neural networks: a survey and new results. *IEEE Transactions on Neural Networks*, *12*(4):929–935.

Azhan, M., & Meraj, S. (2020). Credit card fraud detection using machine learning and deep learning techniques. *In 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)* , (pp. 514-518).

Baesens, et al. (2003). Using neural network rule extraction and decision s for credit-risk evaluation. *Management Science*, *49,* 312–329.

Batool, I., Awais, M., Rehman, K. U., Shafiq, M., & Dar, I. B. (2019). How Social Economy can add Value to State Development?. *Foundation University Journal of Business & Economics, 4*(1), 1-12.

Brans, J. P. (1982). L'ingénièrie de la décision; Elaboration d'instruments d'aide à la décision. La méthode PROMETHEE, in: R. Nadeau, M. Landry (Eds.), L'aide à la décision: Nature. *Instruments et Perspectives d'Avenir, Presses de l'Université Laval, Québec, Canada*, 183-213.

Carter, C., & Catlett, J. (1987). Assessing credit card applications using machine learning. *IEEE Expert*, 71–79.

Chen, W., & Shih., J. (2006). A study of Taiwan's issuer credit rating systems using support vector machines. *Expert Systems with Applications*, *30*(3), 427–435.

Cherif, A., Badhib, A., Ammar, H., Alshehri, S., Kalkatawi, M., & Imine, A. (2023). Credit card fraud detection in the era of disruptive technologies: A systematic review. *Journal of King Saud University - Computer and Information Sciences*, *35*(1),145-174,.

Desai, V. S., Convay, D. G., & Crook, J. N. (1997). Credit scoring models in the credit union environment using neural networks and genetic algorithms. *IMA Journal of Mathematics Applied in Business and Industry*, *8*, 323–346.

Feng, Y., Zhou, M., & Tong, X. (2021). Imbalanced classification: A paradigm-based review. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *14*(5), 383-406.

Frydman, H., Altma, E. I., & Kao, D. (1985). Introducing recursive partitioning for financial classification: the case of financial distress . *The Journal of Finance*, *40*, 269–291.

Gaudreault, J. G., & Branco, P. (2024). A Systematic Literature Review of Novelty Detection in Data Streams: Challenges and Opportunities. *ACM Computing Surveys*.

Hwang, C. L., & Yoon, K. (1981). Multiple attribute decision making methods and applications. *Operations Research & Decision Theory*.

Kim et al. (2003). Constructing support vector machine ensemble. *Pattern Recognition*, *36*, 2757–2767.

Kou, Y. L. (2012). Evaluation of classification algorithms using MCDM and rank correlation . *International Journal of Information Technology & Decision Making*, *11*(1), 197.

Kou, G., Lu, Y., Peng, Y., & Shi, Y. (2012). Evaluation of classification algorithms using MCDM and rank correlation. *International Journal of Information Technology & Decision Making*, 11(01):197-225.

Kou, G., Peng, Y., Shi, Y., & et.al. (2005). Discovering credit card holders' behavior by multiple criteria linear programming . *Annals of Operations Research*, *135*(1), 261–274.

Kou, G., Yang, P., Peng, Y., Xiao, F., Chen, Y., & Alsaadi, F. E. (2020). Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods. *Applied Soft Computing*, *86*, 105836.

Kou., G., Peng., Y., & Wang., G. (2014). Evaluation of clustering algorithms for financial risk analysis using MCDM methods. *Information sciences*, *275*, 1-12.

Leonard, K. J. (1993). Detecting credit card fraud using expert systems. *Computers and Industrial Engineering*, *25*, 103–106.

Maira, A., Mohsin, A., & Amit, Y. (2015). A comparative study of decision tree algorithms for class imbalanced learning in credit card fraud detection. *International Journal of Economics, Commerce and Management*, *3*(12), 86-102.

Maira., A., & Mohsin., A. (2017). Investigating the performance of smote for class imbalanced learning: a case study of credit scoring datasets. *Eur. Sci. J*, *13*(33), 340-353.

Maryeme, T., Hatim, G., & Mahmoud, N. (2019). MCDM method for Financial Fraud Detection: A review. *The 4th International Conference On Big Data and Internet of Things*, (pp. 1-8).

Opricovic, S., & Tzeng, G. H. (2004). Compromise solution by MCDM methods: a comparative analysis of VIKOR and TOPSIS. *European Journal of Operational Research*, *156*(2), 445–455.

Peng, Y., Kou, G., Shi., Y., & et.al. (2008). A multi-criteria convex quadratic programming model for credit data analysis. *Decision Support Systems*, *44*, 1016–1030.

Peng., Y., & Wang., W. (2010). User preferences based software defect detection algorithms selection using MCDM. *Information Sciences*.

Singh, K., Kolar, P., Abraham, R., Seetharam, V., Nanduri, S., & Kumar, D. (2024). Automated Secure Computing for Fraud Detection in Financial Transactions. *Automated Secure Computing for Next-Generation Systems*, 177-189.

Song, Y., & Peng, Y. (2019). A MCDM-based evaluation approach for imbalanced classification methods in financial risk prediction. *IEEE Access*, *7*, 84897-84906.

Tseng, K. J., Liu, Y. H., & Ho, J. (2008). An efficient algorithm for solving a quadratic programming model with application in credit card holder's behavior . *International Journal Of Information Technology & Decision Making*, *7*, 421–430.

Varetto, F. (1998). Genetic algorithms applications in the analysis of insolvency risk. *Journal of Banking & Finance*, *22*, 1421–1439.

Visa, S., & Ralescu, A. (2005). Issues in mining imbalanced data sets-a review paper . *In Proceedings of the sixteen midwest artificial intelligence and cognitive science conference.*

West, J., & Bhattacharya, M. (2015). Intelligent financial fraud detection: a comprehensive review. *Computers & Security*.

Yaqub, A., Rehman, F., Awais, M., & Shafiq, M. (2018). The Impact of Financial Constraints, Dividend Policy and Capital Structure on Share Price Volatility. Foundation University *Journal of Business & Economics, 3*(2), 15-27.

Zhu, M., & Zhang, Y. G. (2024). Enhancing Credit Card Fraud Detection: A Neural Network and SMOTE Integrated Approach. *Journal of Theory and Practice of Engineering Science*, *4*(2), 23-30.