STATE-OF-THE-ART METHODS FOR CREDIT RISK PREDICTION WITH LOW DEFAULT PORTFOLIOS

Kate Xu
Massachusetts Institute of Technology
Cambridge, MA 02139
katexu@mit.edu

August 13, 2019

ABSTRACT

Financial credit risk assessment with Low Default Portfolios (LDPs) is a difficult and open problem due to highly imbalanced datasets that contain significantly more good credit examples than bad ones. Current research uses machine learning methods, but an ideal classification model for credit scoring has yet to be discovered. In this paper, state-of-the-art algorithms for imbalanced data are described, and the best models for further investigation are proposed. The most common modern approaches that can be applied to LDPs are data level methods, algorithm level methods like cost-sensitive learning and anomaly detection, and classifier or hybrid ensembles. Performance metrics that are more appropriate for models on imbalanced data are also discussed. The three proposed models include a bagging ensemble on cost-sensitive decision trees, a boosting hybrid algorithm with undersampling, and an extreme gradient boosting algorithm with Bayesian hyperparameter optimization.

1 Introduction

Credit risk prediction remains a significant challenge for financial institutions such as banks. A borrower goes into default if he or she cannot pay back a lender by the agreed time, and the lender loses the amount of money borrowed and the cost of disruptions to cash flows. Machine learning techniques are often used for credit scoring because they are more effective and reliable [41]. However, it is difficult for algorithms to estimate risk because lenders make careful decisions before distributing credit, so credit data are highly imbalanced [68]. The disparity between the number of good and bad credit examples leads to low default portfolios and causes insufficiently trained models [59]. This problem is an open question in machine learning, and there does not appear to be a singular method that consistently outperforms all others across datasets [10].

The rest of this paper is organized as follows. Section 2 describes the systematic search methodology and the paper bank used for research. Section 3 presents state-of-the-art algorithms for credit risk assessment and imbalanced data by category. Section 4 discusses the results and proposes the best approaches for low default portfolios. Section 5 summarizes the paper and provides the advantages and disadvantages of the proposed models. Section 6 explains the limitations of the research in this paper.

2 Methods

This section is divided into two subsections: the first explains the systematic search process, and the second describes recent works that have similar intentions to this paper but are written by other researchers.

2.1 Systematic search

A systematic search methodology was used to compile relevant research papers because such a procedure enables this paper to be the least biased and most reproducible by applying specific inclusion and exclusion criteria. The search

equation was created using keywords relating to the problem of credit risk prediction with highly imbalanced data, and the online citation and abstract database Scopus was used to conduct the actual search. As shown in Figure 1 below, keywords that describe techniques like "classification", words that describe the application like "credit risk", and those that describe the context like "state-of-the-art" were considered. Buzzwords like "machine learning" and word synonyms like "novel" for "state-of-the-art" were added to the search equation to broaden the scope of the search.

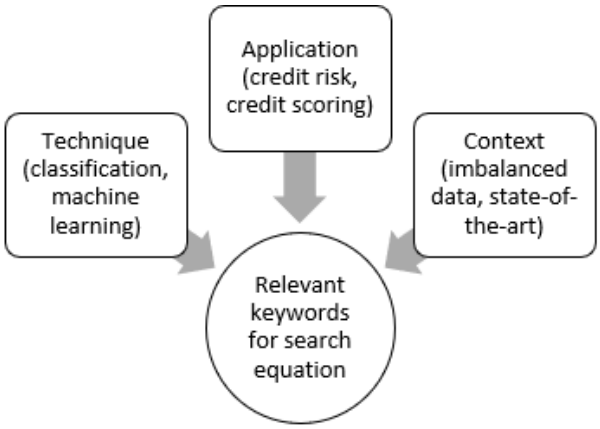

Figure 1. Visual representation of keywords selection to create the search equation

The systematic search was limited to research papers published since 2015 because most literature with a similar purpose as this paper were dated in that year. For example, Lessman et al. wrote the most recent comprehensive state-of-the-art paper on classification algorithms for credit scoring in 2015 [31]. Other papers on related topics were written in that same year or later, but there has not been a thorough study of newer classification methods since 2015. The resulting search equation used on Scopus is the following.

```
ALL(("credit risk" OR "credit scor*" OR "credit rating") AND ("imbalanced data*" OR
"unbalanced data*") AND classif* AND ("deep learning" OR "machine learning") AND ("state-
of-the-art" OR "state of the art" OR novel OR current OR modern OR recent)) AND PUBYEAR >
2014 AND (LIMIT-TO(LANGUAGE, "English"))
```

This equation returned 333 results at the time of the search on June 26, 2019. Figure 2 below illustrates an approximate breakdown of the initial paper bank grouped by category: credit risk or default prediction, fraud detection or e-commerce, anomaly detection, health or medicine, and other or not applicable. The Other/NA category consists of papers that focused on a broad research area or did not have enough papers to form a meaningful group such as the topics of retail and traffic accidents.

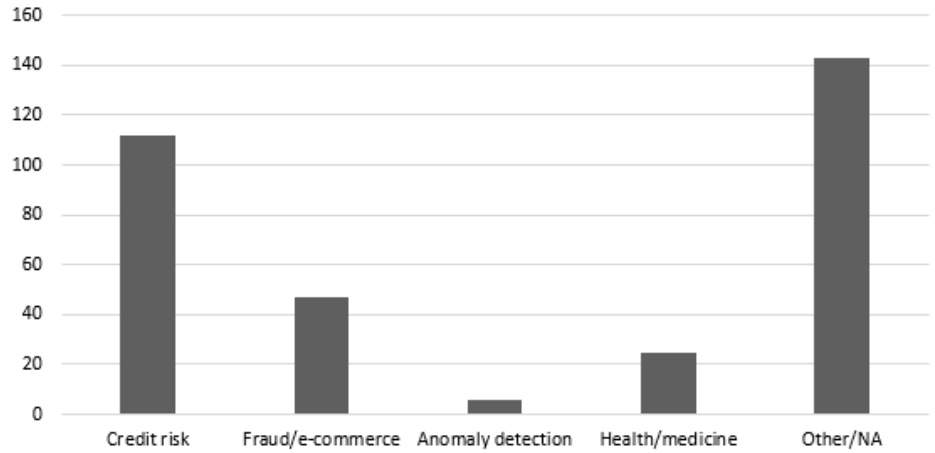

Figure 2. Bar graph depicting breakdown of initial paper bank by category

Many of the results in the initial paper bank appeared to be useful for research on credit risk and imbalanced data. 112 papers were related to credit risk, 47 papers were related to fraud detection or e-commerce, and 6 papers were related to

anomaly detection. These three categories alone accounted for nearly half of the paper bank. However, some papers could be irrelevant, so the 333 search results were exported to an Excel spreadsheet for the next filtering stage.

As shown in Figure 3 below, a three-step exclusion process was used to disregard irrelevant papers. First, papers that focus on topics unrelated to the research problem of credit risk and do not discuss class imbalance were discarded, which reduced the paper bank to 271 papers. Second, 71 more papers were discarded because they did not discuss imbalanced data even though they may have discussed credit risk. After applying these restrictions, the modified paper bank contained 200 papers. Third, some papers were inaccessible with the subscription, so each of the remaining papers was searched online for a full-length version using multiple websites, including ScienceDirect, ResearchGate, and a third source like ArXiv or IEEE. The complete systematic search method resulted in 96 potential papers for research.

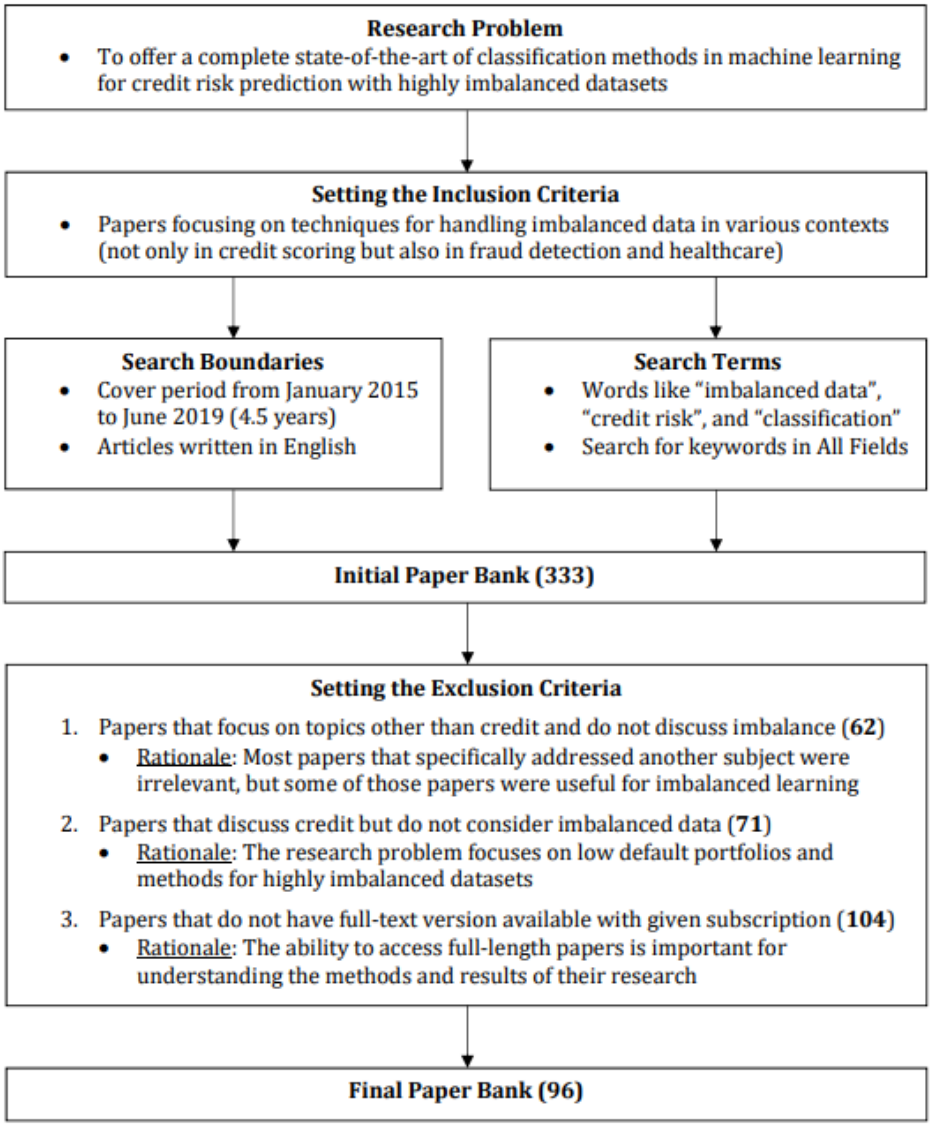

Figure 3. Complete systematic search methodology for compiling relevant papers

2.2 Recent related works

Various researchers have examined recent advances in classification methods for imbalanced data. [31] conducted a benchmarking experiment on state-of-the-art classifiers for retail credit scoring that extends the work of Baesens et al. (2003). Their literature review compares the datasets, classifiers, and evaluation metrics used in papers from 2003 to 2013. They analyzed 41 classification algorithms on seven real-world credit datasets, and they concluded that artificial neural networks can predict the probability of default more accurately than logistic regression, and that homogeneous ensembles outperform individual classifiers.

Many of the existing works published after [31] either consider mainly well-established classifiers with limited emphasis on novel ones or do not focus on applying those techniques to credit risk. [10] provided a review of financial credit risk assessment from its earliest research in 1932 to 2015. [23] studied class imbalance learning using papers from 2006 to 2016. [72] compared eight established classifiers and three newer classifiers on 71 datasets, and they found that stochastic gradient boosting trees, random forests, support vector machines, and extreme learning machines had the best classification performance. [75] described and analyzed five supervised, one unsupervised, and one semi-supervised classification methods for fraud detection. [29] studied papers from 2010 to 2018 that focus on big data and class imbalance, and they noted the similarities between methods for addressing imbalance in traditional and big data. [30] studied machine learning in bank risk management from papers after 2007, and their research explores credit risk, market risk, operational risk, and liquidity risk.

Paper	Years	Focus	Data	Classifiers	Metrics
[11]	1932-2015	Review of financial credit risk assessment since its first appearance in literature	N/A	N/A	N/A
[23]	2006-2016	Recent methods and applications for imbalanced learning and rare event detection	N/A	N/A	N/A
[29]	2010-2018	Techniques for high class imbalance on big data	N/A	N/A	N/A
[30]	1992-2018	Machine learning methods for various types of banking risk management	N/A	N/A	N/A
[31]	2003-2013	Update of Baesens et al. (2003) and classification methods for credit scoring	8 credit scoring datasets	41 algorithms	PCC, AUC, PG, H-measure, BS, KS
[72]	1900-2017	Empirical study of classification algorithms	71 UCI and Keel datasets	11 algorithms including ELM and DL	Accuracy, AUC, running time
[75]	1999-2018	Supervised, unsupervised, and semi-supervised learning for credit scoring and fraud detection	N/A	N/A	N/A

Table 1. Related works by focus and experimental data, if available as part of their research

Table 1 above presents seven related works published in recent years. The rows with N/A represent papers that describe techniques without giving specific examples or comparing classifiers on datasets. The table also emphasizes the need to create a state-of-the-art paper that includes the latest developments in machine learning as discussed in section 2.1. [72] confirms this suggestion as their paper's related works section mentions [31] and three other papers that were published in 2015. As a result, the objective of this paper is to provide an update on the current classification methods for credit risk assessment and class imbalance focusing on 2015 and later.

3 Results

The common strategies for imbalanced data are data-level and algorithm-level methods as seen in Figure 4 below [23, 29, 42]. At the data level, sampling techniques and feature selection and extraction are applied to balance and preprocess the data before training models. Algorithm-level methods modify existing algorithms to adapt them to imbalanced distributions and alleviate the bias towards the majority class. They include cost-sensitive methods, which account for misclassification costs and enhance classification performance by combining several diverse classifiers [10]. Researchers have also used hybrid and ensemble methods that combine data and algorithm level techniques [27, 28, 39].

The following subsections describe these general approaches in more detail. Sections 3.1 and 3.2 discuss data sampling and feature selection and extraction, which are common at the data level. Sections 3.3 and 3.4 explain some algorithm-level methods such as cost-sensitive learning and anomaly detection. Sections 3.5 and 3.6 discuss more algorithms like neural networks, support vector machines, and rule-based classifiers. Sections 3.7 and 3.8 describe ensemble and hybrid methods that can use both data-level and algorithm-level approaches. Section 3.9 explains the performance metrics used to evaluate models for imbalanced data.

These nine subsections do not necessarily represent nine different ways to tackle imbalanced datasets, but rather they split the state-of-the-art methods from the research papers into categories based on their likely appearance in a general credit risk model pipeline. Researchers often first preprocess their data with data-level methods, and then use algorithm-level techniques like individual and hybrid classifiers for training and testing. Lastly, metrics are used to evaluate and improve models [10]. Three models are suggested for future research based on the results of this paper, and an analysis of the proposed models is presented in Section 4.

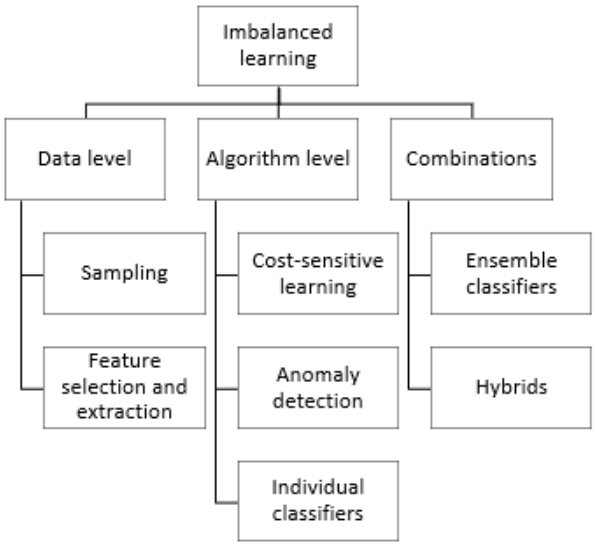

Figure 4. Overview of common imbalanced learning strategies

3.1 Data sampling

Data sampling adds or removes class instances to balance the training data. Sample size is important because some classification algorithms like logistic regression and neural networks show decreasing accuracy on smaller datasets [7]. Adaptive Synthetic (ADASYN) sampling and the Synthetic Minority Oversampling Technique (SMOTE) can be used to generate synthetic data for the minority class. ADASYN reduces bias from class imbalance and uses a density distribution to determine the number of additional examples needed [20]. In particular, it generates examples on the safe and border area based on the data distribution ratio [19]. Meanwhile, SMOTE attempts to balance the dataset by creating artificial minority examples between the real instances rather than oversampling with replacement [44], and it finds the nearest neighbors of data points as a basis for creating synthetic data [19]. As shown in Figure 5 below, a key difference between ADASYN and SMOTE is that the former will focus on the samples that are difficult to classify using the k-nearest neighbors algorithm, while SMOTE does not make any distinction [20]. There is a noticeable difference in the density of the new data points on the right side of the two distributions that illustrates this benefit of ADASYN.

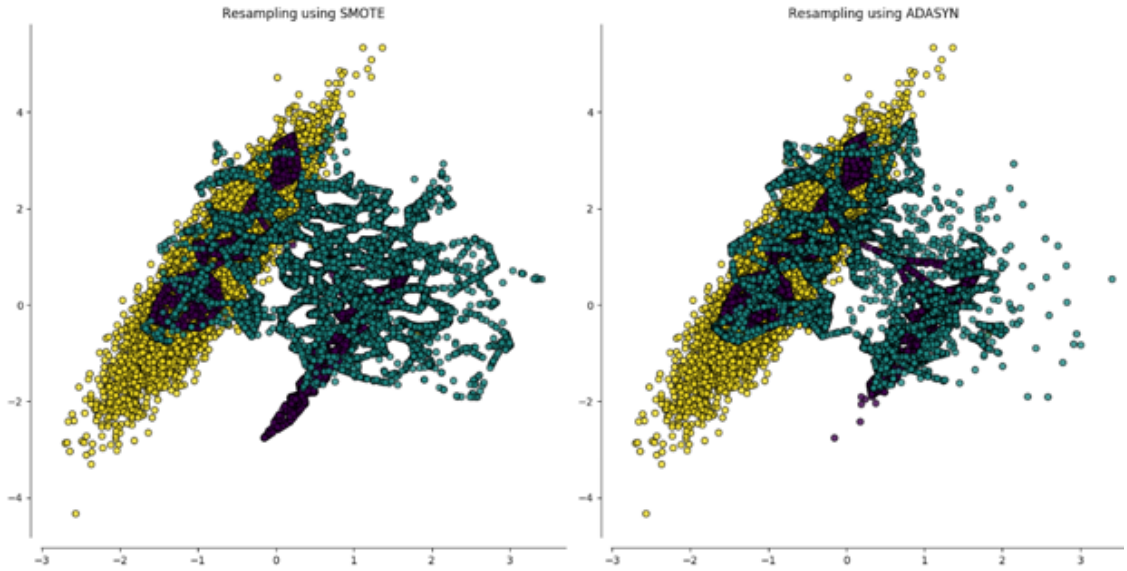

Figure 5. Difference between SMOTE (left) and ADASYN (right). Source: imbalanced-learn.readthedocs.io

Researchers have enhanced the performance of SMOTE for class imbalance [19, 20, 38]. [53] proposed the Enhanced Minority Oversampling Technique (EMOTE), which populates the dataset with minority class misclassified instances,

along with their minority class nearest neighbor from correctly classified instances. For highly dimensional datasets, an alternative SMOTE approach called SMOTE-Subset of Features (SMOTE-SF) that uses a novel distance metric based on relevant features without feature selection is proposed [38]. Researchers have also considered how highly correlated features negatively affect SMOTE when it creates synthetic data. [19] proposed an Attribute Weighted and kNN Hub on SMOTE (AWH-SMOTE) using attribute weighting and occurrence data in the kNN minority class (kNN hub). They found that AWH-SMOTE performs better than other synthetic oversampling methods and that the weighting attribute in kNN provides more representative neighbors. A novel approach to apply variational interference to oversample the minority class has also been proposed by [20]. They used Variational Autoencoders (VAE) for their Variational Oversampling (VOS) technique, which has performed better than SMOTE and ADASYN for fraud detection and tumor images. Adaptive Swarm Balancing Algorithms that combine SMOTE with swarm balancing algorithms also work well on large moderately and highly imbalanced datasets [32].

In addition to SMOTE variants, evolutionary sampling approaches such as Multiobjective Genetic Sampling (MOGASamp) have been proposed for imbalanced data. MOGASamp generates sets of samples for ensembles so that base classifiers have high accuracy and diversity [21]. Another resampling method in literature is undersampling such as random undersampling, which uniformly takes a subset of the majority class. Although oversampling and undersampling techniques generally work well, they can have disadvantages such as excluding important data points. As a result, research has explored hybrid techniques such as SMOTE and Tomek links (SMOTE-TOMEK) and SMOTE and edited nearest neighbor (SMOTE-ENN), which alleviate the individual disadvantages of oversampling and undersampling [41]. A novel hybrid sampling method is SMOTE Oversampling and Borderline Undersampling (SOS-BUS), which introduces synthetic minority class examples and removes some majority class examples but retains all instances near the decision boundary [52]. Another integrated approach combines SMOTE for oversampling and Particle Swarm Optimization (PSO) for undersampling. This hybrid has been experimented for malicious web domain identification models, and the results show that the proposed approach performed better than most of the six other resampling strategies used over four datasets [25].

[46] compared resampling techniques on rule-based classifiers for credit scoring with an imbalanced dataset from a major Iranian bank, and they found that random oversampling yielded better performance for all classifiers. However, data sampling may not be the best option for all classifiers, and the effectiveness of sampling depends on the evaluation metric as well as the classifier. For example, researchers have found that sampling did not significantly affect logistic regression, while C4.5 decision trees and random forests had performances that varied for the three metrics used: AUC, top-decile lift, and maximum profit measure [76].

Paper	Technique	Proposal	Description
[19]	Oversampling	Attribute Weighted and kNN Hub on SMOTE (AWH-SMOTE)	Provides more representative neighbors with attribute weighting. Performs better than SMOTE and ADASYN.
[20]	Oversampling	Variational Oversampling (VOS)	Uses Variational Autoencoders. Performs better than SMOTE and ADASYN for fraud detection and on images.
[21]	Oversampling	Multiobjective Genetic Sampling (MOGASamp)	Evolutionary sampling approach that creates samples for ensembles to improve base classifier accuracy.
[32]	Oversampling	Adaptive Swarm Balancing Algorithm	Combines SMOTE with swarm balancing. Performs well on moderately and highly imbalanced large datasets.
[38]	Oversampling	SMOTE-Subset of Features (SMOTE-SF)	Novel distance metric based on relevant features without feature selection. Achieved largest AUC in experiments.
[53]	Oversampling	Enhanced Minority Oversampling Technique (EMOTE)	Populates dataset with minority class misclassified instances and with correctly classified nearest neighbors.
[25]	Hybrid	SMOTE and Particle Swarm Optimization (PSO) for undersampling	Performs better than most other resampling strategies when tested on four datasets.
[52]	Hybrid	SMOTE Oversampling and Borderline Undersampling (SOS-BUS)	Oversamples and removes some instances from majority class but keeps original instances near decision boundary.

Table 2. Novel oversampling and hybrid techniques since 2015

As shown in Table 2 above, researchers have found that these new sampling techniques perform better than the traditional SMOTE and ADASYN. Results vary depending on the classifiers and datasets used for the experiments, but the general trend for data sampling is that researchers use oversampling more often than hybrid or undersampling techniques. It is best to use data sampling when the size of the dataset is relatively small and when the intended algorithm is known to perform better on balanced class distributions, such as the C4.5 classifier [7].

3.2 Feature selection and extraction

Methods can also select important features and effectively process imbalanced data. In supervised learning, datasets with many features may suffer redundancy and noise, so only relevant variables that add value should be used. Feature selection finds a subset of features that reduces the computational cost of high dimensions while maximizing the prediction power of the classification model. Such techniques can be divided into wrapper, embedded, and filter methods [40]. For financial distress prediction, business experts usually select features according to domain knowledge, while data mining researchers often rely on data alone [74]. Feature extraction reduces dimensionality by using functional mapping to process raw data, and some techniques include Principal Component Analysis and Singular Value Decomposition [23]. The main difference between the two techniques is that feature extraction creates new features from original features, while feature selection returns a subset of the original features. Feature selection and extraction can be used for high-dimensional class imbalance problems where data sampling or algorithm-level methods may not work as well, and these techniques can also help overcome the issue of overfitting [40].

Recent feature selection and extraction methods include the BB (branch and bound)-based hybrid feature selection with imbalance-oriented multiple-classifier ensemble (BBHFS-IOMCE). BBHFS uses a hybrid feature selection based on BB and t-test to determine a feature extraction rate, and it can optimally select feature subsets that maximize classifier performance [62]. Another example of using both approaches is to first transform raw data into real numbers within intervals, and then use the logistic regression significant discriminant to select key indicators and use the Fisher discriminant to weight each feature [55]. Other researchers have incorporated transfer learning to feature selection to develop a group method of data handling neural networks (GMDH), and they have created a method called transferred feature selection based on GMDH (TFSG) to create training data and select feature subsets [77]. The novel feature selection approach multiclass RFE-SOCP (Recursive Feature Elimination with Second-Order Cone Programming) for support vector machines can also outperform other feature selection approaches like Fisher and SVM-RFE on different classifiers [34]. Variance ranking can also be applied to attribute selection, but the data must be discrete or continuous and not categorical [18]. Feature selection techniques are also used for a fuzzy cluster and fuzzy pattern recognition credit rating model, in which indices are divided into quantitative and qualitative indices and then quantitative features are standardized [54].

3.3 Cost-sensitive learning

For imbalanced data, the minority class is outnumbered by the majority class, but researchers are often interested in the minority class instances. As a result, cost-sensitive learning assigns higher costs to the minority class, but it can be difficult to choose an optimal misclassification cost ratio, which means that classification results may not be stable. Three cost-sensitive approaches are (1) making particular classifiers cost-sensitive, (2) using Bayes risk theory to assign each example to its lower risk class, and (3) using meta-models for converting a classification learning dataset into a cost-sensitive one [65]. Cost-sensitive learning can be used for class imbalance when there is a high imbalance ratio and when the cost of false negatives is significantly higher than that of false positives [42].

Example-dependent cost-sensitive decision trees (ECSDT) are another example of cost-sensitive learning that use various example-dependent costs for a new cost-based impurity measure and pruning criteria [8]. These cost-sensitive decision trees use bagging-style methods, where the researchers trained four different base classifiers (bagging, pasting, random forest, and random patches) with three different combination approaches for a total of 12 different models. To make a cost-sensitive algorithm, the researchers used a different cost matrix for each data point to predict either good or bad examples. As shown in Figure below, the cost matrix that they used for credit scoring has the cost of true positives and true negatives as zero. The cost of false positives is equal to the sum of the loss in profit by rejecting a good customer and the false positive cost of giving the loan to an alternative customer. The cost of false negatives is equal to the product of the credit line and the loss given default.

	Actual Positive $y_i = 1$	Actual Negative $y_i = 0$
Predicted Positive $c_i = 1$	$C_{TP_i} = 0$	$C_{FP_i} = r_i + C^a_{FP}$
Predicted Negative $c_i = 0$	$C_{FN_i} = Cl_i \cdot L_{gd}$	$C_{TN_i} = 0$

Figure 6. ECSDT cost matrix for credit scoring

In addition to majority voting as a combination method, [8] proposed two new cost-sensitive combination methods for ECSDT: cost-sensitive weighted voting and cost-sensitive stacking. They evaluated their models using five different

real-world databases and split the data into 50% training, 25% validation, and 25% testing. They found that random patches worked best on base classifiers for cost-sensitive decision trees, but no particular combination method stood out above the rest.

[35] proposed another cost-sensitive modification to decision trees that is cost-sensitive pattern mining with classification by aggregating cost-sensitive patterns (CSPm+CACSP). For their cost-sensitive patterns, they computed the cost of each class as shown in Equations (1) and (2) below, where C_{min} and C_{max} are the misclassification costs of the pattern for the minority and majority class, respectively. S_{min} and S_{max} are the supports of the pattern in the datasets containing the minority and majority class instances. They concluded that CSPm+CACSP results in lower misclassification costs than other common classifiers, and they noted that the best classification results were attained when the cost was equal to the imbalance ratio.

$$C_{min} = S_{min} * C(0,0) + S_{min} * C(0,1)$$

Equation 1. CSPm+CACSP misclassification cost formula for minority class

$$C_{maj} = S_{maj} * C(1,0) + S_{maj} * C(1,1)$$

Equation 2. CSPm+CACSP misclassification cost formula for majority class

Other researchers have experimented with similar ideas of using cost-sensitive methods with bagging and neural networks. [50] compared nine cost-sensitive algorithms on the German credit data and using metrics like error, misclassification cost, sensitivity, specificity, and the Friedman test. Their main findings with respect to cost-sensitive learning were that cost-sensitive approaches reduced cost at the expense of accuracy, and that cost-sensitive bagging algorithms offer the best tradeoff between accuracy and misclassification cost. Moreover, an extended hybrid genetic algorithm with neural networks (HGA-NN) has been proposed that uses cost-sensitive learning [42]. They evaluated their extended HGA-NN on German and Croatian datasets and compared with the original HGA-NN with the same parameter values. They found that their proposed model was better than the standard HGA-NN for all metrics except accuracy.

3.4 Anomaly detection

Anomaly detection is another technique that is related to the problem of credit scoring with class imbalance. Anomalies are unexpected behaviors that differ from most of the data, and the three types are point, collective, and contextual anomalies. These outliers can be identified using visualizations, and they can help solve problems for bank fraud and intrusion detection [43]. Outliers can also lead to biased models, and they may be the consequence of attribute noise and class noise. Some ways to handle anomalies include label-noise robust approaches, label-noise cleansing approaches, and label-noise tolerant approaches [65]. Anomaly detection methods can be used to identify inconsistent data points and classify emerging patterns and trends from the data. These techniques can be applied to credit risk prediction on imbalanced datasets because both anomaly detection and credit risk share the problem of class imbalance [63].

In credit scoring, anomaly detection is useful to estimate default for borrowers. Research shows that linear regression with rule-based classification performs better than with logistic regression for predicting loan default using an anomaly detection algorithm [63]. Entropy-based algorithms such as the Difference in Maximum Entropy (DME) approach have also been used to detect anomalous values for credit scoring. DME can overcome data imbalance because it does not need to be trained on past default instances, and it performs as well as random forests [48].

In fraud detection, a framework called CoDetect can predict fraud by simultaneously using graph-based similarity matrices and feature matrices. Real-world trades usually occur among companies of similar type, so graph matrices with block structures can represent businesses connected by transactions. Graph anomalies like trade between companies of different types are rare, and CoDetect uses this suspicious activity to detect fraud [26]. Another fraud detection approach is Anomaly Prevention using Advanced Transaction Exploration (APATE), which maps past transactions into meaningful features to compare with incoming transactions. APATE uses network-based extensions where edge weights represent the recency of transactions between merchants and credit card holders [67].

Researchers have used the multi-layer perceptron (MLP) with the Levenberg-Marquardt (LM) to modify the current algorithm to account for both cost sensitivity and outliers. They introduce the misclassification cost in their criterion to minimize, and they use label-noise robust approaches to handle anomalies in the data [65]. They tested various combinations of the MLP and LM with and without outliers and cost-sensitive approaches. They concluded that the conjoint use of cost-sensitive weight and robust criterion for outliers improves classifier accuracy.

3.5 Artificial neural networks and support vector machines

Algorithms in recent literature have also used artificial neural networks (ANN) and support vector machines (SVM). These methods are useful on class imbalance problems to obtain a good accuracy and generalize well on classification problems from various domains. However, they may produce models that are biased towards the majority class and do not perform as well on the minority class [2].

For artificial neural networks, associative memories have been used for pattern classification problems on imbalanced environments [14]. A novel example for credit scoring is the Classification Restricted Boltzmann Machine (ClassRBM), which uses a strong classifier and deals with class imbalance [66]. A Restricted Boltzmann Machine is a type of ANN that consists of bidirectional connections between hidden layers and visible layers. Research has also focused on Deep Belief Networks (DBN) that are made of multiple RBMs but are often time consuming [69]. Compared to other machine learning algorithms, DBN achieved a high accuracy rate and low error rate on an imbalanced handwritten dataset [2]. Multilayer perceptrons (MLP) are another type of ANN that are widely used for financial decision support systems, and experimental results show that MLP with five and four hidden layers in the sigmoid activation function perform the best [81]. Some researchers have found that algorithms based on deep learning neural networks do not necessarily outperform tree-based models like random forest and gradient boosting [1], while others claim that neural networks display the best performance on average for higher imbalance ratios [7].

Support vector machines are another type of learner that can perform as well as ANN while overcoming some of their problems [16]. The four types of SVM models are standard, modified, hybrid, and ensembles. The most commonly used SVM for credit scoring is hybrid models, and the most popular credit datasets for SVMs are the German and Australian datasets [22]. SVMs have been used for cost-sensitive learning by assuming that the distribution of examples near the borderline is similar to the global imbalance ratio. Since this may not always be the case, boosted support vector machines have been proposed that select the most important examples and more accurately calculate misclassification costs [79]. There has also been research on profit-based support vector machines for classification on imbalanced data such as the l2l-SVM and l1l-SVM [37]. Support vector machines have been used with multiple kernel learning for binary classification on imbalanced data by applying SMOTE and optimizing the SVM parameters and weights of kernel functions [47]. In addition, SVMs have been used for peer-to-peer lending classification problems. The instance-based entropy fuzzy support vector machine (IEFSVM) can outperform six other state-of-the-art models for loan status prediction [13].

3.6 Rule-based classifiers

Rule-based classification such as associative classification predicts outputs using if-then rules. This approach has been applied to class imbalance problems, but there does not appear to be a specific best use case for such classifiers [5]. When rule-based classifiers like CBA, CBA2, CMAR-C, CPAR-C, and Fuzzy-FARCHD-C are evaluated on imbalanced datasets, researchers found that Fuzzy-FARCHD-C was the most promising in terms of accuracy [5]. Others have compared classification algorithms such as JRip, DT, OneR, ZeroR, Fuzzy Rule, PART, and Genetic Programming (GP) to find the best rule-based method satisfying accuracy, sensitivity, and specificity. Their experimental results show that the GP algorithm obtained the most efficiency for predicting credit risk [58].

3.7 Ensemble classifiers

Classifier ensembles have performed well on imbalanced data, and some examples are bagging, boosting, and stacking. Ensemble learning trains multiple base models and combines hypotheses to optimally predict outputs. Many modern approaches for working with imbalanced datasets use ensemble learning because it can achieve a high classification accuracy [4]. The best use cases for ensemble classifiers are to reduce bias by combining the results from multiple classifiers, to obtain a good classification accuracy, and to overcome some of the disadvantages of data sampling like overfitting and removing data points [51].

Ensembles are created using either different training sets, feature subsets, base classifiers, or combination schemes. Bagging, also known as bootstrap aggregating, learns base models independently and in parallel. Boosting builds ensembles incrementally by learning base classifiers sequentially. There are also heterogeneous weak learners such as stacking that use base models trained on a single dataset but are combined using a meta-model to predict outputs [51]. There are various ways to choose classifiers for such multiple classifier systems (MCS), and research has compared dynamic classifier selection (DCS) and dynamic ensemble selection (DES) methods. Generally, DES techniques like META-DES outperform DCS ones, and dynamic selection improves baseline methods like Majority Voting and ensembles like AdaBoost [15]. Figure 7 below shows the common ensemble learning methods that will be discussed in this section.

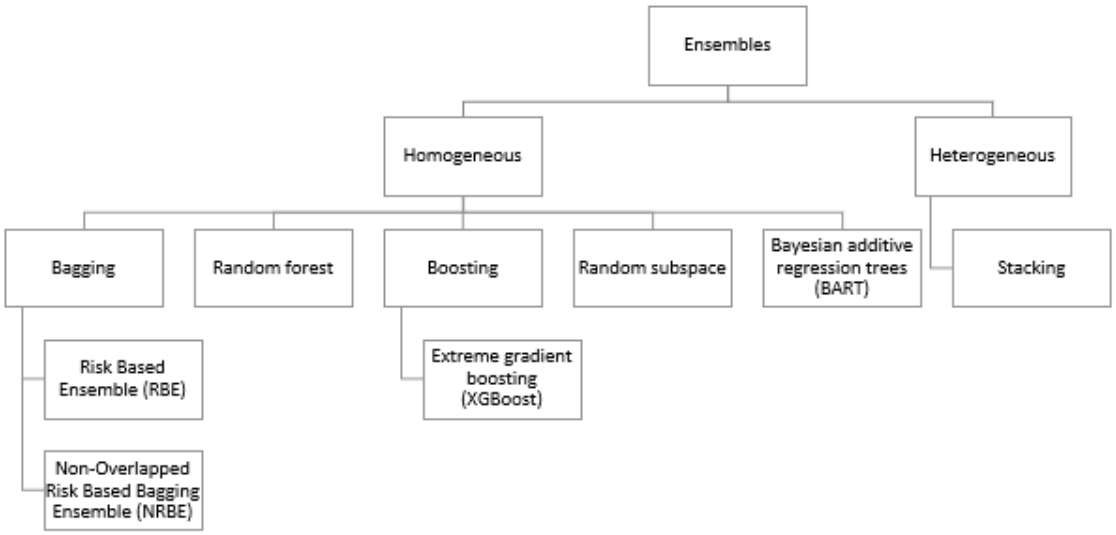

Figure 7. Hierarchy of ensemble learning methods

Random Forest is an enhanced bagging ensemble classifier using multiple decision trees. It is useful because it handles overfitting, reduces variance, and uses independent classifiers. Random forests select random subsets of training data to train decision trees. They then make a final prediction by taking a majority vote of individual tree predictions [9]. Researchers have proposed modifications to traditional random forests such as a class weights voting (CWsRF) algorithm that assigns different weights for each class to better represent the majority and minority classes [78].

Another example is the Risk Based Ensemble (RBE) that modifies regular bagging by using Naïve Bayes as the base learner and creates overlapped bags for training without the problem of underfitting [60]. Furthermore, Non-Overlapped Risk Based Bagging Ensemble (NRBE) extends regular bagging and includes a non-overlapped bag creation method, which replaces the weak learner with Naïve Bayes [61]. In fraud detection, bagging classifiers using decision trees have been used because decision trees perform well on imbalanced data by weighting individual classifier results and reduce overfitting [71]. Random subspace, also known as feature bagging, is another ensemble method similar to bagging. Researchers have found that profit-based models (PBMs) designed with random subspace perform more accurately than those with bagging or boosting [17].

On the contrary, Gradient Boosting reduces bias and variance and uses sequential classifiers, but it could lead to overfitting. It learns from previous mistakes by saving residual errors and makes a final prediction using all prior predictions. The most common boosting algorithm in research is AdaBoost, but there is a new popular method called XGBoost. XGBoost is a powerful gradient boosting system released in 2014 that improves model speed and accuracy [11]. A novel approach using extreme gradient boosting for an ensemble of decision trees has been proposed for bankruptcy prediction. When tested on data from Polish companies against reference classifiers and other ensembles, extreme gradient boosting with new synthetic features produced significantly better results [80].

Other boosting variations include the Grabit model, which extends the Tobit model with trees as base learners for default prediction. They analyzed their model on loan data from Swiss small and medium-sized enterprises using varied sample sizes and imbalance ratios. Grabit outperforms other state-of-the-art approaches like logistic regression, random forest, tree-boosted logistic regression, neural network, and tree-boosted multinomial logistic regression [56]. Another boosting algorithm is MEBoost, which alternately uses decision tree and extra tree classifier as two weak estimators with AdaBoost. MEBoost has shown to perform well against SMOTEBoost, RUSBoost, AdaBoost, DataBoost, EUSBoost, and Easy Ensemble [45].

Bayesian additive regression trees (BART) are an ensemble model similar to gradient boosting that is based on the sum of Bayesian trees. When BART and Random Forest were applied to credit scoring and compared with logistic regression on balanced and imbalanced data, researchers concluded that BART was superior to logistic regression for both datasets and to random forest for the unbalanced sample [6].

Heterogeneous ensembles have also been used for imbalanced learning and default prediction. An ensemble of XGBoost, Deep Neural Network (DNN), and Logistic Regression (LR) individual learners can predict peer-to-peer lending default risk, where XGBoost was the main algorithm and DNN and LR were applied to boost prediction accuracy [33]. Stacking ensembles have also been explored such as calling the StackingC method and using J48, Logistic Regression, and Bagging as the base classifiers. A voting method is then applied to combine individual classifier predictions [51].

Typically, every heterogeneous learner category has only one individual learner, but researchers have also attempted ensemble models with multiple individual learners to compare their performance to those with only one learner. [12] proposed a genetic algorithm-based ensemble learning model (GAEM) that uses greedy selection to find the best combination of individual learners. GAEM resulted in better performance than AdaBoost, Random Forest, and Gradient Boosting. Furthermore, when combining heterogeneous classifiers, a consensus system that uses decisions from all individual classifiers to construct an output has also been proposed [3]. This consensus approach performed better than LR, MARS, and traditional combination methods across five datasets.

3.8 Hybrid approaches

State-of-the-art research on class imbalance commonly combine data level approaches with ensemble learning because individual classifiers have difficulty on imbalanced data. The most popular hybrid is AdaBoost with various sampling techniques. For example, SMOTEBoost merges AdaBoost with the oversampling technique SMOTE. RUSBoost uses AdaBoost and random undersampling to remove majority class examples, and it has been shown to have better performance than SMOTEBoost [27]. EUSBoost is another hybrid ensemble-based model that uses AdaBoost and an evolutionary undersampling approach. Evolutionary undersampling applies random undersampling to imbalanced data, and it evolves the data until they are the best possible.

[44] published their findings on combining AdaBoost with cluster-based undersampling, a data level method that clusters majority and minority class instances. Their CUSBoost performs well against AdaBoost, RUSBoost, and SMOTEBoost on highly imbalanced datasets. Another hybrid algorithm is Tomek-link Undersampling-based Boosting (TLUSBoost) that combines Tomek-link and redundancy-based undersampling and AdaBoost for class imbalance [27]. Researchers have also proposed a resample-based ensemble for class imbalance that consists of a static classifier and multiple dynamic classifiers sliding window [73].

A combination of ADASYN and XGBoost has also been proposed for handling class imbalance for predicting rail defects in the transportation industry, and research shows that XGBoost outperforms SVM, Random Forest, and Logistic Regression [39]. Researchers have also used extreme gradient boosting but have avoided resampling due to a high number of false positives, and their model was able to achieve an accuracy of 99.5% [28]. Others have followed suit and used bagging and boosting ensembles without sampling. The Prudential Multiple Consensus (PMC) model uses novel prudential criteria for fraud detection, and the researchers chose five classification algorithms for their experiments: Multilayer Perceptron, Gaussian Naïve Bayes, Adaptive Boosting, Gradient Boosting, and Random Forests. They concluded that PMC performed better than the five competitor models because it can more correctly classify fraudulent transactions [49]. Moreover, multiple classifier ensembles have also been combined with transfer learning for credit scoring, such as a clustering and selection-based transfer ensemble (CSTE) that uses information from various related source domains for modeling in the target domain proposed by [70].

Other researchers have analyzed the effects of various data sampling techniques on both individual and ensemble learners. [57] used BSMOTE, CNN, ROS, RUS, SLS, SMOTE, and no sampling on C4.5, LR, SVM, AdaBoostM1, DTBagging, and RF. They found that logistic regression performs better with oversampling and C4.5 performs better with undersampling for individual learners; AdaBoostM1 and DTBagging perform better with oversampling and Random Forest worked best with undersampling for ensembles. Overall, oversampling and DTBagging outperformed the other combinations. [64] used random oversampling and SMOTE to balance data, and they discovered that an ensemble model that combines principal component analysis plus discriminant analysis (PCA-DA), Logit, and DT under a consensus rule produced an accurate model.

Table 3 below summarizes the ensemble and hybrid methods in recent research papers that were published since 2015. In general, hybrid classification approaches have similar use cases to ensemble classifiers, but the former has the advantage of balancing the dataset before applying the algorithm [39]. However, multiple classifiers without data-level techniques may work better in some cases such as XGBoost [28].

Paper	Approach	Data	Models	Metrics
[27]	TLUSBoost (Tomek-link Undersampling + AdaBoost)	16 benchmark datasets	TLUSBoost*, EasyEnsemble, BalanceCascade, SMOTEBoost, RUSBoost	Accuracy, F-measure
[28]	XGBoost without resampling	Estonian bankruptcy data (17953/208757)	XGBoost for 12 individual and 1 ensemble model	20+ metrics including Accuracy, AUC, etc
[39]	XGBoost with ADASYN	0.08 proportion are defects of 60 million samples	XGBoost*, SVM, LR, RF with combinations of RFE, SVD, BO, and RS	Accuracy, TPR, TNR, FPR, FNR
[44]	CUSBoost (Cluster-based Undersampling + AdaBoost)	13 from Keel, imbalance ratios from 1.87 to 853	CUSBoost*, AdaBoost, RUSBoost, SMOTEBoost	AUC
[49]	Prudential Multiple Consensus (PMC)	Credit card fraud (492/284807)	Prudential criterion*, complete agreement, majority voting, weighted voting with MLP, NB, AdaBoost, Gradient Boosting, RF, and DT Bagging	Sensitivity, Specificity, Fallout, AUC, Miss Rate
[70]	Clustering and selection-based transfer ensemble (CSTE)	PAKDD (341/2471) and UK credit (323/1225)	CSTE*, Subagging, Subagg-OT, TCA, TrBagg, TrAdaBoost using SVM	AUC
[73]	Resample-based Ensemble Framework for Drifting Imbalanced Stream (RE-DI)	PAKDD, Poker, Give Me Some Credit, Forest Cover Type	RE-DI*, OOB, UOB, LB, ARF	Prequential AUC

Table 3. Novel ensemble and hybrid methods since 2015

3.9 Performance metrics

There are three types of evaluation metrics: threshold, probability, and ranking metrics. Each of these types have a different aim for evaluating classifiers, and threshold and ranking are the most common. Metrics also have different purposes such as (1) evaluate the quality of trained classifier on unseen test data, (2) using metric for model selection, and (3) using metrics to select optimal model from all generated solutions during training which would then be used for testing [24]. Nearly all threshold, probability, and ranking metrics could be used for evaluating models on testing data and for selecting the best models.

Various performance metrics are used to evaluate classification models. Accuracy is often calculated to determine the ratio of correct predictions to total inputs. However, classification accuracy works well only when there is a relatively equal number of good and bad examples. In the case of highly imbalanced datasets, accuracy can be misleading because models can achieve a high accuracy by simply classifying all examples as the majority class [55]. For instance, a model can achieve 99% accuracy on data with a class imbalance ratio of 99:1 by labelling all examples as good credit. This model overfits the data and does not consider the misclassification costs of false positives, which are default. Consequently, some papers use average accuracy rate or other metrics to evaluate classification accuracy.

Instead, performance metrics that are better suited for classification on imbalanced data include a confusion matrix, precision, recall, AUC, and MCC. A confusion matrix displays the number of correct predictions and types of misclassified examples. In Figure 8 below, the positive class refers to bad credit risk, and the negative class refers to good credit risk. If the predicted class matches the actual class, then the result will either be a true positive or true negative. However, if a bad credit example is misclassified as a good one, then the result will be a false negative. It is important to note that some papers may refer to the positive class as good credit risk, and negative class as bad credit risk [10].

Actual class	Predicted class	
	Positive	Negative
Positive	true positive (tp)	false negative (fn)
Negative	false positive (fp)	true negative (tn)

Figure 8. Confusion matrix of prediction results

Precision measures the fraction of true positives that are predicted as positive, while recall measures the fraction of true positives that are actually positive [24]. AUC is short for the Area under ROC (Receiver Operating Characteristic)

Curve, and it is a graphical plot that describes the relation between the true positive rate and false positive rate. Table 4 below shows the mathematical formulas for calculating precision and recall, as well as the true positive rate and false positive rate needed for AUC. MCC stands for the Matthews correlation coefficient between the predicted and actual binary classifications. Some researchers suggest that MCC is the best metric to use if classification error needs to be considered [36].

Metric	Definition	Description
Precision	(tp)/(tp+fp)	Fraction of true positives that are predicted as positive
Recall	(tp)/(tp+fn)	Fraction of true positives that are actually positive, i.e., TPR
True positive rate (TPR)	(tp)/(tp+fn)	Proportion of actual positives which are identified as positive
False positive rate (FPR)	(fp)/(tn+fp)	Proportion of actual negatives which are identified as positive

Table 4. Some performance metrics for imbalanced data

4 Discussion

The commonly used techniques in recent research to predict financial credit risk and classify instances from highly imbalanced data fall into data-level, algorithm-level, and ensemble or hybrid methods. Many of the proposed novel synthetic oversampling methods such as EMOTE, SMOTE-SF, AWH-SMOTE, and VOS populate datasets with more representative minority class instances. Hybrid sampling approaches like SOS-BUS and SMOTE with PSO are less studied among the analyzed research papers, but they have also been effective for balancing skewed data. Novel feature selection methods such as BBHFS, TFSG, and RFE-SOCP have been used in some papers to help maximize classifier performance and minimize dimensionality and highly correlated features. Data sampling is best used for relatively small or highly skewed datasets, and feature selection and extraction methods can reduce the high dimensionality of datasets.

Cost-sensitive learning is an algorithm level technique that has been applied to ensembles of decision trees. Researchers have proposed combination methods like cost-sensitive weighted voting and cost-sensitive stacking in addition to experimenting with different bagging-like ensemble methods. Their main findings are that cost sensitivity reduces cost at the expense of accuracy, and that bagging and random patches worked best for decision trees. However, [8] suggested that future research could use boosting instead of bagging for their example-dependent cost-sensitive trees. Anomaly detection is another technique that can be used in conjunction with cost-sensitive learning. It has been applied to credit card fraud detection but seldom to credit scoring. It is difficult to find papers that predict credit risk by discovering outliers besides [48] and [63], so future research can explore anomaly detection for credit risk prediction on imbalanced data. On low default portfolios, cost-sensitive ensembles retain the benefits of ensemble methods and allow models to make more informed decisions that prioritize the minority class. In general, cost-sensitive methods can be used for class imbalance when there is a high imbalance ratio and when the cost of false negatives is significantly higher than that of false positives. Anomaly detection works best to identify inconsistent data points and classify emerging patterns and trends from the data. These methods can be applied to credit risk prediction on imbalanced datasets because both anomaly detection and credit risk share the problem of class imbalance.

In this paper, individual algorithms like artificial neural networks, support vector machines, and rule-based classifiers are also analyzed, as they have appeared in research for imbalanced learning. Variants of ANNs in recent papers include Multilayer Perceptrons and deep neural networks like Restricted Boltzmann Machines and Deep Belief Networks. Although such neural networks can achieve high accuracies on imbalanced data, they do not necessarily outperform tree-based ensembles [1]. SVMs can overcome some challenges of ANNs, and many SVMs use a hybrid approach and have been applied to cost-sensitive learning. Rule-based classifiers such as the Fuzzy-FARCHD-C and Genetic Programming algorithms are most promising in terms of credit risk prediction. These methods are useful on class imbalance problems to obtain a good accuracy and generalize well on classification problems from various domains.

Ensembles and hybrid models are common in state-of-the-art research for classification with imbalanced data. Many researchers have used random forests and bagging ensembles with Naïve Bayes or decision trees as base learners. AdaBoost is still popular for classifier ensembles and hybrids that combine data sampling with ensembles, and Extreme Gradient Boosting is a recent method that has also shown impressive results for financial distress prediction. Novel hybrids since 2015 include CUSBoost (Cluster-based Undersampling + AdaBoost), TLUSBoost (Tomek-link Undersampling + AdaBoost), and a combination of ADASYN and XGBoost. Some researchers have mixed and matched various sampling techniques on ensemble learners, and their analysis shows that the DTBagging ensemble with oversampling outperformed other tested combinations. A study that uses XGBoost instead of AdaBoost with more sampling methods on highly imbalanced data would be interesting for future research. The best use cases for

ensemble and hybrid classifiers are to reduce bias by combining the results from multiple classifiers, to obtain a good classification accuracy, and to overcome some of the disadvantages of data sampling.

When evaluating classification models, many papers have demonstrated that the German credit dataset from the University of California Irvine (UCI) Machine Learning Repository is often used as a benchmark for testing model performance [22, 42, 50]. This confirms the findings from similar studies like [31], which also noted that relying on the same dataset may introduce bias in the creation of new algorithms. Most papers that compared models used Area under the ROC curve (AUC) and accuracy metrics to analyze performance, which affirms the findings by [24].

This paper provides an update to the related works by analyzing novel machine learning techniques since 2015. The results reinforce the current understanding that state-of-the-art classification methods for highly imbalanced datasets mainly use data level, cost-sensitive, and ensemble approaches. There are some areas and approaches that have been little explored, as well as some techniques that are popular and have worked well for classification on class imbalance problems. The three proposed models that will be implemented and improved upon are cost-sensitive decision tree bagging ensembles like Example-Dependent Cost-Sensitive Decision Trees (ECSDT), hybrid classifiers with data-level and algorithm-level techniques like Cluster-Based Undersampling with Boosting (CUSBoost), and Extreme Gradient Boosting (XGBoost).

The proposed ECSDT will extend the work of [8] by using the bagging ensemble with random patches as the inducer and weighted voting as the combination method. This model was selected because it uses an ensemble classifier and cost-sensitive learning, which would reduce bias and consider the misclassification cost of false positives and false negatives. The proposed CUSBoost is a boosting algorithm that samples the dataset by first clustering the majority and minority instances. While this model uses undersampling, it works best on clusterable data and has shown performance than AdaBoost, SMOTEBoost, and RUSBoost when comparing AUC. Lastly, the proposed extreme gradient boosting model modifies the work of [28] and [39] by using Bayesian hyperparameter optimization. This method does not use data sampling because XGBoost has a `scale_pos_weight` parameter that can consider class imbalance.

The implemented algorithms will use the German credit dataset, which has 700 good credit examples and 300 bad examples, as a benchmark for analyzing the three models. The results will be evaluated using performance metrics such as AUC and MCC. Feature selection and extraction, anomaly detection, and hyperparameter optimization may also be applied to provide insights from the data and improve the models. For example, anomaly detection could be applied to better understand the distribution of the data for feature selection and identify potential customers with poor credit.

5 Conclusion

This paper provides insight into state-of-the-art machine learning methods for the challenge of predicting credit risk with low default portfolios. Techniques for approaching highly imbalanced datasets and classification problems in credit scoring and in other contexts such as credit card fraud detection are considered. An ideal credit risk prediction system is valued by banks and other lenders and could save businesses the substantial cost of default.

A systematic search methodology was used to find relevant papers, which made 96 documents available for the study. Data level, cost-sensitive, and ensemble methods for imbalanced data are discussed. Research shows that combined data sampling and ensemble classifiers are most popular and work well for credit scoring, and avenues for future studies that use ensemble and hybrid approaches with data-level and cost-sensitive learning are recommended. Feature selection and extraction will reduce dimensions to improve running time, and anomaly detection can help provide insights from the data and find emerging patterns.

The three proposed models are a bagging ensemble with cost-sensitive decision trees (ECSDT), a hybrid algorithm that uses data sampling and boosting (CUSBoost), and extreme gradient boosting without data sampling (XGBoost). The advantages of ECSDT are that it reduces cost and bias, but it does come at the expense of accuracy. CUSBoost undersamples the data to overcome class imbalance and has worked well on highly imbalanced datasets in research, but it could remove some data points that would be beneficial to the model from the modified dataset. XGBoost has the potential to obtain a high accuracy but works best on balanced data so the algorithm parameters will need to be carefully tuned.

6 Limitations

The systematic search conducted for this state-of-the-art paper used only the database Scopus to find documents, which may have excluded some relevant papers from the study. There were also restrictions to the subscription used for accessing full-length papers from online academic journals. Future research should apply the search equation to other databases like Web of Science and use an all-inclusive subscription for a more comprehensive study.

References

[1] Addo, Peter Martey Guegan, Dominique Hassani, Bertrand. (2018). Credit Risk Analysis Using Machine and Deep Learning Models. Risks.

[2] A'fifah, A'inur Ismail, Amelia Ritahani Ahmad, Abdullah. (2018). Comparative Performance of Deep Learning and Machine Learning Algorithms on Imbalanced Handwritten Data. International Journal of Advanced Computer Science and Applications.

[3] Alaraj, Maher Abbod, Maysam. (2016). Classiers consensus system approach for credit scoring. Knowledge-Based Systems.

[4] Ali, Shahid Sremath Tirumala, Sreenivas Sarrafzadeh, Abdolhossein. (2015). Ensemble learning methods for decision making: Status and future prospects.

[5] Ali, Zulfiqar Ahmad, Rehan Nadeem Akhtar, Muhammad Hussain Chuhan, Zishan Maria Kiran, Hafiza Shahzad, Waseem. (2018). Empirical Study of Associative Classifiers on Imbalanced Datasets in KEEL.

[6] Alves de Brito Filho, Daniel Artes, Rinaldo. (2018). Application of bayesian additive regression trees in the development of credit scoring models in Brazil. Production.

[7] Andric, Kristina Kalpic, Damir Bohacek, Zoran. (2018). An insight into the effects of class imbalance and sampling on classification accuracy in credit risk assessment. Computer Science and Information Systems.

[8] Bahnsen, Alejandro Correa Aouada, Djamila Ottersten, Björn. (2015). Ensemble of Example-Dependent Cost-Sensitive Decision Trees.

[9] Biau, Gérard Scornet, Erwan. (2015). A Random Forest Guided Tour. TEST.

[10] Chen, Ning Ribeiro, Bernardete Chen, An. (2015). Financial credit risk assessment: a recent review. Artificial Intelligence Review.

[11] Chen, Tianqi Guestrin, Carlos. (2016). XGBoost: A Scalable Tree Boosting System.

[12] Chen, Xiaohui Zhang, Zhiyao Zhang, Ze. (2018). Real-time equipment condition assessment for a class-imbalanced dataset based on heterogeneous ensemble learning. Eksploatacja i Niezawodnosc - Maintenance and Reliability.

[13] Cho, Poongjin Chang, Woojin Song, Jae Wook. (2019). Application of Instance-Based Entropy Fuzzy Support Vector Machine in Peer-To-Peer Lending Investment Decision. IEEE Access.

[14] Cleofas, Laura Sánchez, Josep García, Vicente Valdovinos, Rosa. (2016). Associative learning on imbalanced environments: An empirical study. Expert Systems with Applications.

[15] Cruz, Rafael Sabourin, Robert Cavalcanti, George. (2018). Dynamic classifier selection: Recent advances and perspectives. Information Fusion.

[16] Danėnas, Paulius Garšva, Gintautas. (2015). Selection of Support Vector Machines based classifiers for credit risk domain. Expert Systems with Applications.

[17] du Jardin, Philippe. (2016). A two-stage classification technique for bankruptcy prediction. European Journal of Operational Research.

[18] Ebenuwa, Solomon Saeed Sharif, Mhd Alazab, Mamoun Al-Nemrat, A. (2019). Variance Ranking Attributes Selection Techniques for Binary Classification Problem in Imbalance Data. IEEE Access.

[19] Fahrudin, Tora Buliali, Joko Fatichah, Chastine. (2019). Enhancing the performance of smote algorithm by using attribute weighting scheme and new selective sampling method for imbalanced data set. International Journal of Innovative Computing, Information and Control.

[20] Fajardo, Val Andrei Findlay, David Houmanfar, Roshanak Jaiswal, Charu Liang, Jiaxi Xie, Honglei. (2018). VOS: a Method for Variational Oversampling of Imbalanced Data.

[21] Fernandes, Everlândio de Carvalho, Andre Coelho, André. (2015). An evolutionary sampling approach for classification with imbalanced data.

[22] Goh, R. Y. Lee, Lai Soon. (2019). Credit Scoring: A Review on Support Vector Machines and Metaheuristic Approaches. Advances in Operations Research.

[23] Haixiang, Guo Yijing, Li Shang, Jennifer Mingyun, Gu Yuanyue, Huang Bing, Gong. (2016). Learning from class-imbalanced data: Review of methods and applications. Expert Systems with Applications.

[24] Hossin, Mohammad Sulaiman, M. N. (2015). A Review on Evaluation Metrics for Data Classification Evaluations. International Journal of Data Mining Knowledge Management Process.

[25] Hu, Zhongyi Chiong, Raymond Pranata, Ilung Bao, Yukun Lin, Yuqing. (2018). Malicious Web Domain Identification using Online Credibility and Performance Data by Considering the Class Imbalance Issue. Industrial Management Data Systems.

[26] Huang, Dongxu Mu, Dejun Yang, Libin Cai, Xiaoyan. (2018). CoDetect: Financial Fraud Detection With Anomaly Feature Detection. IEEE Access.

[27] Kumar, Sujit Biswas, Saroj Kr. Devi, Debashree. (2018). TLUSBoost algorithm: a boosting solution for class imbalance problem. Soft Computing.

[28] Kuusik, Jüri Kungas, Peep. (2018). Business Credit Scoring of Estonian Organizations.

[29] Leevy, Joffrey Khoshgoftaar, Taghi Bauder, Richard Seliya, Naeem. (2018). A survey on addressing high-class imbalance in big data. Journal of Big Data.

[30] Leo, Martin Sharma, Suneel Maddulety, K. (2019). Machine Learning in Banking Risk Management: A Literature Review. Risks.

[31] Lessmann, Stefan Baesens, Bart Seow, Hsin-Vonn Thomas, Lyn. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research.

[32] Li, Jinyan Liu, Lian-sheng Fong, Simon Wong, Raymond Mohammed, Sabah Fiaidhi, Jinan Sung, Yunsick Wong, Kelvin. (2017). Adaptive Swarm Balancing Algorithms for Rare-event Prediction in Imbalanced Healthcare Data. PLoS ONE.

[33] Li, Wei Ding, Shuai Wang, Hao Chen, Yi Yang, Shanlin. (2019). Heterogeneous ensemble learning with feature engineering for default prediction in peer-to-peer lending in China.

[34] López, Julio Maldonado, Sebastián. (2015). Robust feature selection for multiclass Support Vector Machines using second-order cone programming. Intelligent Data Analysis.

[35] Loyola-González, Octavio Martínez-Trinidad, José Francisco Carrasco-Ochoa, Jesus Ariel García-Borroto, Milton. (2019). Cost-Sensitive Pattern-Based classification for Class Imbalance problems. IEEE Access.

[36] Luque, A Carrasco, Alejandro Martín, Alejandro de Las Heras-Garcia de Vinuesa, Ana. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. Pattern Recognition.

[37] Maldonado, Sebastián Bravo, Cristián López, Julio Perez, Juan. (2017). Integrated framework for profit-based feature selection and SVM classification in credit scoring. Decision Support Systems.

[38] Maldonado, Sebastián López, Julio Vairetti, Carla. (2018). An alternative SMOTE oversampling strategy for high-dimensional datasets. Applied Soft Computing.

[39] Mohammadi, Reza He, Qing Ghofrani, Faeze Pathak, Abhishek Aref, Amjad. (2019). Exploring the impact of foot-by-foot track geometry on the occurrence of rail defects. Transportation Research Part C Emerging Technologies.

[40] Montiel, Jacob Bifet, Albert Abdessalem, Talel. (2017). Predicting over-indebtedness on batch and streaming data.

[41] Namvar, Anahita Siami, Mohammad Rabhi, Fethi Naderpour, Mohsen. (2018). Credit risk prediction in an imbalanced social lending environment. International Journal of Computational Intelligence Systems.

[42] Oreški, S Oreški, G. (2018). Cost-sensitive learning from imbalanced datasets for retail credit risk assessment. TEM Journal.

[43] Park, W Ahn, H. (2018). A multi-level E-commerce anomaly detection model using OMSVM. Journal of Theoretical and Applied Information Technology.

[44] Rayhan, Farshid Ahmed, Sajid Mahbub, Asif Jani, Md. Rafsan Shatabda, Swakkhar Farid, Dewan Md. (2017). CUSBoost: Cluster-based Under-sampling with Boosting for Imbalanced Classification.

[45] Rayhan, Farshid Ahmed, Sajid Mahbub, Asif Jani, Md. Rafsan Shatabda, Swakkhar Farid, Dewan Md. Chowdhury, Mofizur Rahman. (2017). MEBoost: Mixing Estimators with Boosting for Imbalanced Data Classification.

[46] Sadatrasoul, Seyyed Mohsen Gholamian, Mohammad Shahanaghi, K. (2015). Extracting rules from imbalanced data: The case of credit scoring. Journal of Information Systems and Telecommunication.

[47] Saeed, Sana Ong, Hong Choon. (2019). Performance of SVM with Multiple Kernel Learning for Classification Tasks of Imbalanced Datasets. Pertanika Journal of Science and Technology.

[48] Saia, Roberto Carta, Salvatore. (2016). An Entropy Based Algorithm for Credit Scoring. Lecture Notes in Business Information Processing.

[49] Saia, Roberto Carta, Salvatore Reforgiato Recupero, Diego Fenu, Gianni. (2019). Fraud Detection for E-commerce Transactions by Employing a Prudential Multiple Consensus Model. Journal of Information Security and Applications.

[50] Saidi, Meryem Settouti, Nesma El Habib Daho, Mostafa El Amine Bechar, Mohammed. (2018). Comparison of ensemble cost sensitive algorithms: application to credit scoring prediction. International Conference on Advanced Aspects of Software Engineering.

[51] Salunkhe, Uma R. Mali, Suresh. (2016). Classifier Ensemble Design for Imbalanced Data Classification: A Hybrid Approach. Procedia Computer Science.

[52] Salunkhe, Uma R. Mali, Suresh. (2018). A Hybrid Approach for Class Imbalance Problem in Customer Churn Prediction: A Novel Extension to Under-sampling. International Journal of Intelligent Systems and Applications.

[53] Santhalingam, Babu Ananthanarayanan, N.R. (2017). EMOTE: Enhanced Minority Oversampling TEchnique. Journal of Intelligent Fuzzy Systems.

[54] Shi, Baofeng Chen, Nan Wang, Jing. (2016). A credit rating model of microfinance based on fuzzy cluster analysis and fuzzy pattern recognition: Empirical evidence from Chinese 2,157 small private businesses. Journal of Intelligent Fuzzy Systems.

[55] Shi, Baofeng Wang, Jing Qi, Junyan Cheng, Yanqiu. (2015). A Novel Imbalanced Data Classification Approach Based on Logistic Regression and Fisher Discriminant. Mathematical Problems in Engineering.

[56] Sigrist, Fabio Hirnschall, Christoph. (2018). Grabit: Gradient Tree Boosted Tobit Models for Default Prediction.

[57] Sisodia, Dilip Verma, Upasana. (2018). The Impact of Data Re-Sampling on Learning Performance of Class Imbalanced Bankruptcy Prediction Models. International Journal on Electrical Engineering and Informatics.

[58] Smiti, Salima Soui, Makram Gasmi, Ines. (2018). A Comparative Study of Rule Based Classification Algorithms For Credit Risk Assessment.

[59] Somasundaram, Akila Reddy, U. Srinivasulu. (2016). Data Imbalance: Effects and Solutions for Classification of Large and Highly Imbalanced Data.

[60] Somasundaram, Akila Reddy, U. Srinivasulu. (2017). Risk based bagged ensemble (RBE) for credit card fraud detection.

[61] Somasundaram, Akila Reddy, U. Srinivasulu. (2018). Credit Card Fraud Detection using Non-Overlapped Risk based Bagging Ensemble (NRBE).

[62] Sun, Jie Lee, Young-Chan Li, Hui Huang, Qing-Hua. (2015). Combining BB-based hybrid feature selection and the imbalance-oriented multiple-classifier ensemble for imbalanced credit risk assessment. Technological and Economic Development of Economy.

[63] Surti, Krunal M. Patel, Ashish. (2017). Effective Credit Default Scoring using Anomaly Detection. International Journal of Science Technology Engineering.

[64] Tai, Chung-Ching Lin, Hung-Wen Chie, Bin-Tzong Tung, Chen-Yuan. (2018). Predicting the failures of prediction markets: A procedure of decision making using classification models. International Journal of Forecasting.

[65] Thomas, Philippe. (2015). Perceptron Learning for Classification Problems - Impact of Cost-Sensitivity and Outliers Robustness.

[66] Tomczak, Jakub Zięba, Maciej. (2015). Classification Restricted Boltzmann Machine for comprehensible credit scoring model. Expert Systems with Applications.

[67] Van Vlasselaer, Véronique Bravo, Cristián Caelen, Olivier Eliassi-Rad, Tina Akoglu, Leman Snoeck, Monique Baesens, Bart. (2015). APATE: A Novel Approach for Automated Credit Card Transaction Fraud Detection using Network-Based Extensions. Decision Support Systems.

[68] Wang, Hong Xu, Qingsong Zhou, Lifeng. (2015). Large Unbalanced Credit Scoring Using Lasso-Logistic Regression Ensemble. PLoS ONE.

[69] Xenopoulos, Peter. (2017). Introducing DeepBalance: Random Deep Belief Network Ensembles to Address Class Imbalance.

[70] Xiao, Jin Xie, Ling Liu, Dunhu Xiao, Yi Hu, Yi. (2016). A clustering and selection based transfer ensemble model for customer credit scoring. Filomat.

[71] Zareapoor, Masoumeh Shamsolmoali, Pourya. (2015). Application of Credit Card Fraud Detection: Based on Bagging Ensemble Classifier. Procedia Computer Science.

[72] Zhang, Chongsheng Liu, Changchang Zhang, Xiangliang Almpanidis, George. (2017). An Up-to-Date Comparison of State-of-the-Art Classification Algorithms. Expert Systems with Applications.

[73] Zhang, Hang Liu, Weike Wang, Shuo Shan, Jicheng Liu, Qingbao. (2019). Resample-based Ensemble Framework for Drifting Imbalanced Data Streams. IEEE Access.

[74] Zhou, Ligang Lu, Dong Fujita, Hamido. (2015). The Performance of Corporate Financial Distress Prediction Models with Features Selection Guided by Domain Knowledge and Data Mining Approaches. Knowledge-Based Systems.

[75] Zhou, Xun Cheng, Sicong Zhu, Meng Guo, Chengkun Zhou, Sida Xu, Peng Xue, Zhenghua Zhang, Weishi. (2018). A state of the art survey of data mining-based fraud detection and credit scoring. MATEC Web of Conferences.

[76] Zhu, Bing Baesens, Bart Backiel, Aimée vanden Broucke, Seppe. (2017). Benchmarking sampling techniques for imbalance learning in churn prediction. Journal of the Operational Research Society.

[77] Zhu, Bing Niu, Yongge Xiao, Jin Baesens, Bart. (2016). A new transferred feature selection algorithm for customer identification. Neural Computing and Applications.

[78] Zhu, Min Xia, Jing Jin, Xiaoqing Yan, Molei Cai, Guolong Yan, Jing Ning, Gangmin. (2018). Class Weights Random Forest Algorithm for Processing Class Imbalanced Medical Data. IEEE Access.

[79] Zięba, Maciej Tomczak, Jakub. (2015). Boosted SVM with active learning strategy for imbalanced data. Soft Computing.

[80] Zięba, Maciej Tomczak, Sebastian Tomczak, Jakub. (2016). Ensemble Boosted Trees with Synthetic Features Generation in Application to Bankruptcy Prediction. Expert Systems with Applications.

[81] Zoynul Abedin, Mohammad Guotai, Chi Moula, Fahmida Azad, Asm Shamim Uddin Khan, Mohammed. (2018). Topological applications of multilayer perceptrons and support vector machines in financial decision support systems. International Journal of Finance Economics.