Credit Risk Assessment using Different Outlier Detection Techniques on Tree Based Machine Learning Classifiers: An Experimental Study

Project Report

Submitted By

Purobi Rajbanshi, Roll- 1220CMSH-0042, Reg. No. 121-1212-0733-20

Department of Computer Science
Gour Mahavidyalaya
Malda
August, 2023

BONAFIDE CERTIFICATE

This is to certify that this project report entitled "**Credit Risk Assessment using Different Outlier Detection Techniques on Tree Based Machine Learning Classifiers: An Experimental Study**" submitted to Gour Mahavidyalaya, Malda, is a bonafide record of work done by "Purobi Rajbanshi" under my supervision for the session of 2022-23.

Arijit Bhattacharya
Assistant Professor
Department of Computer Science
Gour Mahavidyalaya
Malda

Declaration by Author(s)

This is to declare that this report has been written by me. No part of the report is plagiarized from other sources. All information included from other sources has been duly acknowledged. We aver that if any part of the report is found to be plagiarized, we are shall take full responsibility for it.

Purobi Rajbanshi

Table of Contents

Credit Risk Assessment using Different Outlier detection Techniques on Tree Based Machine Learning Classifiers: An Experimental Study

1.Introduction:

The research is based on the study of outliers, which are defined as a data point that deviates from the rest of the data. Detection of Anomaly is very emergent issue in many multiple fields like machine learning, deep learning, image processing and statistics which have been researched in the various application area from different domain such as healthcare, agriculture, banking and fraud detection. We shall briefly address outliers with data mining and statistics techniques, application and methods in this paper. Another part of paper is consisting of pros and cons of distinct outlier algorithms and also provides a concise familiarity of discrete types of tools for detecting outliers and reviewed about different databases which we can easily use for outlier analysis. Outlier detection is very crucial aspect in area of data mining because it creates very adverse influence on data set. In today era mostly research becomes data mining oriented, so familiarity with data mining is also essential. The objective of this research is to furnish an understanding of outliers and its methods, application areas to detect anomalies for research [1].

Outliers are observations that deviate significantly from the norm, and their detection has been a critical topic in various research areas and application domains, such as video surveillance, network intrusion detection, and disease outbreak detection. In recent years, deep learning-based techniques have shown to outperform machine learning and shallow approaches for outlier detection in streaming data, which are large and complex datasets. However, developing an effective and appropriate model for outlier detection is challenging due to the dynamic nature and variations of real-world applications and data.
In this research, we propose a novel deep neural network (DNN) model for outlier detection in streaming data. Our model is developed with multiple hidden layers to improve feature abstraction and capabilities. We evaluate our proposed model on four real-world outlier benchmark datasets available at the UCI repository and compare its performance with state-of-the-art approaches. Our experiments demonstrate that our proposed model outperforms both machine learning algorithms and deep learning competitors, resulting in significant performance gains. In particular, our proposed approach achieves much higher accuracy, recall, and f1-score rates of 99.63%, 99.014%, and 99.437%, respectively, compared to other algorithms.
Our proposed deep learning-based approach provides an effective and efficient solution for real-time outlier detection in streaming data [2]. The experimental results show its superior performance compared to existing methods, which makes it a valuable contribution to the field of outlier detection.

Our contribution: A comparative analysis of different Importance based Outlier detection models with the motivation to improve scalability and precision of the algorithms in evaluating retail credit risks has been investigated. The efficiency of the proposed algorithms is evaluated using a real-world German credit and Australian dataset downloaded from the UCI database. Outlier detection of the credit risk evaluation have been investigated using the various classifiers AdaBoost Regressor, Random Forest, Decision Tree, XGBoost Regressor and Light GBM Regressor. Our results are differentiated from in significant ways and indicate that the proposed method attains acceptable results[3].

The current article is organised into five sections. In Section 1, a short introduction to outlier detection techniques were discussed. Afterwards, Outlier importance methods were discussed in Section 3literature review were discussed then followed by the methodology of the research work in Section 4. The utilised parameters and criteria in our model accompanied by results are given in

Section 5. We consider them to analyse the performance of proposed algorithms in providing a loan to a company. We also analyse in detail the impact of outlier detection retained by each classifier on the final results. In Section 6, Experimental results and improvements achieved by each classifier are discussed, and at last, it reveals the studys conclusion[9].

2. Outlier detection Techniques

As the outliers decrease the model performance of the machine learning or deep learning model,By removing outliers to increase the model performance is one of the major part in machine learning based models[10].

Several Outlier detection methods exists in literature, these are-

- **Statistical-based methods** The fundamental idea of statistical-based techniques in labeling or identifying outliers depends on the relationship with the distribution model. These methods are usually classified into two main groups - the parametric and non-parametric methods.
- **Distance-based methods** The underlying principle of the distance-based detection algorithms focuses on the distance computation between observations. A point is viewed as an outlier it is far away from its nearby neighbors.For example,KNN,INNE.
- **Density-based methods** The core principle of these methods is that an outlier can be found in the low-density region, whereas inliers are in a dense neighborhood.For Example,LOF.
- **Clustering-based methods** The key idea for clustering-based techniques is the application of standard clustering techniques to detect outliers from given data. Outliers are considered as the observations that are not within or nearby any large or dense clusters. For example, CBLOF.
- **Projection based methods** Projection methods are simple and easy to apply and can highlight irrelevant values. For example, Isolation Forest,loda.

3.Literature Review:

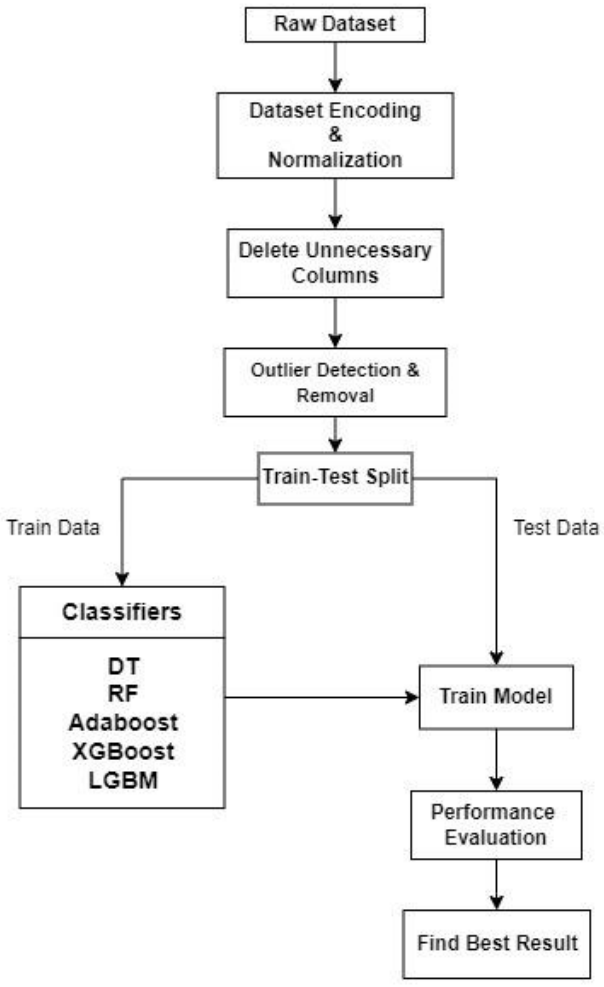
R. S. Sonawane et al.[4] In this literature work, various strategies and techniques of outlier detection are discussed and used machine learning algorithm to detect outliers.Outlier detection is an important information analysis function in its own right. Extended k means algorithm with outlier detection is proposed which focuses on controlling the number of outliers and also focuses on subspace clustering.The proposed method also identifies clusters embedded in subspaces of the original data space. For establishment and convergence an iterative procedure, we have design to optimize the objective function in proposed work. By performing numerical experiments on synthetic data and real data, thus it improves the effectiveness and efficiency in the proposed algorithm.d. For instance, the novel unsupervised approach for outlier detection, this novel unsupervised approach by using a modified clustering algorithm method is detect the outliers and removed these outliers from dataset for these purpose to improve the accuracy of an algorithm.The use of proposed algorithm reduces run time and removes outliers.

HONGZHI WANG et al.[5]] In this literature work, various strategies and techniques of outlier detection are discussed and used machine learning algorithm to detect outliers.Researchers continue to design

robust schemes to provide solutions to detect outliers efficiently.In this survey, we present a comprehensive and organized review of the progress of outlier detection methods from 2000 to 2019. First, we offer the fundamental concepts of outlier detection and then categorize them into different techniques from diverse outlier detection techniques, such as distance-,clustering-, density-, ensemble-, and learning-based methods. The open research issues and challenges at the end will provide researchers with a clear path for the future of outlier detection methods.In this paper, we have provided a comprehensive survey in a structured manner that reviews state-of-the-art methods of detecting outliers by grouping them into different cate-gories.In our discussion section, we discussed their most significant advantages,drawbacks, and challenges.For clustering techniques, since they are generally not considered to be designed explicitly for outlier detec-tion, ensemble techniques, which combine the results from dissimilar models to produce a more robust model, will create a much better result.Therefore, ensemble outlier detection, which shows great potential in enhancing outlier detec- tion algorithms, can be another worthy future research direction[6].

4.Methodology:

Any learning methods employed in this domain are data-driven and computational-based, and they rely less on assumptions about the data, particularly the distribution. While they are seen to be more resilient and better at dealing with complicated non-linear interactions, they are also thought to be difficult to comprehend. Machine learning provides the capacity to find significant patterns in data, and it has become a standard tool for practically any activity requiring the extraction of relevant data from large datasets. Several stages must be completed in order to construct a machine learning model. Following the selection of the problem, appropriate datasets must be gathered from various sources. Following that, datasets must be pre-processed by means of filtering out missing data, and/or encoding, and/or normalizing to improve interpretation. To train the model and validate it using the dataset, a train test split must be performed[7].



Depending on the kind of problem, the appropriate machine learning model must be selected. Following training and testing, performance measures must be used to assess the degree of performance. For credit risk analysis, machine learning methods such as Decision Tree, Random Forest, Adaboost, Xgboost, and LGBM have been addressed in the experimental setup. 80: 20 train test split

has been considered. For validation, 10-fold and 5-fold cross validation is been employed for better insight. Finally, accuracy,precision as well as Roc-AUC score have been recorded. The above figure depicts the framework of our model[8].

Result & Discussion:

In this study, we have utilized two benchmark datasets namely Australian, and German credit data which is the most frequently utilized datasets by the researchers. 5 tree based machine learning classifiers have been utilized, among them one was standalone and others were ensemble classifiers.Thirteen Outlier Detection methods were employed in this study. Table-1, 2, 3, 4, 5 and 6 shows models with highest accuracy, accuracy with 10-fold CV and RoC-AUC Score for Australian and German datasets respectively. Significantly, the performance of hybrid or ensemble classifiers was exceptionally better than other single classifiers. To establish this hypothesis, five tree based classifiers have been considered in this research. Among this, one standalone classifier – DT along with four ensemble classifiers – RF, Adaboost, Xgboost, and LightGBM have been tested in terms of accuracy and RoC-AUC score for two benchmark datasets – German and Australian credit data. As per the experimental result, Random Forest outperformed all other classifiers in terms of accuracy (92.50%), and 10-fold(91.16%)as well as RoC-AUC score (0.984) for Australian dataset.For the second German dataset, LGBM and RF ranked on top for accuracy (82.22%) , RF ranked on top for 10 fold (78.64%), whereas LGBM achieved best RoC-AUC score (0.85). It's been observed that the performances of all the ensemble classifiers are very much significant in contrast to standalone classifiers for both datasets.

Table:1 Accuracy for Australian Credit Data

Outlier Method	Classifier				
	DT	RF	Ada_boost	Grad_boost	LGBM
KNN	0.8833	**0.925**	0.8833	0.9	0.9
ABOD	0.8644	0.8729	0.8644	0.8814	0.8898
ECOD	0.8305	0.9153	0.8136	0.8898	0.8814
COPOD	0.8051	0.8814	0.8305	0.8729	0.8644
CBLOF	0.8475	0.8729	0.839	0.8898	0.8729
FB	**0.8917**	0.9	**0.9083**	**0.9083**	**0.9167**
IFOREST	0.8305	0.8814	0.8644	0.9068	0.8898
LOF	0.8571	0.8908	0.8571	0.8908	0.8992
INNE	0.8051	0.8729	0.8305	0.839	0.839
LODA	0.7712	0.8729	0.8475	0.8814	0.8644
SUOD	0.8235	0.8739	0.8655	0.8739	0.8908
DeepSVDD	0.7881	0.8136	0.839	0.8644	0.8475
ALAD	0.7881	0.8559	0.839	0.8475	0.8644
Best OD technique	FB	KNN	FB	FB	FB

Table:2 10 fold for Australian Credit Data					
Outlier Method	Classifier				
	DT	RF	Ada_boost	Grad_boost	LGBM
KNN	0.8341	**0.9116**	0.8266	0.8758	0.8783
ABOD	0.8003	0.8827	0.8271	0.8525	0.8584
ECOD	0.7931	0.8539	0.7911	0.8219	0.8181
COPOD	0.8096	0.85	0.77	0.8081	0.8136
CBLOF	0.7753	0.8692	0.839	0.8198	0.8301
FB	**0.8766**	0.91	**0.8916**	**0.8983**	**0.8975**
IFOREST	0.8086	0.877	0.8137	0.8652	0.8576
LOF	0.8061	0.8617	0.8035	0.8412	0.8439
INNE	0.7796	0.8079	0.7506	0.8078	0.8221
LODA	0.7399	0.8295	0.7548	0.7876	0.8031
SUOD	0.8375	0.886	0.7696	0.8473	0.8558
DeepSVDD	0.7732	0.8309	0.774	0.8225	0.8223
ALAD	0.762	0.8218	0.7467	0.839	0.8315
Best OD technique	FB	KNN	FB	FB	FB

Table:3 Roc AUC for Australian Credit Data					
Outlier Method	Classifier				
	DT	RF	Ada_boost	Grad_boost	LGBM
KNN	0.8857	0.9814	0.9694	0.97	0.9714
ABOD	0.8595	0.9613	0.9577	0.9648	0.9607
ECOD	0.8331	0.9339	0.9159	0.9579	0.9509
COPOD	0.7945	0.9022	0.8616	0.8994	0.8994
CBLOF	0.8544	0.9113	0.9008	0.937	0.94
FB	**0.9027**	**0.984**	**0.9696**	**0.9765**	**0.9754**
IFOREST	0.8454	0.9382	0.9347	0.9399	0.9381
LOF	0.858	0.9457	0.9216	0.9443	0.9454
INNE	0.8095	0.9025	0.8863	0.9056	0.8994
LODA	0.7648	0.9431	0.9195	0.9475	0.9475
SUOD	0.8145	0.9492	0.9327	0.9606	0.9537
DeepSVDD	0.7755	0.8836	0.9136	0.9104	0.9113
ALAD	0.7865	0.9411	0.9142	0.9391	0.9287
Best OD technique	FB	FB	FB	FB	FB

Table:4 Accuracy for German Credit Data					
Outlier Method	Classifier				
	DT	RF	Ada_boost	Grad_boost	LGBM
KNN	0.6919	0.7838	0.7838	0.7459	0.773
ABOD	**0.7529**	0.7931	**0.8103**	0.8046	0.7931
ECOD	0.7444	0.8222	0.7944	**0.8167**	**0.8222**
COPOD	0.6611	0.7667	0.7611	0.7611	0.7833
CBLOF	0.7	0.7389	0.7611	0.7944	0.7889
FB	0.6667	0.7978	0.7705	0.7923	0.7596
IFOREST	0.7222	0.7944	0.7667	0.7889	0.7944
LOF	0.6813	0.7802	0.7692	0.7582	0.7637
INNE	0.6722	0.7944	0.7444	0.8111	0.8056
LODA	0.6722	0.7611	0.7444	0.7333	0.7611
SUOD	0.663	0.7735	0.7348	0.7569	0.7624
DeepSVDD	0.7167	0.7611	0.7667	0.7556	0.7222
ALAD	0.6444	0.7833	0.7778	0.7556	0.7611
Best OD technique	ABOD	ECOD	ABOD	ECOD	ECOD

Table:5 10 fold for German Credit Data					
Outlier Method	Classifier				
	DT	RF	Ada_boost	Grad_boost	LGBM
KNN	**0.755**	0.7771	0.756	0.7717	0.78
ABOD	0.7508	0.7645	0.7218	0.7695	0.7655
ECOD	0.7	0.7711	0.7566	0.7666	0.7727
COPOD	0.6794	0.75	0.6916	0.7216	0.7333
CBLOF	0.6911	0.7477	0.7183	0.7527	0.7522
FB	0.719	**0.7864**	0.7571	0.7645	0.7745
IFOREST	0.6722	0.7677	0.72	0.7455	0.7405
LOF	0.6622	0.7442	0.7042	0.7247	0.7417
INNE	0.6627	0.7488	0.725	0.7338	0.7488
LODA	0.7055	0.7833	0.735	0.7605	0.7672
SUOD	0.6675	0.7541	0.7015	0.7233	0.7072
DeepSVDD	0.6688	0.7222	0.72	0.7105	0.7116
ALAD	0.7283	0.7811	**0.7722**	0.7744	**0.7816**
Best OD technique	KNN	FB	ALAD	ALAD	ALAD

Table:6 Roc AUC for German Credit Data	
	Classifier

Outlier Method	DT	RF	Ada_boost	Grad_boost	LGBM
KNN	0.6572	0.8399	**0.8324**	0.7794	0.8053
ABOD	**0.7033**	0.8415	0.8232	**0.8346**	0.85
ECOD	0.6894	0.8088	0.7645	0.8102	0.8013
COPOD	0.5833	0.8242	0.7586	0.7785	0.776
CBLOF	0.6456	0.8251	0.8149	0.7963	0.7985
FB	0.5992	0.8018	0.7492	0.788	0.7607
IFOREST	0.6547	0.7967	0.7648	0.7936	0.7645
LOF	0.6097	0.7814	0.7719	0.7703	0.7717
INNE	0.5704	0.8463	0.8205	0.8334	0.8356
LODA	0.6315	0.7844	0.7547	0.7466	0.7372
SUOD	0.5985	0.7542	0.7006	0.7549	0.736
DeepSVDD	0.6376	0.7614	0.7611	0.771	0.7795
ALAD	0.5827	0.819	0.7926	0.7618	0.7821
Best OD technique	ABOD	INNE	KNN	ABOD	ABOD

Conclusion:

The subject of credit risk assessment using statistical, machine learning and heuristic methods were discussed in this article. The goal of this research is to analyze and examine the most recent machine learning algorithms and other techniques for credit risk analysis, as well as classify them based on their performance. As per the study shows that MLPs are typically better than alternative approaches, although the model of choice is dependent on the dataset provided. In general, credit risk analysis is categorized as classification problem but in few studies, it was observed that level of risk also concluded to make it a regression problem. In this work, 46 different datasets have been observed but Australian, German and Japanese dataset were the most utilized. Data pre-processing was carried out in different stages.Various performance methods including Accuracy, Precision, FP-Rate, Recall, f-Measure, MCC etc. were used to access the models performance. Among all the performance metrics- the accuracy, of a machine learning model was utilized in most of the studies and when compared with the other models, it may not be as much to get noticeable but still, the margin of significance is quite influential for the financial institutes. Performance of a model was also increased by reducing the Type-I and Type-II errors. According to the findings, the accuracy of the machine learning technique in dealing with financial issues is superior to traditional statistical methods, particularly when dealing with nonlinear patterns. In most of the cases, Hybrid or Ensemble algorithms have shown to beat conventional algorithms in terms of performance.In most of the cases the models with high performance were trained and tested utilizing very limited datasets. Although some of these methods are highly accurate, further research is needed to determine the best parameters and techniques for achieving better results in a transparent manner.

Finally, the article conducts an analytical study on the cited publications in order to draw conclusions and determines the main tendencies of future research. The experimental setup was carried out for two benchmark datasets. Since most of the earlier research works utilized accuracy as base performance metric, the experiment incorporates accuracy as well as RoC-AUC score as performance metric for comparison. Its been found that ensemble classifiers performed better than the standalone classifiers for both the datasets. Its been validated with the prior research also. Although the performance of all the classifiers were very much significant but scope of improvement still exists. With proper data pre-

processing, Outlier detection and suitable classifier based on the available dataset can significantly improve the performance as well as reliability of the model.

Reference:

[1] E. L. Paula, M. Ladeira, R. N. Carvalho, and T. Marzagáo, ‘‘Deep learning anomaly detection as support fraud investigation in Brazilian exports and anti-money laundering’’ in Proc. 15th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA), Anaheim, CA, USA, Oct. 2016, pp. 954–960.

[2] U. Porwal and S. Mukund, ‘‘Credit card fraud detection in e-commerce: An outlier detection approach,’’ 2018, arXiv:1811.02196. [Online]. Available: https://arxiv.org/abs/1811.02196

[3] K. Alrawashdeh and C. Purdy, ‘‘Toward an online anomaly intrusion detection system based on deep learning,’’ in Proc. 15th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA), Anaheim, CA, USA, Dec. 2016, pp. 195–200.

[4] G. Gebremeskel, C. Yi, Z. He, and D. Haile, ‘‘Combined data mining techniques based patient data outlier detection for healthcare safety,’’ Int. J. Intell. Comput. Cybern., vol. 9, no. 1, pp. 42–68, 2016.

[5] V. J. Hodge and J. Austin, ‘‘A survey of outlier detection methodologies,’’ Artif. Intell. Rev., vol. 22, no. 2, pp. 85–126, 2004.

[6] C. C. Aggarwal and P. S. Yu, ‘‘An effective and efficient algorithm for high-dimensional outlier detection,’’ Int. J. Very Large Data Bases, vol. 14, no. 2, pp. 211–221, 2005.

[7] F. Angiulli, S. Basta, and C. Pizzuti, ‘‘Distance-based detection and prediction of outliers,’’ IEEE Trans. Knowl. Data Eng., vol. 18, no. 2, pp. 145–160, Feb. 2006.

[8] M. Breunig, H. Kriegel, R. T. Ng, and J. Sander, ‘‘LOF: Identifying density-based local outliers,’’ ACM SIGMOD Rec., vol. 29, no. 2, pp. 93–104, 2000.

[9] F. Cao, M. Ester, W. Qian, and A. Zhou, ‘‘Density-based clustering over an evolving data stream with noise,’’ in Proc. SIAM Conf. Data Mining, Apr. 2006, pp. 328–339.

[10] Z. Zheng, H. Y. Jeong, T. Huang, and J. Shu, ‘‘KDE based outlier detection on distributed data streams in multimedia network,’’ Multimedia Tools Appl., vol. 76, no. 17, pp. 18027–18045, Sep. 2017