Comparative analysis of ensemble methods for outlier detection in real estate*

Viktor Khomyshyn[1,*,†], Oleh Pastukh[1,†], Vasyl Yatsyshyn[1,†], Nataliya Zagorodna[1,†] and Ihor Baran[1,†]

[1] *Ternopil Ivan Puluj National Technical University, Ruska str., 56, 46001, Ternopil, Ukraine*

Abstract

The paper evaluates the effectiveness of ensemble methods for outlier detection on a real dataset. The aim of the paper is to implement improved ML approaches in the real estate industry in order to optimize market data analysis. The study covered the real estate market of the city of Ternopil (Ukraine), in particular the sale of apartments and rooms. The prepared dataset contained 760 real estate objects with 12 features. For each real estate object, an anomaly label was assigned by an expert based on its characteristics. Algorithm testing was carried out using two methods of encoding categorical features - Label Encoder and One-Hot Encoder. Data set standardization was carried out using the RobustScaler scaler resistant to outliers. The following ensemble methods were used during the experiments: INNE, LODA, IForest and Feature Bagging. The results of the work were evaluated by three indicators: AUC-ROC, Precision @ Rank n and algorithm execution time. They allowed us to assess the accuracy and efficiency of ensemble algorithms and determine their suitability for real-world problems of anomaly detection in real estate data. The visualization of the results of the algorithms was carried out using PCA and t-SNE dimensionality reduction methods and showed how well each model detects normal and anomalous objects. This study is a sequential stage in building a multi-agent system for working with real estate based on modern innovative machine learning algorithms.

Keywords

machine learning, ensemble methods, outlier detection, anomaly detection, real estate, CEUR-WS [1]

1. Introduction

Numerous studies in the field of machine learning show that two approaches currently dominate it: deep learning and ensemble learning. Deep learning is known for its ability to automatically extract useful patterns from raw data, such as images or audio. Ensemble learning, in turn, has shown high efficiency in building models on structured data that already contain meaningful features. It imitates the natural human habit of referring to several opinions before making an important decision. The basic principle of ensemble learning is to weigh the results of individual models and combine them to achieve a more accurate answer than the one that a single model could obtain [1].

Outlier detection is a key technique in building effective machine learning models for real estate valuation. The accuracy of such models is directly related to the quality of the data [2]. It is important for investors, sellers, and buyers that the price of the object most accurately reflects the current market situation. The non-standard nature of outliers often significantly worsens the forecasting results. Therefore, there are and are used many algorithms that detect outliers based on various criteria: distance, density, angle, etc. More modern approaches involve the use of ensemble machine learning methods for such tasks. They allow to increase the accuracy of outlier detection.

* *CITI'2025: 3rd International Workshop on Computer Information Technologies in Industry 4.0, June 11–12, 2025, Ternopil, Ukraine*

[1,*] Corresponding author.

[†] These authors contributed equally.

homyshyn@gmail.com (V. Khomyshyn); oleg.pastuh@gmail.com (O. Pastukh); vyatcyshyn@gmail.com (V. Yatsyshyn); zagorodna.n@gmail.com (N. Zagorodna); ihor.remm@gmail.com (I. Baran)

0000-0003-4369-501X (V. Khomyshyn); 0000-0002-0080-7053 (O. Pastukh); 0000-0002-5517-6359 (V. Yatsyshyn); 0000-0002-1808-835X (N. Zagorodna); 0000-0002-8153-2476 (I. Baran)

The purpose of our study is to analyze the effectiveness of ensemble machine learning methods for detecting outliers in real estate data, as well as to develop an optimized approach that combines the advantages of different algorithms to increase the accuracy and stability of detection. This will create a theoretical and practical basis for automating data cleaning in real estate data processing systems.

2. Analysis of recent studies and publications

Ensemble approaches in anomaly analysis are much less studied than in classical data mining tasks. This is due to the fact that assessing the quality of individual ensemble components when detecting anomalies is more difficult. This process is also often affected by the lack of accurate labeled data. As a result, decisions about combining and selecting ensembles must be made using intermediate results of the algorithm, rather than specific quality metrics on validation datasets. However, these intermediate results can sometimes be inaccurate estimates of the degree of anomaly. When decisions about selecting and combining ensembles are made in an uncontrolled manner, the probability of making erroneous decisions is much higher. However, although the process of outlier detection is a difficult task for ensemble analysis, these difficulties are not insurmountable.

In [3] different ways of classifying outlier analysis problems are proposed, such as independent or sequential ensembles, and data-driven or model-driven ensembles. The impact of different combination functions and their relationship to different types of ensembles is also discussed. The choice of the right combination function is important, although in the general case it may depend on the structure of the ensemble. In addition, many modern outlier detection methods are compared with different types of ensembles and the possibility of adapting ensemble methods from other data mining tasks to outlier detection tasks is considered.

The paper [4] is devoted to the study of the theoretical foundations of ensemble outlier analysis. Despite the significant differences between the classification and outlier analysis problems, the paper shows that the theoretical foundations of these two problems are actually very similar in terms of the trade-off between bias and variance. The influence of the combination function is discussed and specific trade-offs between the averaging and maximization functions are considered. The authors propose several robust variations of feature bagging methods and two new combination methods based on variance-bias theory.

In [5] a new outlier detection algorithm is proposed – Ensemble Outlier Detection Method Based on Information Entropy – Weighted Subspaces for High-Dimensional Data (EOEH). It first performs a random secondary subsampling of the data, and the detectors are run on different small-scale subsamplings to obtain a variety of detection results. Then, these results are aggregated to reduce the global variance and improve the robustness of the algorithm. Next, information entropy is used to construct a weighting method for dimension spaces, which allows determining influential factors in different multidimensional spaces. This method generates weighted subspaces and dimensions for data objects, reducing the influence of noise generated by high-dimensional data, and improving the detection performance of high-dimensional data. Finally, a new detector – High-Precision Local Outlier Factor (HPLOF) is proposed, which enhances the differentiation between normal data and outliers, thereby improving the detection performance of the algorithm. Experiments using simulated data and UCI datasets have confirmed the feasibility of the proposed algorithm. Compared to popular outlier detection algorithms, EOEH improves detection quality by an average of 6% and runs 20% faster.

In [6], an approach is proposed for automatic optimization of outlier detection ensembles using a limited number of outlier samples (from 1 to 10% of the available outliers). Optimized outlier detection ensembles consist of outlier detection algorithms that provide an outlier estimate and use adjustable parameters. Automatic optimization determines parameter values that improve the discrimination between normal data and outliers. This increases the efficiency of outlier detection.

The study [7] describes a universal framework for outlier detection based on bi-sampling – Bi-Sampling Outlier Detection (BSOD) and provides theoretically justified optimal ratios for row and

column sampling. The BSOD method demonstrates diversity in ensemble construction. As a base, the LOF (Local Outlier Factor) algorithm was used – a classic method for detecting local outliers, which calculates the local density by averaging the density of its neighbors. By implementing LOF within BSOD, a BI-LOF model was built, on which experiments were conducted using 30 synthetic and 17 real datasets. The BI-LOF model was also tested for outlier detection in images. Overall, the experimental results showed high quality and stability of the BI-LOF method.

In unsupervised ensemble outlier detection problems, the lack of labeled outliers it difficult to combine baseline outlier detectors. In particular, existing parallel ensembles do not have a reliable mechanism for selecting effective baseline detectors, which negatively affects the accuracy and stability during model combination. In [8], an approach called Locally Selective Combination in Parallel Outlier Ensembles (LSCP) is proposed, which solves this problem by determining a local region around a test instance through the consensus of its nearest neighbors in randomly selected feature subspaces. Unlike traditional combination methods, LSCP determines the best baseline detectors for each test instance relative to its local environment. The most effective baseline detectors in this local region are selected and combined to obtain the final result. To evaluate the effectiveness, the proposed approach is tested on 20 real-world datasets and demonstrates superiority over baseline algorithms. Four variants of the LSCP model are compared with seven common parallel approaches. The ensemble approach of LSCP AOM shows the best results, achieving higher scores on 13 out of 20 datasets by the ROC-AUC metric and on 14 out of 20 by the mAP metric (average accuracy). The paper also provides theoretical justifications in the context of the trade-off between bias and variance and visualizations that provide a comprehensive understanding of the LSCP technique. Since the LSCP approach demonstrates the promise of using local data, the authors of the study propose to extend this technique by using heterogeneous basic detectors.

A similar approach to the previous one, which also works in the absence of labeled data (without a teacher), is described in [9]. The authors propose a new improved framework for combining anomaly detectors – Dynamic Combination of Detector Scores for Outlier Ensembles (DCSO). Unlike traditional ensemble methods that statically combine detectors, DCSO dynamically determines the most effective baseline detectors for each test case by evaluating their effectiveness in a certain local region. The DCSO algorithm first outlines the local region of the test case by its k nearest neighbors, and then identifies the most effective baseline detectors in this local region. Given the fact that local data connections are crucial for combining anomaly scores, DCSO ranks the quality of individual baseline detectors by their similarity to pseudo-reference data in the local region. To increase the stability of the model and reduce the risks of using a single detector, the study also proposes different variants of ensembles based on DCSO. The performance of DCSO is verified through statistical evaluations on ten real-world datasets. The results confirm that this approach outperforms traditional static combination methods in anomaly detection. In addition to significantly improving detection quality, DCSO is also computationally robust, being compatible with any baseline detector (e.g. LOF or k-NN) and open in demonstrating how outlier estimates are generated for each test instance, given the chosen baseline detector.

In [10], three ensemble methods were compared – LSCP (Locally Selective Combination in Parallel Outlier Ensembles), iForest (Isolation Forest) and FB (Feature Bagging) with eleven widely used anomaly detection methods, including: ABOD, CBLOF, HBOS, KNN, LOF, MCD, OCSVM, PCA, SOS, SOD, AveKNN. The initial data set contained 400 observations with an outlier rate of 0.25. At the first stage, 80 instances were randomly selected from the set, which were divided into approximately equal parts. Two of these parts were used as a training set, and the third as a test set (cross-validation). The results of the three tests were averaged. The three most effective algorithms that showed the best ROC curve values were combined to create a more accurate final model, which was subsequently applied to the full initial data set. The presented approach for forming an ensemble has shown its effectiveness and reliability.

In [11], an ensemble approach called Cumulative Agreement Rates Ensemble (CARE) is proposed, which aims to achieve low error by reducing variance and bias. This method considers

anomaly detection as a binary classification problem of unlabeled data and uses a two-phase aggregation of intermediate results at each iteration to obtain the final result. The two main components of CARE are its parallel and sequential components. The former help to reduce variance by weighted combination of several baseline detectors, and the latter are designed to reduce both bias and variance by using Filtered Variable Probability Sampling (FVPS) and cumulative aggregation. The first stage sequentially eliminates outliers from the original data set to build a better data model on which to estimate the outlier, and the second stage combines results from individual baseline detectors and between iterations. The proposed method was tested on 16 real-world datasets, mostly from the UCI machine learning repository, and showed significant improvement over baseline methods and state-of-the-art ensemble approaches to anomaly detection, demonstrating either superior or close results.

The results of the presented numerous studies show that ensemble technologies are actively developing, providing a significant improvement in the quality of outlier detection, and ensemble analysis is a promising area of research, in particular in anomaly detection problems.

3. Research methodology

3.1. Dataset description

In this study, the city of Ternopil (Ukraine) was selected for the analysis of real estate information. To obtain and maintain the relevance of the dataset, developed software was used in the form of a Windows application, a Microsoft SQL Server database, and a website [12]. Data collection was carried out by parsing the Internet pages of popular real estate portals in the region. Characteristics missing from the advertisement (sources of information) were clarified with the owner.

In general, the study used data that was in the database as of the beginning of 2025. The dataset contained a description of 760 offers for the sale of apartments and rooms in dormitories in the city of Ternopil. To form a dataset that was later used for machine learning, the software was supplemented with a function to export the necessary fields to CSV format. A total of 14 columns were exported from the program (Table 1). The first column is the object ID (unique identifier), which we will use for auxiliary purposes only, since it has no information value. The next 12 columns are characteristics (features) that describe the real estate object. The last column is the anomaly label (1 - anomalous object, 0 - normal object).

The anomaly label of each real estate object in the dataset was assessed by an expert based on an analysis of the object's location, its characteristics, and cost at the stage of adding the object to the database or changing its price or condition. For this purpose, a set of rules was created that allowed detecting anomalous values that differ significantly from the average indicators in the region. For example, objects with too low or high cost compared to other objects of a similar type and location were considered potential anomalies. The analysis took into account that the price of 1 sq.m. of housing is usually higher for small apartments, mainly one-room ones. In addition, atypical combinations of characteristics were taken into account, such as the high cost of apartments in older buildings, mainly from the 80s and earlier. The expert had at his disposal photographs of each object, on average from 5 to 20 pieces, and in some cases a copy of the technical passport, against which the information entered into the database was checked. The dataset prepared in this way contained 10% of the observed anomalous objects. The current state of the database is available at [13].

3.2. Research procedure

The effectiveness of ensemble outlier detection methods was tested in the Python programming language using the Spyder IDE tool included in the Anaconda software package. Data processing was performed using the pandas, numpy, sklearn, and pyod libraries, and visualization was performed using the matplotlib library. Preprocessing of the dataset included: separating the values

of the "id" and "label" fields into separate one-dimensional arrays; removing duplicate rows; removing columns with missing data.

Table 1
Dataset description

Feature ID	Feature name	Feature type	Description
0	id	integer	The ID of the object in the database
1	realty_type	text	Type of real estate
2	district	text	City district (residential area)
3	total_area	integer	Total area (m^2)
4	floor	integer	The floor on which the real estate is located
5	floors	integer	Total number of floors in the building
6	repair_state	text	State of repair
7	wall_material	text	Material of external walls
8	furniture	text	Availability of furniture
9	heating	text	Type of heating
10	build_year	text	Construction years
11	market	text	Real estate market
12	price	integer	The price of the object ($)
13	label	binary	Anomaly label of an object

The list of studied ensemble algorithms and their characteristics is given in Table 2. All of these algorithms are components of the PyOD (Python Outlier Detection) library. By default, Feature Bagging (FB) uses the LOF algorithm as the base estimator. We extended the Feature Bagging study by using the CBLOF, KDE, KNN, OCSVM, and QMCD algorithms as the base detectors instead of LOF. These algorithms showed the best individual results for outlier detection on the studied dataset [14].

For a more comprehensive evaluation, each algorithm was tested on a dataset that underwent various methods of encoding categorical features (Label Encoding and One-Hot Encoding) at the preprocessing stage, as well as with and without data scaling (RobustScaler).

At the same time, the size of the dataset, depending on the method of encoding categorical features, was as follows:

- Label Encoding – dataset of 760 observations and 12 features;
- One-Hot Encoding – dataset of 760 observations and 67 features.

The experiments were conducted on a PC with the Windows 10 Pro x64 operating system, an Intel(R) Core(TM) i3-10105F 3.70 GHz processor, and 32 GB of RAM.

Table 2
The investigated ensemble methods for anomaly detection

Algorithm	Decryption	Type	Year	Multicore	Source
INNE	Isolation-based Anomaly Detection Using Nearest-Neighbor Ensembles	Unsupervised	2018	No	[15]
LODA	Lightweight On-line Detector of Anomalies	Unsupervised	2016	No	[16]
IForest	Isolation Forest	Unsupervised	2008	Yes	[17]
FB	Feature Bagging	Unsupervised	2005	Yes	[18]

The results were evaluated using three indicators:

- AUC-ROC (Area Under the Curve – Receiver Operating Characteristic);
- P@n (Precision @ Rank n);
- algorithm execution time.

For each combination of algorithm and data set, 100 experiments were conducted. The metrics were averaged. The test results are presented in Table 3.

3.3. Preliminary assessment of results

The data in Table 3 indicate that the way categorical features are encoded and the data scaling affects the accuracy of each ensemble differently. We have summarized these results and propose a matrix for selecting the optimal characteristics of the dataset for each ensemble algorithm (Table 4). Let us explain it using the example of the IForest algorithm: the best result is achieved when using the Label Encoder, while the data scaling (yes/no) does not matter.

The quality assessment of models based on the AUC-ROC indicator is as follows [19]:

- AUC-ROC ≥ 0.9 – excellent quality;
- 0.8 ≤ AUC-ROC < 0.9 – good qualit;
- 0.7 ≤ AUC-ROC < 0.8 – acceptable (satisfactory) quality;
- 0.5< AUC-ROC < 0.7 – low quality;
- AUC-ROC = 0.5 – equivalent to random guessing.

The AUC-ROC metric is a balanced global estimate, but it is not sensitive to class imbalance, which is bad for very rare anomalies.

If it is important to focus on the top anomalies (local estimate), the P@n metric is used. It indicates what proportion of objects among the top-n detected by the system are really anomalies.

Among the considered algorithms, the following ensembles have the best accuracy indicators:

- INNE – AUC-ROC=0.863, P@n=0.582;
- LODA – AUC-ROC=0.825, P@n=0.514;
- FB + CBLOF – AUC-ROC=0.832, P@n=0.599;
- FB + KNN – AUC-ROC=0.789, P@n=0.523.

Although the FB + QMCD ensemble showed high results (AUC-ROC=0.880, P@n=0.505), we excluded it from further research due to the fact that the spread of metrics in the experiments was quite significant.

Table 3
Performance comparison of the ensemble algorithms

Algorithm	Data scaling	Categorical feature encoder					
		Label Encoder			One-Hot Encoder		
		AUC-ROC	P@n	Time, ms	AUC-ROC	P@n	Time, ms
INNE	yes	0.809	0.392	193	0.818	0.413	202
	no	0.863	0.582	193	0.863	0.581	205
LODA	yes	0.725	0.327	16	0.566	0.155	18
	no	0.825	0.514	17	0.635	0.218	18
IForest	yes	0.745	0.326	148	0.664	0.147	144
	no	0.748	0.323	146	0.663	0.142	144
FB + LOF	yes	0.787	0.329	92	0.708	0.238	49
	no	0.617	0.193	51	0.615	0.189	46
FB + CBLOF	yes	0.838	0.452	51	0.835	0.468	60
	no	0.832	0.599	48	0.832	0.597	55
FB + KDE	yes	0.814	0.449	396	0.845	0.461	813
	no	0.713	0.249	143	0.713	0.254	295
FB + KNN	yes	0.847	0.447	68	0.830	0.455	55
	no	0.789	0.523	38	0.788	0.521	54
FB + OCSVM	yes	0.804	0.428	375	0.823	0.401	430
	no	0.632	0.217	905	0.652	0.227	956
FB + QMCD	yes	0.752	0.294	33	0.880	0.505	85
	no	0.752	0.295	34	0.880	0.504	85

Table 4
Matrix for selecting optimal data set characteristics

Dataset characteristics	Label Encoder	Label Encoder OR One-Hot Encoder	One-Hot Encoder
Scaling (RobustScaler)	FB + LOF	FB + OCSVM FB + KDE	–
Scaling OR No scaling	IForest	FB + KNN	FB + QMCD
No scaling	LODA	INNE FB + CBLOF	–

3.4. Optimization of algorithms

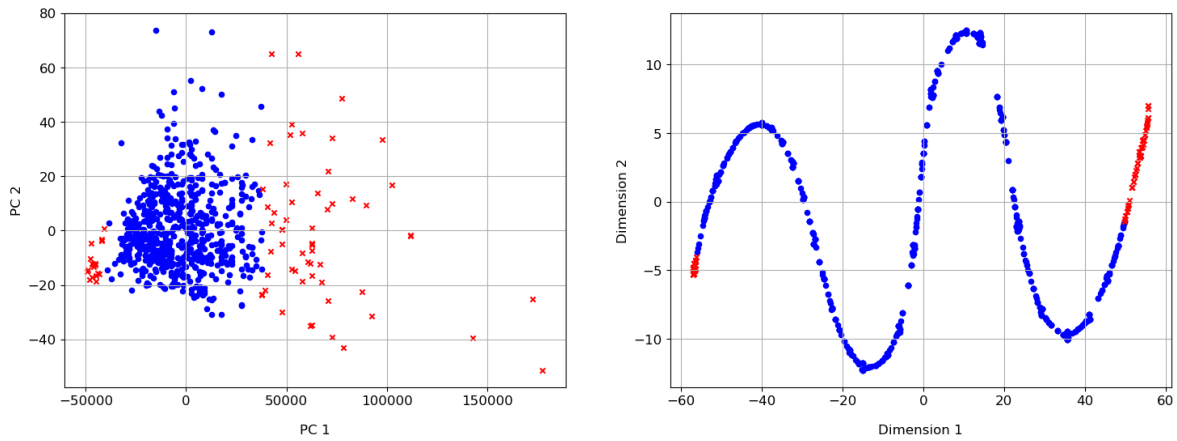
The next step of the study was to optimize the hyperparameters of the selected ensembles in order to obtain the best result that each ensemble algorithm can provide. Currently, there are such popular approaches to solving this problem: grid search, random search and Bayesian optimization. There are several Python libraries for finding optimal hyperparameter settings, including Optuna, Ray Tune and Hyperopt. These libraries simplify and automate the search process and also have the ability to scale in several computing environments to speed up the result [20].

In our study, the Optuna framework was used. This is an open source library that provides efficient optimization by applying modern hyperparameter selection algorithms and effective pruning of unpromising trials. Optuna also integrates with MLflow (a standard format for packaging machine learning models) for tracking and monitoring models and trials [21].

The AUC-ROC metric was chosen as the objective function for optimization. The goal of the optimization was to maximize the value of the objective function. The hyperparameter search space was specified individually for each ensemble (INNE, LODA) or the base algorithm (CBLOF, KNN). The results of hyperparameter optimization for each ensemble are presented in Tables 5-8. Figures 1-4 present visualizations of the results of the optimized outlier detection algorithms using PCA [22] and t-SNE [23] dimensionality reduction methods. Visual evaluation demonstrates high accuracy of anomaly detection by each algorithm.

Table 5
Results of hyperparameter optimization of the INNE algorithm

Settings	Hyperparameter INNE		AUC-ROC	P@n	Time, ms
	Name	Value			
Default	n_estimators	200	0.863	0.582	193
	max_samples	"auto"			
Optimized	n_estimators	155	0.883	0.615	142
	max_samples	3			

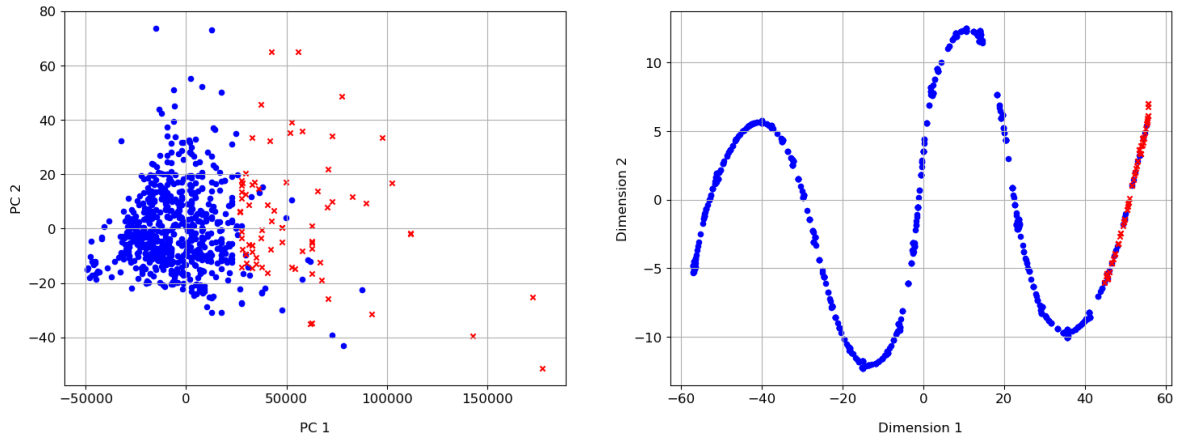


a) PCA transformation b) t-SNE transformation

Figure 1: Visualization of the optimized INNE algorithm performance results.

Table 6
Results of hyperparameter optimization of the LODA algorithm

Settings	Hyperparameter LODA		AUC-ROC	P@n	Time, ms
	Name	Value			
Default	n_bins	10	0.825	0.514	17
	n_random_cuts	100			
Optimized	n_bins	3	0.907	0.564	101
	n_random_cuts	690			

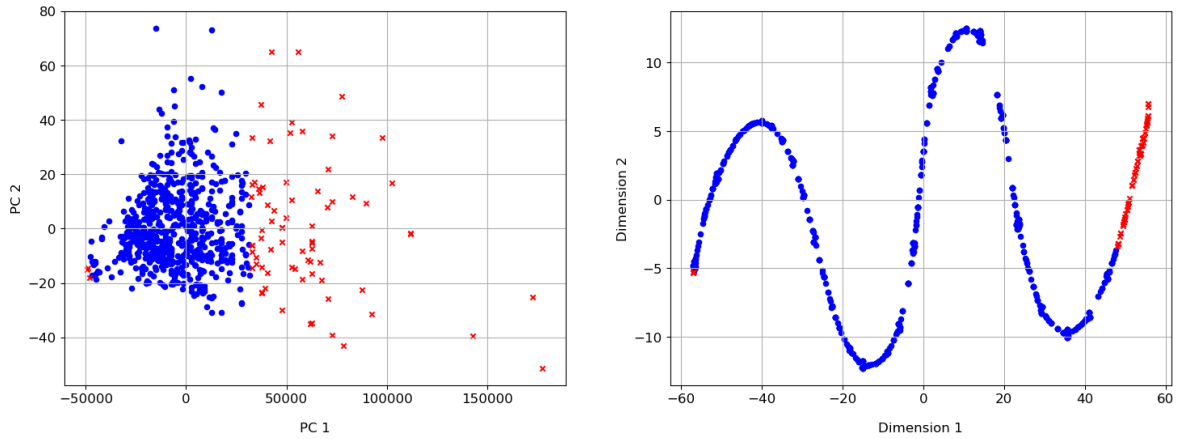


a) PCA transformation b) t-SNE transformation

Figure 2: Visualization of the optimized LODA algorithm performance results.

Table 7
The results of the Feature Bagging algorithm paired with the optimized CBLOF algorithm

Settings	Hyperparameter CBLOF		AUC-ROC	P@n	Time, ms
	Name	Value			
Default	n_clusters	8	0.832	0.599	48
	alpha	0.90			
Optimized	n_clusters	25	0.886	0.660	83
	alpha	0.65			

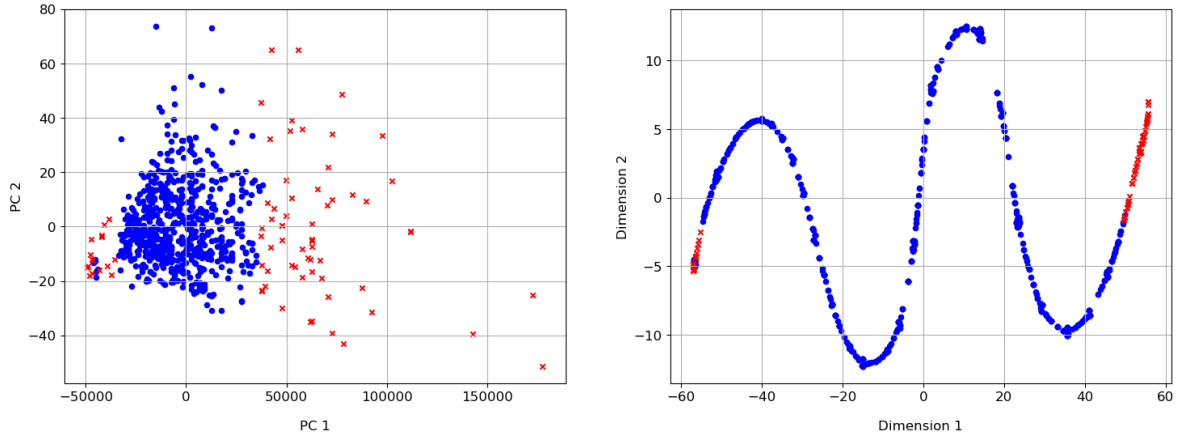


a) PCA transformation b) t-SNE transformation

Figure 3: Visualization of the optimized FB + CBLOF algorithm performance results.

Table 8
The results of the Feature Bagging algorithm paired with the optimized KNN algorithm

Settings	Hyperparameter KNN		AUC-ROC	P@n	Time, ms
	Name	Value			
Default	n_neighbors	5	0.789	0.523	38
	method	largest			
Optimized	n_neighbors	31	0.889	0.632	54
	method	median			



a) PCA transformation b) t-SNE transformation

Figure 4: Visualization of the optimized FB + KNN algorithm performance results.

Figures 5 and 6 respectively present generalized diagrams of the AUC-ROC and P@n metrics of the studied algorithms when detecting anomalies with default settings and after hyperparameter optimization (opt index).

Optimization of the settings provided an increase in the AUC-ROC and P@n metrics. At the same time, the variability (dispersion) of the forecasts decreased and the number of outlier forecasts decreased.

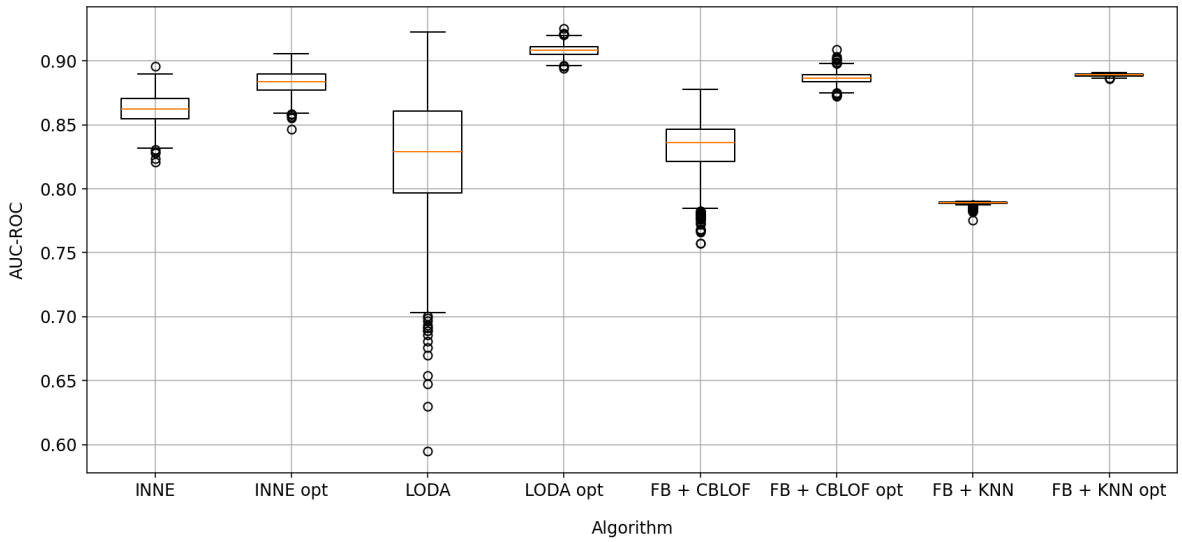


Figure 5: Boxplot of AUC-ROC for the studied algorithms.

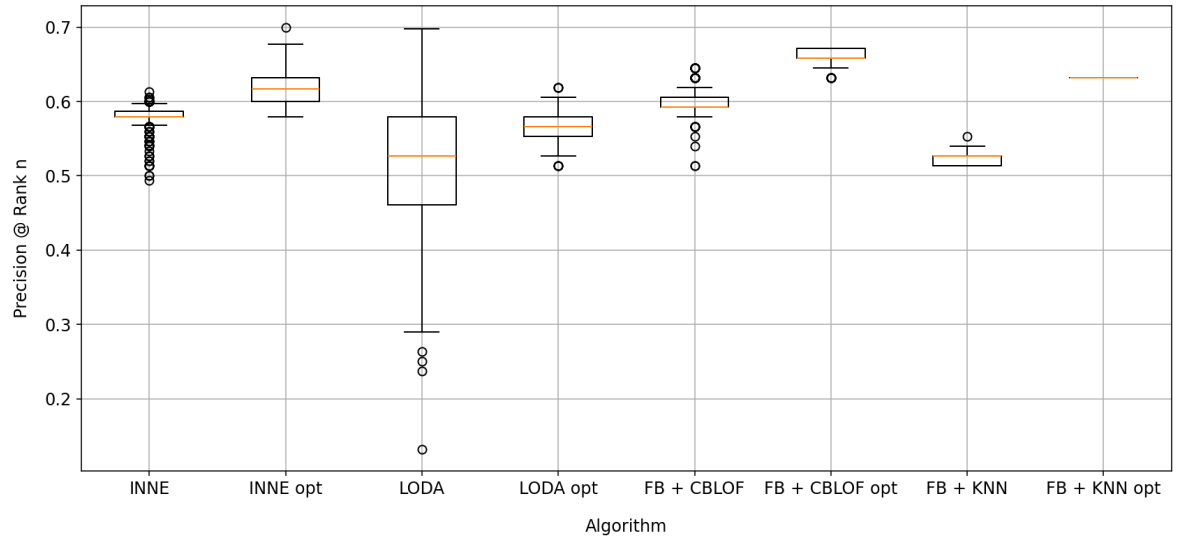


Figure 6: Boxplot of P@n for the studied algorithms.

Conclusions

The conducted comparative study showed high accuracy of ensemble machine learning algorithms in outlier detection tasks in real estate data. The best among all considered options is the combination of the Feature Bagging method and the optimized basic detector CBLOF. At the same time, high metrics (AUC-ROC=0.886, P@n=0.660) and optimal running time (Time=83 ms) are provided. Slightly lower, but quite close indicators are given by the combination of the Feature Bagging method in a pair with the optimized basic detector KNN (AUC-ROC=0.889, P@n=0.632), but this pair works faster than its predecessor (Time=54 ms). The optimized INNE algorithm is also

not much inferior in metrics (AUC-ROC=0.883, P@n=0.615), but is the slowest (Time=142 ms). Regarding the LODA algorithm, for which hyperparameter optimization was also performed, at the highest metric AUC-ROC=0.907, the P@n = 0.564 indicator is the lowest. This is visually confirmed by its graphs, which show that the separation of objects and normal and abnormal is less clearly expressed in comparison with other described algorithms.

Further research on the topic of this work can be aimed at analyzing the effectiveness of other ensemble techniques and algorithms for outlier detection, including their combinations and hyperparameter optimization.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] L. Rokach, Ensemble Learning: Pattern Classification Using Ensemble Methods, 2nd. ed., World Scientific (2019). doi:10.1142/11325.

[2] O. Pastukh, V. Khomyshyn, Using ensemble methods of machine learning to predict real estate prices, in: ITTAP'2024: 4th International Workshop on Information Technologies: Theoretical and Applied Problems, 2024, pp. 438–447. URL: https://ceur-ws.org/Vol-3896/paper26.pdf.

[3] C. Aggarwal, Outlier Ensembles, ACM SIGKDD Explorations Newsletter 14(2) (2012) 49–58. doi:10.1145/2481244.2481252.

[4] C. Aggarwal, S. Sathe, Theoretical Foundations and Algorithms for Outlier Ensembles, ACM SIGKDD Explorations Newsletter 17(1) (2015) 24–47. doi:10.1145/2830544.2830549.

[5] Z. Li, L. Zhang, An Ensemble Outlier Detection Method Based on Information Entropy-Weighted Subspaces for High-Dimensional, Data. Entropy 25(8) (2023). doi:10.3390/e25081185.

[6] N. Reunanen, T. Räty, T. Lintonen, Automatic optimization of outlier detection ensembles using a limited number of outlier examples, International Journal of Data Science and Analytics 10 (2020) 377–394. doi:10.1007/s41060-020-00222-4.

[7] H. Liu, Y. Zhang, B. Deng, Y. Fu, Outlier Detection via Sampling Ensemble, in: International Conference on Big Data (Big Data), 2016. doi:10.1109/BigData.2016.7840665.

[8] Y. Zhao, Z. Nasrullah, M. Hryniewicki, Z. Li, LSCP: Locally Selective Combination in Parallel Outlier Ensembles, in: Proceedings of the 2019 SIAM International Conference on Data Mining, 2019, pp. 585–593. doi:10.48550/arXiv.1812.01528.

[9] Y. Zhao, M. Hryniewicki, DCSO: Dynamic Combination of Detector Scores for Outlier Ensembles., in: ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), Outlier Detection De-constructed Workshop, 2018. doi:10.48550/arXiv.1911.10418.

[10] S. Alexandropoulos, S. Kotsiantis, V. Piperigou, M. Vrahatis, A new ensemble method for outlier identification, in: 10th International Conference on Cloud Computing, Data Science & Engineering, 2020, pp. 786–791. doi:10.1109/Confluence47617.2020.9058219.

[11] S. Rayana, W. Zhong, L. Akoglu, Sequential Ensemble Learning for Outlier Detection. A Bias-Variance Perspective, in: IEEE 16th International Conference on Data Mining (ICDM-2016), 2016, pp. 1167–1172. doi:10.1109/ICDM.2016.0154.

[12] Software for real estate agencies "MyHome", Real estate agency "Leader" Ternopil, 2015. URL: https://lider.org.ua/myhome.aspx.

[13] Sale of apartments and rooms in Ternopil and Ternopil district, Real estate agency "Leader" Ternopil, 2025. URL: https://lider.org.ua/base.aspx?t=kva.

[14] O. Pastukh, V. Khomyshyn. Efficiency research of cluster analysis methods for detecting outliers in real estate market, Herald of Khmelnytskyi National University, Technical sciences 3(1) (2025) 362–381. doi:10.31891/2307-5732-2025-351-45.

[15] T. Bandaragoda, K. Ting, D. Albrecht, F. Liu, Y. Zhu, J. Wells, Isolation-based anomaly detection using nearest-neighbor ensembles, Computational Intelligence 34(3) (2018) 968–998. doi:10.1111/coin.12156.
[16] T. Pevný, Loda: Lightweight on-line detector of anomalies, Machine Learning 102 (2016) 275–304. doi:10.1007/s10994-015-5521-0.
[17] F. Liu, K. Ting, Z. Zhou, Isolation Forest, in: Eighth IEEE International Conference on Data Mining (ICDM '08), 2008, pp. 413–422. doi:10.1109/ICDM.2008.17.
[18] A. Lazarevic, V. Kumar, Feature Bagging for Outlier Detection, in: International conference on Knowledge discovery in data mining (KDD '05), 2005, pp. 157–166. doi:10.1145/1081870.1081891.
[19] D. Hosmer, S. Lemeshow, R. Sturdivant, Applied Logistic Regression, 2nd. ed., Chapter 5 (2013) 173-182.
[20] L. Aleksina, A. Bondarchuk, Hyperparameters optimization for the machine learning, Connectivity 2 (2024) 18–22. doi:10.31673/2412-9070.2024.021822.
[21] Optuna - A hyperparameter optimization framework, Optuna, 2025. URL: https://optuna.org.
[22] K. Pearson, LIII. On lines and planes of closest fit to systems of points in space, The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science 2(11) (1901) 559–572. doi:10.1080/14786440109462720.
[23] L. Maaten, G. Hinton, Visualizing Data using t-SNE, Journal of Machine Learning Research 9(86) (2008) 2579–2605.