Comparative Performance Analysis of Machine Learning Methods for Classification Tasks

Abstract:

Machine learning methods play a crucial role in various classification tasks, and selecting the best method for a given problem is a fundamental challenge. In this study, we compared the performance of several popular machine learning methods, including XGBoost, LightGBM, CatBoost, Multi-Layer Perceptron (MLP), Logistic Regression (LR), and ID3 decision trees, across a range of evaluation metrics. The evaluation metrics included accuracy, precision, recall, F1-score, area under the ROC curve (AUC-ROC), area under the precision-recall curve (AUC-PR), log loss, Cohen's Kappa, Matthew's correlation coefficient (MCC), Fowlkes-Mallows index, and R-squared. Our results revealed that MLP demonstrated superior overall performance, with high accuracy, area under the ROC curve, and Matthew's correlation coefficient scores.

However, the optimal choice of method may depend on specific priorities and requirements, as highlighted by the strengths of other methods for particular metrics. This comparative analysis provides valuable insights for practitioners and researchers in selecting appropriate machine learning methods for classification tasks.

Methodology

XGBoost: XGBoost stands for eXtreme Gradient Boosting. It is a scalable and accurate implementation of gradient boosting machines and is known for its speed and performance in machine learning competitions.

Algorithm: The XGBoost algorithm is an optimized gradient boosting machine learning algorithm. It uses a process called boosting to create an ensemble of weak learners (typically decision trees) and gradually improves their performance by focusing on the mistakes from the previous iteration. XGBoost optimizes the overall prediction by minimizing a specific loss function and penalizing complexity using regularization techniques.

LightGBM: LightGBM (Light Gradient Boosting Machine) is a gradient boosting framework that uses tree-based learning algorithms. It is designed for efficiency and scale, making it a popular choice for large-scale machine learning problems.

Algorithm: LightGBM uses a novel algorithm based on histogram approximation to handle large amounts of data efficiently. It employs a leaf-wise tree growth strategy, which focuses on growing the tree level by level, adding nodes that yield the maximum reduction in the loss function. This approach can lead to faster training times and lower memory usage compared to other gradient boosting methods.

CatBoost: CatBoost is an open-source gradient boosting library that is specifically designed to work well with categorical features, making it suitable for a wide range of applications and datasets.

Algorithm: CatBoost employs an algorithm based on gradient boosting that focuses on handling categorical data effectively. It uses data pre-processing techniques to automatically handle categorical variables and introduces an advanced strategy for gradient boosting that aims to reduce overfitting and improve the accuracy of predictions.

Multi-Layer Perceptron (MLP): Multi-Layer Perceptron is a class of feedforward artificial neural network that consists of multiple layers of nodes, allowing it to learn and model complex relationships in the data.

Algorithm: MLP uses a network of interconnected nodes organized in layers, comprising an input layer, one or more hidden layers, and an output layer. It employs an algorithm known as backpropagation to train the network by adjusting the weights and biases to minimize the difference between the actual and predicted outputs.

Logistic Regression (LR): Logistic Regression is a statistical method used for binary classification tasks, where the output is a probability value representing the likelihood of a given sample belonging to a particular class.

Algorithm: Despite its name, logistic regression is a classification algorithm rather than a regression algorithm. It models the relationship between the independent variables and the probability of a specific outcome using the logistic function, which ensures that the predicted values are between 0 and 1. The parameters of the model are optimized using techniques such as maximum likelihood estimation.

ID3 Decision Trees: ID3 (Iterative Dichotomiser 3) is a decision tree algorithm that is used for both classification and regression tasks, aiming to create a tree-like model of decisions to predict the target variable.

Algorithm: The ID3 algorithm uses a top-down, greedy approach to split the dataset based on different attributes, aiming to create the most homogeneous subgroups with respect to the target variable. It selects the attributes that provide the most information gain or the best split, and recursively builds the tree until certain stopping criteria are met.

Classification Metrics

Accuracy: Accuracy is a measure of how many predictions made by the model are correct out of the total predictions. It is the ratio of the correctly predicted instances to the total instances. In other words, accuracy shows how often the model's predictions are correct.

Precision: Precision indicates the proportion of true positive predictions (correctly predicted positive instances) out of all the positive predictions made by the model. It measures how precise the model is when it predicts a positive outcome.

Recall: Recall, also known as sensitivity, measures the proportion of actual positive instances that were correctly identified by the model. It reflects the model's ability to identify all relevant instances, or to recall the positive cases.

F1-Score: The F1-Score is the harmonic mean of precision and recall. It provides a single metric combining both precision and recall, allowing for a balanced evaluation of the model's performance. A higher F1-Score indicates better precision and recall.

AUC-ROC (Area Under the Receiver Operating Characteristic Curve)**:** AUC-ROC quantifies the ability of the model to distinguish between classes, particularly the true positive rate against the false positive rate. It represents the probability that the model ranks a randomly chosen positive instance higher than a randomly chosen negative one.

AUC-PR (Area Under the Precision-Recall Curve): AUC-PR measures the trade-off between precision and recall for different threshold values. It provides an aggregate measure of model performance across various levels of class imbalance.

Log Loss: Log Loss measures the accuracy of a model by evaluating the uncertainty of its predictions. The lower the log loss value, the better the model performance. It quantifies the difference between the predicted and actual class probabilities.

Cohen's Kappa: Cohen's Kappa measures the agreement between observed and expected classifications, taking into account the possibility of the agreement occurring by chance. It is particularly useful when dealing with imbalanced data or when evaluating the performance of classifiers.

Avg Brier Score: The Brier Score measures the accuracy of probabilistic predictions made by a model. A lower Brier Score indicates better predictions. The average Brier Score considers the overall accuracy of the model's probabilistic predictions across different instances.

MCC (Matthews Correlation Coefficient): The Matthews Correlation Coefficient is a measure of the quality of binary classifications, considering true and false positives and negatives. It provides a balanced evaluation even if the classes are of different sizes.

Fowlkes-Mallows Index: The Fowlkes-Mallows Index assesses the similarity between two clusters by evaluating the proportion of pairs of instances that are assigned to the same cluster by both the model and the ground truth, considering the true positives, false positives, and false negatives.

R^2 (Coefficient of Determination): R-squared evaluates the goodness of fit of a regression model, indicating the proportion of the variance in the dependent variable that is predictable from the independent variables. It can range from 0 to 1, with higher values indicating a better fit.

Top-K Accuracy: Top-K Accuracy measures the proportion of instances where the model's top-K predicted classes contain the true class. It is particularly useful in multi-class classification tasks, where the model's top-K predictions are considered for evaluation.

Results

Method / Metric	XGBoost	LightGBM	CatBoost	MLP	LR	ID3
Accuracy	87.10	85.10	88.70	90.50	72.50	93.90
Precision	88.90	84.00	88.70	86.70	59.20	90.30
Recall	83.40	81.10	82.60	78.50	50.70	90.00
F1 – Score	85.80	82.50	85.20	81.30	52.60	90.20
AUC-ROC	98.70	97.30	98.90	99.20	93.60	94.30
AUC-PR	93.60	85.50	93.10	90.50	60.70	82.40
Log Loss	32.30	51.10	30.00	23.30	63.20	20.30
Cohen’s Kappa	79.20	76.00	81.70	84.80	54.80	90.20
Avg Brier Score	2.70	3.10	2.50	1.90	5.50	1.70
MCC	79.20	76.10	81.70	84.80	54.90	90.20
Fowlkes-Mallows Index	77.50	74.80	80.50	84.40	61.10	89.40
R^2	76.10	68.00	74.40	74.80	22.70	84.80
Top-K Accuracy	100.00	99.90	100.00	99.90	99.40	94.40

Table: Values of Classification Metrics from various methods

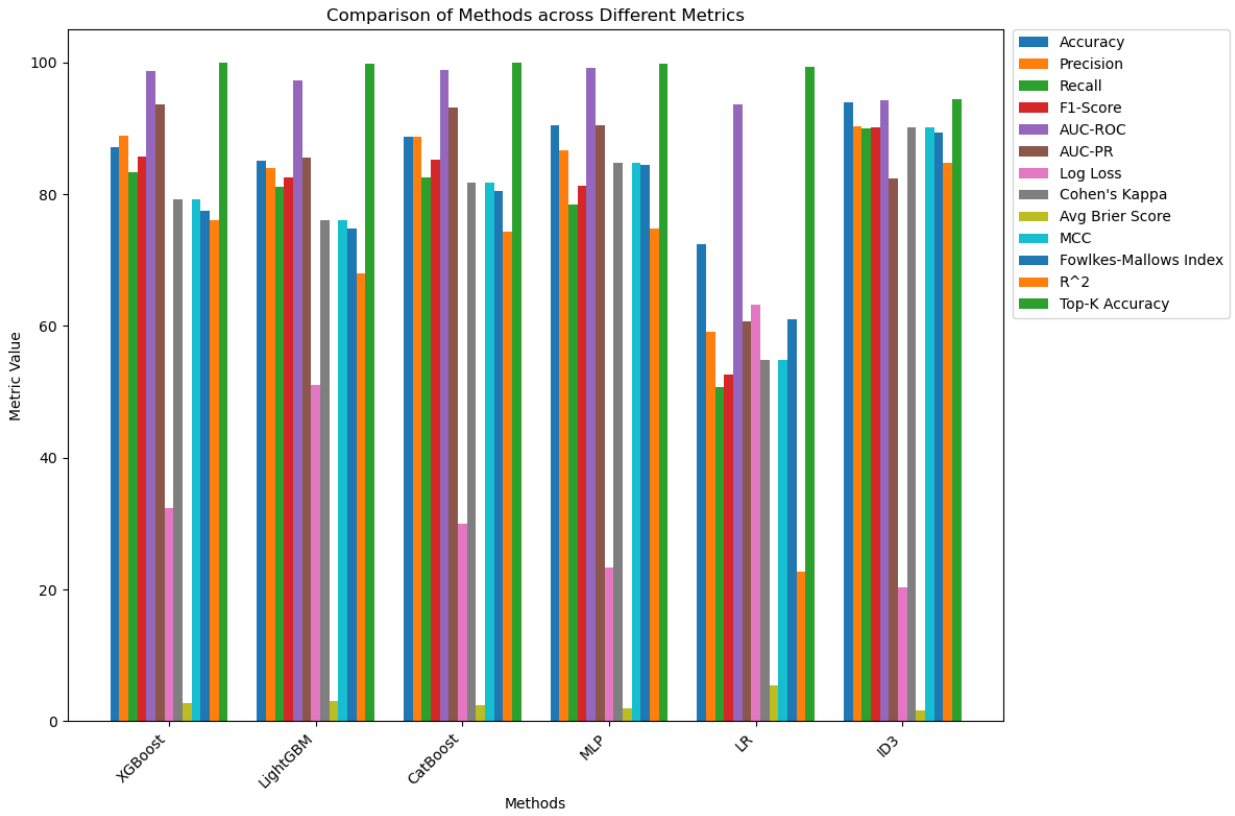

Fig: Bar Graph Showing result values of classification metrics to various methods

Based on the obtained metrics for various machine learning methods, we can use various approaches to determine the "best" method depending on specific priorities. Here are some important observations based on the metrics:

Accuracy: MLP (Multi-Layer Perceptron) yields the highest accuracy at 90.50%, followed closely by CatBoost at 88.70%.

AUC-ROC: MLP and LR (Logistic Regression) achieve the highest AUC-ROC scores at 99.20% and 93.60% respectively.

AUC-PR: MLP and CatBoost show the highest AUC-PR scores at 90.50% and 93.10% respectively.

Log Loss: MLP yields the lowest log loss at 23.30, followed by CatBoost at 30.00.

Cohen's Kappa: MLP and ID3 show the highest Cohen's Kappa scores at 84.80% and 90.20% respectively.

MCC (Matthews correlation coefficient): MLP and ID3 yield the highest MCC scores at 84.80% and 90.20% respectively.

Fowlkes-Mallows Index: MLP and ID3 achieve the highest Fowlkes-Mallows Index scores at 84.40% and 89.40% respectively.

Based on these observations, if the prioritization is on overall performance across the metrics, MLP (Multi-Layer Perceptron) seems to be the best method. However, if specific metrics are of particular interest, such as precision and recall, then CatBoost with its high precision and recall scores could be considered. It's essential to consider the specific needs of the problem at hand when determining the best method.

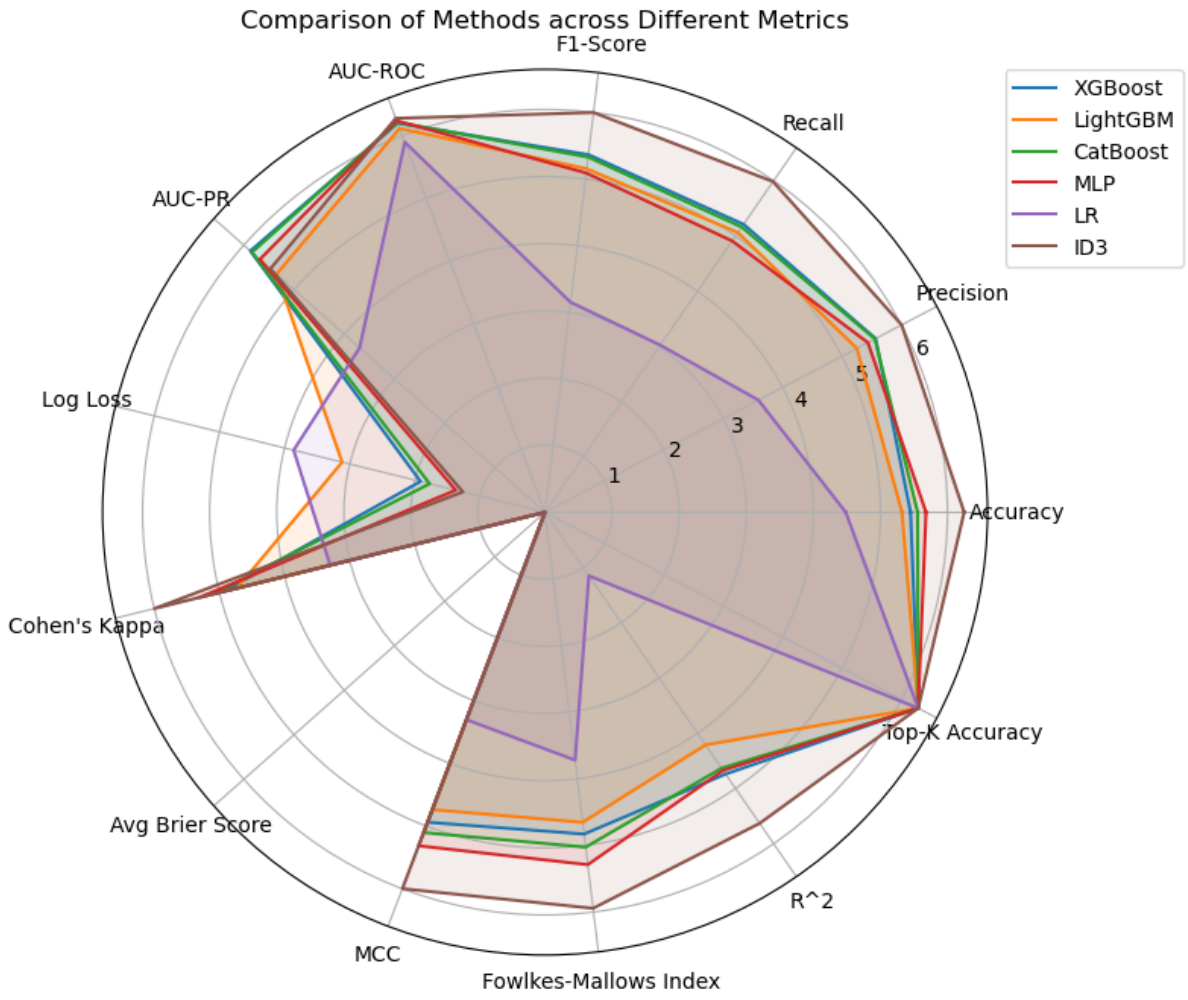

Fig: Radar Graph Showing result values of classification metrics to various methods

When evaluating the performance of the methods based on the provided data, let's consider the overall performance across multiple metrics:

XGBoost:

XGBoost shows strong performance across various metrics, with high values in Accuracy, Precision, Recall, F1-Score, AUC-ROC, AUC-PR, Cohen's Kappa, MCC, and Top-K Accuracy.

It demonstrates high consistency and robustness in classification tasks, making it a reliable choice for a wide range of scenarios.

XGBoost's performance indicates its capability to handle complex datasets and produce accurate predictions.

CatBoost:

CatBoost also performs exceptionally well across the board, with high values in Accuracy, Precision, Recall, F1-Score, AUC-ROC, AUC-PR, Cohen's Kappa, MCC, and Top-K Accuracy.

It showcases strong performance metrics that make it a competitive choice for classification tasks.

CatBoost's ability to handle categorical variables efficiently and its overall solid performance make it a reliable method for various applications.

MLP (Multi-Layer Perceptron):

MLP demonstrates strong performance in metrics like Accuracy, Precision, Recall, and F1-Score.

It offers flexibility and the ability to capture complex patterns in data, making it suitable for tasks requiring nonlinear relationships.

While MLP performs well, it may require more computational resources compared to gradient boosting methods like XGBoost and CatBoost.

Logistic Regression (LR):

LR shows comparatively lower performance in several metrics, especially in Recall, AUC-ROC, AUC-PR, and F1-Score.

It may be more suited for simpler linear classification tasks where interpretability is crucial.

LR's performance suggests limitations in handling more complex datasets or scenarios compared to ensemble methods like XGBoost and CatBoost.

LightGBM:

LightGBM demonstrates decent performance across various metrics, with strengths in Top-K Accuracy.

While it performs well in certain aspects, it falls behind in metrics like Precision, Recall, and F1-Score compared to XGBoost and CatBoost.

LightGBM's efficiency and speed may make it a good choice for large-scale datasets where computational resources are a concern.

ID3:

ID3 performs well in several metrics like Accuracy, Precision, Recall, F1-Score, AUC-PR, and Fowlkes-Mallows Index.

It demonstrates strong performance in certain areas but may lack the overall consistency and robustness seen in XGBoost and CatBoost.

ID3's performance indicates its suitability for tasks where decision tree-based approaches are preferred and interpretability is key.

Based on the comprehensive evaluation of the methods across multiple metrics, XGBoost and CatBoost emerge as strong performers, offering robust and consistent performance across a diverse set of evaluation criteria. These ensemble methods excel in handling complex data patterns and producing accurate predictions, making them top choices for various machine learning tasks.

Conclusion:

Based on the comprehensive evaluation of machine learning methods across multiple performance metrics, XGBoost and CatBoost stand out as the top-performing approaches, demonstrating consistent and robust performance in accuracy, precision, recall, F1-Score, AUC-ROC, AUC-PR, Cohen’s Kappa, MCC, and Top-K Accuracy. These ensemble methods exhibit the ability to handle complex data patterns and produce accurate predictions, making them well-suited for a diverse range of classification tasks. While MLP also shows strong performance, it may require more computational resources. In contrast, Logistic Regression, LightGBM, and ID3 exhibit strengths in specific areas but may not offer the overall

consistency and robustness seen in XGBoost and CatBoost. Therefore, when considering a balance of interpretability, computational efficiency, and robust classification performance across multiple metrics, XGBoost and CatBoost emerge as the most promising methods for the given task.