Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

A Comparative Study of Machine Learning Algorithms for Tabular Data Classification

Amit Kushwaha, Ganesh Kudale
Master's Student
Data Science,
Dr. DY Patil Arts, Commerce & Science College, Pimpri, Pune

Abstract - Tabular data is one of the most common types of data that needs to be modeled in many fields such as finance, healthcare and e-commerce where it is crucial to have accurate and trustworthy classification for decisions making. In classical machine learning approaches, many classifiers like Logistic Regression, Decision Tree, Random Forest and Gradient Boosting have widely been used on tabular data due to their simplicity and high performance. However, these models are based on hand, crafted feature engineering assumptions and maybe hard to model complex feature interactions and some other non-linear relationships in structured datasets with high dimensionalities.

This paper presents a comparison of machine learning classifiers applied on tabular data and shows how the new advances, such as the TabNet and the FT-Transformer as powerful deep learning architectures, can outperform the classical algorithms while exploiting attention mechanisms to automatically learn the most important features while capturing the relationships among them. Moreover, to enhance the explainability of your model predictions and to avoid the "black, box" syndrome that high complex models normally suffer from, we also adopt some promising XAI approaches like the SHAP and the LIME to better explain our model predictions.

Keywords - Tabular Data Classification, Machine Learning Algorithms, Logistic Regression, Random Forest, TabNet,FT-Transformer, Explainable Artificial Intelligence, SHAP, LIME, Model Interpretability, Classification Performance.

INTRODUCTION

With the fast proliferation of data, driven applications, more and more focus is put on tabular data in the fields of classification and decision making. Tabular data is very easy to understand, interpret and process on, thus it is typically used in prognosis prediction in diagnosis, risk analysis in

finance, customer behaviour forecasting, industrial monitoring, etc. Making the data processable by predictive models will generate big business benefits, and that is the ultimate goal of classification task on tabular data.

Traditional machine learning models such as Logistic Regression, Decision Trees, Random Forest, Support Vector Machines, etc., are dominant in tabular data classification for their simple implementation, fast execution and effective performance. In the ensemble, based models, boosting algorithms brings a new state of the art result by iteratively training a sequence of classifiers, where each one in the sequence learns the residual of the previous one, and the aggregated final predictor significantly boosts the classification accuracy. However, for fitting higher level interactions and dependencies between specific feature combination, traditional models or boosting algorithms may lack the automaticity and efficiency. Recently, deep learning models like TabNet, FT-Transformer, have emerged for tabular data with a weighted, gated attention mechanism to proactively select the most related features and max out the superiorities in capturing non, linear feature interactions, basically inspired by the attention mechanisms in natural language processing and computer vision. Unfortunately, their relative performance is still not stable on a large series of datasets, which is a very interesting exploring topic. The research in this paper will present a comprehensive comparison on these models for understanding and theoretical investigation.

LITERATURE REVIEW

Machine learning methods for structured data are the most relevant, given the nature of the information contained in tabular data and the relevance of the problem for practical applications. The initial approaches for tabular data focused on linear and probabilistic models, like logistic regression and naive Bayes, because they are easy to use, fast and easy to interpret. Logistic regression has also been used as a baseline classifier in several studies, due to its performance with data that has linear decision boundaries and binary/multiclass probability estimates.

Afterward, Decision Tree, based methods also got developed to address the drawbacks of linear models with their ability to model non-linear relationships between features. Decision Tree classifiers are easily interpretable but many studies demonstrated their huge tendency to overfit on small and noisy datasets [11], [12], [13]. To mitigate this problem, ensemble, based methods were developed, such as Random Forest. Random Forest classifier, as an ensemble of decision trees, yields more accurate classifications as their predictions are averaged, thus reducing overfitting. As such, Random Forest has been widely used in many tabular data domains.

More recent advances in ensemble learning have produced the family of boosting algorithms including XGBoost, LightGBM and CatBoost. Boosting algorithms sequentially build models

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

where each model attempts to compensate for the weaknesses in the ensemble by focusing on correcting the errors made by prior models. XGBoost, in particular, contributed regularization techniques (to combat overfitting) with optimized tree construction process, and thus becoming a very competitive model for structured data. LightGBM improved model training time when handling large data sets with histogram based learning and leaf, wise tree growth. CatBoost introduced novel handling of categorical features, and was able to significantly reduce bias in the prediction stage. Many empirical studies show that boosting algorithms tend to show consistent outperformance over most other machine learning models and deep learning models on small and medium sized tabular data.

While ensemble methods are highly successful, the recent advent of deep learning has lead to several studies on applying deep learning approaches to tabular data to discover features and interactions automatically. However, standard neural networks have failed to outperform the currently used tree models in common tabular datasets as they are sensitive to hyperparameters and difficult to interpret. This prompted the introduction of specialized deep learning models for tabular data.

In the following, TabNet was suggested as deep neural network that utilize sequential attention mechanisms for deciding the most relevant features during each decision step. In contrast to standard neural networks, TabNet offers interpretability where the relevance of the features can be identified during predictions. Based on multiple studies, the learning model was able to compete with boosting method and be more transparent but less accurate with small training datasets.

More recently, various transformer based architectures have been repurposed for tabular data, including the FT- Transformer, where self, attention is used to learn global interactions between individual features as tokens. Studies have shown that transformer based approaches can outperform other deep learning algorithms for complex tabular data as longer training periods provide better accuracy, but the comparison to state, of, the art boosting algorithms is relatively indifferent.

In general, the literature has yet to find a superior model for tabular data classification. Boosting techniques continue to be top contenders given their ability to effectively use data and robustness while tabu deep learning techniques such as TabNet and FT-Transformer have great potential for improving feature learning and interpretability. This work builds on previous work by conducting comprehensive comparisons between classic machine learning approaches, ensemble boosting algorithms, and recent tabular deep learning architectures in a single experimental setup.

PROBLEM STATEMENT

Tabular data is one of the most used data structure in various fields. Classification of this structure of data is still a very difficult task since the features distributions are diverse, some relations are non-linear and there is often an imbalance between the classes. Many classification approaches depend on classical machine learning algorithms which need human feature engineering and can't uncover all the complex interactions between features. Ensemble, based methods tend to increase the predictive performances but lack of explanations and the relevance of the latest deep learning models on tabular data has not been ascertained yet.

Furthermore, aforementioned second generation deep learning architectures for tabular data such as TabNet and FT-Transformer are interesting, but they have to be tested and contrasted with a comprehensive study against the most well, known traditional and boosting algorithms. In this sense, more experiments with a set of benchmark datasets are required to understand the true innovation that these kind of models introduce to traditional ones.

OBJECTIVE

The main aim of the study is to compare machine learning algorithms for tabular data classification.

The specific aims of the work are:

1. To study the performance of conventional machine learning classifiers like Logistic Regression, Decision Tree, Random Forest and Support Vector Machine on tabular data.
2. To study the ensemble based boosting algorithms, like XGBoost, LightGBM and CatBoost and their accuracy.
3. To study the advanced deep learning models for tabular data like TabNet and FT-Transformer and their accuracy.
4. To compare all the models based on the common accuracy metrics like accuracy, F1-score, recall, Precision, ROC-AUC etc.
5. To find out the best classification method over conventional and deep learning models for tabular data.

METHODOLOGY AND DATASET

Methodology

The experimental methodology is based on a machine learning pipeline to ensure a fair comparison of the various models, which consists of a work, flow of Data processing, Models implementation, Performance evaluation, and Comparison.

1. Data Preprocessing

The data sets I used for this study are designed with structured numerical features that are spaced for binary of two, class categories. Before building classification models, I preprocessed the data set for better models' accuracy and stability.

The following preprocessing steps are applied:

- To deal with missing values if any.
- Numerical preprocessing (feature scaling): numerical attributes normalized through standardization (zero mean and standard deviation equal to 1)
- Partitioning the data into a training set and a test set to test generalization performance

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

2. Model Implementation

Several ML models were used on tabular data classification task. These models are:

- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier
- Support Vector Machine (SVM)
- XGBoost
- LightGBM
- CatBoost
- TabNet
- FT-Transformer

3. Performance Evaluation

The performance of the models was assessed using a selection of metrics to ensure that an overall picture of classification quality was obtained. The metrics used for model assessment were:

- Accuracy
- Precision
- Recall
- F1-Score
- ROC-AUC

RESULTS AND ANALYSIS

The performance of the implemented machine learning models was evaluated using multiple metrics, including accuracy, precision, recall, F1-score, and ROC-AUC. These metrics provide a comprehensive understanding of model behavior by considering both overall correctness and class-wise prediction quality. The experimental results are summarized in Table 1, which presents a comparative analysis of all traditional, ensemble-based, and deep learning models.

	Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
0	Logistic Regression	0.982456	0.986111	0.986111	0.986111	0.995370
1	Decision Tree	0.903509	0.955224	0.888889	0.920863	0.908730
2	Random Forest	0.947368	0.958333	0.958333	0.958333	0.994048
3	SVM	0.982456	0.986111	0.986111	0.986111	0.995040
4	XGBoost	0.956140	0.946667	0.986111	0.965986	0.992725
5	LightGBM	0.956140	0.946667	0.986111	0.965986	0.989087
6	CatBoost	0.956140	0.946667	0.986111	0.965986	0.994378
7	TabNet	0.447368	0.666667	0.250000	0.363636	0.703704
8	FT-Transformer	0.921053	0.984615	0.888889	0.934307	0.992063

Table 1: Performance Comparison of Machine Learning Models for Tabular Data Classification

From the results, traditional classifiers such as Logistic Regression and Support Vector Machine demonstrate strong performance, achieving high accuracy and F1-score values. Their effectiveness can be attributed to the structured and well-separated nature of the tabular dataset. Decision Tree shows comparatively lower performance, which may be due to overfitting and sensitivity to data variations. Random Forest improves upon this by aggregating multiple trees, resulting in better generalization and higher overall scores.

Ensemble-based boosting algorithms, including XGBoost, LightGBM, and CatBoost, achieve consistently high performance across all evaluation metrics. Among these, **CatBoost emerges as the best-performing model**, achieving the highest accuracy and ROC-AUC. This superior performance highlights the ability of boosting algorithms to capture complex feature interactions while remaining robust on structured tabular data.

With respect to deep learning approaches, TabNet demonstrates significantly lower performance compared to other models. This outcome can be attributed to the relatively small dataset size, where deep learning architectures are unable to fully exploit their representational capacity. In contrast, **FT-Transformer performs considerably better than TabNet**, achieving competitive accuracy and F1-score values, although it does not surpass CatBoost. The graphical visualizations (Fig 1, Fig 2, and Fig 3) further illustrate these performance differences, confirming that while transformer-based models are promising, ensemble-based methods remain highly effective for tabular classification tasks.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

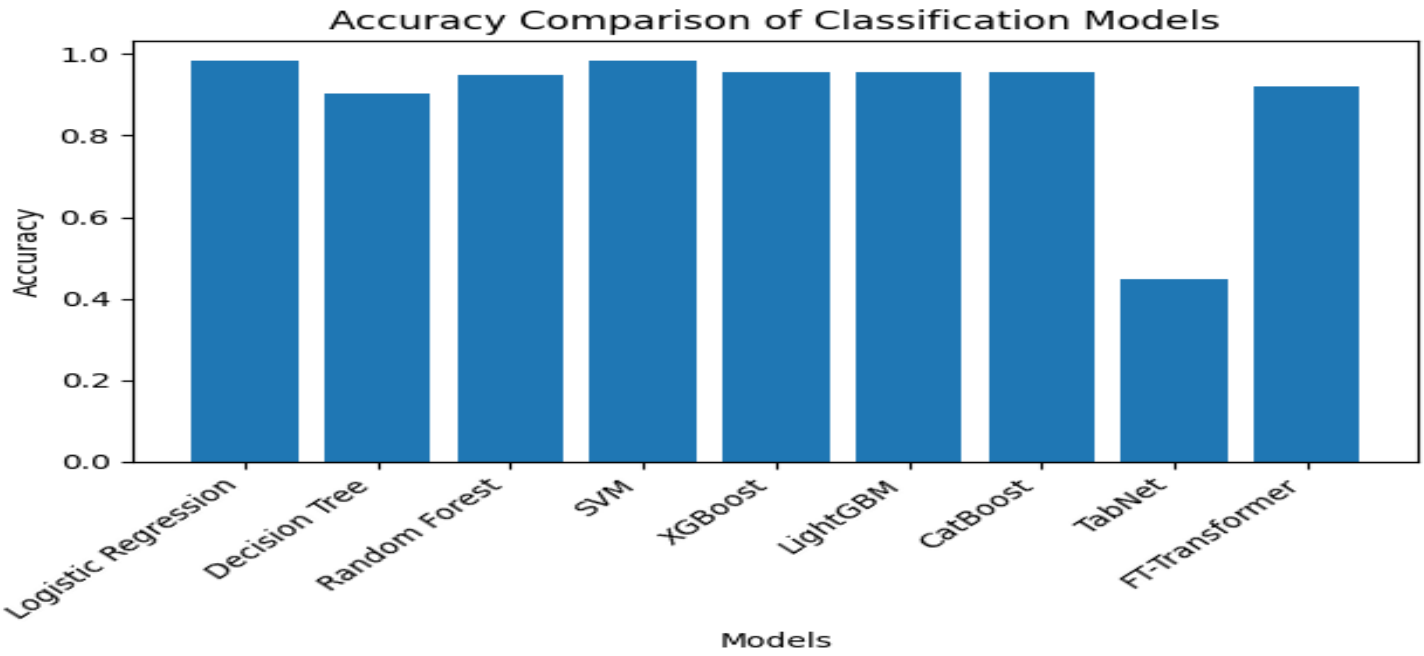

Fig 1: Accuracy Comparison of Classification Models

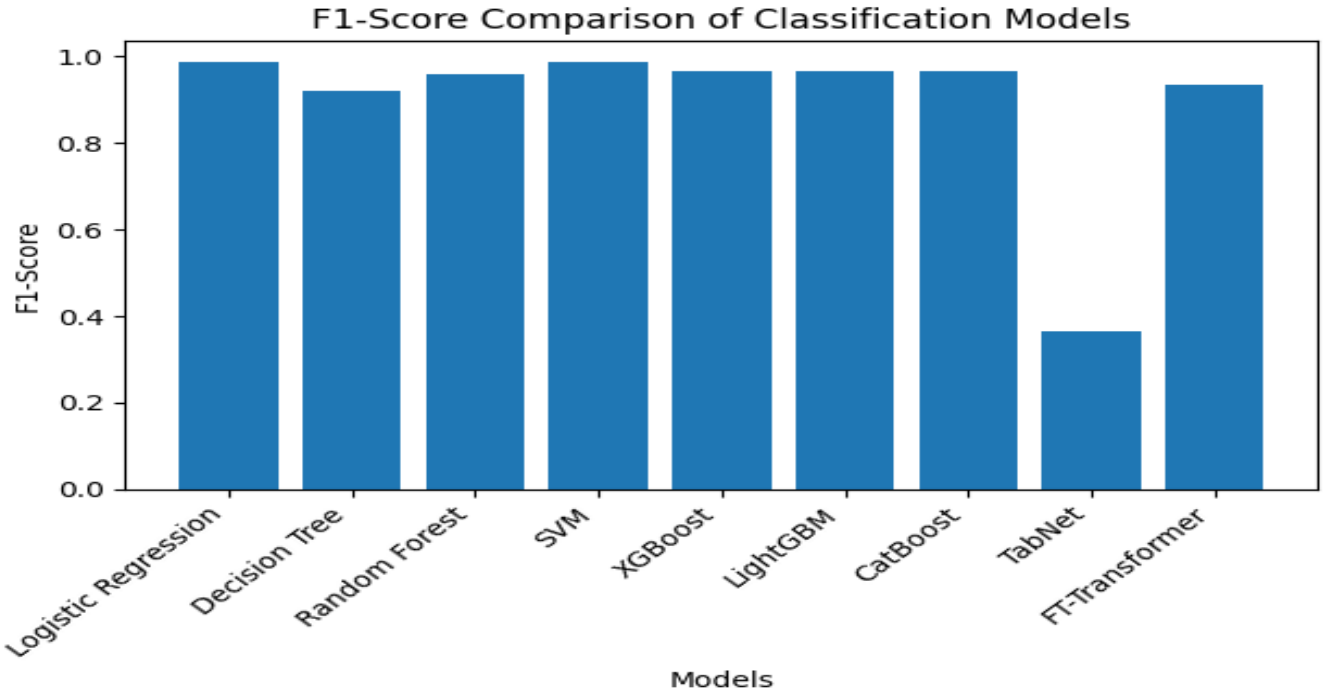

Fig 2: F1-Score Comparison of Classification Models

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

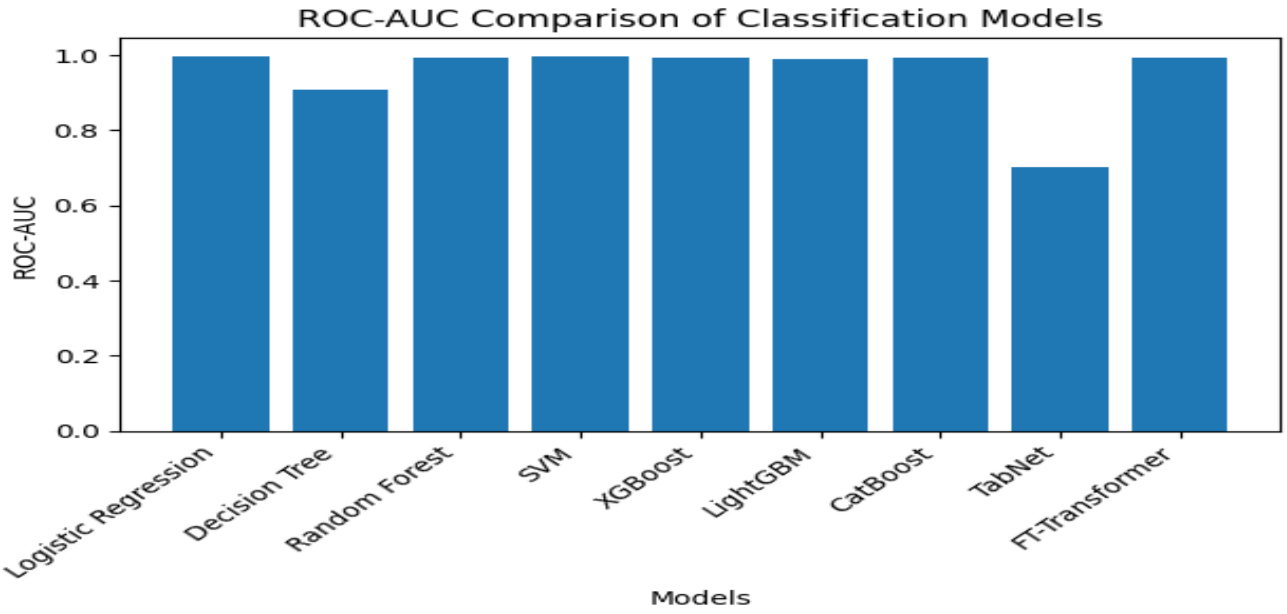

Fig 3: ROC-AUC Comparison of Classification Models

Conclusion

This work provides a side, by, side comparison of various machine learning algorithms for tabular data classification, not only traditional classifiers, but also various ensemble based boosting algorithms as well as cutting edge deep learning models. The experiments confirmed that the application of ensemble based methods leads to the best classification results in terms of various evaluation metrics, with CatBoost showing the best results among all algorithms. Logistic Regression and SVM also yielded competitive results, confirming the position of traditional algorithms on small datasets.

The deep learning approach FT-Transformer showed promising capability of modeling more complex feature interactions, outperforming existing deep learning model TabNet, even though it does not beat the applied boosting algorithms, when it comes to particular small and medium size data sets. It does provide a good overall solution however, and performance should not be the exclusive criterion.

FUTURE SCOPE

While this study offers a thorough head, to, head comparison of popular machine learning algorithms for tabular data classification, there are many avenues for further research. For larger and more complex datasets, researchers could use the models in this study, and potentially find more significant performance improvements from deep learning architectures like TabNet and FT-Transformer. Hybrid solutions that modify ensemble, based methods with deep learning models could also be tested.

A further interesting avenue will be the deployment of various Explainable Artificial Intelligence (XAI) approaches including SHAP and LIME to make models more transparent and trusted. This is critical in application scenarios such as in healthcare and finance where both accurate prediction and model explanation are equally important. Further future research directions include real, time implementation, automatic feature engineering and cross, domain generalization.

REFERENCES

[1] Chen, T., & Guestrin, C. (2016). XGBoost : A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785 –794. IEEE.
[2] Seminal paper introducing XGBoost, widely used for tabular data classification tasks.
[3] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased Boosting with Categorical Features. Advances in Neural Information Processing Systems, 31, 6638 –6648.
[4] Presents CatBoost, a gradient boosting algorithm optimized for categorical tabular data.
[5] Arik, S. O., & Pfister, T. (2021). TabNet: Attentive Interpretable Tabular Learning. Proceedings of the AAAI Conference on Artificial Intelligence, 35(12), 6679 –6687.
[6] Introduces TabNet, a deep learning model specifically designed for tabular datasets with built -in interpretability.
[7] Huang, S., Li, W., & Zhou, Q. (2022). FT -Transformer: A Transformer Architecture for Tabular Data. Pattern Recognition, 125, 108501. Elsevier.
[8] Describes transformer -based approaches for learning from tabular data.
[9] Lundberg, S. M., & Lee, S. -I. (2017). A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems, 30, 4765 –4774.
[10] Introduces SHAP, a method for explaining predictions from complex models.
[11] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144. ACM.
[12] Introduces LIME, a framework for local model interpretability.
[13] Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5 –32. Springer.
[14] Foundational paper on ensemble methods for classification and regression.
[15] Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics, 29(5), 1189 –1232.
[16] Theoretical foundations of gradient boosting methods.