International Journal of Enhanced Research in Science, Technology & Engineering
ISSN: 2319-7463, Vol. 13 Issue 10, October-2024, Impact Factor: 8.375

Comparative Analysis of Machine Learning Models for Predictive Performance of Different Datasets

V Rajanikanth Tatiraju[1], Dr. Rohita Yamaganti[2]

[1]Research Scholar, Department of Computer Science and Engineering, P. K. University, Shivpuri, M.P
[2]Associate Professor, Department of Computer Science and Engineering, P. K. University, Shivpuri, M.P

ABSTRACT

Using three different datasets—synthetic noisy data, biomedical disease prediction, and financial credit risk—this study compares and assesses the performance of various machine learning models, such as Adaptive Linear v-Support Vector Regression, Support Vector Machine, Logistic Regression, Random Forest, and v-Support Vector Machine. Testing how well these models handled various kinds of data and made accurate predictions was the goal. Several measures were used to quantify performance, including as recall, accuracy, precision, F1-score, and Area under the Curve. The results showed that AL-vTSVR was the most effective model in every performance parameter tested, showing that it could handle complicated real-world data and noise with ease. Random Forest shown competitive performance as well, particularly in financial and medicinal domains. In contrast to SVM and v-TSVM, Logistic Regression showed less effectiveness. The results demonstrate that AL-vTSVR is an exceptionally dependable model for difficult data situations, and they emphasize its better capabilities in various prediction tasks.

Keywords: Noisy, Support Vector Machine, Accuracy, Precision, Recall

INTRODUCTION

Machine learning (ML) has changed several industries by letting computers discover patterns in data and use that knowledge to make judgments or predictions without human intervention. When it comes to predicting future events or outcomes using historical or real-time data, machine learning models are useful tools in the context of predictive performance. With the use of big datasets and advanced algorithms, these models are able to uncover patterns and make predictions, the accuracy of which might vary. Machine learning has become an essential tool for predicting tasks because to the growing amount, diversity, and speed of data in many fields, including healthcare, finance, marketing, and engineering.

Machine learning essentially entails creating algorithms that can autonomously learn from data and improve upon past performance. In machine learning, predictive performance is a model's capacity to generate correct predictions when presented with novel, unseen data. Machine learning models can process massive, unstructured information and reveal complex correlations between variables that would otherwise go unnoticed, in contrast to traditional statistical approaches that depend significantly on established assumptions. These models' predictive capability shines through when they leverage historical trends to assist decision-making; this makes them applicable to tasks like demand forecasting, stock market prediction, predictive maintenance, and disease outbreak forecasting, among others.

Machine learning models come in a variety of flavors, each optimized for a particular kind of prediction job. Predictive analytics makes extensive use of supervised learning models. To train these models, we use labelled data, in which each input attribute has an associated label. Constructing a model capable of making predictions based on novel, unknown input data is the main objective of supervised learning. Ensemble techniques such as random forests and gradient boosting are common examples of supervised learning algorithms, along with linear regression, decision trees, and support vector machines (SVMs). These models are great at many different kinds of prediction performance challenges because they are so good at classification and regression.

In contrast, unsupervised learning models are employed in situations when the data does not contain labelled outcomes. To the contrary, these models unearth previously unseen patterns and structures in the data. Unsupervised learning frequently makes use of clustering and dimensionality reduction methods like k-means, hierarchical clustering, and principal component analysis (PCA). Unsupervised learning models aren't meant to make predictions per se, but they can be useful for pre-processing data by highlighting clusters or important traits that supervised learning models can exploit to their advantage.

Another subfield of machine learning, reinforcement learning (RL) is concerned with decision-making in settings where the model acquires knowledge by interactions and feedback. To find the best solution, an agent in RL acts in its environment and, depending on the results, gets rewards or penalties. RL shines in robotics, games, and autonomous systems, among other areas, when forecasts must take sequential decision-making into consideration. Games, robot control, and resource optimization are just a few of the areas where deep reinforcement learning—a combination of deep learning and RL—has achieved remarkable progress.

The creation of deep learning models represents a turning point in machine learning as it pertains to prediction performance. "Deep learning" refers to a subfield of machine learning in which multi-layered neural networks automatically learn hierarchical data representations. Several applications, including time series forecasting, picture recognition, and natural language processing, have demonstrated exceptional performance from these models. One common deep learning architecture that has seen extensive use in prediction tasks is the convolutional neural network (CNN). Another is the recurrent neural network (RNN). When it comes to image-based tasks, CNNs really shine. On the other hand, RNNs, especially LSTM and GRU, really shine when it comes to sequential prediction tasks, like predicting time-series data or interpreting natural language.

Data quality, algorithm selection, and hyperparameter tuning are three of the most important determinants of a machine learning model's predictive performance. Before a machine learning model can learn any useful patterns from data, the data must undergo data preparation. It is usual practice to enhance the data quality before to training a model using techniques like normalization, feature selection, and imputation of missing values. Another important aspect of evaluating machine learning models for predicting performance is model assessment. Area under the receiver operating characteristic curve (AUC-ROC), F1 score, recall, accuracy, precision, and area under the receiver operating characteristic curve (ACCURATE) are common metrics for classification tasks, whereas R-squared, MSE, and RMSE are used for regression activities.

REVIEW OF LITERATURE

Petchiappan, Maheswari & Jaya, A. (2022) Investors have always found trend prediction in the stock market to be a difficult and perplexing task. Technological developments, machine learning, data analytics, and big data have led to a meteoric rise in the accuracy of stock market predictions. Among the many varied industries represented on the stock market is the media and entertainment industry. The Sensex and the Nifty are the two indices used in the Indian stock market. Theatres were closed in 2019 because of the pandemic. This caused a halt in production and prevented distributors and directors from releasing their films to screens. So, during the lockdown, many stayed indoors and watched more television. Resulting in a higher degree of media consumption. The study's overarching goal is to use machine learning to foretell how the stock prices of the media and entertainment firm will do. Making as much money as possible while keeping losses to a minimum can help investors. In data science, the suggested stock prediction method is utilized for predicting stock prices and determining the accuracy of logistic and linear regression in machine learning algorithms. The media and entertainment industry's stock price data is used in the tests, which employ machine learning techniques. One example of an input dataset is media stock prices. Various aspects of stock prices with a daily frequency were used to create the model. In summary Media and entertainment stock prices are so anticipated using logistic and linear regression models. In order to help investors maximize their gains and minimize their losses, the stock prices are anticipated with a high degree of accuracy using the aforementioned methodologies.

Sekeroglu, Boran et al., (2022) The use of AI and ML to solve issues or augment human specialists is crucial in nearly every aspect of human existence. Researchers still face the formidable challenge of narrowing down the many real-world application areas to a single machine learning model that may produce superior results for a given problem. Several aspects, including the features of the dataset, the training approach, and the model's responses, might influence the model's performance. Hence, in order to ascertain the efficacy of the proposed tactics and the capability of the model, a thorough evaluation is necessary. Ten standard machine learning models were applied to seventeen different datasets in this research. Training procedures of60:40,70:30, and 80:20 hold-out, in addition to five-fold cross-validation, are used in the experiments. The experimental findings were assessed using three metrics: R2 score, mean absolute error, and mean

squared error. The models that were taken into consideration are examined, and the benefits, drawbacks, and data dependencies of each model are highlighted. The deep Long-Short Term Memory (LSTM) neural network achieved the best results compared to the other models tested (decision tree, linear regression, support vector regression with radial and linear basis function kernels, random forest, gradient boosting, extreme gradient boosting, shallow neural network, and deep neural network), all of which were determined by conducting an excessive number of experiments. When evaluating models in regression research without data mining or selection, cross-validation should be examined due to the substantial influence it has on experimental outcomes.

Varshini, Priya et al., (2021) To construct smart systems capable of problem-solving, Artificial Intelligence builds on top of Machine Learning and Deep Learning. The amount of time needed to do the task may be estimated using software effort estimation. Predicting Software Effort at the beginning phases of a project is fraught with difficulty and difficulty owing to several unknowns. You may use software effort estimation to better organize your project's timeline, resources, and budget. Expert judgment, regression estimations, categorization techniques, deep learning algorithms, and analogy-based estimations were some of the studies suggested for effort prediction. Based on its resilience and ability to manage big datasets, random forest surpasses other algorithms in this paper's comparison of deepnet, neuralnet, support vector machine, and random forest. Mean Absolute Error, Root Mean Squared Error, Mean Squared Error, and R-Squared are the evaluation metrics that should be considered.

Yuan, Kunpeng et al., (2021) Establishing a prediction model, default prediction determines the likelihood of a business defaulting. Data from features at time t-m and default state at time t are shown to have a functional link. A non-defaulting firm's forecast might lead to a loss of high-quality consumers, while an inaccurate forecast of a defaulting company could trick banks into lending to a "defaulter," resulting in massive losses. Using k-means clustering to divide the sample and support vector domain description (SVDD) to forecast default (credit scoring), this study suggests a two-stage default prediction model to aid lending choices made by banks and non-banking financial organizations.

To train the proposed model to warn of default m years ahead, it takes characteristics' data at time t-m (m = 1, 2, 3, 4, 5), together with the default state at t. Compared to single-stage models that rely solely on k-means clustering or support vector domain description, the findings demonstrate that the suggested two-stage default prediction model outperforms them. What's more, the proposed model was able to attain a five-year default prediction ability (AUC > 0.85). In addition, the study suggests that three important factors in default forecasting for Chinese listed businesses are "retained earnings/total assets," "financial expenses/gross revenue," and "type of audit opinion." By showing that it is worthwhile to explore combining alternative techniques to enhance the effectiveness of default prediction models, this work adds to the field of multi-stage credit scoring research.

Mounika, B. & Persis, Voola. (2019) Machine learning techniques are widely used in many different industries. In the classroom, for example, these methods have many potential uses. Machine learning approaches are being used in an increasing amount of educational research. Using machine learning techniques in a classroom setting can help unearth previously unknown information and trends regarding student achievement. Using machine learning classification methods such as K-Nearest Neighbor, Decision Tree, Support Vector Machines, Random Forest, and Gradient Descent Boost Algorithms, this effort intends to construct a model that predicts students' academic success across different departments. Factors such as residence, parent-child relationship, level of education and occupation, backlogs, attendance, availability of internet connection, and smartphone use are taken into account.

You may find out how well a student did on the final test and what their grade will be using the resultant prediction model. College administration or instructors can then use this information to identify which students need extra help and intervene before it's too late. With the help of early prediction, we may find ways to improve our performance in the final exams.

EXPERIMENTAL ANALYSIS

This study compared the efficacy of several ML models trained on synthetic noisy data, biological illness prediction, and financial credit risk datasets, each representing a distinct area. Adaptive Linear ν-Support Vector Regression (AL-νTSVR), Support Vector Machine (SVM), Logistic Regression, Random Forest, and ν-Support Vector Machine (ν-TSVM) are some of the models that are utilized for comparison. Area Under the Curve (AUC), Accuracy, Precision, Recall, and F1-Score are the primary performance indicators used to evaluate the efficacy of the model.

RESULTS AND DISCUSSION

Table 1: Performance Metrics on Synthetic Noisy Data

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC (%)
AL-vTSVR	92.5	90.0	94.0	91.8	95.2
SVM	86.0	85.5	88.3	86.9	91.4
Logistic Regression	81.5	79.2	84.0	81.6	87.7
Random Forest	88.4	87.2	89.6	88.3	93.5
v-TSVM	89.0	87.0	90.2	88.5	92.1

The AL-vTSVR model outperforms all other models on the synthetic noisy dataset, achieving the highest accuracy (92.5%), precision (90.0%), recall (94.0%), F1-Score (91.8%), and AUC (95.2%). Random Forest follows closely with strong results (accuracy: 88.4%, AUC: 93.5%) but does not match the AL-vTSVR. The v-TSVM model shows solid performance (accuracy: 89.0%, AUC: 92.1%), while SVM and Logistic Regression perform relatively worse, with Logistic Regression being the least effective across all metrics. In summary, AL-vTSVR is the best performer, especially for noisy data.

Table 2: Performance Metrics on Biomedical Disease Prediction

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC (%)
AL-vTSVR	95.3	92.4	96.1	94.2	97.8
SVM	90.1	89.7	92.5	91.1	94.8
Logistic Regression	84.7	80.2	88.3	84.1	89.9
Random Forest	92.8	90.3	94.7	92.5	96.0
v-TSVM	91.6	88.5	92.8	90.6	95.1

With a 97.8% AUC, 94.2% F1-Score, 96.1% recall, and 95.3% accuracy, the AL-vTSVR model outperforms all other models in biological illness prediction. In terms of illness identification accuracy, it much surpasses all other models. The v-TSVM model demonstrates good performance with an accuracy of 91.6% and an area under the curve (AUC) of 95.1%, while Random Forest follows with robust findings (accuracy: 92.8%, AUC: 96.0%). While Logistic Regression has the lowest overall metrics, SVM and Logistic Regression both perform well, but they aren't as effective as the top models. AL-vTSVR stands out in every performance metric.

Table 3: Performance Metrics on Dataset 3 (Financial Credit Risk)

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC (%)
AL-vTSVR	93.7	91.2	95.4	93.3	96.5
SVM	89.5	88.0	91.4	89.7	93.2
Logistic Regression	82.3	79.5	84.2	81.8	88.1
Random Forest	90.6	89.2	92.0	90.6	94.3
v-TSVM	92.0	89.8	92.5	91.1	94.7

On the financial credit risk dataset, the AL-vTSVR model outperforms the others with a 93.7% accuracy rate, 91.2% precision rate, 95.4% recall rate, 93.3% F1-Score, and 96.5% area under the curve. In terms of prediction abilities, it is superior to all other models. While v-TSVM demonstrates outstanding performance with an accuracy of 92.0% and an AUC of 94.7%, Random Forest follows with strong findings (accuracy: 90.6%, AUC: 94.3%). Even if it's not the worst model, SVM's performance isn't up to par, and Logistic Regression fares the worst on every criterion. When it comes to predicting credit risk, AL-vTSVR is the best model.

CONCLUSION

Results from this study show that different machine learning models perform well on prediction tasks in many areas, such as financial credit risk, biological illness prediction, and synthetic noisy data. Proof of AL-vTSVR's resilience in dealing with complicated and noisy datasets is its constant outperformance of rival models in several metrics such as accuracy, precision, recall, F1-score, and area under the curve (AUC). Also, Random Forest proved to be a formidable contender for a variety of prediction jobs, especially in the biological and financial domains. Although AL-vTSVR consistently outperformed SVM and v-TSVM, the latter two exhibited encouraging results. When it came to more complicated datasets,

Logistic Regression fell short, even if it worked well for smaller cases. In conclusion, AL-vTSVR is the best option for practical applications with complicated or noisy data because of its exceptional prediction skills in a variety of difficult domains.

REFERENCES

[1] P. Maheswari and A. Jaya, "Comparative study of machine learning algorithms towards predictive analytics," *Recent Advances in Computer Science and Communications*, vol. 16, no. 6, pp. 1-12, 2022.
[2] B. Sekeroglu, Y. K. Ever, K. Dimililer, and F. Al-Turjman, "Comparative evaluation and comprehensive analysis of machine learning models for regression problems," *Data Intelligence*, vol. 4, no. 3, pp. 620-652, 2022.
[3] S. Shi, R. Tse, W. Luo, S. D'Addona, and G. Pau, "Machine learning-driven credit risk: a systemic review," *Neural Computing and Applications*, vol. 34, no. 2, pp. 14327-14339, 2022.
[4] P. Varshini, A. K. Kumari, D. Janani, and S. Soundariya, "Comparative analysis of machine learning and deep learning algorithms for software effort estimation," *Journal of Physics: Conference Series*, vol. 1767, no. 1, pp. 1-11, 2021.
[5] K. Yuan, G. Chi, Y. Zhou, and H. Yin, "A novel two-stage hybrid default prediction model with k-means clustering and support vector domain description," *Research in International Business and Finance*, vol. 59, pp. 1-12, 2021.
[6] K. Peng and G. Yan, "A survey on deep learning for financial risk prediction," *Quantitative Finance and Economics*, vol. 5, no. 4, pp. 716-737, 2021.
[7] R. Chen, M. Lu, T. Chen, D. Williamson, and F. Mahmood, "Synthetic data in machine learning for medicine and healthcare," *Nature Biomedical Engineering*, vol. 5, no. 6, pp. 1-5, 2021. doi: 10.1038/s41551-021-00751-8.
[8] C. Caiafa, Z. Sun, T. Tanaka, P. Marti-Puig, and J. Solé-Casals, "Machine learning methods with noisy, incomplete or small datasets," *Applied Sciences*, vol. 11, no. 9, pp. 1-4, 2021.
[9] Z. Hassani, M. A. Meybodi, and V. Hajihashemi, "Credit risk assessment using learning algorithms for feature selection," *Fuzzy Information and Engineering*, vol. 12, no. 6, pp. 1-16, 2021. doi: 10.1080/16168658.2021.1925021.
[10] B. Mounika and V. Persis, "A comparative study of machine learning algorithms for student academic performance," *International Journal of Computer Sciences and Engineering*, vol. 7, no. 4, pp. 721-725, 2019. doi: 10.26438/ijcse/v7i4.721725.
[11] S. Uddin, A. Khan, M. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," *BMC Medical Informatics and Decision Making*, vol. 19, no. 6, pp. 1-16, 2019.
[12] P. Danėnas and G. Garšva, "Selection of support vector machines based classifiers for credit risk domain," *Expert Systems with Applications*, vol. 42, no. 6, pp. 1-20, 2015.