ASSESSMENT OF COMBINED MACHINE LEARNING MODELS ON SALEPREDICTION

[1] Sadiku, I. B. S., [2] Aina, O. A. [3] Onalaja. O. O, [4] Ganiyu, A. K
[1,2,3,4] Computer Science Department, Gateway ICT Polytechnic, Saapade, Ogun State,Nigeria
[1]Email: sibsadiku@gmail.com

Abstract:
Machine learning models are combined to output a better skillful model. The output model(s) to use or trust to solve a problem is subject to its record of performances. In this study, ensemble machine learning algorithm is used in anaconda3 environment to combine the predictions from multiple machine learning classification models: K-Nearest Neighbors (KNN), Support Vector Machine Enabled Radial Basis Function (SVM-RBF), Decision Trees (DT), Random forest (RF), Multilayer perceptron (MLP) on sale dataset. The dataset consist of three levels of sale channels (In-person, Instagram and Whatsapp) and three levels of movement restriction period (COVID- 19, nationwide restriction due to national election and no restriction movement). The results revealed that KNN, SVM-RBF, DT, RF, MLP, and the combined model had: accuracy score of 62.99 %, 91.57 %, 52.45 %, 89.93 %, 33.02 % and 49.41 % respectively; the Matthews correlation coefficient (MCC) gave 50.84 %, 88.85 %, 34.76 %, 86.61 %, 6.12 % and 29.95 % respectively; F1 score also gave 59.92 %, 91.24 %, 44.83 %, 89.58 %, 19.58 % and 43.59 % respectively. It can be seen from the results that combining models for machine learning operation does not necessarily give higher performances in all score category but the confidence in individual model to accurately learn on the dataset or from neighbor model on its own.

Keywords: *Skillful, Algorithm, Anaconda3, Sale Channels, Movement Restriction Period*

Introduction

It has been observed that accuracy in prediction is still a challenging task due to the volatility in the market affecting parameters (Sarangi et al. 2021) and machine learning model could predict with certain accuracy. Researchers have been using multiple machine learning models for better accuracy of the combined models. The use of multiple machine learning models is by grouping more than one model to combining or stacking models as a hybrid model with the hope to have better performing model (Shahhosseini et al.2021). According to Buyrukoğlu and Savaş (2023), stacking could be described as one of the most popular ensemble machine learning techniques that could be used to predict multiple nodes to build a new model and improve model overall performance. Stacking enables training of multiple models to solve similar problems and based on their combined output, it builds a new model with improved performance. An ensemble model in machine learning combines the predictions from two or more models. Ensemble learning methods found in machine learning includes bagging, boosting and stacking (Ribeiro and Coelho, 2020). In stacking, various weak learners are ensembled in a parallel manner in such a way that by combining them with Meta learners, better predictions for future can be made (Li et al. 2021). This ensemble technique works by applying the input of combined multiple weak learner's predictions and meta learners so that a better and accurate output prediction model could be achieved. In stacking, an algorithm normally takes the outputs of sub-models as input and attempts to learn how to best combine the input predictions to make a better output prediction. Prediction of sales is essential for estimating future revenue by forecasting the amount of product or serviced a sales unit will make. Sales forecast allow business owners to plan out their much-needed requirements such as their raw materials, workforce, budget, and other logistics-related needs. Accurate sales prediction make sales people and business leaders to make smarter decisions when setting goals, hiring, budgeting, prospecting and other revenue-impacting factors.

Background Study

Researchers apply a number of statistical and machine learning techniques to predict the future expected return of investment. Sarangi et al. (2021) combined Artificial Neural Network (ANN) and Particle Swarm Optimization (ANN-PSO) model to predict future price of gold. The researcher analysed the effectiveness of the ANN-PSO hybrid machine learning model and the result showed that the ANN-PSO model is capable of predicting the future gold price with high accuracy. Mojrian et al. (2020) used extreme learning machine (ELM) classification model integrated with radial basis function (RBF) kernel to form ELM-RBF model to predict the presence of breast cancer. The ELM-RBF model outperforms the linear-SVM model indicating that the ELM-RBF has superior prediction with accuracy, precision, sensitivity, specificity, validation, true positive rate (TPR), and false-negative rate (FNR). Deepa et al. (2018) proposed hybrid machine learning model to protect the SDN controller from DDoS attacks and experimental results clearly manifest that the hybrid machine learning model provides more accuracy, detection rate and less false alarm rate compared to simple machine learning models. Seretis and Sarris (2022) used hybrid model as a combined model for predictions of both measured and simulated data models to predict the signal levels at any location in the environment and it was observed that even where volume of measured data is insufficient the hybrid model still improve the accuracy of the overall

predictions.

In the case of machine learning it is a sub-field of artificial intelligence (AI). The goal is to understand the structure of data. It also fit that data into models that could be understood and utilized by people very well. Although machine learning is a field within computer science, it differs from traditional computational approaches. In traditional computing, algorithms are sets of explicitly programmed instructions used by computers to calculate or solve problem. Machine learning algorithms instead allow for computers to train on data inputs and use statistical analysis in order to output values that fall within a specific range. Machine learning, because of this, facilitates computers in building models from sample data so as to automate decision-making processes based on data inputs. Machine Learning also enables systems to learn on their own instead being explicitly programmed to do so, resulting in more intelligent behavior. It generates data-driven predictions by developing models that discover patterns in historical data and utilize those patterns to generate predictions. The general architecture of machine learning consists of several steps: business understanding (understanding and knowledge of the domain), data acquisition and understanding (gathering and understanding data), modeling (which entails feature engineering, model training, and evaluation), and deployment. K-Nearest Neighbors (KNN) is one of the tools use in the field of artificial intelligence and is easy to implement and can handle complex data. KNN is a supervised learning algorithm for both regression and classification problem (Shah et al. 2020). It works by finding the K nearest points (closest k data points) in the training dataset and uses that proximity to make classifications or predictions about the data set. By default, k-nearest-neighbor algorithm does not learn by iterations (epochs). KNN returns an averaged prediction based on an observation's k nearest (most similar) neighbors (Tripathy et al 2022; Lin and Wu 2023). Radial basis function (RBF) rely on radial function whose value depends on the distance from the point that is not necessarily the point of origin ($x_a = 0$ and $x_z = 0$). RBF kernel (a transformation from one space to another) is similar to K-Nearest Neighborhood Algorithm (Zhang et al. 2020). Support Vector Machine (SVM) is supervised machine learning frequently used for solving classification problems. SVM had been adjusted to solve a regression problem through the use of the Support Vector Regression algorithm (SVR) (Adnan et al. 2022). Decision Trees (DT) could be described as a supervised learning which is used in statistical analysis, data mining and machine learning for the purpose of solving regression and classification problems. DT can be categorical or continuous variable Decision Tree. DT uses a flowchart-like to make decisions based on input data by subjecting data into branches and assigns outcomes to leaf nodes to show the predictions.

Methodology

Stacking generalization (an extended form of the model averaging ensemble technique in which all sub-models equally participate as per their performance weights and build a new model with better predictions) procedure was used for this study (Silbert and Kopelowitz 2020; Zhao et al. 2023). New model was obtained due to stacking of the initial models. Stacked up model includes original (training) data, primary level models, primary level prediction, secondary level model and final prediction (Figure 1).

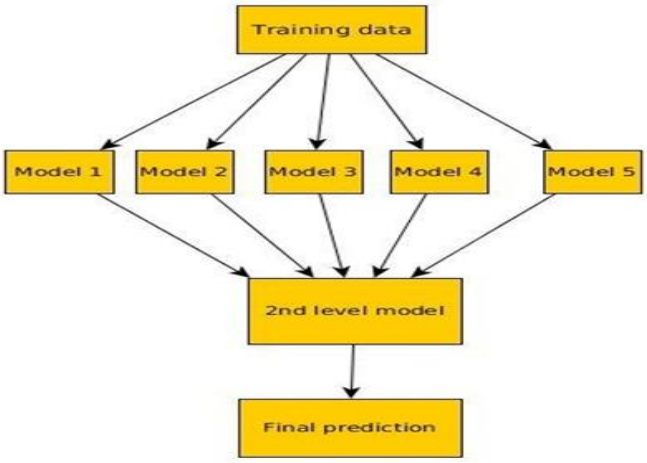

Figure 1: Stacking ensemble method

Prediction of Sale through Different Channel and Period

Sale dataset consist of sale in-person, during COVID-19, nation holiday and online ware obtained from Kaffi store. The dataset was subjected to pre-processing, classification, validation and various accuracy test using python programming in jupyter notebook based on Anaconda3 for development environment. All needed libraries were set in and data frame shape view gave 657 rows × 6 columns (Table 1). Predictions for stack of K-Nearest Neighbors (KNN), Support Vector Machine Enabled Radial Basis

Function (SVM-RBF), Decision Trees (DT), Random forest (RF), Multilayer perception (MLP) using sale dataset were conducted and result presented. Using SALESCHANNELPERIOD as target variable and 0.35 astest_size the X_train.shape, X_test.shape gave (427, 5), (230, 5) respectively while the y_test.value_counts() reveald the following (listing 1)

Listing 1: y_test.value_counts

```
Inperson_N.movement      79
whatsapp_N.movement      66
Instagram_N.movement     38
Inperson_Covid-19        28
whatsapp_Covid-19         8
Inperson_election         6
whatsapp_election         5
Name: SALESCHANNELPERIOD, dtype: int64
```

Results and Discussion

Table 1 showed the result of data frame shape view which gave 657 rows × 6 columns. The results indicates that the dataset was a low to id range dataset capacity.

Table 1: Preprocessed Kaffi dataset view

INDEX	ITEM	QTY	TIME	MARKETPRICE	CUSTOMERPRICE	SALESCHANNELPERIOD
0	0	1	0	7000	6500	whatsapp_election
1	1	3	0	45000	42000	Inperson_Covid-19
2	1	1	0	12000	12000	Inperson_Covid-19
3	1	2	1	6000	6000	Inperson_N.movement
4	1	1	1	12000	12000	Inperson_N.movement
□	□	□	□	□	□	□
652	26	2	0	24000	24000	whatsapp_N.movement
653	26	1	1	12000	11000	whatsapp_N.movement
654	26	1	0	10000	10000	whatsapp_N.movement
655	26	2	0	24000	23000	whatsapp_N.movement
656	26	1	1	12000	12000	whatsapp_N.movement

657 rows × 6 columns without INDEX

The results of predictions for stack of K-Nearest Neighbors (KNN), Support Vector Machine Enabled Radial Basis Function (SVM-RBF), Decision Trees (DT), Random forest (RF), Multilayer perception (MLP) using sale dataset (Table 2) showed that KNN, SVM-RBF, DT, RF, MLP, and the combined model had accuracy score of 62.99 %, 91.57 %, 52.45 %, 89.93 %, 33.02 % and 49.41 % respectively; the Matthews correlation coefficient (MCC) gave 50.84 %, 88.85 %, 34.76 %, 86.61 %, 6.12 % and 29.95 % respectively; F1 score also gave 59.92 %, 91.24 %, 44.83 %, 89.58 %, 19.58 % and 43.59 % respectively.

Table 2: Result of predictions for stack of K-Nearest Neighbors (KNN), Support Vector Machine Enabled Radial Basis Function (SVM-RBF), Decision Trees (DT), Random forest (RF), Multilayer perception (MLP)

Model	**Accuracy (%)**	**MCC (%)**	**F1 (%)**
knn	63	50.84	59.92
svm_rbf	91.57	88.85	91.24
dt	52.46	34.76	44.83
rf	89.93	86.61	89.58
mlp	33.02	6.12	19.58
stack	49.41	29.95	43.59

Conclusion

From the results, it could be concluded that combining models for machine learning operation does not necessarily give higher

performances in all score category. The confidence is on individual model to accurately learn on the dataset or from neighbor model for prediction.

References

Adnan, R. M., Kisi, O.; Mostafa, R. R., Ahmed, A. N. and El-Shafie, A. (2022). *The potential of a novel support vector machine trained with modified mayfly optimization algorithm for stream flow prediction*, Hydrological Sciences Journal 67 : 161-174.

Buyrukoğlu, S. and Savaş, S. (2023). *Stacked-based ensemble machine learning model for positioning footballer*, Arabian Journal for Science and Engineering 48 : 1371-1383.

Deepa, V., Sudar, K. M. and Deepalakshmi, P. (2018). *Detection of DDoS Attack on SDN Control plane using Hybrid Machine Learning Techniques*, 2018 International Conference on Smart Systems and Inventive Technology (ICSSIT) : 299-303.

Li, F.; Zheng, H., Li, X. and Yang, F. (2021). *Day-ahead city natural gas load forecasting based on decomposition-fusion technique and diversified ensemble learning model*, Applied Energy 303 : 117623.

Lin, W. and Wu, S. (2023). *Predicting S&P 500 Index ETF (SPY) During COVID-19 via K-NearestNeighbors (KNN) Algorithm.*, Journal of Accounting & Finance (2158-3625) 23.

Mojrian, S., Pinter, G., Joloudari, J. H., Felde, I.; Szabo-Gali, A., Nadai, L. and Mosavi, A. (2020). *Hybrid Machine Learning Model of Extreme Learning Machine Radial basis function for Breast Cancer Detection and Diagnosis; a Multilayer Fuzzy Expert System*, 2020 RIVF International Conference on Computing and Communication Technologies (RIVF) : 1-7.

Ribeiro, M. H. D. M. and dos Santos Coelho, L. (2020). *Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series*, Applied soft computing 86 : 105837.

Sarangi, P. K., Verma, R., Inder, S. and Mittal, N. (2021). *Machine Learning Based Hybrid Model for Gold Price Prediction in India*, 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO) : 1-5.

Seretis, A. and Sarris, C. D. (2022). *A Hybrid Machine Learning-Based Model for Indoor Propagation*, 2022 16th European Conference on Antennas and Propagation (EuCAP) : 1-5.

Shah, K.; Patel, H.; Sanghvi, D. and Shah, M. (2020). *A comparative analysis of logistic regression,random forest and KNN models for the text classification*, Augmented Human Research 5 : 1-16.

Shahhosseini, M., Hu, G. Huber, I. and Archontoulis, S. V. (2021). *Coupling machine learning andcrop modeling improves crop yield prediction in the US Corn Belt*, Scientific reports 11 : 1-15.

Silbert, O., Peleg, Y. and Kopelowitz, E. (2020). *Model Agnostic Combination for Ensemble Learning*, arXiv preprint arXiv: 2006.09025.

Tripathy, D. S., Prusty, B. R. and Bingi, K. (2022). *A k-nearest neighbor-based averaging model for probabilistic PV generation forecasting*, International Journal of Numerical Modelling: Electronic Networks, Devices and Fields 35 : e2983. *basis function networks for reliability analysis*,IEEE Transactions on reliability 70 : 887-900.

Zhang, D., Zhang, N., Ye, N., Fang, J. and Han, X. (2020). *Hybrid learning algorithm of radial basis function networks for reliability analysis*, IEEE Transactions on reliability 70 : 887-900.

Zhao, J., Hosseini, S., Chen, Q. and Armaghani, D. J. (2023). *Super learner ensemble model: A novel approach for predicting monthly copper price in future*, Resources Policy 85 : 103903.