Machine Learning Algorithms for Predictive Analytics: A Review and New Perspectives

Conference Paper · July 2020

DOI: 10.37896/HTL26.06/1159

2 authors, including:

Dipti Theng
Raisoni Group of Institutions
39 PUBLICATIONS **234** CITATIONS

SEE PROFILE

Machine Learning Algorithms for Predictive Analytics: A Review and New Perspectives

Dipti Theng[1], Mahesh Theng[2]

[1]Assistant Professor, G H Raisoni College of Engineeing, Nagpur, Maharashtra, India

[2]Government College of Engineeing, Nagpur, Maharashtra, India

Abstract

Since predictive analytics is gaining popularity due to the huge amount of availability of real world datasets, the selection of an optimal predictive algorithm is an absolute necessity. Many predictive algorithms are commonly used for predictive analytics, but it is still a challenging task to choose the right algorithm for the given real world dataset and problem under study. This manuscript presents overview of three most popular machine learning algorithms for predictive analytics and their implementation result analysis on real world dataset. These algorithms where evaluated and analyzed by comparative performance metrics such as training time, accuracy, sensitivity, specificity, accuracy, area under curve and error. The purpose of the experimental study presented here is to provide clear guidance for selecting the appropriate predictive algorithm according to the requirement of the problem and the dataset under review in the real world.

Keywords: Machine learning, Predictive analytics, K nearest neighbor, SVM, Naive Bayes.

Abbreviations: NB, Naïve Bayes; SVM, Support Vector Machine; KNN, K-Nearest Neighbor.

I. INTRODUCTION

Predictive analytics are the most common and widely used method of analyzing current and historical data to forecast future events and/or unexpected circumstances with a significant degree of precision [1-3]. This involves a wide range of analytical techniques including data mining, statistical analysis and machine learning. Machine learning a branch of artificial intelligence, which automates the creation of statistical and analytical models allowing systems to learn from information, recognize patterns and make predictions with minimal human involvement. The most popular and widely used machine learning algorithms for predictive analytics are logistic regression, K-nearest neighbor, SVM, decision trees, random forest, and Naive Bayes [4-5]. Many real world application areas demanding machine learning predictive analytics includes health care, industrial solutions, finance, agriculture, education, social media, cyber security, text mining, etc. [2, 6-8]. Machine learning approaches can incorporate large amounts of information into robust predictive analytics architecture; yet do not have any of the drawbacks and limitations of conventional modelling approaches [5].

The manuscript is organized as follows: Section 2 addresses predictive analytical methods, Section 3 presents the implementation of the top three predictive algorithms on the real world dataset, Section 4 presents comparative analytics for results obtained using standard performance

metrics, and Section 5 concludes the manuscript with strong remarks on applications of studied predictive algorithms.

II. RELATED WORK, METHODS AND DATASET

The related research work undertaken by various researchers discussed first in this section, and their most relevant findings have been noted. This will direct the selection for implementation of appropriate predictive algorithms. Followed by this, it addresses the chosen strategies with their advantages and limitations. Finally, a brief overview of the dataset considered for implementation is given.

A. Related work

Authors [1], presented research on opportunities and challenges of predictive analytics to search for targeted consumers, pick marketing strategies for more effective marketing and social media analysis. In [5], intraoperative predictive analytics is performed using machine learning techniques based on existing data from the electronic health record. Authors have implemented eight best machine learning algorithms for the prediction task of post-induction hypotension, and the additional parameter tuning is performed using the best predictor amongst them, before application to the test set. Experimental results found that, in both the training and testing process, gradient boosting exhibited clear discrimination over all other approaches.

In [6], addresses rising prediction issues in the healthcare sector that can be effectively solved using machine learning algorithms. Few to mention as, predictions of cardiovascular diseases, predictions of diabetes, predictions of hepatitis and cancer, etc. The authors have identified common machine learning algorithms used for these predictive tasks as CART, Naïve Bayes, RBF, SVM, Simulated Annealing (SA), C4.5, and ID3. In [7], authors provided a method for identifying production data to be analysed and implemented a predictive analytic model using neural network back propagation to predict sustainability efficiency, specifically power consumption. Some constraints can be considered in future work were discussed, such as the full integration of analytical modelling with big data systems and the extension of implementation to many other significant sustainability outcomes other than power consumption.

B. Machine learning methods for predictive analytics

Widely used machine learning algorithms for predictive analytics tasks noted from above discussion are Naïve Bayes, SVM, and K-nearest neighbours. With the respective predictive analytics tasks, these are studied briefly and discussed as below:

Naïve Bayes (NB): Under the simplest Bayesian network model, it is a "probabilistic classifiers" known as a supervised machine-learning algorithm, which implements theorem of Bayes with naïve assumptions of independence between the attributes. Naïve Bayes classifiers have the advantage of being highly scalable, enabling the number of attributes in a learning problem to be symmetrical to a number of parameters. It was first incorporated into the processing of text and considered a baseline tool for categorizing text and documents, such as scam or licit, with word frequencies as the attribute, and also finds applications in automated medical diagnostics. In [10], authors have applied Naïve Bayes classification algorithm to build a predictive analytics model

in the healthcare, and the results produced have shown that the algorithm operates correctly over the clinical data.

Support Vector Machine (SVM): It is a supervised machine-learning model that analyse the data used for the analysis of binary classification and regression tasks. In [11], uses SVM to construct a predictive model for faculty performance evaluation using multiple kernel approaches, namely sigmoid, linear, polynomial, radial, and Pearson. Out of these, Pearson has shown great accuracy and further explored with new approach. Authors proposed a novel approach to SVM with Pearson VII function, a universal kernel function PUK which has the advantage of a strong mapping power, simplified model building process and is computationally efficient.

K-Nearest Neighbors (KNN): KNN is a simple non-parametric supervised machine-learning algorithm used to solve classification problems as well as regression problems. Implementation is easy which will simply create an imaginary boundary to identify the data and attempt to predict the nearest boundary line for new data points to come in, yet has a major drawback from becoming dramatically slow as the size of that data increases. In [12], authors applied KNN algorithm to forecast meteorological and financial data on histogram time series where Mallows and Wasserstein distance used to identify closest neighbors.

Table 1: Machine learning methods

Strategy	Advantages	Limitations
Naïve Bayes (NB)	If the presumption of independence exists then it operates more effectively than other algorithms. Highly scalable, and can support data that is both continuous and discrete.	Cannot integrate association between variables. Another drawback is the presumption of independent predictors that it is almost impractical to get a set of completely independent predictors in real world.
Support Vector Machine (SVM)	It has good performance in generalization when there is a clear margin of differentiation between categories. Also, more efficient for high dimensional spaces.	It is not optimal for large datasets and does not work very well when the target classes overlap.
K-Nearest Neighbour (KNN)	Simple to implement and understand. The algorithm is robust in search space and takes less time in computation	Works poorly when dataset size grows highly. It's a slow learner, so it takes more time to compute.

C. Dataset description

Algorithms selected for predictive analytics applied to the credit card dataset from finance sector. Dataset has taken from UCI machine learning repository [13]. Dataset characteristics described in below table:

Table 2: Dataset description

Dataset	Type	#Features	#Samples	#Classes	Ratio (instances-to-features)	Domain

Credit Card	Binary Class	24	30000	2	1250.0	Finance
Employee Retention	Binary Class	10	14998	2	1499.8	Industry
Housing Data	Binary Class	18	497	2	27.61	Finance

Default payment (Yes/No= 1/0) is the dependent variable out of 24 attributes, while remaining 23 are predictor variables.

III. IMPLEMENTATION AND RESULTS

This section presents implementation results and comparative analysis of NB, SVM, and KNN on credit card dataset described in previous section. This aims to predict users default payment based on their financial transaction patterns and history.

A. Predictive analytics process

Training set will significantly affect the reliability of an analytical model [9], so preparing a training dataset for machine learning based analytic processing is necessary. Predictive analytics follows the four stages as:

a. pre-processing step in which raw information is collected and pre-processed;
b. data transformation or feature selection phase in which the correct machine learning algorithm is applied to convert pre-processed information into a form (by selecting the most relevant information or attributes as per the problem requirement) that can be easily managed;
c. training phase in which the learning model is constructed using transformed data;
d. testing or prediction phase which uses the previously created learning model to report predictions.

B. Implementation

Training set will significantly influence the reliability of an analytical model [9], so preparing a training dataset for machine learning based analytic processing is necessary. Predictive analytics performed here, follows the four stages as described in fig-1. Three machine-learning algorithms NB, SVM, and KNN implemented by this strategy, each trained on dataset of genomic array as described in secntion-2.

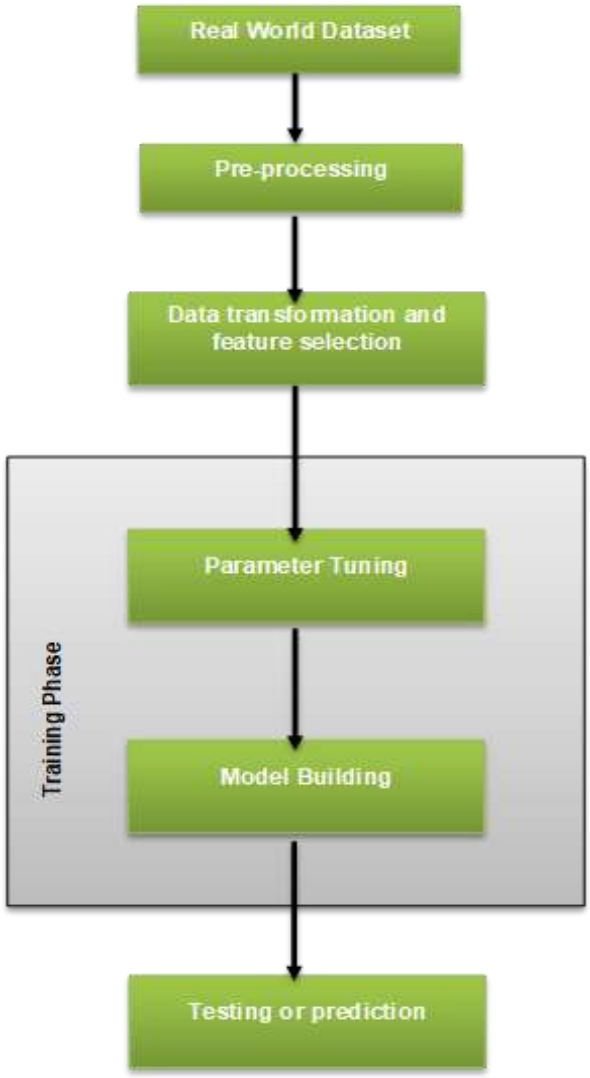


Fig. 1. Predictive analytics process using machine learning approach.

Experimentations carried out in Python programming environment, which is popular among data scientist and machine-learning researchers. It facilitates many pre-processing and data transformation libraries, which makes dataset understandable for learning task and easy for implementation.

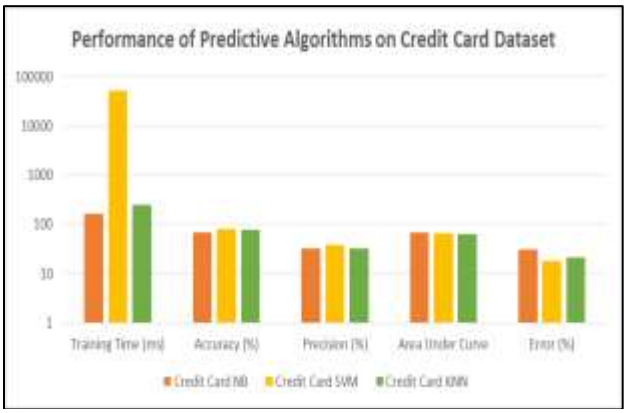
C. Comparative analysis

Implementation results of the predictive algorithms compared with respective various performance metrics is summarised in table-3. As per accuracy score, SVM outperforms over other algorithms. However, training time required for SVM learning algorithm is very high than the other algorithms. Naïve Bayes shows good performance on training time with equal precision power as to KNN.

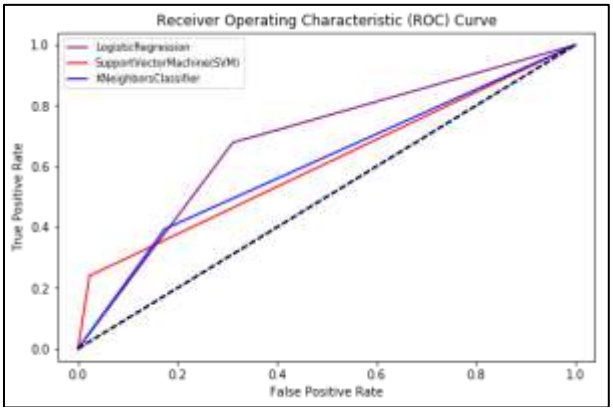
Table 3: Comparative analysis of prediction algorithms

Sr. No	Performance Metrics	Credit Card			Employee Retention			Housing Data		
		NB	SVM	KNN	NB	SVM	KNN	NB	SVM	KNN

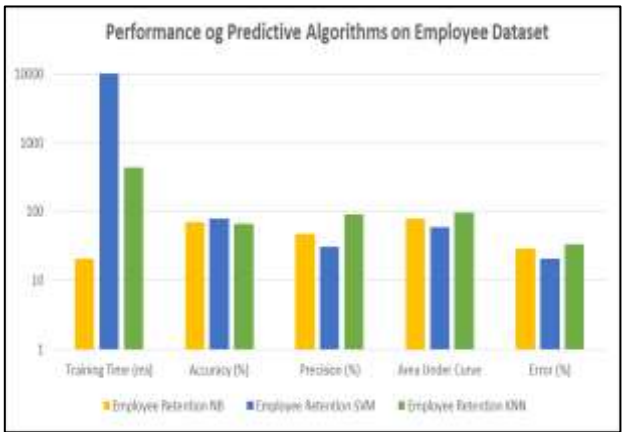
1	Training Time (ms)	167	51400	247	20.40	10000	438	31.20	15.60	3.99
2	Accuracy (%)	68.63	82.60	78.93	70.62	79.21	66.56	61.00	61.00	56.00
3	Precision (%)	32.00	38.00	32.00	47.00	31.00	90.00	53.00	52.00	48.00
4	Area Under Curve	68.35	64.92	62.91	78.00	59.00	97.00	63.00	61.00	56.00
5	Error (%)	31.37	18.25	21.07	29.38	20.79	33.44	39.00	39.00	44.00



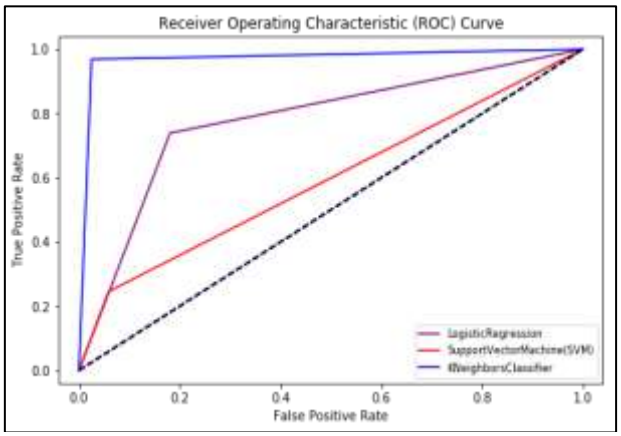
(a)



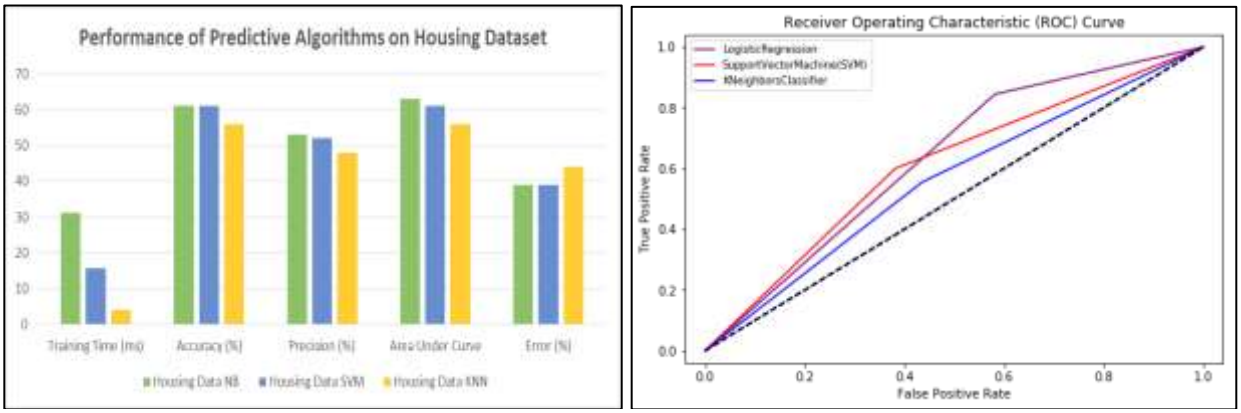
(b)



(c)



(d)



(e) (f)

Fig. 2. Comparative result analysis of predictive algorithms on various datasets

Analysis of algorithms using various performance metrics in table-3 is visually plotted as shown in fig-2.

IV. Applications of machine learning in predictive analytics

Some of the most talked about predictive analytics problems listed below as per problem statement area. For optimal solution, these still need in-depth research.

Table 4: Predictive analytics applications

Area	Tasks/Applications
Banking and Finance	Predict credit scoring. Fraud detection. Predict market risk. Consumer behaviour analysis and customer retention. Customer relationship management.
Health care and Medicine	Early disease prediction and monitoring. Personalised treatment/medicine suggestion. Telemedicine practices. Medical image analysis.
Industry	Environmental sustainability. Predicting cost of manufacturing. Cost effective material suggestion. Plant monitoring. Intellectual climate system for monitoring Industrial environment [14-15].
Entertainment	Content guideline and preferred genre prediction. Most popular content and genre. Genre categorization as per mood and other parameters.

Social media	Predict most influential profile. Sentiment analysis. Spam filtering. Social sustainability.
Education	Analysis of effective teaching methodology. Student's performance prediction. Personalised teaching methodology as per students' performance. Faculty performance evaluation.

V. CONCLUSION

This paper presented predictive analytics using machine learning and their applications on real world dataset. Various works by researchers on predictive analytics using machine learning approach is discussed by listing future scope for the further extension. Three most popular machine learning methods in predictive analytics are implemented and their comparative study is presented using standard performance metrics. Form the experimental results it is observed that SVM outperformed in terms of accuracy and precision over the other learning algorithms. Predictive analytics by SVM can be further explored on many real world problems including classification as well as regression analysis tasks. Many works have been done on kernel function study to further optimize the SVM performance.

VI. FUTURE SCOPE

Some questions still to be addressed in future work include comparative study of different function of the SVM kernel and providing optimal kernel solution with reduced training time and improved accuracy ranking. Another important future research work is to try an ensemble of these best performing predictive algorithms in order to take advantage of each into the final result. Ensemble learning is new future trend in demand because of its simplicity and easy understanding.

References

[1]. Mishra, N., & Silakari, S. (2012). Predictive analytics: A survey, trends, applications, oppurtunities & challenges. International Journal of Computer Science and Information Technologies, 3(3), 4434-4438.

[2]. Kelleher, J. D., Mac Namee, B., & D'arcy, A. (2015). Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies. MIT press.

[3]. Barga, R., Fontama, V., Tok, W. H., & Cabrera-Cordon, L. (2015). Predictive analytics with Microsoft Azure machine learning. Berkely, CA: Apress.

[24. Rajeshkanna, A., Preetha, V., & Arunesh, K. (2019, March). Experimental Analysis of Machine Learning Algorithms in Classification Task of Mobile Network Providers in Virudhunagar District. In International Conference on E-Business and Telecommunications (pp. 335-343). Springer, Cham.

[5]. Kendale, S., Kulkarni, P., Rosenberg, A. D., & Wang, J. (2018). Supervised Machine-learning Predictive Analytics for Prediction of Postinduction Hypotension. Anesthesiology, 129(4), 675-688.

[6]. Nithya, B., & Ilango, V. (2017, June). Predictive analytics in health care using machine learning tools and techniques. In 2017 International Conference on Intelligent Computing and Control Systems (ICICCS) (pp. 492-499). IEEE.

[7]. Shin, S. J., Woo, J., & Rachuri, S. (2014). Predictive analytics model for power consumption in manufacturing. Procedia Cirp, 15, 153-158.

[8]. Lin, J., & Kolcz, A. (2012, May). Large-scale machine learning at twitter. In Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data (pp. 793-804).

[9]. Ratner, B. (2017). Statistical and Machine-Learning Data Mining: Techniques for Better Predictive Modeling and Analysis of Big Data. CRC Press.

[10]. Fernando, Z. T., Trivedi, P., & Patni, A. (2013, August). DOCAID: Predictive healthcare analytics using naive bayes classification. In Second Student Research Symposium (SRS), International Conference on Advances in Computing, Communications and Informatics (ICACCI'13) (pp. 1-5).

[11]. Deepak, E., Pooja, G. S., Jyothi, R. N., Kumar, S. P., & Kishore, K. V. (2016, August). SVM kernel based predictive analytics on faculty performance evaluation. In 2016 International Conference on Inventive Computation Technologies (ICICT) (Vol. 3, pp. 1-4). IEEE.

[12]. Arroyo, J., & Maté, C. (2009). Forecasting histogram time series with k-nearest neighbours methods. International Journal of Forecasting, 25(1), 192-207.

[13]. https://archive.ics.uci.edu/ml/datasets.

[14]. Satpute, P. C., & Theng, D. P. (2013, April). Intellectual climate system for monitoring Industrial environment. In 2013 Third International Conference on Advanced Computing and Communication Technologies (ACCT) (pp. 36-39). IEEE.

[15]. Nirbhay Bhuyar , Samadrita Acharya , Dipti Theng, 2020, Crop Classification with Multi-Temporal Satellite Image Data, International Journal of Engineering Research & Technology (IJERT) Volume 09, Issue 06 (June 2020).