IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 20, 2020, accepted August 3, 2020, date of publication August 13, 2020, date of current version August 25, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3016419

A Novel Technique for Behavioral Analytics Using Ensemble Learning Algorithms in E-Commerce

MOHAMMAD ALOJAIL AND SURBHI BHATIA
Department of Information Systems, College of Computer Sciences and Information Technology, King Faisal University, AL Hufuf 36362, Saudi Arabia
Corresponding author: Surbhi Bhatia (sbhatia@kfu.edu.sa)

This work was supported by the Deanship of Scientific Research at King Faisal University for the financial support through Ra'ed Track under Grant 207006.

ABSTRACT The era of E-commerce and availability of data in every field of operations in an enormous volume that implies to Big Data is one of the biggest sources of competitive advantage for the organizations in this digital world. It provides useful information to grow businesses by posting the advertisement and help consumers find the relevant product according to their preferences. The focus of this research is on the advertisement strategies analysis under which a Business employs several online advertisement strategies in order to appeal to the consumer. This research work will present a detailed analysis in user behavioral to use for business or Online Behavioral advertising and provide the framework of how Enterprise Resource Planning systems track the targeted audience and show their content. The paper's prime objective is to classify and effectively run targeted advertising using the data that shows user's retail behavior. This is where an Enterprise Resource Planning driven data will give rise to behavioral analytics. In addition to this, various data streaming technologies are also emphasized that will help to create a pipeline for the huge amount of data in Enterprise Resource Planning's database.

INDEX TERMS Behavioral analytics, ERP, CRM, OBA, business, ensemble learning.

I. INTRODUCTION

With the increase in the relevance and effectiveness of ads and marketing messages displayed to online audiences, publishers and advertisers have taken advantage of behavioral data for both advertising and personalization. Online advertising [1] can potentially reach anywhere and everywhere, on all sides of the world. The billions of potential customers or clients can be attracted because of the modern technology. Advertising increases the visibility, sales and gains profits in the market [2]. It spreads awareness of the products and services to the wider audience. Based on the cost and the infinite amount of advertising that bombards everyone at all times, one safe assumption is advertising is an effective way for companies to attract customers [3]. The problem of identifying whether the customer is a potential buyer or not is equally important for the service providers. Innovation in Information Systems and Analytics [4] has influenced significantly the business organizations in a better way by providing solutions to the decision-making problems. Business intelligence and analysis of huge information is the need of the hour to expand businesses where Enterprise Resource Planning (ERP) is governing/managing the relationship of customers through wide interconnection of networks. Therefore, in this paper analysis of the data fetched from different sources is achieved using machine-learning algorithms. The analytics model used in the paper (we shall refer to it as "M" from now on for brevity), will work on the data that is being streamed from the ERP database. Primarily, the task is to analyze and classify the users based on their Sales or Search behavior. The following example is described for the analytics model. This task is again broken to a basic sub-class of "hot buyer" and "cold buyer". "The criteria for a profile to be classified is used as a "hot buyer" if both searches and purchases of a particular user are $>$ 100 per month. Similarly, a "cold buyer" is a profile that has searches and purchases $<$ 100 per month".

Data, in today's business and technology world, is indispensable [5]. To reach the full potential of one's business, it is very important to understand the attitudes and behavior of audience. The focus of this research is on the advertisement strategies analysis under which a business employs several online advertisement strategies in order to appeal to

The associate editor coordinating the review of this manuscript and approving it for publication was Zhiwu Li.

the consumer [6]. Examples of what a behavioral ad tries to find out about the consumer includes the consumer's age, gender, and purchase interests. Therefore, in this paper, we will extensively and solely only use the data from the sales and distribution process. The paper will focus on behavioral advertisement analysis as mentioned above, as in individual's preferences, buying habits, or interests. Any Business & E-Commerce Sales and Marketing Strategies consist of several things like Click Fraud, Procedures for Handling Electronic Financial Transactions, Bitcoins, and Driving Business through Meta Tags, Anti-cybersquatting Consumer Protection Act, Search Engine Optimization, and advertisement strategies and so on. Therefore, fetching data from ERP databases needs to be handled sensitively. Identically, issues may arise in continuous data fetching from ERP databases, which may be perceived as an attack, and thus requires exploration into new technologies to do the same. ERP databases are secure and need an API access system to fetch data.

In case of ERP or Customer Relationship Management (CRM) systems, there is a well-defined database for all transactions that take place in a system and user details are readymade, but these databases; however are extremely reactive to any kind of external access. To develop immunity of these databases against hacking attacks such as DDoS, RUDY etc. [7], precautions must be taken. The novelty of this paper is identified in using some data streaming technologies/APIs to get the data for analysis in the model discussed in section 3 in real-time, instead of running a script that periodically manually downloads data from the ERP/CRM database. The prime objectives of this research is to study the existing methods of analyzing behavior online using AI techniques and analyze the outcomes. To propose the analytics model discussed above (referred to as ''M'' from now on for brevity), that will work on the data that is being streamed from the ERP database. To analyze and classify the users based on their Sales or Search behavior, which can be taken as multiple layer binary classification problem. To implement the algorithm for above classification task using ensemble learning. To investigate by applying various permutations and combinations of ensemble algorithms and the one with the highest accuracy will be selected for result as classification. To analyze the effectiveness of the proposed algorithm and findings.

A. MOTIVATION AND CONTRIBUTIONS

The tremendous opportunities associated with data and analysis in different organizations have helped generate significant interest in information analytics, that analyze critical business data to help an enterprise better understand its business and market and make timely business decisions. The need of the hour is to perform behavioral advertisement analysis to track individual's preferences, buying habits, or interests so that Business employs several online advertisement strategies in order to appeal to the consumer. ERP databases are secure and need an Application Programming Interface (API) access system to fetch data. Along with this, issues may be faced in continuous data fetching from ERP databases may also be perceived as an attack, and thus requires exploration into new technologies to do the same. The very idea of the paper is to define a way of using Ensemble-of-Ensembles for the process. The virtue is in the technique used and is explained in the proposed online model. Online model" means a model that is always hot/reactive to inputs from the process. This model doesn't need any external handling to process the data.

The main aim is to create a Behavioral Analytics model (say M) based on data fed into an ERP system and use it to drive targeted advertisement. It will create a stream from ERP Database to M Database using technologies like Kafka, Pubnub, Pusher etc. The paper's prime objective is to classify and effectively run targeted advertising using the data that shows user's retail behavior by using machine-learning techniques.

The man contributions of this research work are:

- To study the existing methods of analyzing behavior online using AI techniques and analysis the outcomes.
- To propose the analytics model discussed above (referred to as ''M'' from now on for brevity), that will work on the data that is being streamed from the ERP database.
- To analyze and classify the users based on their Sales or Search behavior, which can be taken as multiple layer binary classification problem.
- To implement the algorithm for above classification task using ensemble learning.
- To investigate by applying various permutations and combinations of ensemble algorithms and the one with the highest accuracy will be selected for result as classification.
- To analyze the effectiveness of the proposed algorithm and findings.

B. ARTICLE ORGANISATION

The paper has been structured as follows. Section 2 defines the basic concepts and background knowledge with the literature survey. Section 3 explains the Behavioral Analytics Model with the details of the algorithms used for classification. Section 4 analyze the results and evaluate them on certain performance measures and Section 5 concludes the paper.

II. BACKGROUND

Behavioral analytics is none other than a tool to predict the actions taken by the user [8]. It becomes extremely important when an essential quantity of highly sensitive data to be handled. The term ''Big Data'' [9] means to the extremely huge volumes of data generated and available online in digital ecosystem. Applying analytics to this data enables organizations to harness their data and use it to identify new business opportunities. This will help businesses to make smarter and quicker business moves, improved performance, higher profits and happier customers [10]. The raw event data that is gathered on clicking of the user, swiping, and surfing through websites or applications is defined as behavioral data [11].

A. ENTERPRISE RESOURCES PLANNING

The ERP expands to Enterprise Resource Planning is a software suite that allows business and organizations to smoothly run their processes. ERP is used for storing and running business logic in innumerable sectors. Various modules of an ERP are defined by the particular job it does such as Sales and Distribution Module, Human Resource Module, Asset allocation module etc. There can several classes for these processes such as Sales and Distribution Process, Human Resource Development process, Financial Process. These processes can be broadly divided into the following classes. This research will focus on only Sales and Distribution model. It contains various modules as per the business hierarchy levels so that the operations are monitored and calculated correctly as per the flow of the business in the organization. There is no particular pattern that how an ERP works. It will work according to the setups that have been done according to the organization business flow. Sometimes there is need of manipulation in the ERP to attain some custom process which is different from standard process [12]. ERP is just countable on the events that have already occurred, it does not reflect on the occurring events or the future of how the events will happen in near future. Predictive analytics in ERP systems will help in identifying the possible future outcomes or the probability of the event happening. In addition to this, adding analytics to ERP systems will improve efficiency of the production, better risk reduction and fraud detection and Targeted, personalized marketing campaigns [13]. The effective decision making depends on collecting structured/unstructured data commonly referred to as big data [14] from different sources mainly accessed from web, blogs, documents, sensors, call center, surveys etc.

B. ONLINE BEHAVIOURAL ADVERTISING

For OBA is a new, yet interesting fact that tracks people's behavior online and gives benefits to advertisers, consumers, policy makers and scholars. To define OBA more clearly and precisely, it can referred to as ''online profiling'' and ''behavioral targeting'' [12]. The examples may include 'tracking early history through surfing can adjust advertisements' [13] or can be taken as 'a novel personalization technique that help advertisers to post most suitable advertising messages to respective customers through technology' [14]. It is a kind of advertising network, which counts number of times a consumer pays visits to the website [15]. For example the network will identify the theme (entity) in which the person is interested to view (surf through websites). If the person is paying more visits to a website of mobile phones (say IPhone), the network will start displaying more advertisements and images related to phones will be displayed. The two different persons visiting the same website will be shown different advertisements based on the interest by the network (by his findings). According to the statistics shown in the paper [16], the revenues of online advertising is growing at an excellent pace and is going to be on its peak in the future coming years. The boon of OBA has clearly reflected in building records every year [17], [18]. It has been clearly stated that OBA is the prime component that displays the most relevant (boosting up) advertisements according to the interest of the user [19], [20]. Meanwhile, considering the advantages and benefits received to the companies, there is a concern regarding privacy also. Collection of personal data, sharing of personal data connectively raises issues of consumer's privacy. Therefore, US Federal Trade Commission [21] and other Authorities of Europe [22] have imposed some policies and self-regulatory plans to protect consumer privacy in making them aware about displaying and using personal data. As OBA is the future of advertising, advertisers can choose to use and go for more selective targeting [23], [24].

C. ENSEMBLE LEARNING

The Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. There are two categories of machine learning algorithms, supervised and unsupervised algorithms. [25]. Many different algorithms come under broad topic of machine learning. The grouping of these algorithms can be done by their learning style or by their similarity in form or function. Algorithms that are often grouped by similarity in terms of their function (how they work) are mainly tree-based methods and neural network inspired methods [26]. The list continues to regression algorithms, deep learning algorithms, instance based methods, regularization algorithms, Bayesian and clustering algorithms, Association rule learning algorithms, dimensionality reduction algorithms and the list is endless.

Ensemble methods is like using the predictions of small expert models in different parts of the input space. Ensemble Learning comes under the same category where the task is to combine several base models in order to produce one optimal predictive model. Ensemble Learning is basically stacking of some simple or weak models and then combining the result of all models to generate a more powerful model and give an accurate answer. To elaborate, Ensemble means group of musicians. Simple models are to be created. Let them be M_1, M_2,…,M_k. These simple models are called base models. They can be anything like K Nearest Neighbours, Decision Trees or any other model. The combination of these base models into a more accurate model i.e. combining positives of all base models is the objective of ensemble learning. There are three main types of ensembles:

Bagging (Bootstrapped Aggregation): Bootstrap Aggregating (Bagging) is an ensemble generation method [27] that be used with unstable classifiers, that is, classifiers that are sensitive to variations in the training set such as Decision Trees and Perceptron's. Random Subspace is an interesting similar approach that uses variations in the features instead of variations in the samples, usually indicated on datasets with multiple dimensions and sparse feature space.

Boosting: Boosting generates an ensemble by adding classifiers that correctly classify ''difficult samples''. For each

iteration, boosting updates the weights of the samples, so that, samples that are misclassified by the ensemble can have a higher weight, and therefore, higher probability of being selected for training the new classifier. Boosting is an interesting approach but is very noise sensitive and is only effective using weak classifiers [28]. There are several variations of Boosting techniques AdaBoost, Brown Boost, each one has its own weight update rule in order to avoid some specific problems such as noise, class imbalance, etc.

Stacking: Stacking is a meta-learning approach in which an ensemble is used to ''extract features'' that will be used by another layer of ensemble. This method is very useful in real world also. Major understanding of ensemble learning lies in to identify the types of weak learners and to gather different ways in order to combine them.

D. RELATED WORK

The growth of data analytics and data mining techniques to improve quality, performance and decision-making; classifying the users on their sales or search behavior will help ERP systems in predicting the future events and integrating artificial intelligence with business modelling to make recommendations. The predictive analytics is an emerging field, which will help in growing the business. The authors [29] discussed the survey paper by making the researchers aware of how predictive analytics can be used for business and medical industry by listing the techniques and algorithms. The authors detailed the process for predictive analytics modelling but did not outlined the model that be used to initiate the process suitable for big data [30]. The authors [31] discussed the novel application framework by mapping the major domain areas as big data with predictive analytics. The paper discussed a novel architecture and discussed how analysis in predictive modelling helped to achieve in automating the operational decision making. The authors [32] elaborated two main issue faced in ERP dimensions for assessing ERP data quality and the causal relationships between these data quality dimensions. The paper [25] discussed the basics and mathematical theory behind ensemble learning classifiers and concluded that Ensembles can work well as accurate classifiers. The authors [33] discussed the how the ERP systems will become self-moderated by increasing their availability. A framework is developed that works on addressing the post-implementation challenges. The paper highlighted how businesses can benefit by implementing enterprise systems that next-generation breed of ESs that works on catalyzing business performance. The authors [34] carried a project and explored the strategies project managers in the energy sector used to implement successful ERP projects. A brief is given on how the involvement of the stakeholders, strategic practices employed and followed, and other strategies for improvement are carried in ERP systems. The paper [35] explained the diversity of classifiers at a desire level, and how to utilize different learning schemes to build the base classifiers of ensemble. The authors proposed a stacking framework with lazy local learning for building a classifier ensemble learner. The research in the field of analyzing online behavior of customers using different techniques has been given in Table 1.

TABLE 1. Survey of papers.

Papers	Targeted Data	Research Method
Park et al. [36]	Self-reported user profiles and mobile call records	Network analytics, Anomaly detection, Predictive analytics
Chau and Jennifer [37]	User-generated content extracted from blog	User-generated content extracted from blogs, Text and network analytics, Community detection, Network visualization
Abbasi et al. [38]	Financial ratios, and organizational and industrial-level context features	Data analytics, Classification & generalization, Adaptive learning
Aguirre et al. [39]	Facebook data Types of information used included age, gender, location and interest.	Choice theory, psychological ownership theory, and psychological reactance theory.
Tucker et al.. [40]	Targeted and personalized data from Facebook	Improved privacy controls methods, Novel Implications for privacy regulations
Babu et al. [31]	User generated unstructured data from ERP Server	Supervised learning algorithms, Predictive Analytics

III. BEHAVIOURAL ANALYTICS MODEL

Behavioral analytics helps to understand not only what is happening, but also why and how is it happening. It allows understanding the user behavior [41] on website, which is very crucial for online businesses in today's scenario. The website can be seen from the users perspective and track their movements. The process starts with collecting the data, building the model, making predictions and validating the statistical model and revising it from other data sources. Machine learning is data-based. The technique is to classify requires the machine learning model-fitting process to find patterns in the data that precede that line of work. The basic steps involved in building predictive models using machine learning to solve the problem are: Collect examples from the past where events of type X could have happened, along with whether they actually happened or not. Try to construct a function that uses the given information available for each example, and tries to predict the outcome recorded for the example. It can be shown that trying to minimize the error in predicting the recorded outcomes for each example can give you the probability, within the confines of information

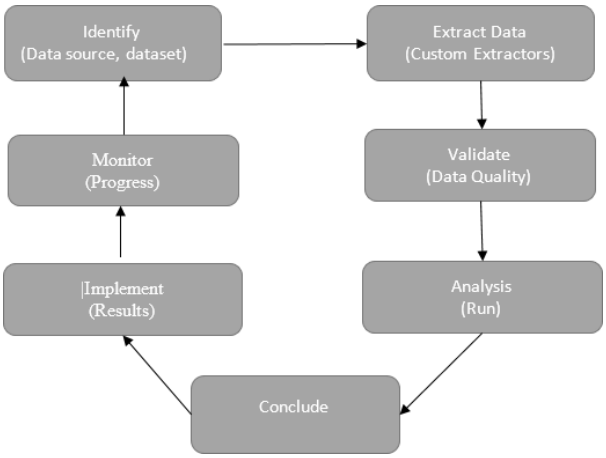

FIGURE 1. Analytical Model.

provided in the examples recorded. Building predictive models is an iterative process, which follows initial hypothesis to solve a business problem and then refine it until the expected outcome is achieved with more accurate results [42], [43]. The tasks elaborated above are explained in Figure 1 given below.

The analytics model used in the paper (referred to as ''M'' from now on for brevity), will work on the data that is being streamed from the ERP database. Primarily, the task is to analyze and classify the users based on their Sales or Search behavior. This task can be taken as a binary classification and first step can be done using basic linear regression.

The second step is to create a multi-class classifier to generate sector-based classification of hot or cold buyer, based on the above example. Say, profile can be marked as – ''hot buyer'' in cosmetics but ''cold buyer'' in electronics. Since, this takes a lot of classes into account, for the sake of more accuracy; help of ensemble learning techniques (bagging and boosting algorithms) is the third step to be used for accomplishing the task. A bag of various permutations and combinations of ensemble algorithms will be applied and the one with the highest accuracy will be selected for result as classification.

A. ENSEMBLE-BLENDING STRATEGY DRIVEN MODEL

This is a novel strategy to take advantage of various ensemble techniques to get better accuracy of the model prediction task; called as ''Ensemble-Blending based Ensemble model'' or ''EBE model'' in short. All ensemble techniques are application of weak learners on bootstrapped datasets from the original dataset, in EBE model. The algorithm for the above classification task using ensemble learning is explained in steps.

- Feed dataset – D.
- Split D into test_D and train_D
- Run Boosting, Bagging, Stacking, GBM and RandomForest classifier models on train_D. Store values as model_bag, model_boost, model_stack, model_gbm and model_rf.
- Predict values by running model_bag, model_boost, model_stack, model_gbm and model_rf with test_D.
- Measure accuracy of each model – store as acc_bag, acc_boost, acc_stack, acc_gbm and acc_rf.
- Measure accuracy of each model – store as acc_bag, acc_boost, acc_stack, acc_gbm and acc_rf. Say, the models are ordered in terms of highest to lowest accuracy as – model_gbm > model_stack > model_boost > model_bag > model_rf.
- All the models will then be blended with each other in a weighted manner, with model_gbm taking the highest weight and model_rf taking the least, but not 0

The choice of Bagging, Boosting, Stacking, GBM and Random Forest is taken randomly by testing and tuning the parameters. The merit of the model is in its idea, configurations are flexible. Models are first sorted in terms of their accuracy, say models are M1, M2 and M3, and M1 has the highest accuracy and M3 has the lowest. So, the blending process is done as: a X M1 + b X M2 + c X M3, where a>b>c. The data is mixed/blended on the basis of higher intrinsic accuracy of each model.

In the proposed model, evaluation of the analysis accuracy is done using Boosting, Bagging, Stacking, GBM and Random Forest.

Boosting is a method of using weak learners used to make models by sequentially running the initial dataset on the learners.

Bagging is a model of using weak learners on bootstrapped dataset from the initial dataset in parallel and then combining the results.

It is important to understand that weak learners used both with Bagging and boosting are homogenous models. Important fact to remember is boosting is a sequential model and bagging is a parallel model.

Stacking is a method of using heterogeneous weak learners for building a Meta model. So, a K nearest neighbor classifier and a Support Vector Machines is taken together to teach a neural network [44] for prediction.

GBM expands to Gradient Boosting Machines; this is also called MART (Multiple Additive Regression Trees). As the name suggests, it use gradient decent in principle and adds regression trees at each step to make the model reach closer to highest prediction.

Random forest is a model where bootstrapped samples with pre/user-defined features are run on deep trees. The results are then combined to form a forest. This method is very similar to that of bagging. Recent development in the field of machine learning used neural deep tree networks for random forests [45].

B. FRAMEWORK OF OVERALL DESIGN

The framework defined of the EBE model has three main components which are discussed below:

1) ERP MODULE

ERP is used for accessing the database. Most ERP databases use Structured Query Language databases and thus are

already set in a pre-defined format. This saves time and resource from the data cleaning tasks. The internal architecture of the ERP system is broadly divided into three major parts:

a: PRESENTATION LAYER

This is basically the Graphical User Interface (GUI) that a user sees and works on. This can be considered as a human machine interface between the user and the logic part of the system.

Application Layer: This layer contains all the business logics, the validation engines and other arithmetic units to effectively process and create insightful relations between the data that is fed through Presentation layer.

b: DATABASE LAYER

The data that is fed through presentation layer and all the results that is generated from processes from Application layer get stored securely in this layer. The data stored in a particular type of schema and has the ability to be accessed in a particular format or layout whenever required.

2) DATA STREAMING

Streaming data is simply a term used in big data for a data stream. Streaming data is usually dealt with or desired, as a real-time, or constant rate source. The logical limit of buffering, of course is called ''downloading'' which is usually not considered streaming. Streaming or media streaming is a technique for transferring data so that it can be processed as a steady and continuous stream.

The use of streaming is becoming extremely important with the popularity of Internet as users are unable to download large multimedia files quickly because of limited access. With streaming, the client browser or plug-in can start displaying the data before the entire file has been transmitted. The data is streamed in real-time from ERP to Analytics Engine. The reason behind this is to condition the EBE model based on the most recent buying trends of a buyer. Since, buying is an action driven by choice and thus the chance of change is very high. Keeping this in view, the proposed model has a streaming pipeline for data collection from ERP for further analysis.

A native apache spark streaming system is used in the paper as it can easily be integrated with an existing data science model. Since it is open source, anyone can replicate out data analysis model. This will increase chances of reproducibility.

3) MODEL

The model is the unit where the data from ERP is streamed and kept. The analytics engine is run for every μ number of data points. So, for example, an analytics model M is created based on μ number of data points. Next, when additional information is fed into the ERP database, say α data points, and it gets streamed into the Model, the database swaps α number of oldest data points with the new α data points in it. This is same as saying that with each new data point that

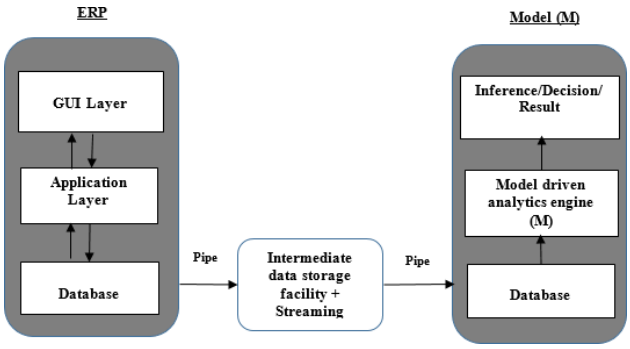

FIGURE 2. Framework Design.

enters the Model's database, the oldest data at that particular time becomes obsolete. In addition, the analytics model runs every time data points equal to 1/4th (this is configurable) of the original μ gets changed. This strategy is used to keep the output of the model relevant at all times. The detailed design of the model is shown in Figure 2.

The data is fed into the analytics model from the ERP database using a streaming API. This data after being received into the model is preprocessed. The formatted data shall then undergo the process as defined in section 3. After the analysis is done, the classified data shall then be pushed as a graphical result in the inference unit. This inference is then used for targeted advertisements. The flow of process is elaborated below.

- Data fed into ERP database.
- Data from ERP is streamed into Model, Let the data set is named as μ with β data points. These data points have a Unique Transaction ID (UTI). The UTI is stored in a separate table.
- This data is used for analysis and inference generation for targeted advertisements.
- The ERP is fed with new customer data.
- Data from ERP is streamed into Model, Let the data have α data points. These new data points will replace old data from μ, while continuing to keep the count of β constant.
- The UTI table constructed at Step 2, is kept as a check on the data updation of μ. As soon as the variation between the UTI of updated μ and stored UTI is more that 25%, the analytics engine is run again and the model is updated. In addition, the UTI table is also updated as per the new μ.

IV. RESULTS AND DISCUSSIONS

The work is based on two different datasets, one consists of samples from the Machine learning [46] (UCI) repository and the other a synthetic database of containing customer sales behavior. The dataset is not changing, however, the results of each blending process will change because each time different ensemble methods are employed in the process. The testing datasets, namely, Autos, credit_rating, pima_diabetes, iris and zoo are publicly available at UCI repository is a

TABLE 2. Information about Datasets.

Dataset	Instances	Attributes	Type	Creator
Autos	205	26	Multivariate	Jeffrey C.
Credit_rating	690	15	Multivariate	Confidential source
Japanese_ vowels	640	12	Multivariate, time series	Mineichi Kudo, Jun Toyama, Masaru Shimbo
Iris	150	4	Multivariate	R.A. Fisher
Zoo	101	17	Multivariate	Richard Forsyth
ERP data	182	12	Multivariate	Confidential source

TABLE 3. Results with RMSE.

RMSE						
Dataset/ Algorithm	Boosting	Bagging	Stacking	GBM	Random Forest	EBE
Autos	0.58	0.3	0.62	0.59	0.35	0.12
Credit_ rating	0.53	0.42	0.54	0.43	0.43	0.29
Japanese_ vowels	0.62	0.63	0.59	0.43	0.4	0.44
Iris	0.71	0.47	0.53	0.48	0.5	0.5
Zoo	0.51	0.51	0.59	0.3	0.3	0.23
ERP data	0.6	0.49	0.63	0.7	0.2	0.19

publicly available [47]. Brief information about each dataset is reported in Table 2.

Analysis was done on various ensemble methods with various standard datasets and ERP dataset. In addition to this, the comparison is done on the different datasets with the proposed EBE model. A few statistical tools like coefficient of determination also called as R^2 and Root mean square Error (RMSE) are used to evaluate the performance of the proposed model given in Equation 1..

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y} - y)^2}{n}} \tag{1}$$

The variation between the predicted and the actual value, represented as RMSE is shown in Table 3.

The variance is calculated on the predicted variable around its mean as Sum square of errors (SSE), Sum square of Regression (SSR), and Sum square of Total (SST).

Square root of difference between predicted and original data divided by total observations) is represented in Equation 2.

Mean SST = Original data – mean (of original) data

Mean SSR = Predicted data – mean (of original) data

TABLE 4. Results: R-squared error (SSE and SST).

R Squared Error (0 being the best, 1 being the worst)						
Dataset/ Algorithm	Boosting	Bagging	Stacking	GBM	Random Forest	EBE
Autos	0.69	0.4	0.51	0.5	0.45	0.25
Credit_ rating	0.63	0.6	0.62	0.63	0.5	0.3
Japanese_ vowels	0.75	0.7	0.51	0.7	0.4	0.33
Iris	0.8	0.45	0.59	0.58	0.5	0.45
Zoo	0.7	0.42	0.61	0.5	0.3	0.29
ERP data	0.8	0.41	0.62	0.6	0.2	0.23

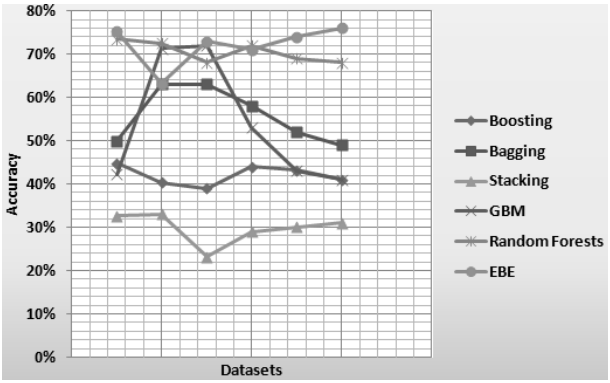

FIGURE 3. Evaluation Results wrt Dataset V/s Algorithms.

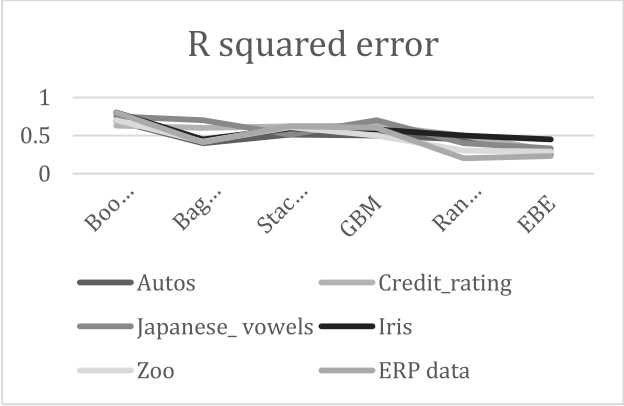

FIGURE 4. R Squared Error on EBE model.

Mean SSE = Original data – predicted data

$$\text{Mean } R^2 \text{ Errors} = \left(1 - \frac{SSE}{SST}\right) \tag{2}$$

The results are shown in Table 4 based on the analysis.

Each model was analyzed based on the above six functions and the conclusion was drawn accordingly.

Accuracy is the proportion of correct predictions for positive and negative class over the total number of observations. The difference between the actual or observed values and

TABLE 5. Results with accuracy.

Dataset V/s Algorithm – Accuracy of Prediction							
Dataset/ Algorithm	Boost-ing	Bagg-ing	Stack-ing	GBM	Random Forests	EBE	Data Source
Autos	45%	50%	33%	42%	74%	75%	https://archive.ics.uci.edu/ml/datasets/automobile
Credit_ rating	40%	63%	33%	71%	73%	63%	https://archive.ics.uci.edu/ml/datasets/credit+approval
Japanese_ vowels	39%	63%	23%	72%	68%	73%	https://archive.ics.uci.edu/ml/datasets/Japanese+Vowels
Iris	44%	58%	29%	53%	72%	71%	https://archive.ics.uci.
Zoo	43%	52%	30%	43%	69%	74%	https://archive.ics.uci.edu/ml/datasets/zoo
ERP data	41%	49%	31%	41%	68%	76%	-

predicted values for the selected model is small and unbiased for train, validation and test data sets

The performance is evaluated and the results are shown in Table 5 and the visual depiction is given in Figure 3.

The Table 5 clearly explains the success of EBE model in case of most datasets.

EBE model closely follows the original ensemble with highest accuracy; this establishes the EBE model mathematically. The R squared error depicting results is given in Figure 4.

V. CONCLUSION AND FUTURE SCOPE

The OBA is thus very important for the E-business concept and have shown a remarkable profit by integrating business and technology strategies. The paper defines OBA through ERP systems by developing a novel framework that uses classification model through streaming technologies. Through analyzing ERP data, this research has provided a better understanding how ERP systems can monitor the content as a means of transforming IT to gain profit among business organizations. The algorithm follows a structured approach and takes advantage of various ensemble techniques to get better accuracy of the model prediction task. To take advantage of overall goodness of various ensemble methods. EBE model is proposed. This model shows how ERP systems can track the targeted audience online through search and surf locally using different web browsers. The strategies used in this paper are driven by the idea of using existing technologies to build novel techniques that give better results. The experimental results have shown an accuracy of 76 % with ERP data using the proposed model. Extensive emphasis is put on the selection and analysis of existing ensemble techniques and emphasis was also put on developing various approaches that could be used to give better results, while keeping the underlying concept same. The merit of the model is in its idea where the configurations are flexible. As said by Melville *et al.*, 2004 [48] that value of IT business can be inculcated though venture of majorly three areas: Process, IT and people In this research mainly focus on the applications of Machine learning algorithms, ignoring the third dimension, i.e. 'the people'. The surveys including analytical personnel is equally critical to take into account for logical and analytical mindset, helping to achieve the business profit from big data analytics. The future research should take the analytical personnel into consideration to create more business value for ERP systems.

REFERENCES

[1] J. Zhang, Z. Wei, Z. Yan, M. Zhou, and A. Pani, "Online change-point detection in sparse time series with application to online advertising," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 49, no. 6, pp. 1141–1151, Jun. 2019.

[2] J. Aguilar and G. Garcia, "An adaptive intelligent management system of advertising for social networks: A case study of Facebook," *IEEE Trans. Comput. Social Syst.*, vol. 5, no. 1, pp. 20–32, Mar. 2018.

[3] R. Wang, Q. Gou, T.-M. Choi, and L. Liang, "Advertising strategies for mobile platforms with 'apps,'" *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 48, no. 5, pp. 767–778, May 2018.

[4] D. Chen, Z. Ding, C. Yan, and M. Wang, "A behavioral authentication method for mobile based on browsing behaviors," *IEEE/CAA J. Autom. Sinica*, early access, Jul. 18, 2019, doi: 10.1109/JAS.2019.1911648.

[5] T. A. Gerace and G. R. Barbour, "Computer method and apparatus for targeting advertising," U.S. Patent 10 510 043, Dec. 17, 2019.

[6] Y. Wang, L. Kung, and T. A. Byrd, "Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations," *Technol. Forecasting Social Change*, vol. 126, pp. 3–13, Jan. 2018.

[7] M. M. Najafabadi, T. M. Khoshgoftaar, A. Napolitano, and C. Wheelus, "Rudy attack: Detection at the network level and its important features," in *Proc. 29th Int. Flairs Conf.*, 2016, pp. 1–6.

[8] P. Zhao, Z. Ding, M. Wang, and R. Cao, "Behavior analysis for electronic commerce trading systems: A survey," *IEEE Access*, vol. 7, pp. 108703–108728, 2019.

[9] H. Wang, S. Ma, and H.-N. Dai, "A rhombic dodecahedron topology for human-centric banking big data," *IEEE Trans. Comput. Social Syst.*, vol. 6, no. 5, pp. 1095–1105, Oct. 2019.

[10] J. J. Sheng, J. Amankwah-Amoah, and X. Wang, "Technology in the 21st century: New challenges and opportunities," *Technol. Forecasting Social Change*, vol. 143, pp. 321–335, Jun. 2019.

[11] T. N. Hewage, M. N. Halgamuge, A. Syed, and G. Ekici, "Review: Big data techniques of Google, Amazon, Facebook and Twitter," *J. Commun.*, pp. 94–100, 2018.

[12] S. H. Sistla, and M. S. P. Babu, "Erp implementation for manufacturing enterprises," *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 3, no. 4, pp. 18–24, 2013.

[13] G. Shmueli and O. Koppius, "Predictive vs. Explanatory modeling in information systems research," in *Proc. Conf. Inf. Syst. Technol.* Hyderabad, India: Indian School Business, 2010, pp. 1–48.

[14] H. Zahid, T. Mahmood, A. Morshed, and T. Sellis, "Big data analytics in telecommunications: Literature review and architecture recommendations," *IEEE/CAA J. Autom. Sinica*, vol. 7, no. 1, pp. 18–38, Jan. 2020.
[15] E. G. Smit, G. Van Noort, and H. A. M. Voorveld, "Understanding online behavioural advertising: User knowledge, privacy concerns and online coping behaviour in Europe," *Comput. Hum. Behav.*, vol. 32, pp. 15–22, Mar. 2014.
[16] C.-D. Ham and M. R. Nelson, "The role of persuasion knowledge, assessment of benefit and harm, and third-person perception in coping with online behavioral advertising," *Comput. Hum. Behav.*, vol. 62, pp. 689–702, Sep. 2016.
[17] S. C. Boerman, S. Kruikemeier, and F. J. Zuiderveen Borgesius, "Online behavioral advertising: A literature review and research agenda," *J. Advertising*, vol. 46, no. 3, pp. 363–376, Jul. 2017.
[18] Internet Advertising Bureau. *First Quarter US Internet Ad Revenues Hit Record-Setting High at Nearly $16 Billion, According to IAB*. Accessed: Jun. 9, 2016. [Online]. Available: https://www.iab.com/news/first-quarter-u-s-internet-ad-revenues-hit-record-setting-high-nearly-16-billion-according-iab
[19] eMarketer. (2010). *How Should Marketers Address Concerns about Ad Targeting?* Accessed: Feb. 16, 2016. [Online]. Available: https://www.emarketer.com/Article/How-Should-Marketers-AddressConcerns-About-Ad-Targeting/1007514
[20] J. Chen and J. Stallaert, "An economic analysis of online advertising using behavioral targeting," *MIS Quart.*, vol. 38, no. 2, pp. 429–449, Feb. 2014.
[21] H. Beales, "The value of behavioral targeting," *Netw. Advertising Initiative*, vol. 1, no. 2010, pp. 1–25, 2010.
[22] M. K. Ohlhausen, "Privacy challenges and opportunities: The role of the federal trade commission," *J. Public Policy Marketing*, vol. 33, no. 1, pp. 4–9, Apr. 2014.
[23] Protection Regulation, "Regulation (EU) 2016/679 of the European parliament and of the council," *Regulation (EU)*, vol. 679, no. 2016, p. 209, 2016.
[24] K. L. Keller, "Unlocking the power of integrated marketing communications: How integrated is your IMC program?" *J. Advertising*, vol. 45, no. 3, pp. 286–301, Jul. 2016.
[25] S. Bhatia, M. Sharma, and K. K. Bhatia, "Sentiment analysis and mining of opinions," in *Internet of Things and Big Data Analytics Toward Next-Generation Intelligence*. Cham, Switzerland: Springer, 2018, pp. 503–523.
[26] S. Bhatia, M. Sharma, K. K. Bhatia, and P. Das, "Opinion target extraction with sentiment analysis," *Int. J. Comput.*, vol. 17, no. 3, pp. 136–142, 2018.
[27] R. K. Roul, A. Bhalla, and A. Srivastava, "Commonality-rarity score computation: A novel feature selection technique using extended feature space of ELM for text classification," in *Proc. 8th Annu. Meeting Forum Inf. Retr. Eval. (FIRE)*, 2016, pp. 37–41.
[28] A. N. Srivastava and J. Han, Eds., *Machine Learning and Knowledge Discovery for Engineering Systems Health Management*. Boca Raton, FL, USA: CRC Press, 2011.
[29] S. Poornima and M. Pushpalatha, "A survey of predictive analytics using big data with data mining," *Int. J. Bioinf. Res. Appl.*, vol. 14, no. 3, pp. 269–282, 2018.
[30] M. Ghahramani, M. Zhou, and C. T. Hon, "Mobile phone data analysis: A spatial exploration toward hotspot detection," *IEEE Trans. Autom. Sci. Eng.*, vol. 16, no. 1, pp. 351–362, Jan. 2019.
[31] M. S. P. Babu and S. H. Sastry, "Big data and predictive analytics in ERP systems for automating decision making process," in *Proc. IEEE 5th Int. Conf. Softw. Eng. Service Sci.*, Jun. 2014, pp. 259–262.
[32] A. Haug, J. S. Arlbjørn, and A. Pedersen, "A classification model of ERP system data quality," *Ind. Manage. Data Syst.*, vol. 109, no. 8, pp. 1053–1068, Sep. 2009.
[33] A. Elragal and H. E.-D. Hassanien, "Augmenting advanced analytics into enterprise systems: A focus on post-implementation activities," *Systems*, vol. 7, no. 2, p. 31, Jun. 2019.
[34] D. K. Arko, "Successful strategies for energy sector enterprise resource planning projects," Tech. Rep., 2019.
[35] H. Homayouni, S. Hashemi, and A. Hamzeh, "A lazy ensemble learning method to classification," *Int. J. Comput. Sci. Issues*, vol. 7, no. 5, p. 344, 2010.
[36] S.-H. Park, S.-Y. Huh, W. Oh, and S. P. Han, "A social network-based inference model for validating customer profile data," *MIS Quart.*, vol. 36, no. 4, pp. 1217–1237, 2012.
[37] M. Chau and J. Xu, "Business intelligence in blogs: Understanding consumer interactions and communities," *MIS Quart.*, vol. 36, no. 4, p. 1189, 2012.
[38] A. Abbasi, C. Albrecht, A. Vance, and J. Hansen, "Metafraud: A meta-learning framework for detecting financial fraud," *MIS Quart.*, vol. 36, no. 4, pp. 1293–1327, 2012.
[39] E. Aguirre, D. Mahr, D. Grewal, K. de Ruyter, and M. Wetzels, "Unraveling the personalization paradox: The effect of information collection and trust-building strategies on online advertisement effectiveness," *J. Retailing*, vol. 91, no. 1, pp. 34–49, Mar. 2015.
[40] C. E. Tucker, "Social networks, personalized advertising, and privacy controls," *J. Marketing Res.*, vol. 50, no. 5, pp. 546–562, Oct. 2013.
[41] L. Li, X. Lin, M. Zhou, and L. Fu, "Sociability-based influence diffusion probability model to evaluate influence of BBS post," *Neurocomputing*, vol. 293, pp. 18–28, Jun. 2018.
[42] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.
[43] H. S. Sastry and M. S. P. Babu, "Cluster analysis of material stock data of enterprises," *Int. J. Comput. Inf. Syst.*, vol. 6, no. 6, pp. 8–19, 2013.
[44] S. Gao, M. Zhou, Y. Wang, J. Cheng, H. Yachi, and J. Wang, "Dendritic neuron model with effective learning algorithms for classification, approximation, and prediction," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 2, pp. 601–614, Feb. 2019.
[45] Y. Yang, I. G. Morillo, and T. M. Hospedales, "Deep neural decision trees," 2018, *arXiv:1806.06988*. [Online]. Available: http://arxiv.org/abs/1806.06988
[46] M. Olson, A. Wyner, and R. Berk, "Modern neural networks generalize on small data sets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 3619–3628.
[47] D. Aha, "UCI machine learning repository: Center for machine learning intelligent systems," Tech. Rep., 2017.
[48] P. Melville and R. J. Mooney, "Diverse ensembles for active learning," in *Proc. 21st Int. Conf. Mach. Learn. (ICML)*, 2004, p. 74.

MOHAMMAD ALOJAIL received the Ph.D. degree in information systems from RMIT University, Australia. He is currently the Chairman of the Department of Information System, College of Computer Sciences and Information Technology, King Faisal University, Saudi Arabia. He is also the Chairman and a member of various committees. He is also the reviewer of international journals and conferences. He has published articles in high indexing databases. He has been an Active Member in the Accreditation team. He has considerable experience in the field of information systems. He has delivered number of talks on leadership and management and successfully led and mange teams to deliver large-scale industrial projects. He has written many scholarly articles in the field of IS and IT outsourcing.

SURBHI BHATIA received the Ph.D. degree in computer science and engineering from Banasthali Vidypaith, India. She is currently an Assistant Professor with the Department of Information Systems, College of Computer Sciences and Information Technology, King Faisal University, Saudi Arabia. She has rich eight years of teaching and academic experience. She has published seven national and international patents. She has published more than 25 articles in reputed journals and conferences in high indexing databases. She has delivered talks as keynote speaker in IEEE conferences and in faculty development programs. She has successfully authored two books from Springer and Wiley. She is also editing three books from CRC Press, Elsevier, and Springer. She has been an Active Researcher in the field of data mining, machine learning, and information retrieval. She is in the Editorial Board Member with Inderscience Publishers in the *International Journal of Hybrid Intelligence* and also in several IEEE conferences.