Big Data Analysis for Revenue and Sales Prediction using Support Vector Regression with Auto-regressive Integrated Moving Average

Gagan K. Sharma*, Sunil Patil
Department of Computer Science and Engineering RKDF University, Bhopal, Madhya Pradesh, India

ABSTRACT
In e-commerce industry, customers' demands get fluctuated throughout the year depending on the purchasing behavior and season. It may be a repetition period in the year, where sales may generally be down, moderate, and whilst some periods are extremely high. Studies reveal that machine learning techniques boosted much e-commerce industry, from supply chain management to business planning. In this paper, a hybrid big data analytical model which integrates Support Vector Regression (SVR) with Auto-Regressive Integrated Moving Average (ARIMA) is proposed to predict product sales and revenues. The simulation results show that the proposed model presents lower relative error rate and higher accuracy that can be utilized for business planning and strategies.

Keywords: Big Data, Predictive Analysis, Machine Learning.

INTRODUCTION
Prediction is the fundamental process for business planning, sourcing, production, product distribution, customer service, and whole supply chain management. Prediction has become more critical and complex due to increasing competition in markets, wide variety of production, and globalization of supply chain. The advanced digital technologies such as cloud computing, Internet of things (IoT), and social media are collecting vast volume of real-time data that is referred to as “big data”. The collected vast volume of data (big data) can be analyzed to predict customers' demands, to improve apprehension of the customers' behavior, and to implement effective supply chain management.

Before the beginning of the sales, customers demands must be predicted for successful business and e-commerce. The under-stocking (when supply of specific products do not fulfill demand of consumers) issues must be avoided by determining the inventory level needs of the customers. The sales predictions have direct impact on customer management, corporate financial market planning, and e-commerce strategies. Sales prediction is indispensable step for modern e-commerce industries and business organizations, but, it experiences many issues and challenges. However, there are many factors that directly or indirectly affect the sales prediction of the business. Figure 1 shows the factors that affect the sales prediction.

The basic economic factors such as national income, interest rates, profit rates, and price level decide the business potential and affect sales prediction. Business factors such as national plans, public opinion, taxation, and loan policies strike the revenues. The internal strategies and plans adopted by business industries, sociological factors (education level, rural-urban habitation, family income, demographic conditions), psycho-logical factors (learning skills, attitudes, habits, perceptions), and competitive factors (quality...
improvement, advertisement, technical competence) create impact on the revenue and sales prediction.\[^{11}\]
Future customers’ demands can be predicted based on historical information and data over certain time-period. Proper consumer demand prediction is an essential process for potential market growth plans, future sales, revenues, decision-making, and strategies. The effectiveness to predict business risks and bankruptcy is essential for financial institutions, as erroneous and inaccurate decisions may lead to financial con-sequences.\[^{13}\] Risky business making decisions can impact on profitability and supply chain management without proper consumer’s demand predictions. An appropriate prediction model can be applied based on the fresh products demands in e-commerce using big data analysis, and thus business can be optimized using future sales prediction.\[^{14}\] The machine learning is applied for monitoring business processes and risk reduction, to predict the degree of risk based on the historical big data corresponding to the risk management.\[^{15}\]

This article is structured as follows; Section II explains the essential impact of machine learning based big data analysis on financial risks and sales prediction in modern e-commerce business. Section III presents related works and literature surveys in the field of revenue and sales prediction using data analysis based on machine learning methods. Section IV presents the proposed integrated “support vector regression (SVR)” with “auto-regressive integrated moving average (ARIMA)” technique, formulation, and overall workflow to evaluate the results. Section V visualizes the python simulation results and comparisons with other methods. Finally, Section VI summarizes the conclusion of the paper with future recommendations.

**Financial Risks and Sales Prediction in E-Commerce**

In the e-commerce industry, financial risks may arise, and they can affect business revenue, funds, manufacturing, supply chain management, distribution cycle, and also social redistri-bution. Financial risks are earnestly influencing e-commerce industries and society. Industries are facing difficulties in the form of financial risks, liquidations and bankruptcies are also raising. Therefore, the predictive analysis and the early cautionary techniques for financial risk management must be perpetually improved.\[^{16}\]
Consequently, Profitable business activities for the economy and society can be maintained using machine learning based big data analysis that can prevent the financial risks effectively.\[^{17}\]

Predictive analysis on vital business risk management can avoid financial loss, and it can maximize the profit. Machine learning can be an auspicious source to business management as the process heavily relies on data-oriented decision-making predictions. The large adoption of machine learning based predictive techniques can be seen in recent year for many business risk management tasks.\[^{18}\]

The risk management is directly associated with the predictive analysis because the effective prediction can significantly minimize the business risks. Based on the previous historical data, the revenues can be efficiently predicted using machine learning-based data analysis. The “Statista” (https://www.statista.com) company specialized in consumer and market data generate predictions about product revenues and sales of upcoming year using machine learning techniques. According to this company, its database contains over 1,000,000 statistical data on above 80,000 subjects from around 22,500 sources and 170 industries. As a sample, Figures 2 and 3 represent apparels and smartphones revenue report of India from the year 2013 to 2026. Based on the previous revenues and sales data from the year 2013 to 2021, the revenues of the year 2022 to 2026 were predicted using machine learning methods.

For financial risk protection in e-commerce business innovation, modern machine learning based big data analytical techniques are the significant landmark. In this paper, “support vector regression (SVR)” with “auto-regressive integrated moving average (ARIMA)” technique is applied to big data of Indian apparels and smartphone sales records for revenue prediction.

**Related Work**

With the progression of Internet technologies, the habit of customers have changed. The online marketing and sales are very common today that create the vast amount of data. Various researchers have proposed their methods to predict business strategies using big data analytics.

For newspaper/magazines’ sales prediction issues, Yu et al.\[^{19}\] used SVR method. They accomplished the lower structural risks in place of the lower empirical risks. Challa et al.\[^{20}\] applied ARIMA to predict risks for companies listed in “Bombay stock exchange (BSE)”.

Using traditional neural networks, Buyar and Abdel-Raouf\[^{21}\] presented a model to predict sales of products. Their research helped business organizations for product stock evaluation, expansion and dividend-price ratio, using future sales prediction. Their model predicted with 3.54% of “mean absolute percentage error (MAPE)” using traditional neural networks and it performed well in training data.

Sentiment analysis and time series analysis-based price prediction model was proposed by Tseng et al.\[^{22}\] for e-commerce products. To enhance the prediction accuracy of their algorithm, they performed text analysis related to product information affected by electronic news. Their outcomes indicate that remarkable events and news influence the e-commerce products sale prices.

Based on the “Stable Seasonal Pattern (SSP)” and “Support Vector Regression (SVR)”, Ye and Eksenaiz\[^{23}\] proposed a hybrid learning technique to enhance the prediction accuracy in stable trend and seasonality constraints. They applied SVR for sales value prediction and applied SSP technique for seasonality index calculation in monthly basis.
Duo et al. [24] applied differential ARIMA for prediction of short-term prices based on current e-commerce conditions. They established “back-propagation neural network” model for prediction of the long-term prices.

Dong et al. [25] established a model for sales prediction using historical e-commerce data. They analyzed the influence of various characteristics that create the immense impact on the product sales. To channelize the business decision-making and to enhance the business profits, their established model is very helpful.

For analyzing the product demands and predictions in e-commerce, Chong et al. [26] used neural network modeling with big data technology. In their analysis, they used variables that are effective sales predictors of online products for better results. Their results reveal that the online customers reviews, promotions, and recommendations extracted from historical data analysis are applicable to develop a model for online e-commerce sales prediction. Their research shows that big data supported techniques to collect large scale data-sets can be applied efficiently, and the results help to predict demands of customers in e-commerce. This research shows that the demands of the products in the business can be predicted through the online marketing. Nowadays, the main platform for product selling is the e-commerce and business companies are focusing to improve product sales prediction.

Palanimalai and Paramasivam [27] discussed the several schemes to analyze Big data which can be exploited to channelize values of the business and modern perceptiveness using higher dimensional big data. For driving sales prediction method, they conducted the experiment using “customer relationship management (CRM)” database that used predicted, target and actual magnitude to aim business
plannings and strategies. The business analytical models help organizations recognize and visualize patterns and trends, which can present perceptive solutions on performance of the modern business. For prediction of product ratings, customers’ sentiments, and recommendations, Woo and Mishra[28] investigated the databases of Amazon customer reviews and applied several machine learning algorithms using Spark and Azure. For conventional machine learning with small datasets as well as massive big data, both Spark and Azure are very effective tools.

To solve the complicated real-world business-to-business (B2B) sales issues, Bohanec et al.[29] demonstrated a novel black-box machine learning model-based application. Their

Table 1: Simulation software

<table>
<thead>
<tr>
<th>Operating System</th>
<th>Ubuntu 22.04 (Linux)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Programming Language</td>
<td>Python 3.9 (Anaconda)</td>
</tr>
<tr>
<td>Dataset</td>
<td>India apparels, smartphone <a href="https://www.statista.com">https://www.statista.com</a></td>
</tr>
<tr>
<td>Framework</td>
<td>PySpark</td>
</tr>
<tr>
<td>Data Preprocessing</td>
<td>Pandas, numpy</td>
</tr>
<tr>
<td>Machine Learning</td>
<td>scikit-learn, tensorflow, auto-arima, forecaster</td>
</tr>
<tr>
<td>Visualization</td>
<td>Matplotlib, seaborn</td>
</tr>
</tbody>
</table>

Table 2: Comparative Analysis For Indian Apparels Sales (2021)

<table>
<thead>
<tr>
<th></th>
<th>SVR with ARIMA Prediction</th>
<th>ARIMA Prediction</th>
<th>ANN Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Revenue (billion US$)</td>
<td>77.10</td>
<td>85.60</td>
<td>56.90</td>
</tr>
<tr>
<td>Relative Error Rate</td>
<td>5.33%</td>
<td>16.94%</td>
<td>22.27%</td>
</tr>
<tr>
<td>Accuracy</td>
<td>94.67%</td>
<td>83.06%</td>
<td>77.73%</td>
</tr>
</tbody>
</table>

Table 3: Comparative Analysis For Indian Smartphone Sales (2021)

<table>
<thead>
<tr>
<th></th>
<th>SVR with ARIMA Prediction</th>
<th>ARIMA Prediction</th>
<th>ANN Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Revenue (billion US$)</td>
<td>40.30</td>
<td>44.80</td>
<td>28.90</td>
</tr>
<tr>
<td>Relative Error Rate</td>
<td>5.49%</td>
<td>17.28%</td>
<td>24.34%</td>
</tr>
<tr>
<td>Accuracy</td>
<td>94.51%</td>
<td>82.72%</td>
<td>75.66%</td>
</tr>
</tbody>
</table>

results demonstrated the usability and effectiveness of the proposed model. A significant recommender system in the product sales strategies and measure action of sellers are the advantages of their proposed model.

Using supervised learning technique, Ma and Fildes[30] proposed a novel “convolution neural network (CNN)” based meta-learner that automatically acquire characteristics from unstructured time-series sales data and their essential factors. Their applied meta-learner significantly improves prediction of the retail sales.

Boldt et al.[31] tested the possibility of precise sales prediction for Nike products using Facebook data and observed the effect of user’s activities of Nike’s Facebook posts on sales revenues. This research can help in business marketing strategies by assessing the information extracted from big social data.

Mishra[32] reviewed and discussed various big data analysis tools developed by computer scientists for significant business intelligent applications in prediction of financial development and management. The business intelligence tools are effective for data-centric business organizations that can create practicable intelligent decision-making results in the analyzable financial scenario.

Niu[33] proposed a novel technique for prediction of “Walmart” sales using “XGBoost” algorithm and feature engineering computation on datasets of Walmart sales. The results depict that this research can effectively extract significant features from data and performed better as compared to “Ridge Regression” and “Logistic regression”. Chen et al.[34] proposed a neural network model for Walmart sales prediction using their database. To train their proposed model, they applied “Tensor Flow” and “Keras”, and the results represent that neural network based model outperforms as compared to support vector machine (SVM) and linear regression algo-rithm. Moreover, they utilized “SHapley Additive exPlanations (SHAP)” framework to examine the significant features for their proposed model.

For fashion products sales prediction, Tehran and Ahrens used data analysis methods. In this research, the simulations were carried out using basic linear regression method. They also exploited the “unsupervised machine learning” techniques for target fitting of datasets.

Sohrabpour et al.[37] applied genetic programming
model to fluctuating export sales record for Middle Eastern business organizations and some considerable important features. It resulted in a genetic programming-based export sales model that assess a prediction for six weeks and sensitivity analysis of the variables. The error metrics mean absolute error, mean squared error, correlation coefficient, and R-squared goodness of fit were calculated to assess the model’s prediction quality.

Qu et al.\(^{[38]}\) proposed a multi-factor “grey wolf optimizer-support vector regression (GWO-SVR)” based model for sales prediction of automobile using external affecting factors and star rating of customers. The effective generalization, basic elementary structure and global optimization are the advantages of SVR model. Significantly, it is well suited for small datasets and multi-dimensional large datasets, that are appropriate in sales prediction considering influencing factors and features. Consequently, the SVR method is applied to train model. The GWO algorithm is further applied to minimize the year 2021. Indian apparels and smartphone sales mainly depends on the seasonal festivals, so the ARIMA based time series analysis with SVR is the best suited method for sales prediction.

**ARIMA**

For prediction in the time series analysis, “Auto-Regressive Integrated Moving Average (ARIMA)\(^{[39,40]}\)” model was established. It basically predicts the future values based on the past values using time-series analysis over a specified time-period. A special statistical method is applied to present the series based on the time-series auto-correlation analytics. With the help of present and past data of time-series, the future values are predicted. Instability of noise data and investigation method’s independency is considered for the ARIMA models prediction process.

The ARIMA \((p,d,q)\) model is expressed in polynomial form using \((l,d)\) process having \(d^{th}\) integer differences that examines the fixed ARMA \((p,d,q)\) process\(^{[41,42]}\) as presented in Equation 1:

\[
T_\tau = \sum_{i=1}^{\phi_1} \phi_i T_{\tau-i} + \sum_{i=1}^{\theta_1} \theta_i \epsilon_{\tau-i} + \epsilon_\tau
\]

(1)

Where \(T_\tau\) = The value of difference time series.
\(\Phi, \Theta\) = The unknown variables.
\(\epsilon\) = Error term with zero mean which is identically and independently distributed.

\(T_\tau\) is usually presented as the past and the current values and also known as the past error term. The three essential conditions of ARIMA model known as stationarity, invertibility and parsimony are applied to identify, estimate and verify the characteristics. Over the time period, mean, variance and co-variance remain constant. This can be completed using differencing the first or second order integral. Using “convergent auto-regressive process”, \(T_\tau\) is calculated in invertibility step.

Finally, the parsimony step produces considerable prediction with additive coefficient. The performance of this model will be better than the parameterized models.

The actual data of previous sales records is denoted by the matrix \(D [a, b]\), where \(a, b\) represents the dimensions of the data, the sales of the companies in \(i^{th}\) year and \(j^{th}\) is denoted by \(S (i, j)\). Where \(i = 1, 2, \ldots, n\) represents the years and \(j = 1, 2, \ldots, 12\) represents the months. So, the total sales in \(i^{th}\) year is calculated as

\[
T_i = \sum_{j=1}^{12} S (i, j)
\]

(2)

The total sales from beginning to \(n^{th}\) year is calculated as

\[
T_n = \sum_{i=1}^{n} \sum_{j=1}^{12} S (i, j)
\]

(3)

The seasonal index \(I_s (i, j)\) of the \(j^{th}\) month in the \(i^{th}\) year is calculated as:

\[
I_s (i, j) = \frac{S (i, j)}{T_i}
\]

(4)

where, \(i = 1, 2, \ldots, n; j = 1, 2, \ldots, 12\). The monthly average seasonal index represents more accurate seasonality if big data of sales record pattern is stable in periodic season. So, the monthly basis seasonal index is calculated as: The seasonal index \(I_s (i, j)\) of the \(j^{th}\) month in the \(i^{th}\) year is calculated as:

\[
I_s (i, j) = \frac{1}{n} \sum_{i=1}^{n} I_s (i, j)
\]

(5)

**Support Vector Regression (SVR)**

The support vector regression (SVR)\(^{[43,44]}\) generally uses the kernel-based transformations. For minimization of regression error margin and maximization of “between class” margin, this kernel-based transformation is used from “original attribute space” to the “higher dimensional space” based on the “Vapnik’s statistical learning theory”.\(^{[45]}\) The non-linear problem can be transformed into the higher dimensional space by applying a selected kernel function.

SVR goes through the “support vector machine” for regression. The basic form of linear regression to express the predicted value \(y\) is:

\[
y (x) = w^{T} \phi (x) + c
\]

(6)

The regular error function \(c\) can be minimized as:

\[
c = \frac{1}{2} \sum_{n=1}^{N} [y_n - \phi (x_n)]^2 + \frac{1}{2} ||w||^2
\]

(7)

The multiple independent variables from input vector \(x\), with the static feature-spaced transformations \(\Phi x\), and the actual data of previous sales records \(D\) are applied to calculate predicted value \(y\).

**Simulation Result and Analysis**

For customers, the Apparel markets in India contain all types of clothing segments which are manufactured, and it is categorized into the main three products: Women’s Apparels, Men’s Apparels and Children’s Apparels. The datasets contain sales record of clothes only, and it does not contain additional personal accessories such as handbags, jewelries,
wristwatches and similar items. Along-with the world’s most popular brands, many small local, middle, and large-scale brands that work successfully in all regional market in India are included in the datasets. The simulations under Ubuntu (Linux) operating system using Python programming language in PySpark framework are performed, the detailed software and tools used to generate results is presented in Table 1.

According to the https://www.statista.com/, in 2022, the expected revenue from the apparel market in India amount to 87.59 billion US$ and the expected annual growth from 2022 to 2026 is 5.1%. The women’s apparel contains the largest portion of the total apparels market volume. The simulations are carried out using big data sales records (datasets) from Indian market of apparels and smartphone.

The results of proposed method is compared with alone ARIMA prediction and “artificial neural network (ANN)” prediction. The Figure 5 represents simulation results that shows the revenue prediction of Indian apparel market. The actual revenue is represented by blue curve, the predicted revenue is represented by the orange curve of proposed model, whereas, green and red curves represent the alone ARIMA and (ANN) based predictions. The Figure 6 represents simulation results to show Indian smartphone market’s revenue prediction.

Similarly, blue, orange, green, and red curves represent actual, proposed, alone ARIMA, and ANN predictions respectively. The alternative bar chart presentation of the result from Figure 5 and 6 is visualized in Figure 7 and 8 respectively. By observing the simulation results from Figure 5-8, it is obvious that the proposed integrated method (SVR with ARIMA) represents more accurate prediction as compared to alone ARIMA and ANN.

Finally, the relative error rate $E_R$ is calculated as:

$$E_R = \left(\frac{|V_A - V_P|}{V_A}\right) \times 100\% \quad (8)$$

where, $V_A$ and $V_P$ denotes the actual and predicted value respectively. The accuracy is calculated as:

$$\text{Accuracy} = 100\% - E_R \quad (9)$$

In the year 2021, the actual revenue generated in Indian apparel sales is 73.2 billion US $, and the actual smartphone sales in India is 38.2 billion US $ (https://www.statista.com). Table 2 and 3 shows the results (relative error and accuracy) generated by ARIMA, ANN, and proposed (SVR with ARIMA) method. By taking Indian apparel sales dataset of the year 2021, the alone ARIMA model predicted 85.6 billion US $ revenue that is 83.06% accurate, the ANN predicted 56.9 billion US $ revenue that is 77.73% accurate, and the proposed model (SVR with ARIMA) predicted 77.1 billion US $ revenue that is 94.67% accurate (Table 2). Similarly, by taking Indian smartphone sales dataset of the year 2021, the alone ARIMA model predicted 44.8 billion US $ revenue that is 82.72% accurate, the ANN predicted 28.9 billion US $ revenue that is 75.66% accurate, and the proposed model (SVR with ARIMA) predicted 40.3 billion US $ revenue that is 94.51% accurate (Table 3). The ARIMA method is perfect for time-series analysis based prediction, and SVR supported ARIMA improves the prediction accuracy significantly. The ANN is very effective method in many fields, however it is not suitable choice for this prediction. As presented in Table II and III, the relative error rate is also minimum for proposed method as compared to alone ARIMA and ANN methods taking Indian apparels and smartphone sales data of the year 2021.

**Conclusion**

The machine learning techniques with big data analysis created significant impact on the progressive growth, strategy, planning, and management in e-commerce business industries. The ARIMA is an effective model to predict time-series seasonal data. This research presented an integrated SVR with ARIMA method to predict sales and revenues of Indian apparel and smartphone for the year 2021. The simulation results represent that the revenues and sales prediction in proposed method are near to the actual record. As compared to alone ARIMA and ANN, the proposed method predicts revenues more accurately with minimum relative error rate.

**References**


Big Data Analysis for Revenue and Sales Prediction using Support Vector Regression


