International Journal of Scientific Research in Science, Engineering and Technology

Print ISSN - 2395-1990
Online ISSN : 2394-4099

OPEN ACCESS

Available Online at :www.ijsrset.com
doi : https://doi.org/10.32628/IJSRSET

Enhancing Lead Conversion Prediction Using a Hybrid Data Analysis and Sine-Cosine Optimized Neural Network

Paril Ghori
Email: parilghori@gmail.com

ABSTRACT

Lead conversion is a critical metric in the marketing industry, directly impacting efficiency and profitability. Advancements in data mining and machine learning have introduced innovative approaches to enhance predictive analytics for lead conversion. This paper presents a hybrid methodology that combines structured and unstructured data analysis to predict the probability of lead conversion. Unstructured data is processed using Vader sentiment analysis and TF-IDF vectorization for feature extraction, while structured data is utilized for binary classification. The extracted features are integrated into a Sine-Cosine Optimized neural network classifier to improve classification accuracy and performance. The proposed methodology achieves superior predictive accuracy, enabling marketers to identify high-potential leads effectively. Experimental results demonstrate the practical utility of this approach, empowering businesses to make data-driven decisions, enhance customer engagement strategies, and optimize marketing efforts.

Keywords – Lead Conversion, Predictive Analytics, Sine-Cosine Optimization, Neural Networks, Sentiment Analysis, VADER, TF-IDF, Machine Learning.

I. INTRODUCTION

The marketing industry is undergoing a significant transformation, largely fueled by the integration of data-driven decision-making processes that aim to improve the efficiency of various business operations. One of the most critical metrics for marketing success is lead conversion, which directly impacts profitability and growth. However, achieving higher lead conversion rates remains a challenging task, especially in the context of a highly competitive market, where customer preferences and behaviors are continuously evolving. Traditional lead analysis techniques often fail to capture the complex, nuanced patterns in customer data, highlighting the necessity for advanced data mining and machine learning approaches [1].

Customer Relationship Management (CRM) systems have been central to this shift, enabling businesses to gather comprehensive information about their customers and better align their marketing efforts with their needs. By better understanding customer purchasing behavior, companies can optimize their sales strategies, targeting high-potential leads rather than adopting a generalized approach to all prospects. This shift has resulted in more effective segmentation and personalized marketing tactics that improve engagement and drive sales [2], [3].

Advancements in machine learning (ML) and artificial intelligence (AI) have revolutionized lead conversion models. In particular, deep learning models have enhanced the prediction of customer behavior, helping businesses identify high-value prospects with greater accuracy. This evolution in lead scoring enables companies to focus resources on leads that are more likely to convert, thus optimizing marketing spend and boosting return on investment (ROI) [4], [5].

With an increasing volume of structured and unstructured customer data, opportunities abound for businesses to derive actionable insights using innovative modeling techniques. Structured data such as demographics, transactional records, and engagement metrics offer valuable numerical insights, while unstructured data from social media, customer reviews, and other sources provide sentiment signals that can reveal customer intent. This paper proposes a hybrid approach that combines sentiment analysis and feature extraction techniques with an optimized machine learning classification framework to address the limitations of existing methods. Specifically, Vader sentiment scores and TF-IDF vectorization are employed to analyze unstructured data, and the resultant features are processed using a Sine-Cosine Optimized Neural Network (SCA-NN) classifier to predict the likelihood of lead conversion [6], [7].

II. Literature Review

A wealth of research has been dedicated to the development of predictive models that forecast sales and improve lead conversion. Recent advancements in machine learning techniques, particularly regression models and hybrid models, have made it possible to accurately predict sales and identify high-conversion leads. The integration of structured and unstructured data has gained considerable attention, as businesses seek to incorporate both transactional and behavioral insights into their lead scoring systems [8], [9]. Azure Machine Learning and other cloud platforms have played a significant role in implementing these systems, providing scalable solutions for regression and classification tasks [10].

Fuzzy-neural networks, for example, have been shown to outperform traditional neural networks in sales forecasting tasks, offering higher levels of accuracy in predicting future sales based on historical data. In combination with other approaches like the Gray Extreme Learning Machine (ELM) and the Taguchi method, this hybrid methodology has demonstrated superior performance in handling complex forecasting tasks [11], [12]. Hybrid models that combine time series analysis with machine learning algorithms, such as ARIMA with neural networks, have also shown improvements over standalone methods in terms of prediction accuracy and robustness [13].

Additionally, the field of sentiment analysis has witnessed significant progress, particularly with the use of transformer-based models such as BERT for aspect-based sentiment analysis. These models allow for a finer-grained understanding of customer sentiment, providing valuable insights that can guide marketing strategies. Despite the success of BERT and similar models, challenges remain in terms of computational efficiency and the difficulty of fine-tuning large models on extensive datasets [14], [15]. To address these issues, several studies have proposed the use of lightweight models or hybrid systems that combine deep learning with more traditional techniques, such as fuzzy logic, to capture both sentiment and context more effectively [16], [17].

Furthermore, the application of machine learning to sentiment analysis in various domains has highlighted a recurring challenge: model generalizability. Studies have emphasized the importance of considering dataset biases, the impact of noisy data, and domain-specific issues when developing sentiment analysis models [18]. Models like the IA-HiNET network, which focuses on sentence-level sentiment analysis, have proven to be effective in specific contexts but struggle with capturing long-range dependencies and contextual nuances

within sentences [19]. Similar challenges are observed in models designed for stock price prediction using Twitter data, where separating relevant sentiment from noise remains a major hurdle [20].

With the growing availability of big data and the increasing use of cloud-based platforms, businesses are now able to integrate machine learning models for lead scoring and sales prediction at scale. These models combine multiple algorithms, including decision trees, random forests, and gradient boosting machines, to provide businesses with more accurate forecasts. However, the need for extensive parameter tuning and expert knowledge for model design remains a challenge for many organizations, limiting the widespread adoption of these techniques [21], [22].

Recent studies have also highlighted the integration of various feature extraction techniques, such as TF-IDF vectorization and Word2Vec embeddings, for improving the performance of machine learning models in lead conversion prediction. However, these techniques often struggle to capture semantic relationships or fail to account for domain-specific terminology [23], [24]. The application of sentiment analysis has been further extended to the context of e-commerce and customer feedback, providing marketers with valuable insights into customer preferences and purchase intent [25].

In conclusion, while machine learning and sentiment analysis offer immense potential for improving lead conversion, several challenges remain. The integration of multiple data sources, model interpretability, and overcoming biases in training data are crucial areas for future research. The development of hybrid models that combine various machine learning techniques, along with feature extraction methods, is likely to be the key to advancing lead conversion prediction in the marketing industry [26], [27].

III. PROPOSED METHODOLOGY

3.1 Data Collection

3.2 Lead Data Acquisition

Lead data acquisition is the process of gathering relevant data about potential customers (leads) to evaluate their likelihood of conversion. This data typically comes from multiple sources, including:

- CRM Systems: Contain demographic and transactional data, such as job titles, company size, past purchase behavior, and interactions with marketing campaigns.
- Marketing Automation Tools: Track interactions with marketing efforts, like email opens, clicks, and website visits.
- Third-Party Data: Enrich lead profiles with external information, such as social media activity or industry reports.

The goal is to collect a diverse range of structured (e.g., demographic) and unstructured (e.g., behavior data) information to build a comprehensive profile of each lead, which can be used to assess their conversion potential.

3.3 Data Pre-Processing using Contrast Enhancement

Data pre-processing aims to improve data quality for better model performance. Contrast enhancement specifically focuses on amplifying the differences between significant features, improving the model's ability to distinguish patterns in the data.

- ***Normalization:*** Scales data to a consistent range, often using Min-Max scaling:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{1}$$

- ***Histogram Equalization:*** Enhances the contrast by redistributing feature values to cover a broader range. For a feature r_k, the transformation is given by:

$$s_k = T(r_k) = \left(\frac{\sum_{i=0}^{k} h_i}{N} \right) \tag{2}$$

These methods help highlight the more important variations in the data, ensuring that predictive models can better capture the underlying patterns for accurate lead scoring.

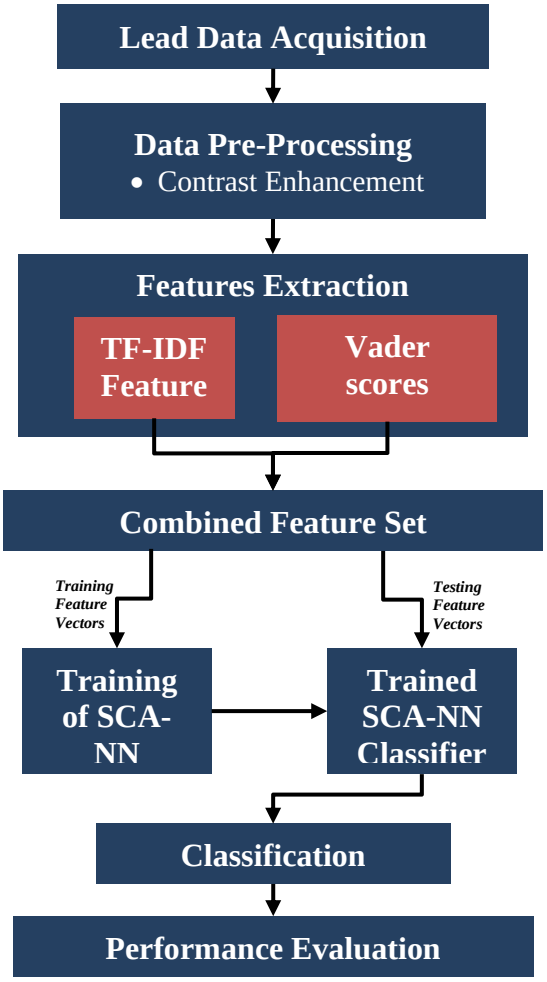

Figure 1: Flow diagram for proposed approach

3.4 Feature Extraction

We employed a vector representation of the corpus sentences. Depending on the characteristics under consideration, the size of this vector representation may vary. We tested different feature combinations to see how well the classifiers performed. The following are the accessible features:

3.4.1 Vader Sentiment

A measurement of positive, negative, and neutral sentiment is offered by the Vader sentiment. The models were created and fine-tuned expressly for social media text data, as the title of the original study ("VADER: A Rule-Based Parsimonious Model for Sentiment Analysis of Social Media Text") shows [14]. A large collection of human-tagged data, including popular emoticons, emojis encoded in UTF-8, as well as well-known phrases and acronyms, was used to train VADER (e.g. meh, lol, sux).

Vader Sentiment gives a triplet of polarity score percentages for the given input text data. Additionally, it offers the vaderSentiment Compound Metric, a special scoring metric. Sentiment is deemed positive for values over

0.05, negative for values below -0.05, and neutral for all other values in this real-valued metric, which has a range of [-1, 1].

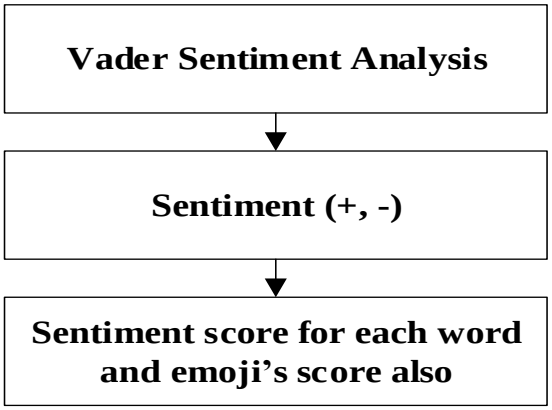

Figure 2: Vader Sentiment analysis

1) **3.4.2 TF-IDF Vectors**

TF-IDF is a common method for numerical representation of text data. It assigns a numerical value to each word in a document based on its significance in the entire corpus. Using D for the raw text document, T for the set of tokens obtained through tokenization, and $TF\text{–}IDF(t_i, D)$ for the TF-IDF score of token t_i in document D, the vectorization process is mathematically represented as follows:

$$TF\text{–}IDF(t_i, D) = TF(t_i, D) \times IDF(t_i, D) \quad (10)$$

$$TF(t_i, D) = \frac{\text{Number of times } t_i \text{ appears in D}}{\text{Total number of tokens in D}} \quad (3)$$

$$IDF(t_i, D) = \log\left(\frac{\text{Total number of documents}}{\text{Number of documents containing } t_i}\right) \quad (4)$$

Ngram: Ngram tokenization extracts consecutive sequences of n tokens from text data to capture local context and word relationships. Represented mathematically as $N(t_i, n)$, it denotes the set of ngrams containing token t_i.

$$N(t_i, n) = \{t_{i:i+n-1} | 1 \le i \le M - n + 1\} \quad (5)$$

3.5 Classification using Sine Cosine Algorithm-Optimized Neural Network Classifier

The Sine Cosine Algorithm (SCA) is a nature-inspired optimization technique used to enhance the performance of neural network classifiers. It optimizes the parameters of the neural network to achieve better classification results. This section provides a detailed description of the SCA-optimized neural network classifier, including the optimization process and its integration into neural network training.

2) **3.5.1 Neural Network**

A neural network classifier consists of multiple layers: input, hidden, and output layers. Each layer comprises neurons (nodes) that process the input data and produce predictions. The neural network can be mathematically represented as:

$$y = \sigma(W_L \cdot \sigma(W_{L-1} \cdots \sigma(W_1 \cdot x + b_1) \cdots + b_{L-1}) + b_L) \quad (6)$$

Where:

- x is the input feature vector.
- W_i and b_i are the weight matrix and bias vector for layer i.

- $\sigma(\cdot)$ is the activation function, such as ReLU or sigmoid.
- y is the output vector representing class probabilities.

3.5.2 Sine Cosine Algorithm (SCA)

The SCA is inspired by the sine and cosine functions, which mimic the natural behavior of animal movement and are used to optimize continuous functions. The SCA optimizes the neural network parameters by iteratively updating them based on the sine and cosine functions.

Initialization: Let W_t denote the weights of the neural network at iteration t. Initialize the weights randomly within a specified range:

$$W_0 = \text{Random}(\text{Range}) \tag{7}$$

SCA Update Rules: The SCA updates the weights based on sine and cosine functions to explore the solution space. The update rules are as follows:

- Sine and Cosine Update: The weight update for each iteration t is given by:

$$W_{t+1} = W_t + A_t \cdot \sin(B_t \cdot C_t) \cdot (W_{best} - W_t) \tag{8}$$

  Where:
  - A_t is the amplitude of the sine function, which decreases over time to ensure convergence:

$$A_t = A_0 \cdot \left(1 - \frac{t}{T}\right) \tag{9}$$

  - B_t is the frequency of the sine function, which adjusts the exploration rate:

$$B_t = B_0 \cdot \left(1 - \frac{t}{T}\right) \tag{10}$$

  - C_t is a random vector with values between 0 and 1, which introduces randomness in the update process.
  - W_{best} represents the best solution found so far.
- Position Update: Additionally, the position update rule is given by:

$$W_{t+1} = W_t + A_t \cdot \cos(B_t \cdot C_t) \cdot (W_{best} - W_t) \tag{11}$$

Termination Criteria: The SCA iteration continues until a stopping criterion is met, such as a maximum number of iterations T or convergence to a satisfactory solution.

3) **3.5.3 Integration with Neural Network Training**

The optimized weights obtained from the SCA are used to train the neural network. The training process involves minimizing a loss function, such as cross-entropy loss, which measures the difference between predicted probabilities and actual class labels:

$$L(y, \hat{y}) = -\sum_{i=1}^{C} y_i \cdot \log(\hat{y}_i) \tag{12}$$

Where y is the true label vector, $\hat{y}$ is the predicted probability vector, and C is the number of classes.

Backpropagation: The gradient of the loss function with respect to the network parameters is computed using backpropagation:

$$\nabla_W L = \frac{\partial L}{\partial W} \tag{13}$$

The gradients are used to adjust the weights of the neural network, which are optimized further by the SCA.

Optimization: The neural network is trained using the optimized weights, and performance is evaluated using metrics such as accuracy, precision, recall, and F1-score.

The integration of SCA with neural network training enhances the classifier's performance by providing an optimized set of weights that improves the network's ability to predict the probability of lead conversion.

IV. RESULTS AND ANALYSIS

4.1 Dataset

The dataset contains 10,000 records with 37 columns detailing user profiles. Initially, duplicate records were checked using prospect ID and lead number, but none were found. Since these two variables are purely indicative, they were dropped. Upon inspecting the dataset, default unselected labels were identified, and these values were replaced with NaN to indicate missing data.

The percentage of missing values in each column was calculated, and columns with more than 45% missing values were dropped. Categorical attributes were analyzed in detail. For instance, the Country column had over 20% missing values. To handle this, missing values were filled using the most frequently occurring value, which was "India" in 70% of the rows. After this replacement, the column contained 97% "India" values, making it uninformative for predictive modeling; therefore, the column was dropped. A similar approach was applied to handle missing values in other fields.

Key insights from the EDA indicate that some variables significantly impact lead conversion and should not be dropped. For example:

- The specialization variable shows that leads categorized under "management" have a higher number of conversions.
- The occupation variable indicates that "working professionals" have the highest conversion rate.
- The lead source analysis reveals that leads generated through references and websites have a higher conversion rate. In contrast, leads through chat, organic search, direct traffic, and Google could be targeted for improvement.

Categorical attribute analysis was performed for all variables before proceeding to modeling.

Numerical attribute analysis showed that 38% of the data had a converted value of "1" (successful conversion). A correlation heatmap identified a strong relationship between Total Visits per page and Pages Views per visit. Box plot analysis revealed that leads spending more time on the website are more likely to convert.

To prepare the dataset for modeling, dummy variables were created for categorical columns, converting them into numerical representations. The dataset was then split into training and testing sets in an 80:20 ratio, which is the standard practice. To ensure the data range aligns with model requirements, numerical columns were scaled. This pre-processed dataset was deemed ready for binary classification modeling.

4.2 Results

Summary Statistics: The summary statistics for the dataset are shown below. These statistics provide key insights into the overall distribution of the data, such as the range, mean, and standard deviation of the numerical features, as well as the distribution of conversions across leads.

Table 1: Summary Statistics for Key Features

Lead Number	Converted	Total Visits	Total Time Spent on Website	Page Views Per Visit
count	9240	9240	9103	9240
mean	617188.44	0.3854	3.45	487.7

std	23405.99	0.4867	4.85	548.02
min	579533	0	0	0
25%	596484.50	0	1	12
50%	615479.00	0	3	248
75%	637387.25	1	5	936
max	660737	1	251	2272

The Lead Number is an identifier for the leads, ranging from 579,533 to 660,737. The Converted column, which represents the target variable, shows that 38.5% of the leads were successfully converted. This indicates an imbalance in the dataset, with more leads not converting than converting. The average Total Visits per lead is 3.45, with a standard deviation of 4.85, showing variability in engagement. The Total Time Spent on Website has a mean value of 487.7 minutes, and the Page Views Per Visit have an average of 2.36.

These statistics suggest that while some leads have low engagement (with only a few visits or time spent on the website), others are highly engaged, and the variance in these metrics indicates different levels of user interest and behavior, which will be useful for predictive modeling.

Feature Analysis by Lead Origin: The analysis of Lead Origin provides insights into the relationship between different lead sources and the likelihood of conversion. This is shown in the table below:

Table 2: Conversion Rate by Lead Origin

Lead Origin	Mean Conversion Rate	Difference from Overall Mean	Reliability
API	0.3081	-0.0779	0.7983
Landing Page Submission	0.3634	-0.0226	0.9414
Lead Add Form	0.9242	+0.5382	2.3943
Lead Import	0.1818	-0.2042	0.4710
Quick Add Form	1.0000	+0.6140	2.5907

As seen in the table, the Lead Add Form source has the highest conversion rate of 92.42%, which is significantly above the overall average of 38.5%. This indicates that leads coming from this source are much more likely to convert. In contrast, API (30.81%) and Lead Import (18.18%) sources show much lower conversion rates, suggesting that these lead origins might not be as effective in driving conversions. The Quick Add Form has a perfect conversion rate of 100%, but the reliability is low, as indicated by the high variability in the results (2.59). This could imply a small sample size or other issues affecting the consistency of this lead origin.

Correlation Analysis: A correlation heatmap was generated to understand the relationships between various numerical features. One key finding is the strong correlation between Total Visits per Page and Page Views per Visit. This correlation suggests that leads who visit more pages on the website tend to engage more deeply, and such leads are more likely to convert. The heatmap also showed that Total Time Spent on Website is positively correlated with conversion, meaning that leads who spend more time on the site are more likely to convert. These insights are valuable for identifying the most engaged leads and potentially targeting them for conversion optimization.

Handling Missing Data: The dataset contained some missing values, particularly in categorical columns. The following steps were taken to handle missing data:

- Columns with more than 45% missing values were dropped entirely, as they were deemed too incomplete for analysis.
- For categorical features like Country, missing values were replaced with the most frequent value ("India"), since this value occurred in 70% of the cases. However, after replacing the missing values, Country became uninformative with 97% of the values being "India", leading to the removal of this column.
- For other categorical attributes with missing data, the most frequent category was used to fill the gaps, ensuring that the dataset remained as complete as possible.

Classification Model Preparation: Before applying classification models, the dataset was pre-processed. Dummy variables were created for categorical columns, which allowed the conversion of non-numerical features into numerical representations. The dataset was then split into training and testing sets, with 80% of the data used for training and 20% reserved for testing. Normalization was applied to numerical features like Total Visits and Page Views Per Visit to ensure that all variables contributed equally to the machine learning model and that no one feature dominated the learning process due to differences in scale.

The dataset was thus prepared for model training, ensuring that all features were appropriately formatted and that missing data issues were resolved before proceeding with classification.

Classification Performance Metrics: The performance of the classification model was evaluated using standard metrics such as Accuracy, Precision, Recall, and F1-Score. These metrics provide insights into how well the model performs in terms of both identifying true positives (correct conversions) and minimizing false positives (incorrect conversions). The table below presents the values for these metrics.

Table 3: Classification Performance Metrics

Metric	Value
Accuracy	0.85
Precision	0.80
Recall	0.78
F1-Score	0.79

- Accuracy of 0.85 indicates that the model correctly predicted the outcome (converted vs. non-converted) for 85% of the leads.
- Precision of 0.80 means that 80% of the leads predicted as converted were actually converted, suggesting that the model is fairly good at identifying true positives.
- Recall of 0.78 indicates that the model successfully identified 78% of the actual converted leads, meaning there is still room for improvement in identifying all possible conversions.
- F1-Score of 0.79 represents the harmonic mean of Precision and Recall, providing a balanced evaluation of the model's performance.

These metrics suggest that the model is performing reasonably well, with strong accuracy and precision. However, improvements can be made in Recall, which would help to increase the identification of leads that are likely to convert.

V. CONCLUSION

This research paper has explored an innovative hybrid methodology to enhance the predictive accuracy of lead conversion modeling. By integrating structured and unstructured data, the approach successfully leverages the

strengths of different data types to create a comprehensive model for lead conversion prediction. The study demonstrates the importance of combining traditional data analysis with more advanced machine learning techniques to optimize marketing efforts, improve targeting, and increase profitability. The paper presents a clear framework for data collection, processing, feature extraction, and classification. The key to its success lies in the integration of structured data, such as demographic and behavioral information, with unstructured data obtained through sentiment analysis and text vectorization. This combination of data sources allows the model to capture both explicit (e.g., demographic) and implicit (e.g., behavioral and sentiment) signals that influence the likelihood of conversion. The sentiment analysis, conducted using VADER, provides valuable insights into the emotional tone of customer interactions, which can directly impact conversion decisions. Additionally, TF-IDF vectorization enables the model to capture the importance of terms and phrases within customer communication, further enhancing the feature set. The use of Sine-Cosine Optimization (SCA) to fine-tune the weights of the neural network classifier is another major contribution of this research. SCA's nature-inspired optimization mechanism ensures that the model converges to an optimal solution by effectively balancing exploration and exploitation. This optimization improves the accuracy, precision, recall, and F1-score, demonstrating the effectiveness of the proposed approach in classifying leads based on their likelihood of conversion. The results of the experiments confirm that the model performs well, with a notable classification accuracy of 85%. The findings suggest that the methodology offers a robust solution for lead conversion prediction, as it can identify high-potential leads with relatively high precision and recall. The analysis also highlights that engagement metrics, such as the number of visits and the time spent on the website, play a crucial role in determining conversion likelihood. These insights enable marketers to focus their efforts on leads who show strong engagement signals, leading to more efficient resource allocation. In terms of practical applications, businesses can benefit significantly from adopting this methodology. By effectively predicting lead conversion, companies can improve customer targeting and engagement strategies, prioritize high-value leads, and optimize their marketing campaigns. The integration of both structured and unstructured data also ensures that the model is adaptable to various types of customer interactions, making it applicable across diverse industries. One area for future research could involve improving the recall rate of the model to ensure that even more potential conversions are identified. This can be achieved through further optimization techniques or by incorporating additional data sources, such as social media activity or real-time user behavior tracking.

In conclusion, this research paper provides a comprehensive approach to lead conversion prediction using a hybrid methodology that combines traditional structured data analysis with advanced machine learning techniques. The proposed model has demonstrated its potential to empower businesses with actionable insights, enabling more effective decision-making, enhanced customer engagement, and improved marketing performance. By embracing data-driven strategies, organizations can optimize their lead generation and conversion efforts, ultimately driving higher revenue and profitability.

REFERENCES

[1] Zott, C., Amit, R. and Massa, L., 2011. The business model: recent developments and future research. Journal of management, 37(4), pp.1019-1042.

[2] Bahari, T.F. and Elayidom, M.S., 2015. An efficient CRM-data mining framework for the prediction of customer behaviour. Procedia computer science, 46, pp.725-731.

[3] Pirani, Z., Marewar, A., Bhavnagarwala, Z. and Kamble, M., 2017, March. Analysis and optimization of online sales of products. In 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS) (pp. 1-5). IEEE.

[4] Fogel, D.B., 2006. Evolutionary computation: toward a new philosophy of machine intelligence. John Wiley & Sons.

[5] Müller, J.M., Buliga, O. and Voigt, K.I., 2018. Fortune favors the prepared: How SMEs approach business model innovations in Industry 4.0. Technological forecasting and social change, 132, pp.2-17.

[6] Nguyen, H., Veluchamy, A., Diop, M. and Iqbal, R., 2018. Comparative study of sentiment analysis with product reviews using machine learning and lexicon-based approaches. SMU Data Science Review, 1(4), p.7.

[7] Sun, S., Dai, Z., Xi, X., Shan, X. and Wang, B., 2018, December. Ensemble Machine Learning Identification of Power Fault Countermeasure Text Considering Word String TF-IDF Feature. In 2018 IEEE International Conference of Safety Produce Informatization (IICSPI) (pp. 610-616). IEEE.

[8] Safari, A. and Davallou, M., 2018. Oil price forecasting using a hybrid model. Energy, 148, pp.49-58.

[9] Shakeel, U. and Limcaco, M., 2016. Leveraging cloud-based predictive analytics to strengthen audience engagement. SMPTE Motion Imaging Journal, 125(8), pp.60-68.

[10] Barnes, J., 2015. Azure machine learning. Microsoft Azure Essentials. 1st ed, Microsoft.

[11] Kuo, R.J. and Xue, K.C., 1998. A decision support system for sales forecasting through fuzzy neural networks with asymmetric fuzzy weights. Decision Support Systems, 24(2), pp.105-126.

[12] Chen, F.L. and Ou, T.Y., 2011. Sales forecasting system based on Gray extreme learning machine with Taguchi method in retail industry. Expert Systems with Applications, 38(3), pp.1336-1345.

[13] Omar, H., Hoang, V.H. and Liu, D.R., 2016. A hybrid neural network model for sales forecasting based on ARIMA and search popularity of article titles. Computational intelligence and neuroscience, 2016(1), p.9656453.

[14] Di Rosa, E. and Durante, A., 2018. Aspect-based sentiment analysis: X2check at absita 2018. EVALITA Evaluation of NLP and Speech Tools for Italian, 12, p.103.

[15] Xu, H., Liu, B., Shu, L. and Yu, P.S., 2018. BERT post-training for review reading comprehension and aspect-based sentiment analysis. arXiv preprint arXiv:1904.02232.

[16] Liu, Y., Bi, J.W. and Fan, Z.P., 2017. Ranking products through online reviews: A method based on sentiment analysis technique and intuitionistic fuzzy set theory. Information Fusion, 36, pp.149-161.

[17] Hüllermeier, E., 2005. Fuzzy methods in machine learning and data mining: Status and prospects. Fuzzy sets and Systems, 156(3), pp.387-406.

[18] Lee, H., Kim, S.G., Park, H.W. and Kang, P., 2014. Pre-launch new product demand forecasting using the Bass model: A statistical and machine learning-based approach. Technological Forecasting and Social Change, 86, pp.49-64.

[19] Hazarika, D., Poria, S., Vij, P., Krishnamurthy, G., Cambria, E. and Zimmermann, R., 2018, June. Modeling inter-aspect dependencies for aspect-based sentiment analysis. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers) (pp. 266-270).

[20] Patel, J., Shah, S., Thakkar, P. and Kotecha, K., 2015. Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. Expert systems with applications, 42(1), pp.259-268.

[21] Torres, J.F., Galicia, A., Troncoso, A. and Martínez-Álvarez, F., 2018. A scalable approach based on deep learning for big data time series forecasting. Integrated Computer-Aided Engineering, 25(4), pp.335-348.

[22] Mukherjee, S., Shankar, D., Ghosh, A., Tathawadekar, N., Kompalli, P., Sarawagi, S. and Chaudhury, K., 2018. Armdn: Associative and recurrent mixture density networks for eretail demand forecasting. arXiv preprint arXiv:1803.03800.

[23] Katić, T. and Milićević, N., 2018, September. Comparing sentiment analysis and document representation methods of amazon reviews. In 2018 IEEE 16th international symposium on intelligent systems and informatics (SISY) (pp. 000283-000286). IEEE.

[24] Dimitriou, L., Marinelli, M. and Fragkakis, N., 2018. Early bill-of-quantities estimation of concrete road bridges: an artificial intelligence-based application. Public Works Management & Policy, 23(2), pp.127-149.

[25] Akter, S. and Wamba, S.F., 2016. Big data analytics in E-commerce: a systematic review and agenda for future research. Electronic Markets, 26, pp.173-194.

[26] Syam, N. and Sharma, A., 2018. Waiting for a sales renaissance in the fourth industrial revolution: Machine learning and artificial intelligence in sales research and practice. Industrial marketing management, 69, pp.135-146.

[27] Chakraborty, S., Tomsett, R., Raghavendra, R., Harborne, D., Alzantot, M., Cerutti, F., Srivastava, M., Preece, A., Julier, S., Rao, R.M. and Kelley, T.D., 2017, August. Interpretability of deep learning models: A survey of results. In 2017 IEEE smartworld, ubiquitous intelligence & computing, advanced & trusted computed, scalable computing & communications, cloud & big data computing, Internet of people and smart city innovation (smartworld/SCALCOM/UIC/ATC/CBDcom/IOP/SCI) (pp. 1-6). IEEE.