ISSN: 2959-6386 (Online), Volume 2, Issue 2

Journal of Knowledge Learning and Science Technology

Journal homepage: https://jklst.org/index.php/home

JKLST
EXPLORING THE INTERSECTION OF KNOWLEDGE, LEARNING, AND SCIENCE & TECHNOLOGY:
UNLEASHING THE POWER OF SYNERGY IN THE 21ST CENTURY

Optimizing Sales Funnel Efficiency: Deep Learning Techniques for Lead Scoring

Kapil Kumar Sharma[1], Manish Tomar[2], Anish Tadimarri[3]

[1]Cisco, USA

[2]Citibank, USA

[3]High Radius, USA

Abstract

Segmenting new commercial leads is a critical endeavor for contemporary businesses operating in highly competitive markets, aiming to unearth lucrative opportunities and bolster their Return On Investment (ROI). Business lead scoring entails attributing a score, representing the likelihood of a lead to make a purchase, to each potential lead generated for the business. These leads' interactions across various marketing channels on the internet yield valuable attributes, including pertinent information such as contact details, lead source, and channel, alongside behavioral cues like response speed and movement tracking. This process aids in evaluating the quality of opportunities and their stage in the purchasing journey. Moreover, an accurate lead scoring mechanism empowers marketing and sales teams to prioritize leads effectively and respond promptly, thereby enhancing the likelihood of conversion. Leveraging machine learning algorithms can streamline this process.

In this study, the authors conducted a comparative analysis of the performance of various machine learning (ML) algorithms in predicting lead scores. The Random Forest and Decision Tree models emerged with the highest accuracy scores, reaching 93.02% and 91.47%, respectively. Notably, the Decision Tree and Logistic Regression models exhibited shorter training times, which can prove pivotal when handling extensive datasets.

Keywords: CRM; Predictive Lead Scoring; Marketing Management; Machine Learning; Artificial Intelligence.

Article Information:
Article history: *Received:01/11/2023Accepted:3/11/2023* *Online: 12/11/2023 Published: 12/11/2023*
DOI: *https://doi.org/10.60087/jklst.vol2.n2.p274*
[i] ***Correspondence author:*** Kapil Kumar Sharma

Introduction

In Business-to-Consumer (B2C) sales activities, lead interactions can be categorized into two main phases: lead generation and lead conversion. The process initiates with the outreach to potential clients across various channels such as websites, social media, and campaigns, aiming to draw them to the business's platform and engage with it. These interactions are systematically monitored by automated systems, contributing to the nurturing of the business's database. Ultimately, a professional sales agent engages with the lead to guide them through the purchasing decision, addressing any obstacles they may encounter and leveraging marketing tactics and financial incentives, such as coupons and discounts, to facilitate the transaction.

However, amidst these stages lies a critical process of identifying valuable prospects for the business. Given the substantial costs associated with sales operations in terms of time and resources, prioritizing the most engaged and suitable leads becomes imperative to enhance and sustain a profitable Return on Investment (ROI). This pivotal process is known as lead scoring, enabling businesses to leverage data analytics to accurately assess the potential value of each lead.Central to the lead scoring process is the determination of weights assigned to each feature. In traditional scoring models, these weights are often derived through a trial-and-error approach with the guidance of marketing experts to ascertain their optimal values. However, with the proliferation of artificial intelligence applications across various domains, lead scoring strategies have evolved, leveraging predictive modeling techniques.Machine learning algorithms play a crucial role in discerning between failed leads and successful ones, i.e., those who ultimately make purchasing decisions. These algorithms analyze common attributes among converted leads to formulate a predictive model that automatically identifies potential prospects for marketing teams.This paper aims to compare the performance of various predictive machine learning models using a publicly available dataset for lead scoring. Section 2 provides a brief overview of the internal workings of B2C processes and the lead scoring mechanism, along with a review of current research in this domain. Section 3 outlines the experiment, detailing the dataset used, the machine learning algorithms employed, and the experimental procedures. Section 4 presents and discusses the experimental results, evaluating the accuracy of the models using different metrics. Finally, the conclusion discusses the practical implementations of such approaches in real-world scenarios and offers insights into future directions for this research.

Paper Context

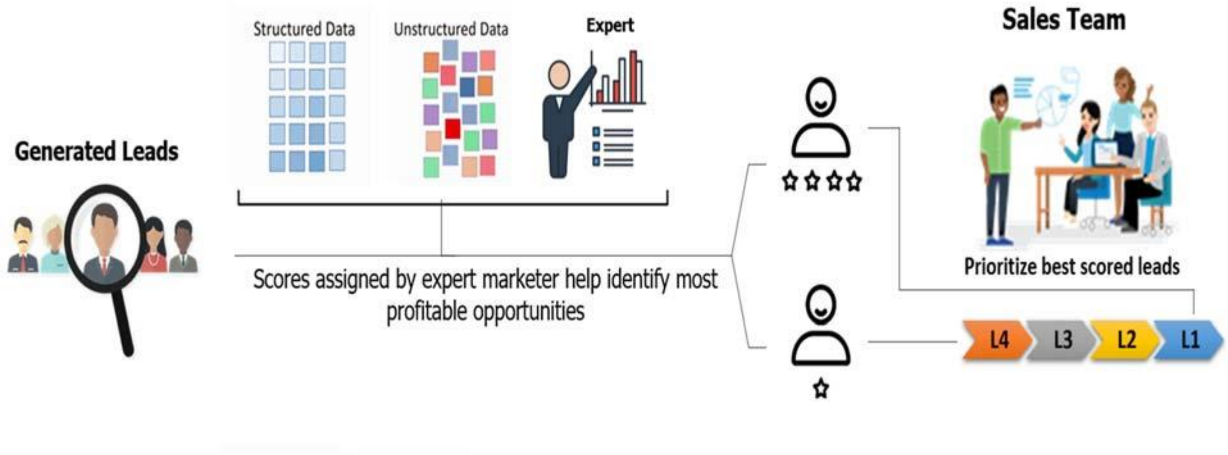
Lead scoring serves as a pivotal marketing tool, aiding decision-makers in pinpointing the most promising potential customers from the pool of generated leads. Consequently, sales professionals can optimize their efforts by focusing on leads with higher conversion probabilities, rather than expending resources on all prospects indiscriminately.

The core concept involves assigning scores to prospects based on the degree of alignment between their attributes and the predefined profile of a converted customer. Leads surpassing a specified threshold are deemed ideal targets. However, the challenge lies in determining the pertinent attributes and their respective weights for profile matching.

Traditional Lead Scoring

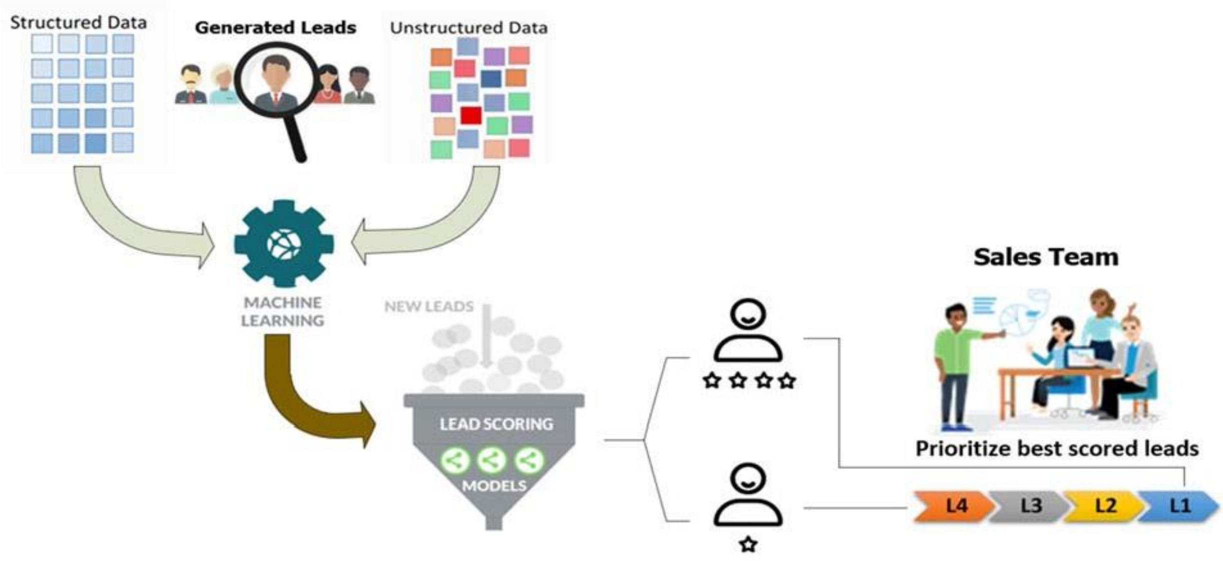
In traditional, manual lead scoring, the task of identifying relevant customer traits and assigning points is typically entrusted to a marketing expert or senior sales manager. Drawing upon their experience, these managers discern the most crucial attributes amidst a myriad of possibilities. Typically, scores are computed based on the lead's alignment with the company's ideal customer profile (demographic details) and their level of engagement (behavioral details). Figure 1 illustrates the architecture of a traditional lead scoring system.

However, due to inherent human biases, sales managers may not always accurately pinpoint the most critical features. Their judgments may be influenced by personal biases or outdated assumptions, leading to suboptimal decision-making and a reluctance to adapt to evolving trends.


Generated Leads
Structured Data
Unstructured Data
Expert
Scores assigned by expert marketer help identify most profitable opportunities
Sales Team
Prioritize best scored leads
L4
L3
L2
L1

Advanced Lead Scoring

Predictive lead scoring employs a statistical technique known as propensity modeling [3] to anticipate the likelihood of specific actions (e.g., purchase, reservation) by visitors.Through the integration of machine learning and data mining, this approach endeavors to forecast the behavior of target audiences and their probability of conversion [4]. Machine learning (ML) algorithms autonomously analyze historical sales data [5][6] to identify relevant attributes and uncover meaningful patterns indicative of a lead's conversion propensity. The resulting model is trained and evaluated to minimize false-positive predictions. Any errors or anomalies detected in predictions can be annotated and incorporated back into the training data, enabling the model to adapt and remain relevant, particularly in dynamic business environments. Figure 2 illustrates the architecture of an ML-based Predictive Lead Scoring System.A key distinction between traditional and predictive lead scoring models lies in their capacity to handle large volumes of data and glean deeper insights for performance enhancement [7]. Furthermore, human perception has its limitations in making sense of thousands of data points and identifying relationships and rules among them. Leveraging machine learning for propensity prediction allows for the augmentation or even replacement of professional expert marketers.



(possibly costly) with an automated system possessing comparable judgment abilities, which will progressively enhance with the adoption of big data [8]. Refer to Table I for significant distinctions between lead scoring systems.

	Traditional Scoring	**Predictive Scoring**
Rules	Subjective rules established by expert marketers	Detected by ML algorithms
Supervision	Requires Manual supervision and regular adjustments and updates	Minimal supervision
Data size	Small datasets and limited processing power	Large datasets (accuracy increase with training data size)
Result	Lead Scores	Conversion Probability

Related Research

The propensity of leads in both B2B and B2C contexts has garnered significant attention from researchers due to its substantial impact on sales efficiency and the optimization of internal workflows in customer management. R. Nygård et al. [9] proposed a supervised learning approach for lead scoring, employing algorithms like Logistic Regression and Decision Trees to predict purchase probability based on prior knowledge and behavioral data. Their findings revealed that the Random Forest model yielded the best performance. S. Singh et al. [10] suggested modeling the search habits of visitors to commercial websites using supervised machine learning algorithms to identify and extract shopping patterns and shifts in trends among these visitors. K. Prasad et al. [11] conducted a comparative analysis between Support Vector Machine (SVM) and Logistic Regression algorithms for building models to predict propensity, evaluating their performance. S. Mortensen et al. [12] utilized structured and unstructured data from a paper and packaging company's IT system to forecast B2C sales success. They compared several algorithms, including Binomial logistic regression and various decision tree approaches, with the best model achieving a propensity accuracy of 80%, along with precision and recall scores of 86% and 77%, respectively. Y. Zhang et al. [13] aimed to identify the most valuable prospects using machine learning. Their comparison between logistic regression and random forest models demonstrated the superior accuracy of the latter. However, logistic regression outperformed random forest in terms of recall rate, F1 score, and Receiver Operating Characteristic (ROC). Y. Benhaddou et al. [14] addressed the challenge of training small datasets by constructing a Lead Scoring model with a Bayesian network, leveraging expertise and applying common heuristics to reduce model complexity.A. Etminan[15] focused on estimating the effect of feature weights by evaluating various feature ranking and judging schemes in a predictive lead scoring scenario. J. Yan et al. [16] proposed a unified, machine learning-based framework for estimating marketing opportunity propensity, addressing challenges specific to the B2B environment. A. Rezazadeh [17] tackled the forecasting of B2B and B2C sales outcomes by introducing a data-driven, machine learning-based pipeline in a cloud environment, demonstrating the superior accuracy and monetary value of decision-making based on ML predictions compared to traditional operator-based approaches. A. Sabbani et al. [18] introduced a novel approach for seller-buyer matching at trade show events using machine learning, advocating for an automated approach to replace syntactic analysis with implicit user feedback on a frontend intelligent application.In summary, this paper contributes by:

- Experimenting with various algorithms (six in total) to confirm the Random Forest as the most suitable choice, as supported by the literature review.
- Employing a diverse range of metrics and validation techniques beyond accuracy alone to evaluate model performance.
- Introducing processing time and computing power as crucial criteria in model selection to ensure stable performance with large datasets.

Materials and Methods

In this study, we adhere to the generally recommended framework for predictive modeling and smart analytics research. Dataset understanding constitutes a crucial initial step, focusing on investigating the dataset and identifying and addressing potential issues. The data preparation process is fundamental, involving handling missing values, detecting outliers, and crafting a relevant feature vector using techniques such as feature selection and extraction to optimize machine learning model construction. Subsequently, multiple models are implemented and evaluated, followed by an analysis and interpretation of each model's performance.

Dataset

Data Description

The primary objective of this study is to showcase the benefits of machine learning in automating the lead scoring process through predictive modeling. To achieve this, we conducted experiments using the "X Education" public dataset, widely utilized in lead prediction processes. The dataset encompasses various variables covering the following aspects:

- The lead outcome (converted or not).
- Visitor interactions on the website (e.g., pages visited, time spent).
- Information collected through website forms (e.g., contact details, newsletter subscriptions).

• Lead source (search engine, referrer, direct).

The dataset comprises 9240 data points with 37 features, each appropriately capturing prospect characteristics. Some features are numerical (6 features), such as website visit duration and frequency, while others are categorical, including search keywords, lead source, and contact preferences.

Data Preprocessing

Using diverse data preprocessing techniques, we extracted 89 features from the raw dataset, which initially contained 37 features. A tailored processing pipeline was applied based on the type and distribution of values for each feature:

• Features with a missing value ratio exceeding 70% were dropped from the dataset due to minimal variance gain and risk of distribution skewness. Similarly, data points with numerous missing attributes were excluded for similar reasons.

• Features with a low missing value ratio underwent feature-specific processing to replace missing values with appropriate values (mean, median, mode, etc.).

• For categorical features (e.g., Lead Origin, Specialization), a One Hot Encoding process was employed to construct a one-hot encoded vector, thereby expanding the number of features from 37 to 89.

The resulting dataset comprised 9074 instances with 89 features and no missing values.

Techniques

Following data preprocessing, we selected six widely used machine learning algorithms from the literature for customer classification tasks and predictive modeling.

K-Nearest Neighbors

KNN is a straightforward, non-parametric classification algorithm initially introduced by E. Fix and J. Hodges in 1951 and further developed by T. Cover. It determines an instance's class label based on a majority vote rule of its nearest neighbors, with the class assigned being the most common among its K nearest neighbors, as measured by a distance metric. Different distance metrics, such as Euclidean, Manhattan, and Minkowski distances, can be utilized based on the specific characteristics of the dataset.

- Euclidean Distance: $\sqrt{\sum}\mathrm{k}\ (x\mathrm{i} — y\mathrm{i})2$

- Manhattan Distance: $\sum k\ |xi — yi|$
- Minkowski Distance: $(\sum k\ (|x — y\ |)q)1/q$

In this paper, Minkowski distance is used as the default distance function in Sklearn.

Naïve Bayes

NB is a simple yet effective classification model based on conditional probability and the Bayesian theorem, making a naive assumption about feature independence. It assigns probabilities to each possible outcome or class based on the features of a given instance. The NB classifier combines the Naive Bayes probabilistic model with a decision rule, typically using the Maximum A Posteriori (MAP) decision rule to select the most probable hypothesis.

Support Vector Machine

SVM is a binary classification algorithm developed by V. Vapnik, aiming to find a hyperplane in a higher-dimensional feature space to separate instances into two distinct classes. It can handle multiclass classifications by employing a "one-against-all" strategy, optimizing multiple binary classifications. The class of an instance is determined based on the decision function output.

Decision Tree and Random Forest

A Decision Tree is a decision-making algorithm represented as a tree structure, where each non-leaf node corresponds to a decision based on an input feature. The tree is constructed by recursively partitioning the dataset into smaller subsets, with classification rules memorized at each level.

Random Forest is an ensemble learning method that combines multiple decision trees. Each tree in the forest independently makes predictions, and the final prediction is determined by the majority vote of all trees, resulting in improved performance over individual decision trees.

Logistic Regression

LR is a binary and linear classification algorithm that maps input features to a value range of [0, 1] using a sigmoid function. Despite its name, LR is widely used for binary and linear classification tasks due to its simplicity and efficiency. The model parameters are obtained by minimizing the loss function, allowing LR to be optimized for large datasets.

Model Evaluation

First Level Indicators

Basic indicators, such as True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN), are computed using the testing dataset and the model's predictions, forming a confusion matrix to evaluate model performance.

Second Level Indicators

Secondary indicators, including Accuracy, Precision, Recall (Sensitivity), F1-Score, and Specificity, are derived from the first-level indicators to provide more nuanced insights into model performance, especially in unbalanced datasets.

K-folds Cross-validation

K-folds Cross-validation is a resampling technique used to assess the generalization ability of an ML model. By splitting the dataset into k equal-sized subsets and iteratively training and testing the model on different combinations of subsets, K-folds Cross-validation helps mitigate bias and variance issues, leading to more robust model evaluation.

Results and Discussion

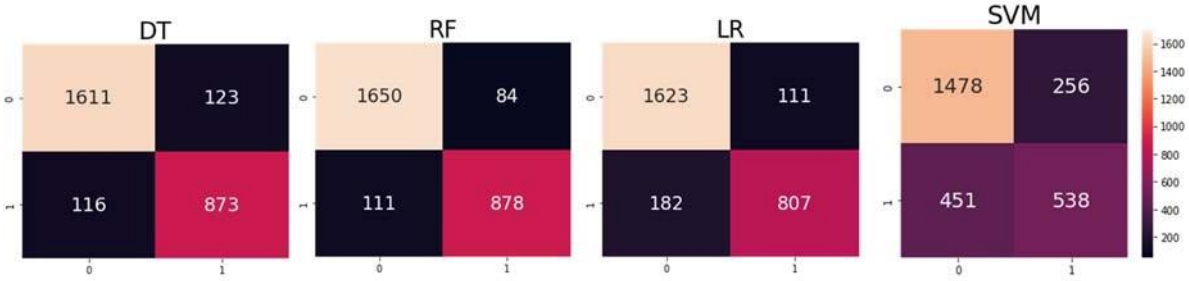
Following the data preparation phase, which included preprocessing and feature selection, the dataset was divided into training and testing sets. Considering various available options, the authors opted for a training-to-testing ratio of 70/30 based on its proven effectiveness in prior studies [24], [25]. Moreover, a cross-validation procedure was employed to assess the models' generalization capability.

All coding was executed in Python, utilizing standard machine learning libraries such as pandas for data preprocessing, Sklearn for model training and testing, and matplotlib for visualization purposes.

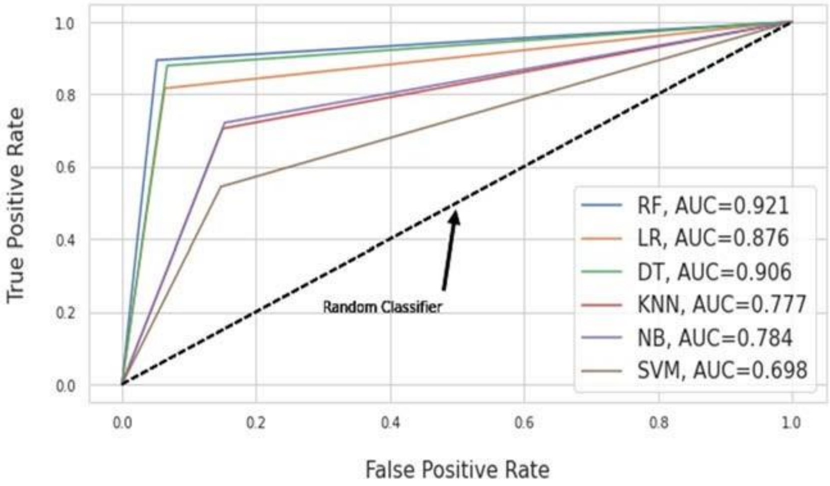
This study employs a pertinent dataset to forecast the likelihood of conversion for website users, leveraging historical user data to discern valuable sales opportunities from a vast pool of website visitors. The experimental outcomes are presented in Table 3.

	Models					
	NB	**RF**	**KNN**	**DT**	**LR**	**SVM**
TP	1468	1650	1472	1611	1623	1478
TN	713	878	694	873	807	538
FP	266	84	262	116	111	256
FN	276	111	252	123	182	451
Training Time (s)	0.01	0.93	0.10	0.05	0.25	8.97
Accuracy (%)	80.18	93.02	80.04	91.47	89.89	73.22
Precision (%)	74.14	92.25	74.56	88.37	89.28	68.25
Recall (%)	73.33	89.05	72.14	87.24	83.34	54.09
Specificity (%)	84.65	95.21	84.69	93.25	93.59	85.23
F1-Score (%)	89.89	84.63	87.97	71.59	72.45	60.34

Upon comparing the outcomes of the constructed models, it is evident that the Random Forest model surpasses all others across all metrics, attaining an accuracy score of 93.02%. Following closely are the Decision Tree and Logistic Regression models, achieving scores of 91.47% and 89.89%, respectively. Meanwhile, SVM and NB, owing to their straightforwardness, achieved scores of 80.18% and 73.22%, respectively. Figure 3 illustrates the confusion matrix of the top four constructed models.

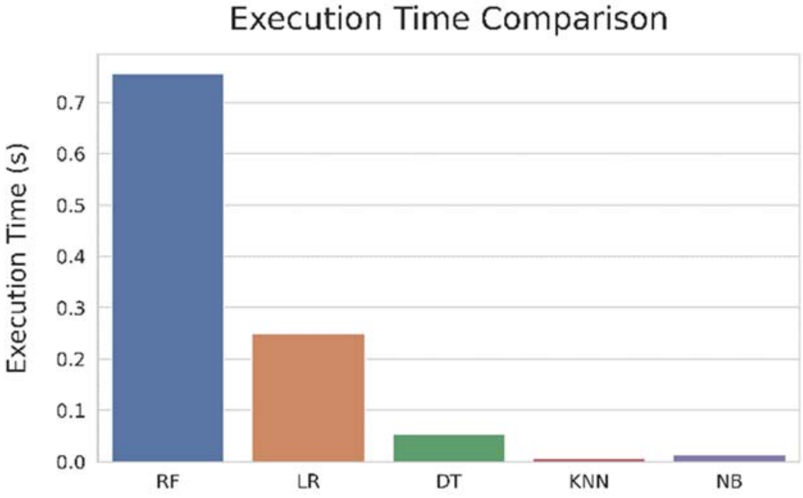


The Receiver Operator Characteristic (ROC) curve serves as a graphical performance metric for binary classification problems. The Area Under the Curve (AUC) quantifies the classifier's capability to differentiate between classes. Refer to Figure 4 for visualization.



The AUC values fall within the range of [0,1], where higher values indicate better ability to distinguish between different classes. In this study, the AUC scores of the compared models corroborated the findings of other metrics. The Random Forest (RF) model exhibited the highest AUC score of 0.92, followed by Logistic Regression (LR) with 0.87, and lastly Naïve Bayes (NB) and Support Vector Machine (SVM) with scores of 0.78 and 0.69 respectively. This further underscores the superior performance of the RF model in this task.

Considering the training time required, the Naïve Bayes model demonstrated the shortest training time at 0.1 seconds, followed closely by the Decision Tree (DT) model at 0.05 seconds. However, the Random Forest model necessitated significantly more time for training, clocking in at 0.93 seconds, followed by Logistic Regression at 0.25 seconds. The SVM model, on the other hand, exhibited a much longer training time, exceeding 8 seconds, likely attributable to the non-linear kernel utilized in the default configuration of the Sklearn model definition. For a visual representation, refer to Figure 5.



In conclusion, the Random Forest (RF) model emerged as the top performer in terms of classification results, demonstrating superior performance. Meanwhile, the Decision Tree (DT) model exhibited commendable performance while maintaining a shorter training time.

To evaluate the models' ability to generalize across the dataset, a cross-validation procedure was conducted, varying the k-value to explore the impact of different train/test size ratios on model bias. The results of K-folds cross-validation, presented in Table 4, mirrored those observed earlier. Across varying numbers of folds (k), both the RF and DT models consistently displayed superior performance and stability over the dataset.

Overall, the RF model maintained a clear advantage over the other models, closely followed by the DT model. However, when considering the criterion of execution time, the DT model becomes a more optimal choice compared to RF.

Conclusion

In conclusion, the advent of AI-driven marketing signifies a transformative shift in sales processes, ushering businesses into a new era of digital success. Through the integration of artificial intelligence technologies, such as machine learning and predictive analytics, organizations can harness vast amounts of data to gain valuable insights into consumer behavior, preferences, and trends.The application of AI in marketing facilitates more personalized and targeted approaches, enabling businesses to tailor their strategies and offerings to meet the individual needs of customers. By leveraging AI-powered tools for lead generation, lead scoring, and customer segmentation, companies can streamline their sales processes, identify lucrative opportunities, and optimize resource allocation.Furthermore, AI-driven marketing empowers businesses to enhance customer engagement and satisfaction through timely and relevant interactions across various digital channels. By leveraging AI-powered chatbots, virtual assistants, and recommendation engines, organizations can deliver personalized experiences, address customer inquiries promptly, and anticipate their needs effectively. Overall, AI-driven marketing represents a paradigm shift in sales processes, empowering businesses to thrive in the digital age by leveraging advanced technologies to drive growth, foster innovation, and deliver exceptional customer experiences. As AI continues to evolve and become increasingly sophisticated, organizations must embrace these technologies to stay competitive and capitalize on the opportunities presented by the digital landscape.

References

[1]. E. Brynjolfsson and K. McElheran, “The Rapid Adoption of Data-Driven Decision-Making,” American Economic Review, vol. 106, no. 5, pp. 133–39, May 2016, doi: 10.1257/AER.P20161016.

[2]. G. Shmueli and O. R. Koppius, “Predictive analytics in information systems research,” MIS Quarterly: Management Information Systems, vol. 35, no. 3, pp. 553–572, 2011, doi: 10.2307/23042796.

[3]. Ö. Artun and D. Levin, “Predictive Marketing,” Predictive Marketing, Aug. 2015, doi: 10.1002/9781119175803.

[4]. J. Järvinen and H. Taiminen, “Harnessing marketing automation for B2B content marketing,” Industrial Marketing Management, vol. 54, pp. 164–175, Apr. 2016, doi: 10.1016/J.INDMARMAN.2015.07.002.

[5]. W. K. Lin, S. J. Lin, and T. N. Yang, “Integrated Business Prestige and Artificial Intelligence for Corporate Decision Making in Dynamic Environments,” Cybernetics and Systems, vol. 48, no. 4, pp. 303–324, May 2017, doi: 10.1080/01969722.2017.1284533.

[6]. C. L. Pan, X. Bai, F. Li, D. Zhang, H. Chen, and Q. Lai, “How Business Intelligence Enables E-commerce: Breaking the Traditional E-commerce Mode and Driving the Transformation of Digital Economy,” Proceedings - 2nd International Conference on E-Commerce and Internet Technology, ECIT 2021, pp. 26–30, Mar. 2021, doi: 10.1109/ECIT52743.2021.00013.

[7]. A. Algi and Irwansyah, “Consumer trust and intention to buy in Indonesia instagram stores,” Proceedings - 2018 3rd International Conference on Information Technology, Information Systems and Electrical Engineering, ICITISEE 2018, pp. 199–203, Jul. 2018, doi: 10.1109/ICITISEE.2018.8721033.

[8]. M. B. Adam, “Improving complex sale cycles and performance by using machine learning and predictive analytics to understand the customer journey,” 2018, Accessed: Nov. 21, 2021. [Online]. Available: https://dspace.mit.edu/handle/1721.1/118010

[9]. R. Nygård and J. Mezei, “Automating lead scoring with machine learning: An experimental study,” in Proceedings of the Annual Hawaii International Conference on System Sciences, Jan. 2020, vol. 2020-Janua, pp. 1439–1448. doi: 10.24251/hicss.2020.177.

[10]. S. Singh, S. Madhwal, G. Datta, and L. Singh, “Modelling Search Habits on E-commerce Websites using Supervised Learning,” in Proceedings of the 8th International Advance Computing Conference, IACC 2018, Jul. 2018, pp. 53–58. doi: 10.1109/IADCC.2018.8692113.