

TOWARD A SMART LEAD SCORING SYSTEM USING MACHINE LEARNING

Aissam Jadli

AICSE Laboratory, ENSAM Casablanca, University Hasan II,
Casablanca, 20700, Morocco
aissam.jadli-etu@etu.univh2c.ma

Mohammed Hamim

AICSE Laboratory, ENSAM Casablanca, University Hasan II,
Casablanca, 20700, Morocco
mohammed.hamim-etu@etu.univh2c.ma

Mustapha Hain

AICSE Laboratory, ENSAM Casablanca, University Hasan II,
Casablanca, 20700, Morocco
mustapha.hain@etu.univh2c.ma

Anouar Hasbaoui

Department of Communication and management, FST Mohammadia, University Hassan II
Mohammadia, 20000, Morocco
anhasbaoui@yahoo.fr

Abstract

The segmentation of new commercial leads is a crucial task for modern and highly competitive businesses, to identify new profitable opportunities and enhance their Return On Investment (ROI). Business Lead scoring involves assigning a score (i.e., a buying probability) to each possible lead generated for the business. The interactions of these leads with the business marketing channels across the internet are converted into multiple attributes, including useful pieces of information (e.g., contact details, lead source, channel) and behavioral hints (e.g., reply speed, motion tracking). This process can help assess the quality of the opportunity and its position in the purchasing process. Furthermore, an accurate lead scoring process can help marketing and sales teams prioritize the selected leads and appropriately respond to them within an optimal time frame, increasing their propensity to become clients. The use of machine learning algorithms can help to automate this process. In this paper, the authors compared the performances of various ML algorithms to predict lead scores. The Random Forest and Decision Tree models have the highest accuracy scores of 93.02% and 91.47%, respectively, whereas the training time of the Decision Tree and Logistic Regression models was shorter, which can be a decisive factor when dealing with massive datasets.

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

1. Introduction

The lead interactions for Business to Consumer (B2C) sales activities can be roughly divided into two categories: lead generation and lead conversion. The process starts by approaching potential clients through different channels (e.g., website, social media, campaigns), with the purpose to attract them to the business's website, and encourage them to interact with it. These actions are monitored by an automated system, which contributes to nurturing the business's database. Finally, the lead is approached by a professional sales agent who will assist the lead to take the purchasing decision and eventually solve any problems and hurdles the client may encounter, he will use all the marketing tactics and financial incentives (e.g., coupons, reductions) to seal the deal.

But somewhere in between, there's a process of understanding the valuable visitors for the business. Since sales operation has a high cost — both in time and money —, it should focus on nurturing only the most engaged and fitted leads in order to improve and maintain a profitable "Return on Investment" (ROI). This process is called lead scoring; it allows the business, using data analytics, to accurately predict the weight of each potential lead [1].

The determination of assigned weight to each feature is a milestone step in the scoring process. In a classical scoring model, these weights are inferred with a "try and guess" approach and the assistance of marketing experts to find their optimal values.

With the rise of artificial intelligence applications in several domains, the lead scoring strategies have taken new horizons using predictive modeling [2]. Machine learning algorithms distinguish between failed leads and successful ones (those who make buying decisions), algorithms look for common information of converted leads in order to determine a formula that will automatically identify them to marketing teams.

In this paper, the authors compare the performances of various predictive ML models. The used dataset is a public leads dataset for lead scoring available online. Section 2 will present a brief review of the B2C internal working and lead scoring mechanism, as well as the current state of research in this field. Section 3 will display the experiment starting with the dataset details, the used ML algorithms, and the experiment proceedings. Section 4 will present and discuss the experimental results along with the assessment of the models' accuracy using different metrics. Finally, the conclusion will propose the implementations of such an approach in the real world and discuss the future forecasts of this work.

2. Paper Context

Lead scoring is a marketing technique that helps decision-makers to identify the more profitably potential customers among the generated leads. As Result, the salesforce professionals will not waste time randomly contacting all prospects and will only concentrate their efforts on more likely converted ones.

The main idea is to assign scores to all prospects based on how their characteristics match with the pre-established profile of a converted customer. The leads that score above a specific threshold are considered as ideal target. The bottleneck of such an approach is the determination of profile-relevant attributes with its respective weights.

2.1. Traditional Lead Scoring

In Traditional (i.e., Manual) lead scoring, the task of finding the most relevant customer traits and points assignation is handled either by a marketing expert or senior sales department manager. In fact, the manager's experience can help to determine the most optimal key attributes among thousands of possible features. Traditionally, scores are calculated based on how well the lead fits the company's client portrait (demographic details) and their commitment (behavioral details). Figure 1 presents the architecture of a traditional lead scoring system.

Due to human nature, the sales manager cannot always define precisely what features are more important, his behavior can be biased or based on prior prejudices, he also tends to trust acquired knowledge and rarely update



Fig. 1. Traditional Lead Scoring System

the selection process.

2.2. Smart Lead Scoring

The concept of predictive lead scoring is based on a statistical approach called propensity modeling[3]. This technique attempts to predict the chances that a visitor will perform certain actions (purchase, reservation, etc.)

It tries -combined with machine learning and data mining- to forecast the behavior of the targeted audiences and the likelihood of successive conversion [4].

Machine Learning (ML) algorithms can detect automatically the relations and hidden rules in historical sales data [5][6] to pick relevant attributes and discover useful patterns that indicate a lead's propensity to conversion. The constructed model is trained and evaluated to give the lowest possible false-positive predictions. If an error/anomaly is perceived in a prediction, it can be annotated and re-added to the training data which allows the model to adjust itself and stay relevant, especially in growing businesses. Figure 2 presents the architecture of an ML-based Predictive Lead Scoring System

The main difference between traditional and predictive lead scoring models is the ability to manage a large volume of data and to gather better insights for performance improvement [7]. Moreover, there's a limit to our perceptiveness and ability to find sense in thousands of data points and detect relations and rules between them. The use of machine learning in propensity prediction allows replacing a professional expert marketer

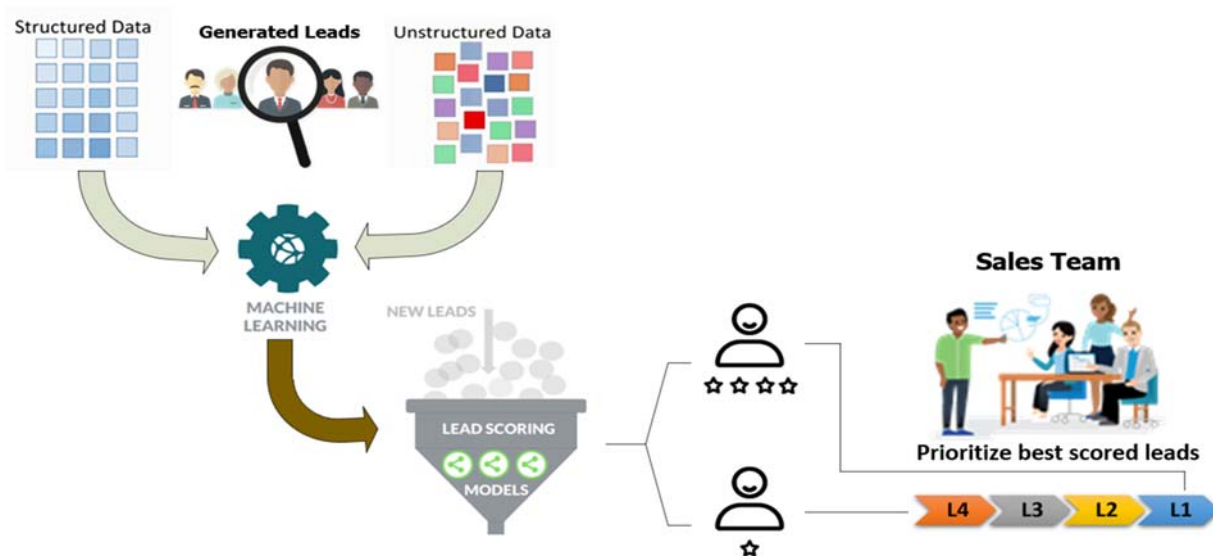


Fig. 2. ML-based Predictive Lead Scoring System

(potentially expensive) with an automatic system with comparable judgment skills which will achieve great improvements over time with big data adoption [8]. See Table I for key differences 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

Table 1. Key Differences between Lead Scoring Systems.

3. Related Works

The propensity of B2B and B2C leads is a research area that recently has been thoughtfully studied by researchers because of its great impact on sales efficiency and internal workflow optimization in customer handling.

R. Nygård et al. [9] suggested a supervised learning approach to lead scoring based on algorithms such as Logistic Regression and Decision Trees to predict the purchase probability (using prior knowledge and behavioral data). The authors found the algorithm that produces the best performances is the Random Forest model.

S. Singh et al. [10] proposed modeling search habits of commercial websites visitors using supervised ML algorithms to detect and extract the shopping patterns, and the shift in trends among these visitors.

K. Prasad et al. [11] proposed a comparative analysis between SVM and Logistic Regression algorithms in building models for propensity prediction and evaluated their performance.

S. Mortensen et al. [12] used structured and unstructured data from a paper & packaging company's IT system to predict B2C sales success. The authors compared several algorithms including Binomial logistic regression with different decision tree approaches (i.e., gradient boosting and random forest). The best-constructed model showed a propensity accuracy of 80%, with precision and recall scores respectively of 86% and 77%.

Y. Zhang et al. [13] proposed to identify the most valuable prospects through the use of machine learning. The authors compared the predictive capacity of logistic regression and random forest, their results showed that the latter model was more accurate than the former one. In regards to the recall rate, F1 score, and Receiver Operating Characteristic (ROC), the Logistic Regression model outperformed the other one.

Y. Benhaddou et al. [14] proposed a solution to the problem of training small datasets through the building of a Lead Scoring model with a Bayesian network. In the context of their paper, the authors suggested the construction of the Lead Scoring model from the expertise and applying usual heuristics (parent divorcing, NoisyOr) to reduce the complexity of the model.

A. Etminan[15] aimed to estimate the effect of feature weights by assessing some feature ranking and judging schemes in a predictive lead scoring scenario.

J. Yan et al. [16] propose a unified, ML-based framework for marketing opportunity propensity estimation, pursuing solutions to several challenges in the B2B environment (e.g., the difference in transactions volume between B2B and B2C companies, noisy data, and the fast-changing market environment).

A. Rezazadeh [17] tackled the problem of forecasting the result of B2B and B2C sales by proposing a data-driven, ML-based pipeline in a cloud environment. The author concluded that decision-making based on the ML predictions is more accurate and brings a higher monetary value than the traditional, operator-based, approach.

A. Sabbani et al. [18] proposed a new approach for seller-buyer matching —when attending a trade show event— based on Machine Learning. The authors suggested an automated approach by replacing the syntactic analysis of the interests of the buyer with implicit user feedback on a frontend intelligent application (website).

Whereas, the contribution of this paper can be summarized as the following:

- The authors tried different algorithms (six in total) to verify that RF is the best one for this task as the literature review states.
- The authors used a variety of metrics and validation approaches than using only accuracy criteria to assess the models' performance.
- The authors introduced the processing time and computing power as a useful criteria in model selection to maintain stable performance over big datasets.

4. Materials and Methods

In this paper, the generally recommended framework for predictive modeling and smart analytics research is adopted. This being said, dataset understanding is an equally important step that focuses on investigating the dataset and identifying and rectifying potential problems within it. Firstly, there is the data preparation process; it consists of dealing with the missing values, outliers detection, and creating a relevant feature vector using different techniques, such as feature selection and feature extraction, this process is very fundamental to optimizing the machine learning model construction. Furthermore, several models are implemented and evaluated. Aftermath, the performance of each model is analyzed and interpreted.

4.1. Dataset

4.1.1. Data Description

The main purpose of this paper is to demonstrate the benefits of machine learning in automating lead scoring process with the implementation of predictive modeling. To do so, we experimented using a public dataset entitled "X Education" widely used in the lead's prediction process. The dataset contains several variables that cover the following aspects:

- The outcome of the Lead (Converted or not).
- Visitor actions on the website (e.g., pages visited, time spent).
- Pieces of information collected through website forms (e.g., contact, newsletter).
- Lead source (search engine, referrer, direct).

The Dataset contains 9240 data points with 37 features. These features are properly stored for each prospect to describe its characteristics. Some features have numerical values (6 features), such as the website's visit

duration and visit frequency, others are categorical such as search keywords, lead source, and contact preferences.

4.1.2. Data Description

Using various data preprocessing techniques, we extract 89 features from the raw dataset which had 37 initial features. Depending on the type and values distribution of each feature, a suitable processing pipeline was utilized:

- Features with a missing value ratio superior to 70% were dropped from the dataset due to little variance gain and distribution skewness risk. Similarly, data points with many missing attributes can be removed for the same reasons.
- Features with a low missing value ratio need feature-specific processing to replace the missing values with an adequate value (mean, median, mode, etc.)
- For categorical features (e.g., Lead Origin, Specialization), a One Hot Encoding process is used to construct a one-hot encoded vector that represents a coded form of the value, thus increasing the number of features from 37 to 89.

The resulting dataset has a shape of (9074, 89) with no missing values.

4.2. Techniques

After the data preprocessing step with several operations and enhancement, we selected six machine learning algorithms, the most widely used in the literature review in customer classification tasks and predictive modeling.

4.2.1. K-Nearest Neighbors

KNN is a simple, non-parametric, and easy-to-use classification algorithm first presented by E. Fix and J. Hodges in 1951, and later expanded by T. Cover [19]. An instance class label is determined by a majority vote rule of its close neighbors, with the class being assigned to be the most common amongst its K nearest neighbors measured by a distance measurement. Different metrics (distances) exist for different use cases and sample distributions, such as:

- Euclidean Distance: $\sqrt{\sum_{i=1}^k (x_i - y_i)^2}$
- Manhattan Distance: $\sum_{i=1}^k |x_i - y_i|$
- Minkowski Distance: $(\sum_{i=1}^k (|x_i - y_i|^q))^{1/q}$

The distance used in this paper was Minkowski's distance which is the default distance function used in Sklearn.

List may be presented with each item marked by bullets and numbers.

4.2.2. Naïve Bayes

NB is a simple, yet effective conditional probability model for classification based on the Bayesian theorem with a naïve assumption about features independence. Given a single instance represented by a vector with n relevant independent features to be classified, the algorithm assigns the probability for each of the K's possible outcomes or class (i.e., win or lost sales opportunity) [20]. Using Bayes' theorem, the conditional probability can be decomposed as (1):

$$\rho(C_k|x) = \frac{\rho(x|C_k)\rho(C_k)}{p(x)} \quad (1)$$

The NB classifier combines the Naive Bayes probabilistic model with a decision rule[21]. The rule used usually is to pick the most probable hypothesis; this is known as the Maximum A Posteriori (MAP) decision rule. Thus, the corresponding Bayes classifier is the function that gives the instance class for some k as follows:

$$y =_{(k \in 1, \dots, K)} \rho(C_k) \quad (2)$$

4.2.3. Support Vector Machine

SVM is a binary classification algorithm constructed first by V. Vapnik[22], which tried to find a hyperplane in a higher N-dimensional features space to separate instances into two distinct classes. It is a preferred classifier for different situations, because it produces significant performances with less computational power.

Even if SVM is a binary classification algorithm, it can perform multiclass classifications by combining several binary classifications in a “one-against-all” strategy. Given k (by the number of classes) similar “one-against-all” optimization tasks:

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2} w^T w + C \sum_{i=1}^I \xi_i \\ & w^T \phi(x_i) + b \geq 1 - \xi_i, \quad \text{if } y_i = m; \\ & w^T \phi(x_i) + b \leq -1 + \xi_i, \quad \text{if } y_i \neq m; \\ & \xi_i \geq 0 \end{aligned} \quad (3)$$

where y_i is the class label of x_i , ϕ is the kernel function, and C is the penalty parameter. Thus, the class of an instance is obtained by (4):

$$c_{SVM} = \operatorname{argmax}_{c \in C} ((w^c)^T \phi(x) + b^c) \quad (4)$$

4.2.4. Decision Tree and Random Forest

A Decision Tree is a family of decision-making algorithms designed as a graph or a “tree” in which each decision (non-leaf) node is labeled given an input feature and a probability. Mathematically, the Entropy for multiple attributes is represented as (5):

$$E(T, X) = \sum_{c \in X} P(c) E(c) \quad (5)$$

The Decision Tree Classifier starts by breaking down a given dataset into smaller subsets, whereas the classification rule in each level is memorized to build incrementally a decision graph[23]. Each node of the tree is labeled with a feature and a probability distribution over the classes, the data set has been classified by the tree into a specific class, based on different metrics such as GINI index and Information Gain.

The Random Forest represents an example of an integrated learning approach that tries to solve a single prediction problem, by building several model combinations. It works by generating multiple classifiers/models that learn and independently make predictions of each other. These predictions are combined into a single prediction, outperforming single classification algorithms.

Ultimately, a Random Forest is a combination of several DT models, while its output is determined by the majority of the categories of individual DT output.

4.2.5. Logistic Regression

Despite its name, LR is a simple yet highly effective binary and linear classification algorithm. It is widely used in industry and research works whenever linear classification is preferred. The algorithm maps a latent features vector to a value range of [0, 1] using a sigmoid function. The training speed of an LR model can be faster than complex algorithms such as Artificial Neural Network (ANN) and Random Forest whereas maintaining decent and comparable performances. Because of its simplicity and its performance, LR is still very appealing to researchers in various scientific areas. In the presence of big data, scientists believe that the algorithms showing promising classification scores manifest a hold-up of training speed. Therefore, LR is the most optimized, comprehensive, and cost-effective classification algorithms.

The formula to calculate the probability of a sample x_i belonging to a category C_i using a multiclassification LR model can be written as (6):

$$\rho(C_i|x) = \frac{e^{w_i^T x + w_{0i}}}{\sum_{j=1}^K e^{w_j^T x + w_{0j}}}, i = 1, \dots, K \quad (6)$$

The bias vector and the weight matrix are the parameters of the constructed model. They can be obtained by minimizing the loss function defined as (7):

$$l(\theta = w, w_0, D) = -\sum_{i=1}^{|D|} \log \rho(y^{(i)} | x^{(i)}) \quad (7)$$

4.3. Model Evaluation

4.3.1. First Level indicators

Different indicators can be used to evaluate the performances of an ML model. Using the testing dataset and the model's prediction for the same dataset, we obtain four basic indicators called first-level indicators. Based on these indicators we construct a matrix called the confusion matrix. Table 2 presents the structure of a confusion matrix.

Confusion Matrix		Actual values	
		Positive	Negative
Predicted Value	Positive	TP	FP
	Negative	FN	TN

Table 2. The structure of the confusion matrix.

The first level indicators give a decent idea about the model's performance and are easy to interpret. Generally, a model needs to enhance its TP and TN scores and minimize its FN and FP scores.

4.3.2. Second Level indicators

For more accurate model performance metrics, researchers extracted more sophisticated scores based on the previous indicators to assess models' performances. These indicators are called secondary indicators and are calculated based on the first-level indicators. The secondary indicators usually used are:

- Accuracy: The proportion of the number of correct predictions among the entire predictions.
- Precision: The proportion of the number of correct positive predictions among the entire positive predictions.
- Sensitivity or Recall: The proportion of the number of correct positive predictions among the entire positives
- F1-Score: The f1-score is a popular classification metric usually preferred over accuracy when data is unbalanced (i.e., when the quantity of samples belonging to a class is significantly greater than in the other classes).
- Specificity: The proportion of the number of correct positive predictions among the entire negative cases.

4.3.3. K-folds Cross-validation

Cross-validation is a popular and easy-to-use statistical method used to evaluate the generalization skill of an ML model. The procedure is based on resampling the entire dataset after splitting it into several groups of the same size (k) and selecting each time a group of them as test data. K-Fold Cross-validation tends to produce a less biased model because it allows every data point to participate in the building of the model.

5. Results and Discussion

After the data preparation step (with data pre-processing and relevant feature selection), the authors split the dataset into a training and testing dataset. After doing running an extensive analysis for available solutions, the authors choose the ratio training/test of 70/30 due to its optimal results found in the literature[24], [25]. Additionally, the authors applied a cross-validation procedure to check the model's generalization ability.

All code was implemented using Python Language with standard Machine Learning libraries (pandas for data preprocessing, Sklearn for models training and testing, and matplotlib for visualization).

This paper uses the relevant dataset to predict the conversion likelihood of website users, based on the historical data of the users, it allows to identify valuable sales opportunities from a large number of website visitors. The results of the experiment are shown 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

Table 3. The Experiment Results.

Comparing the results of constructed models, it is clear that the Random Forest model is outperforming all other models in all metrics, achieving an accuracy score of 93.02%, followed by the Decision Tree and the Logistic Regression models with respectively 91.47% and 89.89%. In the meantime, SVM and NB, due to their simplicity, achieved respectively 80.18% and 73.22%. Figure 3 shows the confusion matrix of the four best-constructed models.

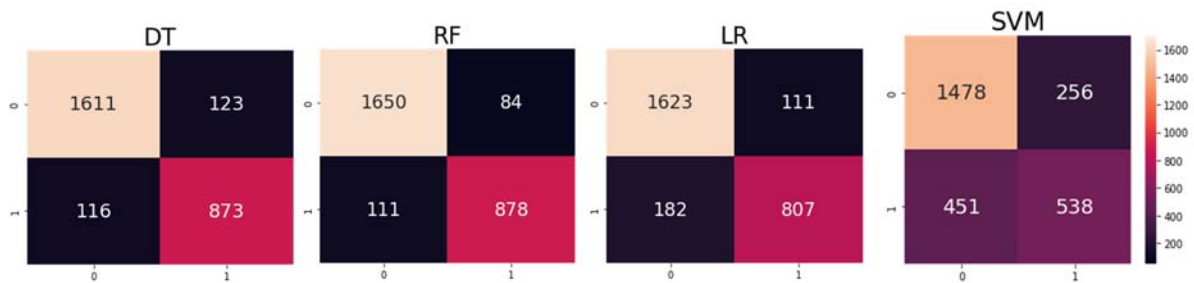


Fig. 1: The confusion matrix of the four best models

The Receiver Operator Characteristic (ROC) curve is a graphical performance metric for binary classification problems. The Area Under the Curve (AUC) is the measurement of the ability of a classifier to distinguish between classes. See Figure 4.

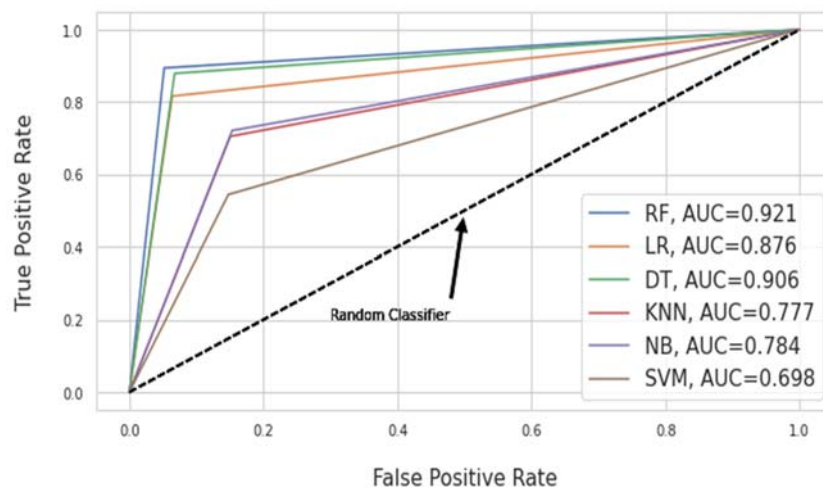


Fig. 4: AUC Comparison

The values of AUC are ranged in [0,1] where the higher the AUC, the better the model can distinguish between the different classes. In this case, the AUC score of compared models confirmed the results showed by other metrics as the RF had the best AUC score (0.92) followed by LR (0.87) and finally NB (0.78) and SVM (0.69), which confirm the superiority of RF model in this task.

When comparing the models using the training time required criteria, the NB model was the fastest to train (0.1s) followed by the DT model (0.05s) while the RF model took a significant additional time (0.93s) for training, followed by the LR model (0.25s). The difference in training time between the models can lead to performance considerations in very large datasets, while the SVM model was too long to train (over 8s) most probably due to the non-linear kernel used in the default configuration of the Sklearn model definition. See Figure 5.

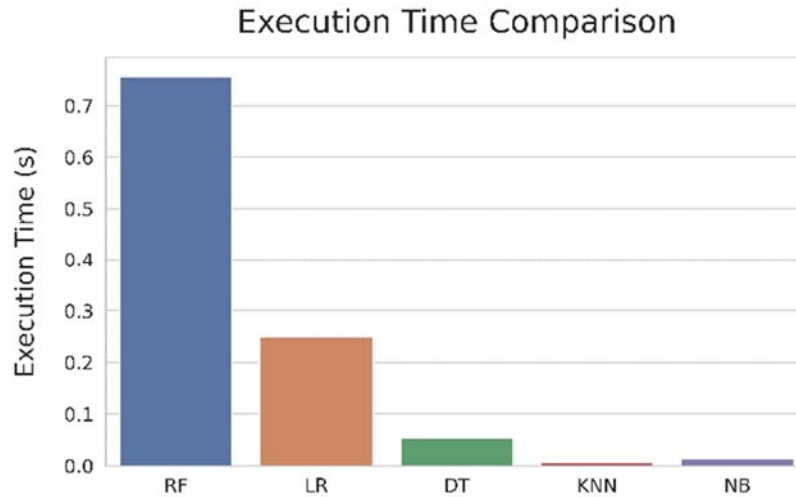


Fig. 5: AUC Comparison

Ultimately, the RF model had the best classification results performance-wise, whereas the DT model was able to provide decent performance and to keep a reduced training time.

A cross-validation procedure was executed on the dataset to assess the model's ability to generalize over the whole dataset. By varying the k-value, we can test the influence of the ratio train/test size on the overall bias of the model. Table 4 presents the results of K-folds cross-validation. The cross-validation procedure gave consistent results with values observed in the first part. Even after varying the number of folds (k), the RF and DT models came on top with superior performance and stability over the dataset. Ultimately, the RF model had a clear advantage over the other models, followed by the DT model. However, when we introduce the execution time criterion, DT becomes more optimal to use than RF.

6. Conclusion

Predictive lead scoring is a hot research topic for businesses exploring new opportunities. In this paper, the authors focused on investigating the use of machine learning algorithms to automate the process of prospect identification in order to replace the traditional scoring system. Different ML algorithms have been used, such as Logistic Regression (LR) and Random Forest (RF), to classify potential leads into qualified and unqualified leads. The two top models (RF and DT) have both achieved good performance, and they can predict, with an accuracy of 93.02% and 91.47% and a precision of 92.25% and 88.37%, respectively, whether visitors will place an order according to historical and behavioral data on the business website. Regarding execution time, decision tree and logistic regression models showed significantly shorter training times, which can be a decisive factor when dealing with massive datasets, especially in the Big Data era. This can help businesses to identify high-quality prospects from ordinary website visitors, thereby helping them to improve their ROI (Return on Investment).

The authors plan to use techniques such as Transfer Learning and Reinforcement Learning in their future work to investigate the use of Deep Learning models to solve other problems regarding other aspects of the scoring process, such as insufficient training data or imbalanced dataset classes.




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.
- [11] K. V. N. K. Prasad and G. V. S. R. Anjaneyulu, "A Comparative Analysis of Support Vector Machines & Logistic Regression for Propensity Based Response Modeling," *International Journal of Business Analytics and Intelligence*, vol. 3, no. 1, Jun. 2015, doi: 10.21863/ijbai/2015.3.1.002.
- [12] S. Mortensen, M. Christison, B. C. Li, A. L. Zhu, and R. Venkatesan, "Predicting and defining B2B sales success with machine learning," Apr. 2019. doi: 10.1109/SIEDS.2019.8735638.
- [13] Y. Zhang, "Prediction of Customer Propensity Based on Machine Learning," in *Proceedings - 2021 Asia-Pacific Conference on Communications Technology and Computer Science, ACCTCS 2021*, Jan. 2021, pp. 5–9. doi: 10.1109/ACCTCS52002.2021.00009.
- [14] Y. Benhaddou and P. Leray, "Customer Relationship Management and Small Data — Application of Bayesian Network Elicitation Techniques for Building a Lead Scoring Model," in *2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)*, Oct. 2017, vol. 2017-Octob, pp. 251–255. doi: 10.1109/AICCSA.2017.51.
- [15] A. Etminan, "Prediction of Lead Conversion With Imbalanced Data : A method based on Predictive Lead Scoring," 2021, Accessed: Oct. 01, 2021. [Online]. Available: <http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-176433>
- [16] J. Yan, M. Gong, C. Sun, J. Huang, and S. M. Chu, "Sales pipeline win propensity prediction: A regression approach," in *2015 IFIP/IEEE International Symposium on Integrated Network Management (IM)*, May 2015, pp. 854–857. doi: 10.1109/INM.2015.7140393.
- [17] A. Rezazadeh, "A Generalized Flow for B2B Sales Predictive Modeling: An Azure Machine-Learning Approach," *Forecasting*, vol. 2, no. 3, pp. 267–283, Aug. 2020, doi: 10.3390/forecast2030015.
- [18] A. Sabbani and A. el Haddadi, "Business matching for event management and marketing in mass based on predictive algorithms," in *Proceedings - 15th International Conference on Signal Image Technology and Internet Based Systems, SISITS 2019*, Nov. 2019, pp. 619–626. doi: 10.1109/SITIS.2019.00102.
- [19] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *American Statistician*, vol. 46, no. 3, pp. 175–185, 1992, doi: 10.1080/00031305.1992.10475879.
- [20] A. Jadli, M. Hain, and A. Hasbaoui, "An Improved Document Image Classification using Deep Transfer Learning and Feature Reduction," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 10, no. 2, pp. 549–557, doi: 10.30534/ijatcse/2021/141022021.
- [21] A. JADLI, M. HAIN, A. CHERGUI, and A. JAIZE, "DCGAN-Based Data Augmentation for Document Classification," in *2020 IEEE 2nd International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS)*, Dec. 2020, pp. 1–5. doi: 10.1109/ICECOCS50124.2020.9314379.
- [22] C. Cortes, "Support-Vector Networks," 1995.
- [23] M. Karim, R. M. Rahman, M. Karim, and R. M. Rahman, "Decision Tree and Naïve Bayes Algorithm for Classification and Generation of Actionable Knowledge for Direct Marketing," *Journal of Software Engineering and Applications*, vol. 6, no. 4, pp. 196–206, Apr. 2013, doi: 10.4236/JSEA.2013.64025.
- [24] J. Tan, J. Yang, S. Wu, G. Chen, and J. Zhao, "A critical look at the current train/test split in machine learning," Jun. 2021, Accessed: Jan. 16, 2022. [Online]. Available: <https://arxiv.org/abs/2106.04525v1>
- [25] B. Vrigazova, "The Proportion into Training and Test Set for the Bootstrap in Classification Problems," *Business Systems Research*, vol. 12, no. 1, p. 2021, doi: 10.2478/bsrj-2021-0015.

BIOGRAPHIES OF AUTHORS



Aissam Jadli    Aissam Jadli is a Ph. D candidate at Ensam Casablanca with Master Degree in Computer Science from University Hassan II in 2018. He obtained Bachelor's Degree in Mathematics and Information Technology from the Faculty of Science of Agadir in 2013. His researches are in the fields of ERP, computer vision, machine learning, and artificial intelligence. He is affiliated with several scientific journals and conferences as an invited reviewer or Board member. Besides, he is also involved in NGOs and student associations. Further details on his ORCID Page: <https://orcid.org/0000-0002-9783-0734>



Mustapha Hain  Mustapha Hain is a Ph.D. research professor at ENSAM Casablanca from University Hassan II and the Director of the Artificial Intelligence & Complex Systems Engineering (AICSE) laboratory in ENSAM Casablanca. His researches are in the fields of digital systems, E-logistics, software modeling, machine learning, and artificial intelligence. He has served as an invited reviewer for several journals and conferences. Besides, he is also involved in NGOs and student associations.



Anouar Hasbaoui, Anouar Hasbaoui is an Economist and a Doctor of Philosophy in Finance working actually with the Moroccan Government. His researches are in fields such as banking and support services, Risk management and stocks prediction.



Mohammed Hamim ^{SC}, Mohammed Hamim is currently a Ph.D. Candidate at ENSAM-Casablanca, University Hassan II. He received his Master's degree in “Big Data & Internet of Things” and his bachelor's in Computer science. His research specialty and interests include Machine Learning, High-dimensional data modeling, Data Analysis, Deep Learning, data visualization, as well as Computer Science.