Vol 5 No 1 (2024) E-ISSN: 2735-1009

International Journal of Management, Finance and Accounting

Enhancing Conversions and Lead Scoring in Online Professional Education

Wen Yang Yim[1], Khai Wah Khaw[1,*],Shiuh Tong Lim[1], XinYing Chew[2]

[*]Corresponding author: khaiwah@usm.my
[1]School of Management, Universiti Sains Malaysia, 11800 Penang, Malaysia
[2]School of Computer Sciences, Universiti Sains Malaysia, 11800 Penang, Malaysia

Abstract

This study seeks to enhance lead conversion for online professional education providers by using supervised machine learning algorithms for lead conversion targeting and lead scoring, including Logistic Regression, K-Nearest Neighbors, Support Vector Machines, Naïve Bayes, Random Forst, Bagging, Boosting, and Stacking. A lead dataset was used to train and test the machine-learning models. The Recursive Feature Elimination (RFE) is used to establish a precise lead profile. The performance of the trained lead conversion models was evaluated and compared using the 10-Folds cross-validation method based on accuracy, precision, recall, and F1-score. The results show that Stacking is the best model with an accuracy of 0.9233, precision of 0.9391, and F1-score of 0.8939. Meanwhile, the Logistic Regression-based lead scoring model demonstrated promising potential for automating lead scoring. The results of the Logistic Regression-based lead scoring model achieved an accuracy of 0.9019, recall of 0.9019, precision of 0.9015, and F1-score of 0.9014. The optimal lead scoring threshold is 0.20, which stroked the optimal trade-off balance between accuracy, sensitivity, and specificity.

Keywords: Machine Learning, Lead Conversion, Lead Scoring

Submitted on 15 June 2023; accepted on 10 November 2023; Published on 29 February 2024.

Published by MMU PRESS. URL https://journals.mmupress.com/index.php/ijomfa/index

1. Introduction

The advancement of information and digital technologies has enabled a bidirectional communication channel and has catalyzed inbound marketing. This marketing approach aims to engage customers via astute and transparent communication with those who have expressed interest in the company's products or services (Jabbouri, 2023). Online education has been propelled by technological advancement, the COVID-19 pandemic, and the industry's demand for skilled labor, which is projected to equip approximately 250 million workers by 2030 (Rahmat et al., 2021; Blyzniuk et al., 2021; Kumar et al., 2017). Moreover, online education is experiencing significant growth in various countries, including China, South Korea, and Malaysia (Kumar et al., 2017). According to Eurostat 2022, 27% of individuals aged between 16 to 74 completed an online course or used online learning resources in 2021.

As with other businesses, online education providers can benefit from incorporating lead scoring into their marketing and sales strategies to capitalize on the trend. Lead scoring can improve overall business performance by lowering costs and increasing profits (Banerjee & Bhardwaj, 2019). Besides, GE Capital improved 30-50% of salespeople's productivity through automated lead scoring (Wu et al., 2023). According to G. N. Kumar and Hariharanath (2021), lead scoring enables online education providers to prioritize and efficiently manage their leads, resulting in improved conversion rates and revenue generation.

The definition of lead management varies across the literature. Generally, it encompasses a range of activities and techniques aimed at managing, processing, and influencing leads, such as lead generation, lead nurturing, and lead conversion (Monat, 2011). The lead management ensures leads are sufficiently nurtured and engaged so leads are more promising to convert into paying customers when approached by the sales department (Banerjee & Bhardwaj, 2019). Hence, effective lead management relies on accurate lead identification, strategic evaluation of leads, and judicious decision-making to meet customers' needs and expectations (Espadinha-Cruz et al., 2021).

Meanwhile, lead scoring is crucial in enhancing the lead funnel's effectiveness and supporting the lead nurturing process. Lead scoring is a marketing strategy that aids decision-makers in identifying and prioritizing the leads with the greatest profit potential (Zumstein et al., 2021). Lead scoring is related to assessing a lead's score, determined by every online and offline interaction with the customer (Nair & Gupta, 2021; Jadli et al., 2022). Consequently, salesforce professionals will not spend time randomly contacting all leads but rather have a sound reference to focus their efforts on the leads with the highest likelihood of conversion or 'hot leads' (Jadli et al., 2022). Poor lead scoring can adversely impact the efficacy of lead management.

In this sense, identifying quality leads is a crucial intermediary step between lead generation and conversion. However, there is a lack of consensus or a weak theoretical foundation to support hot leads characterization and lead scoring practice (Wu et al., 2023; Espadinha-Cruz et al., 2021; Monat, 2011). Conventional lead scoring primarily depends on marketing professionals' and sales executives' experience and understanding (Jadli et al., 2022; Wu et al., 2023). The less statistically supported conventional lead scoring, such as manual scoring and demographic segmentation, may be insufficient in discerning quality leads, resulting in a low lead conversion rate (Zumstein et al., 2021; Banerjee & Bhardwaj, 2019). Moreover, the emphasis on demographic and firmographic data can distort results and fail to adapt to the changing business and online environment without periodic reassessment (Nygård & Mezei, 2020; Eitle & Buxmann, 2019). The lead scoring model, including the features, must be updated periodically to reflect changing customer behaviours and characteristics due to the dynamic nature of online lead data, such as clickstreams, social media, traffic, and sensor data (Alfian et al., 2019).

Furthermore, companies invest heavily in advertisements, web campaigns, and marketing to generate leads and allocate enormous resources to lead nurturing and lead conversion (Nygård & Mezei, 2020; Ohiomah et al., 2019; Espadinha-Cruz et al., 2021). Errors in evaluating lead quality can result in overestimating resource reservations, directing resources toward low-quality leads, and inefficiently organizing marketing and promotional endeavours (Nair & Gupta, 2021). Qualified leads that fail to result in timely sales often slip through, eventually becoming lost revenue opportunities (Wu et al., 2023; Jadli et al., 2022; Espadinha-Cruz et al., 2021). The lead score threshold inaccurately reflects customer behaviour, particularly regarding their interest level and purchasing intentions. In addition, the average of qualified leads is approximately 10.0% and only

1.0% - 6.0% of qualified leads are converted (Gopalakrishna et al., 2022; Wu et al., 2023).

As a result, companies must employ an effective lead management strategy that can cope with the rapidly changing environment. Several studies have stated that applying predictive algorithm-based lead scoring improves overall performance (Zumstein et al., 2021; Haleem et al., 2022; Wu et al., 2023). This study focuses on developing a predictive-based lead management model to target “hot lead” and evaluate lead score using machine learning to manage lead better and ultimately enhance lead conversion for online education.

2. Literature Review

This section reviews recent literature on the context for developing lead conversion targeting and leads scoring models using supervised machine learning techniques.

2.1 Conversion Targeting

“Lead” pertains to an intentional or unintentional recorded indication of interest in a company’s goods or services, whether from a new potential or an existing customer (Monat, 2011; Jabbouri, 2023). Lead is generated when a potential contact satisfies the company-predetermined lead criterion, usually forming interest documentation (Espadinha-Cruz et al., 2021); transformed to cold lead once minimal lead information is collected (Monat, 2011; Priya V, L. 2020). For a lead to be classified as a “marketing qualified lead” (MQL) or “hot lead,” it must undergo a lead funnel conversion process: lead generation, lead nurturing, and lead qualification (Espadinha-Cruz et al., 2021). MQL during the process is determined by predetermined engagement and behavioural focus criteria, usually represented by a predefined score threshold, and then handover to the sales department (Terho et al., 2022; Nygård & Mezei, 2020).

The study’s conversion targeting modelling aims to discern and focus on discovering the use of machine learning algorithms to determine factors (features) contributing to lead conversion and identify hot leads. The weak theoretical consensus delineating hot lead attributes and the complex lead funnel activities can be time-consuming and inefficient. Moreover, lead conversion is influenced by many factors, including but not limited to the marketing environment, lead characteristics, and sales channels (Espadinha-Cruz et al., 2021). Through machine learning, variable importance techniques can identify the most significant attributes or predictor variables to predict the hot lead (Gouveia & Costa, 2022). Marketers can pinpoint and prioritize a lead's actions or behaviours that indicate a higher likelihood of conversion or purchase. It can potentially build a more accurate and statistically backed hot lead profile, enabling precise identification, targeting, and informed decisions.

2.2 Predictive Lead Scoring

Predictive lead scoring uses advanced data-driven analytics to find patterns and insights in lead data conceptualized by propensity modelling (Jadli et al., 2022). The information contained in the lead must be relevant to the factors that influence the lead's purchase timing and intention. By identifying hidden patterns in the big data, the approach estimates the propensity score for each lead (Nygård & Mezei, 2020). Therefore, the propensity score can be considered a lead score (Nygård & Mezei, 2020; G. N. Kumar & Hariharanath, 2021) and an ideal timing indicator for the sales department to initiate contact with the leads. This study posits that machine learning can model the fundamental data pattern that mirrors the lead's features and behaviours to substitute the manual lead scoring, whereby a more statistically supported reflection of lead behaviour to support informed decisions.

Predictive-based lead scoring using machine learning is believed to enhance the accuracy of reflecting the dynamic environment, thus facilitating timely online decision-making (Wu et al., 2023; Jadli et al., 2022; Nair & Gupta, 2021). Machine learning can potentially optimize the data analysis process in near real-time, facilitating rapid delivery of analytical outcomes (Lies, 2019). Additionally, the evaluation of the lead score is executed by employing machine learning algorithms that estimate the probabilities of conversion, which are arguably indicative of the current state of nurturing or level of interest (G. N. Kumar & Hariharanath, 2021; Jadli et al., 2022), and offer valuable feedback to marketers.

2.3 Supervised Machine Learning

Machine learning can capture dynamic patterns of online education leads, such as contact information, website activities, engagement with the company, purchase intent, and more, which improve decision-making and lead nurturing in online education. Machine learning is a subfield of artificial intelligence that uses logic to analyze complex data and extract information for predictions and informed decisions (Verma et al., 2021).

A shred of limited empirical evidence supports lead management scientific research. Supervised machine learning is recently the most frequently used research design (Ni et al., 2020). It seeks a general rule that matches input to output. The expected outputs are labeled by human expertise as the algorithm's ground truth (Choy et al., 2018). Supervised machine learning techniques have demonstrated considerable promise in predicting quality leads (Zumstein et al., 2021; Haleem et al., 2022; Wu et al., 2023). Therefore, eight supervised machine learning algorithms were chosen to build predictive-based lead conversion and lead scoring, primarily for their simplicity and widespread use in classification problems across various fields, including but not limited to fault detection, predictive maintenance, and loan scoring.

Logistic Regression (LR) exhibits several benefits regarding its implementation, computation, training, scaling, and regulation (Ray, 2019). LR is one of the top three famous algorithms for lead-scoring prediction (Wu et al., 2023). As per the research conducted by Kaur and Kaur (2020), LR yielded an accuracy rate exceeding 0.80 in predicting customer churn. Additionally, Itoo et al. (2021) showed that LR obtained an accuracy of 0.96 and a precision of 0.99 when predicting credit faults. LR has the ability to calculate probabilities in nature as performance metrics and cutoff values (Ray, 2019; Shah et al., 2020), suited for modelling scoring problems, assessing the impacts on lead scores, and prioritizing leads (Wu et al., 2023).

K-Nearest Neighbors (KNN) is data distribution independent and well at handling noisy instances or missing attribute values (Uddin et al., 2019). KNN works with non-linear data, outperforming in low-dimensional datasets, and is feasible for multi-class classification (Kravchenko et al., 2020; Wang et al., 2020). In previous research, KNN attained accuracies exceeding 0.85 in customer churn and loan default predictions (Kaur & Kaur, 2020; Hassonah et al., 2019; Hong et al., 2022). Moreover, the study of Jadli et al. (2022) revealed that KNN attained an accuracy rate of 0.80, a precision rate of 0.75, a recall score of 0.72, and an F1 score of 0.88 in hot lead prediction.

Support Vector Machines (SVM) have a robust theoretical foundation and perform well with high-dimension data (Pisner & Schnyer, 2020; McComb et al., 2021; Jain & Salau, 2019; Uddin et al., 2019). Therefore, SVM is less likely to overfit and trap in local optima, good generalize on new and unseen data in numerous real-world applications (Tomasevic et al., 2020; Cervantes et al., 2020; Pisner & Schnyer, 2020). It is also resistant to outliers due to its tolerance margin around the decision boundary and is only affected by support vectors (McComb et al., 2021). The study by Sen et al. (2020) reported that the overall accuracy achieved by SVM was 0.85. In another field setting, SVM based on Radial Basic Function attained an accuracy of 0.93 on credit risk prediction (Alabi et al., 2020). Additionally, previous research in various disciplines in the study of Maskeliunas et al. 2020, Awotunde et al. 2020, and Mebawondu 2020 have proven the overall effectiveness of SVM.

The simple underlying assumption of Naïve Bayes (NB) requires fewer training datasets to estimate the necessary classification parameters and is robust on highly correlated features (Wang et al., 2020; Uddin et al., 2019; Wickramasinghe & Kalutarage, 2020). Despite its easy implementation, NB often outperforms alternative classifiers and is less susceptible to overtraining in small sample sizes (Tomasevic et al., 2020; Wickramasinghe & Kalutarage, 2020). According to Sen et al. (2020), NB demonstrated superior performance in learning speed, classification speed, and handling missing values and achieved an accuracy score of 0.80. Furthermore, the studies of Itoo et al. (2021) demonstrated that NB attained an accuracy exceeding 0.85 and a precision surpassing 0.95 on credit fault prediction.

Ensemble methods are widely used to combine weak learners to obtain more robust outcomes. Random Forst (RF) is widely recognized for its efficacy in classification tasks (Tyralis et al., 2019). According to the studies of Kaur and Kaur (2020), RF recorded an overall accuracy of above 0.84 for customer churn prediction. The study by Ampountolas et al. (2021) revealed that the RF exhibited superior performance to the neural model and recorded an accuracy score of 0.81.

Several previous studies have employed Bagging to improve predictive accuracy. According to Singh & Sivasankar (2019), Bagging could achieve a prediction accuracy of over 0.80. Moreover, the study by Bilal et al. (2022) showed that, in previous research, Bagging in customer churn achieved a minimum accuracy score of 0.72, precision of 0.69, recall of 0.72, and F1-score of 0.62.

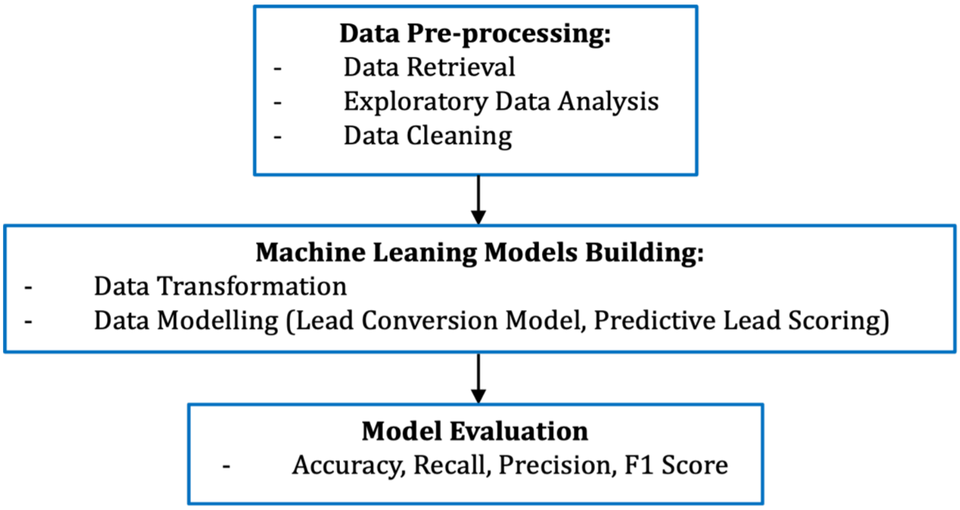
Besides, Bahad & Saxena (2020) conducted a study on predictive maintenance and found that AdaBoost trees recorded an accuracy score of 0.77 using five-fold cross-validation and recorded an accuracy of 0.73, precision of 0.71, recall of 0.45 and F1-score of 0.55 using a single train-test-split. The research of Bilal et al. (2022) and Singh & Sivasankar (2019) demonstrated the various empirical evidence of the efficacy of Boosting customer churn classification and credit default with an average accuracy of 0.80.

The study of Badawi et al. (2019) provided evidence that the stacking model is preferable to the Bagging, boosting, and individual models in terms of accuracy and dependability in various situations. Bokaba et al. (2022) conducted a study that demonstrated the superior performance of the stacking model compared to other models in terms of accuracy, precision, recall, and F1-score in predicting road traffic congestion.

3. Methodology

This section presents the nine stages of this work, as shown in **Figure 1**. These nine stages were displayed in three parts: Data Pre-processing, Machine Learning Models Building, and Model Evaluation.

Figure 1: Flow Chart of Overall Methodology



3.1 Data Pre-processing

In this part, a few processes were carried out to prepare the dataset for data modelling: data retrieval, exploratory data analysis, data cleaning, and data transformation.

3.1.1 Data Retrieval

The X Education Lead Scoring Dataset utilized for this study comprises information regarding the behavioural characteristics of X Education-generated leads. This company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. The dataset is provided by the company via the Kaggle platform. There are 9240 instances and 37 variables included. The CSV file is subsequently transformed into a two-dimensional labelled data frame and imported into Python using Pandas' read_csv(). Each row represents a unique, multifaceted behavioural observation. There are 34 features, excluding the "Prospect ID," "Lead Number," and one target variable.

Table 1: Data Dictionary of X Education Lead Scoring Dataset

Feature	**Description**
Prospect ID	A unique ID with which the customer is identified.
Lead Number	A lead number is assigned to each lead procured.
Lead Origin	The origin identifier with which the customer was identified to be a lead. Includes API, Landing Page Submission, etc.
Lead Source	The source of the lead. Includes Google, Organic Search, Olark Chat, etc.
Do Not Email	An indicator variable is selected by the customer wherein they select whether or not they want to be emailed about the course.
Do Not Call	The customer selects An indicator variable wherein they select whether or not they want to be called about the course.
Converted	The target variable. Indicates whether a lead has been successfully converted or not.
TotalVisits	The total number of visits made by the customer on the website.
Total Time Spent on Website	The total time spent by the customer on the website.
Page Views Per Visit	The average number of pages on the website viewed during the visits.

Last Activity	The last activity performed by the customer. Includes Email Opened, Olark Chat Conversation, etc.
Country	The country of the customer.
Specialization	The industry domain in which the customer worked before. Includes the level 'Select Specialization', meaning the customer had not selected this option while filling out the form.
How did you hear about X Education	The source from which the customer heard about X Education.
What is your current occupation	Indicates whether the customer is a student, unemployed, or employed.
What matters most to you in choosing this course	The customer selected an option indicating their main motto behind doing this course.
Search Magazine Newspaper Article X Education Forums Newspaper Digital Advertisement	Indicating whether the customer had seen the ad in any listed items.
Through Recommendations	Indicates whether the customer came in through recommendations.
Receive More Updates About Our Courses	Indicates whether the customer chose to receive more updates about the courses.

Tags	Tags are assigned to customers indicating the current status of the lead.
Lead Quality	Indicates the quality of the lead based on the data and intuition of the employee assigned to the lead.
Update me on Supply Chain Content	Indicates whether the customer wants updates on the Supply Chain Content.
Get updates on DM Content	Indicates whether the customer wants updates on the DM Content.
Lead Profile	A lead level is assigned to each customer based on their profile.
City	The city of the customer.
Asymmetric Activity Index	
Asymmetric Profile Index	Each customer’s index and score are assigned based on their activity and profile.
Asymmetric Activity Score	
Asymmetric Profile Score	
I agree to pay the amount through cheque	Indicates whether the customer has agreed to pay the amount through cheque.
A free copy of Mastering The Interview	Indicates whether the customer wants a free ‘Mastering the Interview’ copy.
Last Notable Activity	The last notable activity performed by the student.

3.1.2 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a method utilized to understand a dataset, frequently employing visualization that can expose intricate correlations, trends, or irregularities within data at varying levels of detail (U. Singh et al., 2020). The present study utilizes descriptive, univariate, and bivariate analyses and visualizations such as bar charts and box plots to do EDA. Descriptive analysis is undertaken to establish a robust foundation for subsequent analyses (Rahmany et al., 2020).

Descriptive statistics are numerical values that summarize data and describe data's central tendencies, distributions, and proportions within a sample. Pandas describe() function performed the descriptive analysis. The univariate technique involving only one variable uses Seaborn library's counterplot() method to create bar charts for categorical feature univariate analysis. While for the continuous variable, Seaborn's boxplot() is used to analyze the data distribution.

The bivariate analysis involves studying the relationship between features and target variables. In this study, bivariate analysis is performed on categorical features using the counterplot() function with the target variable set as the hue parameter to identify potential variables that can differentiate the target variables and uncover data inconsistencies or trends. The Seaborn library's boxplot() and deploy() functions are used for numerical features. The boxplot visualizes the distribution of observations in relation to the target variable, while the KDE plot represents the data using a continuous underlying probability density.

3.1.3 Data Cleaning

Data cleaning is necessary for ensuring the cleanness of the dataset. Unanswered responses are replaced as missing values. The missing values in each feature were detected using the isnull().sum() function. Features with a 70% or higher null percentage were removed to avoid introducing biases into the dataset. Features with 40% or more null percentages were also removed due to high variability. The remaining features with null values are applied imputation strategies such as mode. Features with 2% and lower null percentages directly removed the null values from the data frame. The outliers are detected using the boxplot() function and removed by excluding data below the 5th percentile and above the 95th percentile. Lastly, the features that provided no additional information for training the model are eliminated from the original data frame after using the drop() function after the EDA. After the data preprocessing, the cleaned data frame comprises 9074 instances and 14 variables, ready for the subsequent modelling phase.

3.2 Machine Learning Models Building

Before constructing the machine learning models, the cleaned dataset was transformed into a valid structure to be trained and tested by the machine learning algorithms.

3.2.1 Data Transformation

Data transformation prepares the dataset in a suitable structure for training the models. The trivial values presented as an issue for interpreting the analysis results, leading to overfitting when handled inadequately. Hence, these trivial values were grouped into the "Others" category to simplify the data and make it more manageable.

Since the chosen machine learning algorithms in this study only accept numerical input, categorical features must be encoded into a value of 0 and 1 with n-1 dummy variables using the Pandas get_dummies() function.

Next, data splitting divides a dataset into train and test subsamples for model training. Joseph (2022) argues that while the 80/20 ratio derives its rationale from the Pareto principle, it is fundamentally a rule of thumb. The researchers proposed dividing the data into ratio $\sqrt{p}$: 1 where p represents the number of parameters to estimate in a linear regression model. The current study has models more than linear regression model. Hence, the 80/20 training-to-test ratio is adopted and implemented using the Train_test_split() from sklearn model_selection library.

Feature scaling standardizes numerical input variables to normalize dataset magnitude discrepancies to minimize magnitude difference issues. Thus, data normalization eliminates large-scale feature dominance. StandardScaler() from sklearn pre-processing library standardizes all numerical features in this study by subtracting the value from the mean and dividing it by the standard deviation.

This study uses the feature elimination method to identify the optimal subset of features to maintain model simplicity and interpretability of variables. The selected feature subset is subsequently used to train and build the lead conversion and lead scoring model. This study eliminates features using LogisticRegression-based Recursive Feature Elimination (RFE). RFE uses the learned model and classification accuracy to select the best feature subset. It achieves this by iteratively eliminating the least significant feature that reduces accuracy (Jeon & Oh, 2020).

3.2.2 Data Modelling

Firstly, eight machine learning algorithms were imported from sklearn to build a lead conversion targeting model. The algorithms are shown below from **3.2.2.1** to **3.2.2.8**.

Next, LR is used to model predictive lead scoring. The predict_proba() function predicts the conversion probability or lead score. According to Zabor et al. (2021), the assumptions must be held to construct a robust LR-based lead scoring model:

1. Each observation is sampled randomly without influence from other observations.
2. The LR is assumed to specify correctly and accurately capture the true relationship between the features and the log odds of the outcomes.
3. No multicollinearity is presented in the dataset, which can lead to unreliable coefficient estimations and large standard errors. Subsequently, the variance inflation factor was evaluated by calling variance_inflation_factor()to confirm the multicollinearity assumption. In the realm of statistical analysis, a variance inflation factor (VIF) value that

exceeds 10 is commonly regarded as a strong indication of the presence of severe multicollinearity. Conversely, a VIF value below 5 is generally deemed acceptable (Craney & Surles, 2002).

4. No extreme values or outliers strongly influence the model presented in the dataset.

A sensitivity analysis is conducted across a range of probability thresholds (0.0 to 0.9) to determine the optimal probability threshold for lead conversion (Kelly et al., 2022; Sharm, 2009). The confusion matrix is computed for each threshold using the confusion_matrix() function, and the accuracy, sensitivity, and specificity values are calculated and depicted using the plot.line() function from Pandas. Sensitivity refers to the ratio of true positives that are accurately identified, while specificity pertains to the ratio of true negatives that are correctly identified, and accuracy is the proportion of instances in which the classifier is correct (Sidey-Gibbons & Sidey-Gibbons, 2019; Sharm, 2009).

3.2.2.1 Logistic Regression (LR)

LR refers to the generalized linear model (GLM) that applies the logit function as the canonical link function and transforms to a LR model when the response variable follows a Bernoulli distribution to address the classification task and estimate the odds ratio of an event's occurrence (Silva et al., 2020). The resultant of the dependent variable is a binary variable, taking on values of either 0 or 1, signifying two potential outcomes (Dastile et al., 2020; DeMaris & Selman, 2013; Kaur & Kaur, 2020). The logistic function or sigmoid function determines the probability of an occurrence ranging between 0 and 1. The link function is defined below.

$$P_i = \frac{1}{1+ e^{-(w^T x_i+b)}} \tag{1}$$

Where:

- ***Pi*** is a function of a vector of explanatory variables ($\boldsymbol{x_i}$)

- $\boldsymbol{w}$ is the associated weight
- $\boldsymbol{b}$ is the constant term
- $\boldsymbol{i}$ is the $\boldsymbol{n}$ sample observation

3.2.2.2 K-Nearest Neighbors (KNN)

KNN is a non-parametric classification and regression algorithm without data distribution assumptions. KNN is commonly called a lazy learner because it learns the training dataset during the testing phase (Wang et al., 2020; Kaur & Kaur, 2020). KNN makes a similarity measure by dividing data points into many classes, categorizing the sample data point, and making predictions based on neighbors' proximity. The *'k'* in the KNN represents the number of nearest neighbors participating in majority voting. One of the famous distance measure techniques is the Euclidean distance (Khan et al., 2018).

$$\boldsymbol{Euclidean} = \sqrt{(x_2 - x_1)^2 +} (y_2 - y_1)^2 \tag{2}$$

3.2.2.3 Support Vector Machine (SVM)

SVM solves classification and regression problems but is primarily employed for classification (Sen et al., 2020). SVM can handle linear and non-linear datasets. SVM maximizes the margin or vertical distance of extreme points or support vectors to generate a hyperplane (n – 1 sub-space) that segregates n-dimensional space (Uddin et al., 2019). SVM uses soft-margin or Kernel Tricks to handle non-linearly separable data. The concept of soft margin pertains to tolerating a certain degree of misclassified data. Besides, SVM uses Kernel Tricks to add another dimension by taking low-dimension input spaces and transforming them to higher dimension space, such as Kernel Radial Basic Function (RBF), Trick: Linear, Polynomial, and Sigmoid (Pisner & Schnyer, 2020; Tomasevic et al., 2020). The RBF formula is applied in this study and is defined below.

$$\boldsymbol{RBF}, f_k(x, y) = \exp\left(-\frac{||x-y||^2}{2\sigma^2}\right) \tag{3}$$

Where:

$\boldsymbol{\sigma}$ is the variance and hyperparameter.

$||\boldsymbol{x}-\boldsymbol{y}||$ is the distance between $\boldsymbol{x}$ and $\boldsymbol{y}$.

3.2.2.4 Naïve Bayes (NB)

NB is a classification algorithm that relies on the principle of conditional probability derived from the Bayes Theorem to determine the conditional probability of one event, given the likelihood of another event. It determines the probability of belonging to a particular class, such as the probability that a record or data collection belongs to a particular class (Tomasevic et al., 2020). NB is applicable for both the binary and multicast classification. As its name suggests, NB's fundamental premise is that the features are conditionally independent and have equal weights (Yang, 2018). The NB formula is defined below.

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} \tag{4}$$

3.2.2.5 Random Forest (RF)

RF is an ensemble learning method that makes predictions by combining multiple decision trees using binary partitioning of predictor variables (Tyralis et al., 2019; Mašetić & Subasi, 2016). RF using randomized subsets of training datasets and predictor variables to generate n-tree bootstrap samples. The trees within RF exhibit similarities to those found in Classification and Regression Trees (CART) (Breiman, 2001). The predictions are obtained by aggregating the results from individual trees and frequently yield superior precision compared to a solitary decision tree model while preserving certain advantageous characteristics of tree models (Speiser et al., 2019).

3.2.2.6 Bootstrap Aggregating (Bagging)

Bagging is an ensemble algorithm that reduces an estimator's variance compared to the weak learner alone (Lee et al., 2020; Latha & Jeeva, 2019). Bagging aggregates the predictions of several homogenous weak learners that trained independently and concurrently on different bootstrap samples of the training set (Bilal et al., 2022; Himeur et al., 2020; Lee et al., 2020; Singh & Sivasankar, 2019). The training set is divided into multiple subsets called bootstrap samples via a process of random resampling with replacement to bootstrap replicate the original training set with few repetitions and omissions (Jafarzadeh et al., 2021; Lee et al., 2020; Singh & Sivasankar, 2019; Latha & Jeeva, 2019). This study will use a Decision tree classifier as it has been commonly used for homogenous weak learners (Jafarzadeh et al., 2021; Singh & Sivasankar, 2019).

3.2.2.7 Boosting

Boosting is similar to Bagging but involves establishing a sequential series of homogenous base classifiers (Jafarzadeh et al., 2021). New subsets are constructed from problematic model components iteratively (Latha & Jeeva, 2019). Each model learns from prior model mistakes by considering each sample's weight to ensure the accurate classification of significant weighted samples. Boosting reduces the weight of correctly identified samples and increases otherwise (Sevinc, 2022; Singh & Sivasankar, 2019). The model parameters for each weak classifier are determined through the loss function minimization of the previous model. Finally, the Boosting algorithm generates a composite model that combines multiple base classifiers, each assigned a weight, resulting in a linear combination (Li & Chen, 2020).

3.2.2.8 Stacking

Stacking involves aggregating heterogeneous base learners in two phases (Wen & Hughes, 2020; Badawi et al., 2019; Džeroski & Ženko, 2004). Stacking constructs a series of base learners in the first layer. The resulting output from this layer is subsequently utilized to build a meta-level learner (Li & Chen, 2020; Džeroski & Ženko, 2004). Stacking aggregates the training set from heterogeneous base learners using a meta-learner, which differs from Bagging and Boosting (Badawi et al., 2019). Meta-learner is commonly constructed by applying a leave-one-out or cross-validation process and voting schema (Džeroski & Ženko, 2004). The present study involves training the meta-level learner through a weighted voting-based voting scheme.

3.3 Model Evaluation

1. **Accuracy:** Accuracy pertains to the percentage of correctly classified instances relative to the total number of predictions made.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (5)$$

2. **Recall:** Recall or sensitivity is the percentage of correct positive instances identified.

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

3. **Precision:** Precision is the percentage of instances predicted as positive is true positive.

$$Recall = \frac{TP}{TP+FP} \quad (7)$$

4. **F1 Score:** The F1 score is a metric that considers precision and recall simultaneously, yielding an average score that strikes a balance between the two.

$$F1\ score = \frac{precision \times recall}{precision+recall} \quad (8)$$

5. **K-Folds Cross Validation:** Cross-validation is a statistical method to test the trained model's generalization ability (Jadli et al., 2022). The dataset is partitioned into k equal-sized groups and resampled. Each iteration selects test data. The model is trained k times on k-1 folds and tested on the remaining fold. The utilization of K-Folds Cross-validation has resulted in a model with reduced bias as it facilitates the involvement of each data point in the model-building process. K-Folds Cross Validation is implemented using cross_val_score, and KFold from sklearn

model_selection library.

4. Results and Discussion

This section presents the results and analysis. Table 2 and Figures 2, 3, 4, 5, and 6 illustrate the exploratory data analysis results. Table 3 presents the precise lead profile-building results using RFE. After that, Tables 4 and 5 present performance comparative analysis results using the confusion matrix and evaluation metrics. Next, the LR-based lead scoring model hypothesis testing and modelling results are presented in Tables 6, and 7 and 8, respectively. Figure 7 depicts the sensitivity plots used to determine the optimal threshold for lead conversion probability.

4.1 Descriptive Statistics

The target variable consists of two categories: "Converted" and "Unconverted." As **Table 2** indicates, 62.14 percent of instances are unconverted. Hence, it is an imbalanced dataset.

Table 2: Descriptive Statistics of Target Variable

Converted	Number	Percentage (%)
Converted	3435	37.86
Unconverted	5639	62.14

4.2 Exploratory Data Analysis (EDA)

During EDA, bivariate visualization enhances comprehension of the dataset and prediction by revealing the hidden pattern between the features and target variable. The discovered patterns represent opportunities and the need for an online professional education provider to tailor their efforts to increase lead conversion rate. Hence, possible business strategies are proposed based on the visualization results. Lead nurturing strategies can be tailored to meaningful patterns. The strategies can focus on providing personalized content and relevant information that align with the insight into lead data, ultimately affecting the lead's subsequent behaviour and increasing the potential to be converted into a customer (Paschen et al., 2020). The online professional education provider can enhance lead conversion likelihood by prioritizing features or behaviours such as lead quality, lead origin, lead source, last notable activities, specialization, occupation, tags, and city-specific strategies.

4.2.1 Categorical Features

Figure 2 indicates that the quality of a lead significantly impacts the likelihood of conversion. The "worst" category had a 98% lead unconversion rate. However, other lead quality categories had higher lead conversion rates. Second, most leads originated from "API" and "Landing Page Submission", with a higher lead unconversion rate of 68.84% and 63.84%, respectively. Lead conversion techniques should prioritize to enhance "API" and "Landing Page Submission" lead sources. Third, "Google" and "Direct Traffic" generated the most leads but had a relatively high unconversion rate of 60.08% and 67.83%, respectively. Besides, lead generated from "Reference" and "Welingak Website" had a high conversion percentage.

Figure 2: Bar Charts of Lead Quality, Lead Origin, and Lead Source based on Converted Classes

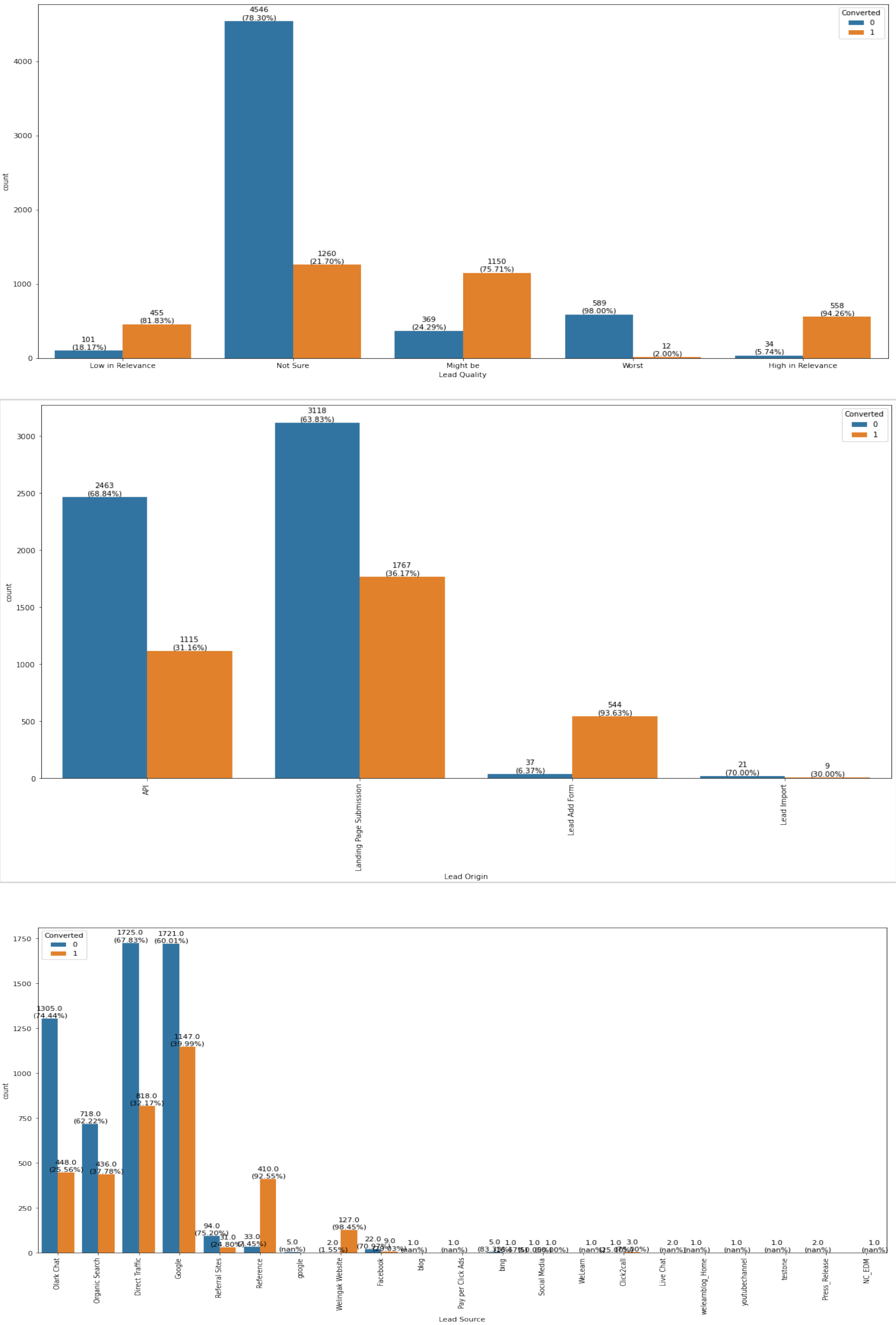
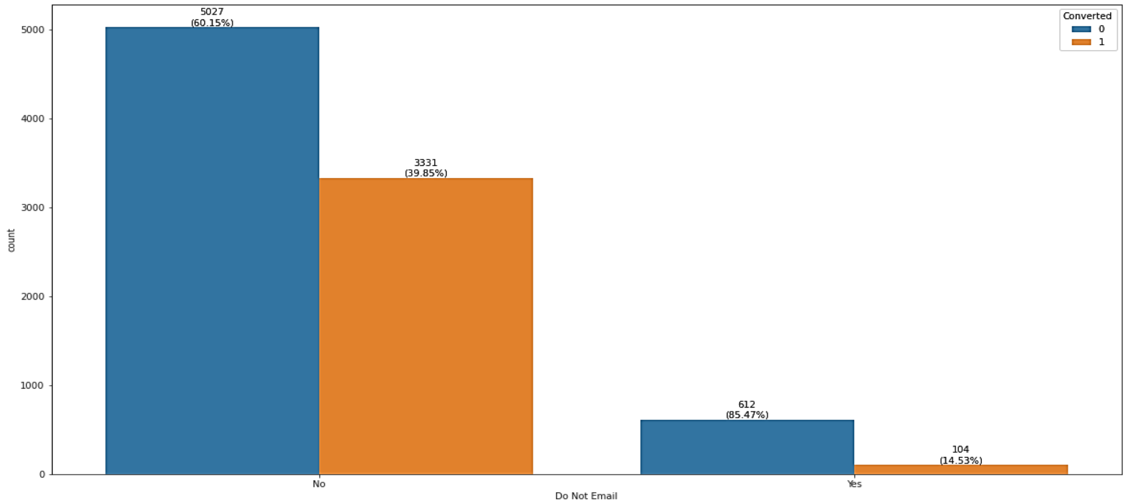
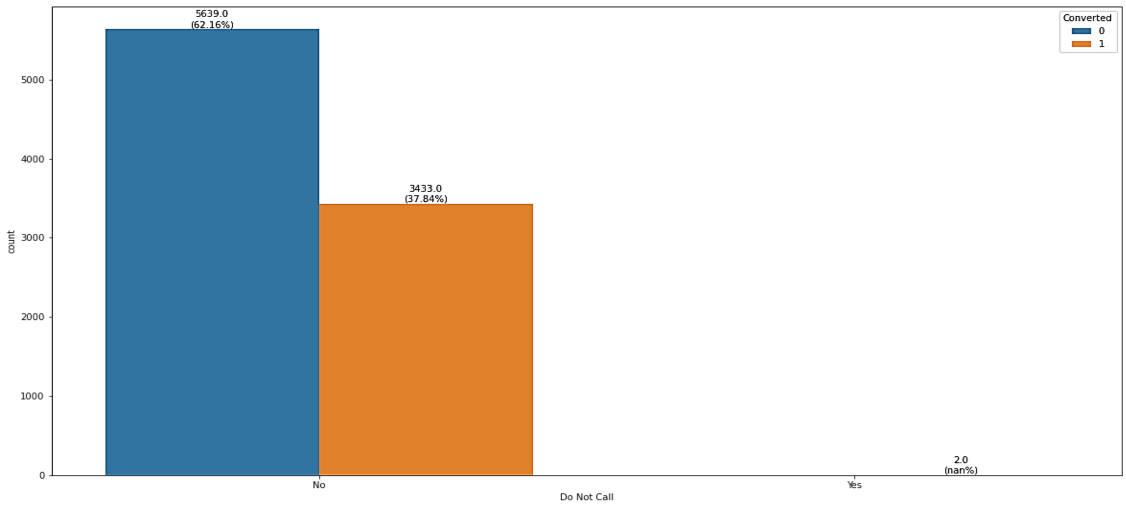
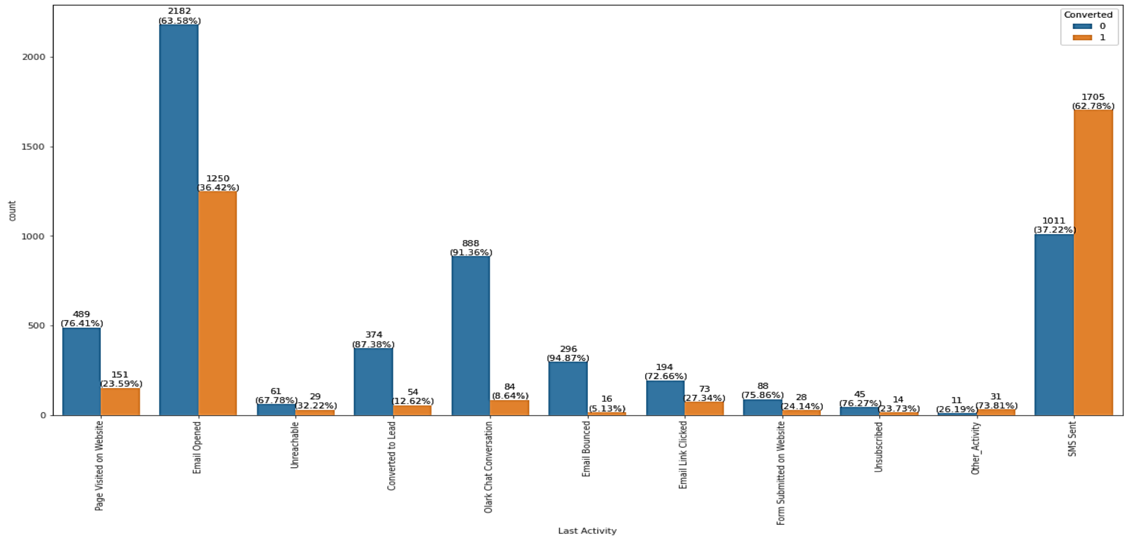


Figure 3 shows that opt-outs for emails and phone calls are prevalent but do not affect conversion due to the limited discriminative information. Next, “Olark Chat Conversation” and “Email Bounced” have poor conversion rates lower than 10 percent, while “SMS Sent” has the highest conversion rate of 63 percent. For the specializations such as “Business Administration”, “Media and Advertising”, “Supply Chain Management”, “Finance Management”, “Human Resource Management”, “Marketing Management”, “Banking, Investment and Insurance”, and “Operation Management” convert at a higher rate than the overall average of 41.36%. Targeting these specializations is believed to increase lead conversion.

Figure 3: Bar Charts of Do not Email and Do Not Call, Last Notable Activity, and Specialization based on Converted Classes







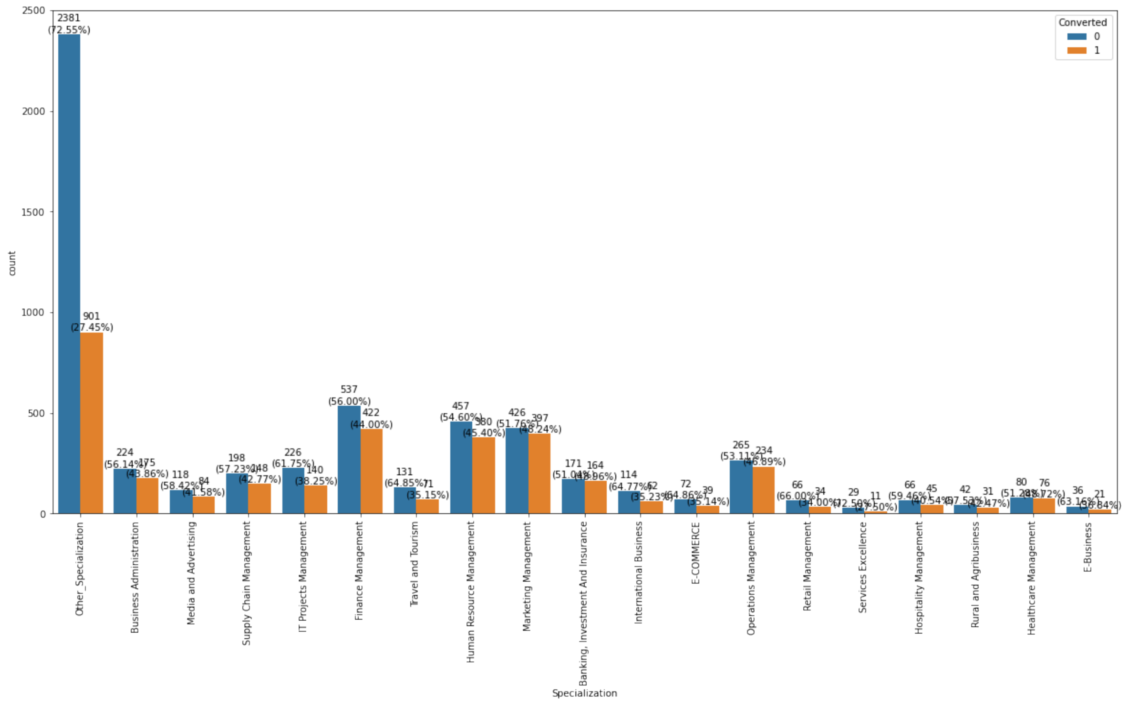
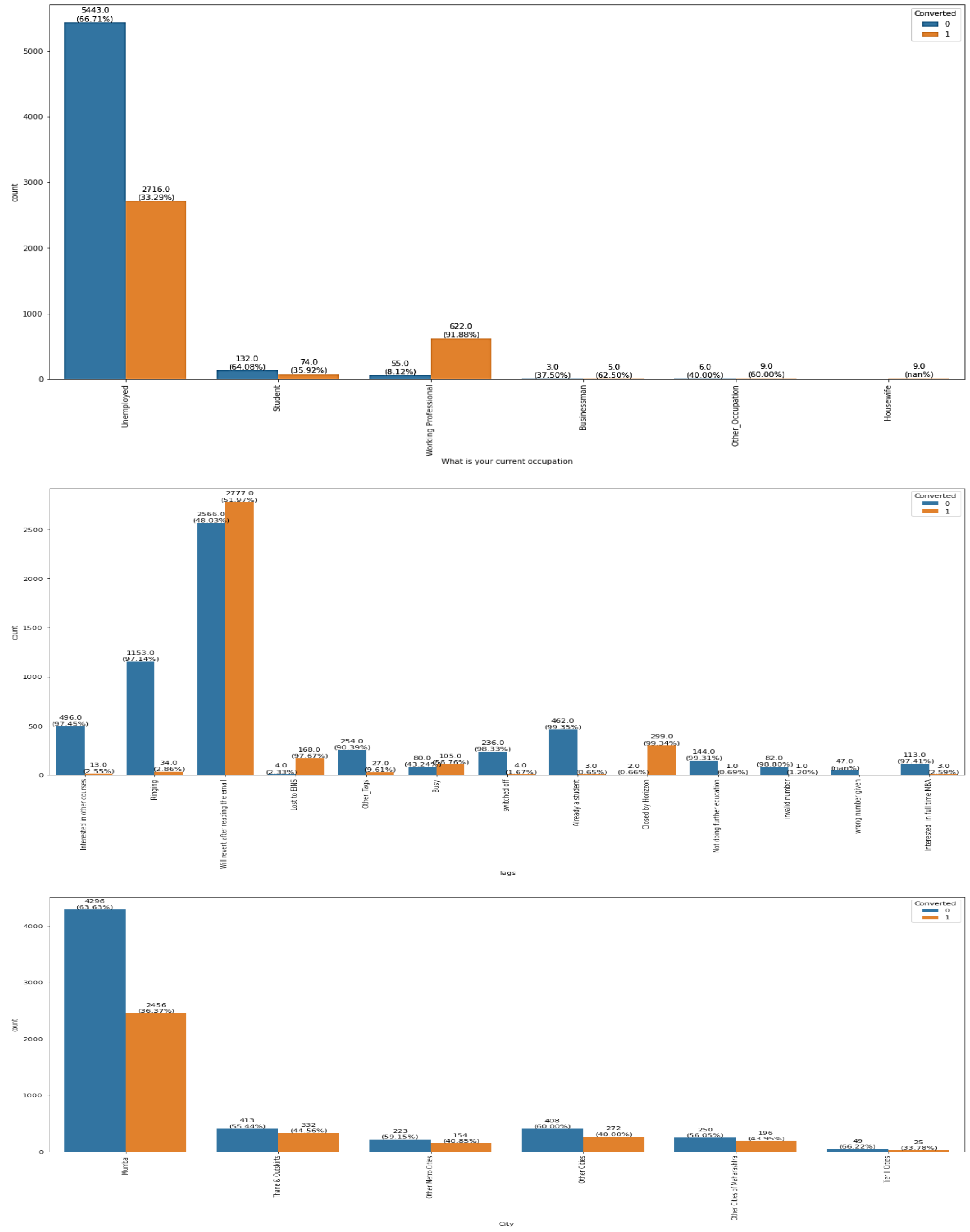


Figure 4 shows that working professionals have a conversion rate that exceeds 90%. Concentrating on this audience can increase the lead conversion rate. “Will revert after reading the email”-tagged leads have a 51.97% conversion rate, while “Closed by Horizon” has a 99.34% conversion rate. Leads containing these tags could be prioritized to improve conversion rates. Next, most leads are from Mumbai, where the conversion rate is 30 percent; tailoring efforts for Mumbai can increase lead conversion.

Figure 4: Bar Charts of Current Occupation, Tags, and City based on Converted Classes



4.2.2 Numerical Features

Figure 5 shows that the median of "TotalVisits" and "Page Views Per Visit" is the same for unconverted and converted leads, indicating no difference between the two classes. However, the "Total Time Spent on Website" boxplot revealed that converted leads had a higher median than unconverted leads, suggesting that leads spending more time on the website are more likely to be converted.

Figure 5: Box Plots of Numerical Features

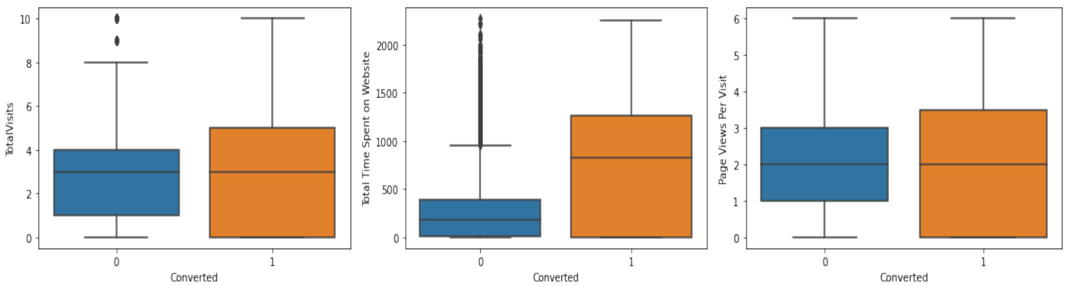
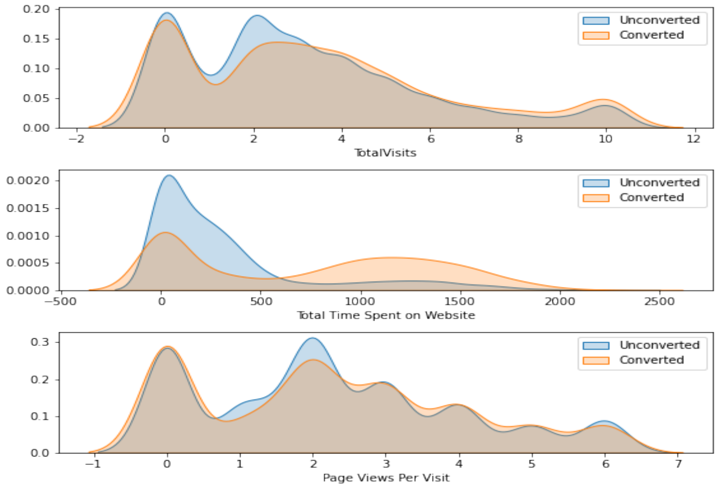


Figure 6 further indicates that "TotalVisits" and "Page Views Per Visit" exhibit similar probability density distributions for the two target classes. This suggests that the unconverted and converted leads exhibit comparable behaviours in relation to these features. However, for "Total Time Spent on Website", unconverted leads are most likely to spend lower time on the website.

Figure 6: Kernel Density Estimation Plots of Numerical Features



As Figure 5 and Figure 6 indicate, there appears to be a tendency for converted leads to spend more time on the website; conversely, unconverted leads display a heightened likelihood of spending less time on the website. The website interface is essential in increasing the lead conversion rate in online professional education. Indeed, Aslam et al. (2020) research indicated that website user interface significantly impacts user loyalty.

4.3 Recursive Feature Elimination (RFE)

Table 3 presents the 15 RFE-supported features. These features are ranked 1 compared to the lowest of 78, indicating that they are among the most significant predictors of the target variable. Consequently, this signifies a precise lead profile that effectively encompasses the attributes associated with the hot lead. These features enable the online professional education provider to effectively focus on the most promising leads whose profiles or behaviours align with such features. One could argue that in the context of the real world, these manifestations represent distinct patterns of behaviour or actions exhibited by those leads.

Table 3: Recursive Feature Elimination Supported Features

Feature	Ranking
Lead Origin_Lead Add Form	1
Lead Source_Welingak Website	1
Last Activity_Email Bounced	1
Tags_Busy	1
Tags_Closed by Horizon	1
Tags_Lost to EINS	1
Tags_Ringing	1
Tags_Will revert after reading the email	1
Tags_invalid number	1
Tags_switched off	1
Tags_wrong number given	1
Lead Quality_High in Relevance	1
Lead Quality_Not Sure	1
Lead Quality_Worst	1
Last Notable Activity_SMS Sent	1

4.4 Performance Result

The trained models are assessed by the confusion matrix, accuracy score, precision, recall, and F1-score to quantify and compare their performance.

4.4.1 Lead Conversion Targeting

The eight trained lead conversion targeting models are compared through a confusion matrix and K-Folds Cross Validation based on accuracy, precision, recall, and F1-score.

Table 4 shows the summary of the confusion matrix of eight models. The results show that SVM, RF, and Bagging best predict the positive class or converted lead, as these three models attained the highest true positive of 577 and the lowest false negative of 94. Arguably, SVM, RF, and Bagging can accurately predict the potential leads and minimize the chances of missing out on them. In contrast, Stacking best predicts the negative class or unconverted lead as Stacking predicted the highest true negative of 1096 and the lowest false positive of 48. Stacking is good for preventing resource misallocation on unpromising leads.

Table 4: Summary of Confusion Matrix of Lead Conversion Model

Model	Positive (Converted)		Negative (Unconverted)	
	True	False	True	False
LR	560	51	1093	111
KNN	567	52	1092	104
SVM	**577**	58	1086	**94**
NB	549	51	1093	122
RF	**577**	58	1086	**94**
Bagging	**577**	59	1085	**94**
Boosting	563	66	1078	108
Stacking	559	**48**	**1096**	112

Table 5 shows the 10-Folds Cross Validation results of eight trained models based on accuracy, recall, precision, and F1-score. The designed value for the number of folds is based on the rule of thumb. The results show that Stacking achieved the highest accuracy score of 0.9233, representing a higher percentage of correctly classified leads into the converted and unconverted classes of Stacking than other models. In contrast, Boosting had the lowest accuracy score of 0.9116.

Table 5: The 10-Folds Cross Validation Results of the Lead Conversion Model

Model	Accuracy Score	Recall	Precision	F1-score
LR	0.9207	0.8539	0.9316	0.8909
KNN	0.9197	0.8489	0.9341	0.8892
SVM	0.9215	0.8546	0.9335	0.8921
NB	0.9142	0.8346	0.9323	0.8807
RF	0.9222	**0.8568**	0.9315	0.8935
Bagging	0.9216	**0.8568**	0.9332	0.8925
Boosting	0.9116	0.8543	0.9083	0.8802
Stacking	**0.9233**	0.8531	**0.9391**	**0.8939**

Furthermore, the RF and Bagging model's recall performance of 0.8568 is better than others. The results suggest that RF and Bagging can reduce the number of false negatives. In contrast, Stacking achieved the highest precision score of 0.9391, indicating its superior performance in minimizing false positives. The results also indicate that the NB and Boosting classifier achieved the lowest recall of 0.8346 and the lowest precision of 0.9083, respectively, suggesting their inferior ability to predict true positives and true negatives, respectively. Lastly, Stacking's F1-score of 0.8939 is the highest, indicating its ability to accurately predict positive instances while minimizing the number of false negative predictions. In contrast, Boosting achieved the lowest F1-score of 0.8802, indicating that the model performs poorly in recall and precision compared to other models.

According to the performance evaluation, the present study has determined that Stacking is the golden model for online professional education lead conversion targeting as it is the best across various metric considerations and reduces misclassification during lead conversion prediction, reducing the costs of false positives and true negatives.

The results confirmed the findings of Badawi et al. (2019) and Bokaba et al. (2022) that Stacking outperformed the alternatives. It was observed that Stacking exhibited superior performance compared to the remaining seven models, especially in the present study applied the cross-validation approach that challenged the generalizability of algorithms. The performance of Stacking is owing to its nature in combining the base learners that possess different strengths and weaknesses to make more accurate predictions (Hosni et al., 2019). Each base learner might capture different data patterns, resulting in different errors. Indeed, Stacking was well-proven in generalization ability and reduced overfitting by previous research, leading to a better F1 score. The favorable performance of the Stacking facilitated the precise targeting of promising leads, which is aligned with the perspectives presented by Eitle and Buxmann (2019). Conversion targeting strives to reduce the inefficient deployment of resources.

4.4.2 LR-Based Lead Scoring

LR-based predictive lead scoring is intended to overcome the conventional lead scoring model shortage. The VIF was used to determine multicollinearity to develop a robust model. The optimal lead scoring threshold is determined via sensitivity analysis on three plots: accuracy, sensitivity, and specificity.

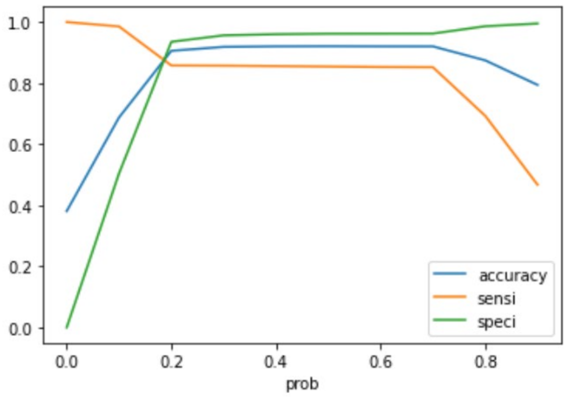
Table 6 shows that the selected features have VIF values below 10, indicating no multicollinearity issues, which means the selected features are statistically independent and can be added to the regression model without bias in the results. Thus, the regression coefficients and their interpretation are statistically reliable.

Table 6: Variance Inflation Factor of Selected Features

Feature	**VIF**
Tags_Closed by Horizzon	1.37
Lead Source_Welingak Website	1.34
Tags_Busy	1.11
Tags_switched off	1.11
Last Activity_Email Bounced	1.06
Tags_Lost to EINS	1.05
Tags_invalid number	1.04
Tags_wrong number given	1.02
Lead Origin_Lead Add Form	0.74
Lead Quality_High in Relevance	0.67
Lead Quality_Worst	0.48
Tags_Ringing	0.30
Last Notable Activity_SMS Sent	0.20
Tags_Will revert after reading the email	0.13
Lead Quality_Not Sure	0.02

False negatives and positives must be considered when determining lead conversion timing. False negatives occur when leads that fit conversion criteria are not identified, whereas false positives occur when leads that do not match conversion criteria are wrongly identified. Reducing false negatives helps identify all quality sales opportunities while eliminating false positives helps avoid allocating resources to low-quality leads. Hence, sensitivity analysis was used to determine the best lead conversion probability threshold that balances false negatives and false positives to optimize accuracy, sensitivity, and specificity.

Figure 7: Accuracy, Sensitivity, and Specificity Pots of Various Conversion Thresholds



According to Figure 7, the lead conversion probability threshold from 0.0 to 0.9 that yielded optimal results was 0.2. Initially, LR takes real-valued inputs and predicts the class membership belonging to the converted or unconverted classes. Predictions are made by setting a probability threshold 0.5 (Jain et al., 2020). The 0.2 threshold is the point at which accuracy and specificity reached their elbow point, while sensitivity was optimized at the expense of a trade-off (Kelly et al., 2022). Hence, the findings imply that the optimal timing for the sales team to convert the lead would be when the conversion probability reaches 0.2.

The lead scoring threshold represents the minimum level of interest that leads need to exhibit and the timing to be considered for lead conversion. Hence, the sensitivity analysis results offer significant insights into the process of lead conversion and can aid the sales team in identifying the most opportune moment to engage leads for conversion. Arguably, lead nurturing management involved both marketers and sales forces are tailored their strategies to nurture leads toward meeting the defined threshold. The lead nurturing process is passed to the sales forces, and a more tailored sales effort can be prepared to attract the audience and convert leads into paying customers (Paschen et al., 2020). Consequently, potential synergistic benefits can be achieved through the collaboration of different departments and the utilization of a LR-based lead-scoring model. It can be argued that implementing more holistic and coordinated approaches to lead conversion is allowed.

Table 7 presents the confusion matrix summary of the LR-based predictive lead scoring model after the probability threshold adjustment. The results indicate that the model accurately predicted 566 converted and 1071 unconverted leads. This information is valuable for the decision-making to focus on the promising leads and avoid resources allocated to leads that are unlikely to result in a successful conversion.

Table 7: Summary of Confusion Matrix of LR-Based Lead Scoring Model

Model	**Positive (Converted)**		**Negative (Unconverted)**	
	True	**False**	**True**	**False**
Predictive Lead Scoring- LR Based	566	73	1071	105

Table 8 presents the performance evaluation of the predictive lead scoring model. The adjusted lead scoring model achieved an accuracy score of 0.9019, indicating the model correctly classified approximately 90.19% of the leads as either converted or unconverted. The recall score of 0.9019 and precision score of 0.9015 highlighted the model's ability to predict the actual converted and actual unconverted lead, respectively. The F1-score of 0.9014 shows the model balanced the precision and recall score well.

Table 8: The Result of Accuracy, Recall, Precision, and F1-Score of LR-Based Lead Scoring Model

Model	Accuracy Score	Recall	Precision	F1-Score
Predictive Lead Scoring- LR Based	0.9019	0.9019	0.9015	0.9014

The performance of the adjusted lead scoring model justified the effectiveness of lead score modelling and using the adjusted probability threshold to predict converted and unconverted leads. The present research's results diverged from Eitle and Buxmann (2019), indicating that predicting the likelihood of leads in their initial stages can hardly be predicted and result in unsatisfied performance. Hence, the utilization of conversion probability as an indicator for optimal timing of lead contact by the sales department and as a means of feedback for marketers is being proposed. This approach aims to prevent missed opportunities and wrong timing engagement by ensuring engagement is made at the most opportune moment. Additionally, the conversion probability offers marketers significant insights into the effectiveness of their deployed strategies. The model extends beyond a basic classification prediction, enabling online education companies to continuously modify and enhance their strategies.

5. Implication

The present study has important theoretical and managerial implications. In terms of theory, the present study guides practitioners in predicting lead conversion and scoring by selecting appropriate models. The efficacy of the ensemble model in capturing the complex patterns of lead conversion prompts further exploration across various domains and problems. While for managerial implications, the lead conversion and lead scoring model help to minimize the costs and enhance the lead conversion rate by focusing on promising leads, lead prioritization, facilitating informed decision-making in the lead nurturing and engagement strategies, and strategies effort evaluation.

6. Conclusion

This study aims to improve lead conversion by using machine learning algorithms to enhance the lead management process on conversion targeting and lead scoring. Timely and accurate information facilitates better-informed decisions that can be achieved using machine learning and historical lead data to make predictions within an online professional education company. RFE determines the precise lead profile, such as the lead origin, lead source, last activity, tags, and lead quality. Following the model training, the models' performances were evaluated using reliable metrics such as accuracy, precision score, recall score, and F1 score. The study's findings indicate that Stacking demonstrates superior performance when evaluating all four-performance metrics in combination, using 10-Folds Cross Validation. Thus, an online professional education company could utilize Stacking for conversion targeting. Furthermore, the present study explored the potential of LR in modelling lead scoring to predict the lead conversion probability and subsequently determine the optimal timing for converting a lead and assessing strategies' effect on leads. The result indicates that the adjusted lead threshold can deliver favourable results and justify its potential. The effectiveness of lead management could be optimized via the combination of conversion targeting and predictive lead scoring, resulting in an increased conversion rate.

Last but not least, this study might overlook all significant variables affecting lead conversion and scoring. The dataset may be missing important features that can be used to train models. Next, it should be noted that the algorithms in this study were implemented using their default settings. Although extensively used and useful as a starting point, hyperparameter default settings are bound within a limited range of values and may not optimize performance across all datasets and issues. Furthermore, the lead dataset often suffers from the issue of an imbalanced class that is skewed toward unconverted leads. Most supervised machine learning algorithms may struggle to handle imbalanced classes effectively, leading to biased predictions.

Author Contributions Statement: The authors worked together for this paper. Conceptualization: W.Y.Y, K.W.K; Methodology: W.Y.Y, K.W.K, S.T.L, X.Y.C; Formal analysis and investigation: W.Y.Y, K.W.K, X.Y.C; Writing - original draft preparation: W.Y.Y, K.W.K; Writing - review and editing: K.W.K, S.T.L, X.Y.C; Resources: K.W.K, X.Y.C; Supervision: K.W.K, X.Y.C. All authors have read and agreed to the published version of the manuscript.

Funding Statement: No funding was received to assist with the preparation of this manuscript.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgement: Special thanks were extended to the School of Management, Universiti Sains Malaysia.

Conflict of Interest Statement: The authors have no competing interests to declare that are relevant to the content of this study.

References

Alabi, K. O., Abdulsalam, S. O., Ogundokun, R. O., & Arowolo, M. O. (2020). Credit Risk Prediction in Commercial Bank Using Chi-Square with SVM-RBF. In *Communications in computer and information science*. Springer Science+Business Media. https://doi.org/10.1007/978-3-030-69143-1_13

Alfian, G., Ijaz, M., Syafrudin, M., Syaekhoni, M. A., Fitriyani, N. L., & Rhee, J. (2019). Customer behaviour analysis using real-time data processing. *Asia Pacific Journal of Marketing and Logistics*, *31*(1), 265–290. https://doi.org/10.1108/apjml-03-2018-0088

Ampountolas, A., Nde, T. N., Date, P., & Constantinescu, C. (2021). A Machine Learning Approach for Micro-Credit Scoring. *Risks*, *9*(3), 50. https://doi.org/10.3390/risks9030050

Aslam, W., Hussain, A., Farhat, K., & Arif, I. (2020). Underlying Factors Influencing Consumers' Trust and Loyalty in E-commerce. *Business Perspectives and Research*, *8*(2), 186–204. https://doi.org/10.1177/2278533719887451

Awotunde, J. B., Jimoh, R. G., Oladipo, I. D., & Abdulraheem, M. (2020). Prediction of Malaria Fever Using Long-Short-Term Memory and Big Data. In *Communications in computer and information science* (pp. 41–53). Springer Science+Business Media. https://doi.org/10.1007/978-3-030-69143-1_4

Badawi, H. M., Azaïs, F., Bernard, S., Comte, M., Kerzérho, V., & Lefèvre, F. (2019). Use of ensemble methods for indirect test of RF circuits: can it bring benefits? In *2019 IEEE Latin American Test Symposium (LATS)*. https://doi.org/10.1109/latw.2019.8704641

Bahad, P., & Saxena, P. S. (2020). Study of AdaBoost and Gradient Boosting Algorithms for Predictive Analytics. In *Algorithms for intelligent systems* (pp. 235–244). Springer, Singapore. https://doi.org/10.1007/978-981-15-0633-8_22

Banerjee, S., & Bhardwaj, P. (2019). Aligning marketing and sales in multi-channel marketing: Compensation design for online lead generation and offline sales conversion. *Journal of Business Research*, *105*, 293–305. https://doi.org/10.1016/j.jbusres.2019.06.016

Bilal, S. M., Almazroi, A. A., Bashir, S., Khan, F. R., & Almazroi, A. A. (2022). An ensemble based approach using a combination of clustering and classification algorithms to enhance customer churn prediction in telecom industry. *PeerJ*, *8*, e854. https://doi.org/10.7717/peerj-cs.854

Blyzniuk, V., Yuryk, Y., Tokar, L. V., Serebrianska, I. M., Bezpalko, O., & Buryk, Z. (2021). Introduction of adult education as a modern educational and economic labour market trend. *Laplage Em Revista*, *7*(1), 304–313. https://doi.org/10.24115/s2446-6220202171726p.304-313

Bokaba, T., Doorsamy, W., & Paul, B. S. (2022). A Comparative Study of Ensemble Models for Predicting Road Traffic Congestion. *Applied Sciences*, *12*(3), 1337. https://doi.org/10.3390/app12031337

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/a:1010933404324

Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., & Lopez, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, *408*, 189–215. https://doi.org/10.1016/j.neucom.2019.10.118

Choy, G., Khalilzadeh, O., Michalski, M., Do, S., Samir, A. E., Pianykh, O. S., Geis, J. R., Pandharipande, P. V., Brink, J. A., & Dreyer, K. J. (2018). Current Applications and Future Impact of Machine Learning in Radiology. *Radiology*, *288*(2), 318–328. https://doi.org/10.1148/radiol.2018171820

Craney, T. A., & Surles, J. G. (2002). Model-Dependent Variance Inflation Factor Cutoff Values. *Quality Engineering*, *14*(3), 391–403. https://doi.org/10.1081/qen-120001878

Dastile, X., Celik, T., & Potsane, M. M. (2020). Statistical and machine learning models in credit scoring: A systematic literature survey. *Applied Soft Computing*, *91*, 106263. https://doi.org/10.1016/j.asoc.2020.106263

DeMaris, A., & Selman, S. H. (2013). Logistic Regression. In *Springer eBooks* (pp. 115–136). Springer Nature. https://doi.org/10.1007/978-1-4614-7792-1_7

Džeroski, S., & Ženko, B. (2004). Is Combining Classifiers with Stacking Better than Selecting the Best One? *Machine Learning*, *54*(3), 255–273. https://doi.org/10.1023/b:mach.0000015881.36452.6e

Eitle, V., & Buxmann, P. (2019). Business Analytics for Sales Pipeline Management in the Software Industry: A Machine Learning Perspective. In *Proceedings of the . . . Annual Hawaii International Conference on System Sciences*. https://doi.org/10.24251/hicss.2019.125

Espadinha-Cruz, P., Fernandes, A., & Grilo, A. (2021). Lead management optimization using data mining: A case in the telecommunications sector. *Computers & Industrial Engineering*, *154*, 107122. https://doi.org/10.1016/j.cie.2021.107122

Eurostat. (2022, January 24). Interest in online education grows in the EU. *Eurostat*. https://ec.europa.eu/eurostat/web/products-eurostat-news/-/edn-20220124-1

Gopalakrishna, S., Crecelius, A. T., & Patil, A. (2022). Hunting for new customers: Assessing the drivers of effective salesperson prospecting and conversion. *Journal of Business Research*, *149*, 916–926. https://doi.org/10.1016/j.jbusres.2022.05.008

Gouveia, B., & Costa, O. (2022). Industry 4.0: Predicting lead conversion opportunities with machine learning in small and medium sized enterprises. *Procedia Computer Science*, *204*, 54–64. https://doi.org/10.1016/j.procs.2022.08.007

Haleem, A., Javaid, M., Qadri, M. A., Singh, R. P., & Suman, R. (2022). Artificial intelligence (AI) applications for marketing: A literature-based study. *International Journal of Intelligent Networks*, *3*, 119–132. https://doi.org/10.1016/j.ijin.2022.08.005

Hassonah, M. A., Rodan, A., Al-Tamimi, A., & Alsakran, J. (2019). Churn Prediction: A Comparative Study Using KNN and Decision Trees. In *Information Technology Trends*. https://doi.org/10.1109/itt48889.2019.9075077

Himeur, Y., Alinier, G., Bensaali, F., & Amira, A. (2020). Robust event-based non-intrusive appliance recognition using multi-scale wavelet packet tree and ensemble bagging tree. *Applied Energy*, *267*, 114877. https://doi.org/10.1016/j.apenergy.2020.114877

Hong, Z., Deng, W., & Gong, X. (2022). Prediction of Car Loan Default Results Based on Multi Model Fusion. *Frontiers in Business, Economics and Management*, *5*(1), 142–149. https://doi.org/10.54097/fbem.v5i1.1515

Hosni, M., Abnane, I., Idri, A., De Gea, J. M. C., & Alemán, J. L. F. (2019). Reviewing ensemble classification methods in breast cancer. *Computer Methods and Programs in Biomedicine*, *177*, 89–112. https://doi.org/10.1016/j.cmpb.2019.05.019

Itoo, F., Meenakshi, & Singh, S. (2021). Comparison and analysis of logistic Regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection. *International Journal of Information Technology*, *13*(4), 1503–1511. https://doi.org/10.1007/s41870-020-00430-y

Jabbouri, J. (2023). The Application of Inbound Marketing to Improve Business Performance: Systematic Literature Review. *www.ijafame.org*. https://doi.org/10.5281/zenodo.7654781

Jadli, A., Hamim, M., Hain, M., & Hasbaoui, A. (2022). TOWARD A SMART LEAD SCORING SYSTEM USING MACHINE LEARNING. *Indian Journal of Computer Science and Engineering*, *13*(2), 433–443. https://doi.org/10.21817/indjcse/2022/v13i2/221302098

Jafarzadeh, H., Mahdianpari, M., Gill, E. W., Mohammadimanesh, F., & Homayouni, S. (2021). Bagging and Boosting Ensemble Classifiers for Classification of

Multispectral, Hyperspectral and PolSAR Data: A Comparative Evaluation. *Remote Sensing*, *13*(21), 4405. https://doi.org/10.3390/rs13214405

Jain, H., Khunteta, A., & Srivastava, S. (2020). Churn Prediction in Telecommunication using Logistic Regression and Logit Boost. *Procedia Computer Science*, *167*, 101–112. https://doi.org/10.1016/j.procs.2020.03.187

Jain, S., & Salau, A. O. (2019). An image feature selection approach for dimensionality reduction based on kNN and SVM for AkT proteins. *Cogent Engineering*, *6*(1). https://doi.org/10.1080/23311916.2019.1599537

Jeon, H., & Oh, S. (2020). Hybrid-Recursive Feature Elimination for Efficient Feature Selection. *Applied Sciences*, *10*(9), 3211. https://doi.org/10.3390/app10093211

Joseph, V. R. (2022). Optimal ratio for data splitting. *Statistical Analysis and Data Mining*, *15*(4), 531–538. https://doi.org/10.1002/sam.11583

Kaur, I., & Kaur, J. (2020). Customer Churn Analysis and Prediction in Banking Industry using Machine Learning. In *Grid Computing*. https://doi.org/10.1109/pdgc50313.2020.9315761

Kelly, J. L., Messaoud, R. B., Joyeux-Faure, M., Terrail, R., Tamisier, R., Martinot, J., Le-Dong, N., Morrell, M. J., & Pepin, J. (2022). Diagnosis of sleep apnoea using a mandibular monitor and machine learning analysis: One-Night agreement compared to in-Home polysomnography. *Frontiers in Neuroscience*, *16*. https://doi.org/10.3389/fnins.2022.726880

Khan, M. M., Arif, R. B., Siddique, M. a. B., & Oishe, M. R. (2018). Study and Observation of the Variation of Accuracies of KNN, SVM, LMNN, ENN Algorithms on Eleven Different Datasets from UCI Machine Learning Repository. In *arXiv (Cornell University)*. Cornell University. https://doi.org/10.1109/ceeict.2018.8628041

Kravchenko, Y., Dakhno, N., Leshchenko, O., & Tolstokorova, A. (2020). Machine Learning Algorithms for Predicting the Results of COVID-19 Coronavirus Infection. *CEUR Workshop Proceedings*, *2845*, 371–381.

Kumar, A., Kumar, P., Palvia, S. C. J., & Verma, S. K. (2017). Online education worldwide: Current status and emerging trends. *Journal of Information Technology Case and Application Research*, *19*(1), 3–9. https://doi.org/10.1080/15228053.2017.1294867

Kumar, G. N., & Hariharanath, K. (2021). Designing a Lead Score Model for Digital Marketing Firms in Education Vertical in India. *Indian Journal of Science and Technology*, *14*(16), 1302–1309. https://doi.org/10.17485/ijst/v14i16.290

Latha, C. B. C., & Jeeva, S. (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. *Informatics in Medicine Unlocked*, *16*, 100203. https://doi.org/10.1016/j.imu.2019.100203

Lee, T., Ullah, A., & Wang, R. (2020). Bootstrap Aggregating and Random Forest. In *Springer eBooks* (pp. 389–429). Springer Nature. https://doi.org/10.1007/978-3-030-31150-6_13

Li, Y., & Chen, W. (2020). A Comparative Performance Assessment of Ensemble Learning for Credit Scoring. *Mathematics*, *8*(10), 1756. https://doi.org/10.3390/math8101756

Lies, J. (2019). Marketing Intelligence and Big Data: Digital Marketing Techniques on their Way to Becoming Social Engineering Techniques in Marketing. *International Journal of Interactive Multimedia and Artificial Intelligence*, *5*(5), 134. https://doi.org/10.9781/ijimai.2019.05.002

Mašetić, Z., & Subasi, A. (2016). Congestive heart failure detection using random forest classifier. *Computer Methods and Programs in Biomedicine*, *130*, 54–64. https://doi.org/10.1016/j.cmpb.2016.03.020

Maskeliunas, R., Lauraitis, A., Damaševičius, R., & Misra, S. (2020). Multi-class Model MOV-OVR for Automatic Evaluation of Tremor Disorders in Huntington's Disease. In *Communications in computer and information science*. Springer Science+Business Media. https://doi.org/10.1007/978-3-030-69143-1_1

McComb, M., Bies, R., & Ramanathan, M. (2021). Machine learning in pharmacometrics: Opportunities and challenges. *British Journal of Clinical Pharmacology*, *88*(4), 1482–1499. https://doi.org/10.1111/bcp.14801

Mebawondu, J. O. (2020). Comparative Analyses of Machine Learning Paradigms for Operators' Voice Call Quality of Service. In *Communications in computer and information science* (pp. 66–79). Springer Science+Business Media. https://doi.org/10.1007/978-3-030-69143-1_6

Monat, J. P. (2011). Industrial sales lead conversion modeling. *Marketing Intelligence & Planning*, *29*(2), 178–194. https://doi.org/10.1108/02634501111117610

Nair, K. S., & Gupta, R. (2021). Application of AI technology in modern digital marketing environment. *World Journal of Entrepreneurship, Management and Sustainable Development*, *ahead-of-print*(ahead-of-print). https://doi.org/10.1108/wjemsd-08-2020-0099

Ni, D., Xiao, Z., & Lim, M. K. (2020). A systematic review of the research trends of machine learning in supply chain management. *International Journal of Machine Learning and Cybernetics*, *11*(7), 1463–1482. https://doi.org/10.1007/s13042-019-01050-0

Nygård, R., & Mezei, J. (2020). Automating Lead Scoring with Machine Learning: An Experimental Study. *Proceedings of the . . . Annual Hawaii International Conference on System Sciences*. https://doi.org/10.24251/hicss.2020.177

Ohiomah, A., Andreev, P. A., Benyoucef, M., & Hood, D. (2019). The role of lead management systems in inside sales performance. *Journal of Business Research*, *102*, 163–177. https://doi.org/10.1016/j.jbusres.2019.05.018

Paschen, J., Wilson, M. W., & Ferreira, J. J. (2020). Collaborative intelligence: How human and artificial intelligence create value along the B2B sales funnel. *Business Horizons*, *63*(3), 403–414. https://doi.org/10.1016/j.bushor.2020.01.003

Pisner, D. A., & Schnyer, D. M. (2020). Support vector machine. *Machine Learning*, 101–121. https://doi.org/10.1016/b978-0-12-815739-8.00006-7

Priya V, L. (2020). Implementing Lead Qualification Model Using ICP for Saas Products. *International Journal of Management*, *11*(10). DOI: 10.34218/IJM.11.10.2020.008

Rahmany, M., Zin, A. M., & Sundararajan, E. A. (2020). COMPARING TOOLS PROVIDED BY PYTHON AND R FOR EXPLORATORY DATA ANALYSIS. *DOAJ (DOAJ: Directory of Open Access Journals)*, *4*(3), 131. https://doi.org/10.56327/ijiscs.v4i3.933

Rahmat, A., Syakhrani, A. W., & Satria, E. (2021). Promising online learning and teaching in digital age: systematic review analysis. *International Research Journal of Engineering, IT and Scientific Research*, *7*(4), 126–135. https://doi.org/10.21744/irjeis.v7n4.1578

Ray, S. D. (2019). A Quick Review of Machine Learning Algorithms. In *International Conference Machine Learning, Big Data, Cloud and Parallel Computing*. https://doi.org/10.1109/comitcon.2019.8862451

Rosário, A. T., Moniz, L., & Cruz, R. M. (2021). Data Science Applied to Marketing: A Literature Review. *Journal of Information Science and Engineering*, *37*(5), 1067–1081. https://doi.org/10.6688/jise.202109_37(5).0006

Sen, P. C., Hajra, M., & Ghosh, M. (2020). Supervised Classification Algorithms in Machine Learning: A Survey and Review. In *Advances in intelligent systems and computing* (pp. 99–111). Springer Nature. https://doi.org/10.1007/978-981-13-7403-6_11

Sevinc, E. (2022). An empowered AdaBoost algorithm implementation: A COVID-19 dataset study. *Computers & Industrial Engineering*, *165*, 107912. https://doi.org/10.1016/j.cie.2021.107912

Shah, K., Patel, H. A., Sanghvi, D., & Shah, M. (2020). A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text

Classification. *Augmented Human Research*, *5*(1). https://doi.org/10.1007/s41133-020-00032-0

Sharm, D. (2009, December 1). *THE CONCEPT OF SENSITIVITY AND SPECIFICITY IN RELATION TO TWO TYPES OF ERRORS AND ITS APPLICATION IN MEDICAL RESEARCH.* https://journals.riverpublishers.com/index.php/JRSS/article/view/22065

Sidey-Gibbons, J. a. M., & Sidey-Gibbons, C. (2019). Machine learning in medicine: a practical introduction. *BMC Medical Research Methodology*, *19*(1). https://doi.org/10.1186/s12874-019-0681-4

Silva, E. C. E., Lopes, I., Correia, A., & Faria, S. (2020). A logistic regression model for consumer default risk. *Journal of Applied Statistics*, *47*(13–15), 2879–2894. https://doi.org/10.1080/02664763.2020.1759030

Singh, B. E. R., & Sivasankar, E. (2019). Enhancing Prediction Accuracy of Default of Credit Using Ensemble Techniques. In *Advances in intelligent systems and computing* (pp. 427–436). Springer Nature. https://doi.org/10.1007/978-981-13-1580-0_41

Singh, U., Hur, M., Dorman, K. S., & Wurtele, E. S. (2020). MetaOmGraph: a workbench for interactive exploratory data analysis of large expression datasets. *Nucleic Acids Research*, *48*(4), e23. https://doi.org/10.1093/nar/gkz1209

Speiser, J. L., Miller, M. I., Tooze, J. A., & Ip, E. H. (2019). A comparison of random forest variable selection methods for classification prediction modeling. *Expert Systems With Applications*, *134*, 93–101. https://doi.org/10.1016/j.eswa.2019.05.028

Terho, H., Salonen, A., & Yrjänen, M. (2022). Toward a contextualized understanding of inside sales: the role of sales development in effective lead funnel management. *Journal of Business & Industrial Marketing*, *38*(2), 337–352. https://doi.org/10.1108/jbim-12-2021-0596

Tomasevic, N., Gvozdenovic, N., & Vranes, S. (2020). An overview and comparison of supervised data mining techniques for student exam performance prediction. *Computers &Amp; Education*, *143*, 103676. https://doi.org/10.1016/j.compedu.2019.103676

Tyralis, H., Tyralis, H., & Langousis, A. (2019). A Brief Review of Random Forests for Water Scientists and Practitioners and Their Recent History in Water Resources. *Water*, *11*(5), 910. https://doi.org/10.3390/w11050910

Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. *BMC Medical Informatics and Decision Making*, *19*(1). https://doi.org/10.1186/s12911-019-1004-8

Verma, S. K., Sharma, R., Deb, S., & Maitra, D. (2021). Artificial intelligence in marketing: Systematic review and future research direction. *International Journal of Information Management Data Insights*, *1*(1), 100002. https://doi.org/10.1016/j.jjimei.2020.100002

Wang, Y., Zhang, Y., Lu, Y., & Yu, X. (2020). A Comparative Assessment of Credit Risk Model Based on Machine Learning ——a case study of bank loan data. *Procedia Computer Science*, *174*, 141–149. https://doi.org/10.1016/j.procs.2020.06.069

Wen, L., & Hughes, M. (2020). Coastal Wetland Mapping Using Ensemble Learning Algorithms: A Comparative Study of Bagging, Boosting and Stacking Techniques. *Remote Sensing*, *12*(10), 1683. https://doi.org/10.3390/rs12101683

Wickramasinghe, I., & Kalutarage, H. (2020). Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation. *Soft Computing*, *25*(3), 2277–2293. https://doi.org/10.1007/s00500-020-05297-6

Wu, M., Andreev, P., & Benyoucef, M. (2023). The state of lead scoring models and their impact on sales performance. *Information Technology & Management*. https://doi.org/10.1007/s10799-023-00388-w

Yang, F. J. (2018). An Implementation of Naive Bayes Classifier. *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*. https://doi.org/10.1109/csci46756.2018.00065

Zabor, E. C., Reddy, C. A., Tendulkar, R. D., & Patil, S. (2021). Logistic Regression in Clinical Studies. *International Journal of Radiation Oncology Biology Physics*, *112*(2), 271–277. https://doi.org/10.1016/j.ijrobp.2021.08.007

Zumstein, D., Oswald, C., Gasser, M., Lutz, R., & Schoepf, A. M. (2021). Lead Generation and Lead Qualification Through Data-Driven Marketing in B2B. *Marketing Automation Report 2021*. https://doi.org/10.21256/zhaw-2402