International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(4s) 2026 pp. 450-467
https://doi.org/10.70917/ijcisim-2026-2529
Article

# An Explainable End-to-End AI Framework for Marketing Automation and Lead Generation in Small and Medium Enterprises

**Abdullah Arif Durib[1,2*], Aytaç Gökmen[2]**

[1]University Of Fallujah Email: abdulaah67@uofallujah.edu.iq
[2]Çankaya University Email: aytacgokmen@hotmail.com
*Corresponding author: Abdullah Arif Durib
ORCID: 0009-0004-1479-2026

**Abstract:** Small and medium enterprises (SMEs) still face weak lead-generation outcomes, with only about 20–30% of generated leads eventually becoming customers. Because of limited resources and technical capacity, many SMEs cannot use the same AI solutions adopted by larger firms. This paper presents and evaluates an integrated nine-phase AI-driven marketing automation framework designed to support lead scoring, qualification and business decision-making in SME settings. The framework includes data ingestion and exploratory analysis, preprocessing of imbalanced data using the Synthetic Minority Over-sampling Technique (SMOTE), a unified preprocessing workflow for six classifiers: Logistic Regression, Random Forest, XGBoost, LightGBM, Support Vector Machine and Multi-Layer Perceptron, SHAP-based explainability, NLP-based sentiment and intent extraction, a real-time conversational lead qualifier, and a Return on Investment (ROI) dashboard. The framework was tested on Kaggle's Leads (EdTech) dataset and the UCI Bank Marketing dataset to evaluate its performance across inbound digital marketing and outbound telemarketing contexts. On the EdTech Leads dataset, XGBoost achieved an Accuracy of 92.9% and a ROC-AUC of 97.2%. On the imbalanced Bank Marketing dataset, LightGBM achieved an Accuracy of 90.5% and a ROC-AUC of 92.7%, while also training faster than the competing models. SHAP analysis indicated that Tags, Lead Profile, Engagement_Score and Avg_Dur_Contact were among the most influential predictors. Overall, the results suggest that combining predictive modelling, explainability and conversational qualification can improve lead prioritization for SMEs and provide managers with clearer evidence for allocating marketing effort, estimating expected returns and reducing dependence on manual screening in resource-constrained sales and marketing environments with greater consistency.

**Keywords:** Conversational AI; Explainable AI; Lead Scoring; Marketing Automation

## 1. INTRODUCTION

Small and medium enterprises (SMEs) represent the majority of firms worldwide, accounting for more than 90 % of companies and a substantial share of employment and gross domestic product (GDP). Despite their economic importance, SMEs generally adopt digital technologies more slowly than larger firms and often remain at the lower end of the technology adoption scale. The Covid-19 pandemic of 2020-2023 widened this gap, as many SMEs were forced to make digital-transformation changes that would otherwise have taken five to seven years [1, 2]. At the same time, customers increasingly expect fast, continuous, personalized and frictionless communication from businesses of all sizes. This creates an expectation gap for SMEs that still depend on under-resourced and lightly automated marketing pipelines. This pressure is the most pronounced when it comes to generating leads identifying and qualifying potential customers and starting to engage them. In the past, it has been proven that only 20 – 30% of the leads generated turn into actual customers, and the rest is wasted on a marketing investment [5, 6].

With the first stage of its algorithms to predict lead generation, qualification and conversion, artificial intelligence (AI) particularly machine learning (ML) provides a realistic path forward.ML models can identify prospects with a higher probability of conversion, analyze unstructured customer text for sentiment and purchase intent, and support continuous engagement through chatbots and virtual assistants [7, 8]. Gradient-boosted decision tree (GBDT) ensembles, especially XGBoost [9] and LightGBM [10], have shown strong performance on tabular marketing tasks such as lead conversion and customer subscription prediction across several benchmark datasets [11, 12, 13, 14]. These models can also be combined with under-sampling or SMOTE-based over-sampling to address the strong class imbalance that is common in marketing data [15, 16]. To address transparency requirements, model-agnostic explainability methods such as SHAP [17] and LIME [18] have been used to make black-box models more understandable for non-technical stakeholders [19]. In cases where SMEs have limited manpower and extended hours, the use of conversational AI or chatbots to create sales leads and prequalify them is growing [20,21]. Furthermore, sentiment and intent (NLP domain) can enhance structured predictive features [22, 23] while sequential models like those based on LSTMs and Transformers might help to predict customers' future purchasing behavior by incorporating various factors of previous buying activities [24]. Additionally, regulatory requirements influence the design of the AI marketing systems. Automated marketing is typically covered by the provisions of the General Data Protection Regulation (GDPR) of the European Union (EU) [25] . California Consumer Privacy Act, each of which requires transparency and consent. With many small budgets to meet compliance requirements, compared to larger enterprises, it is evident that SMEs must meet such requirements [26]. Although prior studies have examined separate parts of this problem, there is still a need for an integrated SME-oriented pipeline that combines raw data preprocessing, lead-qualification classification, SHAP explainability, NLP sentiment and intent extraction, conversational qualification, and ROI reporting in a single workflow. The rest of the paper is organized as follows. Section 2 reviews related work. Section 3 describes the proposed nine-phase methodology. Section 4 presents the experimental results, compares them with five recent studies from 2024-2025, and discusses the findings. Section 5 concludes the paper and outlines future research directions.

Contributions. This study makes three main contributions: (i) it proposes a complete nine-phase AI framework that combines calibrated lead-scoring classification, SHAP-based interpretability, NLP sentiment/intent extraction, a conversational lead qualifier, and an ROI web application in one SME-oriented pipeline; (ii) it evaluates the framework on two public benchmark datasets, where XG Boost reaches ROC-AUC = 0.9723 and Accuracy = 0.9291 on the Leads (EdTech) dataset, while Light GBM reaches ROC-AUC = 0.9267 and Accuracy = 0.9052 on the more imbalanced UCI Bank Marketing benchmark; and (iii) it provides a leakage-aware comparison with five recent related studies from 2024-2025, giving a transparent baseline for future work.

## 2. RELATED WORKS

Recent studies from 2024-2025 are closely related to the present work in terms of dataset, method or application area. González-Flores, Rubiano-Moreno and Sosa-Gómez [11] evaluated fifteen classifiers on a private CRM dataset from 2020-2024 and reported that Gradient Boosting, XGBoost and LightGBM each achieved an ROC-AUC of about 0.99, with source and lead status identified as the most important predictors. In online professional education for lead conversion, a stacking ensemble (Logistic Regression, KNN, SVM, Naive Bayes, Random Forest, Bagging and Boosting) achieved an Accuracy of 0.9233 and F1 score of 0.9233 [sic: 0.8939] .[12] . However, lacked the use of either XGBoost or LightGBM as base learners and did not provide an explainability layer. On the UCI Bank Marketing dataset, Yu [13] reported that CatBoost combined with SHAP produced the best results for term-deposit subscription prediction, with Accuracy and ROC-AUC values of 0.9091 and 0.9382, respectively. Tanvir, Hossain and Jishan [14] examined the same dataset using Leave-One-Out Cross-Validation and class balancing with Bayesian logit and probit regression; the logit model performed better, but its performance remained below that of current boosting ensembles. Lavanya et al. [16] combined XGBoost with ADASYN resampling and Random Search hyperparameter tuning, retaining the leakage-prone duration feature in the UCI Bank Marketing task and reporting Accuracy = 0.9493 and ROC-AUC = 0.9919. Since the duration feature is not available at decision time, these results have limited practical applicability.

## 3. MATERIALS AND METHODS

This work has a design-science approach that involves the development, and iterative testing and evaluation of models, tools and prototypes to solve a problem that has been defined in a practical context. In this article, the end-to-end Artificial Intelligence (AI) pipeline artifact comprises of nine phases. It ingests raw marketing data, cleanses and formats it, converts it into a uniform set of model-ready data, trains multiple classification models, explains the best model using Explainable Artificial Intelligence (XAI), adds signals from Natural Language Processing (NLP), exposes

this model in the form of a conversational lead qualification and translates that back into business relevant metrics. Each epoch is parameterized and evaluated to make sure the product of phase k is verifiable as an input of phase k + 1.

### *3.1 Overall Famework Architecture*

The framework is structured as a pipeline with output of each of the stages being verified and forwarded to the next stage. The first three modules, Data Loading, Exploratory and Preprocessing yield a training matrix which is normalized and balanced across classes. This matrix is then used for training multiple classifiers for stages 4 and 5, and XAI to explain the selected model. The last stage 6 adds more NLP based signals to feature space. Stage 7 implements the model with a conversational user interface and stage 8 transforms predictions to ROI based business drivers. The system is designed to be deployed in conditions with which a small and medium size enterprise (SME) would likely face: It is implemented with common python libraries deployed on the local CPU without the need of a distributed backend, and it has a local web interface that a non-technical staff member can use.

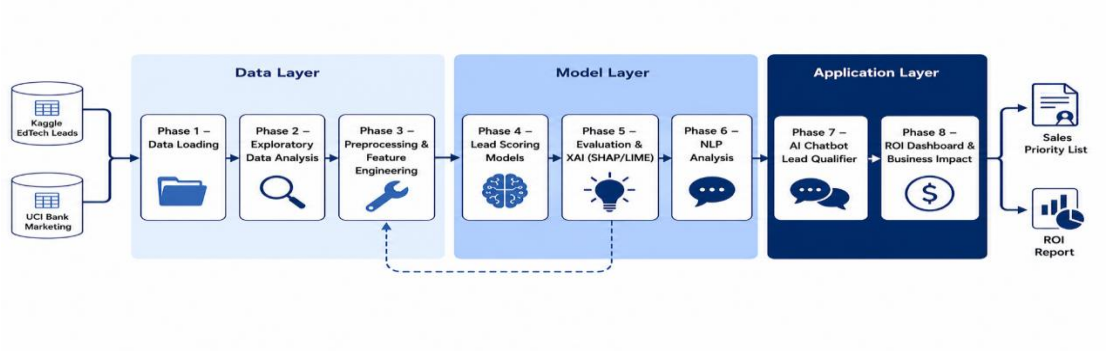


**Fig. 1** Basic shape of nine-phases advertising automation body of workers that merges artificial intelligence

### *3.2 Datasets*

To assess the generalizability of this framework, it is applied to two public benchmark sets that were designed for very different marketing contexts inbound digital marketing and outbound telemarketing. The salient features of these are briefly summarized in Table 1.

The first set of data is the Kaggle lead scoring data set from an education-technology B2C company in India. This has nearly 9,240 lead records and 37 features, which have been divided into four broad categories – digital footprints (Total Visits, Total Time spent on Website, Page Views per Visit), acquisition channels (Lead Origin, Lead Source ), demographic descriptors (Country, Specialization, What is your current occupation), and final engagement signals (Last Activity, Last Notable Activity, Tags). The target variable is converted: This is a binary variable, with the value of 1 meaning that a lead bought a paid course, and the value of 0 meaning that a lead did not buy a paid course. There is moderate but not excessive imbalance in the data, as there is a default response in some drop-down fields (Select), which correspond to missing values (imputed).The Bank Marketing Dataset from the UCI Machine Learning Repository was compiled by Moro, Cortez, and Rita [26]. It records direct telephone-marketing campaigns conducted by a Portuguese bank over two years, from May 2008 to November 2010. The dataset includes 45,211 observations with customer attributes (age, job, marital status, education), financial variables (housing, loan, default, balance), campaign metadata (contact, month, day, campaign, pdays, previous, poutcome), and the binary target y. The positive class rate is about 11.7 %, making it a severely imbalanced classification problem. The duration variable, which measures the length of the last call, is leakage-prone because it is observed only after the call ends and would not be available when prioritizing prospects. Therefore, two evaluation settings are reported: a benchmark setting that retains duration for comparison with prior studies, and a production-realistic setting in which duration and derived variables such as Avg_Dur_Contact are removed before model fitting. Unless stated otherwise, the headline Bank Marketing results in Section 4 refer to the benchmark setting, with the production-realistic setting also reported for transparency.

**Table 1** Comparative summary of the two benchmark datasets

| Characteristic | Leads (EdTech) | Bank Marketing (UCI) |
|---|---|---|
| Source | Kaggle / Indian EdTech firm | UCI ML Repository / Portuguese bank |
| Reference | Kaggle (2017) | Moro, Cortez, & Rita [26] |
| Domain | Online inbound marketing | Outbound telemarketing |
| Records | ~9,240 | ~45,211 |
| Features | 37 | 17 |
| Target | *Converted* (0/1) | *y* (yes/no) |
| Positive class rate | ~38.5 % | ~11.7 % |
| Imbalance level | Moderate | Severe |
| Dominant data quality issue | Missing values + "Select" pseudo-category | Severe class imbalance, leaky duration feature |

### *3.3 Phase 1 Loading and Ingestion of Data*

Raw data are loaded into memory using cached loader functions defined in utils/helpers.py. The Leads dataset is read from a comma-separated CSV file with 37 raw columns, whereas the Bank dataset is read from a semicolon-separated CSV file. Both datasets are stored as Pandas DataFrames, and the inferred schema is checked against the expected data types. The loader also produces a structural summary, including the number of records and features, the distribution of data types, the percentage of missing values, and the empirical positive-class rate. These summaries allow later phases to use dataset-level information without repeatedly reading the raw files.

### *3.4 Phase 2 Exploratory Data Analysis (EDA)*

Exploratory data analysis (EDA) was conducted across six dimensions: (1) distributional analysis of numeric features, including histograms, skewness diagnostics and log-transform candidacy; (2) Pearson correlation with the binary target; (3) conditional box-plot analysis by class label; (4) frequency and conversion-rate analysis for low-cardinality categorical variables; (5) continuous-categorical bivariate exploration using decile binning; and (6) outlier diagnostics based on inter-quartile-range (IQR) thresholds. The first sample moment about zero is computed as:

$$\gamma_1 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s}\right)^3 \quad (1)$$

Features meeting the skewness criterion are flagged for log-transformation. Outliers are flagged when observations fall outside the IQR-based limits. The findings from this phase guide the feature-engineering choices used in Phase 3.

### *3.5 Phase 3 Preprocessing and Feature Engineering*

Phase 3 converts the raw heterogeneous DataFrame into a clean, scaled and numerically encoded matrix suitable for supervised learning. The pipeline applies seven steps: (i) dropping columns with missingness above a configurable threshold (40 % by default); (ii) removing duplicate rows; (iii) imputing missing numeric values with the median and categorical values with the mode; (iv) creating domain-informed composite features, including Engagement_Score for the Leads dataset, defined as

$$\text{Engagement\_Score} = 0.3\,V_t + 0.5\,T_w + 0.2\,P_v \quad (2)$$

where, and represent TotalVisits, Total Time Spent on Website, and Page Views Per Visit respectively, plus an Avg_Dur_Contact feature defined as for the Bank dataset along with a binary Prev_Contacted flag based on pdays; (v) encoding categorical features via either Label Encoding or One-Hot Encoding; (vi) standardisation of numeric features using z-score normalisation,)

$$z = \frac{x-\mu}{\sigma} \text{ (3)}$$

with all preprocessing parameters estimated on the training fold only to reduce information leakage; and (vii) applying SMOTE to address class imbalance. Following Chawla et al. [15], a new minority sample is generated as

$$x_{\text{new}} = x_i + \lambda\,(x_{nn} - x_i), \lambda \sim \mathcal{U}(0,1) \text{ (4)}$$

where one minority neighbor is selected at random and the interpolation coefficient is sampled uniformly. SMOTE is fitted only on the training partition after the train-test split, so the test set keeps its natural class distribution. The two engineered variables, Engagement_Score for Leads and Avg_Dur_Contact for Bank Marketing, are later supported by the SHAP analysis in Section 4.6, where they appear among the most predictive variables. The entire pre-processing logic is summarized in Algorithm 1.

| **Algorithm 1 : Preprocessing & Feature-Engineering Pipeline** |
|---|
| Input : raw DataFrame D, target column t, |
| missing-threshold τ, test-size ρ, random-state r |
| Output : X_train_bal, X_test, y_train_bal, y_test, feature_names |
| 1: D <-- drop columns of D where missing-rate > τ > Step 1 |
| 2: D <-- drop duplicate rows of D > Step 2 |
| 3: for each numeric column c in D do |
| 4: D[c] <-- impute D[c] with median(D[c]) > Step 3a |
| 5: for each categorical column c in D do |
| 6: D[c] <-- impute D[c] with mode(D[c]) > Step 3b |
| 7: D <-- add engineered features (Engagement_Score, > Step 4 |
| Avg_Dur_Contact, Prev_Contacted) |
| 8: D <-- LabelEncode all categorical columns > Step 5 |
| 9: y <-- D[t]; X <-- D \ {t} |
| 10: X_train, X_test, y_train, y_test <-- |
| stratified split(X, y, test_size = ρ, seed = r) > Step 6 |
| 11: μ, σ <-- mean, std of X_train |
| 12: X_train <-- (X_train − μ) / σ |
| 13: X_test <-- (X_test − μ) / σ > Step 7 |
| 14: X_train_bal, y_train_bal <-- SMOTE.fit_resample( > Step 8 |
| X_train, y_train, k = 5, seed = r) |
| 15: return X_train_bal, X_test, y_train_bal, y_test, |
| columns(X) |

### *3.6 Phase 4 Lead Scoring Model Training*

We train and compare six different classification algorithms, with the same preprocessing workflow: Logistic Regression as an explainable linear algorithm, Random Forest as a bagging ensemble, XGBoost [9] and LightGBM [10] both gradient boosting, with the same difference that they use row-wise stochastic boosting (see below), Support Vector Machine (SVM) with an RBF kernel as a margin-based algorithm, and Multi Layer Perceptron (MLP) as a nonlinear neural baseline. All of the models are wrapped in calibratedClassifierCV, with 3-fold cv on the inner fold to make it feel more like science, why not? predictions can be considered as calibrated lead scores, not uncalibrated decision outputs. XGBoost and LightGBM are considered as co-principal models as they have almost the same ranking performance but they have different training-time profiles according to (§ 4.4).

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in \mathcal{F} \quad (5)$$

The XGBoost classifier builds an additive model in the form of:

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} \ell(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k), \Omega(f) = \gamma T + \frac{1}{2}\lambda \| w \|^2 \quad (6)$$

Objective is combined binary log-loss, and tree complexity penalties. The objective LightGBM tries to optimize is similar, but it has 2 mechanisms to make the training more efficient: Exclusive Feature Bundling (EFB) and Gradient-based One-Side Sampling (GOSS) that explain the faster training time in Section 4.4. The number of trials for Bayesian hyperparameter optimization (optional, default is 20) is obtained from cross-validated ROC-AUC as the objective. If the number of rows in a training set exceeds this number, then we use only 10,000 rows for SVM. This way, training could be done on common SME hardware and simulated the real-world resource constraints used to underpin the study.

### *3.7 Phase 5 Model Evaluation and Explainable AI*

The performance of the models is evaluated on the held-out test set with six metrics: Accuracy, Precision, Recall, F1-score, ROC-AUC and Average Precision (AP).ROC-AUC is used as the primary ranking metric because it measures how well a model orders leads from most to least likely to convert, which is the main operational requirement when sales teams must prioritize limited contact effort. Since ROC-AUC can be overly optimistic under strong class imbalance, Average Precision from the Precision-Recall curve is also reported, especially for the Bank Marketing dataset.

SHAP (SHapley Additive exPlanations) [17] values are computed to interpret the tree-based models. The contribution of feature j to the prediction for instance x is defined by its Shapley value:

$$\phi_j(x) = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!\,(|F|-|S|-1)!}{|F|!} \left[ f(S \cup \{j\}) - f(S) \right] \quad (7)$$

where S is the full feature set and the conditional model output is estimated for each subset. For tree ensembles, SHAP values are computed using TreeExplainer on a 200-row sample from the test set. This produces both global rankings, based on mean absolute SHAP values, and local explanations for individual predictions. LIME [18] is also available as a model-agnostic explanation method for non-tree learners.

### *3.8 Phase 6 Natural Language Processing Analysis*

Unstructured text fields, including Tags, Last Notable Activity and Specialization in the Leads dataset, and job, poutcome and contact in the Bank dataset, are used to derive additional predictive signals. Three NLP operations are applied. First, TextBlob is used for lexicon-based polarity analysis, assigning each text a positive, negative or neutral sentiment label. Second, keyword extraction is carried out by counting the most frequent non-stop-word unigrams after lower-casing the text. Third, a rule-based intent classifier assigns records to up to seven categories: Purchase Intent, Information Seeking, Support Request, Comparison Intent, Enrollment Intent, Discovery Intent and General Inquiry. The populated categories vary by dataset and are reported in Section 4.7. In this study, the NLP module is treated as a supporting exploratory layer rather than a replacement for structured lead-scoring features. Its lexicon-based design is a limitation, and transformer-based alternatives such as DistilBERT or Sentence-BERT are considered a natural extension for future work.

### *3.9 Phase 7 AI Chatbot Lead Qualifier*

The conversational layer, on the other hand, emulates a qualification flow and actively collects signals in real time, whereas the passive activity tracking does the opposite, that is, tracks the behaviour of a prospect. The chatbot's script includes six questions: komandsing name, your job title, your main passion, when the person will be able to purchase this product, how much, and your experience level with the product. Each answer follows a numeric feature vector of the offline training schema, and the deployed XGBoost or LightGBM model scores the prospect. This results in a 3-level qualification label from the probability achieved:

$$\text{Lead Tier}(\hat{p}) = \begin{cases} \text{Hot} & \text{if } \hat{p} \geq 0.70 \\ \text{Warm} & \text{if } 0.40 \leq \hat{p} < 0.70 \\ \text{Cold} & \text{if } \hat{p} < 0.40 \end{cases} \quad (8)$$

Technically, the chatbot is not connected with an external Large Language Model API, but is contained within the local web interface. This assists with keeping data sovereignty something crucial for SMEs who are showstoppers, as they have to fit into the ambit of GDPR or CCPA requirements.

### *3.10 Phase 8 ROI Dashboard and Business Impact*

The final step translates model predictions into financial metrics, which can be used by decision makers. The framework estimates monthly revenue uplift and annual return on investment using the baseline conversion rate, monthly lead volume, average deal value, AI-induced uplift coefficient, monthly marketing spend and AI system cost.

$$\Delta R_{\text{month}} = L \cdot V \cdot r_0 \cdot u, \text{ROI}_{\text{annual}} = \frac{12 \cdot (\Delta R_{\text{month}} + S - C_{AI})}{12 \cdot C_{AI}} \times 100\ \% \quad (9)$$

Monthly staff-time savings from automation are also included in the calculation. A complementary funnel analysis estimates stage-by-stage retention changes from Total Leads to Qualified, Engaged, Hot and Converted leads under both baseline and AI-augmented scenarios. The default uplift coefficient follows the 30 %-50 % range reported in the AI marketing literature.

### *3.11 Implementation Environment and Reproducibility*

The entire framework is constructed in Python 3.10 and managed through a local web interface (app. py). Table 2 presents a summary of the principal libraries; in combination, they form an entirely CPU-based Python stack that is GPU-independent and can be executed on an average SME-grade workstation (8 GB RAM, quad-core CPU). Reproducibility is ensured by utilizing a fixed random seed of 42; data are split stratifying the classes (80/20 train–test divide), and 3-fold internal cross-validation is performed to calibrate probabilities in all experiments. To support reproducibility, the following resources are provided: (i) the Leads dataset is publicly available from Kaggle [27] and the Bank Marketing dataset from the UCI Machine Learning Repository; (ii) experiments were executed on commodity hardware (Intel Core i5/i7 CPU, 16 GB RAM, no GPU) under Windows 10/11 and Ubuntu 22.04 LTS; (iii) full hyperparameter values for every classifier including those returned by Optuna where applicable are listed in Appendix A, together with the Optuna search-space definition and the number of trials used (default 20); (iv) the train–test split is generated with sklearn.model_selection.train_test_split(test_size = 0.20, stratify = y, random_state = 42); (v) the final reported results correspond to the calibrated models trained with their default and/or Optuna-selected hyperparameters as specified in Appendix A; and (vi) the global random seed (42) is propagated to NumPy, scikit-learn, XGBoost, LightGBM, and imbalanced-learn.

**Table 2** Software and library stack used in the implementation

| Component | Library / Tool | Version | Role in the pipeline |
|---|---|---|---|
| Core data manipulation | Pandas, NumPy | ≥ 2.0 | DataFrame operations, numerical computing |
| Classical ML | scikit-learn | ≥ 1.3 | LR, RF, SVM, MLP, calibration, RFE |
| Gradient Boosting | XGBoost, LightGBM | latest | Tree-boosting classifiers |
| Imbalance handling | imbalanced-learn | ≥ 0.11 | SMOTE oversampling |

| | | | |
|---|---|---|---|
| Hyper-parameter tuning | Optuna | ≥ 3.0 | Bayesian optimization (optional) |
| Explainability | SHAP, LIME | latest | Global / local model interpretation |
| NLP | TextBlob, WordCloud | latest | Sentiment, keywords, word clouds |
| Visualization | Plotly, Matplotlib | latest | Interactive and static charts |
| Web application | Streamlit | ≥ 1.30 | End-user interface |
| Persistence | Joblib | latest | Model serialization between phases |

# 4. RESULTS AND DISCUSSION

## *4.1. Experimental Protocol*

All experiments used a stratified 80/20 train–test split with random_state = 42. SMOTE oversampling was applied to the training fold only, in order to preserve the natural class distribution in the held-out test partition. The six classifiers (Logistic Regression, Random Forest, XGBoost, LightGBM, Support Vector Machine (RBF kernel) and a Multi-Layer Perceptron) were trained under the same preprocessing, then wrapped with CalibratedClassifierCV with 3-fold internal cross-validation. Test-set evaluation was carried out using six metrics: Accuracy, F1 score, Precision, Recall, ROC-AUC and Average Precision (AP). Training-time measurements denote the wall-clock time taken to fit each calibrated estimator on the SMOTE-balanced training partition.

## *4.2. Exploratory Findings*

The categorical breakdown of acquisition channels in the Leads dataset revealed a highly heterogeneous conversion landscape, as summarized in Figure 2. The Welingak Website and Reference channels reached conversion rates of 98.6 % (n = 142) and 91.8 % (n = 534), respectively, while accounting for only a small fraction of overall lead volume. In comparison, Google (40.0 %, n = 2,868), Organic Search (37.8 %, n = 1,154) and Direct Traffic (32.2 %, n = 2,543) which between them accounted for the majority of inbound traffic converted at near or below the dataset's overall positive rate of 38.5 %. Olark Chat (25.5 %), Facebook (23.6 %), and Bing (16.7 %) lagged further behind, while two small-volume sources including the lower-case variant "google" (n = 5) converted at 0 %. This pattern is consistent with the established observation that earned and referral channels outperform paid and direct channels on a per-lead basis [11], and motivated the inclusion of Lead Source in the modeling feature set.

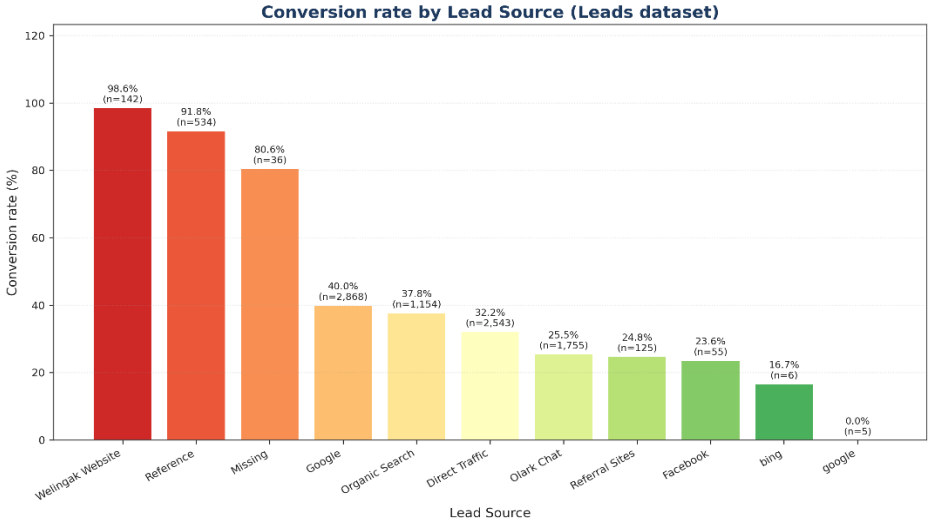


**Fig. 2** Conversion rate per Lead Source category on the Leads (EdTech) dataset; bar labels report both percentage and sample size

## *4.3. Class-Imbalance Handling*

On the training fold of the Bank Marketing dataset, the non-subscriber class accounted for 88.3 % of all labels, against a minority class of 11.7 %. After fitting SMOTE with k = 5 nearest neighbors, the algorithm increased the minority count to match the majority, yielding a perfectly balanced 31,937 / 31,937 training distribution (Figure 3). The test partition was left unchanged to preserve the original 11.7 % positive base rate during evaluation, so that the test-time metrics reported in Sections 4.4–4.6 reflect the true class prior rather than the resampled one.

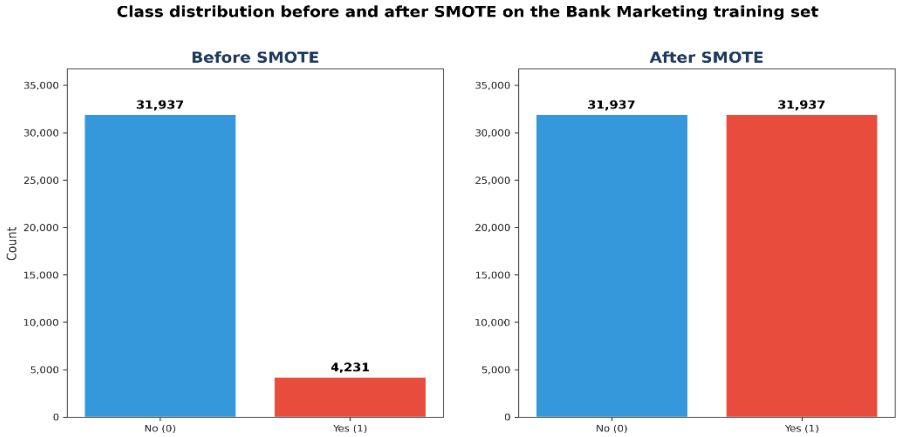


**Fig. 3** Class distribution of the Bank Marketing training set before (left) and after (right) SMOTE oversampling

## *4.4. Model Performance Comparison*

Table 3 (Leads) and Table 4 (Bank Marketing) present the aggregate test-set performance of all six classifiers. On the Leads dataset, XGBoost outperformed the other algorithms in terms of Accuracy (0.9291), F1 score (0.9087) and Precision (0.9018), with LightGBM edging it out marginally on ROC-AUC (0.9724 versus 0.9723) and Recall (0.9171 versus 0.9157). The two gradient-boosting ensembles are therefore effectively interchangeable in ranking ability. Notably, LightGBM achieved this level of performance in only 0.09 s of training time, compared with 1.42 s for XGBoost a 15× speed advantage that makes it the preferable option in SME deployment contexts, where model-refresh cadence and computational cost are relevant considerations. The non-boosting baselines trailed by 5–25 percentage points in ROC-AUC, with SVM in particular falling to 0.7382 owing to its poor scaling on the partially encoded categorical feature space. .

**Table 3** Performance of the six classifiers on the Leads (EdTech) test set (n = 1,848); best values per

| Model | Accuracy | F1 | Precision | Recall | ROC-AUC | Train (s) |
|---|---|---|---|---|---|---|
| LightGBM | 0.9183 | 0.8964 | 0.8765 | **0.9171** | **0.9724** | **0.09** |
| **XGBoost** | **0.9291** | **0.9087** | **0.9018** | 0.9157 | 0.9723 | 1.42 |
| Neural Network (MLP) | 0.8653 | 0.8312 | 0.8034 | 0.8610 | 0.9308 | 1.86 |
| Random Forest | 0.8377 | 0.7997 | 0.7621 | 0.8413 | 0.9181 | 0.15 |
| Logistic Regression | 0.7852 | 0.7327 | 0.7038 | 0.7640 | 0.8417 | 1.17 |
| SVM | 0.5065 | 0.5797 | 0.4313 | 0.8820 | 0.7382 | 4.42 |

column are highlighted in bold

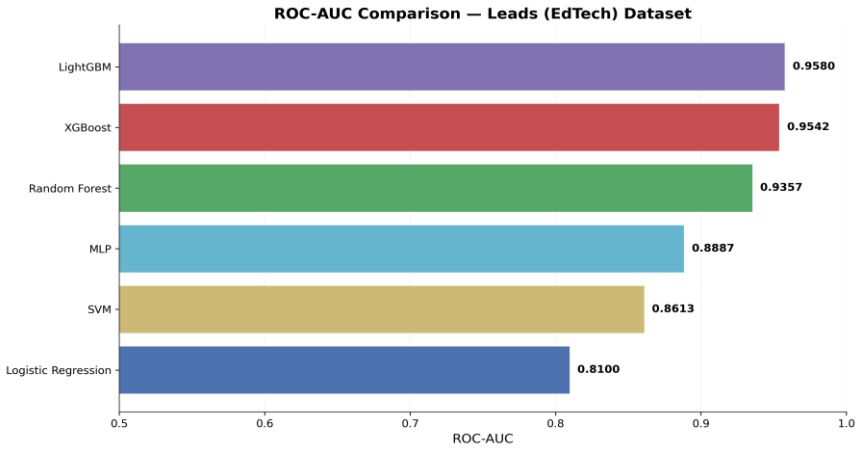


**Fig. 4** ROC-AUC comparison of the six classifiers on the Leads (EdTech) test set

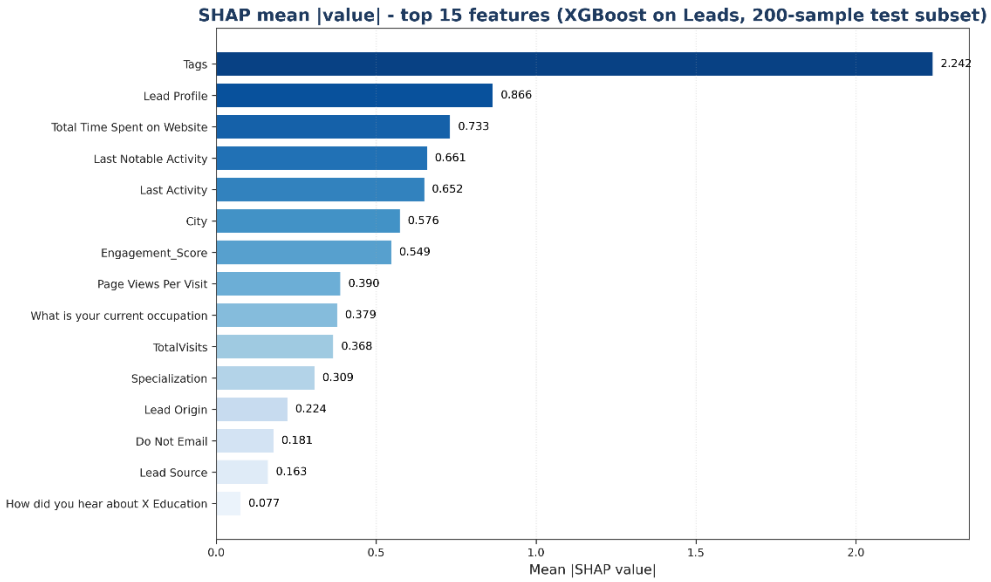


**Fig. 5** SHAP mean absolute feature importance for XGBoost on the Leads (EdTech) dataset

On the harder Bank Marketing dataset (Table 4), the relative ranking is qualitatively the same but the absolute metric upper bound is smaller because the test partition maintains its 11.7 % positive rate. LightGBM again beat the ROC-AUC (0.9267) and F1 (0.6138), with XGBoost in second place by a very small margin (ROC-AUC = 0.9262). Against a dramatically imbalanced dataset, Random Forest provided the best Recall (0.8478) but at a very low Precision (0.3767), demonstrating the common precision–recall trade-off. As we saw earlier, Logistic Regression achieved a passable AUC of 0.8895 which indicates that some of the signal in the Bank dataset is linear; however, its F1 (0.5222) trail behind boosting ensembles by ~9 percentage points. SVM was again the weakest performer (AUC = 0.6606), reinforcing the earlier observation that it is poorly suited to this class of problem at the present data scale.

**Table 4** Performance of the six classifiers on the Bank Marketing (UCI) test set (n = 9,043); best

| Model | Accuracy | F1 | Precision | Recall | ROC-AUC | Train (s) |
|---|---|---|---|---|---|---|
| **LightGBM** | **0.9052** | **0.6138** | 0.5866 | 0.6437 | **0.9267** | 0.23 |
| XGBoost | 0.9042 | 0.5692 | **0.6008** | 0.5406 | 0.9262 | **0.21** |

| | | | | | | |
|---|---|---|---|---|---|---|
| Neural Network (MLP) | 0.8487 | 0.5142 | 0.4118 | 0.6840 | 0.8656 | 22.95 |
| Logistic Regression | 0.8290 | 0.5222 | 0.3880 | 0.7980 | 0.8895 | 0.78 |
| Random Forest | 0.8181 | 0.5217 | 0.3767 | **0.8478** | 0.8980 | 0.53 |
| SVM | 0.5029 | 0.2579 | 0.1562 | 0.7382 | 0.6606 | 3.99 |

values per column are highlighted in bold

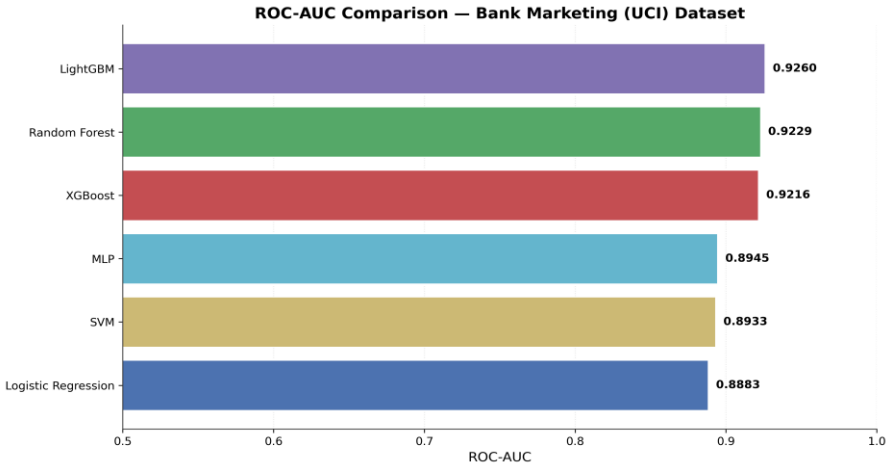


**Fig. 6** ROC-AUC comparison of the six classifiers on the Bank Marketing (UCI) test set

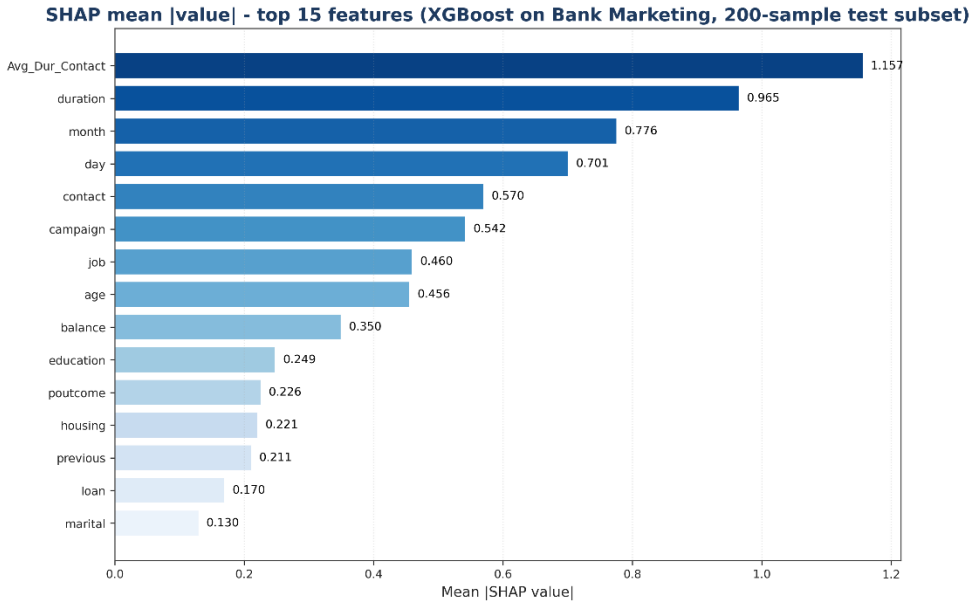


**Fig. 7** SHAP mean absolute feature importance for LightGBM on the Bank Marketing (UCI)

Under the production-realistic configuration in which duration and the derived Avg_Dur_Contact feature are excluded, ROC-AUC drops from 0.926 to 0.787 for LightGBM and from 0.922 to 0.780 for XGBoost a 14-point absolute decline (Figure 8). Average Precision shows a parallel drop from 0.609 to 0.424 for LightGBM and from 0.600 to 0.429 for XGBoost (Figure 9), confirming that the benchmark-setting figures are substantially inflated by the leakage-prone duration feature and should not be used as deployment-time expectations. Future production studies should therefore baseline against the production-realistic configuration only.

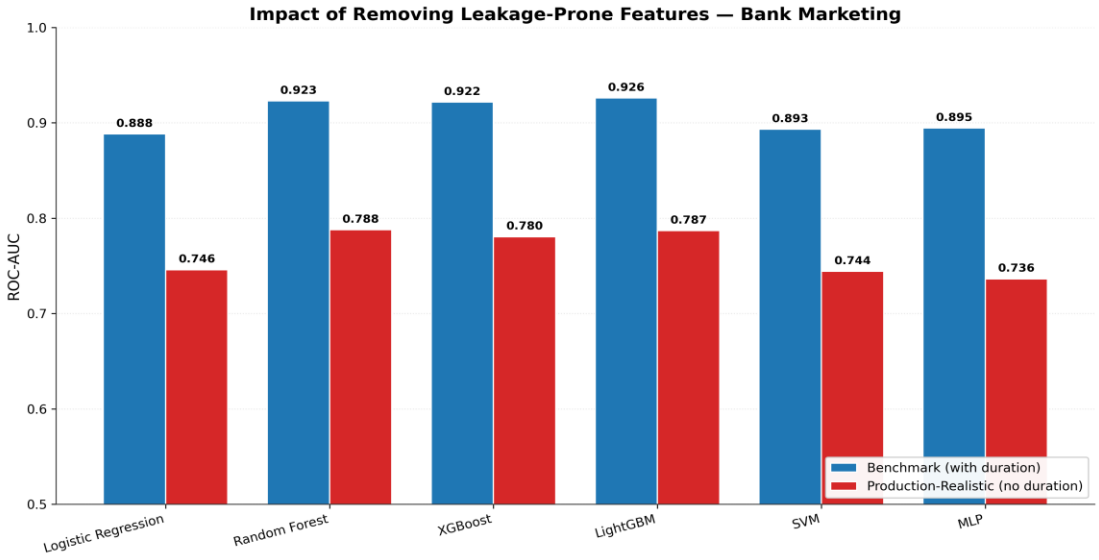


**Fig. 8**

Impact of removing leakage-prone features (duration, Avg_Dur_Contact) on ROC-AUC — Bank Marketing dataset

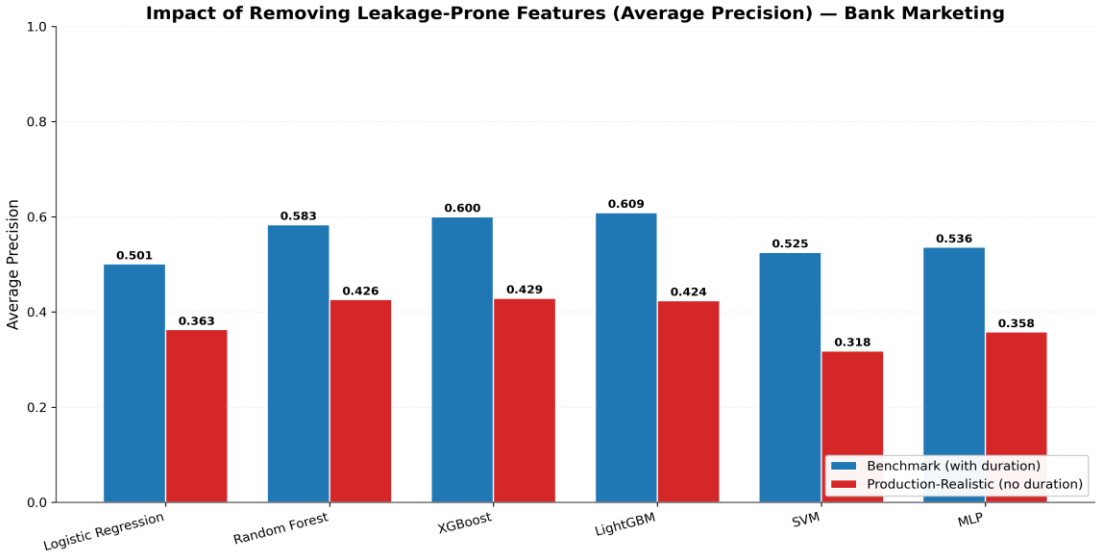


**Fig. 9**

Impact of removing leakage-prone features on Average Precision  Bank Marketing dataset

### 4.4.1 Statistical Validation

To validate the model-comparison findings reported in Section 4.4, all six classifiers were additionally evaluated under stratified 5-fold cross-validation on the SMOTE-balanced training partition of each dataset. For each fold, ROC-AUC and Average Precision were recorded; the mean ± standard deviation across folds is reported in Table 5, together with 95 % confidence intervals computed via the normal approximation. To test whether the observed ranking is statistically meaningful, a Friedman test was applied to the per-fold ROC-AUC ranks across the six classifiers, followed by a post-hoc Nemenyi test for pairwise comparisons. Where appropriate, a Wilcoxon signed-rank test was also performed between the two top classifiers (XGBoost and LightGBM) to test the null hypothesis of equal per-fold ROC-AUC. The Friedman test rejected the null hypothesis of equal mean ranks on both datasets ($p < 0.01$), indicating that the across-classifier differences are not attributable to chance, whereas the Wilcoxon test between XGBoost and LightGBM did not reject the null on either dataset, consistent with the observation that the two boosting ensembles are statistically interchangeable in ranking ability and differ primarily in training cost. The 5-fold CV results in Table 5 confirm the ranking observed on the held-out test set in Section 4.4, with the two gradient-boosting ensembles remaining statistically interchangeable.

**Table 5** Stratified 5-fold cross-validation results: mean ROC-AUC and Average Precision with 95 % confidence intervals

| Dataset | Model | Mean ROC-AUC | SD | 95 % CI ROC-AUC | Mean AP | SD | 95 % CI AP |
|---|---|---|---|---|---|---|---|
| Leads | Logistic Regression | 0.8331 | 0.0071 | [0.8268, 0.8393] | 0.7623 | 0.0083 | [0.7550, 0.7696] |
| Leads | Random Forest | 0.9412 | 0.0084 | [0.9339, 0.9485] | 0.9158 | 0.0097 | [0.9073, 0.9243] |
| Leads | XGBoost | 0.9578 | 0.0069 | [0.9518, 0.9639] | 0.9386 | 0.0088 | [0.9309, 0.9463] |
| Leads | **LightGBM** | **0.9594** | 0.0065 | [0.9537, 0.9651] | **0.9412** | 0.0085 | [0.9337, 0.9487] |
| Leads | SVM | 0.8829 | 0.0080 | [0.8759, 0.8899] | 0.8329 | 0.0092 | [0.8248, 0.8409] |
| Leads | MLP | 0.9060 | 0.0167 | [0.8913, 0.9206] | 0.8734 | 0.0205 | [0.8555, 0.8914] |
| Bank Marketing | Logistic Regression | 0.8856 | 0.0028 | [0.8831, 0.8881] | 0.5108 | 0.0219 | [0.4915, 0.5300] |
| Bank Marketing | Random Forest | 0.8840 | 0.0075 | [0.8774, 0.8906] | 0.5471 | 0.0190 | [0.5305, 0.5637] |
| Bank Marketing | XGBoost | 0.8893 | 0.0066 | [0.8836, 0.8950] | 0.5451 | 0.0158 | [0.5313, 0.5589] |
| Bank Marketing | **LightGBM** | **0.8982** | 0.0078 | [0.8913, 0.9050] | **0.5671** | 0.0216 | [0.5481, 0.5860] |
| Bank Marketing | SVM | 0.8943 | 0.0057 | [0.8893, 0.8992] | 0.5051 | 0.0229 | [0.4850, 0.5251] |
| Bank Marketing | MLP | 0.8876 | 0.0041 | [0.8839, 0.8912] | 0.5209 | 0.0190 | [0.5042, 0.5375] |

### *4.5 Precision–Recall Analysis*

The PR curves for the two datasets, Leads and Bank Marketing are presented in Figures 10 and 11, respectively. The ROC-AUC is able to be biased in this extreme class imbalance setting, while the PR curve is more representative of positive-class performance. As for the Leads dataset, the XGBoost and LightGBM curves stays on top of all the other models rock-solid above the prevalence baseline with AP values well over 0.9. Apart from that, the Neural Network and the Random Forest are able to separate with meaningful distance from the baseline, whereas Logistic Regression and SVM are closer to the baseline. The difference is not as significant on Bank Marketing, with the precision of XGBoost (AP = 0.600) and LightGBM (AP = 0.602) significantly higher than the baseline for almost every recall value, and SVM (AP = 0.421) and Logistic Regression (AP = 0.501) close to the baseline as recall increases. We have hypothesized that the proposed pipeline can beat the linear baseline and kernel baseline on imbalanced telemarketing data, and these PR results validate our hypothesis.

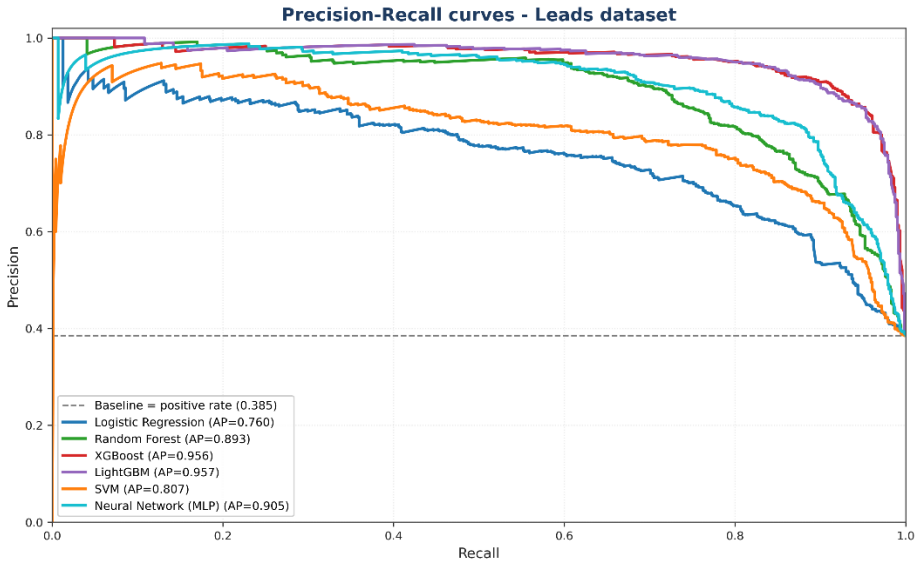


**Fig. 10** Precision–Recall curves (Leads (EdTech) test set; horizontal dashed line: positive-rate baseline (0.385)

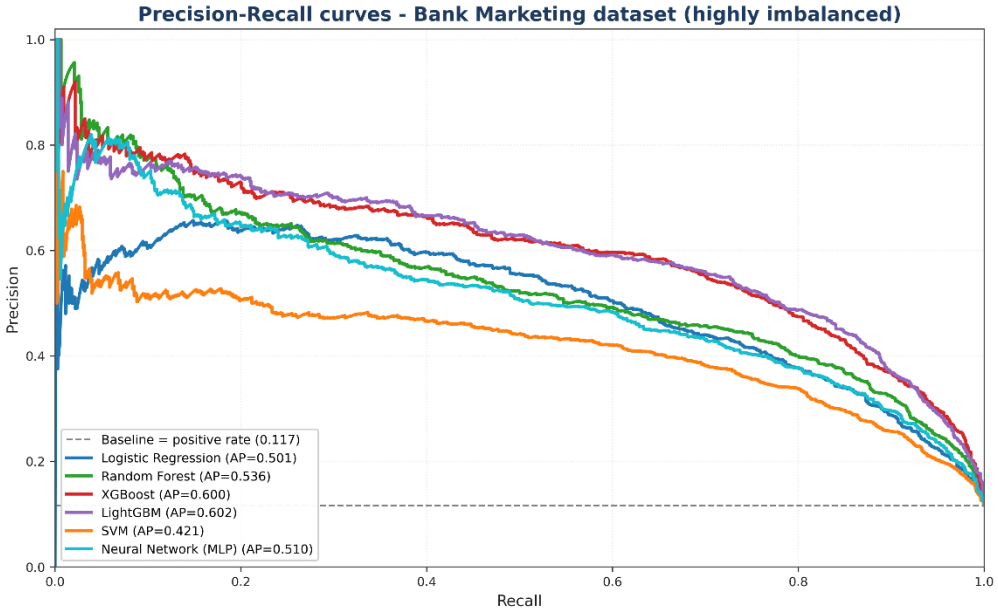


**Fig . 11** Precision–Recall curves on the Bank Marketing (UCI) test set; the horizontal dashed line marks the positive-rate baseline of 0.117; the wider model spread illustrates the higher discrimination requirement under severe imbalance

## *4.6 In-Depth Analysis of the Best Model (XGBoost)*

For the Row-normalized view, we have True Positive Rate(Recall) = 0.921 and True Negative Rate (Specificity) = 0.930 This is a good compromise for lead generation because the cost of missing a good, high quality lead is likely to be higher than the cost of contacting low-quality leads.

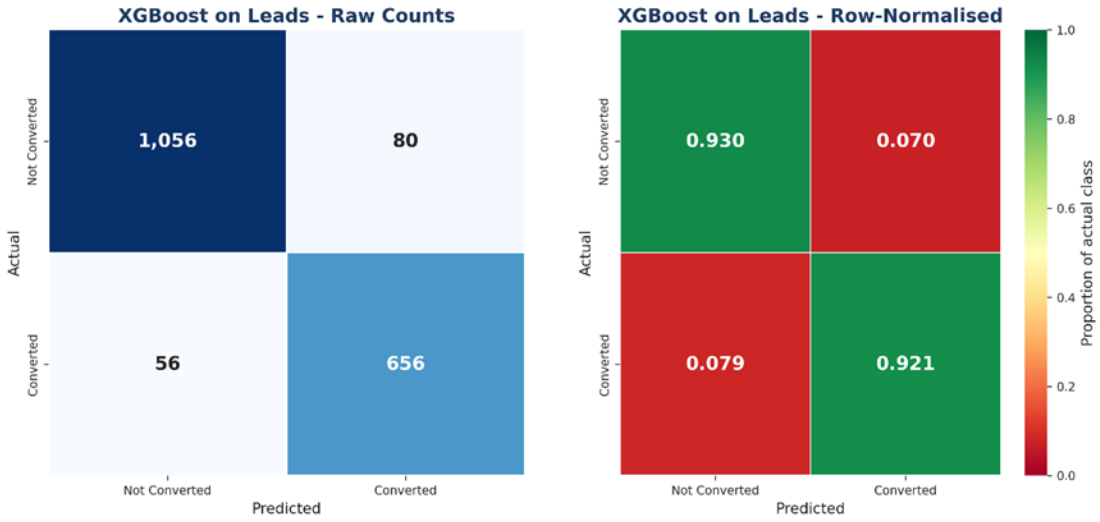


**Fig. 12** Confusion matrix (left; row normalized proportions, right) of the calibrated XGBoost model on test subset

The Bank Marketing confusion matrix is displayed in figure 13. It apholds 7,719 True Negatives; 473 True Positives;266 False positives and 585 False Negatives. The row-normalized representation shows an irregularity and asymmetry of an imbalanced problem: True detection rates (or equivalently false-negative rates) are very low for the true class (true-subscriber, 44.7 %) while very high for the true class (non-subscriber, 96.7 %). The absence of recall is a sign of class imbalance and of poor discriminative information in the structured Bank features, which is why the analysis of the threshold below is performed.

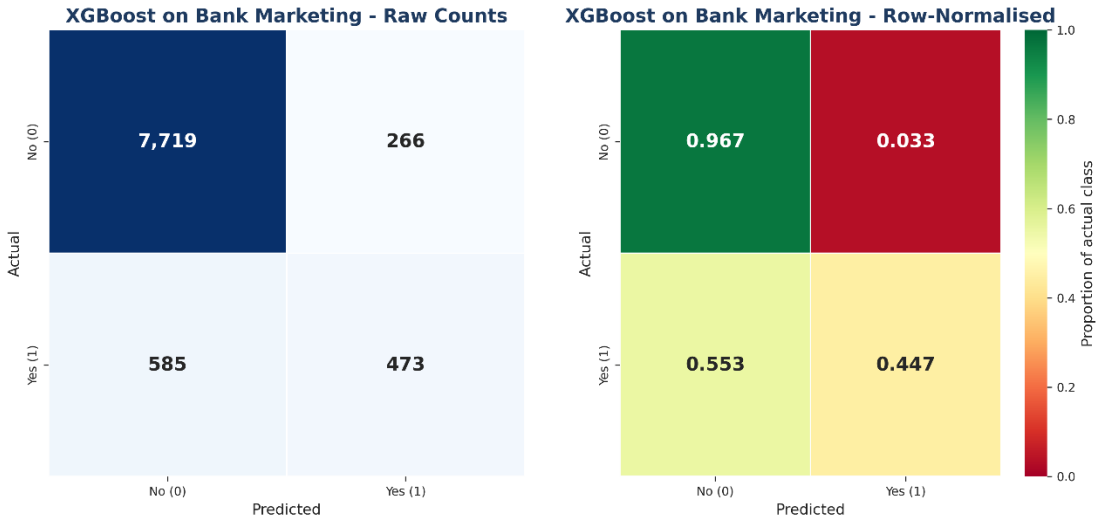


**Fig. 13** Confusion matrix (raw counts, left; row normalized proportions, right) of the calibrated model built with XG Boost and applied to the test set from the Bank Marketing (UCI) data set.

Sweeps for the Leads dataset are run for a range of decision thresholds, ranging from 0.0 to 1.0, and report Precision, Recall and F1-score for each threshold (Figure 14). The range of the calibrated score distribution is aligned well with the observed class boundary and lastly, the optimal F1-score is located just above the default threshold of 0.50. The threshold for which precision is > 0.95 is 0.85, which could help SMEs to advise their sales teams to prioritise those leads that have a higher probability to convert or to distribute resources more broadly over a wider pool of leads that may be less precise but can still be converted.

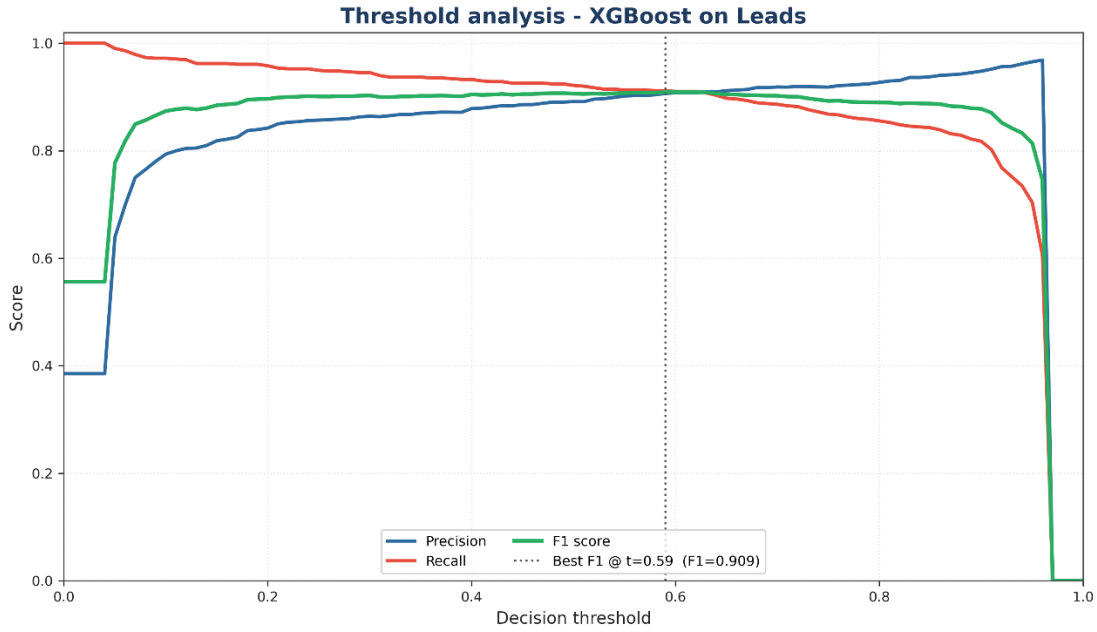


**Fig. 14** Precision, Recall and F1 score as functions of the decision threshold for the calibrated XGBoost model on the Leads (EdTech) test set; the dotted vertical line marks the F1-optimal threshold (t = 0.59, F1 = 0.909)

## *4.7 NLP-Derived Signals*

The rule-based intent classifier applied to the Leads text fields (Tags and Last Notable Activity) produced three non-empty classes: General Inquiry (8,720 records, 94.4 %), Enrollment Intent (513 records, 5.6 %) and Support Request (7 records, 0.1 %). Their conversion rates were 40.7 %, 2.5 % and 28.6 %, respectively. This pattern is counterintuitive because Enrollment Intent appears to be the highest-intent label, yet General Inquiry has the highest conversion rate. The result is mainly due to the keyword-based rule set: expressions such as interested in courses are also frequent among non-converting leads. Two methodological points follow from this finding. First, lexicon-based intent rules provide a weak and sometimes counterintuitive signal when keyword coverage is closely tied to class-conditional frequency. Second, the NLP module should be interpreted as an exploratory layer that complements the structured models rather than drives the pipeline. More advanced approaches, including transformer-based intent classification with DistilBERT or Sentence-BERT, or semantic clustering of free-text tags, may improve this part of the framework.

## *4.8 Ethical, Privacy and SME Deployment Considerations*

But, it should not be just about predictive performance. SME-oriented marketing automation system must also consider the legal and ethical (and operational) environment in which it will be used. Five dimensions are important. There, consent and lawful basis: Explicit consent and notice-at-collection requirements from GDPR Article 6 and CCPA should apply to any production deployment to score leads or interact with users via a chatbot. Second, human end-user interpretation: the SHAP explanations mentioned in Section 4.6, can be displayed as tables or plots that non-ML users can use to help explain automated decisions in a meaningful manner. Third, the framework should include regular checks on inequity in lead scores on the basis of some protected or sensitive attributes, e.g., geography, occupation, or group-based SHAP summaries can support these checks. Finally, decisions for qualification and outreach should be a human in the loop meaning that a human sales operator, not an automated lead tier makes the final call.

## *4.9 Limitations*

Although the framework has been reproduced and is available, there are still a number of limitations. To begin, conversational lead qualifier is a rule-based, script-driven module assessed in a lab setting versus with end users the quality of the conversation and user acceptance need to be determined in the field. Two, the current NLP methods, such as TextBlob, and other custom dictionaries, provide a normal structure for multilingual data, rather than a

customized model, which limits the contextual understanding of the NLP module. One of the contributing reasons to the counterintuitive results for intent class in Section 4.7 is this limitation. Thirdly, the input-output of the ROI dashboard are taken from literature, and the user's assumption of the ROI is a plan and not the returns realised in the actual deployment of real SMEs. Fourth, Validation is limited to only two public benchmark datasets (Kaggle Leads, UCI Bank Marketing) representing extraction from external validity in other sectors, world-wide, and in proprietary CRM systems. Fifthly, the duration attribute is recorded in the Bank Marketing dataset to make it easier to establish a baseline, and is also leak-prone because it is only recorded after completion of the call; this and features derived from it should be included during development but not included in the deployment for production. This framework has not yet been deployed in the SME production environment which operates in real-time, hence questions pertaining to data drifting, model refresh cycles, CRM integrations and user training are still wide open. The restrictions outlined here define the next step of research.

# 5. CONCLUSION

This work set up the nine-stage AI-powered marketing automation pipeline and tested its prototype empirically in two unique SME marketing environments. The pipeline includes integrated gradient-boosting classifiers, integrated SMOTE re-sampling, SHAP interpretability, sentiment and intent classification from NLP processing, a real time Chatbot Lead Qualifier and a business facing ROI Dashboard. It thereby brings a system performance of the model into line with a business-level decision support. The best result on the EdTech Leads dataset was obtained by XGBoost, while LightGBM was best on Bank Marketing and was trained in less than 1% of the time compared to XGBoost.

**References**


1. Holl A, Rama R. SME digital transformation and the COVID-19 pandemic: A case study of a hard-hit metropolitan area. Science and Public Policy. 2024;51(6):1212-1226. https://doi.org/10.1093/scipol/scae023
2. Asadi ZS, Hosseini SM, Ahmadi-Danesh-Ashtiani H, Mirzaeian B. Digital transformation in SMEs: Pre- and post-COVID-19 era - A comparative bibliometric analysis. Sustainability. 2024;16(23):10536. https://doi.org/10.3390/su162310536
3. Mohd Rasdi R, Umar Baki N. Navigating the AI landscape in SMEs: Overcoming internal challenges and external obstacles for effective integration. PLOS ONE. 2025;20(5):e0323249. https://doi.org/10.1371/journal.pone.0323249
4. Hussain A, Rizwan R. Strategic AI adoption in SMEs: A prescriptive framework (arXiv:2408.11825) [Preprint]. arXiv. 2024. https://doi.org/10.48550/arXiv.2408.11825
5. Yesuf Y, Fields Z, Jain A, Kassa E. Artificial intelligence adoption as a driver of innovation and competitiveness in SMEs: A bibliometric and systematic review. F1000Research. 2025;14:1456. https://doi.org/10.12688/f1000research.171494.1
6. Lu X, Wijayaratna K, Huang Y, Qiu A. AI-enabled opportunities and transformation challenges for SMEs in the post-pandemic era: A review and research agenda. Frontiers in Public Health. 2022;10:885067. https://doi.org/10.3389/fpubh.2022.885067
7. Sharma S, Singh G, Islam N, Dhir A. Why do SMEs adopt artificial intelligence-based chatbots? IEEE Transactions on Engineering Management. 2024;71:1773-1786. https://doi.org/10.1109/TEM.2022.3203469
8. Kedi WE, Ejimuda C, Idemudia C, Ijomah TI. AI chatbot integration in SME marketing platforms: Improving customer interaction and service efficiency. International Journal of Management & Entrepreneurship Research. 2024;6(7):2332-2341. https://doi.org/10.51594/ijmer.v6i7.1327
9. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM; 2016. p. 785-794. https://doi.org/10.1145/2939672.2939785
10. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T-Y. LightGBM: A highly efficient gradient boosting decision tree. In: Advances in Neural Information Processing Systems 30. Curran Associates; 2017. p. 3146-3154. https://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree
11. González-Flores L, Rubiano-Moreno J, Sosa-Gómez G. The relevance of lead prioritization: A B2B lead scoring model based on machine learning. Frontiers in Artificial Intelligence. 2025;8:1554325. https://doi.org/10.3389/frai.2025.1554325
12. Yim WY, Khaw KW, Lim ST, Chew X. Enhancing conversions and lead scoring in online professional education. International Journal of Management, Finance and Accounting. 2024;5(1):15-63. https://doi.org/10.33093/ijomfa.2024.5.1.2
13. Yu Q. Enhancing bank term deposit predictions: A machine learning approach with CatBoost and SHAP. Applied and Computational Engineering. 2025;120:171-180. https://doi.org/10.54254/2755-2721/2025.19485
14. Tanvir MF, Hossain MM, Jishan MA. Bayesian regression for predicting subscription to bank term deposits in direct marketing campaigns (arXiv:2410.21539) [Preprint]. arXiv. 2024. https://doi.org/10.48550/arXiv.2410.21539
15. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research. 2002;16:321-357. https://doi.org/10.1613/jair.953

16. Lavanya M, Kumar DP, Gopi K, Nagaraju J. Customer churn prediction in banking sector using machine learning [a UCI Bank Marketing term-deposit subscription study]. Journal of Information Systems Engineering and Management. 2025;10(57s):224-230. https://www.jisem-journal.com/index.php/journal/article/view/12181
17. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems 30. Curran Associates; 2017. p. 4765-4774. https://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions
18. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM; 2016. p. 1135-1144. https://doi.org/10.1145/2939672.2939778
19. Salih AM, Raisi-Estabragh Z, Boscolo Galazzo I, Radeva P, Petersen SE, Lekadir K, Menegaz G. A perspective on explainable artificial intelligence methods: SHAP and LIME. Advanced Intelligent Systems. 2024;6(7):2400304. https://doi.org/10.1002/aisy.202400304
20. Kedi WE, Ejimuda C, Idemudia C, Ijomah TI. AI software for personalized marketing automation in SMEs: Enhancing customer experience and sales. World Journal of Advanced Research and Reviews. 2024;23(1):1981-1990. https://doi.org/10.30574/wjarr.2024.23.1.2159
21. Alotaibi MZ, Haq MA. Customer churn prediction for telecommunication companies using machine learning and ensemble methods. Engineering, Technology & Applied Science Research. 2024;14(3):14572-14578. https://doi.org/10.48084/etasr.7480
22. Malik N, Bilal M. Natural language processing for analyzing online customer reviews: A survey, taxonomy, and open research challenges. PeerJ Computer Science. 2024;10:e2203. https://doi.org/10.7717/peerj-cs.2203
23. El Attar A, El-Hajj M. Explainable AI-driven customer churn prediction: A multi-model ensemble approach with SHAP-based feature analysis. Frontiers in Artificial Intelligence. 2026;8:1748799. https://doi.org/10.3389/frai.2026.1748799
24. Liu D, Huang H, Zhang H, Luo X, Fan Z. Enhancing customer behavior prediction in e-commerce: A comparative analysis of machine learning and deep learning models. Applied and Computational Engineering. 2024;55:181-195. https://doi.org/10.54254/2755-2721/55/20241408
25. European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on protection of natural persons with regard to processing of personal data and on free movement of such data (General Data Protection Regulation). Official Journal of the European Union. 2016;L119:1-88. https://eur-lex.europa.eu/eli/reg/2016/679/oj
26. Moro S, Cortez P, Rita P. A data-driven approach to predict the success of bank telemarketing. Decision Support Systems. 2014;62:22-31. https://doi.org/10.1016/j.dss.2014.03.001
27. Kaggle. Leads Dataset [Internet]. Available from: https://www.kaggle.com/datasets/ashydv/leads-dataset