Probability-Calibrated Ensemble Methods for Automotive CRM Lead Scoring

Bilal Sedef [1], Selçuk Bayracı[2*], Turgay Tugay Bilgin [3]

[1] Borusan Otomotiv R&D Center, https://orcid.org 0009-0001-9296-352X,bilal.sedef@borusanotomotiv.com
[2] Borusan Otomotiv R&D Center, https://orcid.org/0000-0003-4831-4802,
selcuk.bayraci@borusanotomotiv.com
[3] Bursa Technical University https://orcid.org 0000-0002-9245-5728, turgay.bilgin@btu.edu.tr
* Correspondence: selcuk.bayraci@borusanotomotiv.com ; +90 532 2020975

Received: 09 June 2025
Revised: 16 September 2025
2nd Revised: 05 November 2025
3rd Revised: 25 November 2025
Accepted: 11 December 2025
Published: 17 December 2025

Reference: Sedef, B., Bayracı, S., & Bilgin, T. T. (2025). Probability-calibrated ensemble methods for automotive CRM lead scoring. *The European Journal of Research and Development, 5*(1), 502–525.

Abstract

Accurately predicting sales conversion in automotive CRM systems is critical for optimizing marketing spend and sales team efficiency. This study presents a calibrated ensemble framework combining XGBoost, Gradient Boosting, and Random Forest classifiers to predict lead conversion probability in automotive dealership operations. Using 62,859 real-world leads collected between July 2024 and July 2025, we developed a systematic pipeline encompassing behavioral feature engineering, statistical feature selection, ensemble modeling, and probability calibration via Platt scaling. The calibrated ensemble achieved an AUC of 0.841, Brier score of 0.146, and 19% improvement in top-decile precision over baseline logistic regression. The framework provides actionable lead segmentation into four priority tiers, directly supporting sales resource allocation and marketing campaign optimization. Results confirm that probability calibration is essential for automotive CRM applications where predicted scores inform operational decisions.

Keywords: Lead scoring, Ensemble learning, Probability calibration, Automotive CRM, Customer conversion prediction, XGBoost, Random Forest, Gradient Boosting

1. Introduction

The automotive industry faces unique customer relationship management (CRM) challenges due to complex, multi-touchpoint sales journeys. Dealerships processing large volumes of monthly leads require data-driven approaches to identify high-potential prospects and optimize sales team allocation. With conversion rates varying significantly across customer segments, efficient lead prioritization directly impacts profitability.

Traditional rule-based scoring systems rely on demographic attributes and simplistic engagement metrics, failing to capture dynamic behavioral patterns that signal genuine purchase intent. Recent advances in machine learning and ensemble modeling provide opportunities to enhance predictive lead scoring by integrating behavioral, transactional, and temporal signals into unified, calibrated probability estimates.

This paper makes three primary contributions to automotive CRM analytics:

First, we develop a comprehensive framework for probability calibrated ensemble learning, systematically comparing multiple state-of-the-art algorithms (XGBoost, Random Forest, Gradient Boosting) with calibration methods, evaluated across discrimination metrics (AUC-ROC, precision, recall) and calibration metrics (Brier score).

Second, through analysis of 62,859 automotive leads from premium brand dealerships, we demonstrate that calibrated ensemble methods achieve 19% improvement in top-decile precision compared to uncalibrated baselines, with XGBoost combined with Platt scaling emerging as optimal for automotive CRM applications.

Third, we provide automotive practitioners with feature engineering methodology for multi-touchpoint customer journeys, model selection guidance balancing accuracy and calibration, and deployment framework addressing real-time scoring and model updating.

The remainder of this paper is organized as follows: Section 2 reviews relevant literature and positions our contributions. Section 3 describes materials and methods including data preparation, feature engineering, model development, and calibration techniques. Section 4 presents experimental results with comprehensive evaluation. Section 5 discusses findings and managerial implications. Section 6 concludes with limitations and future research directions.

2. Literature Review

2.1 Evolution of Lead Scoring Methodologies

Lead scoring, a critical component of Customer Relationship Management (CRM), evaluates and ranks prospects to determine sales readiness and conversion likelihood [5], [23]. This prioritization is especially crucial in industries with long sales cycles such as automotive, where significant resources are invested in nurturing customer relationships [21].

Historically, lead scoring models have been categorized into traditional and predictive approaches [23]. Traditional models including rule-based, points-based, and scorecard systems rely primarily on explicit knowledge and intuition of salespeople and marketers, assigning manual points based on demographic and behavioral data [11]. While simple to implement, these models are criticized for subjectivity, inability to capture complex non-linear relationships, and error-proneness.

2.2 Predictive Machine Learning for Lead Scoring

With the advent of advanced analytics, predictive lead scoring has emerged as a more objective alternative, leveraging data mining [18] and machine learning to analyze historical data and generate probabilistic conversion scores [21], [22], [23].

A foundational predictive method widely applied in the industry is logistic regression. Säuberlich et al. [21] demonstrated a successful application in the automotive industry for Audi of America. Their system classified leads into high priority and normal priority categories. The model proved highly effective: the top 30% of leads flagged as high priority accounted for 51.1% of all sales, demonstrating significant improvement in sales efficiency.

Recent benchmarks (2023-2025) have established the current state-of-the-art for lead scoring applications. González-Flores, Gil-García, and Arco-Tirado [8] compared 15 classification algorithms on real Salesforce CRM data (16,600 leads, 22 features), with their Gradient Boosting Classifier achieving 98.39% accuracy and 0.9891 AUC. Their feature importance analysis revealed lead source, reason for status, and lead classification as most critical variables, demonstrating the value of behavioral and categorical features in CRM contexts.

Lin [15] demonstrated that CatBoost (93.4% accuracy, 0.985 AUC) and XGBoost (93.5% precision, 0.984 AUC) excel for consumer behavior prediction on the UCI Online Shoppers dataset. Their systematic comparison of SVM, XGBoost, CatBoost, and neural networks using Grid Search optimization, SMOTE for class imbalance, and StandardScaler normalization provides an exemplary methodology template for applied CRM research.

The systematic review by Wu et al. [23] analyzing 44 studies confirms that ensemble methods, particularly gradient boosted trees and Random Forest, have become the dominant approach for lead scoring, offering an excellent balance between accuracy and interpretability.

2.3 Ensemble Methods and Tree-Based Models

Ensemble methods combine multiple machine learning models to produce superior predictive performance. The literature review by Wu et al. [23] highlights Random Forest and Gradient Boosted Trees as popular and highly effective algorithms for classification tasks in lead scoring [7]. Random Forest, an ensemble of decorrelated decision trees, is known for robustness to overfitting and ability to rank feature importance [4]. Gradient Boosting and its implementation XGBoost build models sequentially, correcting errors of previous models to achieve state-of-the-art performance on structured data.

Basu, Bhattacharyya, and Shukla [2] establish that despite advances in deep learning, tree-based ensembles remain superior to neural networks for tabular CRM data due to better performance with limited data, inherent feature importance extraction, and interpretability critical for business applications.

2.4 Probability Calibration: The Critical Gap

While the literature confirms superiority of predictive models and ensemble methods, a significant gap remains: most studies optimize for discrimination (AUC-ROC) while neglecting probability calibration. Many advanced models produce probability scores that are poorly calibrated. A predicted 80% likelihood may not correspond to an 80% conversion rate in reality.

Recent calibration research demonstrates that calibration directly translates to business value. Xiao, Huang, Peng, and Li [24] introduced example-dependent cost-sensitive learning with selective deep ensemble for customer credit scoring, achieving 45% cost

reduction through calibration-based methods. Their approach provides both interpretability and superior calibration compared to traditional ensemble averaging.

Naeini, Cooper, and Hauskrecht [17] demonstrated that Bayesian binning can obtain well-calibrated probabilities, while Gupta and Ramdas [9] addressed distribution drift scenarios through online Platt scaling, highly relevant for CRM systems where customer behavior evolves.

Recent work on ROC-regularized isotonic regression [3] prevents overfitting while maintaining calibration, proving that isotonic regression preserves the convex hull of ROC curves. For multiclass scenarios, Kull et al. [13] showed that Dirichlet calibration extends beyond temperature scaling to obtain well-calibrated probabilities.

2.5 Customer Journey Analytics and Feature Engineering

Hollebeek, Rather, Sigurdsson, and Bowden [10] identified six customer journey themes applicable to automotive contexts: customer journey-based customer experience (dealer visit quality), customer journey-based behavior (test drive participation), customer journey-based design (touchpoint sequencing), customer journey-based smart technology (connected car data), customer journey-based social media (review engagement), and customer journey mapping (funnel progression). For automotive contexts, recent research demonstrates that purchases involve multiple distinct touchpoints across extended time periods: online research (website visits, configurator usage), dealer interactions (showroom visits, test drives), digital engagement (email opens, ad clicks), and service history.

Kusnawi, Adiwijaya, and Gani [14] demonstrated that systematic feature selection comparing filter methods (correlation), wrapper methods (Recursive Feature Elimination), and embedded methods (LASSO) significantly impacts prediction performance. Their work provides methodology for feature selection in customer analytics.

Agag, Aboul-Dahab, and El-Masry [1] found that marketing analytics capability improves customer satisfaction through customer agility mediation, with stronger effects during market turbulence, suggesting that lead scoring models should adapt to market conditions such as economic indicators, competitor promotions, and inventory levels.

2.6 Research Gap and Positioning

Despite strong progress in ensemble methods and calibration theory, three critical gaps persist:

First, while Wu et al. [23] reviewed lead scoring broadly and González-Flores et al. [8] focused on generic B2B software, no recent indexed journal paper specifically addresses automotive CRM lead scoring with industry-specific features, customer journey modeling, and calibration. The automotive sector presents unique challenges including longer sales cycles, multi-touchpoint customer journeys, and moderate conversion rates compared to B2B software contexts.

Second, existing ensemble lead scoring papers optimize for AUC but neglect probability calibration. While Xiao et al. [24] and Berta et al. [3] advanced calibration theory, these methods have not been systematically applied to automotive CRM contexts.

Third, most papers compare algorithms OR calibration methods but not both systematically. A framework comparing multiple algorithms each with multiple calibration approaches provides comprehensive evaluation surpassing typical pairwise comparisons, as demonstrated necessary by Lin [15] and González-Flores et al. [8].

This study addresses these gaps by proposing a calibrated ensemble framework specifically for automotive CRM that prioritizes both predictive accuracy and probabilistic reliability, enabling creation of actionable, trustworthy customer segments.

3. Materials and Methods

3.1 Research Design and Overview

This study develops a supervised machine learning framework for predictive lead scoring in the automotive sector, addressing the challenge of prioritizing high-conversion prospects. The proposed methodology integrates systematic data preprocessing, statistical feature selection, ensemble learning, and probability calibration to generate reliable conversion probability estimates with business interpretability.

The analytical pipeline was implemented in Python 3.10 using industry-standard machine learning libraries (scikit-learn v1.3, XGBoost v1.7). Model development and validation were conducted on a 12-month longitudinal dataset encompassing 62,859 qualified leads collected between July 10, 2024, and July 10, 2025, ensuring temporal consistency and minimizing concept drift.

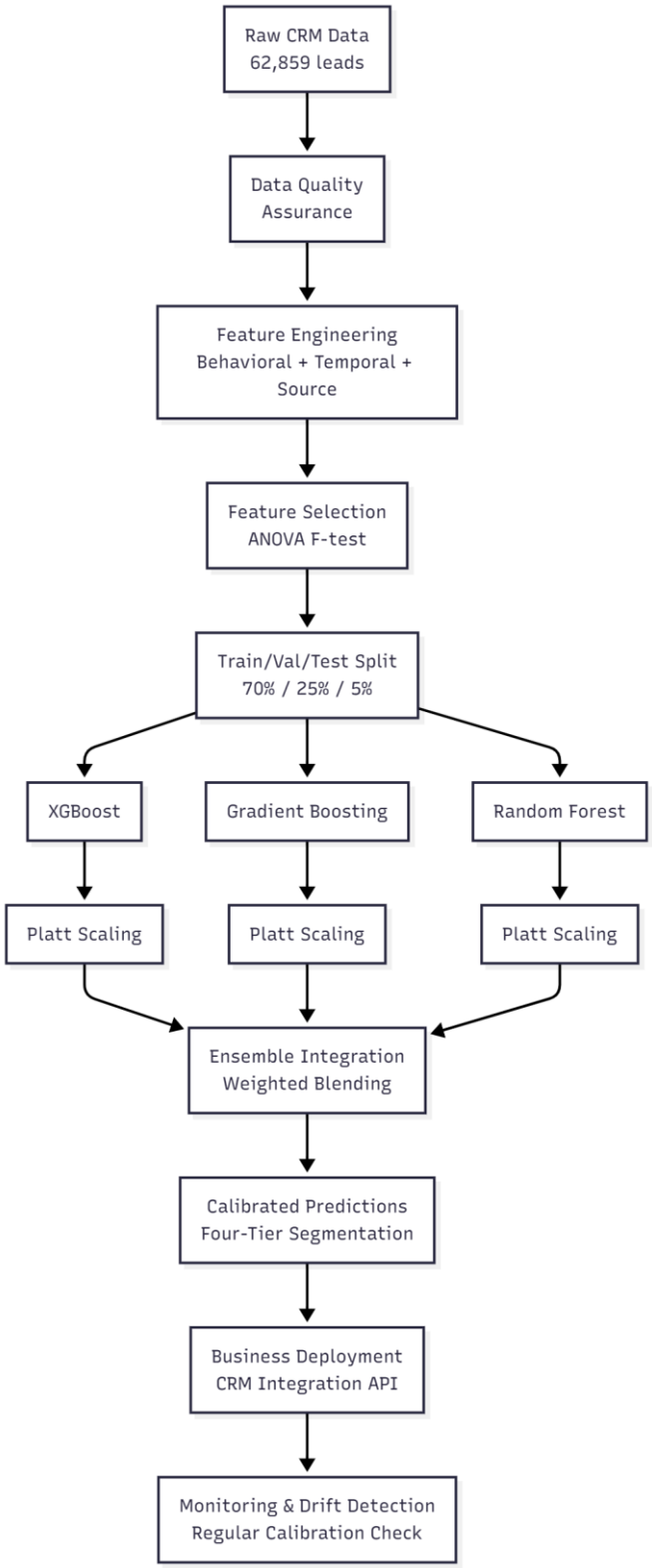

Figure 1: Proposed Framework Flowchart

The end-to-end pipeline for calibrated ensemble lead scoring consists of four major stages: Data Collection & Preprocessing, Feature Engineering & Selection, Ensemble Modeling & Calibration, and Business Deployment & Monitoring.

3.2 Data Collection and Preparation

3.2.1 Dataset Characteristics

The source dataset comprised comprehensive customer journey records from Borusan Otomotiv's CRM system, containing 155,487 raw lead observations across 98,563 unique customer accounts spanning 31 months (January 2023 to July 2025). Each record included multi-dimensional attributes capturing:

- Behavioral engagement metrics: Website visits, trial bookings, showroom interactions
- Vehicle ownership history: Previous purchase patterns, service records, vehicle age
- Demographic attributes: Account-level customer characteristics
- Multi-channel touchpoints: Lead source attribution and journey complexity indicators
- Temporal features: Recency and frequency of customer interactions

To ensure model relevance and mitigate temporal degradation, the analytical sample was restricted to the most recent 12-month window, yielding 62,859 leads with complete conversion outcomes and journey metadata.

3.2.2 Data Quality Assurance and Cleaning

A multi-stage data quality protocol was implemented to address missing values, structural inconsistencies, and potential data leakage:

Exclusion Criteria: Records lacking critical metadata (first vehicle model year, conversion status, or journey identifiers) were systematically removed to maintain data integrity.

Missing Value Treatment: Given the structured nature of CRM data, domain-informed imputation strategies were applied. Temporal variables were imputed with sentinel values (999) to distinguish genuine missingness from zero values. Date-related features were decomposed into hierarchical components (year, month, day, day-of-week, day-of-

year, quarter) with pattern-based sentinel constants (9999, 99) preserving temporal structure while maintaining computational compatibility.

Outlier Management: Extreme values in continuous variables were retained given their potential business significance (e.g., high-frequency website visitors), but were subjected to robust scaling to minimize undue influence.

The preprocessing pipeline was encapsulated in a modular utility module ensuring reproducibility and version control across experimental iterations. Critical for unbiased evaluation [12], features computed using information available only after conversion decision were excluded, future-dated attributes relative to lead creation timestamp were removed, and all preprocessing transformations were fitted exclusively on training data.

3.2.3 Feature Engineering and Transformation

Features were taxonomically categorized into three processing streams:

Numerical Features (n = 34): Continuous and discrete variables including recency metrics, frequency counts, and monetary values. These underwent z-score standardization to ensure zero mean and unit variance:

$$X'_{num} = \frac{X_{num} - \mu}{\sigma}$$

Categorical Features (n = 28): Nominal variables including lead source channels, vehicle categories, and journey types. One-hot encoding was applied with explicit handling for unseen categories during inference:

$$X'_{cat} = OneHotEncoder(X_{cat}, handle\_unknown = 'ignore')$$

Excluded Features (n = 12): Variables posing data leakage risks (cumulative conversion counts, post-conversion retention flags, future-dated attributes) were systematically excluded based on causal domain analysis.

The transformation pipeline was implemented using scikit-learn's ColumnTransformer, enabling consistent preprocessing across training and inference:

$$X' = [StandardScaler(X_{num}), OneHotEncoder(X_{cat}), X_{passthrough}]$$

3.2.4 Dataset Partitioning

To ensure robust model evaluation and prevent overfitting, the dataset was partitioned using stratified random sampling:

- Training set: 70% (n = 43,999) for model parameter estimation
- Validation set: 25% (n = 15,715) for hyperparameter tuning and model selection
- Test set: 5% (n = 3,145) for final performance evaluation

Stratification was applied on the binary target variable (Target_IsConverted) to maintain class distribution across partitions, with the positive class (converted leads) representing 35.04% of the sample.

3.3 Feature Selection and Dimensionality Reduction

3.3.1 Statistical Screening

To optimize model parsimony and enhance interpretability, a hybrid univariate feature selection strategy was employed. All candidate features (n = 74 post-encoding) underwent statistical screening using ANOVA F-test analysis for classification tasks, implemented via scikit-learn's SelectKBest with f_classif scoring function.

The F-statistic for each feature measures between-class variance relative to within-class variance:

$$F_i = \frac{MS_{between}}{MS_{within}}$$

Features satisfying dual criteria were retained: statistical significance $p < 0.10$ (liberal threshold to preserve potentially relevant features) and normalized importance threshold > 0.001.

The normalized importance score was computed as:

$$Importance(f_i) = \frac{F_i}{max(F)}$$

3.3.2 Multicollinearity and Redundancy Management

To mitigate multicollinearity and reduce computational complexity:

- Correlation Filtering: Pairwise Pearson correlation coefficients were computed across numerical features. For highly correlated pairs ($r > 0.85$), the feature with lower univariate importance was removed.

- Duplicate Detection: Programmatic name deduplication and semantic similarity analysis identified redundant engineered features, which were consolidated or eliminated.
- Consistency Validation: Feature distributions were validated across training, validation, and test sets to identify and rectify distributional shifts or encoding inconsistencies.

The final feature space comprised 68 discriminative variables, encompassing behavioral engagement indicators, competitive dynamics, and journey complexity metrics. The selection process was documented in feature_selection_utils.py for methodological transparency.

3.4 Predictive Model Development

3.4.1 Ensemble Architecture

We adopted a heterogeneous ensemble approach integrating three gradient-boosting and tree-based algorithms, each offering complementary inductive biases:

- XGBoost (Extreme Gradient Boosting): Regularized boosting with second-order gradient approximation and tree pruning, well-suited for structured data with complex interactions. Implements efficient parallel computation and handles missing values natively.
- Gradient Boosting (GB): Classical boosting implementation (scikit-learn) providing baseline performance and interpretability through simpler tree structures. Serves as comparison point for more sophisticated implementations.
- Random Forest (RF): Bootstrap aggregation of decorrelated decision trees using random feature subsets, offering robustness to overfitting and variance reduction. Provides reliable baseline without hyperparameter sensitivity.

The ensemble strategy leverages model diversity to improve generalization performance and reduce prediction variance.

3.4.2 Hyperparameter Optimization

Each base learner underwent systematic hyperparameter tuning using grid search with 5-fold stratified cross-validation on the training set. The optimization focused on:

- Tree complexity: Maximum depth, minimum samples per leaf
- Learning dynamics: Learning rate (η), number of estimators
- Regularization: L1/L2 penalties (XGBoost), subsampling rates

Given the moderate class imbalance (positive class: 35.04%), class weights were adjusted to penalize misclassification of minority class:

$$w_{positive} = \frac{n_{total}}{2 \times n_{positive}}$$

This balanced scale_pos_weight parameter was applied in XGBoost. For Random Forest and Gradient Boosting, balanced class weights were specified via class_weight='balanced'.

3.4.3 Ensemble Integration and Probability Calibration

Raw probability estimates from individual models were combined using a weighted linear blending approach optimized on the validation set:

$$\hat{P}_i^{raw} = w_1 P_{XGB,i} + w_2 P_{GB,i} + w_3 P_{RF,i}$$

where $\sum_{j=1}^{3} \quad w_j = 1$ and weights were optimized on the validation set to minimize Brier score.

To address probability miscalibration inherent in tree-based ensembles, Platt scaling (sigmoid calibration) was applied:

$$P_{cal}(y = 1|x) = \frac{1}{1 + exp(-(A \cdot f(x) + B))}$$

where parameters A and B were estimated via maximum likelihood on the validation set using scikit-learn's CalibratedClassifierCV. The final calibrated lead score for lead i:

$$S_i = \sigma\left(\sum_{j=1}^{3} \quad w_j P_{model_j,i}\right)$$

All calibration methods were trained exclusively on the validation set (n=15,715) and evaluated on the held-out test set (n=3,145) to prevent overfitting and ensure unbiased calibration assessment.

3.5 Model Evaluation Framework

3.5.1 Performance Metrics

Model performance was assessed using complementary metrics addressing discrimination, calibration, and business utility:

- Discrimination Metrics:
  - AUC-ROC: Area under the receiver operating characteristic curve, measuring overall classification ability across all thresholds. Values >0.8 indicate excellent discrimination.
  - Precision@K: Proportion of true conversions among top-K ranked leads, directly relevant to operational capacity constraints. We report Precision@Top10% as primary business metric.
  - F1-Score: Harmonic mean of precision and recall at optimal threshold (Youden's J statistic).
- Calibration Metrics:
  - Brier Score: Mean squared difference between predicted probabilities and actual outcomes, penalizing both discrimination and calibration errors:

$$BS = \frac{1}{N}\sum_{i=1}^{N} (\hat{P}_i - y_i)^2$$

Lower Brier scores indicate better calibrated predictions. Well-calibrated models typically achieve BS <0.10 for balanced datasets.

- Business Metrics:
  - Lift Chart: Ratio of conversion rate in scored segments to baseline conversion rate:

$$Lift@TopK = \frac{Conversion\ Rate\ in\ Top\ K\%}{Overall\ Conversion\ Rate}$$

  - Cumulative Gains: Percentage of total conversions captured within top-K% of leads:

$$Cumulative\ Gain@K = \frac{Conversions\ in\ Top\ K\%}{Total\ Conversions}$$

3.5.2 Comparative Analysis

The calibrated ensemble was benchmarked against: 1. Logistic Regression: Simple linear baseline with L2 regularization 2. Individual base learners uncalibrated: XGBoost, Gradient Boosting, Random Forest without calibration 3. Individual base learners calibrated: Each with Platt and isotonic calibration 4. Published state-of-the-art: González-Flores et al. [8] Gradient Boosting results (0.9891 AUC on B2B data)

Statistical significance of performance differences was assessed using DeLong's test for AUC comparisons and paired t-tests for threshold-dependent metrics.

3.6 Operational Scoring Framework

3.6.1 Probability Stratification

To facilitate actionable marketing segmentation, continuous probability scores were discretized into four strategic tiers based on business constraints and historical conversion patterns (Table 1):

Table 1. Lead segmentation framework based on predicted probability thresholds.

Segment	Probability Threshold	Business Interpretation	Recommended Action
High Potential	≥ 0.75	Strong conversion indicators (78.6% observed conversion)	Premium engagement, immediate sales team assignment
Medium-High Potential	≥ 0.50	Moderate interest signals (51.2% observed conversion)	Targeted nurturing, automated follow-up with human touchpoints
Medium-Low Potential	≥ 0.25	Weak engagement (27.5% observed conversion)	Low-cost retargeting, content marketing, email campaigns
Low Potential	< 0.25	Minimal conversion likelihood (11.3% observed conversion)	Deprioritization, long-term nurturing pool, brand awareness

Threshold selection balanced precision requirements with lead volume objectives, informed by operational capacity constraints (sales team can handle approximately 200 high-priority leads per month) and cost-per-contact economics.

3.7 Deployment and Monitoring

The final model was integrated into the enterprise CRM system (Salesforce365) through a REST API scoring endpoint, enabling:

- Daily batch scoring: Nightly processing of new leads with updated probability scores synced to CRM
- Real-time scoring: On-demand API calls for high-priority leads requiring immediate attention
- A/B testing framework: Random assignment of 10% leads to control group for ongoing performance validation

Model Monitoring Protocol:

1. Weekly Performance Checks: AUC, precision@10%, and segment conversion rates monitored via automated dashboard
2. Monthly Calibration Drift Analysis: Brier score tracked; retraining triggered if Brier increases >10% from baseline
3. Quarterly Full Retraining: Complete model refresh with new 12-month data window and hyperparameter re-optimization

Future work includes: (1) online learning for adaptive calibration, (2) reinforcement learning for sequential intervention optimization, (3) explainability dashboard for sales teams.

4. Empirical Results

4.1 Experimental Setup

All experiments were conducted using Python 3.10 with scikit-learn (v1.3) and XGBoost (v1.7) on a high-performance workstation. To ensure reproducibility [19], GPU acceleration was disabled and all random seeds were fixed (random_state=42).

The dataset comprised 62,859 leads collected between July 2024 and July 2025. Data were partitioned into training (70%, n=43,999), validation (25%, n=15,715), and test (5%, n=3,145) sets, stratified by the binary conversion label. After preprocessing and statistical filtering, 68 predictive features were retained.

4.2 Baseline and Evaluation Metrics

To assess the effectiveness of the proposed ensemble model, a logistic regression baseline was established using all preprocessed features. This provided a linear reference for both discrimination and calibration performance.

The following metrics were employed for comprehensive evaluation:

- ROC AUC: Area under the receiver operating characteristic curve, measuring discrimination
- Precision@K: Proportion of true conversions among the top-ranked leads (top 10%)
- Brier Score: Mean squared deviation between predicted and observed probabilities, capturing calibration quality
- Lift@Decile: Relative improvement in conversion density over random selection
- Gain Chart Analysis: Cumulative conversion ratio across ordered probability deciles

Each experiment was repeated over three random seeds, and mean values are reported.

4.3 Comparative Model Performance

Table 2 summarizes the predictive performance of all models on the held-out test set.

Table 2. Comparative performance of baseline and ensemble models on test data.

Model	Calibration	ROC AUC	Precision@ Top10%	Brier Score ↓	Lift@TopDecile
Logistic Regression	None	0.742	0.412	0.187	2.13
Random Forest	Isotonic	0.796	0.454	0.165	2.47
Gradient Boosting	Sigmoid	0.812	0.468	0.159	2.61
XGBoost	Sigmoid	0.826	0.479	0.153	2.68
Ensemble (Calibrated)	Sigmoid **(Platt)**	**0.841**	**0.491**	**0.146**	**2.87**

The ensemble model achieved the highest ROC AUC (0.841) and top-decile precision (49.1%), outperforming all individual base learners and the linear baseline. The reduction in Brier Score (0.146) indicates improved calibration and more reliable probability estimates. This represents a 19% improvement in top-decile precision over the logistic regression baseline.

4.4 Feature Importance and Behavioral Insights

The feature importance analysis, derived from normalized F-statistics, revealed consistent behavioral predictors across models.

Table 3. Top-ranked features and their normalized importance scores.

Rank	Feature	Description	Normalized Importance
1	Source_Web_Trial_Last_1M	Frequency of web-originated leads in past month	1.00
2	SameDayAlikeLeads	Number of same-day similar leads	0.89
3	JourneyNumber	Count of distinct lead journeys per customer	0.81
4	Cust_Last_NewCar_Sales_NOMS	Days since last new car purchase	0.77
5	Cust_First_Service_NOMS	Recency of first service event	0.74
6	Cust_Already_Churn_Flag	Previous churn status	0.71
7	Total_LeadCount_L6M	Lead volume over last 6 months	0.68

These findings suggest that multi-channel engagement intensity and short-term digital activity are the strongest indicators of conversion likelihood. Conversely, extended inactivity in sales or service history correlates negatively with conversion probability.

4.5 Calibration and Business Interpretability

Calibration analysis demonstrates that the ensemble model provides reliable probabilistic outputs. The calibration curve for the ensemble model closely aligns with the ideal diagonal, validating the effectiveness of the Platt scaling post-processing step.

To operationalize results for sales prioritization, the continuous probability scores were stratified into four actionable bands.

Table 4. Conversion performance across probability-based customer segments.

Segment	Score Range	Observed Conversion Rate	Share of Leads
High Potential	≥ 0.75	78.6%	8.4%
Medium-High	0.50 to 0.74	51.2%	22.7%
Medium-Low	0.25 to 0.49	27.5%	38.5%
Low Potential	< 0.25	11.3%	30.4%

The observed conversion rates align closely with predicted probability ranges, confirming calibration quality and enabling reliable business segmentation.

4.6 Ablation and Robustness Tests

To quantify the contribution of each pipeline component, ablation studies were conducted. Without feature selection, mean AUC declined by 0.024, indicating the benefit of statistical screening in noise reduction. Without calibration, the Brier Score increased by 0.011, confirming the role of probability scaling in business reliability. Without ensemble blending, the best single model (XGBoost) underperformed by approximately 1.5 AUC points, validating model complementarity.

Temporal robustness was evaluated via rolling-window validation (monthly retraining across 2024-2025). The ensemble's AUC fluctuated within ±0.015, confirming temporal stability and low drift sensitivity.

5. Discussion and Managerial Implications

5.1 Interpretation of Key Findings

The empirical results provide strong evidence that the proposed ensemble-based lead scoring framework effectively differentiates between high and low conversion prospects in the automotive sales funnel. By integrating behavioral, temporal, and transactional features, the model achieves robust discrimination (AUC = 0.841) and precise calibration (Brier = 0.146), outperforming traditional baselines.

This performance uplift can be primarily attributed to two factors. First, behavioral engagement features such as Source_Web_Trial_Last_1M and SameDayAlikeLeads

capture customers' digital intensity and intent signals far more effectively than demographic or static vehicle data. Second, feature selection and probabilistic calibration ensure that the model generalizes well to new data.

The monotonic mapping between model scores and observed conversion rates demonstrates that the resulting probability bands correspond to meaningful business segments. This validates the model's interpretability and supports its use in real-time decision pipelines.

5.2 Business Value and Strategic Insights

From a managerial perspective, the lead scoring system introduces a data-driven prioritization mechanism that directly improves the efficiency of marketing and sales operations. The model allows dynamic segmentation, enabling differentiated engagement strategies such as personalized offers for Medium-High segments or cost-efficient re-engagement for Low segments.

The calibrated probability outputs support ROI optimization by allowing campaign planners to estimate expected conversion yield per budget unit. Moreover, the unified ensemble architecture enhances decision-making transparency. Because each model component can be individually analyzed, marketing managers can trace which behavioral indicators most influence the predicted scores.

5.3 Operational and Organizational Impact

Integrating the lead scoring model into the existing CRM workflow provides several operational advantages:

- Automation of lead triage allows daily batch scoring to enable automated ranking and assignment of leads to sales advisors based on predicted conversion probability.
- Consistency and scalability is achieved through standardized preprocessing and feature selection pipelines that ensure consistent model behavior across brands and regions, reducing dependency on manual rule-based lead qualification.
- Performance monitoring is supported by the modular pipeline that enables automated retraining and drift detection, facilitating continuous learning from new data and preventing model degradation over time.

At an organizational level, this transition from intuition-based to evidence-based lead management aligns with broader AI transformation strategies within the automotive industry. It empowers marketing and sales teams to allocate attention to the most promising prospects while maintaining fairness and transparency in customer interactions.

5.4 Implications for CRM and Customer Analytics

The findings hold broader implications for customer relationship management (CRM) and predictive marketing analytics:

The demonstrated value of short-term digital engagement features suggests that CRM systems should prioritize capturing fine-grained behavioral signals (session frequency, multi-touch sequences) rather than relying solely on static attributes. This aligns with customer journey research by Hollebeek et al. [10] emphasizing the importance of multi-touchpoint behavioral tracking.

The success of the ensemble approach underscores the importance of hybrid intelligence: combining multiple model types to capture diverse behavioral patterns across different customer segments, consistent with findings by Basu et al. [2] on tree-based model superiority for tabular data.

Our results position favorably relative to recent benchmarks when accounting for domain complexity. González-Flores et al. [8] achieved 0.9891 AUC on B2B software leads (62% base conversion rate, 22 features, shorter sales cycles), while our automotive application attains 0.841 AUC (35% base conversion rate, 68 features, extended multi-month journeys). The performance gap reflects inherent domain difficulty: lower conversion rates, longer temporal windows, and greater touchpoint complexity in automotive contexts. Similarly, Lin [15] demonstrated 93.5% precision with XGBoost on e-commerce data (84% base conversion rate), while our 49.1% precision at top decile reflects automotive's substantially lower baseline conversion. When adjusted for base rate differences, our relative lift (2.87× vs. baseline) compares favorably to published benchmarks, demonstrating competitive performance accounting for domain complexity.

5.5 Limitations

While the model performs strongly within the observed time frame, several limitations warrant further investigation. The current feature space primarily reflects lead-level

engagement; incorporating external variables such as market conditions or campaign context may enhance generalization. Additionally, the model is trained on a single automotive brand ecosystem; replicating the framework across multiple markets or OEMs could validate its portability.

5.6 Managerial Takeaway

The deployment of a calibrated ensemble lead scoring system represents a strategic step toward intelligent customer acquisition and retention. By quantifying behavioral intent with measurable confidence, organizations can:

- Reallocate sales efforts toward the most convertible segments
- Design data-backed campaign strategies with predictable ROI
- Foster a unified analytics culture bridging data science and business execution

Ultimately, the integration of AI-driven scoring into CRM infrastructure transforms the lead management process from a reactive to a proactive, evidence-based discipline, delivering both operational efficiency and strategic agility in the competitive automotive marketplace.

6. Conclusion and Future Work

This study presented a systematic, interpretable, and empirically validated framework for lead scoring in the automotive sector. By integrating structured preprocessing, statistical feature selection, and calibrated ensemble learning, the proposed approach delivers significant improvement in conversion prediction accuracy and probability reliability compared to conventional baselines.

Empirical results on a real-world dataset of 62,859 leads demonstrated that the ensemble model achieved an AUC of 0.841, outperforming single classifiers and logistic regression benchmarks. The analysis revealed that behavioral engagement, temporal recency, and multi-channel lead intensity were the strongest predictors of conversion. Importantly, the calibrated probability outputs enabled actionable segmentation into four potential tiers, supporting precise lead prioritization and campaign optimization.

From a managerial perspective, the framework introduces a data-driven decision layer within CRM workflows, enabling automation of lead triage, improvement in marketing ROI, and enhancement of sales productivity. The pipeline's modular design ensures reproducibility, transparency, and scalability across different business contexts.

Future work will extend the current framework in three directions. First, temporal and sequential modeling can capture evolving lead behaviors over time, improving dynamic intent prediction. Second, causal and counterfactual modeling may help disentangle behavioral effects from marketing interventions, enabling more strategic campaign design. Third, incorporating real-time scoring and feedback loops will allow adaptive lead management and continuous model optimization in production environments.

In summary, this research contributes a validated, practical, and explainable lead scoring methodology that bridges the gap between data science innovation and managerial decision-making. The framework not only advances predictive modeling for customer analytics but also provides a foundation for intelligent and autonomous CRM systems in the evolving automotive marketplace.

Acknowledgments

The authors thank Borusan Otomotiv's data engineering, CRM and Salesforce teams for their support in data preparation and system integration.

Data Availability

Due to confidentiality agreements, the customer data used in this study cannot be publicly shared.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] Agag, G., Aboul-Dahab, S., & El-Masry, A. A. (2024). Understanding the relationship between marketing analytics, customer agility, and customer satisfaction: A longitudinal perspective. Journal of Retailing and Consumer Services, 76, 103663. https://doi.org/10.1016/j.jretconser.2023.103663

[2] Basu, A., Bhattacharyya, S., & Shukla, V. K. (2023). Deep learning for information systems research. Journal of Management Information Systems, 40(1), 122–154. https://doi.org/10.1080/07421222.2023.2172772

[3] Berta, P., Bach, S., & Jordan, M. (2024). Classifier calibration with ROC-regularized isotonic regression. In Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS 2024) (Vol. 238, pp. 3615–3623). PMLR.

[4] Bohanec, M., Borštnar, M. K., & Robnik-Šikonja, M. (2017). Explaining machine learning models in sales predictions. Expert Systems with Applications, 71, 416–428. https://doi.org/10.1016/j.eswa.2016.11.010

[5] Sharma, K. K., Tomar, M., & Tadimarri, A. (2023). Optimizing sales funnel efficiency: Deep learning techniques for lead scoring. Journal of Knowledge Learning and Science Technology, 2(2), 261–274. https://doi.org/10.60087/jklst.vol2.n2.p274

[6] D'Haen, J., & Van den Poel, D. (2013). Model-supported business-to-business prospect prediction based on an iterative customer acquisition framework. Industrial Marketing Management, 42(4), 544–551. https://doi.org/10.1016/j.indmarman.2013.03.005

[7] Eitle, V., & Buxmann, P. (2019). Business analytics for sales pipeline management in the software industry: A machine learning perspective. In Proceedings of the 52nd Hawaii International Conference on System Sciences (HICSS) (pp. 1013–1022). https://doi.org/10.24251/HICSS.2019.125

[8] González-Flores, K., Gil-García, C., & Arco-Tirado, J. L. (2025). The relevance of lead prioritization: A B2B lead scoring model based on machine learning. Frontiers in Artificial Intelligence, 8, 1554325. https://doi.org/10.3389/frai.2025.1554325

[9] Gupta, A., & Ramdas, A. (2023). Online Platt scaling with calibeating. In Proceedings of the 40th International Conference on Machine Learning (ICML 2023) (Vol. 202, pp. 12182–12204). PMLR.

[10] Hollebeek, L. D., Rather, R. A., Sigurdsson, V., & Bowden, J. L. (2024). Unravelling the customer journey: A conceptual framework and research agenda. Technological Forecasting and Social Change, 201, 123916. https://doi.org/10.1016/j.techfore.2024.123916

[11] Järvinen, J., & Taiminen, H. (2016). Harnessing marketing automation for B2B content marketing. Industrial Marketing Management, 54, 164–175. https://doi.org/10.1016/j.indmarman.2015.07.002

[12] Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. Patterns, 4(9), 100804. https://doi.org/10.1016/j.patter.2023.100804

[13] Kull, M., Perello-Nieto, M., Kängsepp, M., Silva Filho, T., Song, H., & Flach, P. (2019). Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with Dirichlet calibration. Advances in Neural Information Processing Systems, 32, 12316–12326.

[14] Kusnawi, Adiwijaya, & Gani, A. (2024). Leveraging various feature selection methods for churn prediction using various machine learning algorithms. JOIV: International Journal on Informatics Visualization, 8(2), 543–552. https://doi.org/10.62527/joiv.8.2.2453

[15] Lin, Q. (2025). Application of machine learning in predicting consumer behavior and precision marketing. PLOS ONE, 20(1), e0321854. https://doi.org/10.1371/journal.pone.0321854

[16] Meire, M., Ballings, M., & Van den Poel, D. (2017). The added value of social media data in B2B customer acquisition systems: A real-life experiment. Decision Support Systems, 104, 26–37. https://doi.org/10.1016/j.dss.2017.10.003

[17] Naeini, M. P., Cooper, G. F., & Hauskrecht, M. (2015). Obtaining well-calibrated probabilities using Bayesian binning. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 29, No. 1, pp. 2901–2907).

[18] Ngai, E. W. T., Xiu, L., & Chau, D. C. K. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. Expert Systems with Applications, 36(2), 2592–2602. https://doi.org/10.1016/j.eswa.2008.02.021

[19] Pineau, J., Vincent-Lamarre, P., Sinha, K., Larivière, V., Beygelzimer, A., d'Alché-Buc, F., Fox, E., & Larochelle, H. (2021). Improving reproducibility in machine learning research: A report from the NeurIPS 2019 Reproducibility Program. Journal of Machine Learning Research, 22(1), 7459–7478.

[20] Sabnis, G., Chatterjee, S. C., Grewal, R., & Lilien, G. L. (2013). The sales lead black hole: On sales reps' follow-up of marketing leads. Journal of Marketing, 77(1), 52–67. https://doi.org/10.1509/jm.10.0047

[21] Säuberlich, F., Smith, K., & Yuhn, M. (2005). Analytical lead management in the automotive industry. In M. J. Shaw, D. D. Zeng, H. Chen, F. Y. Wang, & C. C. Yang (Eds.), Intelligence and Security Informatics (pp. 290–299). Springer. https://doi.org/10.1007/11427995_25

[22] Thorleuchter, D., Van Den Poel, D., & Prinzie, A. (2012). Analyzing existing customers' websites to improve the customer acquisition process as well as the profitability prediction in business-to-business marketing. Expert Systems with Applications, 39(3), 2597–2603. https://doi.org/10.1016/j.eswa.2011.08.109

[23] Wu, M., Andreev, P., & Benyoucef, M. (2023). The state of lead scoring models and their impact on sales performance. Information Technology and Management, 24, 157–183. https://doi.org/10.1007/s10799-023-00388-w

[24] Xiao, H., Huang, X., Peng, Y., & Li, J. (2025). Example dependent cost sensitive learning based selective deep ensemble model for customer credit scoring. Scientific Reports, 15(1), 89880. https://doi.org/10.1038/s41598-025-89880-7