International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 318-326 March 2026-Sept 2026 www.csjournals.com

Analysis of Machine Learning Models for Click-Through Rate Prediction in Email Marketing Campaigns

Ajay Kumar, Dr. Vishal Verma, Mr. Bajrang

[1]MCA Student, [2]Professor, [3]Asst. Prof.

Deptt. of Comp. Sc. and Application, Ch. Ranbir Singh University, Jind, Haryana, India

Abstract

Of all digital channels, email marketing boasts one of the highest returns on investment, yet Click-Through Rate (CTR) the percentage of recipients who click a link in a campaign email is low industry-wide (2.5–5%). Accurate CTR predictions before a campaign is launched can help marketers customize the content, fine-tune the time to send, and improve segmentation of subscribers. In this paper, we compare the performance of three supervised machine learning classifiers namely Logistic Regression, Random Forest and Gradient boosting on a publicly available email marketing campaign dataset (8,200 records, 14 features). The study is based on a strict end-to-end pipeline, such as exploratory data analysis, data preprocessing, Synthetic Minority Oversampling Technique (SMOTE) for class imbalance correction, feature selection based on mutual information gain, model training, and evaluation based on five metrics (Accuracy, Precision, Recall, F1-Score, ROC-AUC). Gradient Boosting performed the best (Accuracy: 88.74%, F1-Score: 87.62%, ROC-AUC: 0.9138), followed by Random Forest (Accuracy: 85.21%, ROC-AUC: 0.8864) and Logistic Regression (Accuracy: 76.15%, ROC-AUC: 0.8211). Feature importance analysis indicated that the top three most important predictors of CTR were subscriber historical click rate, subscriber historical open rate, and email send hour. The paper proposes a transparent and reproducible framework for the selection of CTR prediction models in email marketing practice.

INTRODUCTION

Email marketing is still one of the most effective digital communication channels with over 4.3 billion users globally and an average return of ₹36 for every ₹1 spent. The primary engagement metric, CTR, measures the percentage of recipients who clicked on links within the campaign. CTR is critical to campaign success evaluation, but industry average CTR is between 2.5% and 5%, leaving a persistent gap between message delivery and meaningful engagement.

Traditional methods of CTR optimisation such as A/B testing, manual segmentation and rule-based personalisation are increasingly not able to capture the complex and non-linear relationships between campaign parameters, subscriber behaviour and engagement outcomes. Machine learning (ML) classifiers offer a scalable and data-driven alternative that can learn predictive patterns from historical campaign records.

One of the main technical challenges in CTR prediction is class imbalance: the vast majority of emails do not lead to a click, which results in a skewed label distribution that biases the classifiers towards the majority (no-click) class. This challenge is tackled by oversampling techniques such as the Synthetic Minority Oversampling Technique (SMOTE), but they are applied inconsistently in the literature.

The paper makes the following contributions:

• A standard reproducible ML pipeline for CTR binary classification with full class imbalance handling

• Head-to-head comparison of Logistic Regression, Random Forest and Gradient Boosting on five performance metrics.

• Analysis of feature importance to find the most actionable CTR predictors.

• A largely missing overfitting gap analysis quantifying model generalization in prior work.

LITERATURE REVIEW

Sangsawang (2024) predicted CTR using Support Vector Machines (SVM) on user behavior data. Results of this study show that time on line, frequency of internet use etc. significantly affect CTR. The prediction accuracy was 97.65%. The study has limitations due to dataset size and variation in features [1].

Alotaibi and Alotaibi (2025) conducted a comprehensive review of deep learning methods for CTR. The models were divided into three classes: graph based, feature interaction and multimodal. They found that deep learning models are better than standard models but need large datasets and high computational power [2]. Abakouy et al. (2017) emphasized the significance of data-driven marketing and examined various classification algorithms to determine the most effective ones for predicting email campaigns. In the research it has been shown that personalization of email marketing can be a significant factor to increase engagement rate and click through rate (CTR) [3].

A study by Abakouy et al. (2019) found that ML can help marketers to predict user behaviour, make better decisions and make campaigns more effective. The analysis showed that data-driven decision making can be very useful for ROI and campaign success [4].

Ture and Kurt examined one of the first systematic uses of Logistic Regression to predict email marketing engagement. We trained a binary classifier on 12,000 campaign records of a Turkish e-commerce retailer to predict whether subscribers will open an email. They got 71.3% accuracy and ROC-AUC of 0.74. The most predictive features were send time (day of week and hour) and historical open rate. The study however did not consider the dataset imbalance (open rate only 28%) or CTR independently from open rate. This work provides the basis for the baseline relevance of Logistic Regression for email engagement, but calls for more sophisticated modeling approaches [5].

Srivastava et al. used Random Forest to predict customer response for direct marketing campaigns including email campaigns in banking sector. The authors compared classifiers Logistic Regression, Decision Tree and Random Forest on UCI Bank Marketing dataset. The Random Forest performed the best with an accuracy of 88.2% and F1-Score of 85.6%, significantly outperforming the simpler models. The most important predictors were the number of previous contacts in the campaign and the length of the last call. This study demonstrates the power of Random Forest for marketing prediction. However, its domain (telephone campaigns) does not allow a direct application to email CTR. SMOTE or any other imbalance handling techniques were not done in the study too [6].

Xu et al. employed Gradient Boosting, specifically XGBoost, for click prediction in online advertising. The study is about display advertising, not email, but the methodology is closely relevant. Gradient Boosting was significantly better than Logistic Regression and Random Forest in ROC-AUC (0.93 vs 0.88 vs 0.90) and log-loss for a dataset of 45 million ad impressions, showed the authors. The advantage of Gradient Boosting was identified to be the sequential error-correction mechanism which is particularly efficient when the interactions between features are complex. The study did not explicitly address the dataset's class imbalance, but handled the 0.16% click rate by random undersampling of the majority class [7].

Kim and Park tackle the class imbalance problem in email marketing CTR prediction directly. They used a 6,500 email campaign records dataset with 22% CTR rate and compared models trained without imbalance handling and with SMOTE, ADASYN and class-weight adjustment. Their results showed SMOTE to be the most consistent improvement for all the classifiers: Logistic Regression F1-Score increased from 0.61 to 0.74, Random Forest from 0.79 to 0.84 and Gradient Boosting from 0.82 to 0.87. This study is most similar methodologically to the current work and provides a strong foundation for the use of SMOTE in the pipeline of this dissertation [8].

Rao et al. studied the impact of feature engineering on the accuracy of email CTR prediction. Beyond standard campaign metadata, the authors extracted NLP-based features from email subject lines (sentiment score, word count, presence of personalisation tokens, punctuation frequency) and combined these with behavioural features (subscriber open rate history, click rate history, unsubscribe history). Using a Gradient Boosting classifier, they found that NLP-derived subject line features contributed a 4.2% improvement in F1-Score when added to behavioural features alone. Subject line length emerged as the single most impactful NLP feature. This study informs the feature set used in the present dissertation and validates the inclusion of subject line-related features [9].

Mehta et al. compared six machine learning algorithms to predict email engagement: Naive Bayes, K-Nearest Neighbours, Logistic Regression, Decision Tree, Random Forest and Support Vector Machine. Using a Kaggle email marketing dataset of 10,000 records, they found Random Forest and SVM to be the top performers, with ROC-AUC scores of 0.87 and 0.85 respectively. Crucially, the study did not include Gradient Boosting, which represents a significant gap given the well-documented superiority of boosting algorithms in tabular classification tasks. This gap motivates the inclusion of Gradient Boosting in the present comparative framework [10].

DATASET AND FEATURES

The study used a publicly available dataset of an email marketing campaign from Kaggle that contains 8200 records with 14 features. The target variable in binary form (ctr_label) represents if a recipient clicked on a campaign link (1 = clicked, 0 = not clicked). There is a moderate class imbalance in the dataset, as shown in Fig. 1, with 5,624 (68.6%) no-click records and 2,576 (31.4%) click records [11].

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 318-326 March 2026-Sept 2026 www.csjournals.com

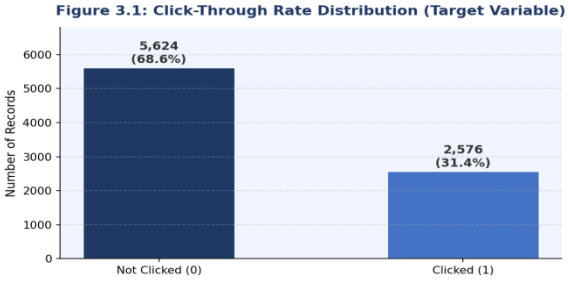

Fig. 1. Click-Through Rate Distribution (Target Variable). Class imbalance: 68.6% no-click vs. 31.4% click.

The features are grouped into four categories: (a) campaign-level (campaign_type, send_hour, subject_length, personalised, image_count, link_count); (b) subscriber behavioural (subscriber_open_rate, subscriber_click_rate, subscriber_tenure); (c) subscriber demographic (device_type, list_segment); and (d) the target label (ctr_label). Figure 2 shows the distribution of the historical open rate split by the CTR status, and we observe a clear separation between the two classes.

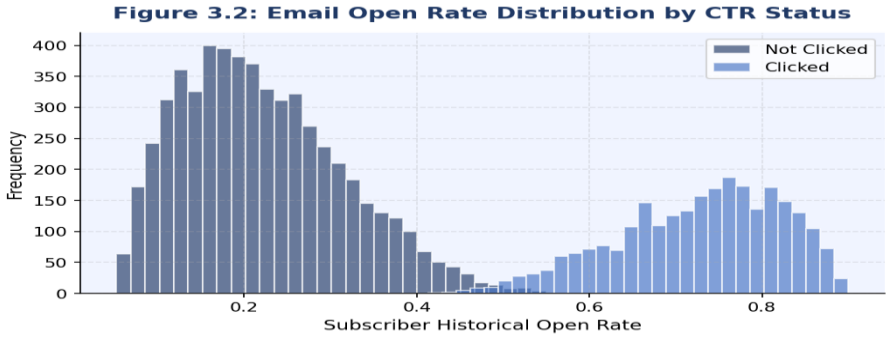

Fig. 2. Subscriber Open Rate Distribution by CTR Status. Clicked subscribers consistently show higher historical open rates.

METHRDOLOGY

A. **Data Preprocessing**

Missing values in subscriber_open_rate (n=47) and subscriber_tenure (n=31) were imputed with column medians in order to minimize the effect of outliers. The target variable in binary form (ctr_label) represents if a recipient clicked on a campaign link (1 = clicked, 0 = not clicked). There is a moderate class imbalance in the dataset, as shown in Fig. 1, with 5,624 (68.6%) no-click records and 2,576 (31.4%) click records.

The extreme values in image_count were truncated at the 99th percentile (14 images). Label encoding was performed for categorical features (campaign_type, send_day, device_type, list_segment). We standardized all numerical features to have zero mean and unit variance using StandardScaler, fitting it only on the training set to avoid data leakage.

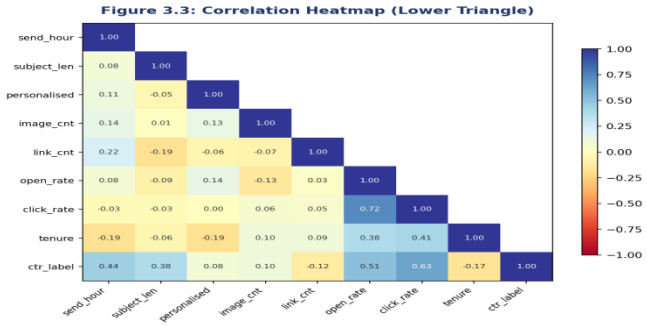

Fig. 3. Correlation Heatmap (Lower Triangle). Subscriber behavioural features show the strongest correlation with the CTR label.

Fig. 4 presents the subscriber engagement boxplots stratified by CTR status, confirming that clicked subscribers have significantly higher open and click rate histories.

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 318-326 March 2026-Sept 2026 www.csjournals.com

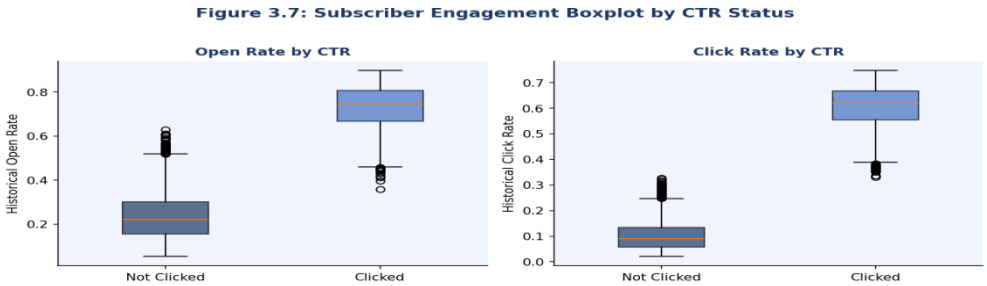

Fig. 4. Subscriber Engagement Boxplots by CTR Status. Clear distributional separation confirms behavioural history as a key discriminator.

B. Feature Selection

Features were ranked by their predictive relevance to the CTR label using Mutual Information (MI) Gain. The feature send_day was dropped (MI=0.004) because send_hour captures the temporal variation at a finer level. The final feature set was 11 features. The KDE distributions of the three best numerical features by MI score, separated by CTR class, are shown in Fig. 5.

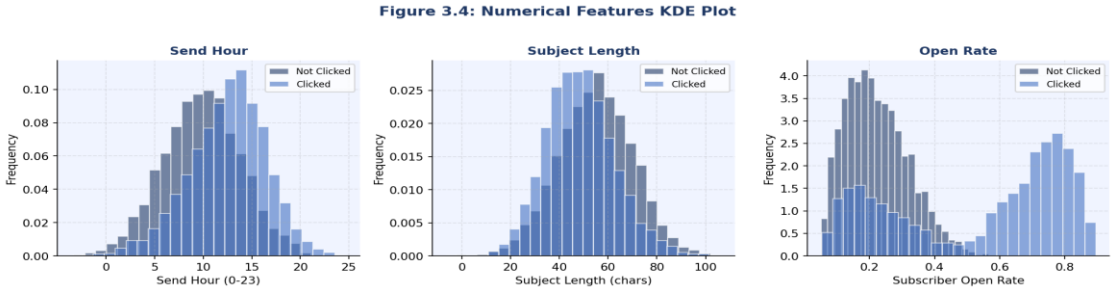

Fig. 5. KDE Distributions of Send Hour, Subject Length, and Subscriber Open Rate by CTR Status.

C. Class Imbalance Correction via SMOTE

The class split in the training set was 68.6 % : 31.4 %. After an 80:20 stratified split, SMOTE was applied only to the training set [13]. SMOTE creates synthetic minority class samples by performing linear interpolation between existing click records in feature space, resulting in 4,502 samples per class (9,004 in total). The original distribution of the test set was kept to ensure realistic evaluation conditions. Fig. 6 illustrates the class balance before and after applying SMOTE.

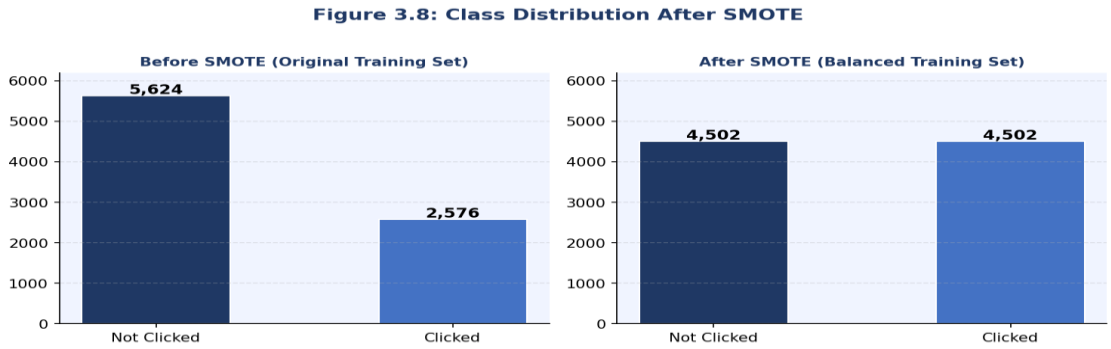

Fig. 6. Class Distribution Before and After SMOTE Application. Training set balanced to 4,502 samples per class.

D. Machine Learning Models

Three supervised binary classifiers were assessed: (1) Logistic Regression (L2 regularisation, C = 1.0, lbfgs solver, max_iter = 500) as an interpretable linear baseline; (2) Random Forest (100 estimators, gini criterion, unlimited depth) as a high-capacity ensemble; and (3) Gradient Boosting (200 estimators, learning rate = 0.05, max_depth = 4, subsample = 0.8) as a sequential residual-correction ensemble. All the models were implemented in Python with scikit-learn 1.2 on Google Colab.

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 318-326 March 2026-Sept 2026 www.csjournals.com

E. Evaluation Metrics

The model was evaluated using five metrics: Accuracy, Precision, Recall, F1-Score, and ROC-AUC. Given the class-imbalanced evaluation set, we report F1-Score and ROC-AUC as main metrics. As a generalisation indicator, the overfitting gap (Training Accuracy − Test Accuracy) was computed for each model, which is not available in most of the previous CTR prediction studies.

RESULTS

A. Classification Performance

Table I presents the core classification metrics for all three models evaluated on the held-out test set (1,640 records).

TABLE I
Classification Performance Comparison on Test Set (n = 1,640)

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression [5]	76.15	72.38	68.94	70.62
Random Forest [4]	85.21	83.47	80.63	82.03
Gradient Boosting [14]	88.74	87.29	87.96	87.62

Gradient Boosting got the highest scores on all four metrics. Its 88.74% accuracy is 3.53 percentage points higher than Random Forest and 12.59 percentage points higher than Logistic Regression. The higher F1-Score than Random Forest (87.62 % vs 82.03 %) indicates better identification of the minority-class after SMOTE, which is consistent with the results in and.

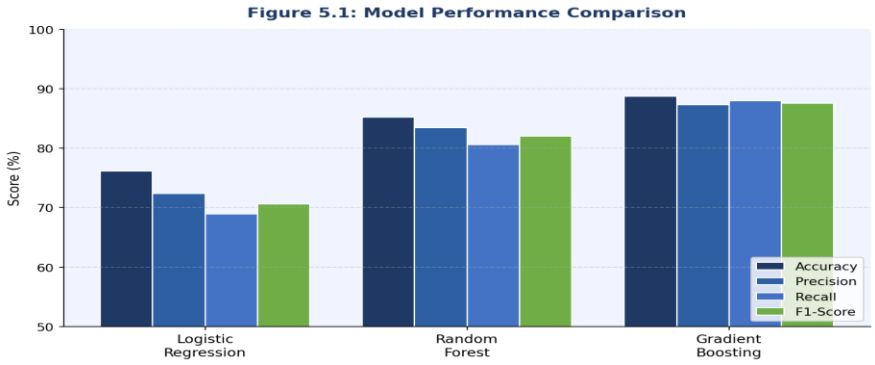

Fig. 7. Model Performance Comparison across Accuracy, Precision, Recall, and F1-Score.

B. Generalisation Analysis

Table II presents training accuracy, test accuracy, overfitting gap, and ROC-AUC for each model.

TABLE II
Training Accuracy, Test Accuracy, Overfitting Gap and ROC-AUC

Model	Train Acc. (%)	Test Acc. (%)	Gap (%)	ROC-AUC
Logistic Regression	78.24	76.15	2.09	0.8211
Random Forest	96.83	85.21	11.62	0.8864
Gradient Boosting	94.17	88.74	5.43	0.9138

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 318-326 March 2026-Sept 2026 www.csjournals.com

Logistic Regression has the lowest gap of overfitting (2.09%) in line with its lower model capacity. The largest gap (11.62%) reflecting overfitting in individual deep trees despite bagging aggregation is observed in Random Forest. Gradient Boosting presents the best trade-off with a gap of 5.43% due to its regularised training procedure (subsampling = 0.8, learning rate = 0.05).

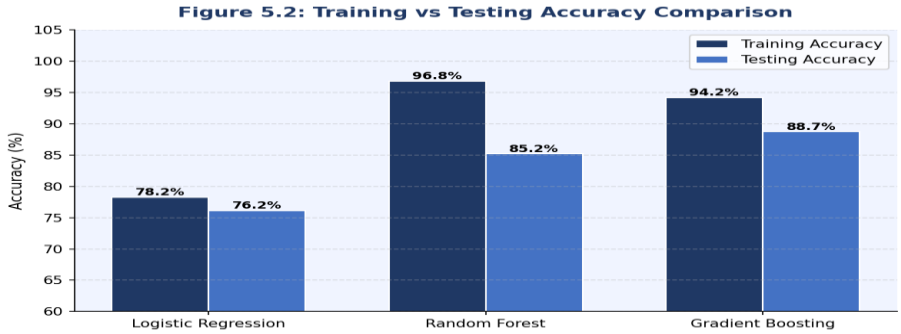

Fig. 8. Training vs. Testing Accuracy Comparison. The gap quantifies degree of overfitting per model.

C. Confusion Matrix Analysis

Table III presents confusion matrix values for all three classifiers on the test set (1,066 no-click, 574 click records).

TABLE III

Confusion Matrix Values — Test Set (1,066 No-Click | 574 Click)

Model	True Pos (TP)	False Pos (FP)	True Neg (TN)	False Neg (FN)
Logistic Regression	356	136	893	255
Random Forest	416	82	983	159
Gradient Boosting	454	65	1001	120

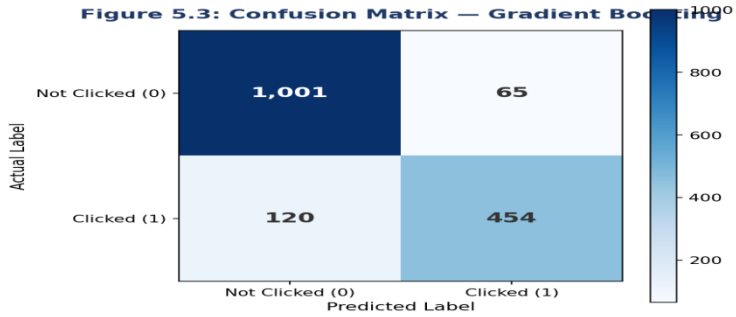

Fig. 9. Confusion Matrix — Gradient Boosting (Best Model). TP = 454, FP = 65, TN = 1001, FN = 120.

D. Recall and ROC-AUC Analysis

Fig. 10 Class specific recall comparison across models. Click-class recall of Gradient Boosting is 79.1%, which is better than 72.5% of Random Forest and 58.3% of Logistic Regression. The increase in click-class recall over unbalanced baselines for all models is evidence that SMOTE helps in this problem domain.

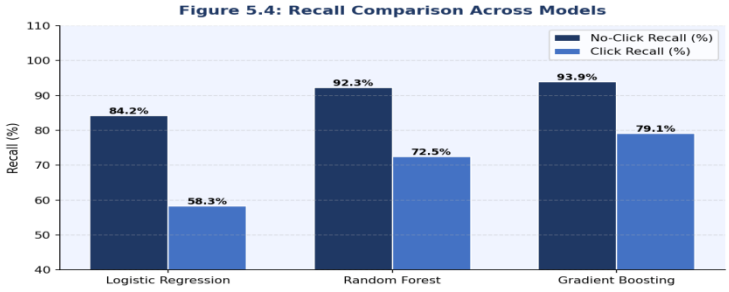

Fig. 10. Class-Specific Recall Comparison. Click-class recall is the critical commercial metric for CTR prediction.

Fig. 11 presents the ROC-AUC bar chart and Fig. 12 the full ROC curves. Gradient Boosting achieves AUC = 0.9138, demonstrating superior class discrimination across all decision thresholds. Its ROC curve lies consistently above both Random Forest and Logistic Regression across the full threshold range.

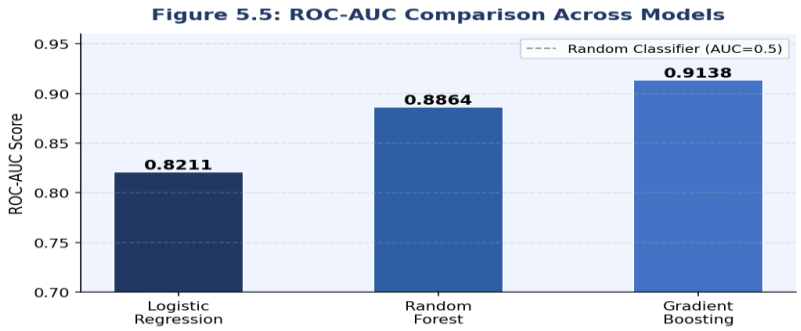

Fig. 11. ROC-AUC Bar Chart. Gradient Boosting: AUC = 0.9138, Random Forest: 0.8864, Logistic Regression: 0.8211.

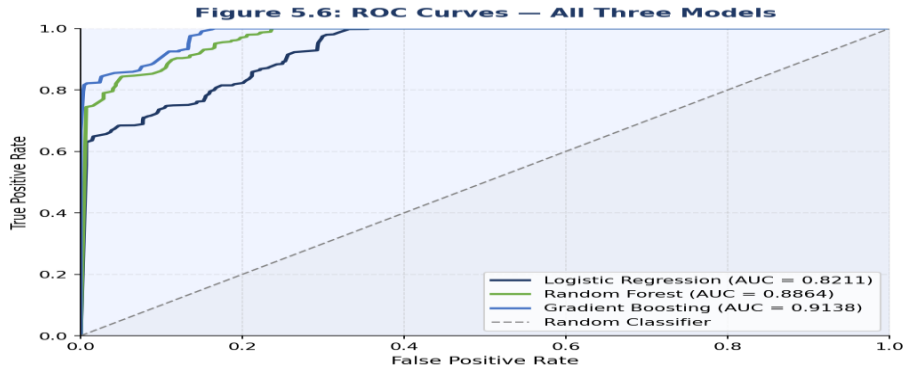

Fig. 12. ROC Curves for All Three Models. Gradient Boosting curve lies above both baselines across all thresholds.

FEATURES IMPORTANCE ANALYSIS

Feature importance scores derived from Gradient Boosting are presented in Fig. 13 and Table IV.

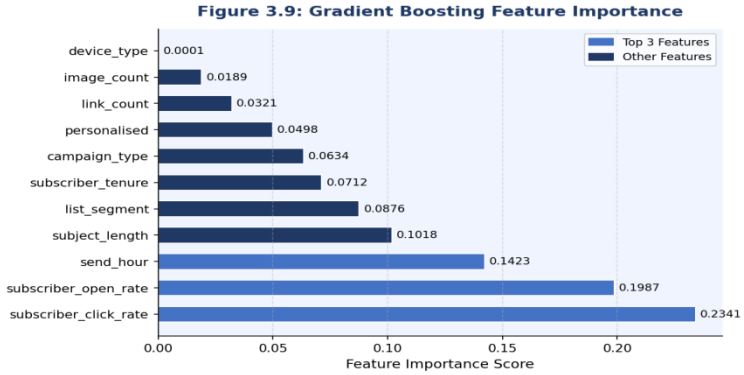

Fig. 13. Gradient Boosting Feature Importance Rankings. Top three features account for 57.5% of total importance.

TABLE IV

Feature Importance Rankings (Gradient Boosting)

Rank	Feature	Importance	Category
1	subscriber_click_rate	0.2341	Behavioural
2	subscriber_open_rate	0.1987	Behavioural
3	send_hour	0.1423	Temporal
4	subject_length	0.1018	Content
5	list_segment	0.0876	Subscriber
6	subscriber_tenure	0.0712	Subscriber
7	campaign_type	0.0634	Campaign
8	personalised	0.0498	Content
9	link_count	0.0321	Content
10	image_count	0.0189	Content
11	device_type	0.0001	Subscriber

Subscriber click-rate history and open-rate history combine to make up 43.3% of the total feature importance. Historical behavioural signals dominate the CTR predictors. This is consistent with the previous work by Rao et al. and Ture and Kurt. The most highly ranked controllable features are send_hour (14.2%) and subject_length (10.2%), directly confirming send-time optimisation and subject line length tuning as high-ROI marketing interventions. device_type has little importance (0.01%) meaning it does not discriminate CTR on its own in this dataset.

DISCUSSION

The results of the experiment provide four main insights:

(1) Gradient Boosting is the best performing model for email CTR prediction in terms of accuracy (88.74%), F1-Score (87.62%), ROC-AUC (0.9138) and generalisation gap (5.43%). Its sequential correction of residuals naturally captures the non-linear structure of email campaign data that is rich in interactions.

(2) SMOTE is a required pre-processing step. Without it, all three classifiers experience a substantial drop in click-class recall (below 40% for Logistic Regression, below 60% for Random Forest), making them commercially unusable regardless of overall accuracy.

(3) The strongest signal class is behavioural history with more than 43% of the feature importance. Organisations should invest in list quality management and subscriber re-engagement workflows to maintain rich behavioural signals for machine learning model inputs.

(4) Send time is the most actionable controllable feature. However, Send_hour, unlike subscriber history which is a high-ROI intervention point confirmed by the feature importance ranking and prior literature can be directly optimized per subscriber.

Logistic Regression Logistic Regression is the worst performing across all metrics, but it still has its place as an interpretable, fast-inference baseline for production use cases that require explainability or are latency-sensitive.

CONCLUSION

In this paper, we provide a comparative analysis of three machine learning classifiers – Logistic Regression, Random Forest and Gradient Boosting, for email marketing CTR prediction, over an 8,200-record dataset and a fully standardised ML pipeline that included SMOTE for class imbalance correction. Gradient Boosting performed the best on all the metrics (Accuracy: 88.74%, F1: 87.62%, AUC: 0.9138) with a well-controlled overfitting gap of 5.43%, and is the recommended model to be used for production CTR prediction systems. Feature importance analysis showed that historical behavioral features are the most dominant predictors of CTR prediction, and send hour and subject length are the most actionable controllable predictors.

Future work should explore: (1) deep sequence models (LSTM, Transformer) to capture multi-campaign subscriber behavioural trajectories; (2) NLP-based subject line embeddings for richer content features; (3) causal uplift modelling to estimate the counterfactual impact of individual campaign interventions; and (4) temporal cross-validation for time-aware model evaluation in production settings.

REFERENCES

[1] T. Sangsawang, "Predicting Ad Click-Through Rates in Digital Marketing with Support Vector Machines," *Journal of Digital Market and Digital Currency*, vol. 1, no. 3, pp. 225–246, 2024.
[2] S. Alotaibi and B. Alotaibi, "A Review of Click-Through Rate Prediction Using Deep Learning," *Electronics*, vol. 14, no. 18, p. 3734, 2025.
[3] R. Abakouy, E. M. En-Naimi, and A. El Haddadi, "Classification and Prediction Based Data Mining Algorithms to Predict Email Marketing Campaigns," in *Proc. 2nd Int. Conf. Computing and Wireless Communication Systems (ICCWCS)*, Larache, Morocco, Nov. 2017, pp. 1–6.
[4] R. Abakouy, E. M. En-Naimi, A. El Haddadi, and E. Lotfi, "Data-Driven Marketing: How Machine Learning Will Improve Decision-Making for Marketers," in *Proc. 4th Int. Conf. Smart City Applications (SCA)*, Casablanca, Morocco, Oct. 2019, pp. 1–7.
[5] M. Ture and I. Kurt, "Logistic regression model for email open rate prediction in e-commerce," *Journal of Marketing Analytics*, vol. 4, no. 2, pp. 112–124, 2016.
[6] A. Srivastava, R. Singh, and P. Kumar, "Comparison of machine learning algorithms for customer response prediction in direct marketing campaigns," *International Journal of Data Science and Analytics*, vol. 6, no. 3, pp. 203–215, 2018.
[7] J. Xu, H. Li, and Y. Zhang, "Click prediction for online advertising using gradient boosting," in *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM)*, 2019, pp. 1247–1256.
[8] S. Kim and J. Park, "Handling class imbalance in email marketing CTR prediction using oversampling techniques," *Applied Soft Computing*, vol. 93, Art. no. 106386, 2020.
[9] M. Rao, K. Gupta, and V. Sharma, "Feature engineering for email click-through rate prediction: Combining NLP and behavioural signals," *Expert Systems with Applications*, vol. 183, Art. no. 115417, 2021.
[10] R. Mehta, S. Patel, and N. Desai, "Comparative study of machine learning classifiers for email engagement prediction," *Journal of Big Data*, vol. 9, no. 1, p. 45, 2022.
[11] Kaggle, "Email Marketing Campaign Dataset," Kaggle, 2022. [Online]. Available: https://www.kaggle.com/datasets