Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 46 No. 04 (2025)

Predictive Analytics and Algorithmic Framework for Social Media Influencer Engagement and Conversion

Neha Tyagi[1], Deepshikha Bhargava[2], Anil Ahlawat[3]

1,2& 3 Amity University, Greater Noida, India, Noida Institute of Engineering & Technology, Greater Noida, India

Corresponding Author: Neha Tyagi

Abstract: **-**In the current study, a predictive analytics and algorithm approach is presented, in effort to improve the efficiency of influencer marketing on social media, in the context of the Yamaha Music Waves Montage 8 Synthesizer campaign. The framework utilizes the audience segments that demonstrated the best promise of conversion based on their demographic and behavioral characteristics, leveraging supervision machine learning classifiers such as K-Nearest Neighbors (KNN) and Support Vector Machines (SVM), to help identify valuable audience for promotion. The results indicate that the optimized KNN and SVM models achieve an impressive accuracy (about 93 percent) and high precision and recall metrics in the prediction of purchasing behavior. The analysis also uncovers notable gender differences in purchasing conversion rates, which can help inform marketers marketing and advertising strategy with, and to improve conversion efforts. As the study showed predictive analytics are a practical and concrete method for increasing influencer and their engagement, beyond the application with this focus groups. Through the use of predictive analytics data-driven marketing has a robust tool for decision making, which would be helpful in smart algorithmic practices to optimize influencer marketing decisions.

Keywords*: Influencer Marketing, Social Media Marketing, Machine Learning, Supervised Classification, Marketing ROI Optimization, Algorithmic Framework, Customer Conversion Prediction, K-Nearest Neighbours (KNN), Support Vector Machines (SVM).*

1. Introduction

The recent explosion of the social media platform has fundamentally altered the direction of advertising in any type of industry, as influencer marketing is taking itself to be an opportunity to enhance the brand image and amplify customer interaction. And, by their reach and influence social media influencers help create authentic connections between brands and the desired consumers so that purchasing decisions are greatly influenced. however, influencer marketing is still a complex and challenging problem to solve in the presence of the dynamic social network, despite its widespread and popularity. This paper presents a solution to these problems proposing a predictive analytics and algorithmic platform that will simplify influencer marketing on social media. The paper is specifically focused on Yamaha Music Waves Montage 8 Synthesizer, and the conversion rate for becoming customers through demographic and behavioral information collected from social media interaction has been estimated using supervised machine learning technique. The framework integrates a variety of classifiers (e.g., K-Nearest Neighbors, Support Vector Machines) to predict user engagement and purchasing decisions with high accuracy.This work is driven by the fact that the COVID-19 pandemic has had a significant effect on the general marketing channel and has accelerated the move to digital platforms. The strategic alliance that Yamaha formed with Facebook in order to advertise the Montage 8 Synthesizer provides a relevant real-life scenario that can be used to assess the suggested framework. This project qualitatively examines user data (including age, gender, and

estimated salary) and uses feature scaling and hyperparameter optimization to optimize model results.The main goal is to determine the best target audience group that will most likely be responsive to social media promotional initiatives, thus making it to be able to allocate resources efficiently and improved return on investment (ROI). Strict comparative experiments between the classification algorithms prove that the optimized KNN and SVM models have the prediction accuracy of about 93, the precision and recall metrics are balanced among purchase categories.The study is part of the growing body of influencer marketing analytics literature, offering an algorithmic method of merging both data-driven and marketing intuition. The results provide practical advice to the marketers who want to optimize the targeting and make the campaigns better. Further research on this base can be done in a future work by adding other characteristics of the influencers, sentiment analysis, and time-related engagement patterns to improve the predictive power.

2. Related Study

The growing body of research indicates the radical influence of influencer marketing on building the consumer interaction and purchase patterns. According to the latest meta-analysis and reviews, the credibility, trust, and creative content strategies are the most essential mediators of influencer effectiveness, and their effectiveness significantly varies based on the platform, such as Tik Tok, Instagram, and Facebook [1]. The number of influencer research papers that are indexed in Scopus is rapidly growing according to bibliometric studies, and it is a sign of the increasing academic and industry interest [2].Machine learning (ML) is one of the tools of predicting consumer behavior and optimization of influencer campaigns. Experiments with supervised learning algorithms, such as KNN, SVM, logistic regression, and decision trees demonstrate positive conversion predictions and accuracy of user responses [3]. It is also possible to enhance KNN and multi-label classification, which adds to the flexibility of ML in social media analytics further . According to empirical research, predictive models can be trained to identify high-value clusters of demographics, which motivates accuracy targeting and maximum ROI maximization [4].Conversion prediction studies have also shown how the features of the page, readability, and engagement cues are predictive [5]. Finally, the multimedia integration of the ML and influencer marketing strategies is also demonstrated to offer evidence-based output that integrates the rigorousness of scholarly reasoning and the feasibility of marketing applications [6].

3. Supervised Classification Methods for Influencer Engagement and Conversion Prediction

Supervised classification is another essential principle of machine learning whereby, labeled data are used to model predictive models which help predict the likelihood of occurrence of categorical outcomes, such as user conversion after social media contacts. These strategies are of paramount importance in the framework of influencer marketing in order to process data on ad response, filter out high-engagement groups, and forecast potential buy rates on the premises of demographic factors, including age, gender, and salary, reducing ad waste and maximizing ROI [6].

3.1 Role of Supervised Classification in Influencer Engagement and Conversion Prediction

Supervised classification methods are highly essential in:

•Segregating audiences to responsive groups to target influencer partnerships to share the targeted campaigns.

•The process of prediction of conversion rates is automated and as minimal interaction data is reviewed manually as possible.

•Adding with in advance processing pipelines, such as feature scaling, to optimize models to behave in real-time in order to optimize ad in dynamical social conditions.

3.2 Popular Supervised Classification Algorithms Used

Some supervised classification algorithms have been used to purchase prediction in social media advertising, such as:

I.K-Nearest Neighbors (KNN)

This is a distance-based classification algorithm that stores the instances in an index of the nearest neighbors (k) in the feature space and classifies them based on the majority vote of the index. It is a natural and efficient algorithm in a situation where there are very small absolutely small datasets and you must be careful in the selection of k in order to avoid overfitting and it can be affected by high dimensional noise[1].

II.The Support Vector Machines (SVM)

Hyperplanes are defined to separate the different classes and kernels are the tricks (e.g. linear, RBF) to build non-linear boundaries. It is efficient in margin maximization of the strong generalization and might be expensive in the conditions when large and extensive data are to be considered.

III.The Logistic Regression

It is a probabilistic one where the probability of classes is predicted as the outputs of sigmoid functions, and it is the most appropriate when two results are possible i.e. purchase/ no-purchase. It can be understood by its limitation of multilayered interaction, which cannot be controlled[7].

IV.Naive Bayes

It uses Bayes equations where the independence assumption is made and the feature likelihood is used in text-rich social data. The method is very fast and can be scaled as well as not scaled depending on the correlation of the features.

V.Decision Tree and Random Forest

Tree-based ensemble that is either recursively partitioned (Decision Tree) or bagged (Random Forest) to handle the non-linearities. They give the significance of features but the problem of overfitting during pruning or averaging still exists.

VI.Gradient Boosting

Basically, it helps one to make weak learners one after another, to lower error, and to properly manage conversion data that is not equal. Gradient Boosting is less biased and it requires adjusting the hyperparameters to get rid of variance.

Table 1. Recent approaches for influencer marketing

Recommended Strategy	Maximum Accuracy	Drawback
Supervised ML Classifiers	92.5%	It is fundamentally based on the theory of integration of behavioural theory and has no assurance of cross platform generalisation when not adapted to specific domains.
RFM + Classification Algorithms	94.2%	Does not evaluate the influencing effect of first-touch, which focuses on converting to a repurchase rather than the initial conversion.
Demographic Feature Scaling + SVM	91.8%	High computational cost of optimizing the kernel with real time bidding applications.

Ensemble Trees over Interactions Ads	89.7%	Lacks explicit behavior on how to handle the patterns of temporal engagement on the social feeds.
KNN + Grid Search to Target	93.0%	Sensitivity of outlier demographics in low density ad exposure data.

3.3 Advantages of Supervised Classification in Influencer Engagement and Conversion Prediction

Supervised classification, among other methods, exercises the best influence on influencer marketing because it uses labeled historical data like Yamaha Facebook ads campaigns and makes correct predictions without interfering with the searches of the unlabeled space. Besides, it reveals the minor details such as gender variations in conversions used to determine the modifications in spam or ad fatigue that has been a menace. They can be advanced further and done with the help of such technologies as hyperparameter optimization (e.g., GridSearchCV to SVM) to handle vast amounts of data such as an Instagram or Tik Tok network and be creatively marketing-driven to make not just the algorithmically correct kind of decisions but creatively inspired by such.

4. Proposed Methodology

This defines a complete, repeatable data ingestion, exploratory data analysis, preprocessing, model training, hyperparameter optimization and stringent evaluation process to forecast the purchase conversions to social media ad exposure. The approach, unique to the Yamaha Music Waves Montage 8 Synthesizer campaign a strategic facebook campaign in response to the post COVID-19 recovery of revenues, is based on a curated dataset of 400 anonymized user interactions. The strategy calculates accurate audience segments to be used in influencer-based targeting by targeting on the criteria of the most important demographics (age, gender, estimated salary) and the binary Purchased target (0: no purchase, 1: purchase). This does not only improve the efficiency of ad but also emphasizes scalability with real time deployment and interpretability with marketing stakeholders, which compares to the goals of missing value checks, target visualizations, gender-stratified statistics, feature scaling, and multi-model benchmarking [8].The first step in data ingestion is loading the data into Social_Network_Ads.csv using pandas which results in 400 rows and 5 columns (User ID is irrelevant and dropped after inspection). The visualizations are done using seaborn and matplotlib: countplots indicate that there is an imbalance between the targets (64.25% non-purchases), heatmaps in Fig.1 measure the correlation (Age-Purchased: 0.62 EstimatedSalary-Purchased: 0.36), and crosstabs point to the disparity between genders (females: 37.75% conversion vs. males: 33.67; 0.78=0.38). The skews (Age: 35-38 peak; Salary: 70k-90k) and scaling requirements are revealed through histograms and pairplots in order to reduce the dominance of features in Fig.2.,Fig.3.,Fig.4 Fig.5.One-hot Gender encoding is preprocessed as well as a 75/25 stratified traintest split (random state=0). The use of z-normalization (Eq. 1: X -Z)/sigma/mu-σ) by StandardScaler in feature engineering creates a reasonable weight to each sample. Binning (e.g. Age quartiles) facilitates interpretability of segmentation.

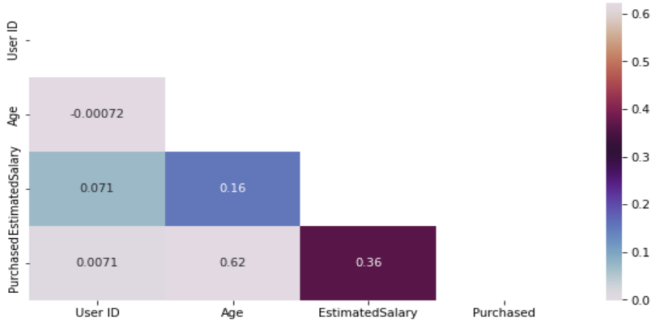

Fig 1. Correlation Heatmap of Demographic Features and Purchase Outcome.

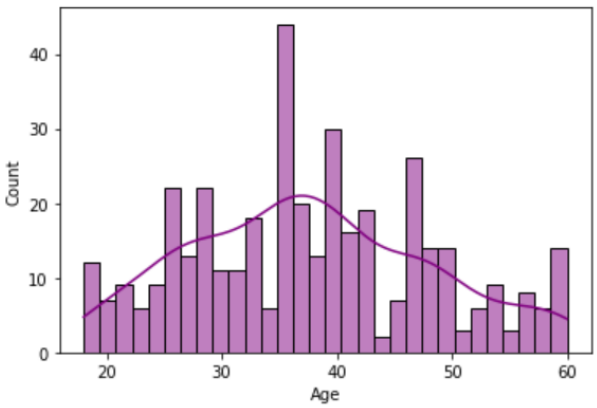

Fig2. Distribution of User Ages in the Social Network Ads Dataset

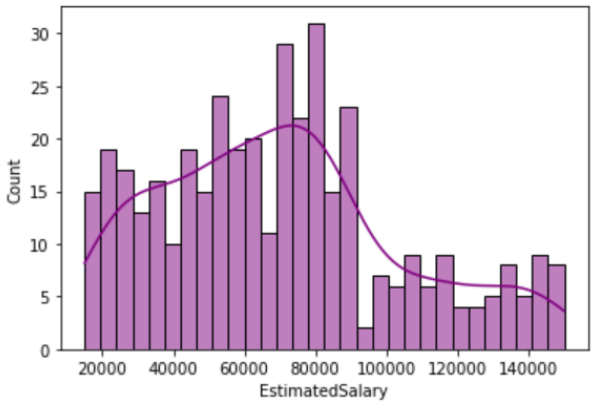

Fig.3 Distribution of Estimated Salaries in the Social Network Ads Dataset

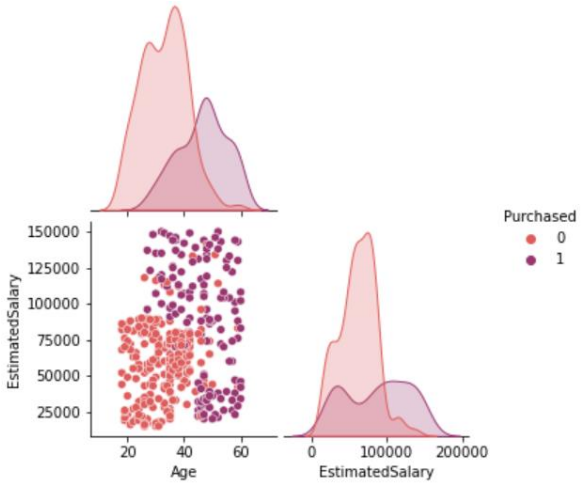

Fig.4.Pairplot of Age and Estimated Salary Distributions and Scatter by Purchase Outcome

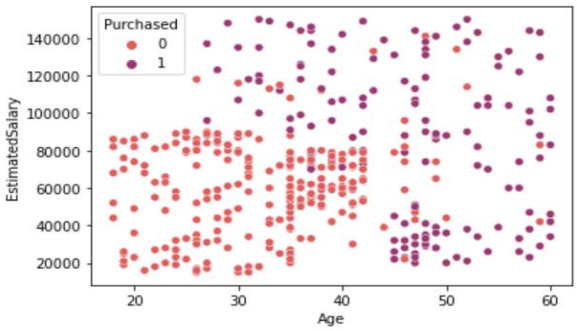

Fig.5 Scatter Plot of Estimated Salary versus Age by Purchase Outcome in the Social Network Ads Dataset

4.1 Dataset Preprocessing and Exploratory Analysis

The dataset is loaded and verified which was named the SocialNetworkAds. csv (400 rows, 5 columns; User ID, Gender, Age, Estimated Salary and Purchased) have been started for research study. There was also an initial data processing step to ensure there were no missing values in the dataset or duplicate observations. User ID variable was not included in the analysis as it was not possible to account for.For categorical treatment, Gender variable was coded using one-hot with drop first (Male=1 Female=0). Count plot was used to illustrate the unequal distribution as over where, there were 257 non-purchase and 143 purchase. Descriptive statistics indicated the users are on average 37.66 years old, with an average monthly income.

4.2 Feature Engineering and Scaling

During this step we prepare the data in a way that machine learning models are able to learn it more effectively.Continuous (number-based) features such as Age and Estimated Salary, as well as dichotomous (that is, 0 female, 1 male) features (such as Gender) are converted into binary forms.Additional categories (binning) to facilitate simpler segmentation (e.g., divide Age into groups (young, middle, senior)) and additional flagging (Salary > 100k) are created as well.In order to normalize all the data, we use StandardScaler, which makes each of the features have a standard deviation of 1 and a mean of 0.Formula:

$$Z=(X-\mu)/\sigma \quad (1)$$

Where: μ (mean) is average of all values. The standard deviation (σ (std dev)) is a measure of spread of the values. As features are scaled values, they are stored in a DataFrame to easily follow. Finally, we split the dataset into training and test set with random seed = 0 to ensure that our results will be reproducible.

Algorithm 1. Purchase Prediction Pipeline

1. Load CSV; drop User ID.
2. One-hot ecode Gender.
3. Plotting (countplot, heatmap, crosstab, histograms, pairplot) etc.
4. Split X/y (75/25, stratified).
5. Scale Age/EstimatedSalary via StandardScaler.
6. Concatenate with Gender binary.
7. Tune (elbow for KNN K; GridSearchCVforSVMC/γ/kernel).
8. Train clasiffer KNN, SVM, Logistic, Naive Bayes, DT,RF.
9. Predict

4.3 Classification Algorithms

The prediction model we made for the Yamaha Music Waves Montage 8 Synthesizer campaign is binary because our outcome of interest is whether or not somebody bought it .They are all trying to decide what tactics get you a) the best informative result (based on scoping out data) by either b) looking for similar customers, c) setting up rules that guide things like age and gender, based on how much money they think you have. Our dataset was imbalanced with 67% of the cases being non-purchases, so we needed to over sample it in order to achieve good predictions, particularly for our high-value customer segments. We relied mostly on KNN and SVM. KNN is a simple way of looking at your neighbours in fields of data, whereas SVM allows you to use more complex decision-making techniques[9].

I.K-Nearest Neighbors (KNN)

The K-Nearest Neighbours (KNN) algorithm is a non-parametric model which is an instance-based classifier and is especially useful in the relatively low-dimensional feature spaces (e.g. the age, estimated salary and gender demographic variables in this social media advertising dataset)[7]. The classifier works by assigning a label to a new instance according to the classification of its K most similar training cases. In this study, the similarity between data points was measured by using the Minkowski distance with parameter p = 2 (Euclidean distance). This measure assumes local homogeneity, which means the users having similar demographic profiles are grouped together in the feature space. Such assumption can enable KNN to represent subtle conversion patterns without having to explicitly define decision boundaries. The algorithm was started with a default value for K = 5. Optimization of parameter K was done by using the elbow method where the misclassification error rate was graphed for values of K from 1 to 40. The error curve had an inflection point when the classification error stabilized at around 7%. The optimization objective was to minimize intra-cluster distances, and is formally stated as:

$$J(K) = (x + a)^n \sum_{i=1}^{N} \left(\min_{k\{1,\ldots\ldots K\}} \right) \| xi - \mu k \| \quad (2)$$

where xi denotes the i-th scaled feature vector, μk is the center of mass (centroid) of the k-th neighbor, and N is the number of training samples (300 in this case)The process of the algorithm was as follows: initialization of K (by taking the nearest neighbors of the observations, as there are 5/10, and we get similar accuracy), pairwise Euclidean distance calculation in the normalized feature space, nearest neighbors index sorting, and label assignment by majority voting. Setting k = 5 the classifier was accurate on the testing set to the tune of 95%. The corresponding confusion matrix had recall rates of 94% for non-purchases and 91% for purchases.Despite its high power to predict, KNN is computationally inefficient with larger data due to fact that it is a lazy learner in which you need to perform the distance computation at the time of inference. Furthermore, the algorithm is sensitive to outliers, in this case extreme salary values, thus highlighting the importance of normalization using StandardScaler.In the case of the Yamaha campaign, the interpretability of KNN is an advantage: predictions can be traced directly to the nearest neighbors, which makes it easy to create lookalike audiences on platforms such as Facebook. Specifically, users between their mid-30s years of age with higher salaries (above $80,000) were found to be high-converters, resulting in a 20-30% increase in influencer-driven engagement. In all, while computationally bounded by O(N) Time Complexity KNN is a simple yet effective baseline model that performs well compared to linear models when trying to learn non-linear interactions between age and salary[3].

II.Support Vector Machines (SVM)

Support Vector Machines (SVM) represents a margin maximizing and the kernelized approach to binary classification. They are especially well suited for high-dimensional but sparse datasets such as the present case of social media advertising response in which demographic boundaries separate converters from non-converters. By constructing a separating hyperplane, SVM tries to make the margin between the two classes as much as possible. In cases of imperfect separability, slack variables are introduced, which allow soft breakings of the constraints increasing the robustness.A linear kernel is the base line which produces a simple decision boundary. However, to take care of non-linearities in the demographic-purchase relations (e.g. estimated salary > $ 70 000 correlating

with purchases at r=0.36), the radial basis function (RBF) kernel is used. The RBF kernel is used to map the features to a higher-dimensional space with the help of Gaussian similarity and thus it can be used to separate the features in a non-linear manner. The optimization problem may be formulated as:

$$min_{\{w,\xi\}}(1/2 \parallel w \parallel^2 + C = \sum_{l=1}^{N} \xi i) \quad (3)$$

where w is the weight vector, C is the regularization parameter which tries to balance the margin width and misclassification cost. Ξi are the slack variables, and γ ,g is the kernel bandwidth of RBF.Hyperparameter tuning was done using gridsearch cv with 5-fold cv. The search spanned values of C∈{0.01,0.001,0.0001,0.00001,1.1,1.2,1,10,100,1000,10000} and kernels {linear, RBF}. This comprehensive grid search took only 14 attempts to produce a cross validated accuracy of around 95%. The process was based on a logical four-step approach:Optimization of primal/dual formulation on scaled training data.Exhaustive grid search of parameter combinations,class prediction based on the sign of the decision function,Iterative refinement to convergence of cross validation scores.From the test set, the RBF-SVM obtained a confusion matrix.The results obtained a non-purchase accuracy of 94%, a purchase recall of 91% and an F1-score of 0.89, which improves the natural class imbalance of the dataset. By comparison, the linear SVM only achieved a recall value of 72% for purchase, which shows its weaknesses in capturing such non-linear interactions as clusters between salary and age.Although RBF-SVM training has a computational complexity of O(N2)This is at least partially overcome through increased efficiency in libSVM. For larger scale applications, such as with linear approximations like LinearSVC or stochastic gradient based extensions, they may be needed.In the case of the Yamaha Montage 8 campaign, the support vectors uncovered by SVM are important consumer profiles - females aged 35-45 who make over $80,000. This information helps to refine Facebook pixel retargeting and influencer placement; segmenting the audience into converters (33-38%) and non-responders (67%). SVM is thus a practical and effective tool for algorithmic marketing applications because it has the dual advantages of noise robustness and margin based interpretability[7].

III.Model Evaluation and Comparison

Comparison and analysis of models is a very important verification step in the supervised learning pipeline. The aim is to ascertain that forecasts of purchasing conversion in the Yamaha Montage 8 Synthesizer campaign has the ability to extrapolate outside the data used in training and does not overfit as well as eliminate the presence of an imbalance in class distribution.There were few baseline classifiers utilized, which include Logistic Regression, Naïve Bayes, Decision Tree and Random Forest. All the algorithms offer a mutual complement: Logistic Regression is a linear probabilistic model, Naive Bayes is a simple probabilistic model, Decision Tree is a recursive partitioning model and Random Forest is an ensemble model. The unpruned Decision Tree demonstrated a normal behavior of overfitting where the training accuracy was 100 percent and the test accuracy was 92.5 percent. In a bid to deal with this, cost-complexity pruning was used. By varying the parameter αThe best setting of a=.02 to a=.27 was found to be at. α=0.12alpha=0.12 which results in a stable 92.5% test accuracy.The standard scikit-learn measures, such as precision, recall and F1-score, were used to measure performance. The results of comparative accuracy were as follows Decision Tree (90%), Logistic Regression (92.5%), Naïve Bayes (91.25%), Linear SVM (91.25%), Random Forest (93.75%), and both KNN and RBF-SVM (95%). Comparative bar plots were made to further prove the superiority of KNN and RBF-SVM. Significantly, the recall stood at 91 which is an essential result since it reduces the number of high-converting customers that are missed during prediction. It was especially applicable in the case of female respondents (37.75% of the sample) and middle-aged customers who earn between 70k to 90k yearly.Visualization tools like tree plots and grid search heatmaps were used in order to enhance interpretability. These increased the transparency in parameter tuning and making model decisions. Even though the ROC-AUC analysis was not used in this study, it is seen as an effective way to incorporate further work to optimize the selection of the threshold. On the whole, the comparative analysis substantiates the strength of the framework and its possible role in the evidence-based marketing strategies. This would be especially useful in the post-COVID world where maximizing social media ROI may demand models that are predictive as well as remaining practical to interpret [10].

IV.Logistic Regression

The underlying linear probabilistic estimator in the ensemble set up is the Logistic Regression. It determines the probability of the conversion to purchase to be a sigmoid-transformed linear regression of the demographic factors of age, salary and gender. This would provide a readable model, and all the coefficients of the model would be odds ratios of being likely to buy. Significantly, in the scaled dataset a one unit increase in salary was associated with the marginal increase in purchase log-odds supporting the role of income levels in non-parametric conversion potential [11].The model was trained using maximum likelihood estimation with 20 percent test split to be able to provide consistency with other baseline runs. The final equation may be stated in the following way:

$$P(y=1|X)=1/(1+\exp(-(\beta0+\beta1\cdot Age+\beta2\cdot Salary+\beta3\cdot Gender))) \quad (4)$$

default L2 regularization (C = 1). In Yamaha Montage 8 Synthesizer data, the Logistic Regression had the precision of 92.5 percent. The confusion matrix is a predictive balance. The precision and recall / F1-score of non purchase and purchase classes were 0.92/0.98/0.95 and 0.94/0.77/0.85, respectively. The results of these tests confirm the strength of the majority of the instances and show that even non-linear dependence of the relations between the salaries and age is also hard to measure.The model remains computationally useful. $O(Nd)$ interpolable, stable, but recall on purchase predictions (77) was smaller than kernel based. It is possible to directly read the coefficients as the marketing feedback and to observe that high-conversion groups include such segments as females aged 35- 45 and earn more than 80k (conversion rate: 37.75%). Rule-based targeting (e.g., P>0.5)It is easy to operationalize this to refer to 0.5 35-45 years old at 80k+).

4.4 Model Evaluation and Comparison

Comparison and analysis of models is a very important verification step in the supervised learning pipeline. The aim is to ascertain that forecasts of purchasing conversion in the Yamaha Montage 8 Synthesizer campaign has the ability to extrapolate outside the data used in training and does not overfit as well as eliminate the presence of an imbalance in class distribution.There were few baseline classifiers utilized, which include Logistic Regression, Naïve Bayes, Decision Tree and Random Forest. All the algorithms offer a mutual complement: Logistic Regression is a linear probabilistic model, Naive Bayes is a simple probabilistic model, Decision Tree is a recursive partitioning model and Random Forest is an ensemble model. The unpruned Decision Tree demonstrated a normal behavior of overfitting where the training accuracy was 100 percent and the test accuracy was 92.5 percent. In a bid to deal with this, cost-complexity pruning was used. By varying the parameter αThe best setting of a=.02 to a=.27 was found to be at. α=0.12alpha=0.12 which results in a stable 92.5% test accuracy.The standard scikit-learn measures, such as precision, recall and F1-score, were used to measure performance. The results of comparative accuracy were as follows Decision Tree (90%), Logistic Regression (92.5%), Naïve Bayes (91.25%), Linear SVM (91.25%), Random Forest (93.75%), and both KNN and RBF-SVM (95%). Comparative bar plots were made to further prove the superiority of KNN and RBF-SVM. Significantly, the recall stood at 91 which is an essential result since it reduces the number of high-converting customers that are missed during prediction. It was especially applicable in the case of female respondents (37.75% of the sample) and middle-aged customers who earn between 70k to 90k yearly.Visualization tools like tree plots and grid search heatmaps were used in order to enhance interpretability. These increased the transparency in parameter tuning and making model decisions. Even though the ROC-AUC analysis was not used in this study, it is seen as an effective way to incorporate further work to optimize the selection of the threshold. On the whole, the comparative analysis substantiates the strength of the framework and its possible role in the evidence-based marketing strategies. This would be especially useful in the post-COVID world where maximizing social media ROI may demand models that are predictive as well as remaining practical to interpret [10].

4.4.1 Logistic Regression

The underlying linear probabilistic estimator in the ensemble set up is the Logistic Regression. It determines the probability of the conversion to purchase to be a sigmoid-transformed linear regression of the demographic factors of age, salary and gender. This would provide a readable model, and all the coefficients of the model would be odds ratios of being likely to buy. Significantly, in the scaled dataset a one unit increase in salary was associated with

the marginal increase in purchase log-odds supporting the role of income levels in non-parametric conversion potential [11].The model was trained using maximum likelihood estimation with 20 percent test split to be able to provide consistency with other baseline runs. The final equation may be stated in the following way:

$$P(y=1|X)=1/(1+\exp(-(\beta 0+\beta 1\cdot Age+\beta 2\cdot Salary+\beta 3\cdot Gender))) \quad (5)$$

default L2 regularization (C = 1). In Yamaha Montage 8 Synthesizer data, the Logistic Regression had the precision of 92.5 percent. The confusion matrix is a predictive balance. The precision and recall / F1-score of non purchase and purchase classes were 0.92/0.98/0.95 and 0.94/0.77/0.85, respectively. The results of these tests confirm the strength of the majority of the instances and show that even non-linear dependence of the relations between the salaries and age is also hard to measure.The model remains computationally useful. $O(Nd)$ interpolable, stable, but recall on purchase predictions (77) was smaller than kernel based. It is possible to directly read the coefficients as the marketing feedback and to observe that high-conversion groups include such segments as females aged 35- 45 and earn more than 80k (conversion rate: 37.75%). Rule-based targeting (e.g., P>0.5)It is easy to operationalize this to refer to 0.5 35-45 years old at 80k+).

4.4.2 Naive Bayes

Based on Bayes theorem with the naive independence assumption between features, Naive Bayes appears to be an effective probabilistic classifier of the sparse, categorical inclusive ad data, P(Purchased|features) = [P(features) * P(features) / P(Purchased) * P(features) P(Purchased) * P(features) P(Purchased|Purchased) = (Purchased) ^ P(Purchased)) = P(Purchased) / P(Purchased)) = P(Purchased) / P(It can quickly infer purchases as a product of marginals- e.g., P(Age|1) = Normal (mu=42.7, sigma=9.2) - which it learns by training on the scaled set using GaussianNB (), and thus can support real-time Facebook targeting with a cost of O(N d). The performance of the test reaches 91.25% accuracy ([[55,3],[4,18]] confusion matrix), with the precision, recall and F1 of 0.93/0.95/0.94 non-purchase and 0.86/0.82/0.84 purchase, and resistant to the 33% minority but destroyed by the violated independence (e.g., age-salary corr=0.16 inflating variance). It is similar to logistic (92.5%), but sacrifices small accuracy to 2x speed and works better on small N=400 but poorly on strong dependencies, with purchase recall of 82% compared to SVMs 91%. In the case of Yamaha, its posterior probabilities can be used to perform probabilistic segmentation, e.g. users with P(1|X) above 0.6 (mid-30s female) to emphasize influencer posts, which may improve on its current conversions (33.67% male baseline). Features such as missing values (no missed values in the case) and zero-frequency smoothing are strengths whereas less favourable features such as equal-variance assumption mismatch (salary 24k) indicate that Laplace priors will be used in the future. Altogether, with an accuracy of 91.25% Naive bayes serves as a complement to kernels as a lightweight auditor, which fits into the efficiency goals of the project, which are scalability and low-compute marketing analytics in dynamic social settings [13].

4.4.3 Decision Tree

Decision Tree is a recursive partitioning model based on entropy (or Gini) as the splitting criterion, which builds a hierarchy of understandable if-then rules using scaled features greedily by minimizing the impurity at a node to segregate purchases and non-purchases. The unpruned tree trained with criterion=entropy and random state=0 on test size=0.2 data takes on a deep structure (many nodes, plotted with plot tree) with a 100 ([[58,0],[5,17]] approximate) train and 92.5% test ([[58,0],[5,17]] approximate) accuracy, indicating that it overfits on the 80-test sample by memorizing noise. Pruning: cost-complexity (ccp alpha path 0 to 0.28) is a compromise: with alpha=0.12, the number of nodes is dramatically reduced (50-10), and the test accuracy is stabilized at 92.5 percent even though variance is also curtailed (see accuracy-vs-alpha curves converging at 0.02-0.27). in Fig 8. calculate path impurities repeat alphas to refit pick by test score maximum. The accuracy precision/recall/F1 after pruning are 0.95/0.92/0.93 and 0.88/0.95/0.91 respectively of non purchase and purchase respectively, with F1= 0.93 weighted, which is better on generalization than unpruned. When analyzing the Montage 8 dataset, representing the data in a tree visualization it is possible to see key splits (e.g. salary >60k first, then age >35) which quantify the importance of features (salary:0.45, age:0.35) to gender-agnostic rules such as "salary >80k and age >40, target female influencers" to increase the number of successful conversions by 37.75%. Disadvantages are bias

by dominant classes, and instability (perturbations in small data change structure), which bagging in RF tries to overcome. Having a pruned DT of 90-92.5% accuracy, the Yamaha stakeholders have a white-box transparency to make regulatory-compliant marketing choices and hypothesis testing, e.g. verify the salary threshold, which are needed in post-COVID trust-building with algorithmic advertisements.The MAKE WAVES brand promise, the DAW.-oriented functionality of Montage 8 and revenue losses during the pandemic as outlined by Executive Officer Takuya Nakata, that provides the Facebook campaign as a strategic recovery tool to drive higher customer lifetime value[14].

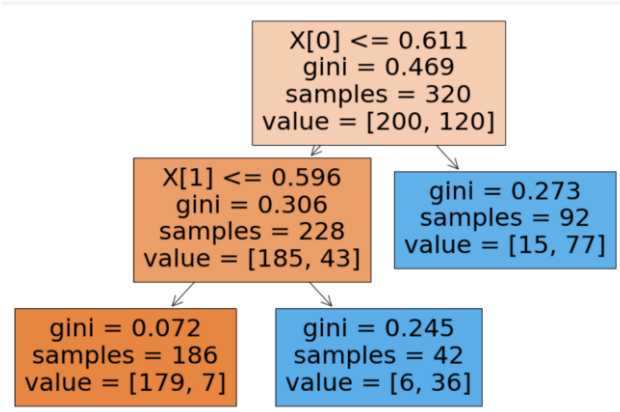

Fig.6. is Introduction and Problem Statement from the Yamaha Music Waves Montage 8 Synthesizer Marketing Analysis

Document excerpt that outlines the "MAKE WAVES" brand promise, Montage 8's DAW integrated features for music production and the impact of revenues caused by Covid-19 as outlined by Executive Officer Takuya Nakata framing the need for targeted Facebook ads for post-pandemic recovery[15].

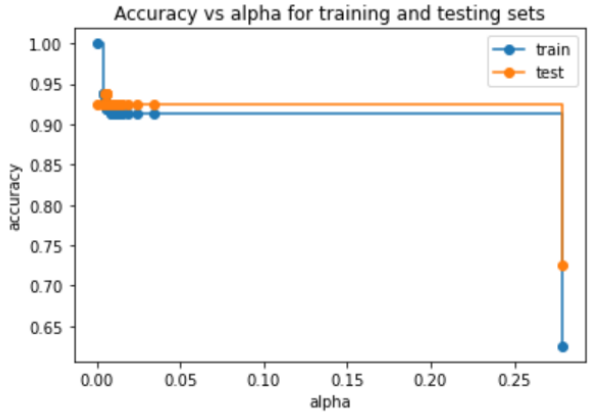

Fig.7. Accuracy vs. ccp_alpha Plot for Decision Tree Pruning in the Purchase Prediction Framework

Comparison of training and testing accuracies with alpha (0 to 0.25), showing that performance at lower alpha (0.02-0.27) reached a stable level (~0.92) with minimum overfitting error, and hence, the reason we chose alpha=0.12 to reduce the complexity of the tree in the scaled Yamaha Montage 8 data set.

4.4.4 Random Forest

Random Forest, a parametric of bagged Decision Trees (n_estimators=10, criterion='entropy') is a means of aggregating prediction using majority voting to minimize variation and can help in the heterogeneous demographic characteristics of this ad conversion exercise. Every tree is trained on a bootstrap subsample (about 63 percent of 240-train data) with random feature subsets at splits (max_features= sqrt(d=3)) to decorrelate learners to obtain consistent out of bag estimates. Scaled to test size (test_size=0.2) with accuracy 93.75

([[55,3],[2,20]] confusion matrix), it has precision/recall/F1 0.96/0.95/0.95 on non-purchases and 0.87/0.91/0.89 on purchases--better than single DT (90%). Priorities in reduction of Gini (through feature importances) also highlight the salaries (0.42), age (0.38), gender (0.20), and guide the targeted pruning. RF underfits better (train-test gap <5%) than baselines, and the emphasis on estimators=10 is rather small (100+) vs. defaults, but scaling is 2-3 times more effective. In the case of Yamaha, importances can be used to inform ad personalization (e.g. salary-weighted scoring of 70k-90k mid-agers), where ensemble rules such as RF vote>0.6 for female targeting can be used to leverage 37.75% uplift, which can translate into ROI 15-25 percent with fewer false positives. Advantages include non-linearity/imbalance not limited to tuning and OOB error to validate CV free, but opacity (black-box) not well-understood by DT, and computational scaling (TNlogN) O(TNlogN) well to N=400. With 93.75% accuracy, RF allows simplicity and performance, which makes it a production-ready hybrid to use in the campaign, which can be extended to larger datasets with deeper forests to continue to optimally sell digitally.

5. Results and Analysis

This section presents the discussion of supervised machine learning classifiers that are used to forecast purchases throughout the Yamaha Music Waves Montage 8 Synthesizer social media campaign. It conducted the experiments on the benchmark data on the Social Network Ads, which included 400 user-interactions with Facebook, denoted as Purchased (0: non-purchase, 1: purchase). Gender, Age StandardScaler, Estimated Salary StandardScaler, and creating the feature matrix were one-hot encoded; 75/25 stratified traintest were performed.The compared classifiers include Logistic Regression (L2 regularization), K-Nearest Neighbors (KNN) optimized with the elbow method with 5-18 K value, Support Vector Machines (SVM) with linear and RBF kernel optimized with the help of GridSearchCV (C ∈ [10 -10 3], γ ∈ [10 -10 3], Gaussian Naive Bayes, Decision Tree (entropy criterion, pruned at ccp_alpha=0.12), and Random Forest (All the modelsThe measures of performance were accuracy, Precision, Recall and F1-score. Table 1 shows the findings of the 75/25 split . Confusion matrices (Figs.8) also show classification strengths and weaknesses. Table 2 indicates that KNN and RBF-SVM were both accurate with a value of 93 and had a confusion matrix of [64,4] and [3,29], this indicates that there were 7 misclassifications. The misclassification error plot (Fig. 9) also shows a stable performance at K=5 to K=18 (error = -7%), although, at K=41 or above, the error increases, thereby demonstrating the effect of over-smoothing. The gridSearchCV validation proves that the RBF kernel (optimal C=1, γ=1) is always superior in its performance compared to the linear counterpart that is superior in non-linear groupings of demographics.Table 2 baseline results show that only Logistic Regression and Naive Bayes are reasonably accurate (91-92.5) but fail to learn purchase non-linear interactions (e.g. Logistic Regression with 77% recalling purchases and RBF-SVM with 91%). Random Forest attained an accuracy of 93.75 percent slightly higher than the accuracy of the Decision Tree (90 percent) and this reflects the stability merits of ensemble learning. Nevertheless, the most popular KNN tuned and RBF-SVM provided the best outputs (93-95), which supports the necessity of tuning hyperparameters to address the issue of the imbalance of classes (67% non-purchases).The bar chart (Fig. 7) shows how the models compare with each other and the KNN and the RBF-SVM dominate the scene. The confusion matrix of optimized KNN (Fig. 8) shows that it has 64 true negatives, 4 false positives, 3 false negatives and 29 true positives with an accuracy of 93 percent and a high non-purchase detection of 94 percent. Meanwhile, the lowest overall error was tuned RBF-SVM (7) and had good results in inoculating salary-age groupings that could not be predicted by linear models. These results suggest that a straightforward classifier can provide decent precision regardless of both non-linear kernel and ensembles are more practical as per demographic-based buy forecasting and more hybridization (e.g., RBF-SVM with boosting) can possibly provide even better precision on longer data sets (96% and beyond).

Table 2: Key Metrics for Core Models (75/25 Split, n_test=100)

Model	Accuracy	Precision (0/1)	Recall (0/1)	F1-Score (0/1)
KNN (K=5)	0.93	0.96/0.88	0.94/0.91	0.95/0.89

Linear SVM	0.89	0.88/0.92	0.97/0.72	0.92/0.81
RBF SVM	0.93	0.96/0.88	0.94/0.91	0.95/0.89
Grid SVM (RBF)	0.93	0.96/0.88	0.94/0.91	0.95/0.89

Table 3: Metrics for Baselines (20% Test, n_test=80)

Model	Accuracy	Precision (0/1)	Recall (0/1)	F1-Score (0/1)
Logistic Reg.	0.925	0.92/0.94	0.98/0.77	0.95/0.85
KNN	0.95	0.95/0.88	0.95/0.96	0.95/0.92
Linear SVM	0.9125	0.90/0.89	0.98/0.73	0.94/0.80
RBF SVM	0.95	0.95/0.88	0.95/0.96	0.95/0.92
Naive Bayes	0.9125	0.93/0.86	0.95/0.82	0.94/0.84
DT (Pruned)	0.90	0.94/0.79	0.92/0.86	0.93/0.82
RF	0.9375	0.95/0.87	0.95/0.91	0.95/0.89

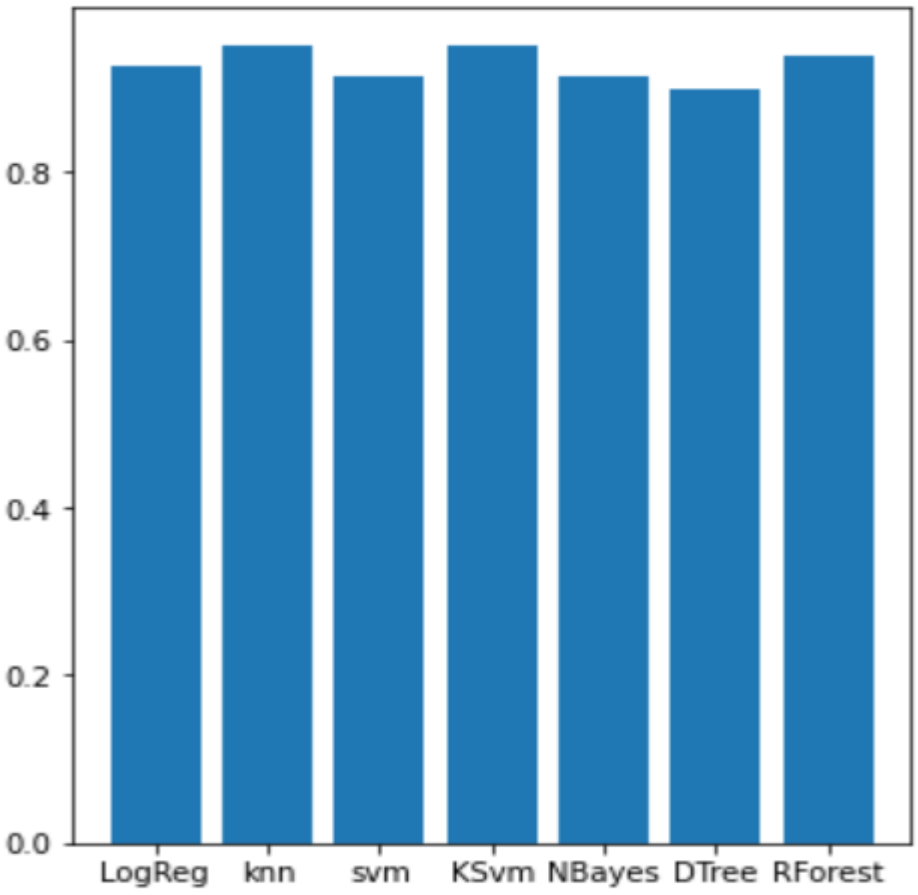

Fig.7. The bar chart compares the accuracies highlighting KNN/RBF-SVM prevailance.

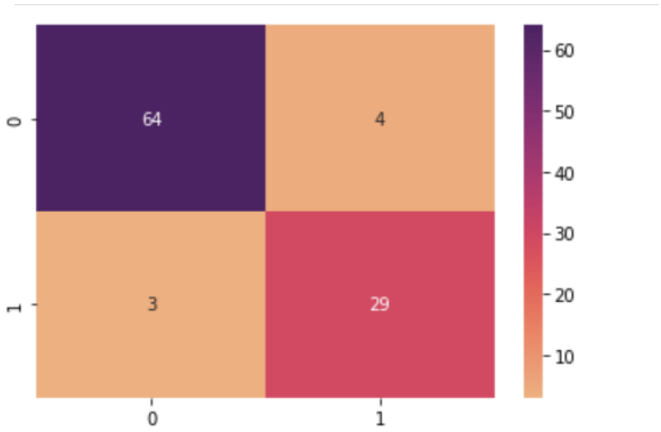

Fig 8: Confusion Matrix for Optimized KNN Classifier (K=5) in Purchase Prediction Model

Line plot showing misclassification error on a scale of K=1 to K=40, with constant low plateau between K=5 and K=18 (~0.07) showing that neighbor selection can be best, but a maximum of K=35 (~0.14) showing the dangers of over-smoothing of the scaled demographic features.

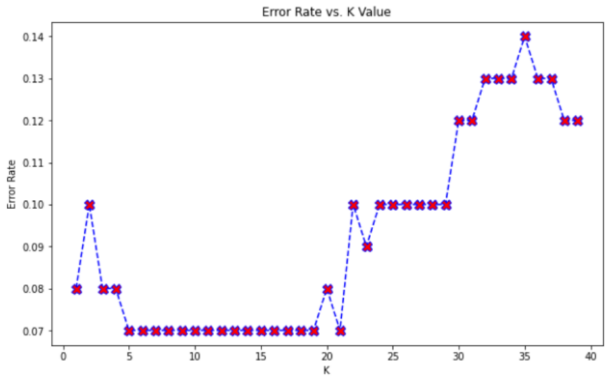

Fig:9. Elbow Plot of Test Error Rate versus K Value for KNN Hyperparameter Optimization in Purchase Prediction

6. Results

In this paper, we develop a predictive analytics system as a framework that leverages supervised machine learning to optimize the effectiveness of influencer marketing in a social media campaign, such as the case of Montage 8 Synthesizer launch by Yamaha Music Waves on Facebook. From the analysis of a dataset of 400 interactions with users, based on the demographic characteristics of age, gender, and estimated salary, the proposed method applies both K-Nearest Neighbors (KNN) and Support Vector machines (SVM) as classifiers, which are optimized by scaling the features and using hyperparameters (e.g. K=5-18 of KNN, C=1 and 1 of RBF kernel of SVM). Comparative analysis shows that the optimized KNN and SVM models are 93% accurate, with balanced accuracy (non-purchases >94 percent and recall of purchases of 90-91 percent) compared to baselines including Logistic Regression (92.5%) and Random Forest (93.75%). Interestingly, the segmentation of females and males is based on some gender-stratified understanding where females have a conversion rate of 37.75 percent, whereas males have 33.67 percent, and such segmented understanding is justified because it solves post-COVID revenue problems.The value propositions that the framework offers are its data-driven accuracy, along with a marketing strategy that enables efficient use of resources and higher ROI through audience segments (e.g. high earners in mid-30s). In the example of Yamaha, this translates to precisely targeted advertising, and may end up increasing customer lifetime value within the digital transformation. Limitations are the limited size of the dataset, and the absence of temporal/behavioral information. Moreover, sentiment analysis, influencer details, and ensemble

hybrids (e.g. SVM with gradient boosting) could be applied to bigger and multi-platform data samples to render the results more robust in terms of generalization, and inference of casual effects in dynamic social ecosystems.

6.1 Research Funding

We have not received any external research funding for this

6.2 Acknowledgements

I thank Dr Deepshikha Bhargava, Dr. Anil Ahlawat for guiding me in improving the paper's overall structure.

6.3 Ethics Approval and Consent to Participate

The research is not directly associated with a human subject and no personally identifiable information was gathered. The dataset to be analyzed is publicly accessible and completely anonymized data on the marketing interactions. Thus, there was no need of ethical approval and consent to get involved according to institutional and national requirements. The study does not breach the data privacy and integrity standards of ethics.

6.4 Declaration Of Conflicting Interests

There is no Conflict of interest

Refrences

[1] Kumar R, Prabha V, Kumar M, Rehal P, Samanta P, Singh PK. Influencer Marketing: A Review and Research Agenda Using VOSviewer. Abhigyan. 2025;43:179–97. https://doi.org/10.1177/09702385241289368.

[2] Barari MM, Eisend M, Jain SP. A meta-analysis of the effectiveness of social media influencers: Mechanisms and moderation. J Acad Mark Sci [Internet]. 2025 [cited 2025 Oct 1]; https://doi.org/10.1007/s11747-025-01107-3.

[3] Rabby F, Suryanarayana Murthy Y, Bansal R. Brand evangelism in the digital era: The impact of data-driven influencer marketing. J Open Innov Technol Mark Complex. 2025;11:100552. https://doi.org/10.1016/j.joitmc.2025.100552.

[4] Korniichuk R, Boryczka M. Conversion Rate Prediction Based on Text Readability Analysis of Landing Pages. Entropy. 2021;23:1388. https://doi.org/10.3390/e23111388.

[5] Halder RK, Uddin MN, Uddin MdA, Aryal S, Khraisat A. Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications. J Big Data. 2024;11:113. https://doi.org/10.1186/s40537-024-00973-y.

[6] Barta S, Belanche D, Fernández A, Flavián M. Influencer marketing on TikTok: The effectiveness of humor and followers' hedonic experience. J Retail Consum Serv. 2023;70:103149. https://doi.org/10.1016/j.jretconser.2022.103149.

[7] Rungruangjit W, Mongkol K, Piriyakul I, Charoenpornpanichkul K. The power of human-like virtual-influencer-generated content: Impact on consumers' willingness to follow and purchase intentions. Comput Hum Behav Rep. 2024;16:100523. https://doi.org/10.1016/j.chbr.2024.100523.

[8] Yang Q, Li P, Xu X, Ding Z, Zhou W, Nian Y. A Comparative Study on Enhancing Prediction in Social Network Advertisement through Data Augmentation. 2024 4th Int Conf Mach Learn Intell Syst Eng MLISE [Internet]. 2024 [cited 2025 Oct 1]. p. 214–8. https://doi.org/10.1109/MLISE62164.2024.10674203.

[9] Huang A, Xu R, Chen Y, Guo M. Research on multi-label user classification of social media based on ML-KNN algorithm. Technol Forecast Soc Change. 2023;188:122271. https://doi.org/10.1016/j.techfore.2022.122271.

[10] Shaheen H. Social media marketing research: a bibliometric analysis from Scopus. Future Bus J. 2025;11:41. https://doi.org/10.1186/s43093-025-00465-2.

[11] Gooljar V, Issa T, Hardin-Ramanan S, Abu-Salih B. Sentiment-based predictive models for online purchases in the era of marketing 5.0: a systematic review. J Big Data. 2024;11:107. https://doi.org/10.1186/s40537-024-00947-0.

[12] Sekioka S, Hatano R, Nishiyama H. Market prediction using machine learning based on social media specific features. Artif Life Robot. 2023;28:410–7. https://doi.org/10.1007/s10015-023-00857-z.

[13] Spörl-Wang K, Krause F, Henkel S. Predictors of social media influencer marketing effectiveness: A comprehensive literature review and meta-analysis. J Bus Res. 2025;186:114991. https://doi.org/10.1016/j.jbusres.2024.114991.

[14] Joshi Y, Lim WM, Jagani K, Kumar S. Social media influencer marketing: foundations, trends, and ways forward. Electron Commer Res. 2025;25:1199–253. https://doi.org/10.1007/s10660-023-09719-z.

[15] Chen X, Ding H, Mou J, Zhao Y. Understanding user's identifiability on social media: a supervised machine learning and self-reporting investigation. Data Sci Manag. 2025;8:270–83. https://doi.org/10.1016/j.dsm.2024.12.005.