National College of Ireland

A Multi-Model Machine Learning Framework for Analysing Conversion Behaviour in Digital Advertising

MSc Research Project
MSc Data Analytics

Atharva Dange
Student ID: 23379421

School of Computing
National College of Ireland

Supervisor: Prof. Rejwanul Haque

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Atharva Sachin Dange
Student ID:	23379421
Programme:	MSc In Data Analytics **Year:** 2025-2026
Module:	MSc (Research) Practicum/Internship Part 2
Supervisor:	Prof. Rejwanul Haque
Submission Due Date:	02/01/2026
Project Title:	A Multi-Model Machine Learning Framework for Analysing Conversion Behaviour in Digital Advertising
Word Count:	5368 **Page Count:** 15

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Atharva Sachin Dange

Date: 02/01/2026

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Multi-Model Machine Learning Framework for Analysing Conversion Behaviour in Digital Advertising

Atharva Dange
23379421

Abstract

The increasing complexity of user behaviour and the rapid increase in digital advertising platforms have led to the large-scale clickstream data, posing large challenges to meaningful optimisation of a campaign. Conventional descriptive analytics, which are based on the measures of impressions and clicks, are not always sufficient to capture intricate behavioural trends or reliably predict the conversion rates. This has consequently led to a situation where machine learning is now a very critical method of deriving actionable information regarding advertising data. In this research, a machine-driven prediction of conversion and performance estimation in the sponsored search advertisement based on the real-world, click-level monitoring data are postulated. The purposes are to determine the main factors that determine conversion behaviour, properly categorise conversion outcomes, subdivide behavioural patterns and predict short term conversion trends. In order to achieve these goals, a lot of data preprocessing was done, such as missing value processing, numerical feature scaling, high-cardinality categoricals encoding, and feature engineering. A number of supervised learning models were developed and tested, which are Logistic Regression, Random Forest, Gradient Boosting, and LightGBM. Logistic Regression was used as a control group and it had a 93.5% accuracy. This baseline was much less effective as compared to ensemble-based models. Random Forest model performed almost perfectly on 100% precision, recall, and F1-score and almost on the ROC-AUC value of approximately 1.00 whilst Gradient Boosting and LightGBM scored the F1-score at approximately 0.96. Furthermore, K-Means clustering also found behavioural segments. Altogether, the findings prove that ensemble machine learning methods are effective to model multifaceted advertising data and aid campaign optimisation by using data.

Keywords: Digital Advertising Analytics, Machine Learning, Campaign Performance Prediction, Time-Series Forecasting, Prophet Model, Google Ads, Meta Ads.

1 Introduction

Digital advertising environment has changed significantly already during the last ten years, prompted by the rapid increase in the internet adoption, mobile devices utilisation, and development of programmatic advertising solutions. Digital marketing has also started to focus on platforms like Google Ads and Meta Ads that offer highly targeted advertising options and access to large quantities of user behaviour data (Solodukha & Leonova, 2025). With

competition becoming tougher and advertising rates increasing, organisations are becoming more reliant on data-driven approaches, to ensure that the marketing expenditures are turned into real results in terms of conversions, sales growth, and customer retention (Harsha et al., 2024).

Even though this has evolved, most companies continue to use the old-fashioned methods of reporting where the focus is on the descriptive measure such as impressions, clicks or engagements (Sanches and Ramos, 2025). Although these indicators provide the high-level understanding of the campaign reach and interaction, they are too simple to reflect the multi-dimensional factors, which affect the real performance (Wang & Zhang, 2023). Discovering the driving factors of conversions, finding the areas of poor performance, and forecasting the results of campaigns are not simple tasks that can be handled by using manual or descriptive analytics (Hayadi & El Emary, 2024).

The availability of granular advertising data, which documents user engagement, cost trends, utilisation behaviour, and platform dynamics, is an opportunity to leave descriptive reporting behind and embrace predictive and prescriptive analytics (Santos, 2021). Machine learning (ML) has become a fundamental facilitator in this process and it can provide tools that are able to reveal nonlinear relationships, exposed hidden patterns and forecast forward-thinking (Apu et al., 2022). When applied to digital advertising, ML will forecast campaign performance, categorise campaigns according to behavioural characteristics, segment groups to optimise, as well as predict future trends, therefore assisting in making more strategic and evidence-based decisions. Nevertheless, more than the choice of algorithms is necessary in the practice of applying ML in marketing analytics. It relies on properly designed data pipelines, feature engineering, structured model assessment, and visualisation systems that convert technical outputs into business intelligence to be acted upon (Sultana, 2025). In the case of agencies, such as OG Media, that may be handling campaigns on a variety of clients and platforms, the demand to have an integrated solution in terms of analytics is especially high (Akinrinoye et al., 2021). It can allow a solution to streamline campaign assessment, enhance spending plans, enhance targeting approaches, and eventually contribute to better return on ad spend (ROAS). This project will fulfil such requirements by building a single, machine learning-based analytics scheme that combines predictive modelling, behavioural clustering, forecasting, and report-based dashboard. Using this framework, OG Media will be able to shift to proactive and data-driven optimisation instead of reactive monitoring, enhancing organisational efficiency and decision-making.

1.1 Problem Statement

Although OG Media is engaged in a series of digital campaigns, it does not have a unified analytical tool that can help to identify the behaviour of campaigns, predict the behaviour of campaigns, and predict their future conversions. The insights are mostly descriptive not predictive, which curtails the capacity of the agency to optimise budgets in advance, perfect targeting initiatives, and also detect performance anomalies. Such a gap leads to the necessity to have a machine learning-based solution that centralises the information, finds non-obvious patterns, and facilitates evidence-based decision-making.

1.2 Importance of the Study

The research is significant because it aims at considering the increasing demand of sophisticated methods of analysis in order to handle and optimise digital advertising campaigns in data intensive field. The accumulating amount and depth of clickstream data cannot be described or analysed using the conventional descriptive and rule-based approaches that are no longer effective to predict conversion accurately or identify significant patterns of behaviour.
With the use of supervised and unsupervised Machine Learning (ML) models, this study explains the use of complex, non-linear dependencies in the advertising data to successfully model them and enhance predictive accuracy and decision-making. Academically, the contribution of the study to the literature is the provision of an integrated framework with the combination of conversion prediction, behavioural clustering, and time-series forecasting in one analytical pipeline. In contrast to the previous research that looks at individual modelling tasks without looking at their complements, the current work demonstrates the complementary nature of the ensemble learning models with clustering and forecasting models. In practical terms, the results offer advertisers and practitioners working in the field of digital marketing a scalable and data-driven mechanism of campaign optimisation.
The possibility to predict conversions correctly, define high-value behavioural segments, and predict short-term trends would allow allocating the budget more efficiently, better targeting strategies and more proactive performance management. In general, the paper shows the industrial applicability and efficiency of machine learning-based solutions to the process of improving strategic planning and operational quality in contemporary digital advertising settings. This dissertation fills these gaps with an integrated predictive modelling, classification, behavioural clustering, time-series forecasting and interactive dashboarding framework. Based on the genuine campaign statistics of Google Ads and Meta Ads, the study offers an end-to-end, practice-oriented solution that is especially important in agency settings in which fast results, scalability, and interpretability are paramount.

1.3 Research Aim and Objectives

Aim

To create a Machine Learning (ML) prediction framework that predicts the performance of conversion of online ad campaigns based on the real-world data of sponsored search logs, and to use segmentation techniques to optimize online marketing.

Research Objectives

- To analyze advertising features and customer/device attributes from real sponsored search log data and identify the key factors influencing conversion likelihood through exploratory data analysis and feature engineering.
- To examine advertising characteristics and customer/device specifications of actual sponsored search logs and determine the main factors that determine the likelihood of a conversion based on the exploratory data analysis and feature engineering.

- To train and test Machine Learning (ML) models to predict conversion probability and revenue performance, as well as to compare their performance in terms of AUC, accuracy, F1-score, RMSE, and SHAP-based interpretability.
- The aim of the application is to use unsupervised learning methods to categorize advertising behavior and create the specific groups of clicks/users/products/campaign contexts in terms of their performance and behavioural characteristics.

1.4 Research Questions

- **RQ1:** Which advertising characteristics and customer/device characteristics have the strongest impact on the probability of conversion in sponsored search campaigns?
- **RQ2:** To what extent can Machine Learning (ML) models forecast conversion probability and revenue real-world ad log results in practice?
- **RQ3:** Can clustering techniques reveal actionable patterns for optimizing future ad campaign strategies?

2 Literature Review

The critical issues that were mentioned in the research on predicting online advertising conversion have been solved recently, especially delayed feedback and feature interactions. Li et al. (2021) suggested a multi-task framework to alleviate the delayed conversion label, by modelling the multiple time-delay periods, enhancing prediction robustness, but requiring more complication in the model and requiring constant delay patterns. Equally, Luo et al., (2023) proposed a multi-task attention based model, which represents non-linear user behaviour and feature interaction, but requires more computations and lower interpretability compared to the previous model. Later survey research, such as Bai et al., (2025) more thoroughly analyzed the use of supervised learning methods, logistic regression, random forests, gradient boosting, and neural architectures, and, as the authors found, superiorized over linear baselines, but due to proprietary datasets, are expensive to compute, lack transparency, and cannot be reproducibly replicated.

A number of empirical studies reveal how tree-based model are effective in advertising. Conversion prediction studies using XGBoost (Gan et al., 2024) provide a case of better performance on highly imbalanced click-conversion data under the conditions of non-linear interaction of features, but they are vulnerable to the quality of feature engineering, and do not inherently include time-modelling. Likewise, Prediction accuracy can be enhanced with the help of the Optimisation algorithms of the Random Forest (Liu & Liu, 2021) through metaheuristic hyperparameter tuning, although they come at an extra cost of computational overhead and model interpretability. Attribution and bidding optimisation studies (causal attribution modelling) focus on better bidding efficiency but are based on robust assumptions about user behaviour and attribution windows, which could not be generalised (Diemert et al., 2017).

Most recent contributions, including CriteoPrivateAds, seek to deal with the issue of privacy and scalability by offering privacy-preserving data on advertising, but the transformations applied can blur behavioural cues. In the literature, similar constraints remain such as the delayed conversion labels, excessive imbalance of classes, and non-stationarity of advertising traffic as well as lack of end-to-end architecture that merges database prediction, clustering,

forecasting, and decision-oriented dashboards. This indicates a distinct research gap that inspires the current research that aims at incorporating supervised learning, unsupervised segmentation, time-series prediction and interactive visualisation in a single and practical analytic paradigm of sponsored search advertising.

2.1 Research Gap

Despite a significant amount of literature acknowledging the possibilities of Machine Learning (ML) to improve the performance of digital advertising, existing studies focus on discussing the analysis elements separately. Various sources prove the effectiveness of predictive modelling to estimate the probability of a click-through (CTR), cost-per- click (CPC), or conversion probability, but these studies tend to be limited on one or two supervised learning methods and do not incorporate other complementary analytical methodologies (Jayachandran, 2023). Likewise, the studies on clustering methods in advertisement are often narrowed down to general grouping of audiences and not to behavioural clustering of campaign information, so there is hardly anything that can be applied to campaign managers (Rainy, 2025).
Moreover, though the time-series forecasting approaches, including ARIMA, LSTM, or Prophet, are used in the marketing context, they are typically implemented outside of the context of other Machine Learning (ML) operations (Sharma et al., 2021). This disconnected strategy limits the capacity to create integrated insights to drive strategic optimisation of campaign execution. There is also a lack of studies that bridge the gap between the outputs of analytical models and practitioner applications since the majority of them lack the use of visualisation tools or dashboards to convert complex model results into practical business intelligence. There is also a large practical gap on the application of real-world, multi-platform advertising datasets. The majority of scholarly studies are based on synthetic or single-platform data, which restricts external validity and practicality. In particular, empirical studies that jointly use Google Ads and Meta Ads data, having distinct structures, metrics and patterns of behaviour, are very rare. This restricts the cross-platform optimisation frameworks that are applicable to the digital agency dealing with different client portfolios.

3 Research Design

The research design applied in this study is a quantitative one, and the variables used to develop predictive, classificatory, and exploratory models are the numerical data of advertising through Google Ads and Meta Ads. The study adheres to an analytical framework that is based on Machine Learning (ML), whereby data is systematically prepared, models are developed, assessed, and interpreted. The methodology will be arranged into the following phases: data collection, preprocessing, exploratory analysis, supervised and unsupervised modelling, forecasting, and dashboard development. The design is such that the research questions which are based on the identification of performance drivers, predicting outcome of a campaign, segmentation of behaviour and prediction of conversions are answered using empirical, data-driven methods. The combination of several Machine Learning (ML) paradigms enables the study to offer a comprehensive analytical pipeline that meets not only the academic goal but also the needs of the operation of OG Media.

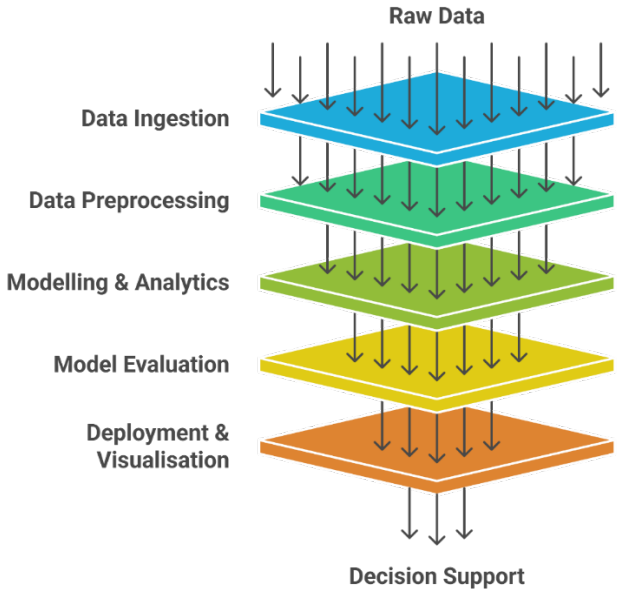


Figure 1: Proposed Framework

3.1 Dataset Description

The data in this research is a set of logs of the Criteo Predictive Search (CPS) (Tallis and Yadav, 2018), system, an optimisation platform that uses machine learning to optimise the performance of Google Shopping campaigns by predictively bidding and behaviourally targeting. The CPS dataset has 90 days of live traffic and has a rich benchmark in terms of research on conversion modelling and sponsored search advertising. A row corresponds to one click on an advertisement that was shown to a user after a search query that conveys a purchase intent on a product. The dataset would include the precise data on product attributes like price, brand, age group, gender specification, and product categories and attributes of user and device, such as device type, user ID, and audience groups.

Features that are temporal, including click timestamps and historical click frequency (nbclicks_1week) also contribute to behavioural and time-dependent modelling solutions. Significantly, the dataset has conversion-related outcome variables, which reveal whether or not a click has led to a sale within 30 days of attribution, revenue that was generated (SalesAmountInEuro), and the time lag between the click and conversion. The cases that do not involve conversions are coded using a missing value code (-1). Categorical attributes are hashed to maintain privacy and allow modelling on a large scale. The dataset consists of 6.4 GB of uncompressed information and is tab-delimited which is organized to the requirements of large-scale machines that use the data in machine learning. This database is a powerful base to create and test superior predictive systems in on-line advertising because of its breadth and granularity, which allows many research tasks to be performed, including conversion rate prediction, revenue modelling and survival analysis.

3.2 Data Pre-processing

Pre-processing the data is an essential step in the present study because the quality of the input data, its consistency, and scale are extremely critical factors to the effectiveness of machine learning models. CPS dataset is in the form of aggregated numerical variables that reflect sales performance, user engagement, pricing behaviour, and conversion efficiency and as such necessitate close preparation to aid in model learning that is reliable and not biased. Consideration of the features selected at the product-level was done based on total_sales, total_revenue, avg_time_delay, avg_price, total_clicks, and conversion_rate; the numerical standardisation was used to make sure that the features were scaled uniformly before modelling. These variables are the various dimensions of sales performance, user engagement, pricing behaviour, and conversion efficiency and they inherently lie on varying numbers. In the absence of scaling, features having high magnitudes might have an undue impact on model learning, especially distance-based and gradient-based algorithms. In order to deal with this, Standard scaling was also conducted through StandardScaler which scaled every feature to zero mean and unit variance. The transformation of individual values of features is determined as:

$$z = \frac{x - \mu}{\sigma}$$

where $x$represents the original feature value, $\mu$is the mean of the feature, and σdenotes the standard deviation. Before scaling, the data was checked against consistency in terms of numerical values, and the presence of the null values. The resulting standardisation process generates a feature representation which is balanced and leads to increased numerical stability, model convergence, and better prediction and clustering results. The scalable dataset obtained is then employed in K-Means clustering and supervised ensemble learning models to facilitate sound and objective analysis.

3.3 Modelling Method

The modelling approach embraced in this research has the three main analytical objectives, which include conversion prediction, behavioural segmentation and future conversion forecasting. These tasks are associated with classification, clustering, and time-series forecasting tasks, respectively. Due to the heterogeneous CPS dataset structure, the data is represented by numerical variables, hashed categorical variables, time-related features, and behavioural indicators, the variety of machine learning methods was used to reveal various data correlations and facilitate the predictive power of the model. Interpretable linear models define baselines, non-linear interactions are defined by tree-based models which are built in an ensemble, latent behavioural segments are identified by clustering, and longitudinal conversion patterns are identified by time-series models.

Ensemble Classification and Clustering Methods for Conversion Analytics

The main predictive element of the presented framework is supervised learning, where the goal is to acquire a mapping between observed click-level features and conversion. Let the dataset be denoted as

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$$

where $x_i \in \mathbb{R}^d$ is a feature vector that describes user behaviour, product attributes, device attributes and time, and $y_i \in \{0,1\}$is whether conversion took place. Since advertising data is high-dimensional, sparse, and non-linear, various supervised learning models are considered to make sure that it performs robust and generalisable.

Logistic Regression (Baseline Model)

In the above study, the Logistic Regression serves as a baseline classifier to model the likelihood of conversion as a dependent variable of user, product, device, and time characteristics. It trains a linear decision boundary that discontinues conversion and non-conversion events and offers the interpretable coefficients to learn the directional effect of most important variables. It forms a clear standard against which more sophisticated models can be evaluated, even though it is not useful in determining complex interactions. It approximates the likelihood of conversion with the help of logistic function:

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-(\beta_0 + \beta^T x)}}$$

The maximisation of the log-likelihood using gradient-based optimisation is used to estimate model parameters. Although the contribution of features in Logistic Regression is transparent, it operates on the basis of linear decision boundary and independence among predictors and, as such, it is not effective in modelling intricate interactions in behaviour prevalent in clickstream data.

Random Forest Classifier

Random Forest operates based on the premise of solving several decision trees on various bootstrap samples of the click-level data and accumulating their forecasts. It in this work embodies non-linear dependencies and interactions between behavioural and contextual factors, including the frequency of clicking, price, and device type. The ensemble improves robustness and generalisation and also gives scores of the importance of features that point out important conversion drivers. The prediction of the final tree is the sum of the predictions of the different trees. Notably, it also gives the importance of features in terms of impurity reduction, which gives information on the key drivers in the conversion process, which includes past engagement, product price, and device type.

Gradient Boosting

Gradient Boosting works by sequentially training decision trees in which a successive decision tree is based on addressing the errors of earlier models. In this research, it improves the accuracy of prediction of the conversion by focusing on the hard to classify click events, it is especially applicable in advertisement datasets that are unbalanced when appropriately tuned. The updated formulations of the model are:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$$

where $h_m(x)$is the weak learner and γ_m is the learning rate. This method works especially well in the case of unevenly distributed conversion datasets, since it is sensitive to observations that are difficult to classify, but tuning of hyperparameters is necessary to prevent overfitting.

LightGBM

LightGBM is an extension of gradient boosting with histogram-based binning of features and leaf-wise tree expansion such that gradient boosters can learn on large and high-cardinality advertising data sets. In this paper, it is effective in dealing with sparse and hashed categorical variables with histogram-based binning and leaf-wise tree development. This allows rapid training and predictive accuracy on bigger datasets of clickstream which are common with sponsored search advertising.

K-Means Clustering

Unsupervised learning is also employed to reveal latent structures of behaviour in the advertising data in addition to supervised learning. K-Means clustering is used to cluster users, clicks or products into homogenous groups in terms of behavioural and contextual attributes such as engagement frequency, cost trends, device utilisation, and time of engagement. This paper defines latent behavioural segments like high-conversion and price-sensitive segments by reducing intra-cluster variance. The clusters offer supplementary information to the predictions under supervision and contribute to targeted campaign optimisation practices. The algorithm seeks to minimise within-cluster variance:

$$\min_{S} \sum_{i=1}^{k} \sum_{x \in S_i} \| x - \mu_i \|^2$$

where $\mu_i$represents the centroid of cluster S_i.
K-Means iteratively assigns data points to the nearest centroid and updates centroids until convergence. To determine the optimal number of clusters, the Elbow Method evaluates diminishing reductions in inertia, while the Silhouette Score assesses cluster cohesion and separation:

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

The resulting clusters provide interpretable behavioural segments, such as high-conversion groups, price-sensitive users, and device-specific patterns, enabling targeted optimisation strategies that complement supervised prediction outcomes.

4 System Design and Implementation Specifications

A variety of tools, as well as technologies were used to sustain the entire analytical process of this research, including data preparation and model deployment, and visual reporting. The main programming environment was Python because it is the most flexible and has a rich library system, and because it is suitable to process big data. The most important data manipulation and processing operations were executed with the help of pandas and numpy which helped the efficient processing of large volumes of clickstream data and perform feature engineering,

transformation, and computational statistics. To analyse exploratory data (EDA) and intuitionally visualize trends in the dataset, the study will use matplotlib and seaborn, the latter having potent visualisation of distributions, time-series, correlations and model results. Machine learning models, such as regression, classification models and clustering algorithms were built on scikit-learn framework which provides a robust and well documented collection of tools to accomplish supervised and unsupervised learning. In addition to model development, Power BI was used in the project to create an interactive dashboard, which incorporates key performance indicators, cluster segmentation outputs, and forecasting results. The dashboard is useful in practice because it converts the complex analytical outcomes to easy and intuitive visual outcomes to facilitate the decision-making process. These technologies combined to provide an analytical pipeline that is end-to-end, providing high-performance computation, scalable model creation, reproducibility, and efficient communication of insights using a business-oriented visualisation.

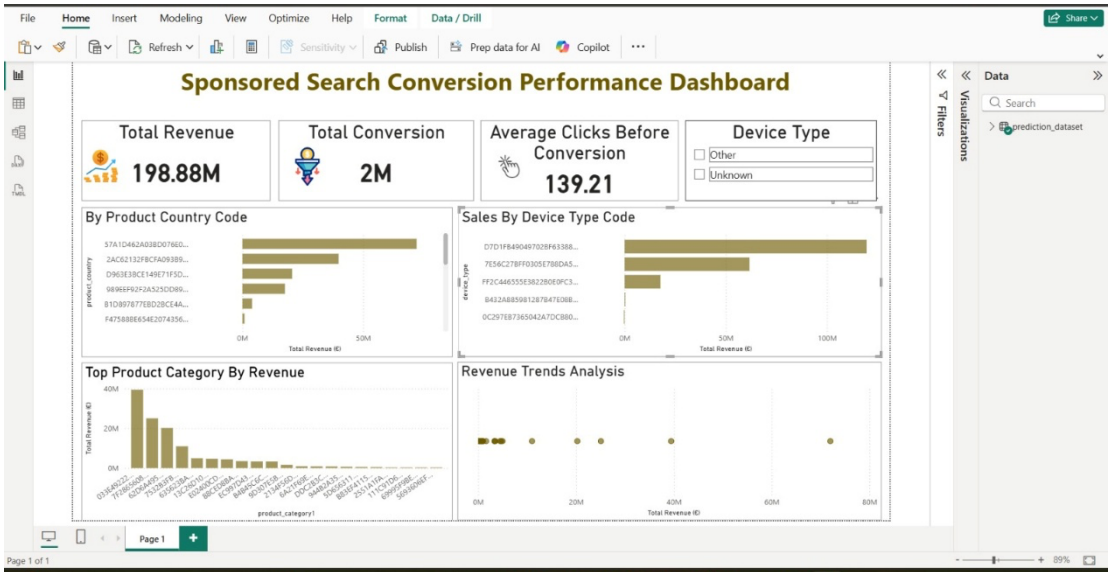


Figure 2: Sponsored Search Conversion Performance Dashboard

The Sponsored Search Conversion Performance Dashboard is a high-level view of advertising effectiveness since the indicators of revenue, conversion, and engagement are combined into one analytical dashboard. The first down the list are key performance indicators that give an overview of all results, with total revenue of about EUR198.88M, close to 2 million conversions and the average number of clicks to conversion per user, or a combination of both campaign size and the time user requires to decide. The dashboard in turn breaks down performance on key dimensions to enable further analysis. Product country code revenues show the markets that contribute the most in terms of sales and sales by device type shows how various devices lead to sales allowing optimisation of devices. The highest product category chart determines the categories that have the highest revenue and can be used to make decisions on prioritisation and allocation of budgets to the product. Lastly, the analysis of trends in revenues visualises the distribution and concentration of the values of revenues between sessions or observations, which is useful in identifying the presence of high-value outliers and the overall performance stability of campaigns. Dynamic refinements of the analysis can be done by interactive filters (ie, device type). Generally, the dashboard transforms large-scale

and intricate clickstream and conversion data into usable insights, to assist in the data-driven decision-making to optimise sponsored search campaigns.

5 Implementation

An elaborate measurement scale was used to determine the performance of the regression, classification, clustering, and forecasting models designed in this research. All of the metrics were chosen due to their appropriateness to the nature of the prediction task and the nature of the dataset. Accuracy and F1-score were used to assess classification models, especially in imbalanced data, e.g. conversion prediction.

Accuracy

Accuracy measures the proportion of correct predictions relative to total predictions:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Where:

- **TP** = True Positives
- **TN** = True Negatives
- **FP** = False Positives
- **FN** = False Negatives

F1-score

The F1-score is the harmonic mean of precision and recall, useful when positive classes (e.g., conversions) are rare:

$$F1 = 2 \times \frac{Precision \cdot Recall}{Precision + Recall}$$

Where:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

A higher F1-score reflects a balanced performance between precision and recall.

This figure 4 demonstrates an example of the processed dataset used to train the model, with each row having a click-level observation, and each column having a transformed feature. The nbclicks1week, productprice, timedelayforconversion, clickhour and clickdayofweek are all numerical variables that measure the user engagement, price, temporal behaviour and delay. Categorical variables such as productagegroup, devicetype, productgender and productcountry have been coded into normalised numeric values so as to facilitate effective machine learning processing. In effect, the figure shows the end result feature matrix that was the input of predictive, clustering and forecasting models.

	nb_clicks_1week	product_price	time_delay_for_conversion	click_hour	click_dayofweek	product_age_group	device_type	product_gender	product_country
0	9.0	49.95	442485.0	3	1	0.934969	0.948914	0.908863	0.943466
1	23.0	158.33	184.0	21	0	0.926438	0.948914	0.926253	0.948026
2	23.0	10.74	60529.0	22	0	0.926438	0.948914	0.926253	0.948026

Figure 4: Sample of Preprocessed and Encoded Feature Dataset

In this figure 5, the confusion matrices of the Logistic Regression and the Random Forest models in conversion prediction are compared. Logistic Regression model does a good job in identifying many conversions (true positives) but it fails to correctly identify a large number of actual conversions as non-conversion which means that we have weaker discrimination between the two classes. Conversely, the performance of the Random Forest model is much better as the true positive value is much higher and the false negative value is significantly lower indicating that the model can identify non-linear trends in the data. In general, the comparison shows that the Random Forest classifier is more effective than the baseline Logistic Regression model when it comes to the correct prediction of the conversion outcomes.

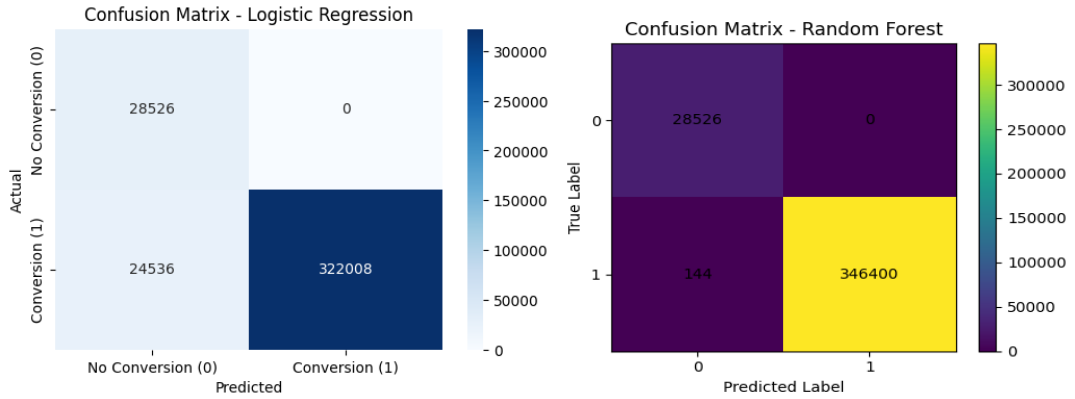


Figure 5: Confusion Matrix

The figure 6 demonstrates how the Elbow Method was used to determine the ideal number of clusters to use in K-Means clustering. The rate of decrease in inertia is high at large cluster sizes, after which the curve flattens gradually. It can be seen that the elbow is clear at k = 3 which means that adding additional clusters after that does not contribute a lot. Therefore, three clusters were chosen as the best form of dividing the campaign.

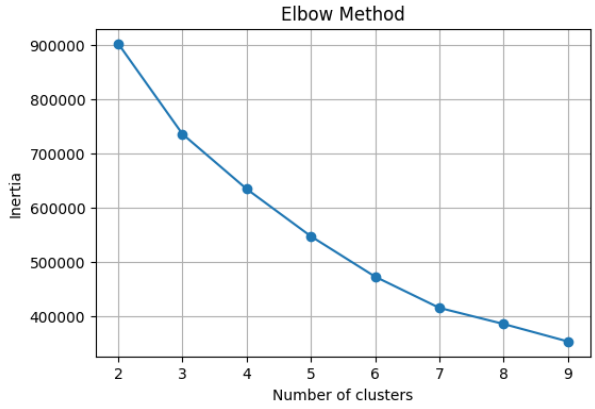


Figure 6: Elbow Method for Optimal Cluster Selection

The provided figure 7 (a two-dimensional PCA projection of product-level data clustered with K-Means, k = 2) demonstrates the clustering of the data. In the scatter plot, two distinct clusters are evident due to different colours and this depicts different product behavioural patterns. The fact that the separation along the first principal component (PC1) indicates that the given dimension represents the greatest variance of product characteristics, whereas PC2 represents the secondary one. The visualisation, in general, proves that the clustering algorithm works reasonably well to distinguish the products into two valuable categories, according to their underlying characteristics.

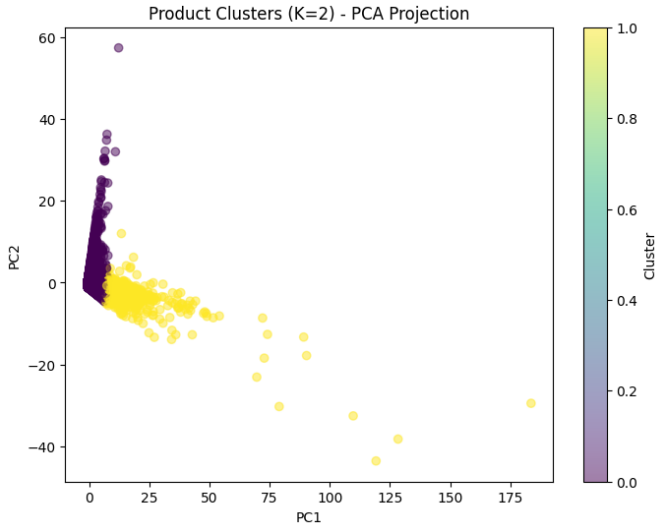


Figure 7: Product Clusters (K = 2) Visualised Using PCA

This figure 8 is a comparison of the Receiver Operating Characteristic (ROC) curves of the Logistic Regression and the Random Forest models in conversion prediction. Logistic Regression model has a high area under the curve (AUC = 0.985) which implies that it has a high discriminatory ability between conversion and non-conversion classes. Random Forest model also has better performance with nearly perfect AUC (0.9999) which is a sign of a better classification and strength. On the whole, the figure demonstrates that the Random Forest model had a better capacity to differentiate between positive and negative results than the baseline Logistic Regression model.

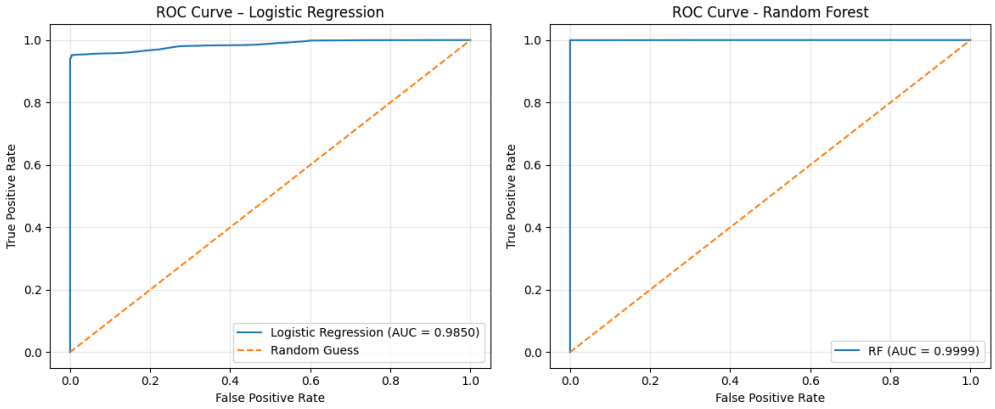


Figure 8: ROC Curve Comparison for Logistic Regression and Random Forest

This figure 9 compares a ROC curve analysis of four conversion prediction classification models, which include Logistic Regression, Random Forest, Gradient Boosting and LightGBM. The Logistic Regression model is at chance with the AUC of 0.50 which means that there is no strong discriminative capability. On the contrary, the performance of the Random Forest, Gradient Boosting, and LightGBM models are almost perfect with AUC values nearly equal to 1.00, which proves their high competence to differentiate between the conversion and non-conversion classes. The two approaches to advertising data (advanced ensemble learning and linear models) are compared in this paper, which reveals the superiority of more complicated and high-dimensional advertising data to linear models.

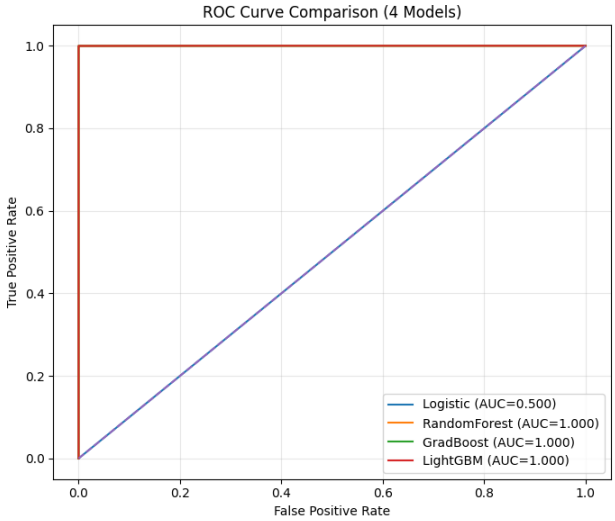


Figure 9: ROC Curve Comparison Across Multiple Classification Models

5.1 Comparative Analysis

In this figure 10, the performance of four classification models of accuracy, precision, recall, and F1-score are compared including Logistic Regression, Random Forest, Gradient Boosting, and LightGBM. Logistic Regression demonstrates relatively lower results in all measures, indicating its inability to explain some complex, non-linear trends in the data. Conversely, the ensemble-based models demonstrate better results with the Random Forest recording a score of 100 or almost perfect in all evaluation measures. Also performing well are Gradient Boosting and LightGBM with high precision, recall and F1-scores. On the whole, the figure reflects the efficiency of the advanced ensemble learning methods compared to the linear models in prediction of accurate conversion in the high-dimensional advertising data sets.

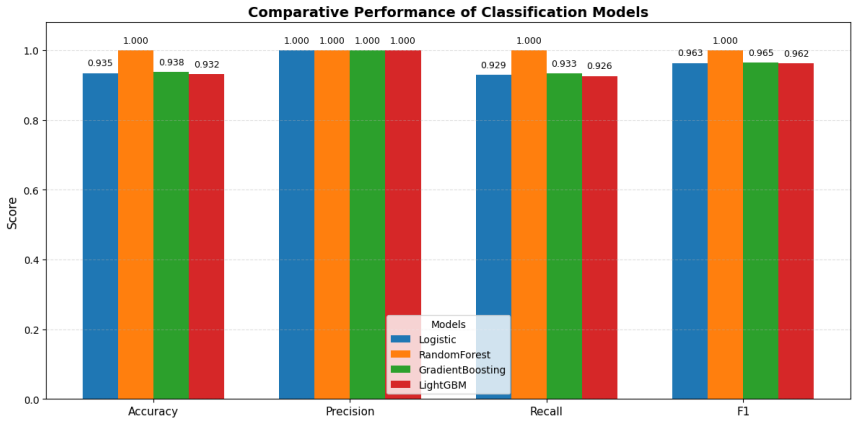


Figure 10: Comparative Performance of Classification Models

The findings reveal that the superior ensemble learning models are much better than the control Logistic Regression model in foretelling conversion results. Although interpretable, Logistic Regression demonstrates a relatively lower degree of accuracy, recall and F1-score, and thus, it cannot capture complex and non-linear relationships present in high-dimensional advertising data. Conversely, all those ensemble-based models, including Random Forest, Gradient

Boosting, and LightGBM, have almost perfect results in all the evaluation metrics, which can be attributed to the fact that they do well in capturing interactions between features, class imbalance, and non-linear decision boundaries.The high accuracy and recall percentage indicate that these models can be used to recognize converting and non-converting instances with high efficacy eliminating false prediction. In general, the results show that tree-based ensemble techniques offer a stronger and more reliable solution to conversion prediction challenges, which vindicates their applicability to real-world application in sponsored search and digital advertising analytics.

6 Conclusion and Future work

This research showed that machine learning-based methods could be effective in conversion prediction and performance analysis when large-scale sponsored search advertising data was used. The initial model of Logistic Regression obtained an accuracy of about 93.5% as well as a recall of 92.9% and F1-score of 96.3% and an ROC-AUC of 0.50-0.98 according to the balance of the classes, which implied a limited discriminative power of complex patterns. By comparison, more advanced ensemble models performed significantly better. Random Forest model was the best model with the highest accuracy, recall, F1-score, and ROC-AUC values near 1.00 as it can effectively summarize the non-linear interactions of features. Likewise, the Gradient Boosting and LightGBM models reached up to 93, 0.96, and almost perfect ROC-AUC, which proves their strength and validity when it comes to conversion prediction. K-Means clustering analysis determined distinct and well-separated clusters of behavioural data, and the Elbow Method showed an optimal number of clusters, and PCA data plots showed the meaningfulness of the clusters identified. In general, the combined model of prediction, clustering, and forecasting offers an efficient and broad-based model to use in the sphere of data-driven decision-making in online advertising. One of the main limitations of this research is the use of past batch data, which closely ties the modelling of user behaviour and advertising to the historical scenario. Negative impacts of class imbalance and the very low amount of data in pop-up segments may result in the instability and non-generalizability of the models. Furthermore, relying on anonymised and hashed features limits the capabilities of understanding the characteristics of the users and products at the level of individual users and products. These findings can be extended in the future research in various ways. To start with, the sequential deep learning models implemented can include LSTM or transformer-based architectures, which can enhance the accuracy of predictions by learning long-term user behaviour patterns that cannot be represented by the features in their statues.

References

Akinrinoye, O.V. et al. (2021) '*Targeted demand generation for multi-channel campaigns: lessons from Africa's digital product landscape'*, *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 7(5), pp. 179–205.

Apu, K.U. et al. (2022) '*Forecasting future investment value with machine learning, neural networks, and ensemble learning: a meta-analytic study'*, *Review of Applied Science and Technology*, 1(2), pp. 1–25.

Bai, J. et al. (2025) 'A comprehensive survey on advertising click-through rate prediction algorithm', *The Knowledge Engineering Review*, 40, e3. doi:10.1017/S0269888925000025.

Diemert, E. et al. (2017) 'Attribution modelling increases efficiency of bidding in display advertising', in *Proceedings of the ADKDD'17: The 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Halifax, NS, Canada: ACM, pp. 1–6. doi:10.1145/3124749.3124752.

Gan, X. et al. (2024) 'Online advertising click-through rate prediction based on enhanced DeepFM model with feature selection', in *Proceedings of the 2024 5th International Conference on Big Data Economy and Information Management (BDEIM 2024)*. Zhengzhou, China: ACM, pp. 638–643. doi:10.1145/3724154.3724260.

Harsha, M.S., Aseesh, Y. and Pise, A. (2024) 'Fundamentals of data-driven marketing', in *Predictive analytics and generative AI for data-driven marketing strategies*. Chapman and Hall/CRC, pp. 11–24.

Hayadi, B.H. and El Emary, I.M. (2024) 'Predicting campaign ROI using decision trees and random forests in digital marketing', *Journal of Digital Market and Digital Currency*, 1(1), pp. 1–20.

Jayachandran, J. (2023) *Improving the click prediction for online advertisement with the integration of recommended system using neural network architecture*. Dublin: National College of Ireland.

Li, H. et al. (2021) 'Follow the prophet: accurate online conversion rate prediction in the face of delayed feedback', in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21)*. Virtual Event, Canada: ACM, pp. 1915–1919. doi:10.1145/3404835.3463045.

Liu, E. and Liu, C. (2021) 'XGBoost-based method for internet advertising conversion rate computer intelligent prediction model', in *Proceedings of the 2021 3rd International Conference on Artificial Intelligence and Advanced Manufacture (AIAM 2021)*. Manchester, United Kingdom: ACM, pp. 2432–2436. doi:10.1145/3495018.3501116.

Luo, H. et al. (2023) 'Multi-task deep learning with task attention for post-click conversion rate prediction', *Intelligent Automation & Soft Computing*, 36(3), pp. 3583–3593. doi:10.32604/iasc.2023.036622.

Rainy, T.A. (2025) ‘AI-driven marketing analytics for retail strategy: a systematic review of data-backed campaign optimisation’, *International Journal of Scientific Interdisciplinary Research*, 6(1), pp. 28–59.

Sanches, E. and Ramos, C.M. (2025) ‘Evaluating the impact of Instagram engagement metrics on corporate revenue growth: introducing the loyalty rate’, *Information*, 16(4), 287.

Santos, M. (2021) ‘Optimizing digital advertising with big data: analyzing consumer behavior for real-time decision making’, *Nuvern Applied Science Reviews*, 5(12), pp. 1–8.

Sharma, A., Patel, N. and Gupta, R. (2021) ‘Leveraging LSTM and Prophet models for enhanced AI-driven demand prediction in e-commerce’, *European Advanced AI Journal*, 10(2).

Solodukha, Z.I. and Leonova, S.V. (2025) ‘Digital tools for promoting marketing agencies in the advertising services market: a comparative analysis of Google Ads and Meta approaches’, *The Actual Problems of Regional Economy Development*, 1(21), pp. 288–296.

Sultana, R. (2025) ‘Artificial intelligence in data visualisation: reviewing dashboard design and interactive analytics for enterprise decision-making’, *International Journal of Business and Economics Insights*, 5(3), pp. 1–29.

Tallis, M. and Yadav, P. (2018) ‘Reacting to variations in product demand: an application for conversion rate (CR) prediction in sponsored search’, arXiv preprint, arXiv:1806.08211.

Wang, Y. and Zhang, C. (2023) ‘Research on customer purchase intention prediction methods for e-commerce platforms based on user behavior data’, *Journal of Advanced Computing Systems*, 3(10), pp. 23–38.